REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12/28/09	Final Technical Report	9/30/07 - 9/30/09

4. TITLE AND SUBTITLE

The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Collaboration

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-07-1-1152

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Opava,S; Abney,K; Barjami,S; Bellardo,J; Bekey,G; Black,M; Cardinal,T; Clark,C; Clague,D; Derickson,D; Fernsler,J; Frame,S; Gillen,K; Griffith,E; Hall,G; Hazelwood,S; Jin,X; Lehr,C; Lundquist,T; Lin,P; McDonald,R; Mitra,N; Moss,R; Niku,S; O'Halloran-Cardinal,K; Self,B.; Sharpe,J; Sungar,N; Pohl,J; Rahman,S; Saunders,K; Schwartz,P; Szlavik,R; Tomanck,L; Vorst, K; Yu,X-H; miscellaneous student authors.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Cal Poly Corporation, Sponsored Programs Office
Bldg. 38, Rm. 102
San Luis Obispo, CA 93407-0830

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Clifford W. Anderson
Office of Naval Research
875 Randolph St.
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Available for public release

13. SUPPLEMENTARY NOTES

20091231175

14. ABSTRACT

The primary purpose of this project is to carry out applied research and development projects and build research capacity in areas of interest to the Department of Defense and the Office of Naval Research. Research areas include communications, computing, command and control, sensors, coastal monitoring, force protection and performance, bio- and chemical-hazard detection and mitigation, vulnerability assessment, new materials and devices, data acquisition, optical and radar imaging, autonomous vehicles and robots, alternative energy sources and energy efficiency.

15. SUBJECT TERMS

Liquid crystals, electrooptics, AUVs, UAVs, robotics, situational awareness, LIDAR, SGDBR lasers, path optimization, biofuels, compliant-core materials, tissue engineering, injury repair, bone mass, bio/chemical sensors, bio-MEMS devices, LEDs, solar cells, photonic lattices, knowledge management, computational electromagnetics, satellite constellations, denial of service attacks, ethics.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Susan C. Opava, Ph.D.
U	U	U	UU		19b. TELEPHONE NUMBER *(Include area code)* 805-756-1508

The California Central Coast Research Partnership: Building Relationships, Partnerships, and Paradigms for University-Industry Research Collaboration

FINAL REPORT
ONR GRANT NO. N00014-07-1-1152
September 30, 2007 to September 30, 2009

Principal Investigator:

Susan Opava, Ph.D.
Dean of Research and Graduate Programs
California Polytechnic State University
San Luis Obispo, CA

December 18, 2009

Table of Contents

I. Title of Project and Principle Investigator	1
	1
II. Summary of Project	1
III. Relevance to ONR Objectives	1
A. Relevant Partners	1
B. Relevant R&D Focus	2
C. University-Industry-Government Partnership	3
D. University Strengths	3
IV. Summary of results During the Period of Performance	5
A. General	5
B. Development of New Research Capacity	7
1. Instrumentation	7
2. Infrastructure	9
C. Detailed Research Reports	9
1. Collaborative Agent Design Research Center (CADRC)	10
Knowledge Management Project	11
2. Other Research Projects	
Quenched Random Disorder in Soft Condensed Matter	24
Communication within Satellite Constellations	41
Construction of an Efficient Microbial Peptidase Delivery System to Treat Celiac Disease and Maximize Human Health	51
Building a laboratory to investigate injury –repair in skeletal muscle and its vasculature	66
Single Cell Impedance Sensing for Pathogen	78
Multi-AUV Path Optimization for improved Ocean Model Forecasting	87
LIDAR Applications Enabled by Fast Wavelength-Tuning Single-Chip Wavelength Tunable SGDBR Lasers	93
An experimental and theoretical study of the electro-optical response of a new type of ferroelectric liquid crystal	121
Feature Selection and Boosted Classification Algorithms for Pedestrian Detection	134
Design and construction of magneto-optical trap for experimental investigation of atomic dipole traps for quantum computing	142
Upgrading a fleet of GPS-tracked ocean surface drifters for improved performance and extended coverage area	148
Shock and Vibration Hardened Data Acquisition Device	158
Bone Mass Preservation and Fracture Risk Assessment with Bisphosphonate Therapy during Space Flight	183
Investigation of Photonic Lattice Based Gallium-Nitride Light Emitters	196

Algae Lipid Charaeterization and Extraction	256
Autonomous Military Roboties: Risk, Ethies, and Design	269
Unmanned Aireraft Systems; Situational Awareness for Small Taetieal Unmanned Aireraft and an Autonomous Paekage Delivery Coneept Study for the US Marines	385
Sandwieh Composite Report	391
U.S. –China Collaborative Soil-Strueture –Interaetion Researeh	421
Implementation and Evaluation of Physiologie Conditions for a High Throughput "Blood Vessel Mimie" Model System	430
Vulnerability Assessment of Water Distribution Networks due to Insuffieient Fire Flows	441
A Portable New Chemieal/Biologieal Sensor	452
Solar Transportation: Sunlight to Eleetrieity to Motion	457
Galvanie Vestibular Stimulation Applied to Flight Training	469
Optieal forees and the effeets of partiele proximity in optical trapping	479
Impaet of the Neural Toxin Paraoxon on the Eleetrophysiology of the Neuromuseular Junetion	487
Environmental Proteomies: the Minimal Stress Proteome in the Marine Model Organisms Ciona Intestinalis and C. Savignyi-Networks of Co-Expressed Proteins	492
Adaptation of the Bardo Airway to the Intraoral Mask: Innovative Airway Management Devices Working in Coneert	499
V. Appendix	
A. University of Washington Consultant Biographies	567
B. Project Related Thesis and Relevant Publications	570

I. Title of Project and Principal Investigator

The California Central Coast Research Partnership: Building Relationships, Partnerships and Paradigms for University-Industry Research Collaboration; Susan C. Opava, Ph.D.

II. Summary of Project

The mission of the California Central Coast Research Partnership (C^3RP) is to facilitate the exchange of technical knowledge and skills between the higher education sector and the private sector in San Luis Obispo County, and to encourage the growth of high-tech companies in the region, thereby enhancing economic development and quality of life. Since its inception, the project has focused on technologies of relevance to the Department of Defense. The partnership is a long-term plan to create a dynamic and self-supporting university-industry-government partnership that capitalizes on the strengths and mutual interests of the educational and technology-based business sectors. The plan recognizes the key role of higher education in preparing a highly skilled work force and transferring new knowledge to practical uses. The outcomes of this partnership, when fully realized, will be the creation of a robust and self-sustaining base of University R&D activities; the development of existing technology-based businesses and the creation of new ones; and the generation of opportunities for job training and research and development activities for University and Community College students and faculty.

The project also includes the construction (with non-DOD funding) of a technology park on the California Polytechnic State University campus, which will provide state-of-the-art space for private technology companies engaged in research and development activities, as well as a business incubator that will provide all of the support services needed by start-up, technology-based companies. The aspect of the program supported by this ONR grant is the continued development of a strong base of applied research at Cal Poly, through university-government-industry partnerships designed to optimize the application of the strengths of each of these sectors to problems of mutual interest. The management team, operational since January '02, continues to lead the project and develop the collaborative relationships between the educational and private sectors that are essential to realizing long-term goals and securing the financial base that will allow full-scale project development.

III. Relevance to ONR Objectives

A. Relevant partners.

C^3RP represents a coalition of educational institutions, local, state and federal government, and private businesses that have worked together in unprecedented fashion to advance the common goals inherent in the proposed university-industry partnership. The current partners in the project and their contributions include:

California Polytechnic State University

- committed the land for the first building in the Cal Poly Technology Park construction project, valued at ~$1.5 million
- provided assistance in financial management of the project
- contributed $90,000 for a pre-feasibility study by Bechtel Corporation
- committed several hundred thousand dollars of in-kind contributions of senior management time and effort over several years and continues to do so
- invested ~$1,000,000 in efforts to raise additional funds for the project; secured sufficient private and other federal funding to construct the first building in the Cal Poly Technology Park

CENIC (Corporation for Educational Network Initiatives in California; association of Internet2 universities in CA)

- works with Cal Poly to provide high-bandwidth internet access to support C^3RP research projects

City of San Luis Obispo

- in partnership with Cal Poly developed a carrier-neutral fiberoptic ring around the city that benefits both Cal Poly and technology-based businesses

Housing and Urban Development

- provided funds toward construction of the pilot technology park building.

Economic Development Administration

- has provided funds toward construction of the pilot technology park building.

Efforts are ongoing to secure new partners, including:

- Major corporations
- Small technology-based businesses

B. Relevant R&D focus.

The research programs that were supported are relevant to seven of the eight "thrust areas" of ONR's Code 30 Science and Technology Program. The projects involved basic research in these areas, as well as applied research and development leading to more immediate technological applications. The seven areas of relevance and the more specific focus areas to which the research contributed are listed below:

Command and Control, Computers, Communication: situational awareness; communications; knowledge management; computational electromagnetics; autonomous systems; LIDAR; data acquisition; satellite constellations; reconnaissance; denial of service attacks.

Force Protection: post-impact/explosive force stability assessment; bio- and chemical-hazard detection and mitigation; risk/vulnerability assessment; new materials; collision avoidance; autonomous robots.

Mine Countermeasures: coastal monitoring; IED detection; situational awareness.

Human Performance, Training and Survivability: cognitive performance enhancement; physical performance enhancement; smart materials; sensors; biological stress reactions; biomarkers; injury repair; improved materials and processes for use on military bases and in the field.

Intelligence, Surveillance and Reconnaissance: data acquisition; sensors; satellite constellations; autonomous vehicles; optical and radar imaging.

Logistics: alternative energy sources; new materials.

Maneuver: advanced design and materials for vehicles (land and water).

C. University-industry-government partnership.

The primary focus of this long-term initiative is to forge a strong link between private sector R&D and University applied research to speed the development of new knowledge and the transfer of technology to the public and private sectors. San Luis Obispo has become a draw for technology businesses (with a heavy concentration of software development companies) from both the LA Basin and Silicon Valley. For example, SRI (Stanford Research Institute), International operates a "software center of excellence" in the city. Lockheed-Martin has a research and development group in nearby Santa Maria. Two local companies manufactured critical components for the Mars rovers, and other companies, e.g. California Fine Wire, Aeromech, and CDM Technologies are suppliers to the military. Also located on the Central Coast are branches of two major biotechnology companies: Promega Biosciences and Santa Cruz Biotechnology.

D. University strengths.

Cal Poly is a State university that has achieved national distinction as a polytechnic university, with engineering and computer science programs ranked among the very best undergraduate programs in the country. Its strengths have led it to orchestrate the research partnership effort and the consortium of partners described herein. Cal Poly also has affiliations with CSA (California Space Alliance) and with Vandenberg Air Force Base, where it has offered an M.S. in Aerospace Engineering by distance learning. Cal Poly has the capability to offer many more academic programs by distance learning to remote locations. In particular, through possible collaborative agreements at cable-head locations around the world (including Asia and Europe) our programs can be made available to military personnel stationed almost anywhere in the world.

Over the past several years, the University's faculty has been turning over at a rate of about 10% a year. The University has responded to this opportunity by hiring research-oriented faculty and promoting applied research and development. With as many as 50-60 new faculty hires per year in the past six years, the University is positioned to undertake significant R&D projects for government and industry. C^3RP has provided needed support and infrastructure for these faculty, which has enabled them to develop ongoing research programs and secure >$7 of competitive funding for each $1 of C^3RP funding invested in them.

Cal Poly also has a highly qualified student body with entering credentials comparable to students who attend the highest ranked campuses of the University of California. Our students gain valuable experience working with faculty on externally sponsored research projects.

A hallmark of Cal Poly is its extensive network of industry partners. The President's Cabinet consists of more than 30 major corporate and business leaders. Each college, and each department within the college, has its own industrial advisory board. Until recently these connections were not exploited to attract industry-sponsored R&D to campus; hence, one of the goals of the C^3RP partnership is to use these existing relationships with industry to garner support for our R&D efforts, wherever possible in a three-way partnership with government entities, so as to leverage the assets of each partner for the benefit of all. Cal Poly's College of Agriculture, Food and Environmental Science has successfully demonstrated this kind of partnership through its Agricultural Research Initiative. Through this initiative, a consortium of four campuses in the CSU garnered $5 million a year in on-going funding from the State of California to support agricultural research of interest to the State, with a pledge to raise matching funds from industry. This State and private funding has leveraged additional support from the federal government. Similar new CSU-wide initiatives include the Council on Ocean Affairs, Science and Technology (http://www.calstate.edu/coast) and the Water Resources and Policy Institute.

As will be seen in the remainder of this report, Cal Poly has extraordinary interdisciplinary technical assets that can be brought to bear on the science and technology issues of importance to ONR.

In summary, the California Central Coast Research Partnership has taken advantage of a confluence of factors, including existing and potential relationships, fortuitous technological and economic developments in the region, the particular strengths and expertise of the Collaborative Agent Design Research Center at Cal Poly, and a meshing of the research and development interests of the University, the Office of Naval Research, and the private sector. C^3RP is the vehicle for fully realizing the benefits of the common goals and synergies of the partners and their respective resources.

IV. Summary of Results During the Period of Performance

A. General.

The C^3RP program was originally funded through an award from ONR in FY '02. This report covers an award that began on 9/30/2007 and ended on 9/30/2009. General accomplishments are summarized below. Detailed reports are presented later in the document.

An overview of accomplishments during this project period follows:

- Research carried out by the **CADRC (Cooperative Agent Design Research Center)**, of particular interest to ONR and the Marine Corps, was again funded. A detailed report on this project is provided in Section IV.C.1 of this report.

- **New research** has been developed and some research has been continued, including some with industry collaboration. Projects address topics highly relevant to defense and national security, such as data acquisition, imaging and analysis, energy efficiency, communications, command and control, reconnaissance, autonomous vehicles, bio/chemical sensors, robotics, and risk/vulnerability assessment. Detailed reports of the results of these projects are presented in Section IV.C.2 of this report.

- Since January 1, 2003, C^3RP-supported faculty have **received ~$45 million ($44,987,351) in competitive funding from other sources**.

- **New research capacity was developed,** including new instrumentation and enhanced infrastructure (detailed below in Section IV.B).

- Funds were also provided to support small **student research projects** through collaboration with Cal Poly's Honors Program. Talented, high-achieving students in Cal Poly's selective Honors Program were given the opportunity to work on research projects with a faculty member for 1-2 academic quarters and to present their results at a campus symposium at the end of the academic year. C^3RP-supported project reports are presented in Section IV.C.3 of this report.

- Two **consultants** from the University of Washington spent two days on campus meeting with various faculty and academic administrators to discuss our **research and development and technology transfer capability.** The bios of the two consultants are appended (Appendix A).

- **Information technology infrastructure support** was provided. **Internet2** connectivity was initially acquired for the campus in November 2001, to support current and future research efforts. Internet2 membership and connectivity has continued during this grant period.

- A **database of >70 technology–based companies** that are potential partners in the project and potential research collaborators has been updated. A series of on-campus research forums initiated last year was continued. Companies were invited to campus two times during the last year to learn about specific University research projects and identify potential areas for collaboration. Several collaborative relationships have developed.

- The first **research and development company to be located on campus** in anticipation of the construction of the pilot building for the technology park continues to flourish in the campus environment and has developed research collaborations with faculty in several different disciplines and colleges. These have resulted in two federal research grants (USDA and NIH) in the areas of alternative fuels from biomass and oral vaccines, respectively. The company, Applied Biotechnology, Inc., specializes in the use of genetically modified plants to produce non-food products, for example, industrial enzymes, biochemical reagents and oral vaccines. The presence of the company has spurred faculty to develop research in this area and a specialized research greenhouse supports this developing work. A second company has been recruited to campus, specializing in object-oriented software development.

- **Cooperative relationships** have been established or renewed with technology companies that are potential research collaborators, including: Cascade Designs (water treatment systems), Stellar Explorations (aerospace), Toyon (radio frequency and electronics), Phycotech (algae for biofuel), Visual Purple (virtual realilty), Moch International (catalyst manufacturer), Horsepower LLC (computer applications), Couto Solutions (software), ARB Green Power (hybrid vehicles), Electricore (R&D consortium), Discovery Life Sciences (life sciences research), Medusa LLC (R&D consortium), EFuel (alternative fuels), Vetel Diagnostics (life sciences), Rantec (power systems), NextIntent (aerospace systems).

- The **web site for the project** (www.c3rp.org), which presents C^3RP as an interface between Cal Poly and business/industry for the purpose of facilitating R&D relationships, was updated.

- Efforts continue to develop industry partners in the area of **alternative energy and energy efficiency** for the purpose of developing research in this field. To this end we have worked with Phycotech, First Solar, Continental WindPower, EFuel, Energy Alternative Solutions, Inc., Pacific Gas & Electric, Rey Energy, and Blue Aqua Solutions.

- The project's leaders have continued to work with other private and government partners to advance the project and to attract research collaborators and support, including the Institute for Energy Efficiency at the University of California, Santa Barbara and the Naval Facilities Engineering Command at Port Hueneme. During this project period, we continued to use funding provided by the Economic Development Administration for **construction of the first building of the**

technology park. Construction began in November 2008 and will be completed in Spring 2010; construction is about 45% complete. We are currently negotiating leases with technology-based companies that will become the first tenants in the building. Criteria for tenant selection include significant R&D activity in areas that complement Cal Poly's research strengths and a commitment to collaboration with faculty.

B. Development of new research capacity

One of the goals of the project was to increase the capacity of the organization to carry out state-of-the-art research in the areas of interest. To this end, specialized instrumentation was acquired and infrastructure was developed, as detailed below.

1. Instrumentation.

We acquired the following major equipment/systems. Other minor instrumentation, acquired for use on individual projects, is described in the reports for those projects.

NanoTest System

This is a fully flexible nano-mechanical property measurement system, capable of measuring hardness, modulus, toughness, adhesion and many other properties of thin films and other surfaces, as well as single dynamical load and displacement ranges. Sample environment and impacting conditions can be set up to closely replicate conditions that these materials actually see in-use. The NanoTest is a fully modular system that allows the user to configure the system to meet his/her individual needs, making it particularly suitable for the shared use that we engage in.

This system has been used by C^3RP researchers in various departments (Materials Engineering, Chemistry, Aerospace Engineering, Mechanical Engineering) who work on surface properties of coatings, composites and other materials. One special feature of the system is that it includes a higher temperature environment than any other currently available competitive product, allowing the study of materials in a high-temperature environment. Current projects on thin films, tissue engineering and properties of bone have benefited from the availability of this resource. The military has shown interest in medical research on bone physical properties.

Nanoflow HPLC and Chip/MS (ion trap) Analytical System

The acquired system is the combination of a nanoflow HPLC and a Chip/MS (ion trap) analytical system. Analysis of molecules by this system consists of powerful resolution by HPLC (separation of myriads of peptides in the case of proteomics, or high resolution of any other chemicals), followed by tandem mass spectroscopy obtained by ion-trap methodology. In research on chemical components and proteins, this is the state of the art methodology for identification of molecules, whatever their origin. The Agilent 1200 Series HPLC combined with the Chip/MS system is a new microfluidic chip-based

technology for nanospray LC/MS. Based on the HPLC-Chip and HPLC-Chip Cube MS interface, it provides a new level of nanospray MS sensitivity, robustness, ease of use and reliability.

Cal Poly entered the field of proteomic analysis after the first wave of scientific and technical advances had resulted in a second generation of protocols and instrumentation. Proteomics is the simultaneous analysis of all the proteins in a living system, microbe or organ. The analytical system we acquired provides the flexibility needed to address the diverse scientific needs of the users at Cal Poly for molecule identification. For example, the instrument has a very high sensitivity which is necessary for the quality of research we are conducting in the field of proteomics, but can also easily be used by chemists working with pure substances. Very important is its ease of use and maintenance. The separation/chromatography chemistry can be modified by changing a 'chip' (the size of a credit card), which is very convenient in an environment of diverse research work.

The analytical system has been used to support a number of C^3RP projects, among them:

- Protcomic analysis of milk fat globule membrane proteins as biologically active compounds. The instrument is used to characterize proteins of the milk fat globule membrane in an ongoing C^3RP-supported project to determine the beneficial effects of these unique proteins on human health and performance.

- Reactive coatings. The system is used to characterize synthetic molecules being created in the lab. These molecules are being immobilized in polymers and on surfaces for the neutralization of chemical and biological warfare agents.

- Post-translational modifications in enzymes and proteins. Using this instrument, kinase/phosphatase activity is monitored for important biological processes, such as signaling in malarial diseases. This tool allows us to determine the location and extent of phosphorylation in enzymes important to this disease.

- Novel marine natural products. This instrument is a vital component of characterization of new products currently being isolated in the lab. The ability to confirm both molecular mass and structure provides invaluable information on their nature and potential uses.

- Tracking environmental change in marine organisms. This C^3RP project uses the instrument to complement other proteomics capabilities that exist in the Department of Biological Sciences (MALDI-TOF-TOF). The electrospray ion trap enables separation of proteins by liquid chromatography, allowing the subsequent use of MuDPIT (Multi-Dimensional Protein Identification Technology), a powerful protein-separating technique that enables a high-throughput mode. Importantly, the MALDI-TOF-TOF offers excellent mass accuracy that is used for peptide mass fingerprinting, while the electrospray ion trap offers excellent capabilities for *de novo* sequencing of peptides and for the identification of post-translational modifications.

Other Instrumentation

Total Organic Carbon Analyzer – used for biofuels research, specifically to conduct carbon mass balances on algae ponds.

Kjeldahl Apparatus – used for nitrogen analysis in a variety of research applications, including biofuels research.

Single and Multi-cell Impedance Sensor on Chip – a custom-designed instrument created for bio/chemical sensing.

Differential Interference Contrast Microscopy System – used to study the effects of neurotoxins (nerve gases) in a model cellular system.

Reactive Ion Etcher System – this enhancement to our microfabrication laboratory uses a plasma (rather than environmentally dangerous chemicals) to etch silicon, silicon dioxide and silicon nitride and is used to fabricate microfluidic devices (lab-on-chip) and microscale electrostatically actuated mirrors; and in some solid-state lighting applications (white LEDS).

2. Infrastructure.

In the infrastructure area, the following projects were supported.

Development of a laboratory to investigate injury-repair in skeletal muscle and vasculature after traumatic injury and ischemia. Equipment was acquired to support this research which uses state-of-the-art molecular and genetic techniques to study muscle repair and angiogenesis. More detail is provided in section IV.C.2.

Development of a Biosafety Level II Facility for biosecurity research. The facility includes an autoclave, biosafety cabinets and incubators and meets the standards for conducting BL-II research. The facility will be used for biosecurity research related to protecting the food supply by monitoring and detecting pathogens.

Development of an Auto-Control System for remote operation of a Yamaha RMAX helicopter. The system has been developed to remotely operate a research-scale helicopter donated to the program by Northrup Grumman. It will run piloting software and sensing systems. More detail is provided in section IV.C.2.

C. Detailed research reports

The remainder of this report contains detailed individual reports of the technical results of the research projects carried out during this project period. They are presented in the following order:

1. Collaborative Agent Design Research Center (CADRC) project report
2. Other research project reports
3. Honors-student project reports

Representative publications resulting from this work are included with the individual reports. Documents that supplement the reports are included in Appendix B.

Knowledge Management Project
Collaborative Agent Design Research Center (CADRC)

Project Investigator:

Pohl Jens
Department of Architecture
California Polytechnic State University
San Luis Obispo, CA

Knowledge Management Project

Collaborative Agent Design Research Center (CADRC)

Executive Summary

With the rising popularity of distributed computing (e.g., Web Services), the process of mapping data within and between data models has become a primary focus of computerized information system development. Though methodologies have been developed to overcome the syntactic and semantic obstacles to the interoperability of heterogeneous data models, too often these methodologies are labor-intensive, prone to error, and require the application of considerable technical and domain-functional expertise. In addressing this growing problem, researchers and students at the Collaborative Agent Design Research Center (CADRC) have developed the Intelligent Mapping Toolkit (IMT). IMT is a software system that provides decision support for mixed-initiative semantic mapping of metadata and instance data between relational data models.

To automate specific aspects of the semantic mapping problem, IMT employs a variety of *reasoning* paradigms in the form of software agents to support human analysis. These paradigms are drawn from the field of artificial intelligence (AI). IMT utilizes a novel federation of software agents capable of leveraging an open-ended set of reasoning methods, as well as derivatives of these software agents to correct data quality issues stemming from network-centric computing.

Introduction

In its previous research efforts, the CADRC has studied the interoperability of heterogeneous data models in the domain of logistical decision-support. The application of this research led to the design and implementation of the Intelligent Mapping Toolkit (IMT) in the 2006-2007 research cycle. This initial version of IMT proved to be an effective multi-agent semantic mapping tool with demonstrated utility for large-scale problems and the potential to dramatically improve mapping throughput by USTRANSCOM analysts dealing with data quality issues in support of military logistics. Through the IMT software, the CADRC has demonstrated that combinations of intelligent agent comparators are more effective than comparators operating independently.

During the 2007-2008 project year, researchers and developers at the CADRC, under the auspices of the California Central Coast Research Partnership, extended IMT's initial data analysis support and its semantic-based mapping capabilities to offer a wider range of services including:

- New intelligent agent comparators to realize greater mapping effectiveness.

- Data cleansing capabilities for the correction of flaws in operational enterprise information.
- Intelligent data querying capabilities.

These capabilities can be employed to help relieve difficulties associated with interoperability among the underlying data of the many disparate systems within the United States Department of Defense (DoD). This report describes the evolution path to the current version of the IMT software, as well as future development goals.

Project Description

The IMT technology evolved to transcend the original vision of supporting analysts in mapping the data representations of legacy systems to a common shared representation. This was achieved by providing Intelligent Reference Data Management (IRDM) services that support semantic mapping, data analysis, operational data cleansing, and intelligent queries. All of these services were designed to be the building blocks of a problem-specific software artifact. The following section describes the outcome of current efforts to enhance this IMT technology foundation.

Semantic Mapping

Data mapping is the process of logically correlating data elements among heterogeneous data models. Because these heterogeneous data models can exhibit data discrepancies, IMT technology exploits semantic relationships and suggests the semantic mappings necessary for addressing and resolving data discrepancies. The foundational mapping services and their similarity agents define the core of IMT technology.

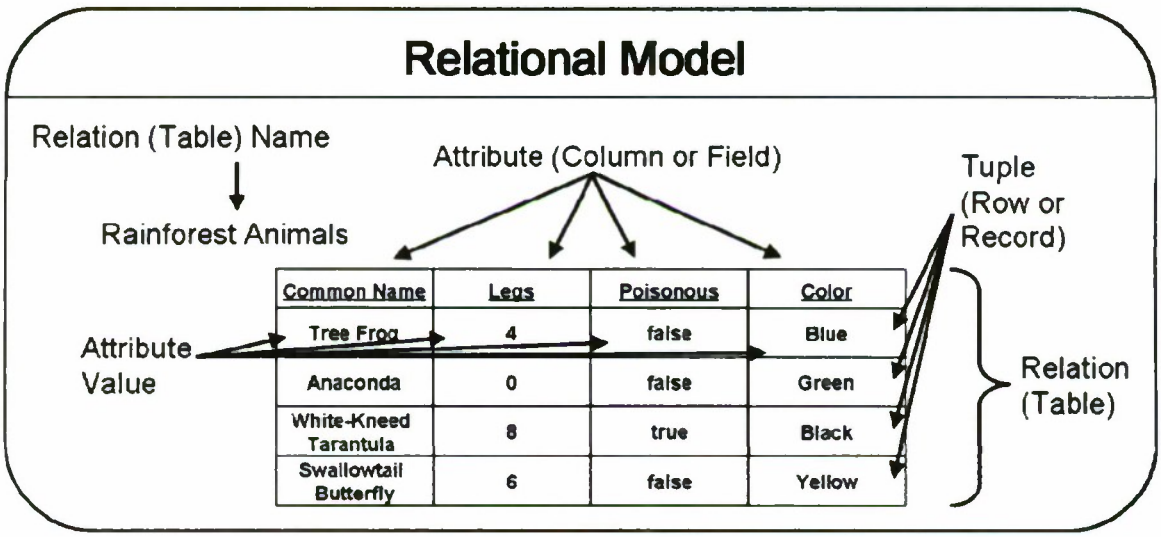


Figure 1: Example illustration referencing key elements of the Relational Model.

IMT technology assumes data models involved with mapping conform to the relational model as illustrated in Figure 1. Translator modules are used to marshal external data sources to relational model compliant data sources during run-time mapping (e.g., data is

translated to a commonly shared representation as mapping takes place). Similarly, data brokering modules are used to marshal external data artifacts to relational model compliant data artifacts prior to mapping (e.g., mapping occurs over data elements of a commonly shared representation).

In review, the original tasks that the IMT was built to carry out include:

1. ***Metadata mapping***: This entails discovery of tables and fields (e.g., structural data elements) in a target database, which correspond to tables and fields in the source database. Documentation of this relationship is referred to as metadata mapping. The IMT documents two types of metadata mappings:
    a. *Relation mapping* - the relationship between a relation or table from one database to a relation or table of another database.
    b. *Attribute mapping* - the relationship between an attribute or column from one relation to an attribute or column of another relation.
2. ***Instance-data mapping***: This entails the discovery of records (e.g., value data elements) in a target relation, which correspond to records from the source relation. Documentation of this relationship is called an instance-data mapping.

Comparison Category	Comparison Agent
Text	• Semantic Text (Inexact match) • Key (Exact match) • Phonetic

Location	• Geo-Spatial Location
	  ○ Latitude-Longitude
	  ○ Zip Code

Quantity	• Weight • Volume • Dimension
Numeric	• Real Number • Date • RGB Color
Generic	• Data-Value

Table 1: The IMT Similarity Agents.

Note that the word *correspond* in the previous context is used loosely. The criteria leveraged to promote candidate mappings come from two sources:

1. ***Software Similarity Agents***: Agents are geared toward solving domain-specific problems. If an agent specializes in comparing colors, then given characteristics of two colors the agent will produce an objective value that can be utilized to gauge similarity. Table 1 lists agents currently available to IMT.

2. ***Humans***: Human reasoning is an integral part of determining correct data mappings, since agent reasoning is limited to the information available about data elements. While the results of several agents can be aggregated in order to attempt to suggest the most accurate data mapping, a human may have domain-specific knowledge that the agent does not have. For this reason a human should always be allowed to participate in the mapping process.

The following sections describe enhancements to the IMT technology's semantic mapping capabilities that were designed and implemented during the past project cycle.

Data Value Agent

Often metadata embedded in multiple domains is found to be of such poor quality that insufficient semantic context is available to existing IMT metadata mapping agents to provide useful mapping suggestions.

Unlike the original IMT metadata mapping agents, the Data Valuc Agent does not rely on the characteristics of attributes but rather assists in creating metadata mappings by determining the similarity between two relational attributes based on their values in a set of reference data (Figure 2). If the values are of a comparable data type, then instance-based agents (currently utilizing term classification with text and numeric similarity comparators) are employed to provide a similarity metric for the attributes.

In order to apply the Data Value Agent, a set of reference data must exist, supporting each attribute in question. From a sample of each reference data set the Data Value Agent determines the type of data associated with the attribute (i.e., alphanumeric, numerical, a single character, and so on) and assigns similarity agents accordingly. The results of these agents can be aggregated with results of other mctadata mapping agents to increase the accuracy of the metadata mapping suggestions.

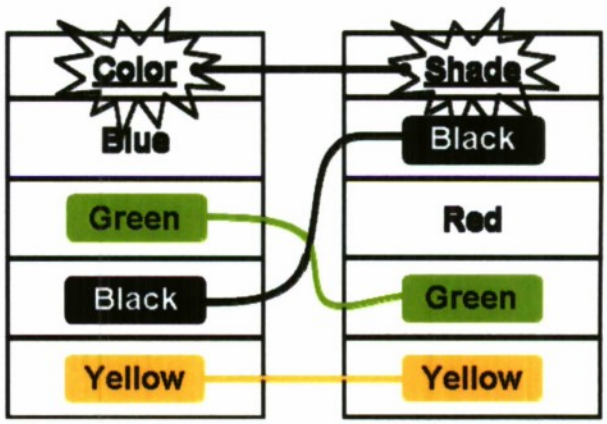


Figure 2: A comparison is made between instances of two attributes.

Phonetic Agent

One of the subjects of research and development in 2008 involved the expansion of IMT's collection of agents to include the Phonetic Agent. The Phonetic Agent brings IMT technology closer to the world of natural language processing by utilizing sounds involved in human specch to determine the similarity betwcen words and phrases.

The Phonetic Agent employs a Double Metaphone algorithm to classify words by their pronounced sound (Philips 2000). For example the terms "anaconda" and "annahkanda" will both transform to "ANKNT" and be considered equivalent by the Phonetic Agent. To produce a similarity metric, the Phonetic Agent classifies every two-gram derivative of the reduced terms and employs the Case-Based Classification engine.

Semantic mapping services offered by IMT technology can readily take advantage of the Phonetic Agent by assigning a weighting factor that allows the results produced by the agent to contribute to the overall similarity metric of the data elements for both metadata and instance-data mapping suggestions.

Color Agent

An instance-data-based agent that has recently been added to the IMT technology's suite of similarity agents is the color agent. The color agent determines the similarity between two Red-Green-Blue (RGB) color values. A color can be broken up into its red, green, and blue components, each ranging from zero to 255. For example, a very bright red color would have a red component of 255, a green component of zero, and a blue component of zero. The simple but effective approach taken by the Color Agent is to plot the red, green, and blue values for any two colors in three-dimensional space on x, y, and z axes, respectively. A traditional three-dimensional distance formula is then applied to produce a metric that represents the similarity between the two colors, relative to the other colors in the data that are being considered.

Because a typical set of data will not contain the red, green, and blue integer values for a color attribute (i.e., a database of people might list a person's hair color as *brown*), the color agent supports the loading of synonyms prior to determining color similarity. A synonym passed to the color agent is expected to map the name of a color (i.e., *green*) to a hexadecimal value (i.e., 0x00FF00). The red, green, and blue values are extracted from the hexadecimal value and are used in place of the data wherever the name of the color is found.

Intelligent Query

A direct derivative of the mapping technology is the intelligent or semantic query capability. The existing IMT instance-data mapping agents allow for the performance of inexact queries across any type of character-encoded data. Every reference data set attribute specified within data managed by IMT is associated with an agent that will perform instance-data similarity comparisons to achieve a powerful *query by similarity* capability.

The scope of database query facilities desirable for a semantic search capability far exceeds traditional database management system (DBMS) functions. They presuppose a level of embedded intelligence that has not been available in the past. Some of these desirable features include: conceptual searches instead of factual searches; automatically generated search strategies instead of predetermined search commands; multiple database access instead of single database access; analyzed search results instead of direct (i.e., raw) search results; and, automatic query generation instead of requested searches only.

A traditional DBMS typically supports only factual searches. In other words, users and applications must be able to define precisely and without ambiguity what data they require. In complex problem situations users rarely know exactly what information they require. Often they can define in only conceptual terms the kind of information that they are seeking. Also, they would like to be able to rely on the DBMS to automatically

broaden the search with a view to *discovering* information. Traditional query technology utilizes exact match and Boolean logic constraints to identify and retricve an item of interest (e.g., name = Joe or length < 100). In traditional search engines, Boolean logic keywords are commonly used for identifying items of interest (e.g., description contains 'Joe'). IMT technology offers the ability to search inexactly on free form text.

An effective query mechanism is broken into two major components:

1. ***Term Parser***: All known documents and query criteria are divided into terms. Terms can represent paragraphs, sentences, phrases, words, or even pieces of words, such as n-grams and phonetic parts. IMT technology is built on n-gram (specifically tri-gram) and word terms. Due to the abundance of terms, an efficient data structure that supports rapid maintenance and retrieval should be used. A typical data structure for this purpose is some variant of a B-tree.
2. ***Similarity Comparator***: After the terms of a documents have been extracted and indexed in a data structure, the next step in the information retrieval process involves comparing terms from the query criteria to the indexed terms. This comparison leverages IMT's instance-data mapping agents.

Intelligent query is an evolving capability of IMT technology that is proving useful for USTRANSCOM endeavors. This technology will receive a great deal of focus for 2009 efforts involving Document Search and Agent-Based Semantic Search capabilities.

Data Cleansing

Data cleansing is the process of validating error prone operational data against reference data, finding invalid or questionable attribute values, and presenting suggestions for correction. The end-result is a set of higher quality data with few or no errors.

There are two common practices that introduce errors into data.

- ***System-to-system data transfer***: When data are transferred from one system to another system, there is a chance that data can be altered due to data model translation mistakes. When this cycle is repeated, the problem compounds resulting in partially invalid data.
- ***Human input***: When data are manually entered into a system, the human user can make mistakes and enter invalid data. Many times these errors will have a pattern such as neighboring keys on the keyboard being pressed instead of the correct key. Homoglyphs, as described previously, can also add to errors whenever data are entercd manually.

Guaranteeing the accuracy of operational data is crucial precisely because such data are commonly used as an input to systems that rely on valid data in a standard format. Most systems will fail when provided with slight alterations of an expected data format, even though the data may be meaningful to a human operator.

In order to cleanse data, a great deal must be known about thc reference data that acts as an oracle in the cleansing process. By making use of the semantic mapping tools offered by the IMT tcchnology, metadata and instance-data relationships between data elements

are first exposed. For a simple example, suppose that operational data eontains the name of a rainforest animal. In order for thc data cleansing proeess to eheek if the animal name is present in the Rainforest Animals referenee data shown in Figurc 1, a metadata mapping must exist between the operational and reference attributes. Onee the proper metadata and instanee-data mappings are established, then the following types of validations ean oeeur:

Identifier Validation: Identifiers are distinguished by known attribute mappings from an operational data table to a reference data table. If an identifier value in the operational data is not found in the reference data, then thc operational valuc is deemed invalid. For example, suppose that the operational data eontain an identifier attribute eorresponding to the Common Name attribute shown Figure 1 with the attribute value of "Polar Bear." IMT teehnology determines that "Polar Bear" is not a valid Common Name for a rainforest animal beeause it is not part of the referenee data.

Intra-Instance Field Validation: Given an identifier, a reference data table may eontain attributes expressing eharaeteristies of the identified entity (i.e., a tree frog has four legs). The operational data may eontain the same attributes, thus it is possible to eorrelate reeords based on an identifier attribute and validate supporting attribute values. Suppose the operational data speeifies that a tree frog has six legs. While it is entirely possible for a Rainforest Animal to have six legs, it is not valid for a tree frog to have six legs. In this case the operational value is deemed invalid and sinee the IMT teehnology has sueeessfully eorrelated operational data with referenee data to determine that the combination is invalid, the eorreet referenee data values ean be aequired as suggested eorreetions. This type of validation utilizes only a single referenee data set and no instanee-data mappings.

Inter-Instance Field Validation: This type of validation produees the same results as the Intra-Instanee Field Validation; howcver, it works aeross multiple referenee data sets tied together by instanee mappings.

Constraint Validation: This technique does not rely on referenee data but rather on domain-specific constraints. An operational attribute value is deemed valid if it eonforms to a Boolean logie eonstraint sueh as: less than; starts with; or, matches a regular expression.

Complete Record Validation: The IMT teehnology validation serviee ineludes a module that performs eomplex SQL queries to eorrelate operational data with reference data and determines the minimal number of required ehanges to a reeord in order to deem it entirely valid. This eapability relies heavily on the other validation teehniques and is not useful for determining what makes a reeord invalid, but suggests how to fix it.

Suggestions: Suggestions for the correction of attribute values to establish validity are the result of two processes:

- *Intelligent Query*: Continuing examples with the rainforest animals reference data, suppose that identifier validation deems the attribute value "annahkanda" invalid. Leveraging IMT technology's intelligent query capability in conjunction with the Phonetic Agent, the data cleansing technology will suggest the correction "anaconda".
- *Instance Correlation*: Intra-Instance Field Validation and Inter-Instance Field Validation rely on enumerated valid candidate values to determine validity of attribute values. If an attribute value is deemed invalid then the enumeration of valid candidate values is simply returned as suggestions.

To understand a comprehensive example of data cleansing, please first review Figure 3. The operational data contains seven attributes: stFirstName (a student's first name); stLastName (a student's last name); profFirstName (a professor's first name); profLastName (a professor's last name); section (a class section); grade (a student's letter grade for a class section); and, Grade Points (a student's awarded grade points for a class section). Five reference data sets exist: Students, Professors, Class Sections, Letter Grades, and Grade Points. Many metadata and instance-data mappings exist among these data sets, exposing the relationships within and among the data.

The following validation can take place as shown in Figure 3:

1. Leveraging Identifier Validation reveals if a student with the specified *first name* exists.
2. Leveraging Identifier Validation also reveals if a student with the specified *last name* exists.
3. Leveraging Intra-Instance Field Validation reveals if there exists a student with both the specified *first name* and *last name* (e.g., the combination is valid).
4. Leveraging Inter-Instance Field Validation reveals if the specified student belongs in the specified *class section.*

Validation will continue in a similar fashion until every operational attribute value has been validated.

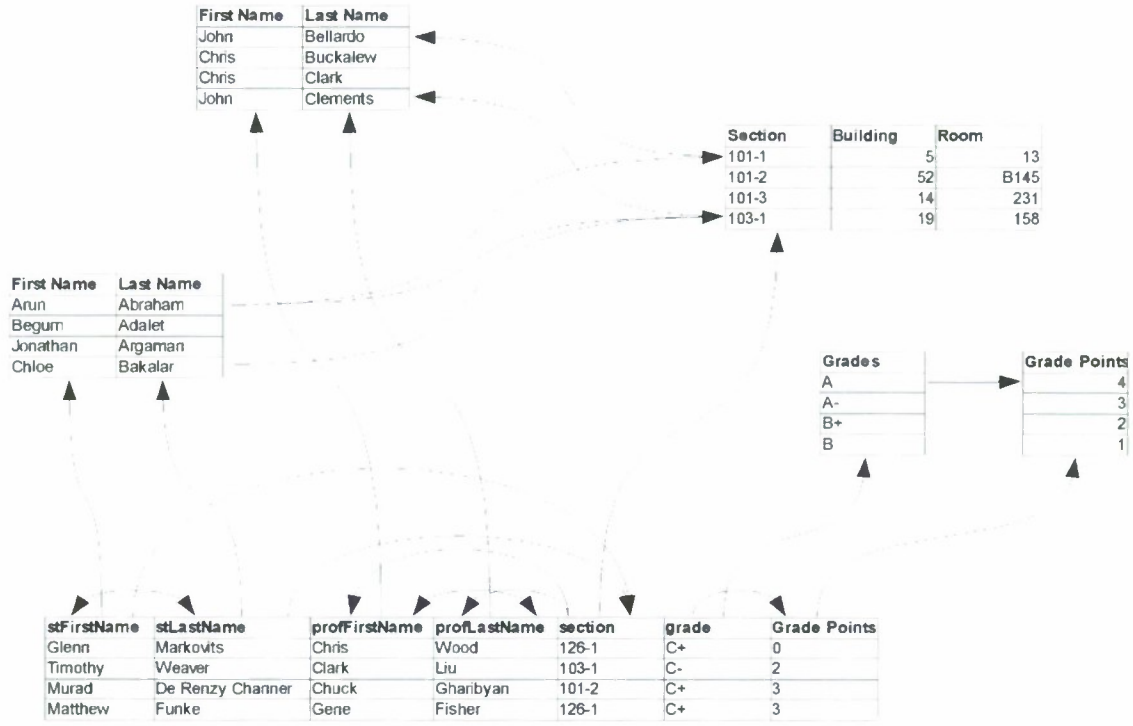


Figure 3: An example illustrating the relationships among reference and operational data sets.

Service Orientation

IMT was designed and implemented as a stand-alone single-user system. Diverging from the monolithic software paradigm, IMT has applications in the Software as a Service (SaaS) domain. Services, particularly Web Services, expose IMT's core capabilities over a network, such as the Internet, for use by other software applications in completing their work. SaaS services take advantage of concepts drawn from the Service-Oriented Architecture (SOA) paradigm and are implemented as components of an Enterprise Service Bus (ESB), thereby enabling seamless integration into diverse systems.

Service interfaces offered by IMT technology rely on XML-based messaging between a client and server. The primary interface utilizes the World Wide Web Consortium (W3C) Simple Object Access Protocol (SOAP) standard to achieve intelligible communication between the IMT services and clients (World Wide Web Consortium 2007, April 27). In addition to the SOAP interface, the popular but non-standardized Representational State Transfer (REST) architecture is also supported for XML message exchange. The popularity of REST stems from its use by web browsers to request web page content over the Internet.

Support for the service interfaces is incomplete without resources describing the data model utilized for establishing message content. IMT technology describes its service interface data model via the W3C endorsed Web Service Description Language (WSDL) and XML Schema (World Wide Web Consortium 2007, June 26). Figure 4 illustrates the service interfaces provided by the Intelligent Mapping Facility, which are described in a WSDL document.

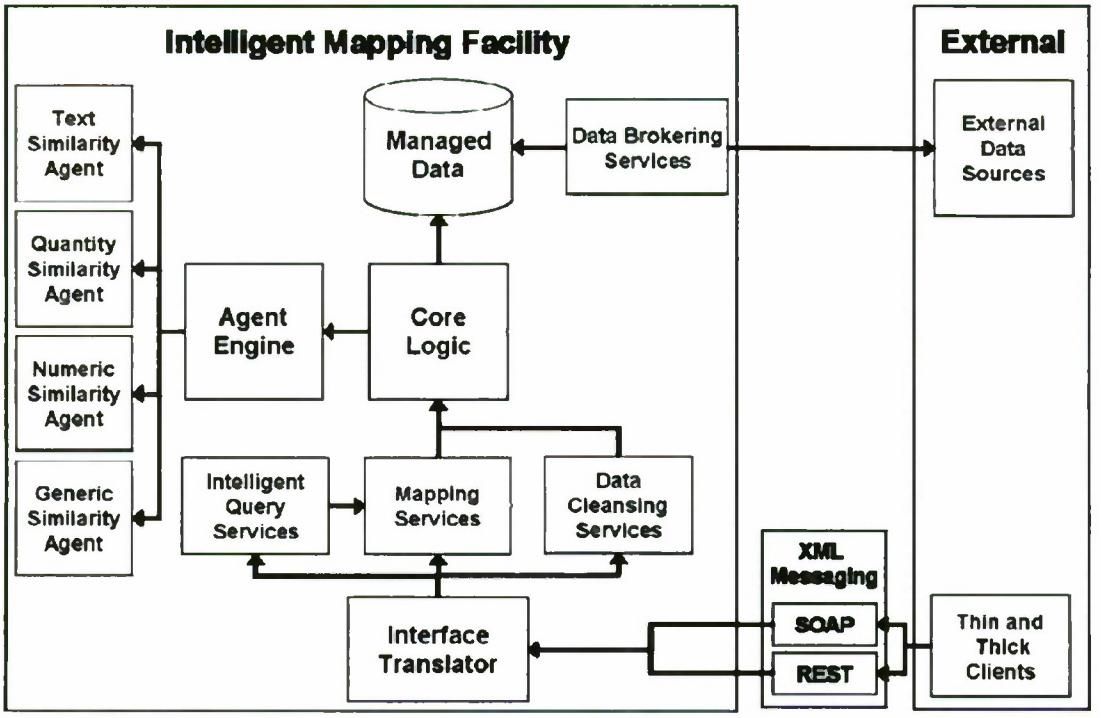


Figure 4: Service Oriented Architecture of the Intelligent Mapping Facility.

In September 2008 the CADRC began discussions with the J6 division of USTRANSCOM relating to the establishmcnt of a Knowledge Management Laboratory (KML) as a *field research unit* of the CADRC Center. The KML facility would serve as a test bed for current USTRANSCOM-J6 efforts to transition to an information-centric SOA-based knowledge management environment, under a CRADA contract. The goal of the laboratory is to showcase Software as a Service (SaaS) concepts and provide a test bed for service integration. IMT technology was featured as a primary player in the initial test bed presentations.

Second Research Cycle (focus for 2009)

Document Search

The focus for IMT technology research in 2009 will be the further evolution of the intelligent query capabilities to support document search.

Proposed Areas of Research:

- *Research of Indexing Technology*: The current intelligent query mechanisms adopted from the IMT semantic mapping technology offer a linear order of complexity for term classification. For very large data sets, searching complexity of this magnitude is not feasible when confronted with a time sensitive problem. Indexing technology for efficient information retrieval typically involves variants of an Inverted Index and B-tree with a logarithmic worst case cost in locating a record.
- *Information Retrieval*: To avoid the overhead of constructing data structures necessary to search on every service request, the proposed new indexing technology would construct the data structure on the server's hard-disk drive. Data structure content on-disk would not be loaded into memory when needed; instead, it would use fast disk searching algorithms for information retrieval.
- *Term Parsing for Standard Document Formats*: To accommodate indexing of documents, terms must first be extracted from the document's content. The goal of a term parser is to locate the content for term extraction and then employ a standardized tokenizing algorithm (e.g., n-gram tokenization). The process of accessing required content can vary among document formats, necessitating the need for specialized term parsers.
- *Concept Extraction*: To infer categorization by extracting the meaning from a document. For example, a document describing "trees" could refer to Biology or Computer Science because the keyword can have different meanings.
- *Workload Distribution and Clustering*: To allow independent instances of IMT services to work together in solving a problem. This feature may involve partitioning indexed terms among servers.
- *Identification of a Test Environment*: To select an appropriate case study testing domain. Initial efforts for locating a suitable testing environment focused on working in the domain of a library; particularly, the Kennedy Library at the California Polytechnic State University (Cal Poly) in San Luis Obispo. An ideal approach for integration and testing of the search technology would entail enhancement of the library's recordkeeping activities to supply an inexact full-content search. Other possible participants include the Cal Poly College of Architecture and Environmental Design's Media Resource Center.

Other Areas of Research

- *Service Security Layer*: Enforce integrity and confidentiality on IMT service messaging. Initial research into this subject should revolve around the Open Oasis specification WS-Security (WSS). WSS describes how to attach signatures and encryption headers to SOAP messages.
- *Natural Language Processing*: Convert human language into a more formal and consistent representation for manipulation by computer algorithms. This subject of research may involve determining both the semantics of a document,

or extracting additional types of terms from a document (i.e. Phoneme Extraction Agent).

- *WordNet Integration*: Enhance agent growth by utilizing external resources such as WordNet, a publicly available linguistic ontology (Fellbaum 1998). WordNet will strengthen an agent's ability to identify occurrences of terminological variations due to conceptual abstraction. For example, WordNet's hyponym relationships (is-a-type-of) between concepts (i.e., a Swallowtail Butterfly is-a-type-of insect) can be exploited.

References

Fellbaum C. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Philips L (2000). The Double Metaphone Search Algorithm. *Dr Dobb's Journal.* Retrieved from http://www.ddj.com/cpp/184401251, 1 June.

World Wide Web Consortium (2007). SOAP Version 1.2 Part 0: Primer (Second Edition). Retrieved from http://www.w3.org/TR/2007/REC-wsdl20-primer-20070626, 27 April.

World Wide Web Consortium (2007). Web Services Description Language (WSDL) Version 2.0 Part 0: Primer. Retrieved from http://www.w3.org/TR/2007/REC-soap12-part0-20070427, 26 June.

Quenched Random Disorder in Soft Condensed Matter

Project Investigator:

Saimir Barjami
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

Project title: Quenched Random Disorder in Soft Condensed Matter.
Investigator: Saimir Barjami, Physics Department, Calpoly, San Luis Obispo, Ca.
Project Goal: The goal of this project was twofold:

1. To establish a world-class experimental facility in the Physics Department, in the College of Science and Math, California Polytechnic State University for the study of order and disorder phenomena in condensed matter, the Order-Disorder Phenomena Laboratory.
2. To study the physics of quenched random disorder effects on the phases and phase transitions of soft condensed matter and complex fluids (liquid crystals).

Summary of Results:
I will present the summary of results from this project based on the two goals presented in the "Project Goal" part of this report.

1. **Building and operation of the AC Calorimeter for the Order-Disorder Phenomena Laboratory.**

Two students were hired beginning Spring quarter and Summer quarter in the construction and testing of AC Calorimeter. A block diagram of the design and electronic equipment of the ac calorimeter that was built from scratch and will be used in this work are presented in figures 1 and 2. The work to built and test the AC Calorimeter extended over a period of 6 months, from March 2008, till August 2008, due to the very complex nature of the calorimeter.

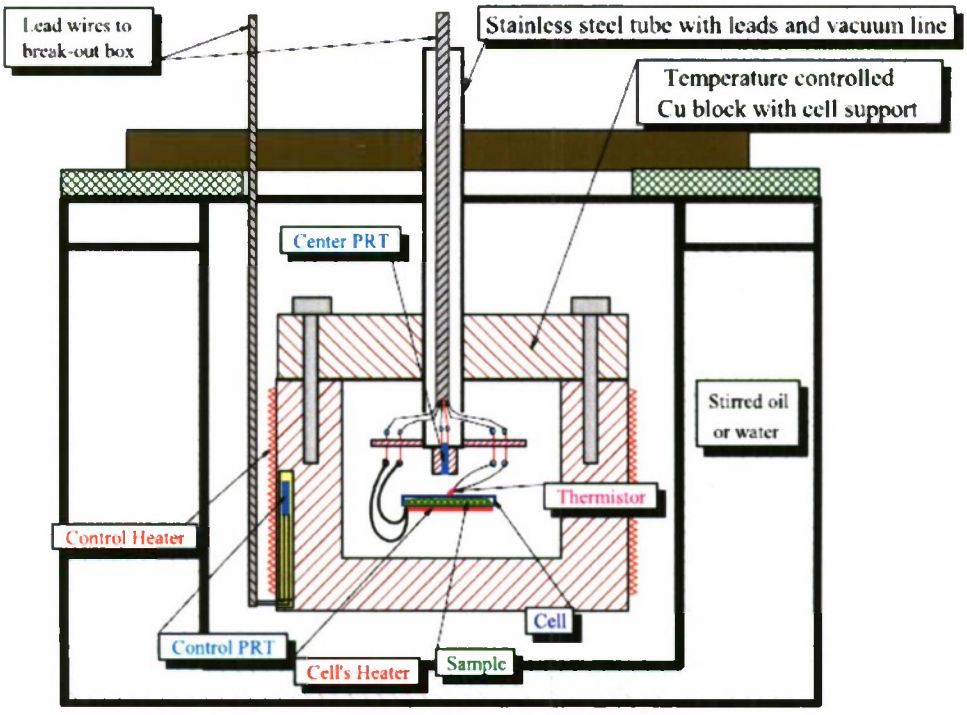


Figure 1: Schematics of the *AC* Calorimeter build in our Order-Disorder Phenomena lab. The sample lies inside a massive copper block, which is temperature controlled to better than $1mK$ in the stepwise mode.

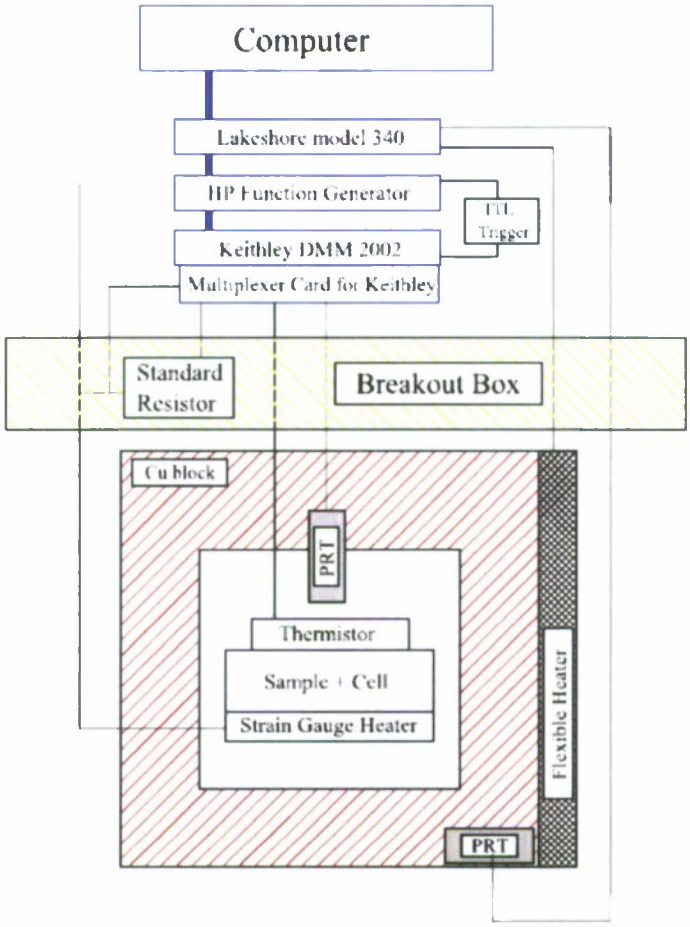


Figure 2: Block diagram of Calorimeter built in our Order-Disorder Phenomena Lab., showing all the connections to the instruments. The whole control and data acquisition is computer controlled through a GPIB interface.

A detail list of equipments and accessories that were bought to construct the calorimeter will be attached to this report upon request.

Initial Results in Testing the AC Calorimeter.

The basics of the AC calorimetry technique consist of applying periodically modulated sinusoidal power to the material to be studied, and monitoring the resulting sinusoidal temperature response of the material. After sending the sinusoidal power to the material we monitored the temperature response oscillations attached to the sample. Such oscillations are presented in Figure 3.

The data shown in Figure 3 represents the raw temperature oscillations in the sample, without any averaging or smoothing routine used, which highlights again the very high - resolution capabilities of our calorimeter.
Sinusoidal response of the sample shown in Figure 3 is a direct indicator of the successful operation of the new, built from scratch AC Calorimeter in our new Order-Disorder Phenomena Laboratory, which we employed in the study of quenched random disorder effects.

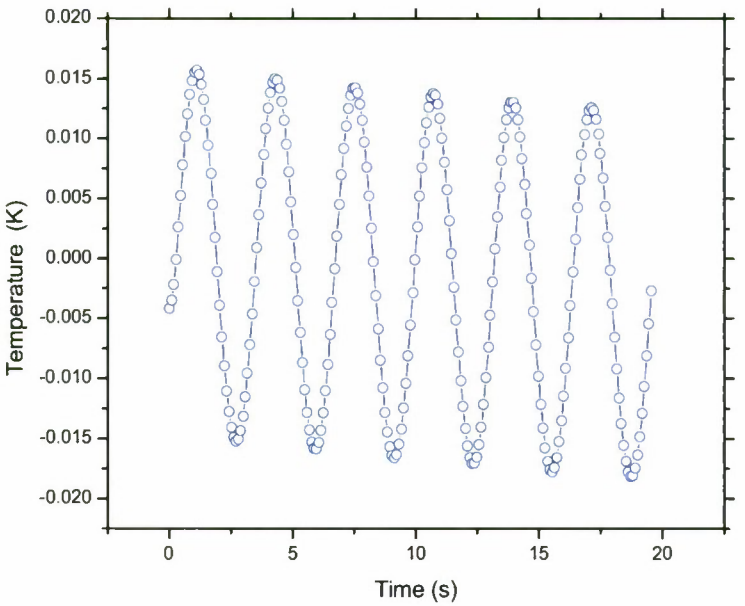


Figure 3: Sinusoidal response of AC Calorimetry. The data shown represents the raw temperature oscillations in the sample, without any averaging or smoothing routine used, which highlights the very high - resolution capabilities of our calorimeter.

2. Quenched Random Disorder in Soft Condensed Matter.

The isotropic (*I*) to nematic (*N*) phase transition is weakly first-order and the orientational order is characterized by a tensor order parameter. In the presence of dispersed surfaces as the source of the Quenched Random Disorder (QRD), the disorder couples to both the magnitude and the direction of the orientational anisotropy axis (the director). Thus, in general, both field-like and elastic couplings may occur. In Liquid Crystals (LC) + aerosil systems, the soft nature of the gel partially anneals the stronger elastic coupling reveling the random-field coupling to the magnitude of the order parameter.
In this experimental research work we focus on a liquid crystal – 4' – n – Pentyloxy – 4 – Cyanobi – phenyl (5OCB). We prepared several samples of 5OCB with quenched random disorder (QRD) incorporated into them. QRD in Liquid Crystals (*LC*) requires the inclusion of fixed random solid surfaces at all possible length-scales up to the sample size. This was achieved by "dissolving" type 300 aerosil (*SIL*) into the host *LC*. The *SIL* is comprised of SiO_2 (silica) spheres of diameter about 7 *nm*, coated with (*-OH*) hydroxyl group. The coating enables the spheres to hydrogen bond and form a thixotropic10, fractal gel, in an organic solvent, through a diffusion limited aggregation process.

Several AC Calorimetry temperature scans were carried out for bulk 5OCB and for several densities of aerosil particles in 5OCB:

$\rho_S = 0.078\frac{g}{cm^3}$; $\rho_S = 0.220\frac{g}{cm^3}$; $\rho_S = 0.489\frac{g}{cm^3}$; $\rho_S = 0.647\frac{g}{cm^3}$. This work was carried out from September 20 until December 15.

In order to determine the excess heat capacity associated with the phase transitions, an appropriate background was subtracted. The total sample heat capacity over a wide temperature range had a linear background, *Cp*(background), subtracted to yield:

$$\Delta C_p = C_p - C_p\left(background\right)$$

In Figure 4 we present excess specific heat, ΔCp, as a function of temperature for bulk *5OCB* liquid crystal taken with our new AC Calorimeter, over a wide range of temperatures, from 305 K to above 360 K. A Sharp Isotropic to Nematic phase transitions is observed for the bulk at 337 K., as shown clearly in Figure 4. A double calorimetric feature is evident in the I+N coexistence region from the excess specific heat data as seen in Figure 4, and is a signature of the first order character of this transition.

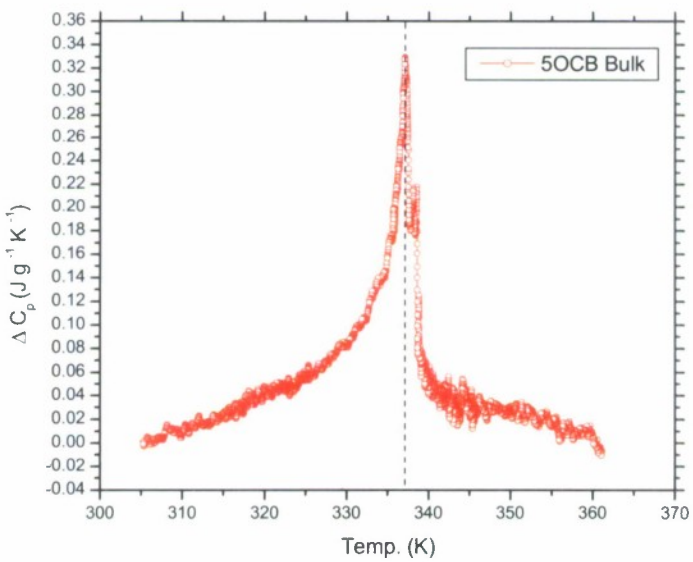


Figure 4: Excess specific heat, ΔCp, as a function of temperature for bulk *5OCB* liquid crystal. The vertical dashed line indicate the *I-N* transition. Isotropic to Nematic phase transition is observed for the bulk at 337 K.

An expanded view of the temperature scan data about the Isotropic to Nematic phase transition as a function of temperature is shown in Figure 5. A double calorimetric feature is evident in the I+N coexistence region from the excess specific heat data as seen in Figure 5, and is a signature of the first order character of this transition. This is a clear indication that the release of the latent heat is affected by the disorder, and follows a two-step process. Disorder seems to affect both the magnitude and the dynamics of the latent heat release.

Previous work done on the Isotropic to Nematic phase transition for CCN47 + aerosil disordered systems, presented in the Project Proposal, showed a similar double calorimetric feature for this transition and offers compelling evidence that the I-N transition with weak quenched random disorder proceeds via a two - step process, in which random-dilution is followed by random-field interactions on cooling from the isotropic phase, a previously unrecognized phenomena. We believe the double calorimetric feature evident in the Isotropic to Nematic phase transition for bulk 5OCB taken with our new Calorimeter, confirms that result.

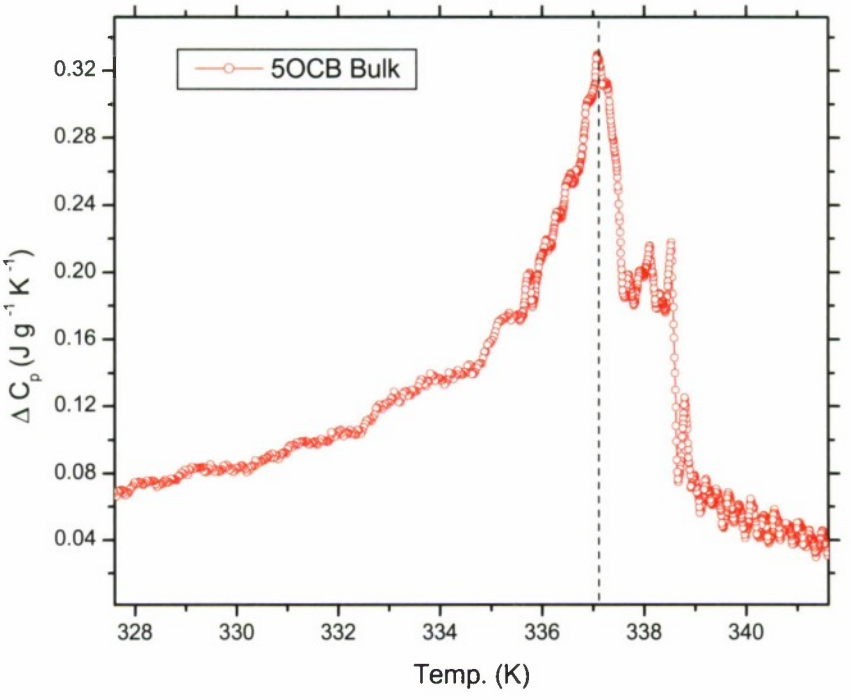


Figure 5: Expanded view of the temperature scan data about the nematic to isotropic phase transition as a function of temperature for bulk *5OCB* liquid crystal. The vertical dashed line indicate the *I-N* transition. Isotropic to Nematic phase transition is observed for the bulk at 337 K.

After bulk 5OCB data, we then proceeded with several AC Calorimetry temperature scans for 5OCB + aerosil disordered systems for several densities of aerosil particles in 5OCB: $\rho_S = 0.078\frac{g}{cm^3}$; $\rho_S = 0.220\frac{g}{cm^3}$; $\rho_S = 0.489\frac{g}{cm^3}$; $\rho_S = 0.647\frac{g}{cm^3}$.

In Figure 6 we present different temperature scan data taken for the above samples and for bulk 5OCB.

Isotropic to Nematic phase transitions for different samples are clearly indicated in the Figure 6 by different peaks. Temperature shifts and the suppression of the Isotropic to Nematic phase transition peaks are clearly seen in Figure 6 as the density of silica is increased from 0.078 to 0.647 g/cm^3.

The evolution of the first order Isotropic to Nematic phase transition peak is clearly indicated from the picture starting from the bulk, getting bigger for 0.078 g/cm^3 sample, getting supprcssed for the 0.220 g/cm^3, more suppressed for 0.489 g/cm^3

sample and almost smeared out for 0.647 g/cm^3 sample, which is a clear signature of the Quenched Random Disorder effects of aerosil particle in 5OCB liquid crystal.
The effect of quenched random disorder is clearly showing in the data: the shift in transition temperatures, the broadening of the transition region, the suppression of the peak with increasing silica density, and even the complete disappearance of the peak for the 0.647 g/cm^3 sample.

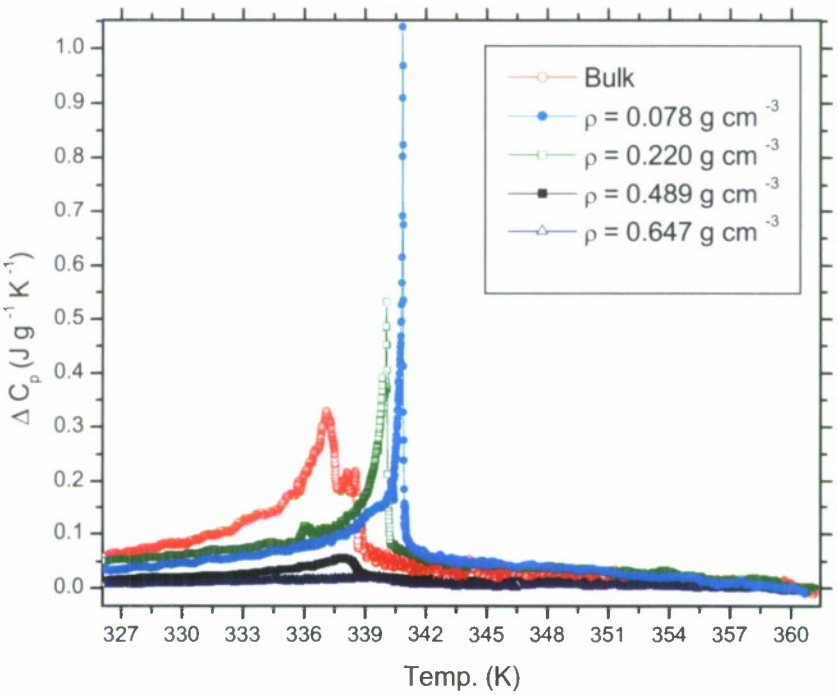


Figure 6: Excess specific heat, ΔCp, as a function of temperature for bulk *5OCB* and *5OCB* + aerosil samples at different silica densities taken with our new AC Calorimeter. Isotropic to Nematic phase transition regions are clearly seen in the figure. See figure insets for definition of symbols.

The I-N Transition Enthalpies:
For first-order transitions finding the I-N transition enthalpy is complicated by the presence of a two-phase coexistence region, as well as a latent heat ΔH. The total enthalpy change through a first-order transition is the sum of the pre-transitional enthalpy and the latent heat. In an *AC*-calorimetric measurement, ΔCp values observed in the two-phase region are artificially high and frequency dependent due to partial phase conversion during a *Tac* cycle. The pre-transitional enthalpy δH is typically obtained by substituting a linearly truncated ΔCp behavior between the bounding points of the two-phase coexistence region into:

$$\delta H = \int \Delta C_p dT$$

An independent experiment is required to determine the latent heat ΔH. A direct integration of the observed ΔCp yields an effective transition enthalpy δH^* and this

contains some of the latent heat contributions, thus $\delta H < \delta H^* < \Delta H_{total} = \delta H + \Delta H$. We will conclude this analysis at a later time.

We conclude pointing out the clear success of our project. We have designed, constructed and successfully tested the AC Calorimeter, which did require an extensive work during six months and participation in the project or two undergrad students, Aaron Jahoda, Physics major, and Ryan Koether, Math major. We have investigated the Quench Random Disorder effects in *5OCB* liquid crystal + aerosil samples at different silica densities.

This work and its results are part of the very important applications of Liquid Crystals and their many effects in many areas of science and engineering, as well as device technology, such as Optical and Radar Imaging, with a tremendous interest to **the Office of Naval Research** . Applications of Quenched Random Disorder in Liquid Crystals are still being discovered and continue to provide effective solutions to many different problems.

These results will be submitted for publication in one of Physical Review Journals at a later time.

Administrative Details

- The PI took 2 units release time in Spring Quarter 2008.
- The PI received Summer salary from this project.
- Two students were hired during Summer Quarter 2008.

Communication within Satellite Constellations

Introduction

The CubeSat's small, standardized form-factor has been one of its biggest assets. It allows the satellites to be flown economically, enabling a myriad of applications and experiments that were previously not conducted due to high launch costs. However, this same form-factor poses one of the biggest engineering challenges to designing, building, and deploying CubeSats. The size and power constraints that result from such a small enclosure limit the complexity of the satellite's payload and corresponding experiments.

There are two general approaches to working within these constraints. The first is to push the miniaturization envelope, creating ever-smaller payloads. The second approach is to fly multiple satellites in a constellation (or cluster) that cooperate with each other, each performing a piece of the overall mission. To date the cluster approach has been rarely used with extremely small satellites, however the operational concept is well known and discussed amongst CubeSat developers.

Current CubeSat communication protocols are built on the assumption that the ground station communicates directly with the satellite. This simplifies protocol design in a number of ways. For instance, if the Cal Poly ground station has a packet to send to a CubeSat (CP3), it simply waits for CP3 to pass overhead and sends the packet. This is easier than sending the CP3 packet to the first satellite that passes overhead, CP4, and then depending on CP4 to relay the packet to CP3. This relaying capability is a fundamental component of cluster communication.

Satellites within a constellation, especially those launched in different P-PODs, will be in periodic communication range of each other as determined by their orbits. This periodic connectivity requires a dramatically different routing strategy than is employed in fixed, "always on" connections. Existing CubeSat ground stations resolve periodic connectivity through the use of human operators and orbital predictions. The operators use the predicted time of the next pass, prepare commands for the satellite in advance, and then perform the actual communication as the satellite passes overhead. Satellite constellations, on the other hand, have no onboard human operator responsible for communication. All the scheduling, relaying, and routing must be implemented by onboard software.

Initial development of this networking software, the corresponding protocols, and the necessary tools to evaluate them before flight, was the focus of this work. The remaining sections in this report provide an overview of the work done and the work yet to do in four important areas, including how this work aligns with The Office of Naval Research (ONR) goals, the communication protocol, the protocol simulator, and initial protocol analysis.

Alignment with ONR Goals

Satellite constellations can be used to solve a number of different and important problems. The research in this proposal is a fundamental enabling technology for constellations. This section describes how constellation technology could be applied to three different categories of specific interest to ONR and DOD: national security applications, command and control / communication, and intelligence / reconnaissance.

This work has two primary national security applications. First, a large part of national security is ensuring that our critical infrastructure is protected against failure when attacked. While it is beyond the scope of this work to identify exactly what constitutes critical infrastructure, this research will develop and evaluate different technology that provides the most basic form of failure protection, redundancy. In addition, advanced satellite technology is considered a strategic advantage, and this research will advance the state of the art in satellite communication technology.

This research also falls squarely into the ONR's interests of command and control, computers, and communication. The broad categories of computers and communication are easily meet by the nature of the research. Specific example applications include deploying a disruption tolerant satellite network that can carry mission critical data securely (command and control, etc). The tolerance is obtained from the redundancy in the system, however the security goal is not as obvious. The underlying technology needs to be designed and tested such that is operates correctly even when under computer-based attacks, like Denial-of-Service attacks, that don't cause any physical harm to the satellite.

Finally, the ability to amortize the cost of a single, high bandwidth link across many satellites makes this technology ideal for intelligence, surveillance, and reconnaissance applications when it is necessary to deploy a large number of resource constrained sensing nodes and related infrastructure.

Protocol

The first step in designing a routing protocol is determining the overriding behavioral categories for the protocol. These categories have a pronounced impact on the details of the design. In multi-hop wireless networks, there are two decisions to make: Should the protocol be proactive or reactive and should the protocol be link-state or distance vector.

Proactive protocols determine the path through the network before any packet is actually transmitted. This is used, and works very well, in static networks (like the Internet) and multi-hop routing protocols for networks whose mobility pattern results in nearly static networks.

The decision was made to use a proactive protocol for satellite cluster networks because of two factors. First, the orbits are predictable. The satellites themselves

can make high quality predictions of when they will come in communication range of each other. This makes the network properties closer to static than dynamic. The second reason is that exposure time, the time when one satellite is in communication with another satellite, is relatively small and reactive protocols tends to have much higher initial overhead in every exposure period. These two factors lead us to design a proactive protocol.

The second major design decision is selecting between link-state and distance vector. In a link-state protocol each node builds a virtual map of the entire network, and then uses graph algorithms to select the best route through the network. In a distance vector network each node only learns about the nodes its neighbor can reach, and then selects the best neighbor for the current packet. Given the disconnected nature of CubeSat constellations, we decided that a link-state protocol is more appropriate. It enables the protocol to use the aforementioned exposure predictions in a straightforward manner. A distance vector protocol would have also resulted in much higher protocol overhead due to the dynamic network.

The link state table, which holds the information necessary for the routing protocol, contains more information than a traditional link state protocol. In addition to storing details about which satellites are within communication range of each other, it also stores information that enables the satellites to predict when they will no longer be in range, and when they with subsequently fly in range again.

Since the link state information is reasonable concise, sending it between satellites enables each satellite to predict exposure time and transmission opportunities for the entire network.

Types of Nodes

This work divides the responsibilities of network node into three different categories: Low Latency Mule (LLM), High Bandwidth Mule (HBM), and Sensor. The LLM allows small amounts of information to be transmitted between the Sink and the sensors. This is primarily intended for command and control data from the ground station to the satellites. Since ground station exposure is small, roughly 15-20 minutes per pass with only a few passes a day, this connection is not able to transmit large quantities of data.

The HBM primarily used for bulk data transfers, enabled by deploying a higher powered and larger satellite in capable of higher downlink bandwidth. The primary use of the HMB is to move large amounts of payload data, like images or complex data tables, from sensors down to Earth. The HBM has a longer period between exposures to each satellite, but stays in range of that satellite for several hours each time. The HBM will also have a much higher bandwidth connection with the sensors because the range will be much smaller and with less atmospheric interference than from a sensor straight to the ground station. In addition to increased bandwidth, this can also save power on the sensor because each sensor will be sending the data a shorter distance with a lower power radio.

The HBM will then relay this information down to the ground station much faster than an individual sensor can. This is possible because the HBM is a larger satellite with more power and a larger antenna, which enables it to achieve higher data rates than a sensor can. In the case of the low earth orbits (LEO; like CubeSats), the sensors see the ground station 40 minutes a day, but will only see the HBM every one or two months.

The final type of node is the sensor itself. The sensors can be gathering any sort of data from the payload to be transmitted to the ground station.

Satellite Modes

Each node in the network may be in one of three different modes: discovery, power-save, and active.

Discovery Mode

Discovery mode is used to learn about the orbits and exposure times of peer satellites in the network. This mode is entered immediately after satellite deployment. In this mode the satellite sends out broadcast packets to discover all of the satellites in the network. A broadcast packet is sent periodically until a node obtains enough topology information from the other satellites. This requires a satellite to pass and contact to each other satellite at least twice to calculate the relative orbit times. Once this information is obtained a satellite leaves Discovery Mode and enters Power-save Mode.

Power-Save Mode

Power-save mode is used, as the name implies, to save power when no communication is possible for a long time. This mode is entered when a node determines that it will not be in communication range with another satellite for at least one minute. While in this mode a satellite will still respond to discovery mode broadcast packets, but will not generate them. A satellite leaves power save mode and enters active mode when it expects to be in communication range of another satellite within the next minute.

Active Mode

Active mode is used to exchange both link state information and data packets. A node enters active mode when it anticipates coming within communication distance of another satellite in the next minute. Active mode is further divided into three phases: handshake, link state exchange, and data exchange.

Active nodes begin in the handshake phase. During the handshake phase they periodically broadcast and respond to special handshake packets. Once a node has received a handshake response from another satellite, both satellites know they are in communication range and proceed to the link state exchange phase.

During the link state exchange phase the satellites send the contents of their link state table to each other. This information is necessary to the operation of the routing protocol, so it is transferred before any actual data. Since this information is only used within the routing protocol, it is all considered protocol overhead.

To minimize this overhead, the information transmitted during the link state exchange phase is kept intentionally small. Each satellite has a unique identifier (a small integer) assigned during construction that is used to describe the endpoints of a connection. The type of both of the satellites is also exchanged. This allows the satellites to differentiate between sensors, LBM, and HBM so data can be routed through the most appropriate channel. The last time the satellite was "seen" is exchanged. This is the time that the satellites last came within range of one another and started to communicate. The duration of the exposure is exchanged and used in conjunction with the radio bit rate to calculate the maximum amount of data that can be transferred during a connection. The final piece of data is the time until the two satellites will be in range again. The link state information is required for power save mode and routing to work correctly.

After all the link state information has been exchanged for every satellite in the network, the node begins the data exchange phase. In this phase any data that has been previously queued the peer satellite is transmitted. This includes both data generated at the node and data the node has received and agreed to forward on behalf of another node in the network. The receiving node acknowledges each packet successfully received, which facilitates increased reliability in the network.

The data exchange phase ends when the satellites are no longer in communication range. At that point the individual satellites will enter power save mode or active mode as appropriate.

Routing

Routing is the process of determining which path, from source to destination, the data should take to traverse the network. This routing algorithm considers four important pieces of information when routing a packet: the type of packet (command and control or data), the type of satellite (HBM vs. LBM), the remaining capacity of the connection, and the projected amount of time to transmit the data. More specifically, the algorithm considers all possible paths for the current data and select the set of paths that result in the data arrive as soon as possible at its destination. It is possible, especially if the quantity of data to transfer is large, that the data gets split into smaller pieces that are routed independently. This allows the real possibility that each piece of data takes a separate path through the network.

Simulator

Understanding the performance implications of a protocol before deploying it is critical, especially in a network composed of satellites. As part of this work, a discrete event simulator was developed to help analyze the protocol.

The simulation engine has a few important features that enable different scenarios. First, each satellite in the simulation is given its own three line element (TLE). TLEs describe the orbit of a satellite. There are a number of publically available TLEs for existing CubeSat satellites. Using actual CubeSat orbits improves the realism of the simulations.

The simulation also provides the ability to adjust both the both the bit rate and range of the radio used to communicate between the satellites. This enables us to understand the minimum performance requirements for the satellite-satellite radio.

The simulator also includes an implementation of the aforementioned protocol. This implementation is able to transfer data from orbiting sensor nodes to the HBM, measuring the performance of protocol. The implementation was also designed to be as separate from the simulator as possible, to make it much easier to port to a CubeSat platform.

Initial Protocol Performance

The first performance metric analyzed was the theoretical benefit of using a satellite constellation with aggregated downlink capacity. The theoretical results consider only the maximum exposure time in a year between a sensor satellite and both the ground station and a HBM and the amount of data that can be transmitted during those intervals. Figure 1 shows this analysis:

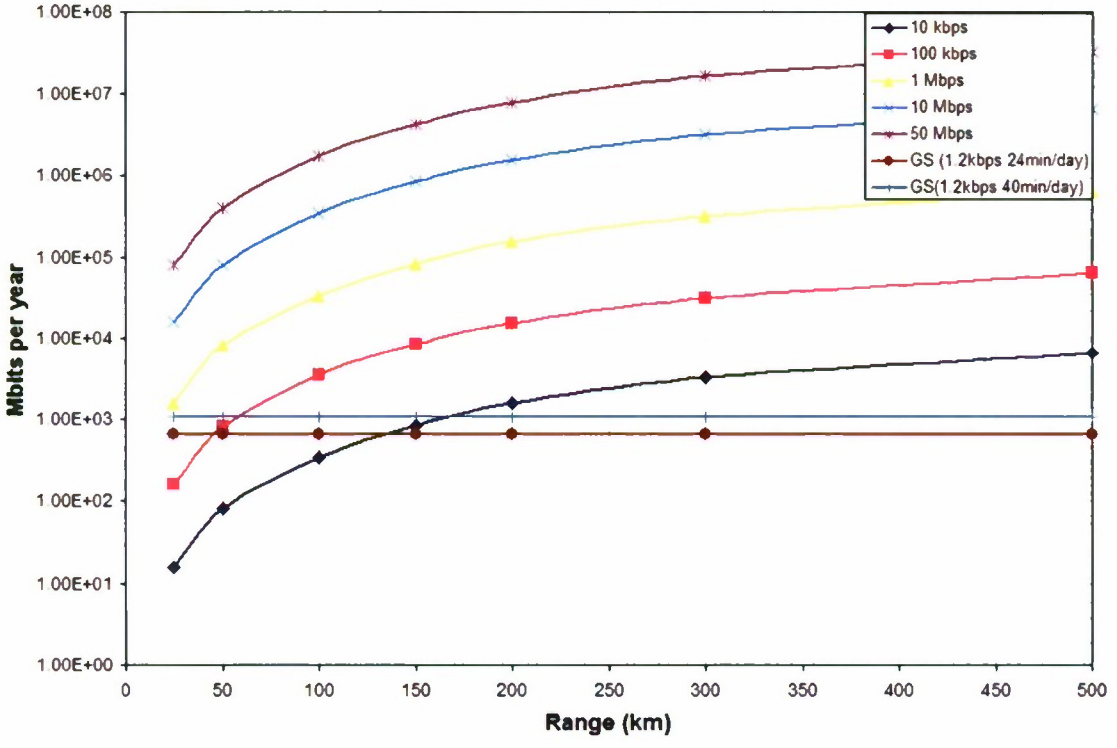


Figure 1

The curved lines are satellite-to-satellite capacities, while the straight lines are direct to ground station. The different lines represent different radio bit rates. The x-axis is the range of the satellite-to-satellite radios. Notice that a 10kbps radio with a range of 200km is the minimum necessary to improve on the existing direct-to-ground station communication.

Figure 2 analyzes the overall throughput of the network per year, given a variable number of HBM satellites and fixed radio performance (1mbps, 150km range). It shows that total network capacity increases almost linearly as additional HMBs are added to the network.

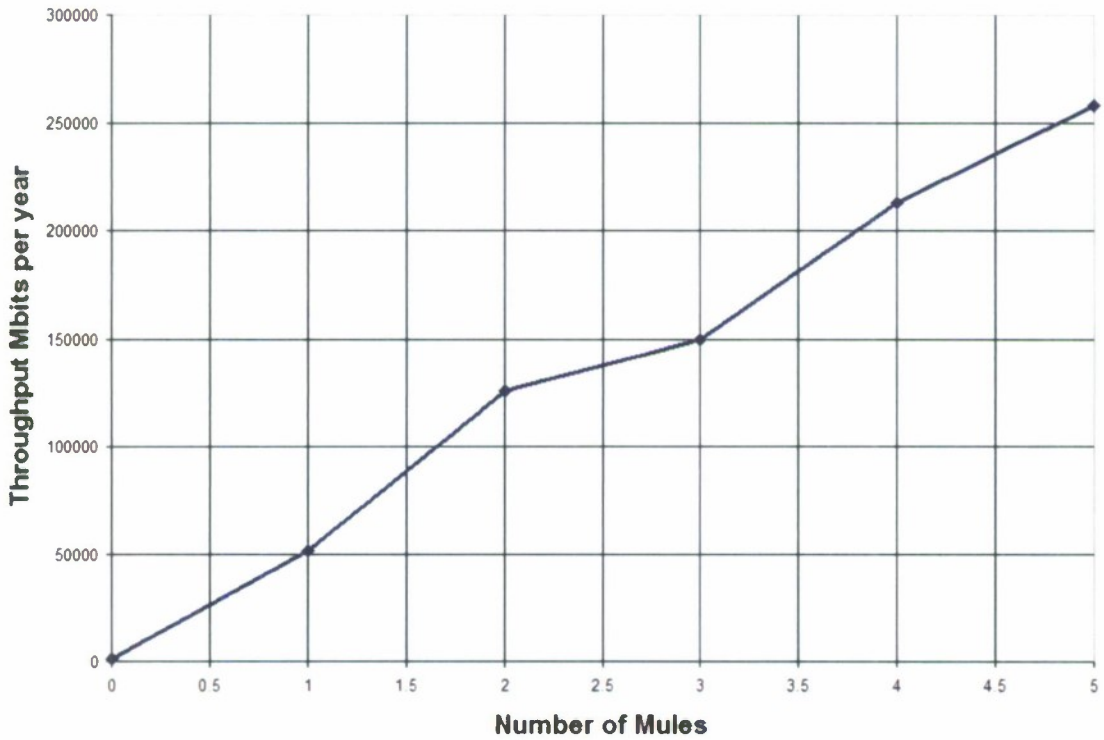


Figure 2

Another important performance consideration is the amount of time it takes to get data back to earth from the network. There are two considerations here. The first is the time is takes, and the second is the amount of data transferred. Table 1 shows the maximum latency, or the length of time, until the data arrives given one, two, and three sensors in the network.

1 Sensor / 1 HMB	2 Sensors / 1 HBM	3 Sensors / 1 HMB
78 Days	41 Days	20 Days

Table 1: Maximum latencies given a varying number of sensor nodes.

Figure 3

With only one sensor talking to one HBM, 78 days occur between each connection and therefore it can only send data back to the ground station once every 78 days. The two and three sensor times get drastically better making the maximum latency reduce to a maximum of 20 days.

These results seem poor when comparing the protocol to the maximum latency of communicating directly with the ground station of one day, but it is also important to take into account the amount of data that can be transferred during these times. Figure 3 shows the average amount of data that can be transferred per day depending on the bit rate of the communication between the sensors. As can be seen it takes only a communication speed of 5 kbps in the worse case to send more data per day on average than is currently possible, despite the much larger latency.

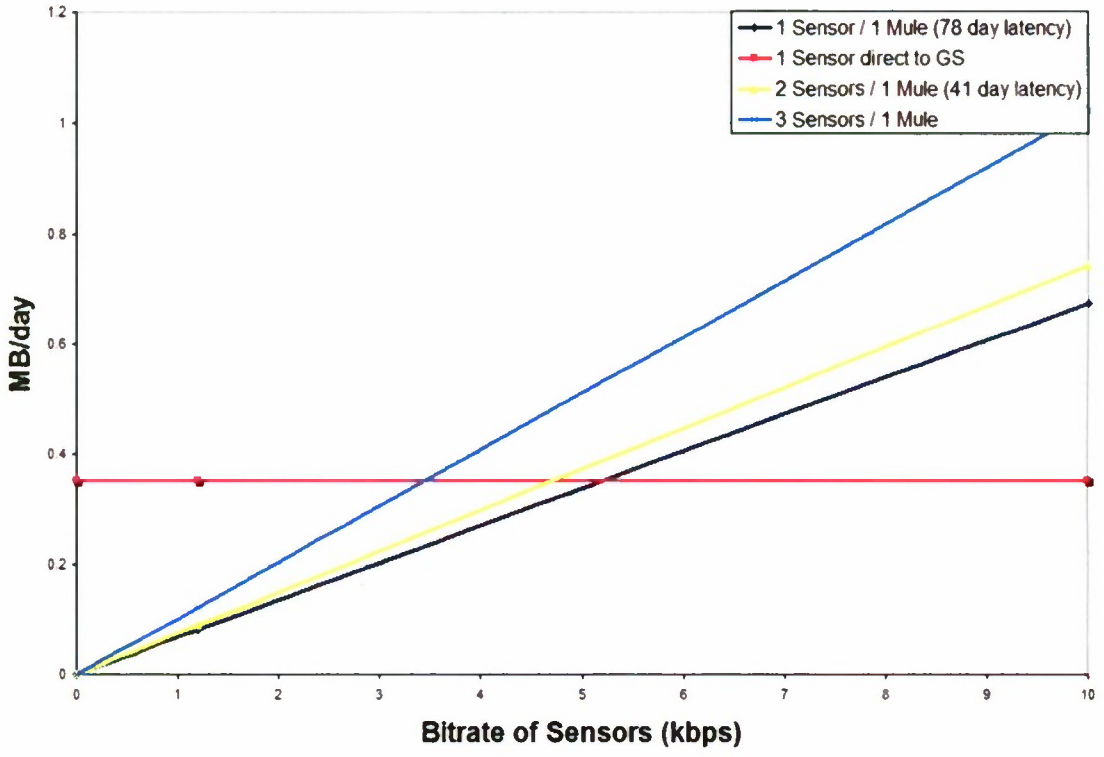


Figure 3

This discrepancy between latency and bandwidth underscores the importance of classifying data before sending it. 78 days is too long for command and control data, but isn't unreasonable for bulk data.

Future Work

Simulator

The simulator created gives a good idea of how the protocol will perform, but is by no means perfect. The ground station side of the simulator is not yet complete, so attempting to measure the data going from the HBMs down to the ground stations can not yet be tested and must be left to theoretical evaluation. This is an important aspect to be considered before attempting to build prototypes of the system so in the next revision of this project it is important to incorporate this into the simulator.

Protocol

As in any new protocol there are many places for improvement and optimization. Some future details that could be improved include creating a more advanced medium access protocol in order to more fully utilize the available bandwidth. One such possibility is to make the transfer of data fully synchronized to further limit the number of collisions and achieve near 100% bandwidth use. The protocol can also be extended to work with larger varieties of satellites and could be used to connect, pico satellites, LEO satellites, and even GEO satellites into a larger network. This is beyond the scope of this work, but would be an interesting and beneficial future addition.

Conclusion

In conclusion the protocol begun by this work is a plausible and effective solution to creating a sensor network in space, where the orbits of the satellites are predictable and the density is sparse. This work supports these conclusions through both theoretical analysis and simulation results. The preliminary throughput simulations show improvement over what is currently in use in the CubeSat project.

Contributors

Dr. John M Bellardo and Mr. Trevor Koritza were the primary contributors for this work and the final report.

Communication within Satellite Constellations

Project Investigator:

John M. Bellardo
Department of Computer Sciences
California Polytechnic State University
San Luis Obispo, CA

Communication within Satellite Constellations

Introduction

The CubeSat's small, standardized form-factor has been one of its biggest assets. It allows the satellites to be flown economically, enabling a myriad of applications and experiments that were previously not conducted due to high launch costs. However, this same form-factor poses one of the biggest engineering challenges to designing, building, and deploying CubeSats. The size and power constraints that result from such a small enclosure limit the complexity of the satellite's payload and corresponding experiments.

There are two general approaches to working within these constraints. The first is to push the miniaturization envelope, creating ever-smaller payloads. The second approach is to fly multiple satellites in a constellation (or cluster) that cooperate with each other, each performing a piece of the overall mission. To date the cluster approach has been rarely used with extremely small satellites, however the operational concept is well known and discussed amongst CubeSat developers.

Current CubeSat communication protocols are built on the assumption that the ground station communicates directly with the satellite. This simplifies protocol design in a number of ways. For instance, if the Cal Poly ground station has a packet to send to a CubeSat (CP3), it simply waits for CP3 to pass overhead and sends the packet. This is easier than sending the CP3 packet to the first satellite that passes overhead, CP4, and then depending on CP4 to relay the packet to CP3. This relaying capability is a fundamental component of cluster communication.

Satellites within a constellation, especially those launched in different P-PODs, will be in periodic communication range of each other as determined by their orbits. This periodic connectivity requires a dramatically different routing strategy than is employed in fixed, "always on" connections. Existing CubeSat ground stations resolve periodic connectivity through the use of human operators and orbital predictions. The operators use the predicted time of the next pass, prepare commands for the satellite in advance, and then perform the actual communication as the satellite passes overhead. Satellite constellations, on the other hand, have no onboard human operator responsible for communication. All the scheduling, relaying, and routing must be implemented by onboard software.

Initial development of this networking software, the corresponding protocols, and the necessary tools to evaluate them before flight, was the focus of this work. The remaining sections in this report provide an overview of the work done and the work yet to do in four important areas, including how this work aligns with The Office of Naval Research (ONR) goals, the communication protocol, the protocol simulator, and initial protocol analysis.

Alignment with ONR Goals

Satellite constellations can be used to solve a number of different and important problems. The research in this proposal is a fundamental enabling technology for constellations. This section describes how constellation technology could be applied to three different categories of specific interest to ONR and DOD: national security applications, command and control / communication, and intelligence / reconnaissance.

This work has two primary national security applications. First, a large part of national security is ensuring that our critical infrastructure is protected against failure when attacked. While it is beyond the scope of this work to identify exactly what constitutes critical infrastructure, this research will develop and evaluate different technology that provides the most basic form of failure protection, redundancy. In addition, advanced satellite technology is considered a strategic advantage, and this research will advance the state of the art in satellite communication technology.

This research also falls squarely into the ONR's interests of command and control, computers, and communication. The broad categories of computers and communication are easily meet by the nature of the research. Specific example applications include deploying a disruption tolerant satellite network that can carry mission critical data securely (command and control, etc). The tolerance is obtained from the redundancy in the system, however the security goal is not as obvious. The underlying technology needs to be designed and tested such that is operates correctly even when under computer-based attacks, like Denial-of-Service attacks, that don't cause any physical harm to the satellite.

Finally, the ability to amortize the cost of a single, high bandwidth link across many satellites makes this technology ideal for intelligence, surveillance, and reconnaissance applications when it is necessary to deploy a large number of resource constrained sensing nodes and related infrastructure.

Protocol

The first step in designing a routing protocol is determining the overriding behavioral categories for the protocol. These categories have a pronounced impact on the details of the design. In multi-hop wireless networks, there are two decisions to make: Should the protocol be proactive or reactive and should the protocol be link-state or distance vector.

Proactive protocols determine the path through the network before any packet is actually transmitted. This is used, and works very well, in static networks (like the Internet) and multi-hop routing protocols for networks whose mobility pattern results in nearly static networks.

The decision was made to use a proactive protocol for satellite cluster networks because of two factors. First, the orbits are predictable. The satellites themselves

can make high quality predictions of when they will come in communication range of each other. This makes the network properties closer to static than dynamic. The second reason is that exposure time, the time when one satellite is in communication with another satellite, is relatively small and reactive protocols tends to have much higher initial overhead in every exposure period. These two factors lead us to design a proactive protocol.

The second major design decision is selecting between link-state and distance vector. In a link-state protocol each node builds a virtual map of the entire network, and then uses graph algorithms to select the best route through the network. In a distance vector network each node only learns about the nodes its neighbor can reach, and then selects the best neighbor for the current packet. Given the disconnected nature of CubeSat constellations, we decided that a link-state protocol is more appropriate. It enables the protocol to use the aforementioned exposure predictions in a straightforward manner. A distance vector protocol would have also resulted in much higher protocol overhead due to the dynamic network.

The link state table, which holds the information necessary for the routing protocol, contains more information than a traditional link state protocol. In addition to storing details about which satellites are within communication range of each other, it also stores information that enables the satellites to predict when they will no longer be in range, and when they with subsequently fly in range again.

Since the link state information is reasonable concise, sending it between satellites enables each satellite to predict exposure time and transmission opportunities for the entire network.

Types of Nodes

This work divides the responsibilities of network node into three different categories: Low Latency Mule (LLM), High Bandwidth Mule (HBM), and Sensor. The LLM allows small amounts of information to be transmitted between the Sink and the sensors. This is primarily intended for command and control data from the ground station to the satellites. Since ground station exposure is small, roughly 15-20 minutes per pass with only a few passes a day, this connection is not able to transmit large quantities of data.

The HBM primarily used for bulk data transfers, enabled by deploying a higher powered and larger satellite in capable of higher downlink bandwidth. The primary use of the HMB is to move large amounts of payload data, like images or complex data tables, from sensors down to Earth. The HBM has a longer period between exposures to each satellite, but stays in range of that satellite for several hours each time. The HBM will also have a much higher bandwidth connection with the sensors because the range will be much smaller and with less atmospheric interference than from a sensor straight to the ground station. In addition to increased bandwidth, this can also save power on the sensor because each sensor will be sending the data a shorter distance with a lower power radio.

The HBM will then relay this information down to the ground station much faster than an individual sensor can. This is possible because the HBM is a larger satellite with more power and a larger antenna, which enables it to achieve higher data rates than a sensor can. In the case of the low earth orbits (LEO; like CubeSats), the sensors see the ground station 40 minutes a day, but will only see the HBM every one or two months.

The final type of node is the sensor itself. The sensors can be gathering any sort of data from the payload to be transmitted to the ground station.

Satellite Modes

Each node in the network may be in one of three different modes: discovery, power-save, and active.

Discovery Mode

Discovery mode is used to learn about the orbits and exposure times of peer satellites in the network. This mode is entered immediately after satellite deployment. In this mode the satellite sends out broadcast packets to discover all of the satellites in the network. A broadcast packet is sent periodically until a node obtains enough topology information from the other satellites. This requires a satellite to pass and contact to each other satellite at least twice to calculate the relative orbit times. Once this information is obtained a satellite leaves Discovery Mode and enters Power-save Mode.

Power-Save Mode

Power-save mode is used, as the name implies, to save power when no communication is possible for a long time. This mode is entered when a node determines that it will not be in communication range with another satellite for at least one minute. While in this mode a satellite will still respond to discovery mode broadcast packets, but will not generate them. A satellite leaves power save mode and enters active mode when it expects to be in communication range of another satellite within the next minute.

Active Mode

Active mode is used to exchange both link state information and data packets. A node enters active mode when it anticipates coming within communication distance of another satellite in the next minute. Active mode is further divided into three phases: handshake, link state exchange, and data exchange.

Active nodes begin in the handshake phase. During the handshake phase they periodically broadcast and respond to special handshake packets. Once a node has received a handshake response from another satellite, both satellites know they are in communication range and proceed to the link state exchange phase.

During the link state exchange phase the satellites send the contents of their link state table to each other. This information is necessary to the operation of the routing protocol, so it is transferred before any actual data. Since this information is only used within the routing protocol, it is all considered protocol overhead.

To minimize this overhead, the information transmitted during the link state exchange phase is kept intentionally small. Each satellite has a unique identifier (a small integer) assigned during construction that is used to describe the endpoints of a connection. The type of both of the satellites is also exchanged. This allows the satellites to differentiate between sensors, LBM, and HBM so data can be routed through the most appropriate channel. The last time the satellite was "seen" is exchanged. This is the time that the satellites last came within range of one another and started to communicate. The duration of the exposure is exchanged and used in conjunction with the radio bit rate to calculate the maximum amount of data that can be transferred during a connection. The final piece of data is the time until the two satellites will be in range again. The link state information is required for power save mode and routing to work correctly.

After all the link state information has been exchanged for every satellite in the network, the node begins the data exchange phase. In this phase any data that has been previously queued the peer satellite is transmitted. This includes both data generated at the node and data the node has received and agreed to forward on behalf of another node in the network. The receiving node acknowledges each packet successfully received, which facilitates increased reliability in the network.

The data exchange phase ends when the satellites are no longer in communication range. At that point the individual satellites will enter power save mode or active mode as appropriate.

Routing

Routing is the process of determining which path, from source to destination, the data should take to traverse the network. This routing algorithm considers four important pieces of information when routing a packet: the type of packet (command and control or data), the type of satellite (HBM vs. LBM), the remaining capacity of the connection, and the projected amount of time to transmit the data. More specifically, the algorithm considers all possible paths for the current data and select the set of paths that result in the data arrive as soon as possible at its destination. It is possible, especially if the quantity of data to transfer is large, that the data gets split into smaller pieces that are routed independently. This allows the real possibility that each piece of data takes a separate path through the network.

Simulator

Understanding the performance implications of a protocol before deploying it is critical, especially in a network composed of satellites. As part of this work, a discrete event simulator was developed to help analyze the protocol.

The simulation engine has a few important features that enable different scenarios. First, each satellite in the simulation is given its own three line element (TLE). TLEs describe the orbit of a satellite. There are a number of publically available TLEs for existing CubeSat satellites. Using actual CubeSat orbits improves the realism of the simulations.

The simulation also provides the ability to adjust both the both the bit rate and range of the radio used to communicate between the satellites. This enables us to understand the minimum performance requirements for the satellite-satellite radio.

The simulator also includes an implementation of the aforementioned protocol. This implementation is able to transfer data from orbiting sensor nodes to the HBM, measuring the performance of protocol. The implementation was also designed to be as separate from the simulator as possible, to make it much easier to port to a CubeSat platform.

Initial Protocol Performance

The first performance metric analyzed was the theoretical benefit of using a satellite constellation with aggregated downlink capacity. The theoretical results consider only the maximum exposure time in a year between a sensor satellite and both the ground station and a HBM and the amount of data that can be transmitted during those intervals. Figure 1 shows this analysis:

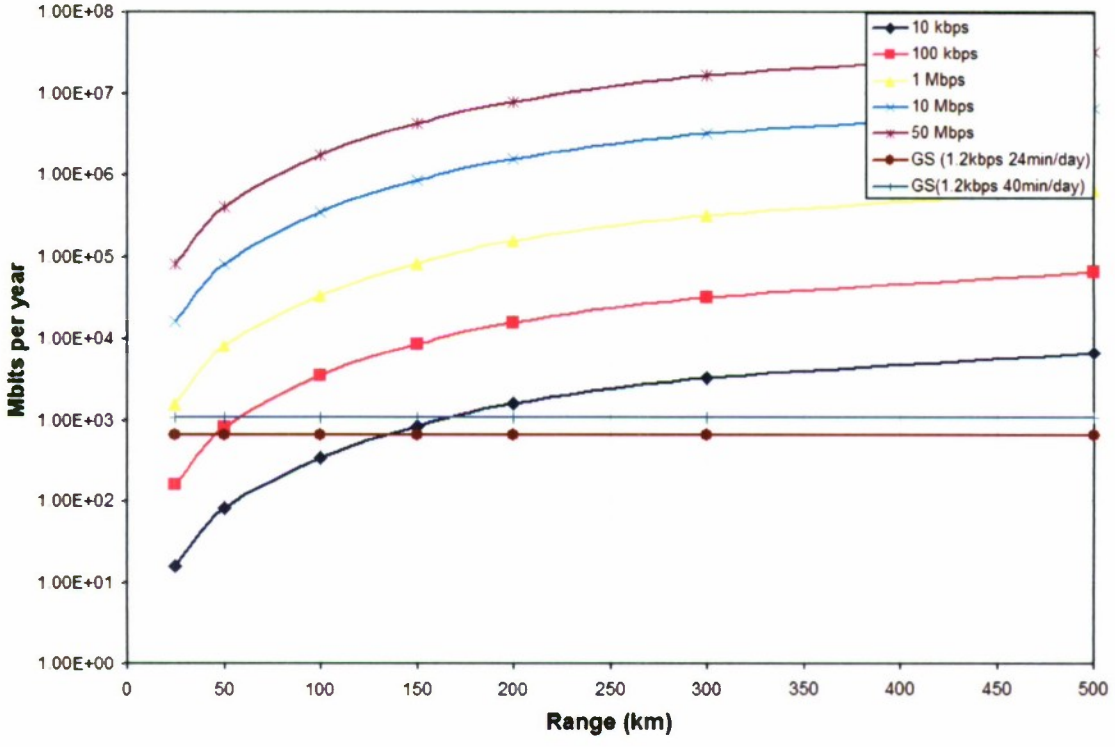


Figure 1

The curved lines are satellite-to-satellite capacities, while the straight lines are direct to ground station. The different lines represent different radio bit rates. The x-axis is the range of the satellite-to-satellite radios. Notice that a 10kbps radio with a range of 200km is the minimum necessary to improve on the existing direct-to-ground station communication.

Figure 2 analyzes the overall throughput of the network per year, given a variable number of HBM satellites and fixed radio performance (1mbps, 150km range). It shows that total network capacity increases almost linearly as additional HMBs are added to the network.

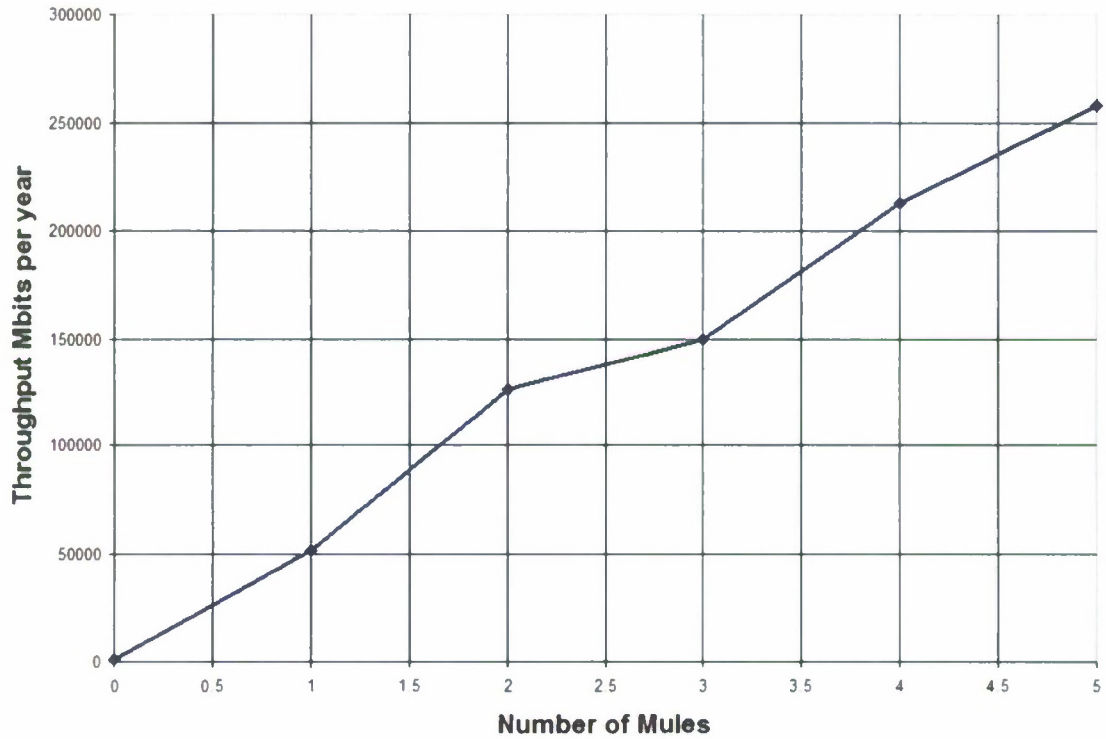


Figure 2

Another important performance consideration is the amount of time it takes to get data back to earth from the network. There are two considerations here. The first is the time is takes, and the second is the amount of data transferred. Table 1 shows the maximum latency, or the length of time, until the data arrives given one, two, and three sensors in the network.

1 Sensor / 1 HMB	2 Sensors / 1 HBM	3 Sensors / 1 HMB
78 Days	41 Days	20 Days

Table 1: Maximum latencies given a varying number of sensor nodes.

Figure 3

With only one sensor talking to one HBM, 78 days occur between each connection and therefore it can only send data back to the ground station once every 78 days. The two and three sensor times get drastically better making the maximum latency reduce to a maximum of 20 days.

These results seem poor when comparing the protocol to the maximum latency of communicating directly with the ground station of one day, but it is also important to take into account the amount of data that can be transferred during these times. Figure 3 shows the average amount of data that can be transferred per day depending on the bit rate of the communication between the sensors. As can be seen it takes only a communication speed of 5 kbps in the worse case to send more data per day on average than is currently possible, despite the much larger latency.

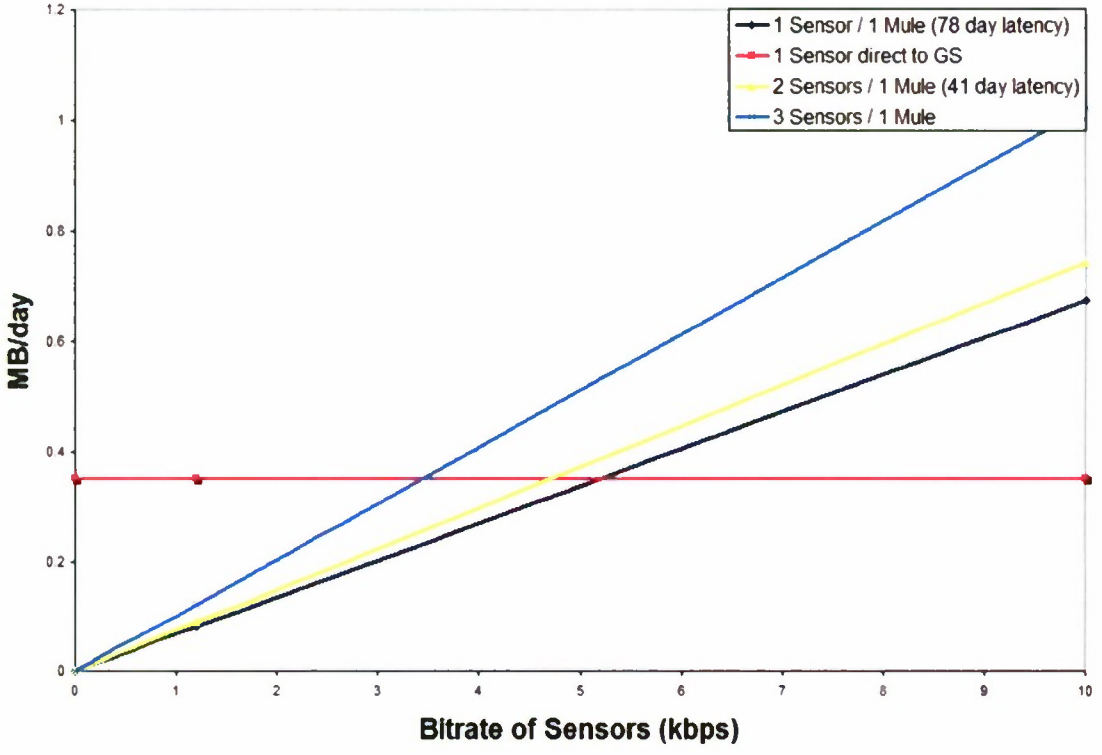


Figure 3

This discrepancy between latency and bandwidth underscores the importance of classifying data before sending it. 78 days is too long for command and control data, but isn't unreasonable for bulk data.

Future Work

Simulator

The simulator created gives a good idea of how the protocol will perform, but is by no means perfect. The ground station side of the simulator is not yet complete, so attempting to measure the data going from the HBMs down to the ground stations can not yet be tested and must be left to theoretical evaluation. This is an important aspect to be considered before attempting to build prototypes of the system so in the next revision of this project it is important to incorporate this into the simulator.

Protocol

As in any new protocol there are many places for improvement and optimization. Some future details that could be improved include creating a more advanced medium access protocol in order to more fully utilize the available bandwidth. One such possibility is to make the transfer of data fully synchronized to further limit the number of collisions and achieve near 100% bandwidth use. The protocol can also be extended to work with larger varieties of satellites and could be used to connect, pico satellites, LEO satellites, and even GEO satellites into a larger network. This is beyond the scope of this work, but would be an interesting and beneficial future addition.

Conclusion

In conclusion the protocol begun by this work is a plausible and effective solution to creating a sensor network in space, where the orbits of the satellites are predictable and the density is sparse. This work supports these conclusions through both theoretical analysis and simulation results. The preliminary throughput simulations show improvement over what is currently in use in the CubeSat project.

Contributors

Dr. John M Bellardo and Mr. Trevor Koritza were the primary contributors for this work and the final report.

Construction of an Efficient Microbial Peptidase Delivery System to Treat Celiac Disease and Maximize Human Health

Project Investigator:

Michael W. Black
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA

Construction of an Efficient Microbial Peptidase Delivery System to Treat Celiac Disease and Maximize Human Health

I. Background:

Celiac disease (coeliac disease, celiac sprue, CD) is an auto-immune enteropathy resulting from sensitivity to epitopes contained within gluten, the dominant protein portion of many common grains including wheat, barley and rye. Exposure to these proteins leads to inflammation and widespread atrophy of intestinal epithelia in individuals with CD. Symptoms of CD are generally similar to other chronic gastrointestinal disorders such as Crohn's disease and Ulcerative colitis, thus it is commonly mis-diagnosed in both children and adults (Picarelli et .al., 2000; Corazza, 1996) . While the symptoms of CD may be tolerable or even sub-clinical in some individuals, the associated conditions and long-term effects of unmanaged CD can be severe or even fatal. These include; skin disorders, ulcerative colitis, lymphocytic colitis, liver disease, thyroid disease, anemia, diabetes, bone metabolism, malignancy and even neurological and psychological disturbances (Ciclitira et. al., 2005). Since these grains are ubiquitous in the average western diet, ths disease greatly detracts from the quality of life. The prevalence of CD is approximately 1:100 among the Caucasian population, making it the most commonly diagnosed chronic gastrointestinal disorder (Ciclitira et. al., 2005). The goal of this project is to develop a strain of probiotic lactic acid bacteria that will colonize the small intestine and supplement the intestine with prolyl peptidase activity. This activity will allow for increased break down and absorption of dietary proteins and removal of potentially harmful metabolites.

Rationale for Prolyl Peptidase Supplementation

Proline is the most abundant amino acid in beta-casein and a major component of gluten. Proline is unique in that the amine group involved in peptide bonding is covalently bonded to its side chain. This secondary amine produces tight kinks in protein secondary structure such that the linkages are resistant to common digestive enzymes. The result is an accumulation of the proline rich peptides, most of which are not absorbed by the body and are excreted through the urine or feces. Incomplete utilization of these proteins not only predisposes people to the development of CD, it also limits the nutritive value of two staples of the western diet: milk and bread.

Prolyl Peptidases in Bacteria

Prolyl peptidase activity is conspicuously absent from the human digestive tract. These enzymes however can be found throughout the microbial community, including those that are residents of the human gastrointestinal tract. Presently the most abundant incidence of prolyl peptidase activity can be found in the genus *Lactobacillus*. Due to their inability to synthesize many amino acids, LABs are dependent upon extracting their amino acid requirements from the environment. For this reason LABs (especially *Lactobacilli*) maintain an extremely complex and diverse proteolytic system. Prolyl peptidase activity has already been characterized in several species in the LAB family including, *Lactobacillus helveticus*, *Lactobacillus sakei*, *Lactobacillus reuteri* and *Lactococcus lactis* (Degraeve et. al., 2003; Sanz et. al., 2001; Rollan et. al., 2001; Xu et. al. 2001). Prolyl peptidase expression has also been documented in bacteria outside of the LAB family such as, *Flavobacterium meningosepticum*, *Sphingomonas capsulate*, *Myxococcus xanthus* (Shan et. al. 2004). We have confirmed that prolyl peptidase activity is present in several species of *Lactobacillus* including *L. helveticus*, *L. reuteri*, *L. casei*, *L. acidophilus* and *L.*

delbruckei. The procedure developed to assay prolyl peptidase activity uses the substrate Proline-*para*-Nitroaniline (Pro-*p*NA) to measure proline hydrolysis at the carboxy side of the proline. Cleavage at this site releases the yellow *p*NA product that can be quantitatively measured by spectrometry (λ = 405 nm). We have observed the prolyl peptidase activity in lactic acid bacteria is primarily confined to the cytoplasmic milieu of the cell. Cell hydrolysates (prepared by sonication) showed substantially higher prolyl peptidase activity than intact cells indicating that very little prolyl peptidase is secreted into the extracellular environment (table 1). We will continue these experiments on cultures that demonstrated differences in protein expression when grown in media containing milk components to determine if there is a correlation between this proteolytic activity and the presence or absence of supplemented milk components.

Table 1: Qualitative and Quantitative Analysis of Ala-Pro-pNA hydrolysis by cell lysates.

Strain	Genus species	OD @ 0 min	15 min	30 min	Qualitative
Blank	Blank	0.096	0.106	0.112	-
GR-1	*Lactobacillus rhamnosus*	1.874	1.921	2.01	+/-
7469	*Lactobacillus rhamnosus*	1.962	2.047	2.179	+/-
LC10	*Lactobacillus casei*	1.128	1.104	1.092	-
RO49	*Lactobacillus rhamnosus*	0.878	0.868	0.866	-
MR220	*Lactobacillus helveticus*	1.325	1.521	1.767	++
4356	*Lactobacillus acidophilus*	2.012	3.313	4	++++
23272	*Lactobacillus reuteri*	1.927	2.348	2.776	+++
NCK388	*Lactobacillus helveticus*	2.118	2.454	3.041	+++
San	*Lactobacillus delbrueckei ssp. Lactis*	0.564	2.179	3.115	++++
MR120	*Lactobacillus delbrueckei ssp. bulgaricus*	1.133	1.492	1.914	+++
RO11	*Lactobacillus rhamnosus*	1.672	1.692	1.722	+/-
393	*Lactobacillus casei*	1.462	1.57	1.807	+

(Optical density measured at λ=405 nm)

Probiotic bacteria as enzyme delivery vehicles to manage CD

Bacteria that stabilize the intestinal tract and promote healthy digestion are termed probiotic. Although the term probiotic is relatively new the idea of beneficial bacteria is as old as the study of our defenses against harmful bacteria. Classically, probiotic bacteria are primarily members of the lactic acid family of bacteria (LAB). LABs are gram positive, non-spore forming bacteria that produce lactic acid from the fermentation of sugars. *Lactobacillus* is the largest genus of LABs consisting of more than 50 species (Stiles & Holzapfel,1997; Tannock, 2004). Lactobacilli are important for the production of foodstuffs. Most notable among these organisms are *Lactobacillus delbrueckii ssp. bulgaricus* and *Lactobacillus sanfranciscensis* used in the commercial production of yogurt and sourdough bread respectively. While their primary role in food production is the fermentation of sugars to lactic acid, non-commercially important LABs are also responsible for producing a wide variety of compounds beneficial to humans including; anti-microbial peptides, exopolysaccharides and other metabolites (Ross, Morgan & Hill, 2002).

An important role of probiotic bacteria in the human intestinal tract is the regulation of intestinal homeostasis. This involves protecting the mucosal epithelium from pathogens, delivery

of regulatory signals to the immune system and neuromuscular system, as well as inhibition of pathogenic bacteria through nutrient competition, binding translocation and production of anti-microbial peptides (Shanahan, 2004). The genus *Lactobacillus* has been shown to be effective in the treatment and prevention of gastrointestinal disorders such as infectious diarrhea and inflammatory bowel disease (Saavedra, 2000; Servin, 2004; Shanahan, 2004).

Several of the bacteria discussed in the previous section that possess prolyl peptidases are able to colonize the intestines, including the small intestine. A bacterial strain that can colonize the small intestine and constitutively express and secrete prolyl peptidase enzymes at a high level of activity would effectively supplement the area with enzymes capable of detoxifying gluten. Furthermore, if a probiotic bacterium such as those described above were used, the individual would not only alleviation of CD symptoms due to prolyl peptidase activity, but also enjoy the health benefits attributed to intestinal colonization by probiotics. Unfortunately, no probiotic microorganism yet studied expresses prolyl peptidase enzymes on the outer surface of the cell. However, due to the advancements in molecular biology technology, it may now be possible to engineer strains capable of surface expression of prolyl peptidases. This hypothetical strain would be able to colonize the small intestine and express prolyl peptidases on the surface of the cell, thus effectively localizing enzyme activity at the site of celiac pathology. The remainder of this document describes the initial research undertaken to achieve this single goal, to provide the first non-dietary treatment option for celiac disease.

Construction of S-layer Directive Reporter/Expression Vectors

The microbial S-layer is a two-dimensional crystalline matrix that completely covers some members of the Eubacteria, and nearly all members of the Archaea, domains of life. One extensively studied S-layer protein is the SlpA protein of *Lactobacillus acidophilus*. The SlpA protein is composed of three components, a N-terminal leader sequence that directs secretion, a crystallizing domain and a C-terminal anchoring domain. The construction of several new S-layer detection and expression vectors has begun in our lab in order to assess the efficiency of the S-layer extractions, measure induction of the S-layer promoter system in *Lactobacillus acidophilus*, and to express endogenous prolyl peptidases at the outer surface of the cell (as a fusion with the S-layer protein) or secreted from the cell. The objective in re-localizing these peptidases to the exterior of the cell is to determine if the digestion of proline-rich proteins is enhanced, thus increasing nutrient availability and destroying potentially harmful immunogenic peptides.

II. Results

Construction of Expression Cassettes

To first determine the suitability of the expression system for localizing fusion proteins, such as those expressing prolyl endopeptidases, a green fluorescent reporter was used to track the localization of the protein. All components necessary to build the cassettes shown in figure 1 were amplified from extracted *Lactobacillus acidophilus* NCFM genomic DNA (ProL, Pro, Aterm and Term) or plasmid pMB293 (GFP) using the high fidelity pyrococcus-like enzyme, Phusion DNA polymerase (Finnzymes).

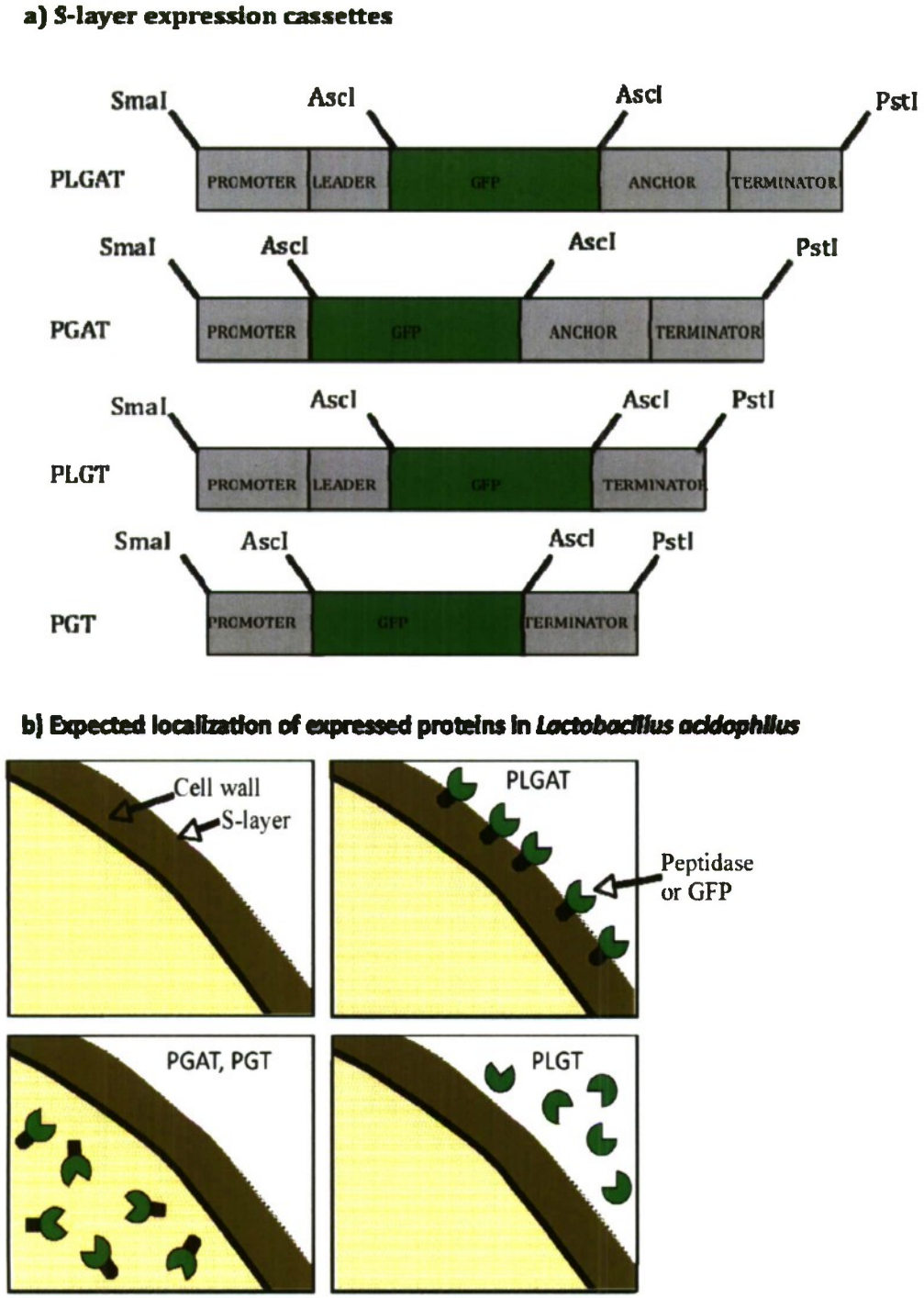


Figure 1: S-layer GFP expression cassettes and expected localization. (A) S-layer cassette constructs with engineered restriction sites for cloning and replacement of the GFP reporter with prolyl peptidases of interest (AscI). (B) the expected localization of the fusion protein product from each expressed cassette (A) in *L. acidophilus*.

Digestion independent cloning of S-layer GFP cassettes into pGKMCS

In order to efficiently clone the cassettes into pGKMCS, a digestion independent cloning (DIC) technique was developed. DIC allows for orientation specific cloning using sequence homology

between the insert and vector (see Figure 2). For DIC, insert and vector DNA containing homology at each end was mixed together in a Phusion PCR without primers and taken through 5 cycles of denaturation and extension.

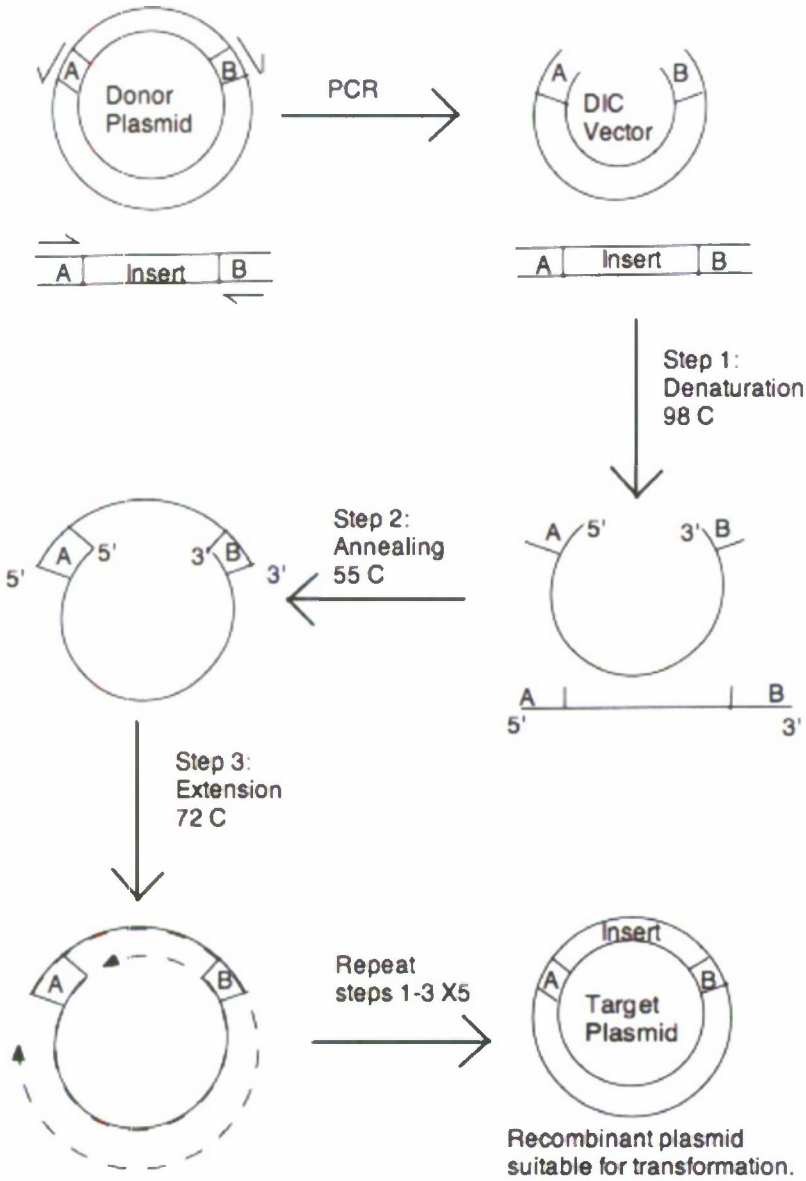
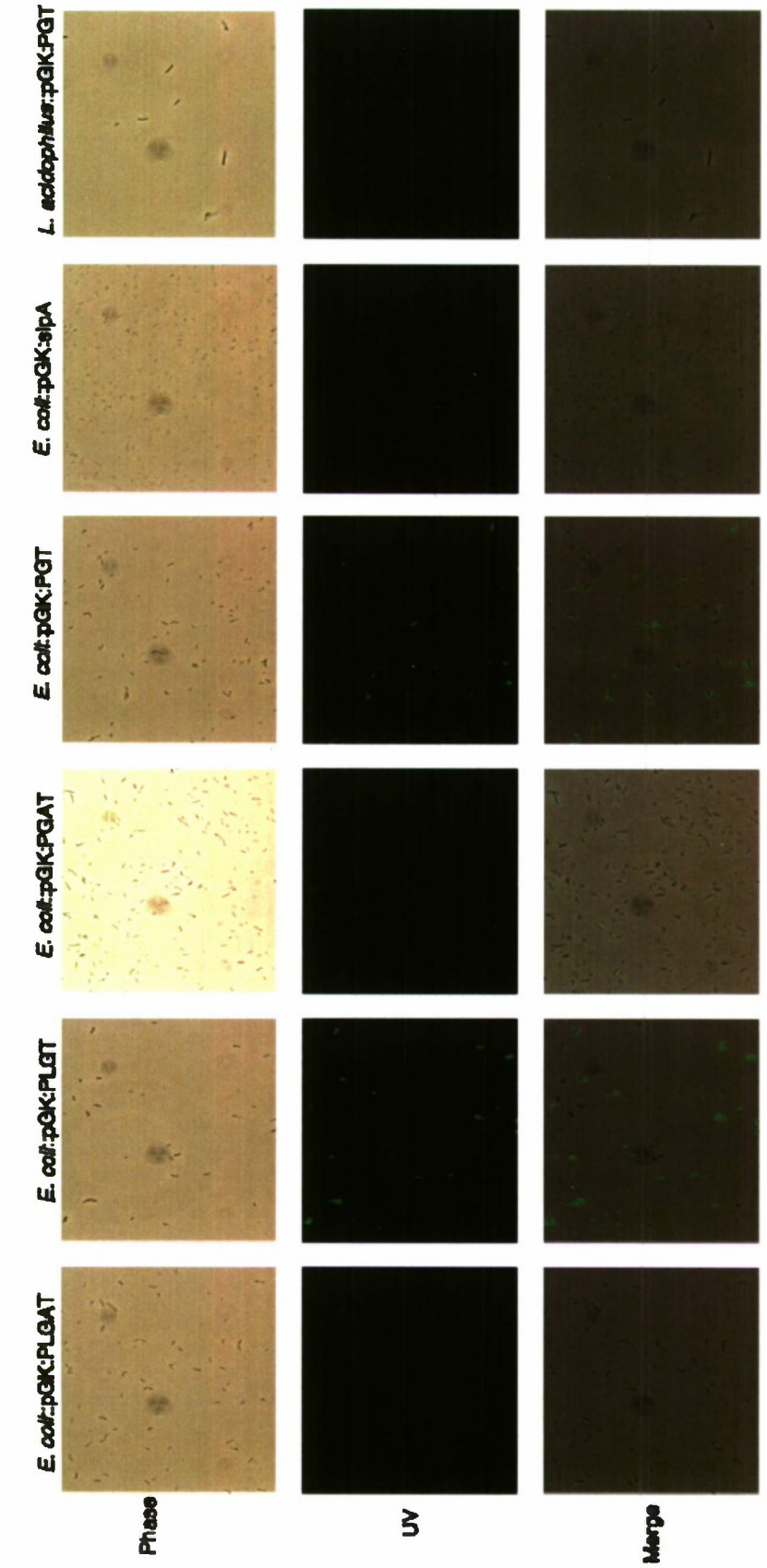


Figure 2: Digestion Independent Cloning (DIC) of expression cassettes into suitable shuttle vector for propagation and expression in *E.coli* and *Lactobacillus sp.*

A relatively high degree of success was found with this method, allowing the recovery of >100 potential clones for each cassette. Recovered clones were screened by CFU PCR to identify those that contained their respective inserts. After CFU PCR identified the clones, they were screened for GFP using fluorescence microscopy and western blot analysis (see Figure 3).


a)
Phase
UV
Merge
E. coli:pGK:PLGAT
E. coli:pGK:PLGT
E. coli:pGK:PGAT
E. coli:pGK:PGT
E. coli:pGK:slpA
L. acidophilus:pGK:PGT

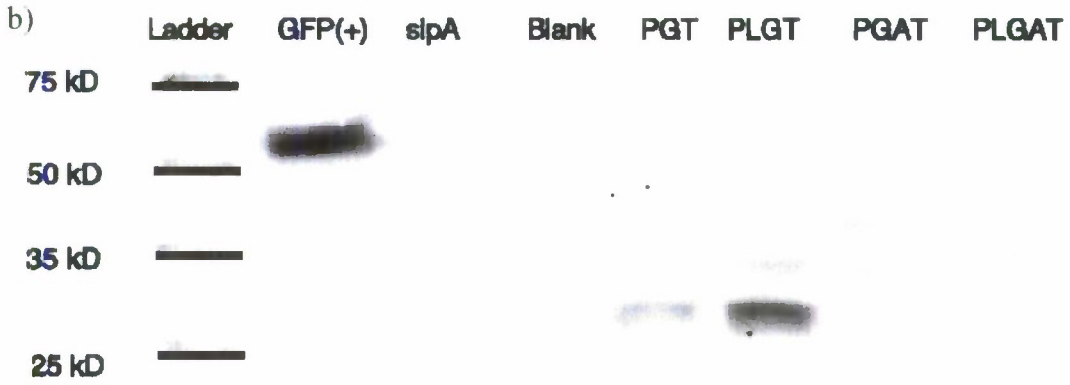


Figure 3: Assessment of expression from fusion cassettes in *E.coli*. (A) GFP expression assessed by fluorescence microscopy for plasmids carrying PLGAT, PLGT, PGAT, and PGT. (B) Western blot analysis of expression cassettes using anti-GFP antibodies. Expected sizes for each product are: GFP(+) from yeast expression plasmid = 52.8kD; PGT = 27kD; PLGT = 31kD; PGAT = 40kD; PLGAT = 45kD.

In order to be able to discern possible reasons why PLGAT is not expressed in *E. coli* and some of the expression cassettes had limited fluorescence under the microscope, the inserts from three plasmids pGKMCS:PLGAT, pGKMCS:PLGT and pGKMCS:PGT were sequenced. Sequencing results for the PLGAT cassette showed a frameshift deletion of two nucleotides at places 347 and 348 . The PLGT cassette acquired a single nucleotide deletion resulting in a frame-shift mutation at nucleotide 352. For the available sequence, the PGT insert appeared to be as expected, however, poor sequence data was obtained for the 5' end of each insert, including the sigma factor binding sites.

This investigation has yielded the successful construction of a broad host range shuttle vector with a robust multiple cloning sequence to allow relatively easy cloning and expression of DNA in *Lactobacillus* and other organisms in which the pGK12 plasmid will replicate. This work has also revealed toxicity of unknown origin which resulted in the acquisition of mutations in the inserts, despite the fact that a proofreading polymerase was used for each amplification. Furthermore, we have developed a novel cloning technique, which once optimized, may allow for direct cloning of these cassettes into *Lactobacillus*, thus bypassing the toxicity problem in *E. coli*. While it is likely that the anchor sequence plays a role in this toxicity, it cannot be the sole culprit as the PLGT insert was also mutated. Until it is determined whether or not DIC can be utilized for this application, inducible expression systems for *E. coli* was developed to prevent selection for mutants during cloning and allow for analysis of the *slpA* localization signals in both organisms.

Development of an Inducible Expression System to Evaluate Fusion Protein Toxicity

Inducible plasmid-based expression systems are often used to allow expression of inserts that may be toxic to their host. The previous section has shown that the expression and/or localization signals derived from the *slpA* gene of *Lactobacillus acidophilus* NCFM are toxic in *E. coli*, leading to the inability to recover clones bearing the expected inserts. In order to be able to effectively assess the efficacy of using the *slpA* localization signals to guide surface layer expression of prolyl peptidase activity, we selected the *E. coli* expression system, pET30, which

exploits the IPTG controlled expression to provide easily inducible expression of heterologous proteins. The current investigation describes the utilization of pET30 based plasmids to assess the viability of using *slpA* signals to drive prolyl peptidase localization in *E. coli*.

Initially, two pET30 based plasmids were constructed using the SlpA leader ("L"), a truncated version of the original SlpA anchor sequence ("S"), and the Xaa-prolyl peptidase from *L.reuterii* ("X"), named pET30:LS and pET30:LXS. In order to determine if toxicity is associated with either insert, growth was monitored after induction with IPTG. The pET30:LS plasmid showed a significant halt in growth after 15 minutes of induction and even a slight drop in OD600 after 1.5 hours compared to the uninduced control (see Figure 4). Interestingly, the pET30:LXS only exhibited only a small degree of growth impairment when compared to the uninduced control.

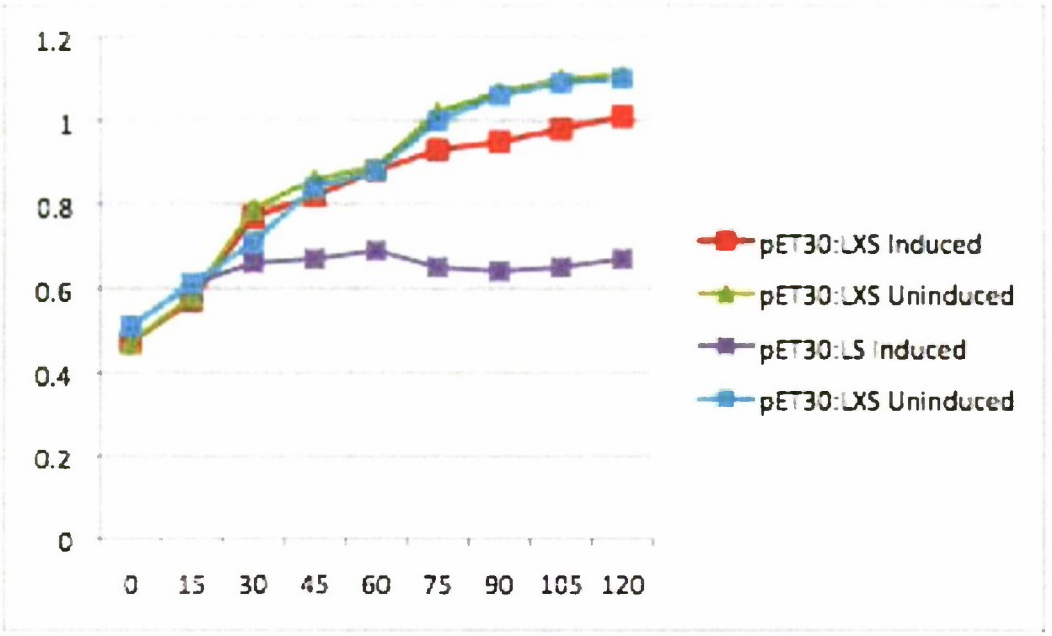


Figure 4: Assessment of toxicity associated with the SlpA leader and anchor sequences. Growth curves from *E.coli* harboring the pET30:LS or pET30:LXS plasmids. OD600 measurements were taken every 15 minutes from cultures either induced with IPTG or not.

The two pET30 plasmids were built in such a manner as to preserve the downstream His-tag as part of the inducible open reading frame, thus allowing for determination of full length expression of the inserts by detection of the His-tag using western blot. The His-probe analysis showed full-length LXS expression (101.1 kD) localized to the insoluble pellet fraction of the cell lysate. Expression of LS (12.7 kD) was much lower and was also localized to the pellet fraction (see Figure 5). No peptidase activity could be detected in whole cell lysates prepared from the induced pET30:LXS cultures despite the evidence of full-length expression of the insert by His-tag detection (data not shown).

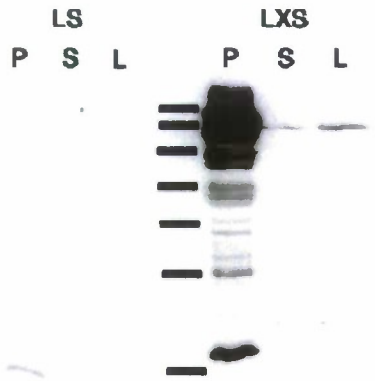


Figure 5: Localization of LS and LXS in *E.coli*. The C-terminal His-tag of the 1.5-hour induced gene products from cells harboring pET30:LS and pET30:LXS. L = lysate, S = cell supernatant (medium), P = insoluble pellet from lysate. Molecular weight standards (from bottom to top): 15kD, 25kD, 35kD, 50kD, 75kD, 100kD, 150kD. Expected sizes: LS = 12.7kD, LXS = 101kD.

To further investigate the possibility that these fusions were localized to inclusion bodies in the cells, induced cultures were examined by light microscopy. Distinctive morphological differences were observed between induced cultures of *E. coli* harboring pET30:LXS and uninduced cultures. Induced cultures showed drastically elongated cellular morphology with irregular staining. Uninduced cultures displayed the individual 1 uM by 2 uM rods characteristic of *E. coli* (see Figure 6).

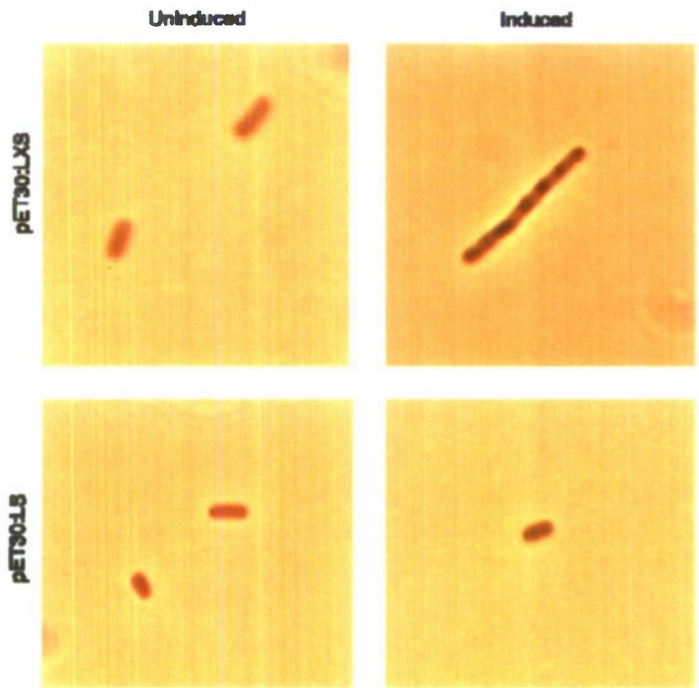


Figure 6: Morphological characteristics of induced vs. uninduced cultures. The presence of inclusion bodies were clearly visible in the pET30:LXS plasmid-bearing strain upon IPTG induction.

III. Discussion:

The extremely well characterized pET30 expression system allowed for the continued investigation of the potential for using *slpA* secretion and anchoring signals in fusion with peptidase genes to drive localization. The examination of these gene fusion products in *E. coli* was undertaken to provide conceptual evidence for the experimental design. The pET30:LS plasmid played a dual role in this investigation. This expression vector was initial constructed as a negative control to be used in the pET30:LXS peptidase activity assays, however it also allowed for direct assessment of toxicity associated with the S-layer localization components. Upon induction with IPTG the pET30:LS culture displayed no growth after 15 post-induction. This observation provides undeniable evidence that the *slpA* leader and/or anchor are extremely toxic in *E. coli*. The observed growth characteristics are consistent with the murein hydrolase activity previously prescribed to the full size anchor, but was thought to have been abolished by replacement of this component with the truncated anchor.

It was interesting to observe the distinctive difference in toxicity between the LS insert and the LXS insert. There were three possible explanations for the attenuated toxicity associated with LXS as compared to LS:

1) **The full-length LXS cassette may not be expressed in *E. coli*.** It was possible that the Xaa-Pro peptidase derived from *L. reuteri* may not be expressed in *E. coli* due to codon bias associated with the coding region of this gene. If a rare codon was used in this open reading frame then this might lead to truncation in *E. coli* and thus effectively delete the anchor domain from the gene product. If the anchor were indeed the source of toxicity, then this would abolish the lethal effects of this protein. In order to determine if this was the case, HRP conjugated NTA was used to probe for the C-terminal His-tag. Since the cloning scheme allowed for only the C-terminal His-tag of the pET30 expression vector to be in frame with the recombinant protein, only full length proteins would be detected using this method. The blot clearly shows evidence of expression of the entire insert yielding the expected 101.1 kD protein (Figure 3-4). Therefore, truncation cannot be the explanation for decreased toxicity associated with this insert as compared to LS.

2) **LXS is aggregating to form inclusion bodies in the cytoplasm of *E. coli*, thus is not able to elicit its toxic effects in the periplasmic space.** This hypothesis may be supported by observed differences in cellular morphology between induced and uninduced LXS cultures. The induced culture showed drastically different morphology and contained what may be inclusion bodies within the cells. Meanwhile the induced LS cells showed no morphological differences from the uninduced control (Figure 3-5). The His-tag probe results may also support this assertion, as the vast majority of protein was associated with the insoluble cell lysate pellet. By definition, inclusion bodies are comprised of insoluble protein aggregates and would therefore localize to the pellet fraction after cell lysis and centrifugation. An alternative explanation for the presence of the expressed protein in the insoluble fraction is that the anchor is functioning properly in *E. coli* and is anchoring within peptidoglycan layer of this organism. If the protein is being expressed as an inclusion body the enzyme may still be able to be purified, re-solubilized and re-folded to determine enzymatic activity. Also, the concentration of IPTG used for induction could be dropped in order to achieve a level of induction that is not high enough to promote aggregation.

3) **The Xaa-Pro domain abolishes the toxicity associated with the other domains.** The peptidase domain of the fusion protein may result in the inability of the leader peptide to be recognized or properly processed in *E. coli* leading to lack of secretion, which may ease the toxic affects of the insert. Analagously, if toxicity is associated with the anchor domain then the peptidase domain may result in a conformation that abrogates any catalytic activity associated with this domain. Thus leading to attenuated toxicity when compared to LS. Unfortunately, the current study could not address how the Xaa-Pro domain influences the S-layer signals.

It is worth noting that while possibility one described above can be discounted, possibilities two and three are not mutually exclusive and both may be partially responsible for the lack of toxicity observed after induction of LXS. In order to properly assess these affects as well as the exact source of toxicity in *E. coli* (Leader or Anchor or both), new plasmids allowing for every possible permutation of these components, are currently being built. These new plasmids (pET30:LX, XS, X, L, and S) would not only allow for the determination of the exact source of toxicity in *E. coli* but would also provide insight into the reason why there is no Xaa-Pro peptidase activity detectable in induced pET30:LXS cultures whole cell lysates. The lack of peptidase activity may be a result of at least two possibilities. The peptidase domain may not be able to fold correctly when fused to the leader and/or anchor components, and /or this protein may require cellular components not present in *E. coli* that are required for protein folding or enzymatic activity. Furthermore it is likely, based on microscopic and His probe analysis that aggregates of LXS form inclusion bodies, which are unlikely to show activity. Construction and analysis of the aforementioned vectors would allow for progress to be made toward understanding what exactly is curtailing peptidase activity in the present study and may elucidate a solution to this problem.

IV. Concluding Remarks and Future Goals:

The work described herein provides a solid technological foundation for the genetic engineering of Lactobacilli. Among these advancements is the construction of an *E. coli-Lactobacillus* shuttle vector with a robust MCS as well as general procedures describing molecular techniques optimized for these bacteria. A notable achievement is the development of a novel method of digestion independent cloning (DIC), which, once fully characterized may allow for direct cloning of cassettes into *Lactobacillus*. This research also highlighted the unique difficulties inherent within a non-model system. Incompatibility between the *Lactobacillus* derived localization components and the *E. coli* cloning host in this study resulted in the inability to recover the desired shuttle plasmids and resulted in a significant amount of lost time and resources. In order to prevent this same mistake from occurring again, a three stage experimental plan has been developed.

Stage 1: Characterization of all components in the *E. coli* model. This initial investigation will allow for preliminary conceptual evidence for the utilization of prolyl peptidases to hydrolyze immunogenic peptides when fused to secretion and anchoring signals.

Stage 2: Determination of the functionality of S-layer localization components in the intermediate host, *L. acidophilus*. This stage utilizes the organism from which the S-layer components were originally isolated to determine if they are still functional when expressed in recombinant form and fused to the peptidase.

Stage 3: Assess the efficacy in the destination organism, *L. reuteri*. The final stage will ascertain whether or not a bacterial prolyl peptidase delivery vehicle is a viable treatment option for CD.

By following this experimental design, future research is expected to be more likely to succeed. In this investigation, it is likely that the toxicity witnessed in the cloning host is a result of fundamental differences between the cellular physiology of these organisms that are not yet completely understood. Indeed the exact function of the SlpA protein in *Lactobacillus acidophilus* has not yet been completely defined.

This underscores a fundamental limitation imposed by the nature of applied genetic research. The genomic revolution has provided an incredible wealth of information from a vast number of organisms with extremely diverse ancestry. The challenge of sifting through this information and effectively mining the data for products that may be of therapeutic importance is laid upon the shoulders of researchers in the post-genomic era. The step-wise approach described above provides a framework for investigators in this field. While the exact stages will differ depending on the specific goal, it is important to be able to move through models to prevent the over pursuance of false leads and maintain a steady course towards the final goal.

This ongoing research presented in this report aims to create an entirely new form of treatment, not only for celiac disease but other disorders in which enzyme supplementation through commensal microorganisms might be useful. This goal represents a paradigm shift in the way scientists view the human body. It is no longer seen as a collection of discrete parts that make up a whole, but as a single, integrated biological system in which microorganisms play a fundamental role.

References

Cavaletto, M., M. G. Giuffrida, *et al.* (2004). "The proteomic approach to analysis of human milk fat globule membrane." Clinica Chimica Acta **347**(1-2): 41-48.

Cellier, C., Delabesse, E., Helmer, C., Patey, N., Matuchansky, C., Jabri, B., Macintyrem E., Cerf-Bensussan, N., Brouse, N. (2002) Refractory sprue, coeliac disease, and enteropathy associated T-cell lymphoma. *Lancet* 356, 203-208.

Ciclitira, P., Johnson, M., Dewar, D., Ellis, J. (2005) The pathogenesis of celiac disease. *Molecular Aspect of Medicine* 26; 421-458.

Corrao, G., Corrazza, G.R., Bagnardi, V., Bruseo, G., Giacci, C., Cottone, M., Sategna, Guidetti, C., Usai, P., Cesari, G., Pelli, M.A., Loperfido, S., Volta, U., Calabro, A., Certo, M., (2001). Mortality in patients with coeliac disease and their relatives: a cohort study. *Lancet* 358; 356-361.

De Angelis, Maria, Rizzello, Carlo G., Fasano, Alessio, Clemente, Maria G., De Simone, Claudio, Silano, Marco, De Vincenzi, Massimo, Losito, Ilario, Gobbetti, Marco (2006). VSL#3 probiotic preparation has the capacity to hydrolyze gliadin polypeptides responsible for celiac sprue. *Biochimica et Biophysica Acta* 1762; 80-93.

Degraeve, P., Martial-Gros, A. (2003) Purification and partial characterization of X-prolyl dipeptidyl aminopeptidase of *Lactobacillus helveticus* ITG LH1. *International Dairy Journal* 13; 497-507.

Dupont, L., B. Boizet-Bonhoure, M. Coddeville, F. Auvray, and P. Ritzenthaler. 1995. Characterization of genetic elements required for site-specific integration of *Lactobacillus delbrueckii* subsp. *bulgaricus* bacteriophage mv4 and construction of an integration-proficient vector for *Lactobacillus plantarum*. J. Bacteriol. 177:586-595.

Hausch, F., Shan, L., Santiago, N.A., Gray, G.M., Khosla, C. (2002) Intestinal digestive resistance of immunodominant gliadin peptides. *American Journal of Physiology: Gastrointestinal and Liver Physiology*. 283; 996-1003.

Heng, N. C. K., H. F. Jenkinson, and G. K. Tannock. 1997. Cloning and expression of an endo-1,3-1,4-ß-glucanase gene from *Bacillus macerans* in *Lactobacillus reuteri*. Appl. Environ. Microbiol. 63:3336-3340.

Marteau, P., Minekus, M., Havenaar, R., Huis In't Veld, J. (1997) Survival of lactic acid bacteria in a dynamic model of the stomach and small intestine: validation and the effects of bile. *Journal of Dairy Science* 80; 1031-1037.

Mukai, T., Kaneko, S., Matsumoto, M., Ohori, H. (2004) Binding of *Bifidobacterium bifidum* and *Lactobacillus reuteri* to the carbohydrate moieties of intestinal glycolipids recognized by peanut agglutinin. *International Journal of Food Microbiology* 90; 357-362.

Mamone, G., S. Caira, et al. (2003). "Casein phosphoproteome: Identification of phosphoproteins by combined mass spectrometry and two-dimensional gel electrophoresis." Electrophoresis **24**(16): 2824-2837.

Naidu, A. S., Bidlack, W.R., and Clemens, R.A. (1999). "Probiotic spectra of lactic acid bacteria." Critical review in food science and nutrition. **39**: 13-126.

O'Donnell, R., J. W. Holland, et al. (2004). "Milk proteomics." International Dairy Journal **14**(12): 1013-1023.

Rojas, M., F. Ascencio, et al. (2002). "Purification and Characterization of a Surface Protein from *Lactobacillus fermentum* 104R That Binds to Porcine Small Intestinal Epithelia.

Rollan, G., Font de Valdez, G. (2001) The peptide hydrolase system of *Lactobacillus reuteri*. International Journal of Food Microbiology. 70; 303-307.

Ross, P., Morgan, S., Hill, C. (2002) Preservation and fermentation: past, present and future. *International Journal of Food Microbiology* 79(1-2); 3-16.

Ruben, C.F., Brandborg, L.L., Flick, A.L., et. al. (1962) Biopsy studies on the pathogenesis of coeliac sprue in intestinal biopsy. In: Wolstenhome, G.E.W., Camera, M.P. (Eds.), London Ciba Foundation Study Book 14.

Sanz, Y., Toldra, F. Purification and characterization of an X-prolyl-dipeptidyl peptidase from *Lactobacillus sakei*. *Applied and Environment Microbiology* 67(4); 1815-1820.

Saavedra, Jose. (2000) Probiotics and infectious diarrhea. The Journal of Gastroenterology 95(1).

Servin, Alain. (2004) Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiology Reviews* 28; 405-440.

Shan, L., Marti, T., Sollid, L., Gray, G., Khosla, C. (2004) Comparative biochemical analysis of three bacterial prolyl endopeptidases: implications for coeliac sprue. *Biochemistry Journal* 383; 311-318.

Shan, L., Molberg, O., Parrot, I., Hausch, F., Filiz, F., Gray, G.M., Sollid, L.M., Koshla, C. (2002) Structural basis for gluten intolerance in coeliac sprue. *Science* 297; 2275-2279.

Shanahan, Fergus. (2004) Probiotics in inflammatory bowel disease-therapeutic rationale and role. *Advanced Drug Delivery Reviews* 56; 809-818.

Stepniak D., et al. Highly efficient gluten degradation with a newly identified prolyl endoprotease: implications for celiac disease (2006). *Am. J. Physiol. Gastrointest Liver Physiol.* **291**:G621–G629.]

Stiles, M.E., & Holzapfel, W.H. (1997). Lactic acid bacteria of food and their current taxonomy. *International Journal of Food Microbiology*, 36(1), 1-29.

Tannock, G.W. (2004). A special fondness for Lactobacilli. *Applied and Environmental Microbiology*, 70(6), 3189-3194.

Valeur, Nana, Engel, Peter, Carbajal, Noris, Connolly, Eamonn, Ladefoged, Karin (2004). Colonization and immunomodulation by Lactobacillus reuteri ATCC 55730 in the human gastrointestinal tract. Applied and Environmental Microbiology, 70(2); 1176-1181.

Wang, J. J., D. F. Li, et al. (2006). "Proteomics and its role in nutrition research." Journal of Nutrition 136(7): 1759-1762.

Wu, Chi-Ming, Tung-Ching Chung. Green fluorescent protein as a reliable reporter for screening signal peptides functional in *Lactobacillus reuteri*. *Journal of Microbiological Methods*.

Xu, M., Li, Y., Jie, L., Min, D., Liu, J. (2002) An X-prolyl dipeptidyl aminopeptidase from Lactococcus lactis: Cloning, expression in Escherichia coli, and application for removal of N-terminal Pro-Pro from recombinant proteins. *Protein Expression and Purification* 24; 530-538.

Building a laboratory to investigate injury –repair in skeletal muscle and its vasculature

Project Investigator:

Trevor Cardinal
Department of Biomedical & General Engineering
California Polytechnic State University
San Luis Obispo, CA

Building a laboratory to investigate injury-repair in skeletal muscle and its vasculature

The long-term goal of my research program is to participate in the development of molecularly-targeted therapeutic strategies for patients with injury or disease in the peripheral limbs. Example injuries of the peripheral limb include muscle strains or bone fractures while example diseases of the peripheral limbs include peripheral artery disease, osteoporosis, or sarcopenia. This research effort also has important implications for the possibility of limb regeneration, as injury-repair involves cellular processes that are similar to those that control limb development in the embryo. The role that we will play in this overall effort towards more personalized treatments for these injuries and diseases is in advancing the understanding of the molecular regulation of injury-repair. Specifically, we are interested in how blood vessel function is affected by injury and disease. The blood vessels we are most interested in are the small arteries and arterioles, whose primary function is to regulate tissue blood flow through changes in their diameter- a process known as vasoactivity. Although it is well known that injury-repair in skeletal muscle of the limb is associated with impaired vasoactivity, the molecular mechanism (i.e. the proteins) underlying this impairment is poorly understood. This is critical because tissues will not function properly without sufficient oxygen delivery by blood.

Given the critical role of blood flow control in maintaining normal tissue function, the original goals of this proposal were to better understand the impact of injury-repair on vasoactivity and skeletal muscle function by defining the tissue conditions associated with impaired vasoactivity and reduced skeletal muscle function using a chronic ischemia (insufficient blood flow due to arterial occlusion) model of injury-repair. Preliminary data indicated that impaired functional hyperemia is not a ubiquitous feature of ischemia, but occurs only in those muscles that are hypoxic (low O2) or contain recently grown blood vessels. Therefore, the objectives of this research are to:

1. Determine the effects of hypoxia and angiogenesis (growth of new capillaries) on vasoactivity.
2. Determine the effects of hypoxia and angiogenesis on muscle force production.
3. Characterize the hypoxic, angiogenic, and muscle regeneration response to ischemic injury in previously un-described skeletal muscles in the hindlimb

These experiments will further our understanding of the specific aspects of ischemia that cause impaired vasoactivity and skeletal muscle function and thus provide the foundation for future investigations directed towards uncovering the specific proteins underlying these impairments. However, **the focus of this project was solely to acquire some of the necessary equipment and supplies to enable this research and did not involve actual experimentation.**

Background & Significance

The ability for skeletal muscle to increase its blood supply during contraction (termed functional hyperemia) is impaired during injury-repair) [1] and disease (e.g. obesity, hypertension, diabetes, etc) [2]. This observation is significant because skeletal muscle function is required for many essential activities, such as postural maintenance and locomotion. Muscle contraction lasting longer than several seconds requires an increase in O_2 delivery to support the elevated metabolism. The increased O_2 delivery is primarily mediated by increased blood flow through resistance vessel vasodilation (increase in blood vessel diameter), which as stated above, is impaired during a wide variety of

injury-repair and disease states.

Although impaired functional hyperemia during injury-repair and disease are well described in rats [3], rabbits [4], and humans [5], the proteins responsible for this impairment are not known. Understanding the molecular mechanism (i.e. the proteins) underlying this impairment is necessary for the development of specific, efficacious, and personalized therapies designed to restore functional hyperemia during injury-repair and disease. Presently, impaired functional hyperemia in patients (who present with symptoms of intermittent claudication, or pain in the legs during walking) is treated with exercise, dietary changes, and/or pharmacologic agents that are prescribed on a largely trial and error basis [6].

Therefore, to improve the specificity and effectiveness of injury rehabilitation and disease treatment, the goal of my research program is to uncover the proteins that cause the impairment in functional hyperemia during injury-repair and disease. The first step towards achieving this goal is to expand upon previous research in this area by describing the effects of injury-repair on functional hyperemia in the mouse. The importance of this step lies in the fact that mice are the only research animals routinely available for genetic manipulation- targeted genetic disruption and transgenesis. Targeted genetic disruption describes the removal of a specific gene from the mouse genome while transgenesis is a technique used to enhance the expression of a specific gene. Using animals developed with these genetic manipulations, I plan to determine the role of specific gene products (proteins) in modulating functional hyperemia during injury-repair and disease. However, because thousands of the ~30,000 genes present in the mouse genome are expressed in skeletal muscle, genetic manipulation cannot be the sole experimental approach used to uncover the gene products causing impaired functional hyperemia during injury-repair. Viz., it is not practical to selectively modify (remove or over-express) every gene present in skeletal muscle until we find one that causes the impaired functional hyperemia observed during injury repair and disease.

Therefore, we will start the process of identifying candidate proteins involved in vasodilatory dysfunction by assessing the anatomical, physiological, and biochemical impact of injury-repair on vascular function tissue under different experimental conditions. Completing this first step will allow us to define the specific features of injury-repair that lead to impaired functional hyperemia. For this research, we are interested in identifying the particular aspects of ischemic injury (induced by interrupting arterial blood flow) that cause a reduced skeletal muscle blood flow response during muscle contraction.

To determine the effect of ischemic injury on functional hyperemia, we will use chronic hindlimb ischemia in young, healthy mice as a model of injury-repair. In addition to modeling acute, traumatic injury, hindlimb ischemia also serves as a simplified model of peripheral artery disease- therefore the results from this work will provide the foundation for future research directed towards understanding the mechanisms underlying impaired functional hyperemia in more complex human disease models, e.g. aged mice with hypercholesterolemia.

Two different models of chronic hindlimb ischemia will be employed; both models involve interrupting blood flow through the femoral artery (**Figure 1**), which injures the limb and causes muscle fiber death due to diminished oxygen and nutrient delivery. The subsequent repair process involves both muscle fiber regeneration and vascular growth. These two models allow us to examine

the effects of different injurious stimuli and a variety of repair processes.

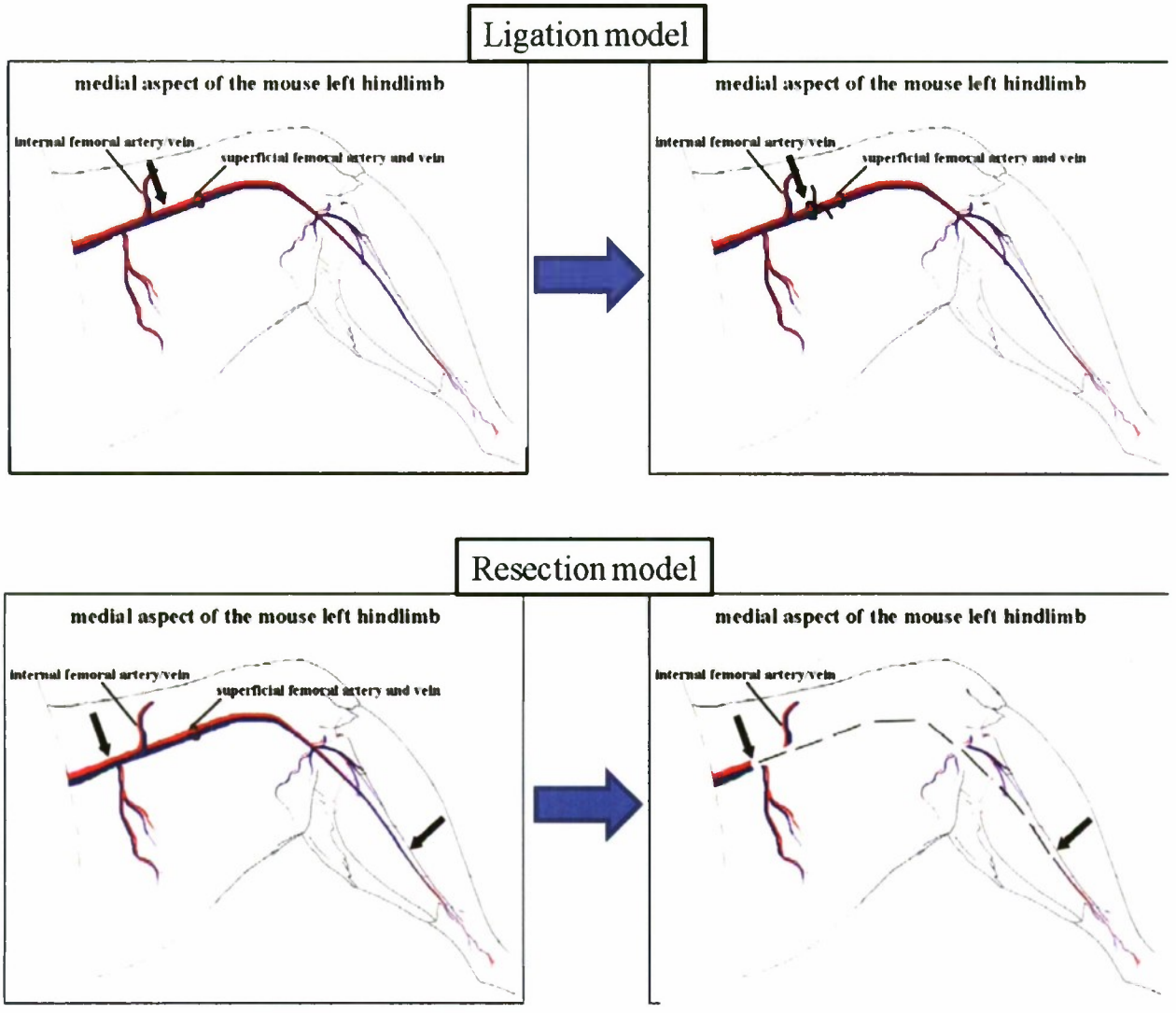


Figure 1. Hindlimb ischemia diagrams

For example, in the ligation model, a single ligation (tie off with suture) will be made in the femoral artery in thc thigh (**Figure 1, top**). This ligation results in collateralization (outward growth of arteries and arterioles) in the thigh and hypoxia in the calf (**Figure 2**) and allows us to examine how two injurious stimuli- hypoxia and ischemia as well as the collateralization aspect of tissue repair affect skeletal muscle physiology. In the resection model, a long segment of the femoral artery will be resected (removed) from the hindlimb (**Figure 1, bottom**). This resection results in hypoxia and angiogenesis (growth of new capillaries) in the calf (**Figure 3**), allowing us to examine the injurious stimuli of hypoxia and ischemia, as well as how microvascular growth affects functional hyperemia during repair.

Additionally, different muscles in each of these models contain different muscle fiber types and also undergo differing levels of injury and subsequent regeneration. This heterogeneous response

provides us an opportunity to examine the influence of muscle fiber properties on the repair process, as well as on physiological function following repair. For example, muscles with different fiber types or that undergo differing levels of fiber damage may also experience differing levels angiogenesis or impairments to blood flow.

Using these two models, we will dissect the contribution of different aspects of the repair process on impaired functional hyperemia and muscle force production.

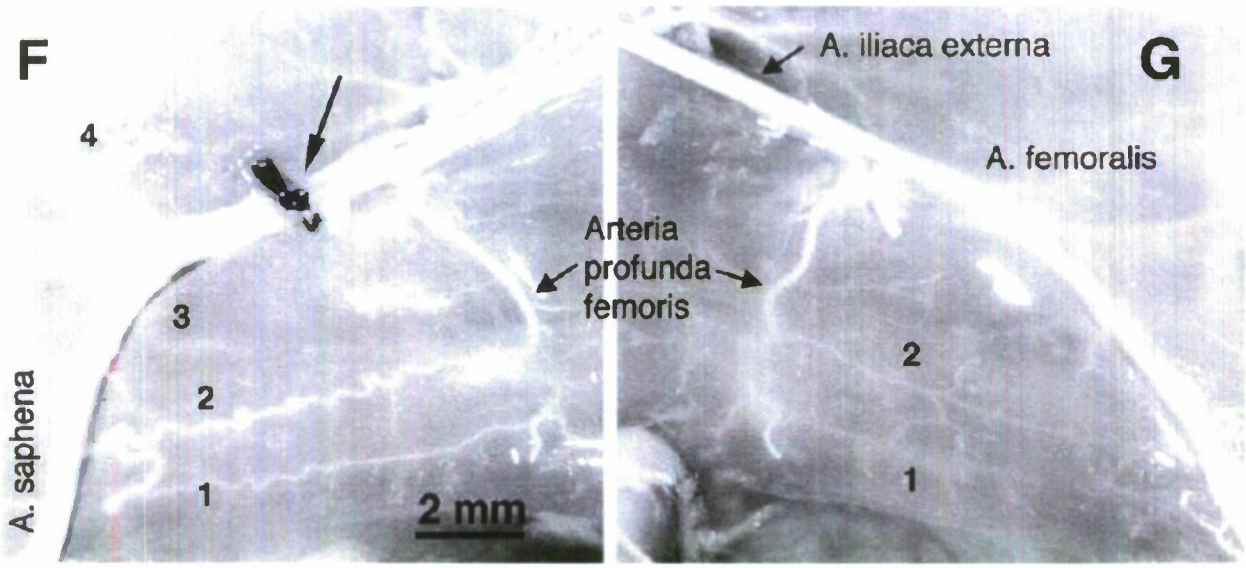


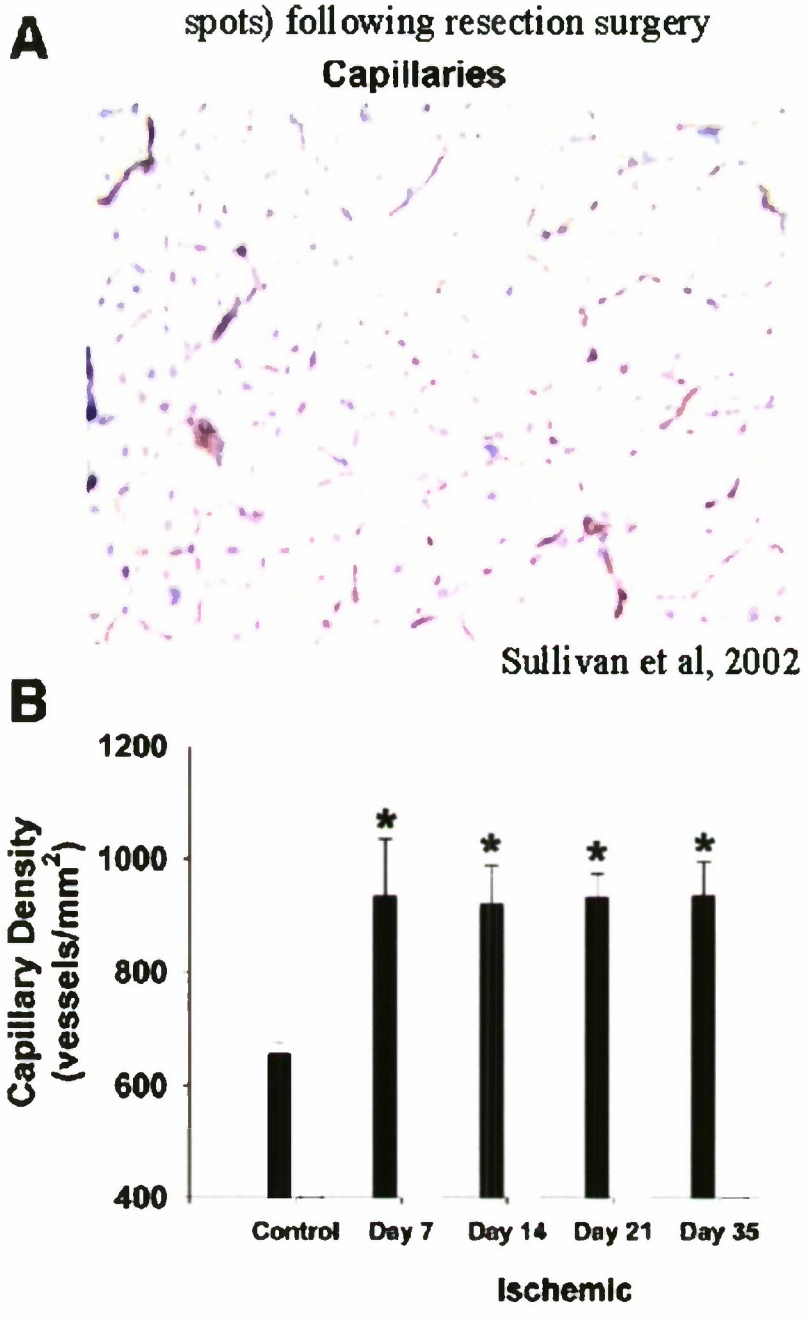
Figure 2. Collateral arteries forming after ligation surgery

For example, if functional hyperemia to the calf is impaired only in the resection model, then we will have learned that angiogenesis during repair is an important contributor to reduced functional hyperemia, and the injurious stimulus of hypoxia does not substantially affect blood flow control.

In summary, the long-term goal of my research program is to identify the proteins responsible for impaired skeletal muscle function during ischemic injury-repair and disease. Using the equipment purchased through this project, we will be able to take the initial steps towards achieving this goal by using a combination of physiological, anatomical, and biochemical assessments to determine the impact of ischemic-injury on vascular function.

gation surg

Figure 3. Capillary formation in calf cross section (angiogenesis, brown spots) following resection surgery



The results of this research may result in the identification of target proteins involved in modulating skeletal muscle function during injury-repair and disease, thus providing targets for specific biotechnology and biomedical engineering-based therapeutic and rehabilitation strategies. Additionally, the results from this work are relevant to the Office of Naval Research with respect to the recovery of soldiers from traumatic injury. By understanding the proteins that affect proper functioning in repairing tissues, specific therapies can be designed to enhance the recovery of soldiers with injuries that affect peripheral limb function by activating those proteins involved in

restoring proper physiological function, or conversely by inhibiting those proteins involved in impairing physiological function. Furthermore, in addition to providing targets for biotechnology-based therapeutics, understanding the proteins involved in mediating normal peripheral function will also provide targets for gauging the effectiveness of physical therapies, such as exercising training, at the molecular level. Thus, identifying the proteins involved in modulating skeletal muscle functional hyperemia will allow the development of specific exercise-training regimens that utilize or are based on biomarkers (mRNA or proteins) to assess fitness and training program effectiveness.

Infrastructure Development & Equipment Use

The first major piece of equipment that I was able to purchase through this award is a micropressure system (**Figures 4 & 5**). The purpose of the micropressure system is to measure blood pressure in blood vessels that are too small to catheterize with a fluid-filled polymer catheter (typically vessels less than 300μm in diameter). The micropressure system is composed of three main components- a glass microelectrode, a pressure regulator, and an amplifier/signal processor (main unit). The microelectrode (a glass capillary tube heat-stretched to a tip diameter of 1-3μm) is filled with a concentrated electrolyte solution (e.g. 3M KCl) that has a very low resistance to current flow. This resistance is continuously monitored by the main unit.

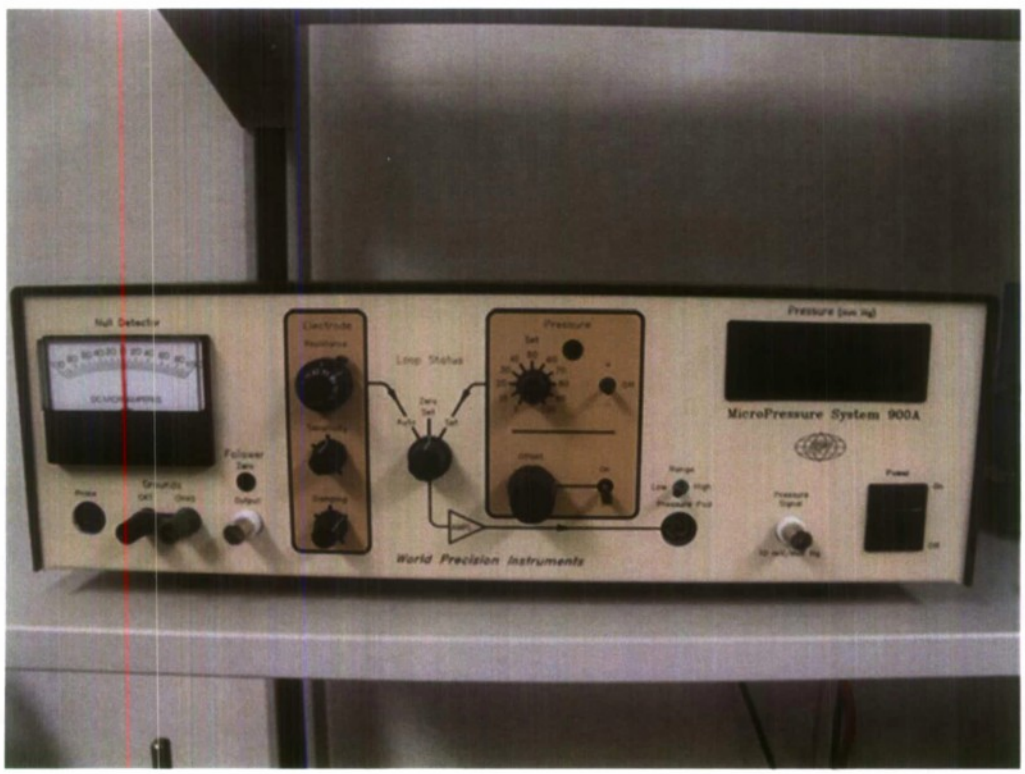


Figure 4. Micropressure system main unit

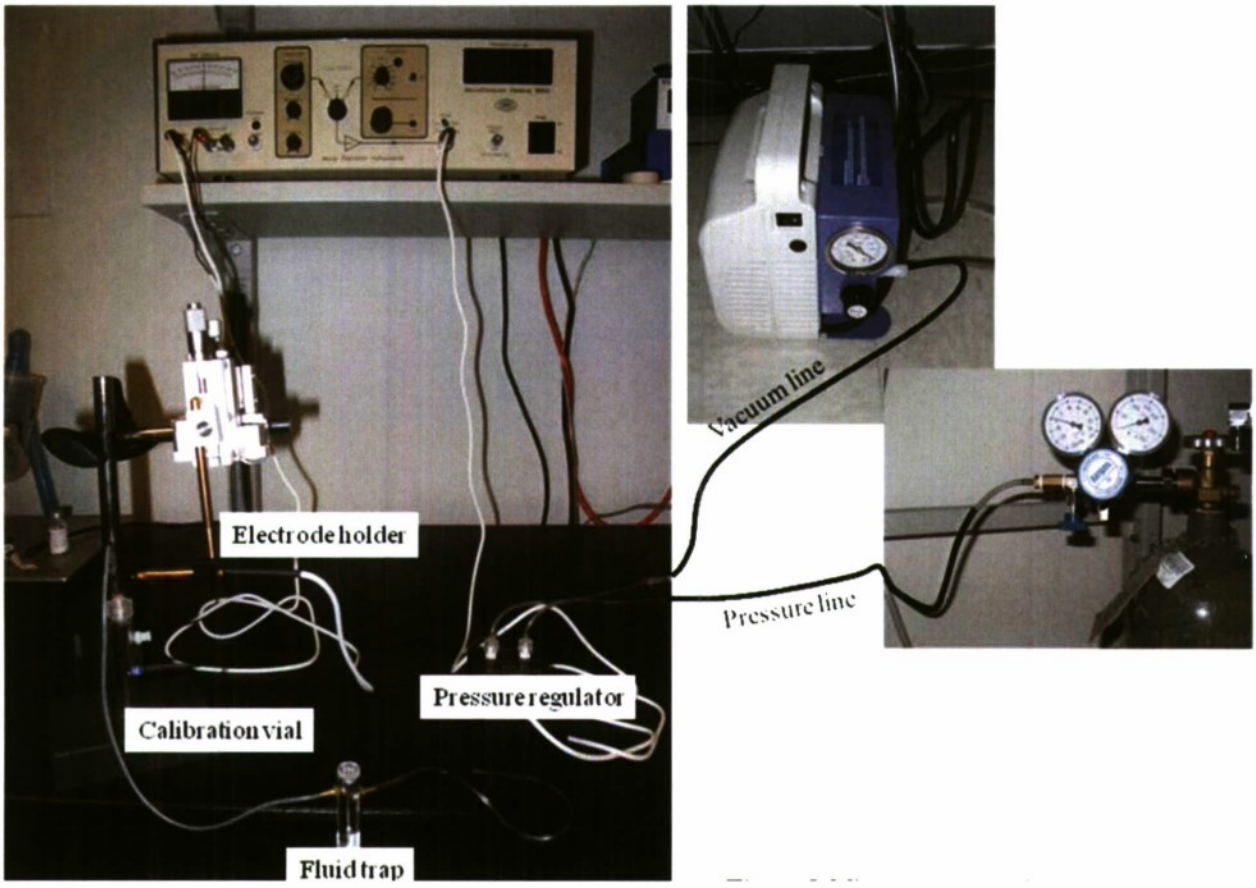


Figure 5. Micropressure system

Once the microelectrode is placed inside a blood vessel, the hydrostatic pressure of blood will cause isotonic plasma to enter the electrode tip. The entry of plasma into the electrode tip will increase the resistance to current flow as it dilutes the concentrated electrolyte solution within the microelectrode. This change in resistance is detected by the main unit, which then signals to the pressure regulator to increase pressure. The pressure regulator is connected to both positive and negative pressure sources (**Figure 5**); if an increase in current resistance as measured by the main unit (due to entry of plasma into the electrode tip), it will signal the pressure regulator to increase the net pressure it allows to pass to the microelectrode by increasing the electrodes' exposure to the positive pressure, decreasing the electrodes' exposure to the negative pressure source, or both. This net pressurization of the electrode will force the isotonic plasma back out of the microelectrode tip and return the current resistance back to its low, baseline level. The arterial pressure is then determined as the net pressure (applied to the microelectrode by the pressure regulator) required to maintain low current resistance (i.e. prevent plasma entry) in the microelectrode.

Blood flow is determined by both the resistance of a vascular network and the pressure gradient across the vascular network. We were previously able to measure blood flow and vessel diameter (the main contributor to resistance). Using this methodology we will now be able to determine pressure gradient across a vascular bed and determine how ischemic injury impacts this driving force of blood flow.

In addition to experimental procedures involving laboratory mice, our research program is also attempting to further our understanding of ischemia and limb injury on vascular physiology and limb function through numerical modeling strategies. A second use of the micropressure system described above will be allow us to measure the arterial pressure within skeletal muscle to provide the boundary conditions for our model calculations as well as validate the pressure predictions made by our model.

The second major piece of equipment that I was able to purchase through this award is a biosensor system (**Figures 6 & 7**). The purpose of the biosensor system is to measure the concentrations of biological molecules that readily participate in oxidation-reduction reactions. The Biosensor is composed of two main components- an amperometric electrode and an amplifier/signal processor. A specific Poise voltage (that is the voltage that induces maximum reduction of a particular biomolecule) is continuously applied to the solid microelectrode from the main unit. However, the electrode is part of an incomplete circuit and therefore current will not effectively flow through the electrode unless that biomolecules is present. When the biomolecules (that is effectively reduced at the specific Poise voltage applied to the electrode) is present, an oxidation-reduction reaction occurs that completes the electrode current and allows current to flow. Therefore, the resistance to current flow in the electrode is directly proportional to the amount of biomolecule in the solution. The specific biomolecules that can be assessed with this device include oxygen, nitric oxide, hydrogen peroxide, hydrogen sulfide, and glucose- the first three of which are critical factors involved in controlling vasodilation and functional hyperemia.



Figure 6. Biosensor system main unit

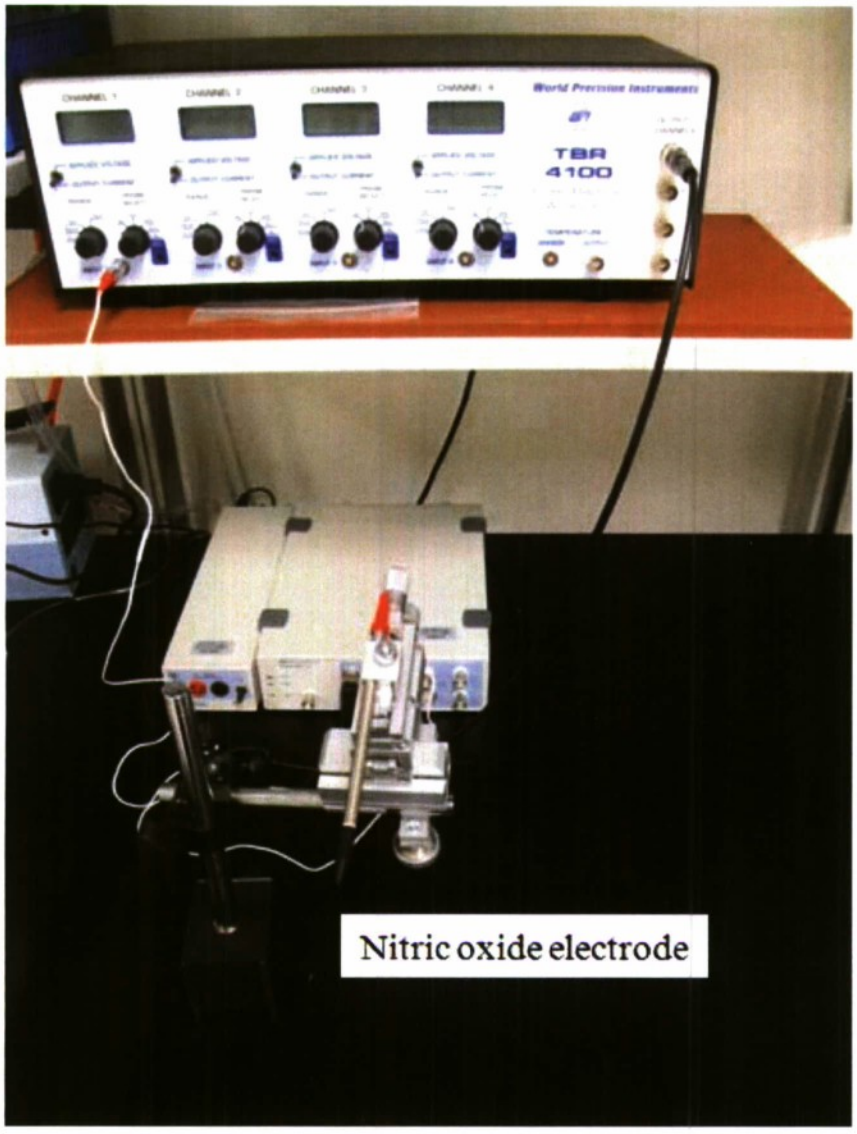


Figure 7. Biosensor system

Although we were unable to use the biosensor for in vivo measurements, we were able to perform bench-top calibrations of the nitric oxide electrode (**figure 8**).

The ability of arterial vessel to dilate and increase downstream blood flow depends on the proper signals being sent to the smooth muscle cells to instruct these cells to contract and decrease blood flow (vasoconstriction) or relax and increase blood flow (vasodilation). As mentioned in the introduction, there is not one single signaling molecule responsible for making smooth muscle cells relax. Fortunately, with our biosensor system we will be able to measure the concentrations of three molecules produced by the endothelium that signal the smooth muscle cells to relax- nitric oxide, hydrogen peroxide, hydrogen sulfide. Measuring how ischemic injury-repair affect the production of the signaling molecules will bring us much closer to determining the molecular cause of impaired hyperemia following ischemic injury. Furthermore, the ultimate regulator of both vasodilation/vasoconstriction and microvascular growth is oxygen.



Figure 8. Nitric oxide bench-top calibration with biosensor system

As oxygen concentrations can be measured with our biosensor, we will be able to describe the effects of ischemic on this critical regulatory molecule- better understanding what features of ischemia (hypoxia, inflammation, tissue growth) are most critical in affecting tissue function.

In conclusion, the primary interest of my research program is in understanding how injury-repair and disease affect tissue function. Specifically, we are interested in microvascular function because the efficient exchange of nutrients and wastes is critical to tissue function. Unfortunately (given its critical importance), microvascular function is impaired following injury and disease. Therefore, the ultimate aim of our laboratory is to describe the molecular mechanism for the impairment of vascular function by ischemic injury. This proposal allowed use to make significant movement towards the completion of this goal through the purchase of equipment that will allow us to assess both hemodynamics, cellular signaling, and tissue health (oxygenation) in healthy animals and following ischemic injury.

References

1. Hudlicka, O., et al., *Effect of long-term electrical stimulation on vascular supply and fatigue in chronically ischemic muscles.* 1994. **77**(3): p. 1317-1324.
2. Frisbee, J.C. and D.W. Stepp, *Impaired NO-dependent dilation of skeletal muscle arterioles in hypertensive diabetic obese Zucker rats.* 2001. **281**(3): p. H1304-H1311.
3. Yang, H.T., R.F. Dinn, and R.L. Terjung, *Training increases muscle blood flow in rats with peripheral arterial insufficiency.* 1990. **69**(4): p. 1353-1359.
4. Walder, C.E., et al., *Vascular endothelial growth factor augments muscle blood flow and function in a rabbit model of chronic hindlimb ischemia.* 1996. **27**(1): p. 91-98.
5. McDermott, M.M., et al., *Lower extremity ischemia, calf skeletal muscle characteristics, and*

functional impairment in peripheral arterial disease. 2007. **55**(3): p. 400-406.

6. Aronow, W.S., *Management of peripheral arterial disease.* 2005. **13**(2): p. 61-68.

7. Couffinhal, T., et al., *Mouse model of angiogenesis.* 1998. **152**(6): p. 1667-1679

8. Murohara, T., et al., *Nitric oxide synthase modulates angiogenesis in response to tissue ischemia.* 1998. **101**(11): p. 2567-2578.

9. Sullivan, C.J., T. Doetschman, and J.B. Hoying, *Targeted disruption of the Fgf2 gene does not affect vascular growth in the mouse ischemic hindlimb.* 2002. **93**(6): p. 2009-2017.

10. Buschmann, I., et al., *Influence of inflammatory cytokines on arteriogenesis.* 2003. **10**(3-4): p. 371-379

11. Paoni, N.F., et al., *Time course of skeletal muscle repair and gene expression following acute hind limb ischemia in mice.* Physiol Genomics, 2002. **11**(3): p. 263-72.

12. Hourde, C., et al., *Sustained peripheral arterial insufficiency durably impairs normal and regenerating skeletal muscle function.* J Physiol Sci, 2006. **56**(5): p. 361-7.

Single Cell Impedance Sensing for Pathogen

Project Investigator:

David S. Clague
Department of Biomedical and General Engineering
California Polytechnic State University
San Luis Obispo, CA

Single-Cell Impedance Sensing for Pathogen Detection and Disease Quantification

Purpose

The motivation behind this work was to provide early detection of pathogen or toxin-related abnormalities at the cellular level with the particular goal to make this technology available to front-line war-fighters, with a secondary application to clinical diagnostics for example, cancer screening. Why single cell detection? Most diagnostics indicators are based on macroscopic symptoms, e.g., fever, swelling, wheezing, and sneezing. For such symptoms to surface, typically thousands of cells are infected; hence, a person exhibiting such symptoms could be in serious or terminal condition and may need to be quarantined. If, however, there were a deployable system that could be used to perform routine checks on personnel via throat swab or blood sample, the war-fighter's health could be assessed at treatable stages.

The operational concept is a system that can be deployed in a first-aid vehicle and personnel would be spot-checked to enable early detection. The desired assay should have the following features:

i) Minimally invasive sample collection
ii) Ease of use
iii) Capability to detect single cells or small groups of cells
iv) Rapid detection
v) Portability

Under these concepts of operation, a MEMS-based single cell impedance sensor was proposed, together with low cost, easy to use support equipment.

During the duration of the C^3RP grant, we developed a Bio-MEMS device to capture single cells and groups of cells. The resultant concept design is shown below in Figure 1.

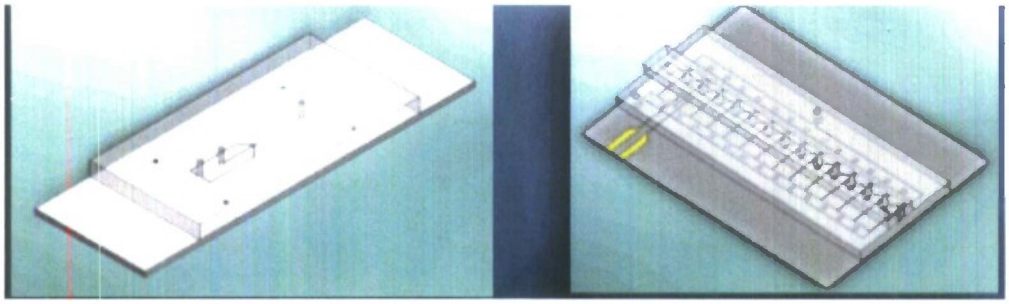


Figure 1. On the left is a multi-chamber, single-cell design to enable simultaneous statistical analysis. On the right is a single chamber to demonstratc the concept.

As shown above, the multiple- ehamber design on the left is optimal to enable rapid prediction of average Impedance Spectra. In this design, cells are introduced into multiple chambers that "dead-end" in a detection chamber. The deteetion chambers are in suction mode during loading. This design ensures (100%) that cells will be captured in cach chamber. To demonstrate the concept of a "dead-end" or sure-fire capture chamber, a single capture chamber was manufactured that is representative of the chambers on the multi-chamber system. (It is important to note that the multi-chamber design was also manufactured and tcsted. It did have some re-design issues that were beyond the scope and timing of the MS student working on the project.) The resultant single-ehamber design was instantiated in PDMS and is shown in Figure 2.

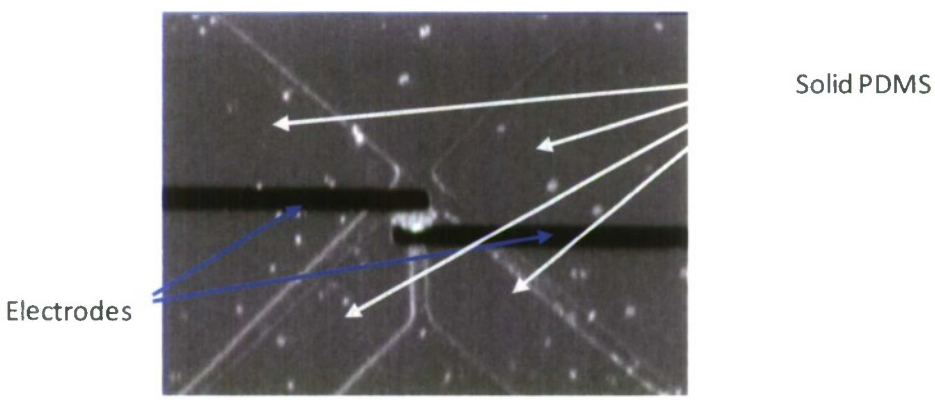


Figure 2. A collection of cells in the PDMS (PolyDimethylSiloxane, translucent rubbcr) eapture chamber with Au electrodes.

As shown above, a collection of cells are positioned between the clectrodes. The geomctric features of the channel walls and chamber walls appear as whitc lines. The

central channel leading vertically below the cells is a suction channel to help position and hold the cells in the chamber where the electrodes meet. The two channels that are at 45 degrees and meet near the top electrode pair are included to enable backflush or expel extra cells in the capture chamber. The speckled white dots in the PDMS regions are merely optical artifacts and do not represent anything significant.

The cell capture chamber was designed for 10 micron diameter cells; hence, if the cell diameter is smaller than 10 microns there will be an accumulation of cells in the capture chamber. If however the cell size is on the order of 10 microns, only a single cell can be captured, see figure 3 below.

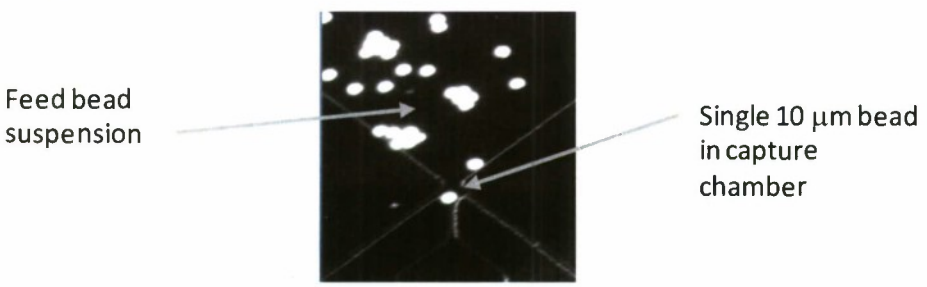


Figure 3. Capture of a 10 micron polystyrene bead in a single-cell capture chamber

A bead suspension is introduced into the chamber and via suction, a single bead is positioned into the capture chamber. To ensure that only a single bead is in the chamber, a back flow is introduced in the 45 degree channels, as described above, with continued suction in the primary capture channel, and the excess beads are flushed out back into the feed suspension chamber (also see Figure 2).

Additionally, as shown in Figure 2., the capture chamber is equipped with parallel electrodes. Once positioned in the chamber, the cell or cells are subject to an AC signal at progressively larger field frequencies. These same electrodes simultaneously create the AC field and read the input current at the various frequencies and the resultant impedance is back-calculated.

To drive the electrodes and to analyze the impedance response, a system was set up using National Instruments cards for the electronics, coupled with a Labview™ interface. In Figure 4, the fluidic and the electronic interface are shown.

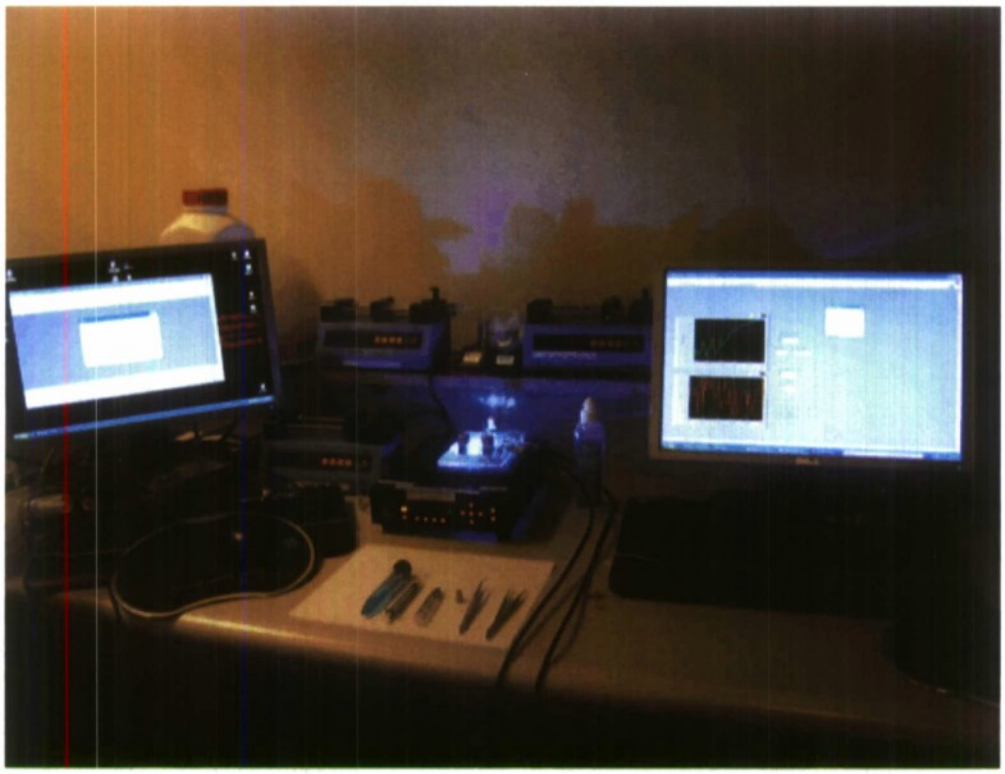


Figure 4. The experimental setup with Labview™ and mierofludic interface.

As described above, the device enabled capture and isolation of a single cell based solely on fluid flow control. In parallel with this effort, the electronic circuit was designed and implemented to enable swept-frequency AC-signal and Impedanec sensing. The circuit design was implemented using a National Instruments Signal Generator and data acquisition cards. The packaged device was designed to enable fluid and eleetrical connection. The connection was designed to enable connection from the macroscopic world to the microscopic. This assembly was then integrated with a LabSmith™ inverted microscope utilizing the Dell XPS with the National Instruments cards and Labview™ interface. The XPS served as a work-station to both drive the circuit and to collect video data of the cell capture process. The electrieal and fluidie eonnections are shown in packaged assembly below in figure 5.

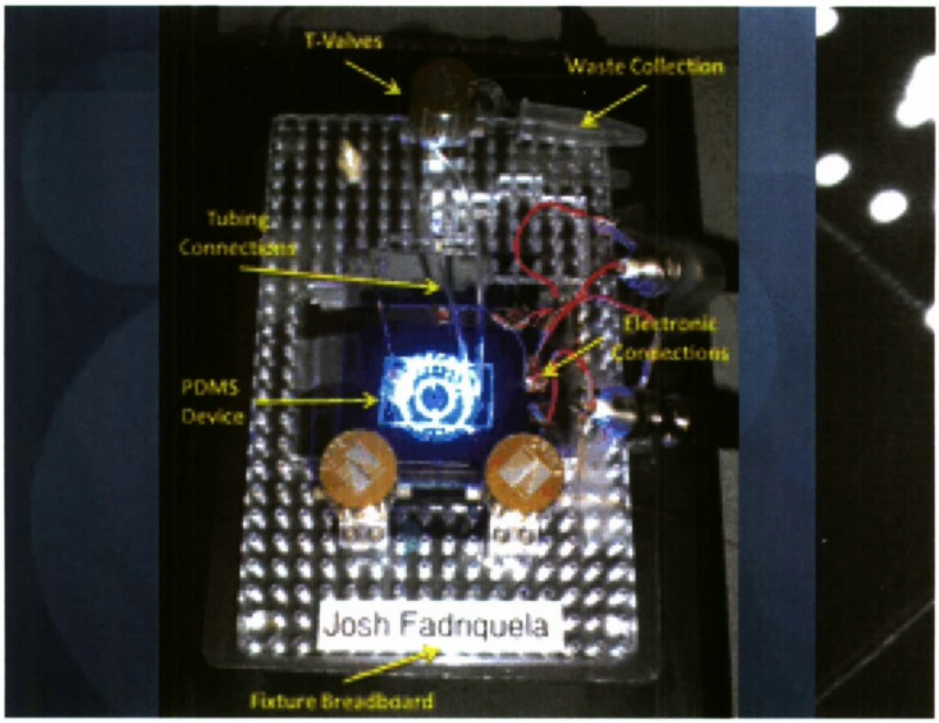


Figure 5. The packaged cell capture chip. On the right the electrical connections are shown. In the center is the chip illuminated by the inverted microscope light source.

The support structure for the assembly shown above is a plexiglass breadboard that enables mounting of Labsmith™ microfluidic valves and plumbing. The outline of the chip can be seen in the middle, and below is the LED microscope light source.

Another significant contribution to this capability is the software interface to enable simultaneous signal generation and data acquisition. The interface between the National Instruments function generator and data acquisition cards was accomplished through a Labview™ program developed by MS student, Stephanie Hernandez. The Labview™ program enabled use-defined signal production/control, data acquisition and analysis. Additionally, a Labview™ interface was developed to enable an effective human interface for experimentation. The interface is shown below in Figure 6.

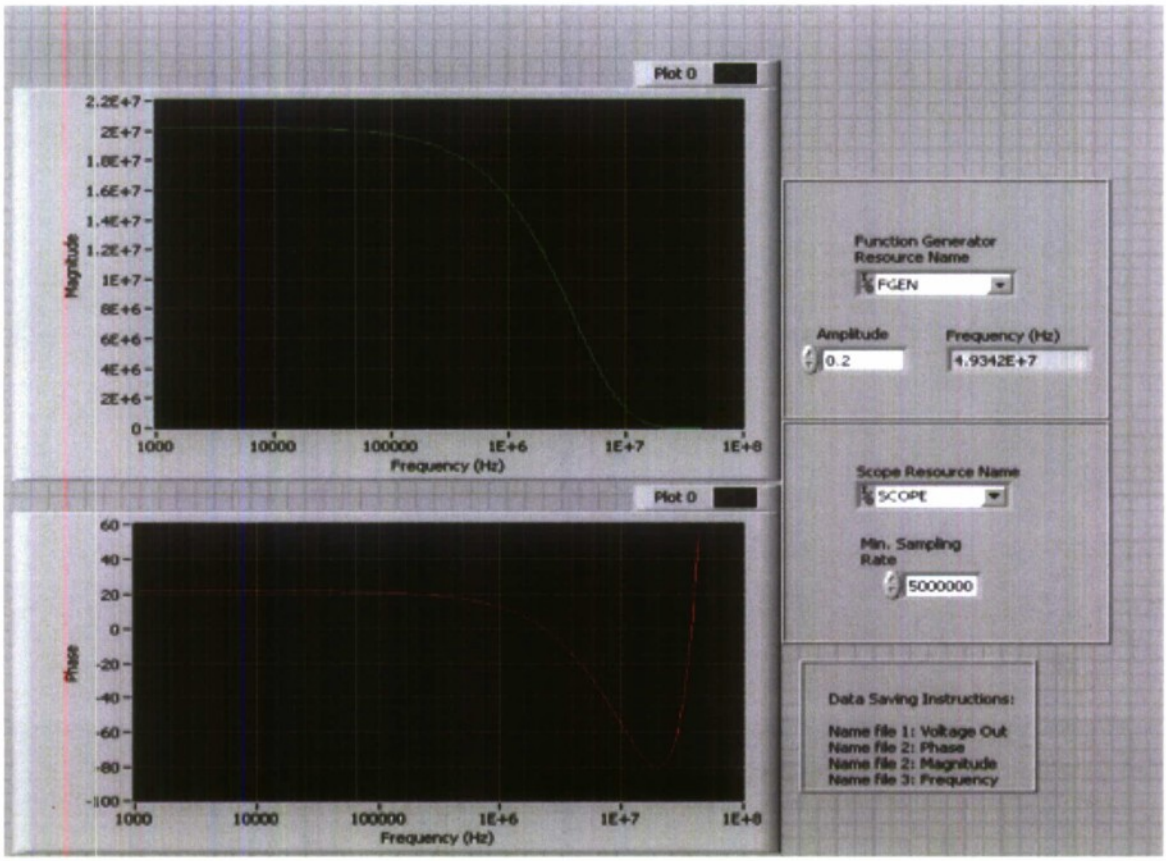


Figure 6. Labview™ interface. The graphs display the measured signal (above/left) and phase (below/left). On the right, the user can input the signal source, the measurement source, the signal amplitude and the signal frequency.

Using the system and interface described above, impedance spectra were developed for air, saline solution, a 10 μm polystyrene bead in saline and a collection of yeast cells with diameters less than 10 μm in saline. The results are compared in Figure 7.

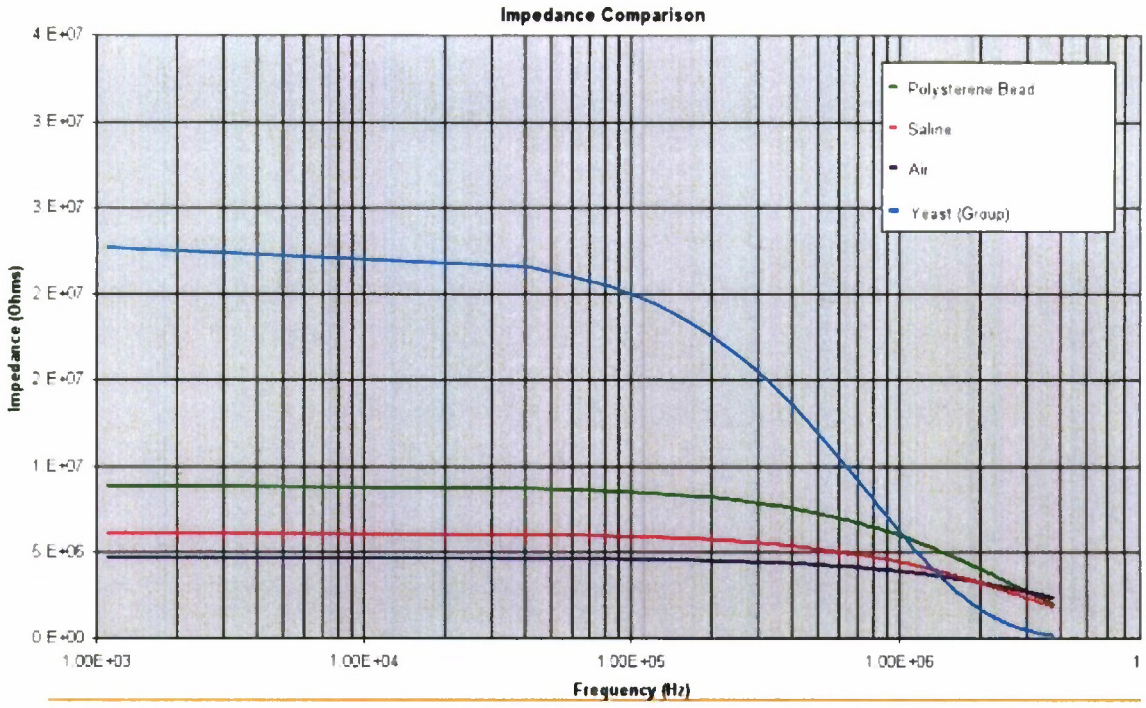


Figure 7. Impedance spectra for air (dark blue), saline solution (magenta), polystyrene bead (green) and yeast cells (turquoise). The frequency ranged from 10 Hz to 40 MHz.

The spectra above cover the desired range necessary for cellular characterization, 100 Hz to 30MHz. Through these measurements, a critical piece of information, the crossover frequency, can be characterized for various cell types, healthy cells and diseased cells. More specifically, the crossover frequency is the frequency at which the target spectrum is equal to the pure saline or background frequency. As shown above, the crossover frequency for a collection of yeast cells, in our system, is approximately 5MHz.

Major Accomplishments

This work has resulted in the following major accomplishments:

- MS degree Josh Fradriquela (Microfluidic single Cell Capture Device) in progress
- MS degree Stephanie Hernandez (LabviewTM interface and Impedance Experiments) in progress
- Design and development of a microfluidic single-cell capture device.
- Design and development of a system to conduct impedance spectra experiments
- Poster presentation at the "2009 Lab on a Chip World Congress"

Future Plans

A start-up company in Minnesota has expressed interest in working with us and possibly hiring students associated with this project. Additionally this system will be adapted for stem-cell research. Specifically the system can be used to characterize cellular modification via chemical stimulation, e.g., differentiation, up-regulation of proteins, cell death, etc.

Current activities

MS student Stephanie Hernandez is still taking data on yeast cells using the system. In particular, she is examining "dead" yeast cells to compare with live yeast cells to demonstrate a demonstrative change in spectra using the same cell line. Additionally MS student Tom Harper will carry on Stephanie's work and adapt the system to explore spectra for cell differentiation. A new graduate student will be recruited to work along side Tom Harper.

The resulting work has been proposed to DARPA as an approach to detection of sub-micron particles, e.g., virus collection and detection from saliva samples. In the proposed scheme, polystyrene beads will be functionalized to create a surface layer of small species, e.g., proteins, viruses, etc. The functionalized beads will be characterized prior to exposing the beads to a sample solution containing the target macromolecule. Once the bead has captured the target, it will be interrogated via the Impedance Sensor System and the spectra will be compared to bead without target. The resulting net spectrum will be used to 1) detect the presence of the target in the sample, and 2) to develop conductivity and permittivity data to characterize these properties of the target, that is, properties yet to be definitively measured.

Multi-AUV Path Optimization for improved Ocean Model Forecasting

Project Investigators:

Christopher M. Clark, Department of Computer Sciences
Mark A. Moline, Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA

Multi-AUV Path Optimization for improved Ocean Model Forecasting

Primary Investigators: Christopher M. Clark, Mark A Moline

Introduction

Oeeans provide us with some of our most valuable resources. To manage and eonserve them rcquires us to understand them. This can be difficult because their characteristic parameters (e.g. temperaturc, currents, etc.) have large spatial and temporal variability.

Ocean modeling provides an essential tool in prcdieting such characteristic parameters with a limited set of measurements. Moreover, models can bc used to foreeast future values of the parameters. Unfortunately, the limited availability of sensor measurements for building such models may not be adequate for generating aeeurate foreeasts.

To remedy this problem, a system of multiple Autonomous Underwater Vehicles (AUVs) ean be deployed between foreeasts to obtain measurements in the specifie loeations which will lead to thc greatest improvement in aeeuraey of the ocean model forecast.

Thc core of this project was to devclop an algorithm that constructs AUV trajeetories that are optimal in the sensc that their measurements will minimize errors in tracking points subject to the practical constraints associated with a real AUV (e.g. maximum velocity, turning radius, ocean currents etc.) In this eontext, the points to be tracked by the AUV are those that minimize error in oeean parameter forecasting.

Along thcse lines, thc investigators have developed a new algorithm to plan paths for multiple AUVs that must visit designated loeations of interest. Unlike previous work, this algorithm takes into consideration 1) kinematic constraints of the AUV, 2) dynamic model of the AUV when caleulating path costs, and 3) the presenec of ocean eurrents.

Algorithm Development

This research addresses the problem of allocating targets to multiple autonomous underwater vehicles (AUV) in the presencc of constant ocean eurrents. The main difficulty of this problem is that the non-holonomic vehieles are constrained to move along forward paths with bounded curvaturcs. The Dubins model is a simple but effcctivc way to handle the kinematic characteristics of AUVs. It gives eomplete eharacterization of the optimal paths between two eonfigurations for a vehicle with limitcd turning radius moving in a plane at constant speed.

In the algorithm developed, Dubins paths arc modified to include ocean currents, resulting in paths defined by curves whose radius of curvaturc is not constant. To determine the time required to follow such paths, an approximate dynamic model of the AUV is queried due to the computational eomplexity of the full model. The lower order model is built from data obtained from sampling the full model. Thc full model is used in evaluating the final tour times of the sequences generated by the proposed algorithm to validatc thc results.

Results

The proposed algorithm solves the task allocation problem with market-based auctions that minimize the total travel time to complete the mission. The novelty of the research is the path cost calculation that combines a Dubins model, an AUV dynamic model, and a model of the ocean current. Simulations were conducted in Matlab to illustrate the performance of the proposed algorithm using various numbers of task points and AUVs. The task points were generated randomly and uniformly close together to highlight the necessity for considering the curvature constraints. Sample task point sequences and the resulting paths are shown in Figure 1.

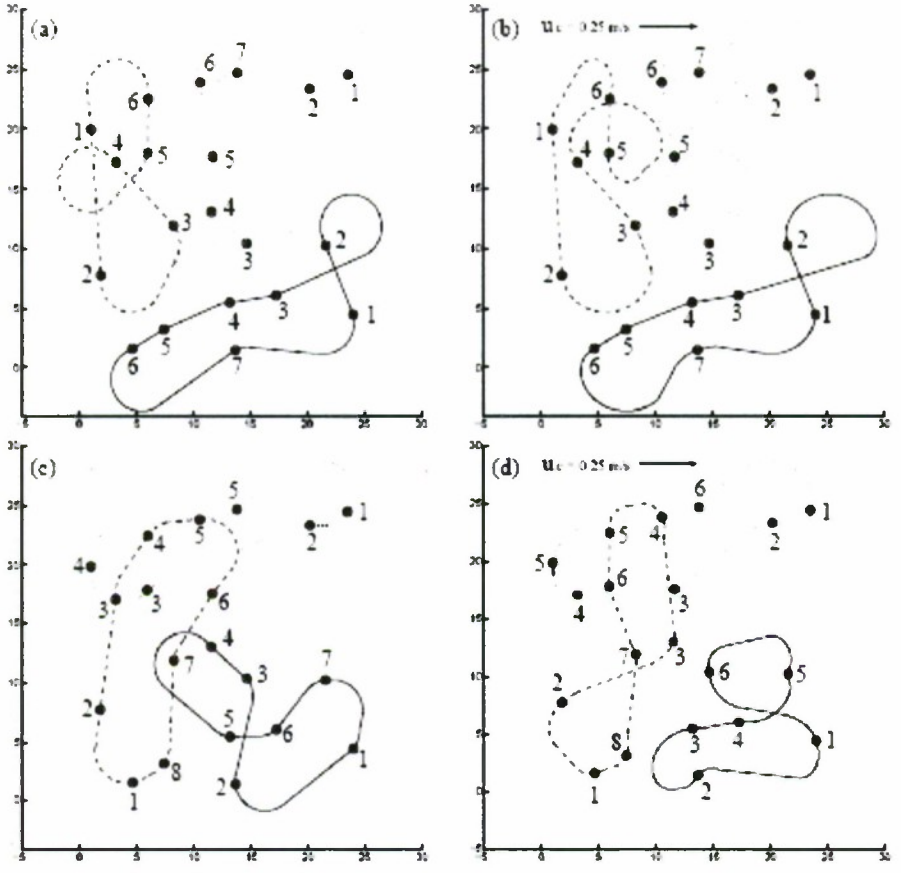


Figure 1: Task point sequences generated by an alternating algorithm that minimizes Euclidean distance path costs (a and c), as well as by the proposed algorithm that considers kinematics when planning (b and d). Additionally currents of velocity 0.25 m/s in the positive x direction were added in (b and d).

For a sufficiently dense set of points, it becomes clear that the ordering of the Euclidean tours (created with an Alternating algorithm) are not optimal in the case of the Dubins multiple travelling salesmen problem. This is due to the fact that there is little relationship between the Euclidean and Dubins metrics, especially when the Euclidean distances are small with respect to the turning radius. An algorithm for the Euclidean problem will tend to schedule very close points in a successive order, which can imply long maneuvers for the AUV. This is clearly demonstrated by the numerous loops that become problematic with dense sets of points. The algorithm proposed in this research does not rely on the Euclidean solution and therefore, even in the presence of ocean currents, can create paths that are feasible for curvature bounded vehicles. To note, path times were decreased considerably in the simulations when using the proposed algorithm as compared to the Alternating algorithm (see Table 1).

	No current		With current	
	Alternating Algorithm	Proposed Algorithm	Alternating Algorithm	Proposed Algorithm
T_{max} (s)	89.9	58.4	101.2	59.8
T_{avg} (s)	75.1	54.5	88.1	57.1

Table 1: Comparison of path completion times for the alternating algorithm and the proposed algorithm.

Field tests were also conducted on an Iver2 AUV at the Avila Pier in California to validate the performance of the proposed algorithm in real world environments, (see Fig 2). Missions created based on the sequences generated by the proposed algorithm were conducted to observe the ability of an AUV to follow paths of bounded curvature in the presence of ocean currents.



(a) (b)

Figure 2: The Center for Coastal Marine Science located at Avila Pier (a) and the Iver2 AUV (b).

Results show that the proposed algorithm generated paths that were feasible for an AUV to track closely, even in the presence of ocean current. In Figure 3, a sample set of thee AUV paths are shown that were planned for by the proposed algorithm and tracked by the Iver2 AUV.

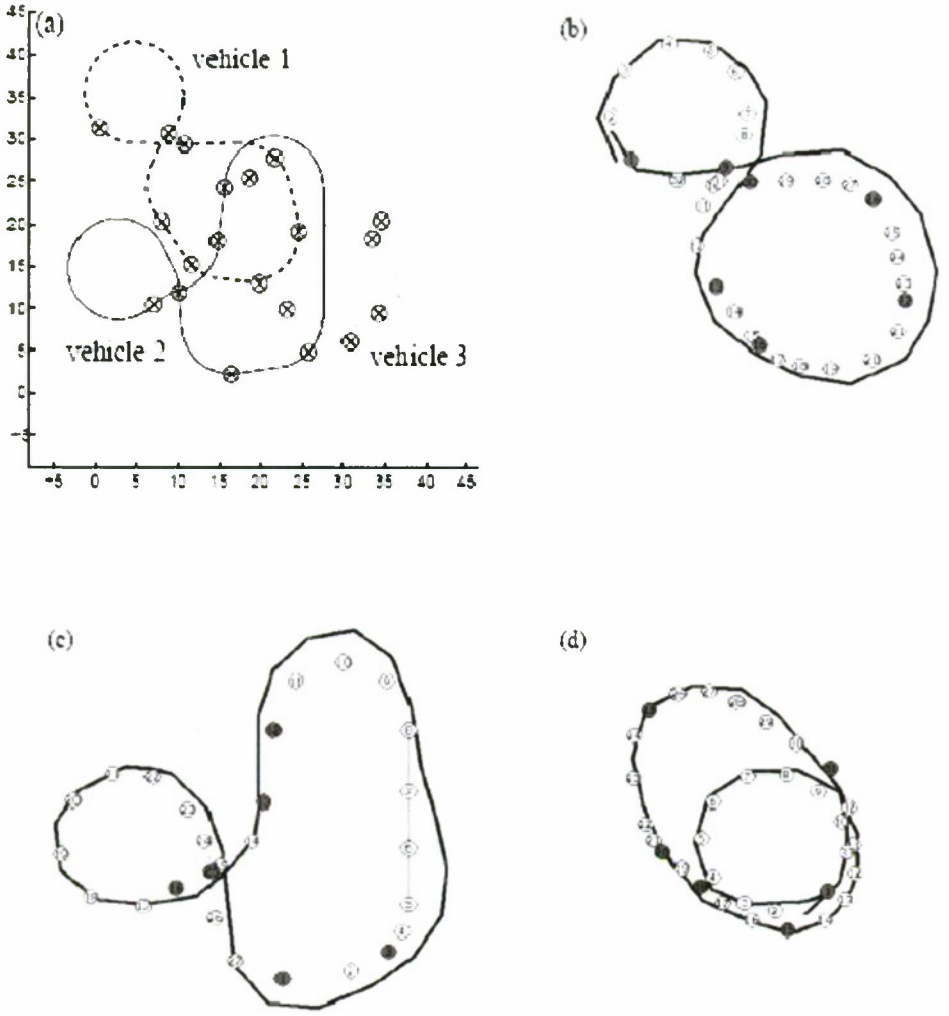


Figure 3: Example paths planned to be track 20 points by 3 AUVs (a). Actual paths (solid line) tracked by the Iver2 AUV in Avila Bay are shown in (a), (b), and (c).

Outcomes

Currently, the work is being published in several forms. Student Beverly Chow's Master's Thesis [1] has been published which details both the algorithm development and ocean results. The experimental results [2] will be published and presented at the International Symposium on Unmanned Untethered Submersible Technology (UUST). Algorithm details and simulation results were submitted in the form of a conference paper [3] to the IEEE/RSJ International Conference on Intelligent RObotics Systems (IROS). Finally, a journal paper is under preparation [4].

As a broader outcome, the algorithm developed will be used for planning within an ocean modeling system that is scheduled to be up and running in by fall. The model will cover the Avila Bay area, enabling forecasting of ocean parameters including temperature and currents in 3 dimensions.

Relevance to ONR

This work falls directly in line with ONR's Code 311- Autonomous Systems, in which "Research is conducted in motion planning algorithms and behavior of teams of agents".

Publications

[1] Chow B. "Assigning Closely Spaced Targets to Multiple Autonomous Underwater Vehicles", Master's Thesis, University of Waterloo.

[2] Chow B, Clark CM, Huissoon JP. 2009. "Assigning Closely Spaced Targets to Multiple Autonomous Underwater Vehicles" To appear in the 2009 International Symposium on Unmanned Untethered Submersible Technology (UUST).

[3] Chow B, Clark CM, Moline MA, Huissoon JP. 2009. "Assigning Closely Spaced Targets to Multiple AUVs", submitted to the IEEE/RSJ International Conference on Intelligent RObotics Systems (IROS).

[4] Chow B, Huissoon JP, Moline MA, Clark CM. 2009. "AUV Path Planning in the presence of Currents" In preparation for submission to the Journal of Field Robotics.

LIDAR Applications Enabled by Fast Wavelength-Tuning Single-Chip Wavelength Tunable SGDBR Lasers

Project Investigator:

Dennis Derickson
Department of Electrical Engineering
California Polytechnic State University
San Luis Obispo, CA

Title of project: **LIDAR Applications Enabled by Fast Wavelength-Tuning Single-Chip Wavelength Tunable SGDBR Lasers**

Investigator(s) and Department(s)

Dennis Derickson

Assistant Professor
Electrical Engineering

Contents:

Report Section 1: 100 kHz Axial Scan Rate Swept-Wavelength OCT using Sampled Grating Distributed Bragg Reflector Lasers

Report Section 2: Generation of High Speed, Linear Wavelength Sweeps Using Sampled Grating Distributed Bragg Reflector Lasers – LIDAR APPLICATIONS

Report Section 3: Microwave Signal Generation Using Single-Chip Fast Wavelength-Tunable Sampled Grating Distributed Bragg Reflector Lasers.

Report Section 4: Summary

Report Section 1: 100 kHz Axial Scan Rate Swept-Wavelength OCT using Sampled Grating Distributed Bragg Reflector Lasers

Shane O'Connor, Michael A. Bernacil, Andrew DeKelaita, Ben Maher, and Dennis Derickson
California Polytechnic State University, 1 Grand Avenue, San Luis Obispo, CA 93407
Contact person: Dennis Derickson, ddericks@calpoly.edu, 805-756-7584

ABSTRACT

Fast wavelength tunable sampled grating distributed Bragg refleetor (SG-DBR) lascrs are used to gcnerate fast, linear, eontinuous wavelength sweeps. High resolution wavelength sweeps in excess of 45 nm are demonstrated at a 100 kHz repetition rate. The front mirror, baek mirror and phase segment tuning segments can be modulated at very fast rates, whieh allows for very fast wavelength ramp rates. This sweep is generated through three time synehronized current versus time waveforms applied to the back mirror, front mirror and phase sections of the laser. The sweep consists of fifty separate mode-hop-free tuning segments which are stitched together to form a near continuous wavelength ramp. The stitching points require a maximum of 60 ns for amplitude, wavelength, and thermal settling time to allow the laser to equilibrate. Wavelength tuning non-linearities, output power wavelength depcndency, and wavelength discontinuities are defects in the wavelength sweep that result from properties of the wavelength tuning meehanism as well as limitations of the signal gcnerators that produee the time varying bias currents. A Michelson Interferometer is used to examine the effeets of these defects for optical coherence tomography (OCT). The OCT measurements demonstrate spectral broadening of the souree and interference signal reduction as the penetration depth increases. However, these effects are not very severe for delay differences less than 2 mm even without eorreetion for sweep nonlinearities.

Keywords: optieal eoherenee tomography, tunable semiconductor lasers, distributed Bragg refleetor lasers

INTRODUCTION

Sampled grating distributed Bragg reflector (SG-DBR) lasers demonstrate wavelength switehing times that are faster than any other laser in its class. Because monolithic cavity SG-DBR lasers utilize a much smaller cavity length than fiber-optie ring lasers[1] and external cavity lasers[2], SG-DBR lasers are capable of generating wavelength sweeps at much faster sweep rates. Experimental results demonstrate that SG-DBR lasers are capable of stepped-wavelength switching times on the order of 10 ns for the phase section of the laser[3]. Furthermore, the large and continuous tuning range of the SG-DBR laser in the C-band, whieh exeeeds 45 nm[4], is another useful feature of these devices. SG-DBR lasers with adjaeent wavelength band coverage can be concatenated for even larger wavelength eoverage. However, creating a eontinuous wavelength sweep requires implementing a complicated tuning meehanism as deseribed in [4].

To implement this tuning meehanism, three synehronized arbitrary waveform generators are used. The arbitrary waveforms are gencrated through a mapping of the wavelength as a function of current into the front mirror, baek mirror and phase sections of the laser. From this wavelength map, a set of currents versus time are synthesized for eaeh of the wavelength tuning segments. The result provides for a continuous wavelength sweep from 1523.317 nm to 1570.078 nm. Repetition rates as fast as 100 kHz for the 47 nm wavelength sweep are demonstratcd. However, wavelength stitehing discontinuities in the wavelength sweep become apparent at high repetition rates. Experimental analysis show that the sweep rate and performanee limitations are due to bandwidth limitations of the arbitrary waveform generators.

A Michelson interferometer is used to experimentally test this high bandwidth, high speed wavelength sweep for high resolution OCT applieations. The results demonstrate that the spectral content of the measurements begin to degrade after approximately 2 millimeters of delay differencc for the 47 nm wavelength sweep operating at a sweep rate of 100 kHz. In this OCT experiment, no attempt was made to eompensate for sweep non-linearities.

2.1. Constructing the Wavelength Sweep

Figure 1 illustrates the block diagram used to generate a fast wavelength sweep. Three arbitrary waveform generators, synchronized by a trigger input, are required to drive the three wavelength tuning segments of the laser. The gain and SOA section of the laser are D.C. biased. Each voltage input is converted to a current input through the use of a series current limiting resistor. The temperature of the laser is held constant at 22° C using a Thermoelectric Cooler (TEC) controller. The optical isolator at the output prevents reflections from causing undesired instability and linewidth broadening within the laser. Though the laser contains an internal isolator, experimental tests proved this isolation to be inadequate.

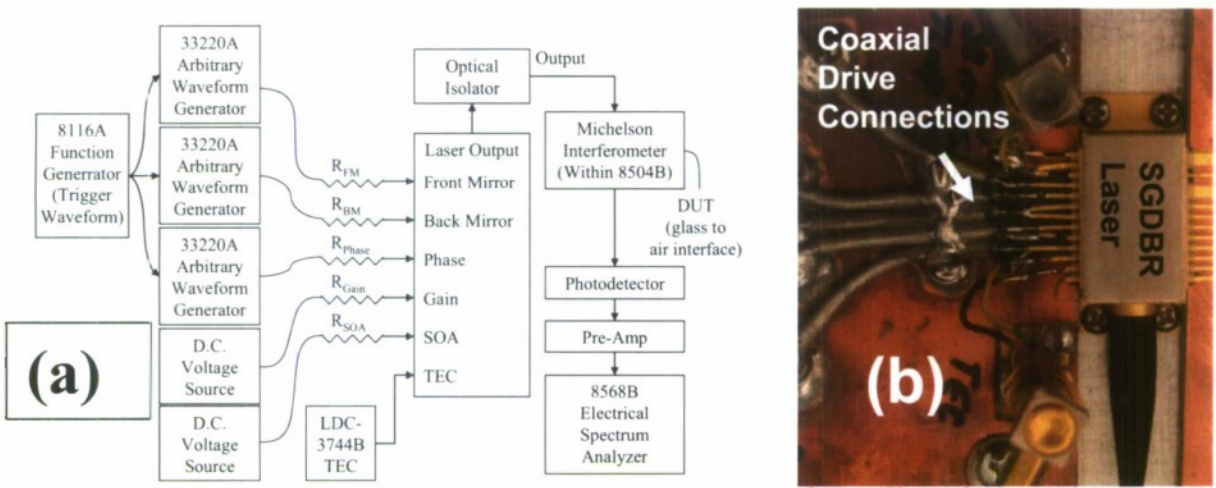


Figure 1: (a) The block diagram required for a fast wavelength sweep generation is shown. Three arbitrary waveform generators are used to drive the three wavelength control connections to the laser (Front Mirror, Back Mirror, and Phase Sections). The resulting ramp of the laser wavelength over a 1523 nm to 1570 nm wavelength range is accomplished in less than 10 microseconds. (b) This is a picture of the SGDBR laser package and the high speed coaxial drive connections.

In order to create a linear, continuous, wavelength sweep, the time varying bias currents to the front mirror, back mirror, and phase sections must be configured appropriately. The DC tuning map of the laser was generated by varying the current into the front mirror and back mirror for the set up in Figure 1. An example tuning map is shown in Figure 2.

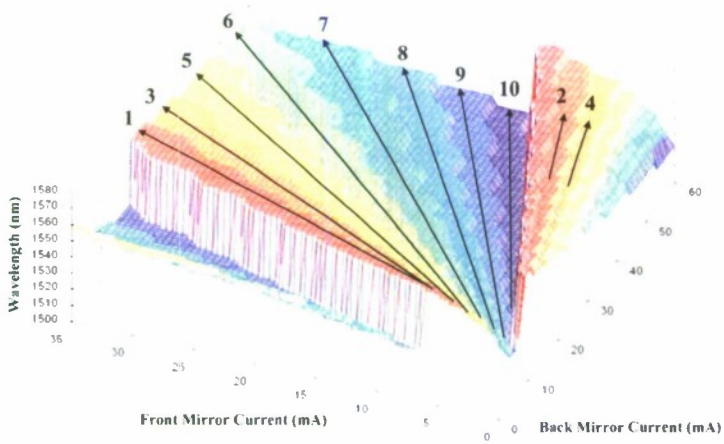


Figure 2: Mode map demonstrating wavelength as a function of front mirror and back mirror bias currents. From this mode map, a tuning strategy is obtained to generate a continuous and linear wavelength sweep from 1523.317 nm to 1570.078 nm. The currents are chosen to follow the paths 1 through 10 that are illustrated in the diagram in order to cover the full laser wavelength range.

In order to tune the laser from the shortest to longest available wavelengths, a series of paths needs to be followed on the tuning curve shown in figure 2. By sequencing through a series of paths from 1 to 10, the entire wavelength range of the SGDBR lasers can be covered. In addition to the basic tuning map of Figure 2, additional tuning information for the phase current and the exact route to take on each tuning path is required. Figure 3 shows an illustration of one of the paths and the optimum route to take on this path.

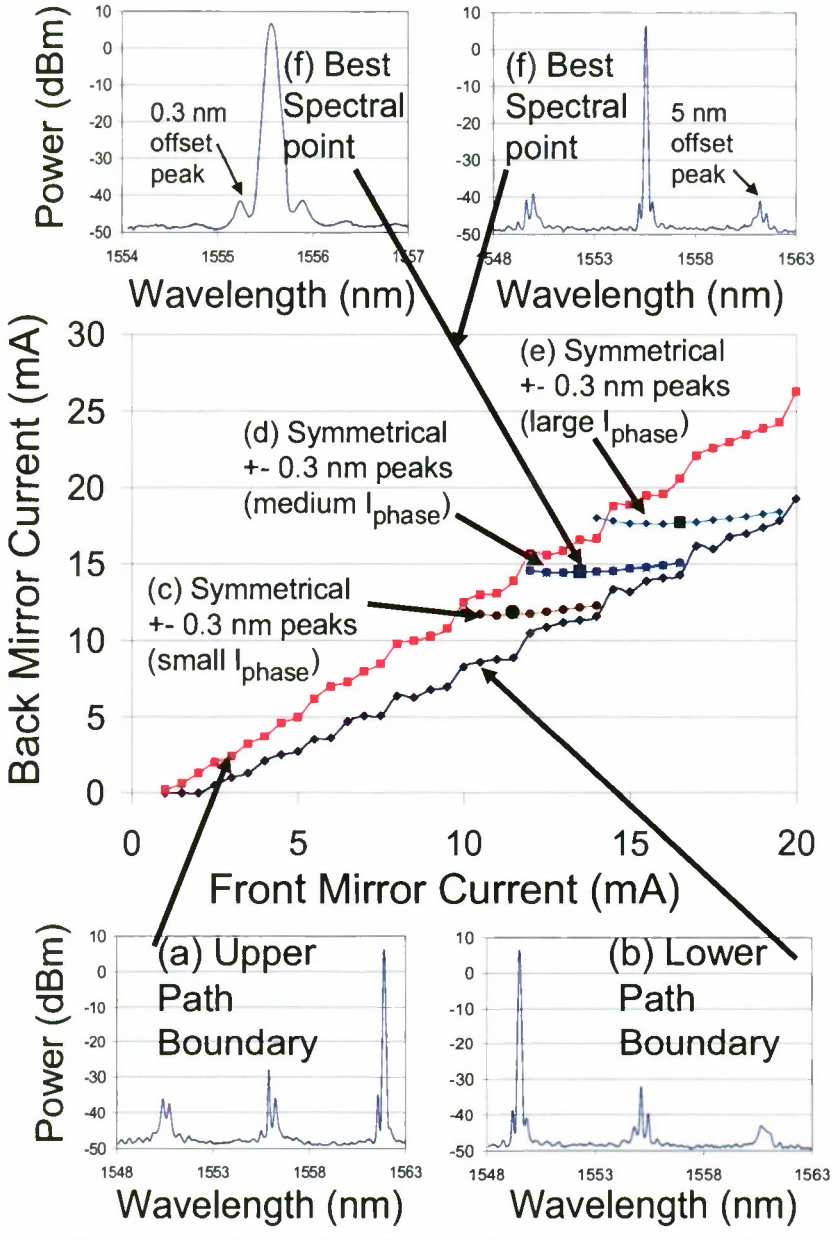


Figure 3: The middle plot illustrates path 6 found in figure 2 in more detail. (a) The upper path boundary is characterized by a +5 nm increase in wavelength. (b) The lower path boundary is characterized by a -5 nm decrease in wavelength. (c) For a small value of phase current, a locus of points is shown where the +- 0.3 nm offset peaks are at a symmetrical level. (d) For a medium value of phase current a locus of points is shown where the +- 0.3 nm offset peaks are at a symmetrical level (e) For a large value of phase current a locus of points is shown where the +- 0.3 nm offset peaks are at a symmetrical level. (f). The best spectral shape bias point for the medium value of phase current is shown. Both a narrow span and wide span spectral trace show the bias point with best spectral shape. A 0.07 nm resolution bandwidth was used for all of the spectral measurements shown in this figure.

In figure 3a/b the boundaries of an example tuning path with spectral examples are given. Three example bias points with low medium and high phase currents are shown in figure 3 c/d/e. An example spectral plot for a best bias point for the front mirror, back mirror, and phase sections is shown in figure 3f. A computer algorithm was generated to sweep the laser through each path and come up with the best set of discrete bias points for best rejection of unwanted spectral side lobes. A computer program was then used to interpolate between each of these best spectral points and generate a table that could be used by the arbitrary waveform generators shown in figure 1. Figure 4 shows an example set of bias values that were sent to the arbitrary waveform generator.

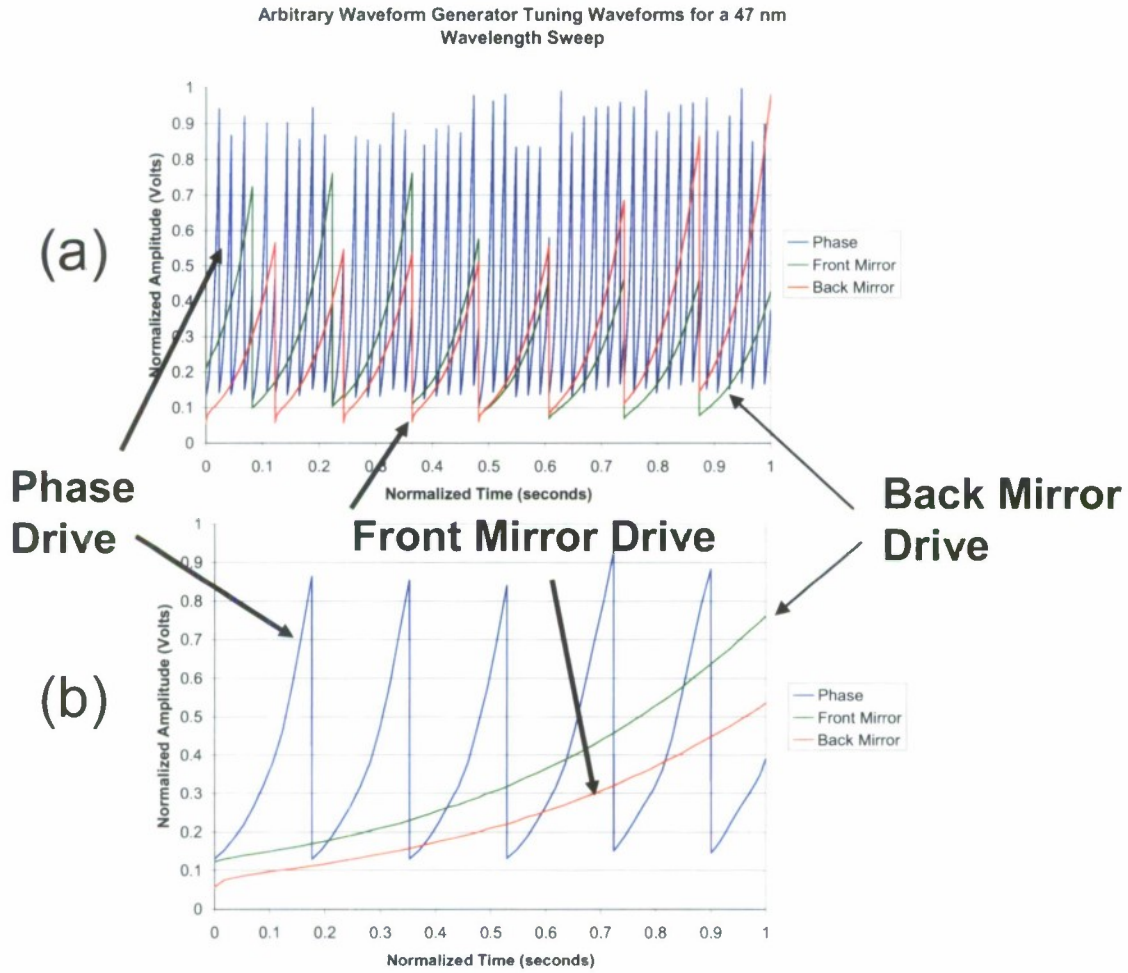


Figure 4: Example drive voltages versus time for the back mirror, front mirror, and phase tuning sections. These voltages are programmed into the arbitrary waveform generators of figure 1. (a) The tuning voltages for a 47 nm wide sweep are shown. (b) The tuning voltages for a narrow sweep are shown.

The waveforms shown in figure 4 show a saw-tooth shaped set of drive voltages. The front mirror and back mirror currents repeat approximately every 7nm of the tuning curve. The phase current repeats approximately every 1 nm. The point in the waveform where the slope is near vertical is called a wavelength ramp stitching point. At the stitching point the current of the laser is changed abruptly but the wavelength before and after the abrupt laser current change should be identical. During the wavelength stitching events, the laser is going trough a mode change and the laser output is not useful during that time interval. Later sections of this paper show that the time duration of these stitching events is about 50 ns for this prototype system. The sweep that is constructed from this laser is then a series of 50 continuous tuning segments stitched together with 50 ns transition times between the segments.

2.2. Wavelength Sweep Analysis

The current waveforms of Fig. 4 were repeated at a series of repetition rates between 0.1 Hz of 1 MHz. The time-averaged spectrum for a 100 kHz repetition rate is shown in Figure 5. This sweep exhibits an optical sweep width of approximately 47 nm. The output power of the source did vary with position in the sweep. The power of the source varies along each path shown in figure 2 if the gain section and semiconductor optical amplifier segments are at a constant bias level. Amplitude flatness can be obtained by providing a current amplitude control to the semiconductor optical amplifier segment as the laser tunes across the band.

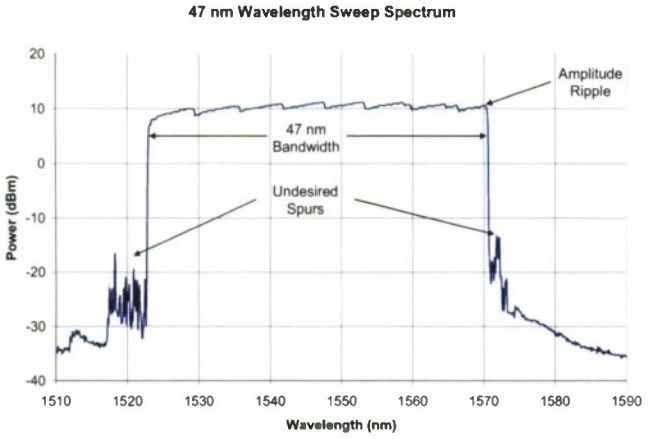


Figure 5: This shows the time averaged power versus wavelength for the laser tuning over a 47 nm wavelength range at a 100 KHz repetition rate. A 0.1 nm resolution bandwidth is used in this measurement.

An experiment was set up to test the linearity of the wavelength sweep and the time duration of the wavelength stitching points. Figure 6 shows the measurement apparatus for characterizing wavelength versus time.

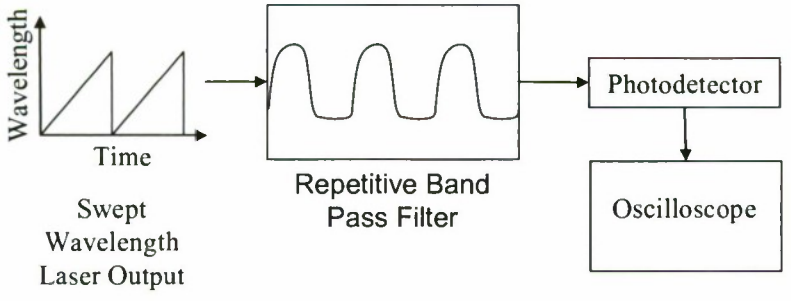


Figure 6: A repetitive band pass filter is used to assess the linearity of the wavelength versus time.

A Fabry-Perot filter with a 0.4 nm free spectral range was place at the output of the tunable laser. The periodic transmission versus wavelength of the Fabry-Perot filter provides wavelength markers that repeat as the tunable laser is scanned over its tuning range. The band pass filter provides an FM to AM conversion of the laser output, since the amplitude of the transmitted signal at the output of the filter varies as a function of the input wavelength. The wavelength ramp at the output of the laser, which provides a linear change in wavelength as a function of time, will then follow the shape of the band pass filter over the width of the wavelength sweep. This optical signal is converted to an electrical signal with a high speed photodetector and applied to an oscilloscope.

Fig. 7 shows the time resolved output of the measurement apparatus in fig. 6 for sweep repetition rates between 1 Hz and 1 MHz. The horizontal axis is time normalized to the period of the sweep. For a 1 MHz repetition rate, the time coverage would be 0 to 1 microseeond. The sweep eovers a 3.2 nm range.

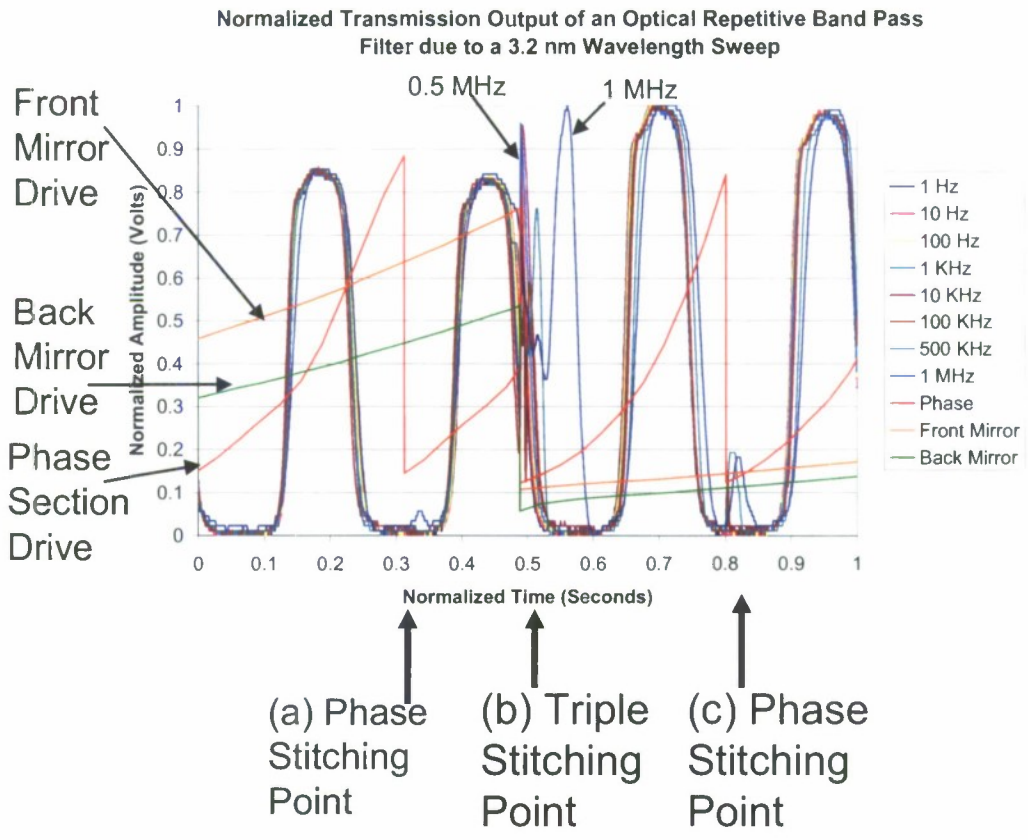


Figure 7: An example output for the measurement apparatus of figure 6 is shown. The output of the repetitive band pass filter demonstrates the performanee of the wavelength sweep as a funetion of sweep repetition rate from 1 Hz to 1 MHz. Two types of switehing points are illustrated for the laser. Figure 7a and 7e show stitehing points where the phase eurrent is rapidly switehed between two values. Figure 7b shows a triple stitehing point where front mirror, back mirror and phase seetion drives are abruptly ehanged. The ideal measurement result would show a smooth ehange in filter output as a funetion of time that replieates the transmission versus wavelength funetion of the Fabry-Perot filter. For low sweep rates, the wavelength diseontinuities at both phase and triple stitehing points are very small. For high sweep rates at 500 kHz and above the laser does not stabilize at the stitehing points and large stitehing errors oeeur.

The drive voltage for the front mirror, baek mirror and phase sections is shown on the same axis as the output of the detector. This wavelength ramp measurement shows two types of wavelength stitching errors. The first phase stitch in figure 7a is found at 0.3 normalized seconds. For low sweep rates, the output of the filter is smooth indicating a small stitching error. For frequeneies above 500 kHz, there is not sufficient time for the laser to stabilize and a wavelength coverage error occurs. The stitching errors are more pronounced when a triple stitching point occurs. A triple stitching point oeeurs when the phase, front mirror, and back mirror drive voltages all change abruptly.

The time duration of the stiteh transient was studied in detail. The SGDBR laser system takes approximately 40 ns to equilibrate to the identical wavelength with a different laser longitudinal mode during a phase stitching transition. The stitehing point at 0.5 normalized seeonds represents the worst ease triple stiteh situation where the phase seetion, front mirror, and baek mirror eurrents are changed simultaneously. In this case, the wavelength, power transient, and thermal transient are more severe but the entire stitching time is still less than 60 ns.

2.3. Experiment: Michelson Interferometer OCT measurements

A Michelson Interferometer is used to test the fast generated wavelength sweep in an OCT test environment. This optical signal is split into two paths; a reference path and a path to image a device under test. The reference path contains a moveable reflector that enables the ability to change the path length difference between the reference path and the path to the DUT. The cleaved end of a single mode fiber optic cable provides 4% power reflection at the glass to air interface and serves as the device under test. The two optical signals are mixed at the output of the interferometer, providing an optical beat signal that is incident on a photodetector. The photodetector converts the envelope of the beat signal into an electrical signal that can be measured by a spectrum analyzer. This experimental set up, illustrated in Figure 8, creates a simulated OCT experiment in which the change in the delay difference between the reference and the DUT allows for the simulation of different penetration depths within the DUT.

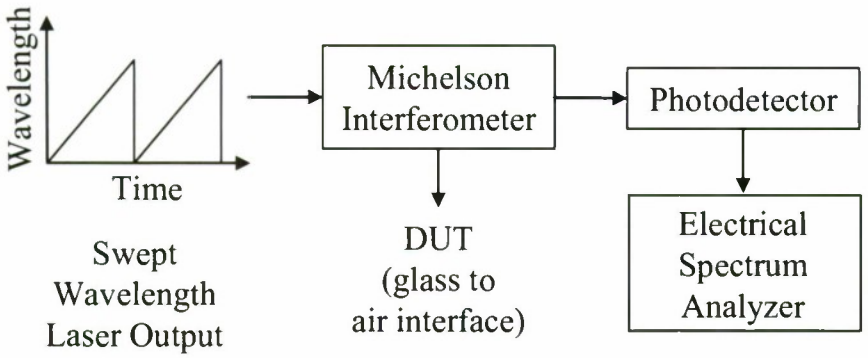


Figure 8: Block diagram of the Michelson Interferometer experiment simulating an OCT measurement.

For a swept wavelength source, the beat signal that results at the output of the interferometer is caused by the wavelength (or frequency) difference between the two signals. This difference frequency is related to the delay difference between the two paths. Furthermore, the difference frequency increases proportionally to an increase in the delay difference between the two paths. Figure 9 demonstrates the OCT measurement for various delay differences using the 47 nm wavelength sweep of Fig. 3 at a repetition rate of 100 KHz. This 47 nm sweep provides a distance resolution of approximately 25.6 μm^5. The experimental results demonstrate that the spectrum of the beat signal broadens and the amplitude decreases as the delay difference between the two paths increases. For delay differences of 2 mm or less, these undesirable effects are not especially severe. No efforts were made to correct for the nonlinearity of the wavelength sweep in this measurement.

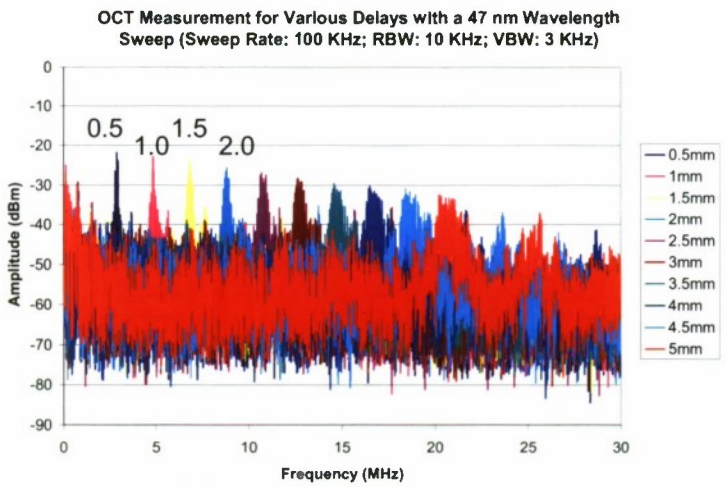


Figure 9: OCT measurements for the 47nm wavelength sweep at a repetition rate of 100 KHz.

3. CONCLUSION

Fast wavelength sweeps with SG-DBR lasers have been demonstrated by applying three synchronized arbitrary waveforms to the respective tuning inputs of the laser. The short cavity length and small capacitance of the SG-DBR laser device allow for a very fast wavelength-swept source for optical coherence tomography applications. A 47 nm continuous wavelength sweep was demonstrated at user adjustable repetition rates between 0 to 1 MHz. The 47 nm wavelength sweeps consists of a concatenation of fifty 1nm wide continuous tuning segments. The fifty concatenated sweep segments each have a 50 nS wait period between sweep segments during which the output of the SGDBR laser is stabilized. An overview of the tuning paths that are needed to drive the laser was given resulting in a series of saw tooth waveforms being applied to the front mirror, back mirror and phase sections of the laser. Useable sweep rates up to 1 MHz were demonstrated for the 3.2 nm narrow sweep, and up to 100 kHz were demonstrated for a 47 nm wavelength sweep. An Optical Coherence Tomography measurement was made with a Michelson interferometer measuring the air to glass interface reflection between an optical fiber and air. The OCT measurement demonstrates that the performance of this wavelength sweep is adequate for imaging up to 2 millimeters in depth into a sample with no correction for wavelength non-linearity in the sweep.

5. ACKNOWLEDGEMENTS

This work was sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152."

6. REFERENCES

[1] Han Young Ryu, Won-Kyu Lee, Han Seb Moon, Ho Suhng Suh, "Tunable Erbium-Doped Fiber Ring Laser for Applications of Infrared Absorption Spectroscopy," *Optics Communications*, 275, pp. 379-384, March 2007.

[2] Kevin S. Repasky, Gregg W. Switzer, and John L. Carlsten, "Design and Performance of a Frequency Chirped External Cavity Diode Laser," *Review of Scientific Instruments*, Vol. 73, No. 9, pp. 3154-3159, September 2002.

[3] Michael A. Bernacil, Shane O'Connor, Ben Maher, Andrew DeKelaita, and Dennis Derickson, "Microwave Signal Generation Using Single-Chip Fast Wavelength-Tunable Sampled Grating Distributed Bragg Reflector Lasers," IEEE *International Microwave Symposium: IMS 2008,* paper WE4D-05, June 2008

[4] Dennis Derickson, Michael Bernacil, Andrew DeKelaita, Ben Maher, Shane O'Connor, Matthew N. Sysak and Leif Johanssen, "SGDBR Monolithic Wavelength Tunable Lasers for Swept Source OCT," *Proceedings of SPIE,* Vol. 6847, paper 6847-97, January 2008.

[5] M. Musa and S. Salous, "Ambiguity elimination in HF FMCW radar systems," Department of Electrical Engineering & Electronics, University of Manchester Institute of Science & Technology, 12 Jan. 2000.

Report Section 2: Generation of High Speed, Linear Wavelength Sweeps Using Sampled Grating Distributed Bragg Reflector Lasers

Shane O'Connor, Michael A. Bernacil, and Dennis Derickson

California Polytechnic State University (Cal Poly), 1 Grand Avenue, San Luis Obispo, CA, 93407, USA

ABSTRACT – Wavelength-tunable sampled grating distributed Bragg reflector (SG-DBR) lasers are used for telecommunications applications in which the laser is set to a communication channel and changed infrequently. SG-DBR lasers can be tuned to any wavelength over a 50 nm tuning range with fast transition times using a set of three control currents. This paper demonstrates generation of fast linear wavelength ramps covering the entire tuning range of the laser. Continuous and linear wavelength sweeps are achieved by applying three time synchronized waveforms to the front mirror, back mirror, and phase sections of the laser. Continuous wavelength coverage is achieved by appending 50 separate mode-hop-free tuning segments. The wavelength stitching transitions require a maximum of 60 ns for amplitude, wavelength, and thermal settling time to allow the laser and drive electronics to equilibrate. Full band wavelength ramps with 100 kHz repetition rates have been demonstrated. An example FMCW LIDAR application of this fast wavelength ramp is given.

***Index Terms* – Distributed Bragg reflector lasers, light detection and ranging.**

I. Introduction

Sampled grating distributed Bragg reflector (SG-DBR) lasers demonstrate wavelength switching times that are faster than external cavity tunable diode lasers or fiber ring lasers due to their very short optical cavity length. Experimental results demonstrate stepped-wavelength switching times on the order of 10 ns for the phase section of the laser[1].

Continuous wavelength sweeps are achieved by mapping the wavelength as a function of current into the front mirror, back mirror, and phase sections of the laser. The wavelength maps are used to find tuning paths that can be concatenated together into a continuous coverage wavelength ramp. Arbitrary waveform generators are used to drive the front mirror, back mirror, and phase sections of the SGDBR laser and create the associated linear wavelength ramp.

A Mach-Zehnder Interferometer is used to experimentally test this wavelength sweep for high speed light detection and ranging (LIDAR) applications[2]. The results reveal that short range LIDAR measurements at sweep rates between 10 and 80 kHz demonstrate a distance resolution of ± 45 μm or less.

II. Constructing the Wavelength Sweep

To create a continuous and linear wavelength sweep, the three tuning sections of the laser must be controlled with very accurate, time synchronized current waveforms. These waveforms are obtained by first analyzing the wavelength of the laser as a function of the front and back mirror bias currents as shown in figure 1.

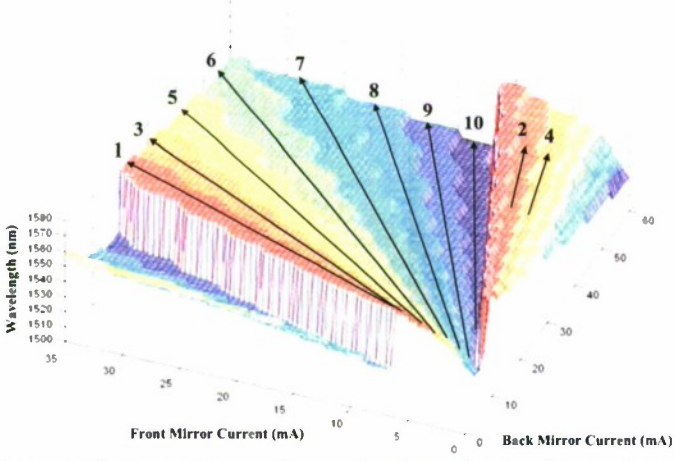


Figure 3: Mode map demonstrating wavelength as a function of front mirror and back mirror bias currents. From this mode map, a tuning strategy is obtained to generate a continuous and linear wavelength sweep from 1523.317 nm to 1570.078 nm.

Figure 1 illustrates 10 different paths that ean be utilized to eover the full wavelength span of the laser. On eaeh path the front mirror and baek mirror eurrents are nearly proportional. The phase seetion tuning eurrent is not ineluded in the map of Fig. 1. The wavelength of the laser would jump in 0.2 nm steps if the phase seetion eurrent of the laser were eonstant. The 0.2 nm step size eorresponds to the longitudinal mode spaeing of the laser. The phase seetion ean be utilized to electrieally streteh the eavity length in order to ereate 1.0 nm wide mode-hop free tuning segments. An example tuning segment of the laser for path 5 is given in figure 2.

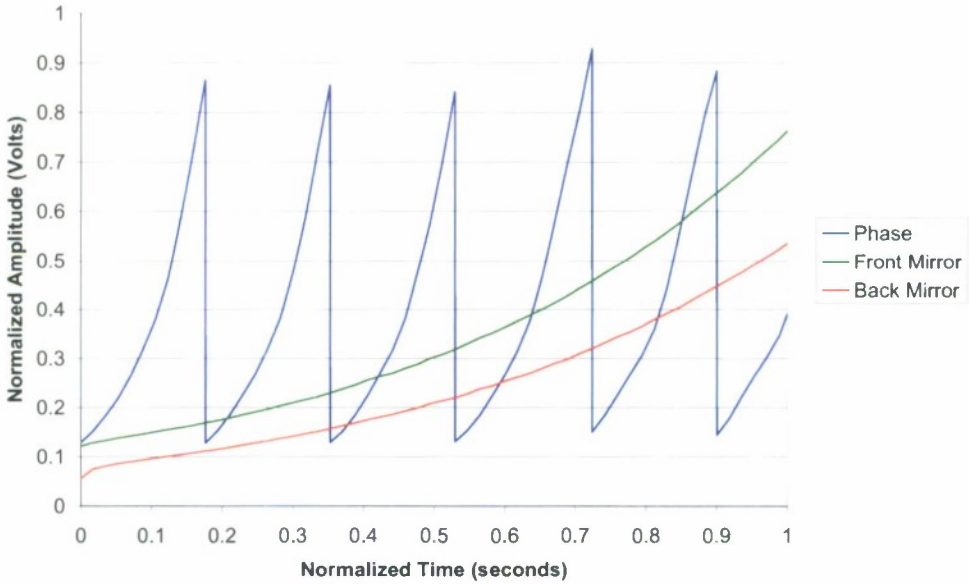


Figure 4: Front mirror, back mirror, and phase section waveforms that drive the laser along path 5.

The front and baek mirror eurrents are proportional but parabolie eurrent versus time funetion is required to get a linear wavelength versus time output. The phase seetion is eleetrieally stretehed and reset 6 times over the duration of this tuning path.

Eaeh path illustrated in figure 1 has a path width. The tuning waveforms are designed to ensure that the laser is tuned along the eenter of eaeh tuning path to prevent mode hops throughout the sweep. The side mode suppression ratio of the tunable laser wavelength ramp was optimized by eharaeterizing the optimal position along the tuning path.

In order to aehieve wavelength sweep linearity, the three synehronized tuning eurrents are mapped in 0.1 nm ineremenls. 1 pm wavelength resolution was aehieved by linear interpolation between the bias points. Finally, by eoneatenating tuning paths in the speeifie order shown in Figure 1, a eontinuous 47 nm wavelength sweep is aehieved from 1523.317 nm to 1570.078 nm.

These three tuning waveforms are uploaded into three Agilent 33220A Arbitrary Waveform Generators. The waveform generators are time synehronized using an external trigger input. Frequeney limitations and finite switehing times of the waveform generators result in wavelength glitehes at eaeh tuning segment eoneatenation point. Thermal settling times intensify the magnitude of the wavelength gliteh at these transition points. Experimental analysis demonstrated that a maximum of 60 ns is required for the laser and drive eleetronies to equilibrate.

III. LIDAR EXPERIMENT

Fig. 3 shows an experimental set up to observe the performanee of the wavelength sweep for LIDAR applieations. The swept wavelength output of the laser is split into two paths; a short referenee path, and a long delay path. These two paths are then eombined to ereate a beat signal that relates to the length differenee between the two paths.

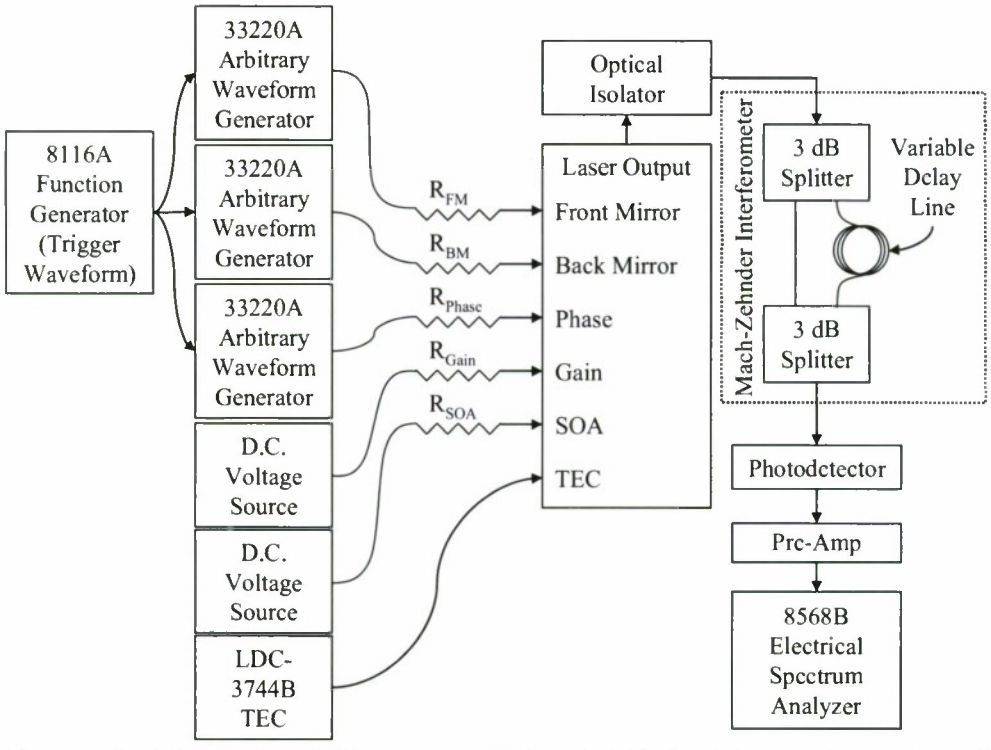


Figure 5: The laser is powered by the D.C. sources and the wavelength sweep is controlled by the arbitrary waveform generators. The Mach-Zehnder Interferometer provides a fiber based LIDAR environment. The frequency of the beat signal, which relates to the length difference between the reference path and delay path, is measured by a spectrum analyzer.

The envelope of the optical beat signal is converted to an electrical signal at the output of a photodetector. This signal is then amplified and measured using a spectrum analyzer.

Figure 4 demonstrates several spectral measurements of the beat signal as a function of the repetition rate. At the 10, 20, 50, and 80 KHz repetition rates, the beat signal demonstrates a resolution of ± 45, 40, 35, and 25 millimeters respectively. Though the theoretical distance resolution of the 47 nm wavelength sweep is approximately 25.6 μm^3, the spectral bandwidth of the beat signal dominates the distance resolution. Furthermore, the spectrum of the beat signal due to a 50 KHz and 80 KHz repetition rate demonstrates an unusual shape. This distortion is most likely attributed to both the non-linearities in the wavelength as a function of time, which is intensified by faster sweep rates, and the bandwidth limitations of the arbitrary waveform generators.

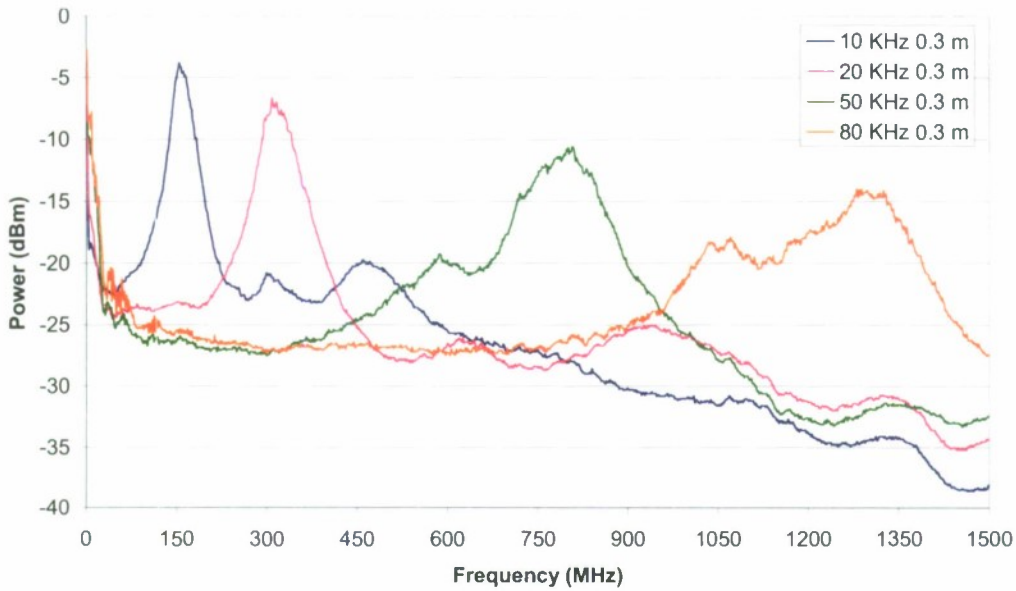


Figure 6: LIDAR measurements for a 0.3 m delay line at sweep rates ranging from 10 KHz to 80 KHz (left to right respectively). The LIDAR measurements are contained within optical fiber using an FMCW architecture.

IV. Conclusion

Fast wavelength sweeps with SG-DBR lasers have been demonstrated by applying three synchronized arbitrary waveforms to the respective tuning inputs of the laser. Mapping the wavelength as a function of the three tuning currents and appending the tuning segments results in a 47 nm continuous wavelength sweep. Performing LIDAR measurements using a Mach-Zehnder Interferometer demonstrates that short range distance measurements at update rates between 10 and 80 KHz provides distance resolutions between ± 45 and ± 25 micrometers.

V. ACKNOWLEDGEMENTS

This work was sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152."

References

[1] Michael A. Bernacil, Shane O'Connor, Ben Maher, Andrew DeKelaita, and Dennis Derickson, "Microwave Signal Generation Using Single-Chip Fast Wavelength-Tunable Sampled Grating Distributed Bragg Reflector Lasers," IEEE *International Microwave Symposium: IMS 2008,* paper WE4D-05, June 2008.

[2] Jesse Zheng, "Optical frequency-modulated continuous-wave interferometers," *Applied Optics*, Vol. 45, No. 12, pp. 2723-2730, April 2006.

[3] M. Musa and S. Salous, "Ambiguity elimination in HF FMCW radar systems," Department of Electrical Engineering & Electronics, University of Manchester Institute of Science & Technology, 12 Jan. 2000.

REPORT SECTION 3: MICROWAVE SIGNAL GENERATION USING SINGLE-CHIP FAST WAVELENGTH-TUNABLE SAMPLED GRATING DISTRIBUTED BRAGG REFLECTOR LASERS

Michael A. Bernacil, Shane O'Connor, Ben Maher, Andrew Dekelaita, and Dennis Derickson

ABSTRACT — MICROWAVE SIGNAL GENERATION USING SINGLE-CHIP FAST WAVELENGTH-TUNABLE SAMPLED GRATING DISTRIBUTED BRAGG REFLECTOR (SG-DBR) LASERS IS DEMONSTRATED. MICROWAVE SIGNALS ARE ESTABLISHED BY A SELF-HETERODYNE TECHNIQUE. THE LASER FREQUENCY IS SQUARE-WAVE MODULATED BACK AND FORTH BETWEEN TWO CLOSELY SPACED WAVELENGTHS. THESE TWO WAVELENGTHS ARE MADE TIME COINCIDENT USING A FIBER MACH-ZEHNDER INTERFEROMETER. THE DIFFERENCE FREQUENCY IS DETECTED AND AMPLIFIED. MICROWAVE SIGNALS UP TO 12 GHZ HAVE BEEN MEASURED BY FREQUENCY MODULATING THE PHASE SECTION OF THE SG-DBR LASER. MILLIMETER WAVE DIFFERENCE FREQUENCIES ARE EASILY AVAILABLE FROM THE SG-DBR LASER. MICROWAVE SIGNAL SPECTRAL WIDTHS AS NARROW AS 10 MHZ HAVE BEEN ACHIEVED FOR LOW BACK MIRROR CURRENT INPUTS. SPECTRAL WIDTH RESULTS AS A FUNCTION OF DEVICE DC BIAS CONDITION ARE PRESENTED. TIME RESOLVED FREQUENCY STEP MEASUREMENTS OF A HIGH-SPEED PACKAGED DEVICE HAS SHOWN 40NS WAVELENGTH SWITCHING TIMES DOMINATED BY THE SPEED OF THE ELECTRICAL DRIVING SOURCE.

***INDEX TERMS* — DISTRIBUTED BRAGG REFLECTOR LASERS, DISTRIBUTED FEEDBACK LASERS, MICROWAVE GENERATION, MICROWAVE OSCILLATORS, MILLIMETER WAVE GENERATION, OPTICAL SELF-HETERODYNING, SEMICONDUCTOR LASERS, WAVELENGTH TUNABLE LASERS.**

1. INTRODUCTION

Tunable diodc lasers are currently the focus of much attention in the telecommunications industry, as thcy arc expected to bccomc essential components for thc next generation of dcnse wavelength division multiplexing systems [1]. Sampled-grating distributed Bragg reflector (SG-DBR) single-chip tunable lasers have emerged as a dominant solution to wavelength-agile fibcr optic communication systcms. These 3 mm long diode lascr chips can tune over the cntirc 1525 to 1565 nm tclecommunication band with three current inputs controlling the laser wavelength. SG-DBR lascrs offer many desirable features such as fast switching times, good side mode suppression ratios (ratios > 40 dB), broad-wavelength tuning ranges of approximately 50 nm (1520 – 1570 nm or 1570 – 1610 nm), as well as lasing stability and rcpeatability. Prcvious research into SG-DBR lasers has led to demonstrations of fast wavelength switching capabilitics of fewer than 5-ns for a 64-channel laser with switching accuracy of ±12 GHz [2].

The fast wavelength switching characteristics and widc tuning range of the SG-DBR laser can bc utilized for microwave signal gcneration. Fig. 1 shows a block diagram illustrating the basic approach that is used. The SG-DBR laser has its opcrating frequency determined by currents applied to the laser. A low frcqucncy square wave modulation with fast transitions is applied to the laser rcsulting in a square wave modulated optical frcquency versus timc. Thc square wave modulated optical signal is applied to a Mach-Zehnder interferometer. The long delay path of thc interferometer is chosen to be equal to half thc period of the squarc wave modulation signal.

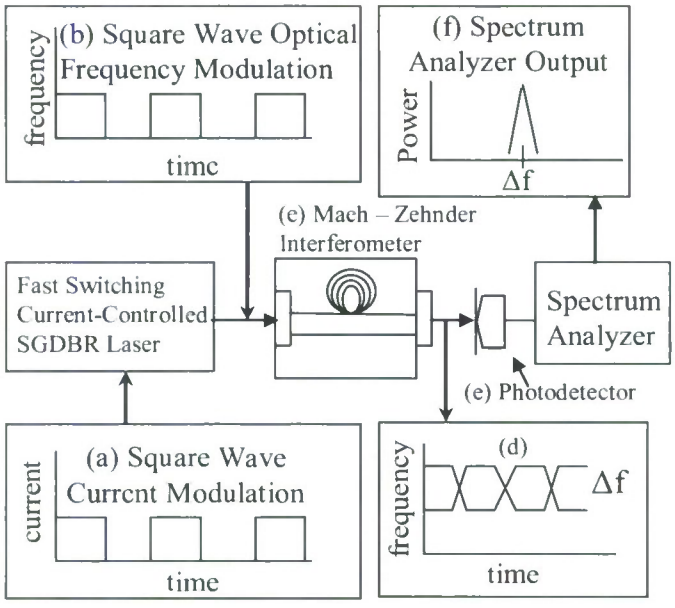


Fig. 1. The block diagram used to generate microwave signals using SG-DBR lasers is shown. The laser is current modulated (a) with a square wave current waveform. A square wave optical frequency versus time (b) results. The output of the interferometer (c) has the two optical frequencies present at the same time. The difference frequency (d) is photodetected (e) and applied to the electrical spectrum analyzer for analysis (f).

Using this self-heterodyne approach, microwave and millimeter wave signals can be generated using a low frequency square wave. Photodetector bandwidth measurements can be made using this self-heterodyne signal as the excitation source. This compact source can frequency hop with sub-microsecond update rates over arbitrary frequency patterns from DC to the millimeter wave bands. The frequency range and power of the microwave and millimeter wave signals are limited by the photodetector characteristics. Photodetector structures with several hundred GHz bandwidth should enable signal generation beyond the range of traditional microwave sources and frequency multiplier circuits. A limitation on the usefulness of these sources in applications is presently the spectral width of the optical sources used for self-heterodyne. Section II will contrast this work to previous examples of signal generation by self-heterodyne methods.

II. MICROWAVE SIGNAL GENERATION USING SELF-HETERODYNE REVIEW

The self-heterodyne method for microwave signal generation described in this paper was initially described in references [3], [4], and [5]. References [3] and [4] used a distributed feedback laser in a self-heterodyne configuration similar to that shown in Fig. 1. Triangular and rectangular current waveforms were injected into the laser bias. The distributed feedback laser produced both intensity and frequency changes in response to the current waveforms. Microwave signals up to 4 GHz were reported. Reference [4] reported on the use of SG-DBR lasers using the self-heterodyne configuration of Fig. 1. Frequencies as with R (FM) that was applied to the Mach-Zehnder interferometer. The optical frequency difference frequency was photodetected and the generated microwave signals were observed with an electrical spectrum analyzer. Reference [4] demonstrated self-heterodyne microwave generation with a full-band wavelength tunable SG-DBR laser. The SG-DBR approach utilized only FM modulation and showed dramatically reduced microwave signal spectral width. Self-heterodyne of a dual-wavelength fiber laser has also been reported [5].

Optically injection locked semiconductor lasers have recently been used in photonic microwave generation [6]. A common method of generating microwave signals from optical components utilizes a master-slave laser arrangement. In a master-slave configuration with sufficient injection power, the optical frequency of the slave laser is injection locked to that of the master laser. By properly adjusting the injection strength and the frequency detuning, instability occurs through Hopf bifurcation into the period-one dynamical state. The result is a tunable microwave oscillation dependent on the optical power of the slave laser [6].

A third eategory used to ereate mierowave signals is based on an external modulation teehnique. The external modulator ean either be an eleetro-optie intensity modulator or a phase modulator. A system that eould generate millimeter-wave signals using an external intensity modulator was proposed in 1992 by O'Reilly et al. [7]. A frequency-doubled eleetrieal signal was optieally generated by biasing the intensity modulator to suppress the optieal earrier and the even-order optieal sidebands. A 36 GHz millimeter-wave signal was generated when the intensity modulator was driven by an 18 GHz mierowave signal. A key advantage of this approaeh is that frequeney-doubled signals ean be generated with a relatively low-speed external modulator. However, similar to approaehes in the first eatcgory, a high quality mierowave referenee souree is required [7].

III. SG-DBR LASER CHARACTERISTICS

The laser ehosen for this work was the JDSU Ine. MKS-063-B widely tunable laser that was internally modified by the manufaeturer to allow for fast wavelength switehing. Fig. 2 shows a top view of the SG-DBR laser diode ehip used in the experiments. This laser ehip has five segments requiring separate eurrent biases. The gain seetion provides for the amplifieation needed to obtain stimulated emission. The baek mirror and front mirror form eurrent-eontrolled wavelength dependent refleetivity to tune the wavelength of the laser. The phase seetion performs fine "eurrent-eontrolled stretehing" of the laser eavity to make small adjustments in the wavelength of the laser. The semieonduetor optieal amplifier (SOA) adjusts the laser output power.

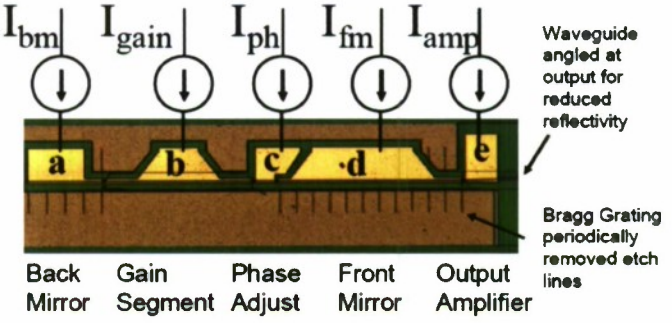


Fig. 2. Top view of the SG-DBR laser utilized in this experiment: The magnitude of currents driving the back mirror (a), front mirror (d), and phase segment (c) control the wavelength of the laser. The gain section (b) and semiconductor optical amplifier (e) control the laser output power.

SG-DBR single-ehip monolithie tunable lasers were originally developed for teleeommunieation applieations. These deviees allow for a wide wavelength tuning range over the erbium-doped fiber amplifier operational band of 1525 to 1565 nm, or the extended L-band of 1565 to 1605 nm. Referenees [1], [2], and [8] provide illustrations of the deviee design and the meehanism of wavelength eontrol in the SG-DBR laser. Similar single ehip tunable lasers sueh as the Y-Junetion variant are also suitable for this applieation [9].

Fig. 3 illustratcs tuning eharacteristics for the deviee used in the experiments of this paper. The wavelcngth as a function of the front mirror and baek mirror eurrents are given. The phase seetion eurrent is held eonstant at 5.5 mA in the measurement. In order to get a smooth ehange of frequeney as a funetion of bias, one must ineremeni the front mirror and baek mirror eurrents simultaneously. A series of paths where the laser wavelength can be continuously tuned is shown in Fig. 3. One has to concatenate paths in order to cover the complete frequency band of the SG-DBR laser. The microwave signals generated in this work involve tuning between wavelengths on a single path of the laser-tuning curve.

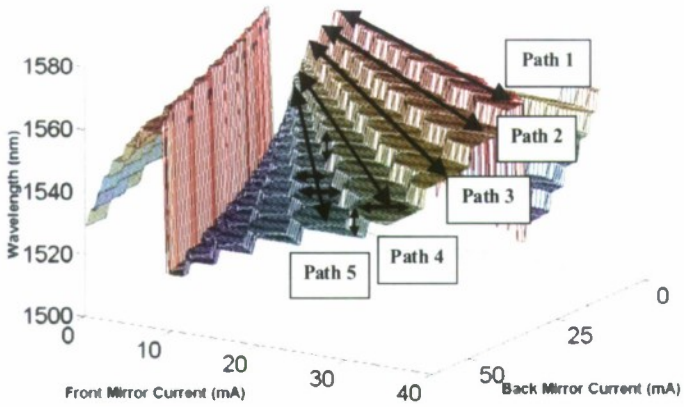


Fig. 3. A Three-dimensional view of the SG-DBR wavelength tuning characteristics is given. The horizontal axes represent the current drive to the front mirror and back mirror segments of the laser. The phase current is held constant at 5.5 mA. The illustrated paths show regions where a continuous change in wavelength is available. Path 1: 1565 – 1560 nm; Path 2: 1560 - 1555 nm; Path 3: 1555 – 1550 nm; Path 4: 1550 - 1545 nm; Path 5: 1545 - 1540 nm. The horizontal arrows represent the 1 nm 'tiled' regions between wavelength paths.

Any laser output wavelength from 1525 to 1565 nm can be produced by generating a wavelength tuning-look up table that utilizes current changes in the front mirror, back mirror and phase segments of the SG-DBR laser. For microwave signal generation, only a small wavelength change is necessary. The entire laser tuning range corresponds to a frequency difference of 5 THz. This self-heterodyne method will utilize very small wavelength changes to provide difference frequencies in the microwave and millimeter wave bands. The front mirror, back mirror and the phase section are the wavelength control segments of the laser. Fig. 4 shows the DC tuning curves of the front mirror, back mirror and phase section of the laser. Both the optical power and frequency are shown as a function of DC bias. The front mirror and back mirror currents show the 5 nm wide stair-step wavelength changes that occur when traversing the tuning curve in Fig. 3 with only one current changing. The phase-section tuning curve is of particular interest in that only a small wavelength range is covered. The phase section output power is also only slightly dependent on the phase section bias current. The output power changed by only 0.2 dB over most of the phase section tuning range. Microwave signal generation using self-heterodyne has been accomplished using the front mirror, back mirror and phase sections. Modulation of the phase section is the focus of this paper. Square wave current modulation of only a few milliamps in the phase section is required to produce difference frequencies in the microwave and millimeter frequency ranges up to 60 GHz.

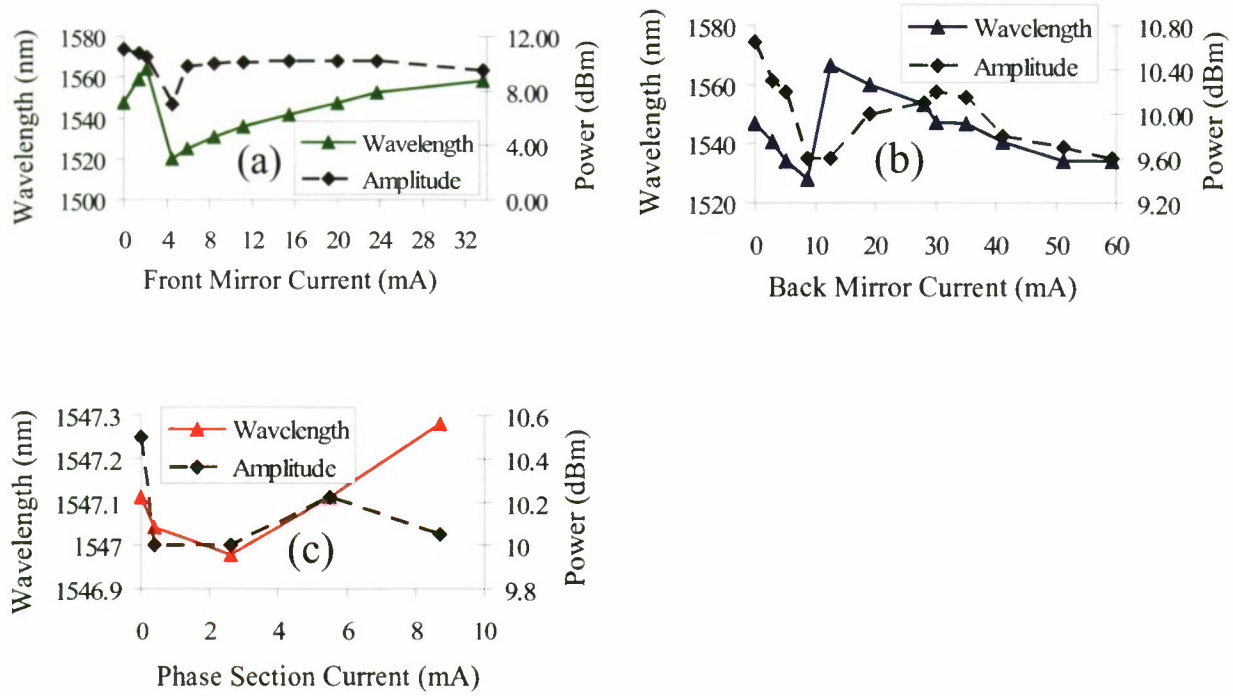


Fig. 4. (a) The wavelength and optical power as a function of current applied to the front mirror is shown. The other laser segments are biased as follows; back mirror: 30 mA, phase section: 5.5 mA, gain: 100 mA, SOA: 150 mA. (b) The wavelength and optical power as a function of current applied to the back mirror is shown. The other laser segments are biased as follows; front mirror current: 20 mA, phase section: 5.5 mA, gain: 100 mA, SOA: 150 mA. (c) The wavelength and optical power as a function of current applied to the phase section is shown. The other laser segments are biased as follows; front mirror current: 20 mA, back mirror current: 30 mA, gain: 100 mA, SOA: 150 mA. These three graphs illustrate the course tuning (front mirror and back mirror) and fine tuning control (phase section) of the SG-DBR laser. The back mirror, front mirror and gain segments primarily control the wavelength and produce a small amount of change in the optical power. The phase section offers continuous wavelength tuning over a narrow range. The front mirror and back mirror offer discontinuous tuning (typically in 5 nm steps) over the full range of the laser.

IV. Microwave signal generation

Fig. 6 shows the measurement diagram and initial results for microwave signal generation using self-heterodyning of an SG-DBR laser. The output of the laser is coupled into a single mode fiber pigtail and sent into a Mach-Zehnder interferometer with a 3.5 ps path time delay difference. A square wave drive current with a repetition frequency of 1/(7ps) was applied to the phase section of the SG-DBR laser. The square wave current modulation causes the laser to switch back and forth between two adjacent optical frequencies. The long delay path of the interferometer builds up a 0 to 3.5 □s time record at λ_1. The current drive then forces the laser to λ_2 for the 3.5 ps to 7 ps portion of a modulation period. The output of the Mach-Zehnder interferometer therefore has λ_1 and λ_2 overlapping in time. The interferometer beats the two switching optical frequencies together and the resulting output of the photodetector/preamplier is a sinusoid at the difference frequency [4].

The linewidth of the laser source is a key parameter controlling the spectral width of the microwave difference frequency output. Fig. 7a shows the results of a homodyne measurement of the laser linewidth for the case of no applied square wave modulation. The homodyne difference frequency is centered at 0 Hz. The – 3 dB bandwidth of the signal centered at DC is 28.7 MHz. This unmodulated width represents a limit to the width of microwave signals generated in the self-heterodyne case since laser linewidth typically increases under modulation.

Fig. 6b and 6c presents examples of microwave signals generated via self-heterodyning the laser output. The measurement set up for these results is shown in figure 6d. All segments except for the phase section are DC biased. The phase section has a 143 kHz square wave current modulation with varying levels of amplitude. The highest frequency microwave signals measured were at 12 GHz. This frequency was limited by our photodetector and amplifier inventory.

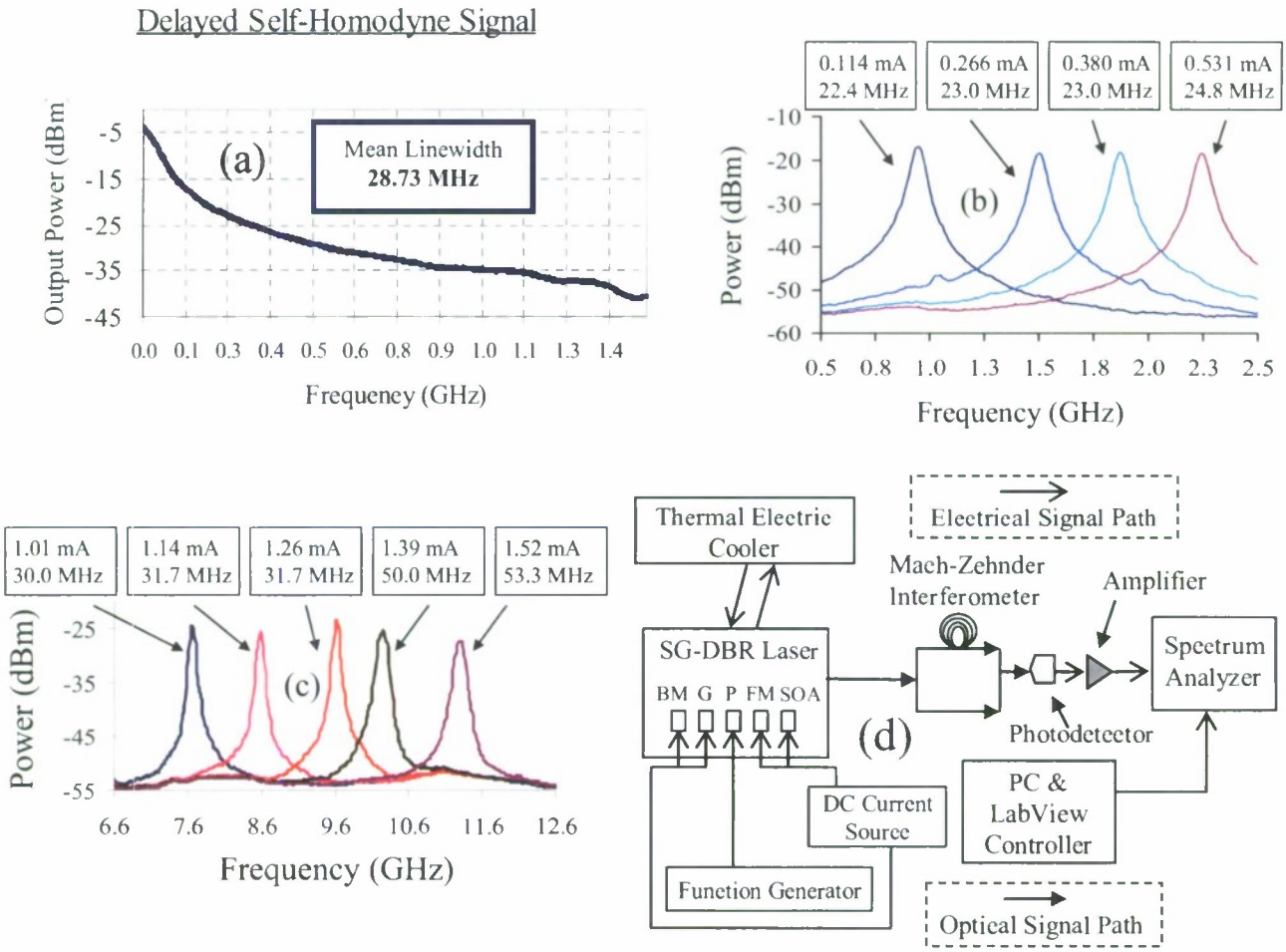


Fig. 6. (a) The self-homodyne linewidth of the signal is shown. back mirror: 30 mA, front mirror: 20 mA, phase: 5.5 mA, gain: 100 mA, SOA: 150 mA. (b) Low frequency microwave signals generated via delayed self-heterodyning. The top number in each box is the DC bias current on the phase section and the bottom number is the 3 dB spectral width. (c) High frequency microwave signals generated via delayed self-heterodyning (d) Experimental setup for obtaining delayed homodyne or self-heterodyne spectrum from the Mach-Zehnder interferometer output. The spectrum analyzer resolution bandwidth is 3 MHz for all measurements.

The average linewidth of the laser measured from the low frequency microwave signals is approximately 23 MHz, and the average linewidth measured from the high frequency microwave signals is approximately 40 MHz. The data shows that the higher the modulation current amplitude, the broader the linewidth. However, these results show

consistency between the unmodulated and modulated linewidth measurements. The 28 MHz measured linewidth from the homodyne signal of fig. 6a is roughly between 23 and 40 MHz.

It is expected that the linewidth of the laser will be increased under modulation compared to the unmodulated case. Linewidth broadening can occur due to low frequency temperature transients during wavelength switching, broadband noise coupling into the frequency modulation segments and the stability of the current sources switching the wavelength of the laser. The linewidth of the laser is also dependent on the DC bias condition of the laser. Fig. 7 shows a map of the variations in measured linewidth and its dependency on front and back mirror current tuning. The linewidth was measured using the homodyne method shown in figure 6a. The narrowest linewidth signals are found when small currents are applied to the front and back mirror segments.

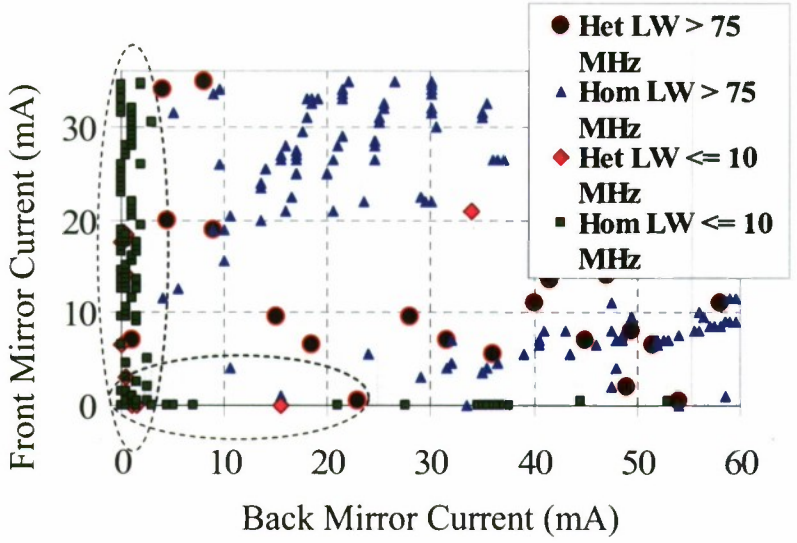


Fig. 7. Two-dimensional tuning map identifying areas where laser linewidth is greater than 75 MHz and less than 10 MHz. Both unmodulated homodyne results (Hom LW) and modulated self-heterodyne (Het LW) results are indicated. The regions enclosed by the dashed ovals had the narrowest linewidth results

A heavy concentration of 10 MHz measured linewidths occurs when the DC bias of the back mirror is between 0 and 5 mA. The linewidth is less dependent on front mirror bias condition. The microwave signal generation method discussed in this work only requires a small wavelength change around a DC operation point established by the front mirror and back mirror bias. Fig. 7 shows that there are specific bias points for narrowest spectral width microwave signal generation.

V. High speed package and response measurements

SG-DBR lasers were originally designed for telecommunication applications that did not utilize the fast wavelength switching capabilities of the laser. Most telecommunication applications utilize integrated low pass filters into the laser package in order to minimize the potential for externally introduced noise affecting the spectral content of the laser. However, for applications requiring fast wavelength switching, such as microwave signal generation, it is necessary to remove the low pass filters. Fig. 8 shows a close up of the packaged SG-DBR laser assembly. Semi-rigid coaxial cables are connected to the pins of the butterfly package to maximize coupling of fast switching time square wave waveforms to the laser.

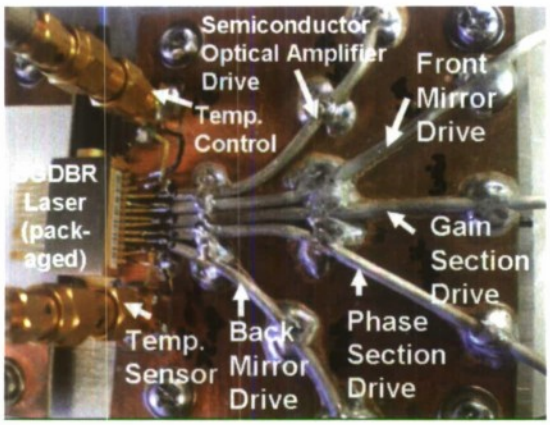


Fig. 8. A modified 14-pin butterfly-packaged packaged laser with internal low pass filters removed is shown. The leads are fed with 0.085-inch diameter semi-rigid coaxial cable to allow high-speed (modulation) drive of the device.

The butterfly leads were suspended over an RT-Duroid dielectric surface to create an environment close to 50 Ω in the butterfly package leads. For heat sink purposes, the entire unit was mounted onto an aluminum block.

The frequency modulation response of the laser was tested to determine the potential wavelength switching speed. Fig. 9 shows the apparatus that was used to measure the FM frequency response of the packaged laser. A sinusoidal AC current waveform was applied to the phase section of the laser inducing sinusoidal frequency modulation at the laser output. The laser output was then applied to a 0.1 nm passband width band pass filter. The laser output was aligned to the skirt of the filter lower in wavelength than the pass band peak. As a result, the optical wavelength was offset from the wavelength of maximum transmission. An FM to intensity modulation converter is formed by the band pass filter skirt. By changing the frequency of the sinusoidal modulation current waveform, the frequency modulation response of the laser was obtained. The photodetector senses the intensity modulation and applies it to an electrical spectrum analyzer. The FM frequency response measurement shown in Fig. 10 indicates that the laser package can be efficiently modulated up to 100 MHz.

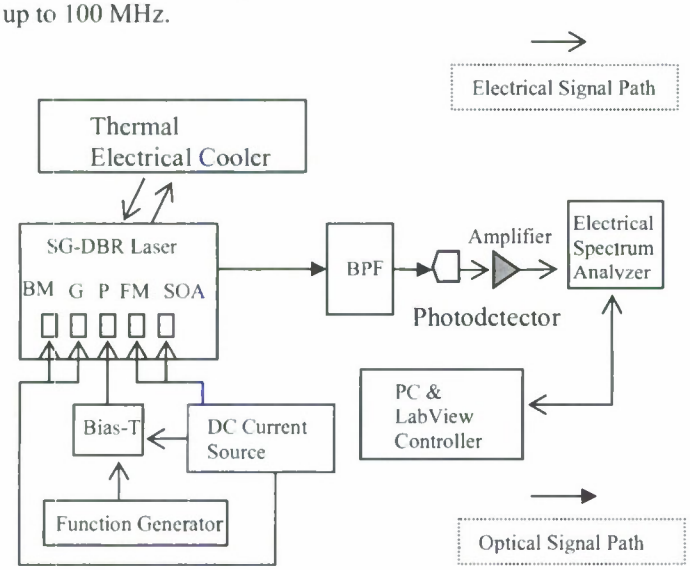


Fig. 9. The experimental setup used to measure FM modulation efficiency is shown. A wavelength tunable band pass filter (BPF) was adjusted so that the laser wavelength was on the skirt of the filter. The BPF serves as a wavelength discriminator converting

frequency modulation into intensity modulation. The frequency deviation as a function of the sinusoidal current modulation to the phase section was measured.

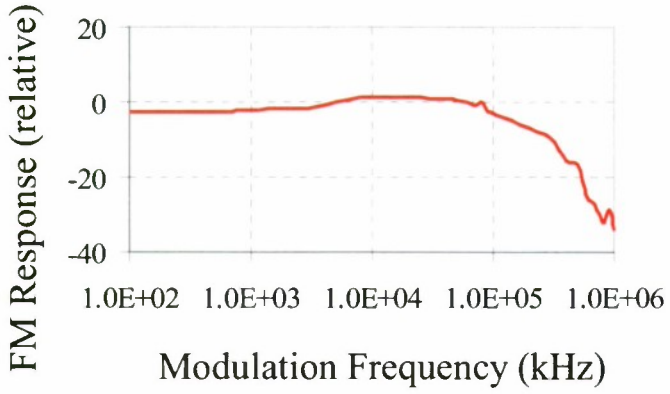


Fig. 10. The frequency modulation (FM) efficiency of the phase segment of the SG-DBR laser with respect to modulation frequency is shown. Fig. 9 showed the measurement set up. Efficient modulation frequencies up to 100 MHz were demonstrated.

VI. time resolved frequency performance

Linewidth broadening under square wave modulation is a fundamental concern for narrow spectral width microwave signal generation. A time resolved frequency step measurement was performed to quantify the time interval over which the microwave signal is off, and to identify a mechanism for microwave signal broadening above the unmodulated linewidth performance of the laser. The experimental setup for the time resolved experiment was similar to the frequency respone measurement setup shown in Fig. 9 with an oscilloscope replacing the spectrum analyzer. The square wave current modulation was directly coupled to the phase section of the laser.

A time resolved frequency measurement result is shown in Fig. 11. The laser output was placed at a wavelength that was longer than that of the filter passband for measurement A. The bandpass filter acts as a wavelength discriminator turning frequency changes into inensity changes that are photodetected and measured on the oscilloscope. The process was repeated with the laser wavelength placed shorter than the filter passband in position B. In position C the laser was placed at the center of the passband where FM to intensity modulation is minimized. The circles in fig. 11 highlight thermal transients effecting the frequency switching time of the laser. The transients have been identified to be on the order of 200 ns in duration. The presence of thermal transients broadens the width of the microwave signal generated by the self heterodyne technique.

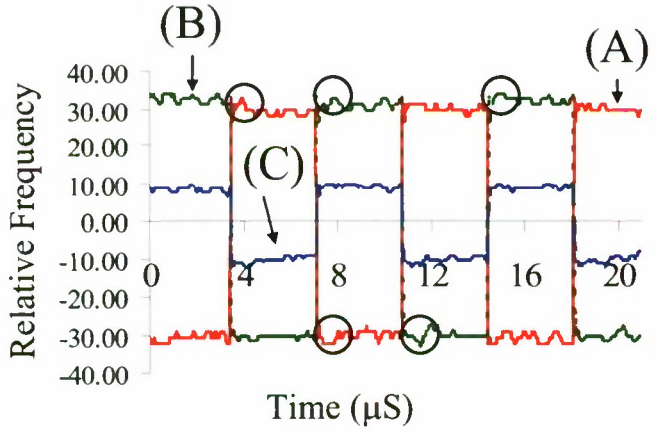


Fig. 11. Optical frequency as a function of time for a square wave current modulation at 143 kHz is shown. The circled regions highlight the overshoot locations of the optical frequency caused by thermal transients. Trace A is for the case of the laser wavelength

positioned longer that the filter passband. Trace B if for positioning shorter than the filter passband. Trace C is for centering of the laser on the filter passband.

The switching time of the SG-DBR laser was also observed from the time resolved step measurement. Fig. 12 shows the transition time between the two optical wavelengths. A wavelength switching time of approximately 40 ns was observed. This time is dominated by the the rise time of the applied current modulation waveform.

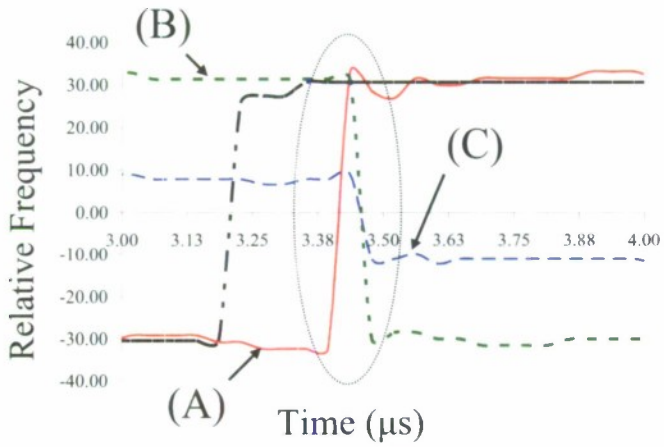


Fig. 12. The frequency versus time is shown near the transition point of the square wave modulated laser. The circled area shows the switching time between sequential optical frequencies. Trace A is for the case of the laser wavelength positioned longer that the filter passband. Trace B if for positioning shorter than the filter passband. Trace C is for centering of the laser on the filter passband.

Recall from [2], switching times of 5 ns have been achieved for this class of diode lasers. However, these results were obtained only under pre-emphasis switching conditions where amplitude overshoot or undershoot conditions of the RF current square waveform were nonlinearly emphasized. Recalling the frequency response measurements presented in Fig. 10, switching times on the order of 10 ns (1/100MHz) should be possible. The microwave signals generated by this method thus are continuous wave with the exception of the the 40 ns switching time of the laser wavelength output.

VII. CONCLUSION

Microwave signal generation using a single wavelength tunable SG-DBR laser has been demonstrated. This technique does not require a master laser synchronized to a slave laser, an external microwave drive source, or a mode-locked laser. Microwave signals up to 12 GHz have been demonstrated. Linewidth measurements have been taken from the delayed self-homodyne signals showed spectral widths between 22 and 53 MHz. Extensive linewidth characterization was performed to show optimal bias conditions for the SG-DBR laser segments in order to generate microwave signals.

A high frequency laser package was designed to minimize the transition time during wavelength switching. Frequency domain and time-resolved frequency response experiments were performed to characterize the packaged laser response. Wavelength switching times of 40 ns were produced.

ACKNOWLEDGMENT

The authors thank Xiaomin Jin and Susan Portugal for help with the tunable band pass filters. Special thanks also goes to Michael Biller for supplying low frequency monolithic amplifiers, and Kyle Woolrich for consultation with high & low frequency amplifier design. Thanks go to Greg Fish at JDSU for SG-DBR assistance. Leif Johansen and Matt Sysak UC-Santa Barbara provided help with SGDBR laser chips.

5. ACKNOWLEDGEMENTS

This work was sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152."

REFERENCES

[1] Gert Sarlet, Geert Morthier, and Roel Baets, "Control of Widely Tunable SSG-DBR Lasers for Dense Wavelength Division Multiplexing," *IEEE Trans. Journal of Lightwave Technology,* vol. 18, no. 8, pp. 1128-1138, August 2000.

[2] John E. Simsarian, Michael C. Larson, Henry E. Garrett, Hong Xu, and Timothy A. Strand, "Less Than 5-ns Wavelength Switching With an SG-DBR Laser," *IEEE Photonics Technology Letters*, vol. 18, pp. 565-567, February 2003.

[3] Gabor Kovacs, Tibor Berceli, "A Novel Approach to Microwave Signal Generation Utilizing DFB-Laser Wavelength Chirp," *2nd European Microwave Integrated Circuit Conference*, pp. 528-531, October 2007, Munich Germany.

[4] Michael A. Bernacil, Shane O'Connor, "Microwave Signal Generation Using Self-heterodyning of a Fast Wavelength Switching SG-DBR Laser," IEEE *International Microwave Symposium: IMS 2008,* paper WE4D-05, June 2008.

[5] Xiangfei Chen, Zhichao Deng, and Jianping Yao, "Photonic Generation of Microwave Signal Using a Dual-Wavelength Single-Longitudinal-Mode Fiber Ring Laser," *IEEE Trans. Microwave Theory and Techniques*, vol. 54, pp. 804-809, Febuary 2006.

[6] Sze-Chun Chan and Jia-Ming Liu, "Frequency Modulation on Single Sideband Using Controlled Dynamics of an Optically Injected Semiconductor Laser," *IEEE Journal of Quantum Electronics*, vol. 42, pp. 699-705, July 2006

[7] J.J. O'Reilly, P.M. Lane, R. Heidemann, and R. Hofstetter, "Optical generation of very narrow linewidth millimeter wave signals," *Electron. Lett.*, vol. 28, no. 25, pp. 2309-2311, 1992.

[8] Dennis J. Derickson and Michael Bernacil, "SGDBR Monolithic Wavelength Tunable Lasers for Swept Source OCT," *Biomedical Optics: BIOS 2008,* SPIE Vol. 6847, January 2008.

[9] J.-O. Wesström, G. Sarlet, S. Hammerfeldt, L. Lundqvist, P. Szabo, P.-J. Rigole, "State-of-the-art performance of widely tunable modulated grating Y-branch lasers" *Proc. OFC 2004*, paper TuE2, Los Angeles, CA, February 2004.

Michael Bernacil received his M.S.EE degree from California Polytechnic State University. He now works for Raytheon Space and Airborne Systems in El Segundo, CA. He received his B.S. from the University of California, Davis in 2005 in the field of Optical Science and Engineering. His research experience includes extensive work for the National Ignition Facility Program at Lawrence Livermore National Laboratory.

Shane O'Connor received is MSEE from California Polytechnic State University. He now works for Raytheon Space and Airborne Systems in El Segundo, CA. His interests include test and design of analog and RF hardware. Shane will begin an exciting career in the RF field with Raytheon Space and Airborne Systems in El Segundo, CA.

Ben Maher is an M.S. graduate from California Polytechnic State University.
Andrew Dekelaita is an M.S EE graduate from California Polytechnic State University.

Dennis Derickson (S'78 -M'83 -SM'06) is an assistant professor of electrical engineering at California Polytechnic State University. He has edited two books "Fiber Optic Test and Measurement" Prentice Hall 1998 and "Digital Communications Test and Measurment" Prentice Hall 2008.

Summary:

This report gave a summary of the work done by Dennis Derickson, Brandon George, Ben Maher, Shane O'Connor, Mike Bernacil and Andrew Dekclaita in over the period of Jan 2008 to Dec. 2008. Table 1 below gives a high level summary of the work with 4 areas highlighted. The significance to the office of Naval Research is also summarized. This report essentially contains the results of four papers that were published over this time period. If a more detailed account of the work is desired, a set of four Master's thesis documents are cited below.

Table 1: The key outcomes of this work are highlighted in the following summary table along with the associated relevance to the project sponsor.

Item	Discussion	Significance to the Office of Naval Research
LIDAR	The initial proposal for this project was to demonstrate Sample Grating Distributed Bragg Reflector Laser's applications for LIDAR systems. Section 2 of this report highlights the LIDAR results that were achieved and reported at the IEEE LEOS annual meeting.	A small size, inexpensive, fast update rate, swept wavelength lasers source was demonstrated for the LIDAR application. This device should be important for portable LIDAR systems where fast update rate is essential.
OCT	This same high repetition rate source was used for optical coherence tomography measurements. Section 1 of this report highlights the investigation results. This work was reported at SPIE's photonics west conference in January of 2009.	OCT is becoming an important method of non-invasive examination of reflectivity versus distance in biological samples. The significance of this work is demonstrating that an inexpensive portable laser source can work in OCT applications.
Microwave Signal Generation	Section 3 of the report shows that these fast tunable lasers have application to generation of microwave and millimeter wave signals.	Optical methods of generating millimeter wave signals with a delayed self heterodyne technique are shown. This source along with high speed optical detectors could be a viable method for signal generation in the 100s of GHz region.
Optical Sensing Interrogation	The fast sweep time of the SGDBR tunable laser allows for 10 μS time resolution of optical sensors.	This work allows for a significant advancement in time-resolved measurements of optical sensors which rely on changes in the frequency response.

Publications that were generated as part of this project form the basis for the report. The references are as follows:

1. Shane O'Connor, Michael A. Bernacil, Andrew DeKelaita, Ben Maher, and Dennis Derickson "100 kHz Axial Scan Rate Swept-Wavelength OCT using Sampled Grating Distributed Bragg Reflector Lasers" in *Coherence Domain Optical Methods and Optical Coherence Tomography in Biomedicine XIII*, edited by Joseph A. Izatt, James G. Fujimoto, Valery V. Tuchin, Proceedings of SPIE Vol. 7168 (SPIE, Bellingham, WA 2009).

2. Shane O'Connor, Michael A. Bernacil, and Dennis Derickson "Generation of High Speed, Linear Wavelength Sweeps Using Sampled Grating Distributed Bragg Reflector Lasers" 2008 IEEE LEOS Annual Meeting, Newport Beach, CA, paper TuB 2

3. Michael A. Bernacil, Shane O'Connor, Ben Maher, Andrew DeKelaita, and Dennis Derickson, "Microwave Signal Generation Using Single-Chip Fast Wavelength-Tunable Sampled Grating Distributed Bragg Reflector Lasers," IEEE *International Microwave Symposium: IMS 2008,* paper WE4D-05, June 2008

4.Derickson, D., Bernacil, M., DeKelaita, A., Maher, B., O'Connor, S., Sysak, M. N., Johanssen, L., "SGDBR single-chip wavelength tunable lasers for swept source OCT" in *Coherence Domain Optical Methods and Optical Coherence Tomography in Biomedicine XII*, edited by Joseph A. Izatt, James G. Fujimoto, Valery V. Tuchin, Proceedings of SPIE Vol. 6847 (SPIE, Bellingham, WA 2008) 68472P.

More Information on this Work is Available in the Following MS Thesis Documents at Cal Poly

1. **Shane O'Connor,** *" High Speed Wavelength Tuning of SG-DBR Lasers for Light Detection and Ranging and Optical Coherence Tomography"* MS Thesis, California Polytechnic State University, May 2008

2. **Mike Bernacil**, " *Microwave Signal Generation Using Self-Heterodyning of a Fast Wavelength Switching SG-DBR Laser*", MS Thesis, California Polytechnic State University, May 2008

3. **Ben Maher**, "*High Speed Wavelength Tuning for Optical Coherence Tomography Applications*", MS Thesis, California Polytechnic State University, Dec 2008

4. **Andrew DeKelaita**, *"Open-loop Tuning of SG-DBR Widely Tunable Lasers for Optical Coherence Tomography Applications"* MS Thesis, California Polytechnic State University, Dec 2007

Publications that were generated as part of this project form the basis for the report. The references are as follows:

1. Shane O'Connor, Michael A. Bernacil, Andrew DeKelaita, Ben Maher, and Dennis Derickson "100 kHz Axial Scan Rate Swept-Wavelength OCT using Sampled Grating Distributed Bragg Reflector Lasers" in *Coherence Domain Optical Methods and Optical Coherence Tomography in Biomedicine XIII*, edited by Joseph A. Izatt, James G. Fujimoto, Valery V. Tuchin, Proceedings of SPIE Vol. 7168 (SPIE, Bellingham, WA 2009).

2. Shane O'Connor, Michael A. Bernacil, and Dennis Derickson "Generation of High Speed, Linear Wavelength Sweeps Using Sampled Grating Distributed Bragg Reflector Lasers" 2008 IEEE LEOS Annual Meeting, Newport Beach, CA, paper TuB 2

3. Michael A. Bernacil, Shane O'Connor, Ben Maher, Andrew DeKelaita, and Dennis Derickson, "Microwave Signal Generation Using Single-Chip Fast Wavelength-Tunable Sampled Grating Distributed Bragg Reflector Lasers," IEEE *International Microwave Symposium: IMS 2008,* paper WE4D-05, June 2008

4.Derickson, D., Bernacil, M., DeKelaita, A., Maher, B., O'Connor, S., Sysak, M. N., Johanssen, L., "SGDBR single-chip wavelength tunable lasers for swept source OCT" in *Coherence Domain Optical Methods and Optical Coherence Tomography in Biomedicine XII*, edited by Joseph A. Izatt, James G. Fujimoto, Valery V. Tuchin, Proceedings of SPIE Vol. 6847 (SPIE, Bellingham, WA 2008) 68472P.

More Information on this Work is Available in the Following MS Thesis Documents at Cal Poly

1. **Shane O'Connor,** *" High Speed Wavelength Tuning of SG-DBR Lasers for Light Detection and Ranging and Optical Coherence Tomography"* MS Thesis, California Polytechnic State University, May 2008

2. **Mike Bernacil**, *" Microwave Signal Generation Using Self-Heterodyning of a Fast Wavelength Switching SG-DBR Laser"*, MS Thesis, California Polytechnic State University, May 2008

3. **Ben Maher**, *"High Speed Wavelength Tuning for Optical Coherence Tomography Applications"*, MS Thesis, California Polytechnic State University, Dec 2008

4. **Andrew DeKelaita**, *"Open-loop Tuning of SG-DBR Widely Tunable Lasers for Optical Coherence Tomography Applications"* MS Thesis, California Polytechnic State University, Dec 2007

An experimental and theoretical study of the electro-optical response of a new type of ferroelectric liquid crystal

Project Investigators:

Jonathan Fernsler, Saimir Barjami,
Karl Saunders
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

Project title: An experimental and theoretical study of the electro-optical response of a new type of ferroelectric liquid crystal

Project aim: The aim of this collaborative project is to conduct a threefold investigation of a technologically promising class of ferroelectric liquid crystals, known as de Vries liquid crystals, which exhibit a colossal electro-optical response. The unusual electro-optical properties of these de Vries materials have generated significant interest for their potential use in liquid crystal displays and photonic devices. This project will simultaneously proceed on three fronts, two experimental and one theoretical. Dr Fernsler will experimentally investigate the relationship between the electro-optical response of the materials and their orientational order. Dr Barjami will conduct calorimetric measurements on different materials to correlate thermodynamic nature of the smectic A-C transition with the size of the optical response. Dr Saunders will use an established theoretical model to gain insight to the interconnected relationships between the optical response, orientational order and thermodynamic nature of the A*-C* transition. Combining the results of these three investigations will provide unprecedented insight to these materials, particularly in their incorporation and optimization for electro-optical devices

Summary

Results

Optical Response and Orientational Order (Dr Fernsler)

Considerable progress has been made in the electro-optic measurements of de Vries liquid crystals and some surprising results have led us to pursue additional investigations. Three Cal Poly undergraduate students have built the apparatus' and conducted experiments: Danial Staines, Reymundo Ortiz, and Austin Havens.

Preparation of Liquid Crystal Cells

We have built custom-designed equipment to prepare liquid crystal (LC) cells for experimentation on our project. LC cells consist of glass plates, coated with a transparent conducting layer of indium tin oxide (ITO), followed by a rubbed-nylon alignment layer, with a ~4-12 μm thick LC layer sandwiched between plates. This geometry allows us to observe aligned LC materials and apply electric fields to these samples between the plates (in a similar fashion to a parallel plate capacitor) to measure electro-optic response. To accomplish this fabrication, students built the following devices: A custom-made dip-coater applies a layer of nylon to glass-slides; a rubbing device, which rubs the nylon-coated cells to align the liquid crystal molecules; a system to measure cell thickness analyzes interference fringes using a spectrometer on an unfilled cell.

Birefringence

Measurements of the birefringence, Δn, versus temperature have been accomplished on two "de Vries"-type liquid crystal materials: an achiral fluorinated compound, 8422, and a siloxane material, TKSiKN65. We also plan to examine a chiral homologue of 8422, 8422[2F3]. The results of these measurements showed behavior that is consistent with Dr. Saunders's predictions from his previous model (see Fig. 1): birefringence increases while cooling from the high-temperature isotropic

phase in the SmA phase, but near the transition to SmC, the birefringence drops. This indicates increasing orientational order when cooling high in the SmA phase, followed by an unusual decrease in orientational order near the transition. In the SmC phase, the presence of arbitrarily-ordered domains makes measurement of birefringence impossible without an aligning electric field. A third de Vries material, W415, showed completely different behavior: the birefringence increased almost linearly with cooling throughout the SmA and SmC phases. This behavior was not expected in a de Vrics liquid crystal, and we wish to investigate this material further. However, the sample we obtained has chemically degraded which has limited our present inquiries: we hope to obtain more of this material and a "standard" smectic material to compare with our other de Vries materials in the future.

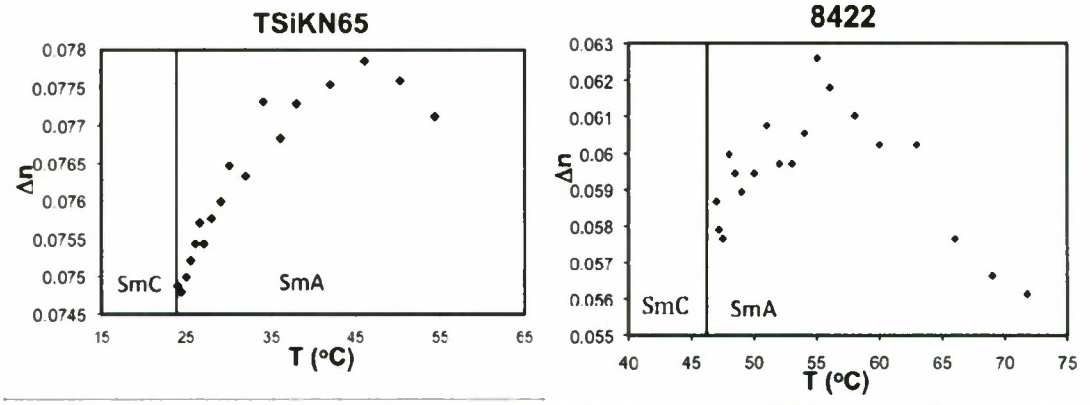


Figure 1 Birefringence, Δn, versus temperature for TsiKN65 (left) and 8422 (right) in the SmA phase.

While our data supports the overall trends predicted by Dr. Saunders's theory, the precise behavior near the SmA-SmC transition may indicate whether these transitions are near the tricritical point as supported by Dr. Barjami's data below. We are implementing an interference filter and a photometric calculation from our camera system to obtain more precise measurements of birefringence for these and other materials.

Our camera system supplied by this grant has allowed us to record photo-micrographs of liquid crystal textures. Figure 2 shows the material 8422 doped with 5 wt% of its chiral homologue, 8422[2F3], in the SmA (66°C) and SmC (36°C) phases.

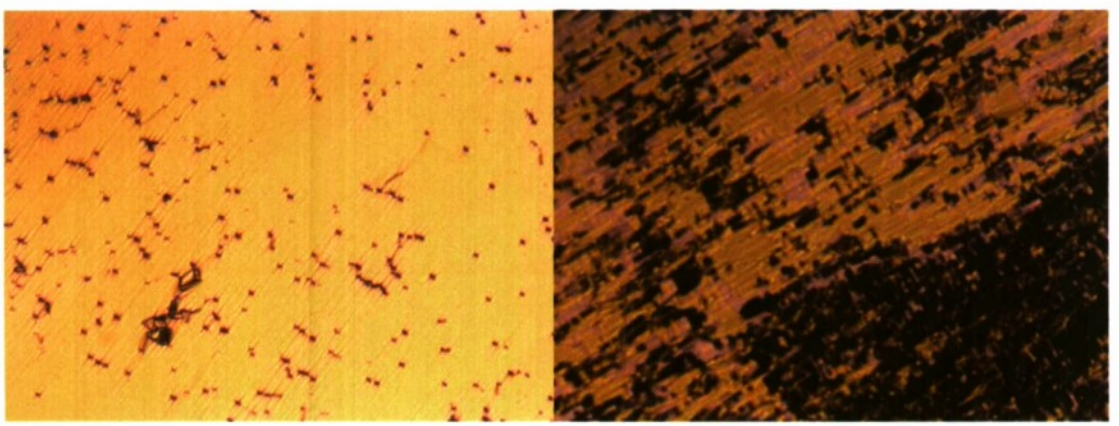


Figure 2 8422 doped with 5 wt% of its chiral homologue, 8422[2F3], at 66°C in the SmA phase (left) and 36°C in the SmC phase (right).

In the SmA phase, the material is well-aligned with a uniform texture. Upon cooling into the SmC phase, the molecular director (average direction of the long axis of

molecules) tilts and domains with different orientations are observed (see Figure 2). An electric field can orient the domains in the SmC phase along one orientation, producing a uniform texture.

Electro-optics

The SmC phase of chiral liquid crystals is ferroelectric: a spontaneous polarization develops that is coupled to the orientation of the molecular director (the average long axis of the molecules). We have applied a DC electric field to our samples to examine the effects on birefringence and molecular orientation. The material 8422 doped with 5 wt% of 8422[2F3] was switched to saturation at an extremely low field of ~1 V/μm: this extremely high susceptibility is a hallmark of de Vries materials that is particularly impressive considering this is an achiral material doped with only a small amount of chiral material. Although no changes were observed with an applied electric field high in the SmA phase, near the SmA-SmC transition we observed an induced tilt within ~3 °C of the transition and small changes in the birefringence (see Figure 3). Our present measurements of birefringence were insufficiently sensitive to record these pretransitional changes, but we expect that our new system should be able to observe these field effects. Below the SmA-SmC transition, the birefringence and tilt angle, θ, increase rapidly and appear to saturate at ~ 37 °C. This large increase in birefringence is consistent with the predictions of Dr. Saunders's theory.

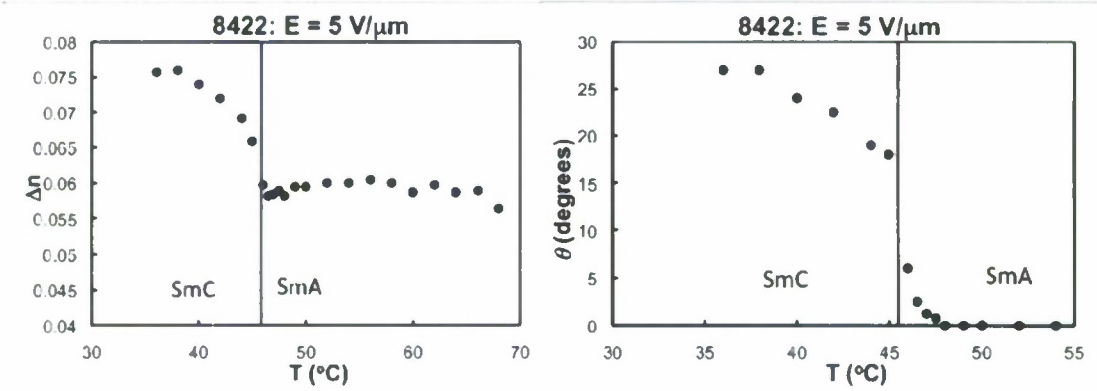


Figure 3 8422 and 5 wt% 8422[2F3] with an applied field of 5V/mm. At left, the birefringence vs. temperature shows the pretransitional drop, followed by a rapid rise in the SmC phase. Tilt angle (at right) was measured to be zero to with 3 °C of the transition, and in the SmC phase it rapidly increases, saturating at $\theta \sim 27°$.

After implementation of our new birefringence technique, we will conduct detailed measurements of 8422 and our other materials with a DC fields.

We applied an alternating field to liquid crystal cells to investigate the time-dependent nature of the molecular reorientation. In the ferroelectric SmC phase, an applied field reorients molecules and switches the polarization-induced surface charge from one side of the cell to the other. This produces a "polarization current" which can be measured by recording the current flowing through the cell as a function of the applied voltage. We built a circuit to measure polarization current and employed a function generator producing a triangle wave, then amplified the signal to produce a 90V peak-to-peak output. An oscilloscope recorded the current and applied voltage, shown in Figure 4 for 8422 doped with 5 wt% 8422[2F3]. We measured peaks in the

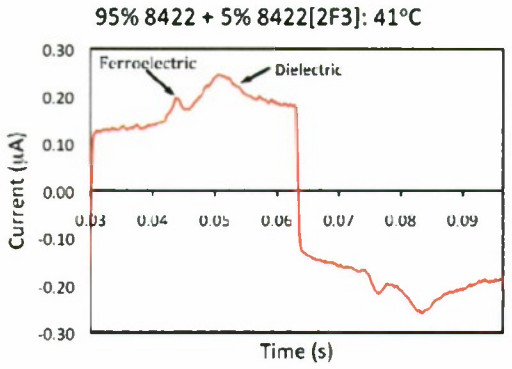


Figure 4 Current vs. time through a cell of 8422 and 5 wt% 8422[2F3] with an applied 45V triangle wave. Two peaks are observed originating from the switching of Ferroelectric polarization and induced dielectric polarization.

current due to ferroelectric switching (seen only in the ferroelectric SmC phase) and dielectric switching (seen throughout the smectic phases). The rest of the current signal is due to the flow of ions in the liquid crystal sample. We can measure the polarization density of our samples by integrating the polarization current of our AC signal. We are developing a routine to compute this polarization density, and we will compare our results in the vicinity of the SmA-SmC transition to the predictions of Dr. Saunders's new theory.

Thermodynamic Nature of A*-C* Transition (Dr Barjami)

A. Building and operation of the AC Calorimeter for the Order-Disorder Phenomena Laboratory.

Two students were hired beginning Spring quarter and Summer quarter in the construction and testing of AC Calorimeter. A block diagram of the design and electronic cquipment of the ac calorimeter that was built from scratch and will be used in this work are presented in Figures 5 and 6. The work to built and test the AC Calorimeter extended over a period of 6 months, from March 2008, till August 2008, due to the very complex nature of the calorimeter.

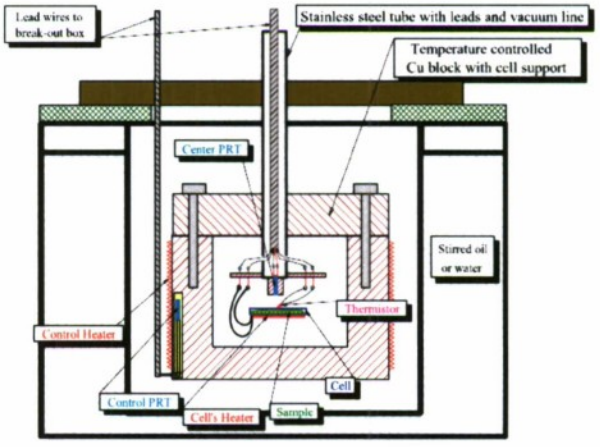


Figure 5: Schematics of the *AC* Calorimeter build in our Order-Disorder Phenomena lab. The sample lies inside a massive copper block, which is temperature controlled to better than $1mK$ in the stepwise mode.

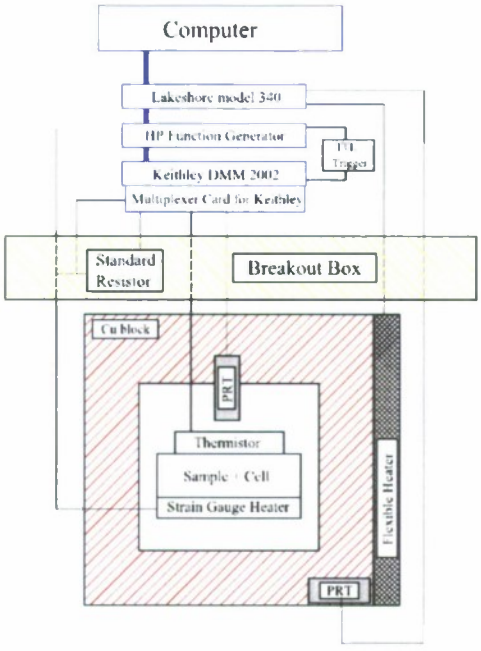


Figure 6: Block diagram of Calorimeter built in our Order-Disorder Phenomena Lab., showing all the connections to the instruments. The whole control and data acquisition is computer controlled through a GPIB interface.

Initial Results in Testing the AC Calorimeter.

The basics of the AC calorimetry technique consist of applying periodically modulated sinusoidal power to the material to be studied, and monitoring the resulting sinusoidal temperature response of the material. After sending the sinusoidal power to the material we monitored the temperature response oscillations attached to the sample. Such oscillations are presented in Figure 7.

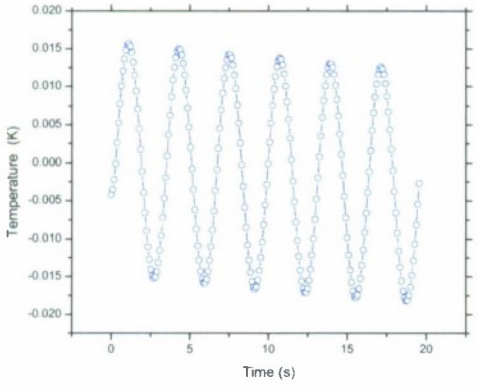


Figure 7: Sinusoidal response of AC Calorimetry. The data shown represents the raw temperature oscillations in the sample, without any averaging or smoothing routine used, which highlights the very high - resolution capabilities of our calorimeter.

Sinusoidal response of the sample shown in Figure 7 is a direct indicator of the successful operation of the new, built from scratch AC Calorimeter in our new Order-Disorder Phenomena Laboratory, which we employed in the study of de Vries liquid crystals. We use this sinusoidal response of the sample to calculate the total heat capacity of the sample and cell, given by:

$$C = \frac{P_0}{\omega T_{ac}}$$

where P_0 is the amplitude of the applied heating power, ώ is the angular frequency, $C = C_s + C_c$ is the total heat capacity of the sample and cell, and T_{ac} is the amplitude of the temperature oscillations.

B. Calorimetric measurements on different de Vries materials to correlate thermodynamic nature of the smectic A-C transition with the size of the optical response.

A new class of chiral smectics, known as de Vries smectics, shows great potential for Ferroelectric Liquid Crystals devices as they exhibit an electroclinic effect with unusually large optical responses, with little or no layer contraction. The application of a moderate electric field (5 V/micron) in the A* phases (5 V/micron) yields a rotation of the optical axis of up to 31^0 with layer contraction of less than 1%! Another attractive feature of these de Vries smectics is that these fast and colossal optical responses occur over a relatively wide temperature range. Finally, cooling into the C* phase shows none of the defects and loss of contrast associated with layer contraction.

Research to date indicates that the unique and desirable features of de Vries materials are directly related to their low orientational order and the thermodynamic nature of the A*-C* transition, which will be the goal of this project.

The material under study: TSiKN65 was synthesized by J. Naciri, Center for Bio/Molecular Science and Engineering, Naval Research Laboratory. It has a significantly large electroclinic effect and exhibits a small layer contraction in the Sm-A phase even when a large electric field is applied across the cell, as well as from Sm-*A* to Sm-*C** phase in the absence of a field. Also, the order parameter is indeed found to be very small compared to that obtained for conventional Sm-*A* or Sm-*C** even when a large electric field is applied across the cell, and is not well understood because no detailed direct information about the local molecular structure has so far been reported.

In order to determine the excess heat capacity associated with the phase transition, an appropriate background was subtracted. The total sample heat capacity over a wide temperature range had a linear background, *Cp*(background), subtracted to yield:

$$\Delta C_p = C_p - C_p(background)$$

as the excess *Cp* due to the *SmA-SmC* phase transition. The resulting ΔCp data are shown for pure TSiKN65 in Fig. 8 over a wide temperature range about, where the units are J K^{-1} per gram.

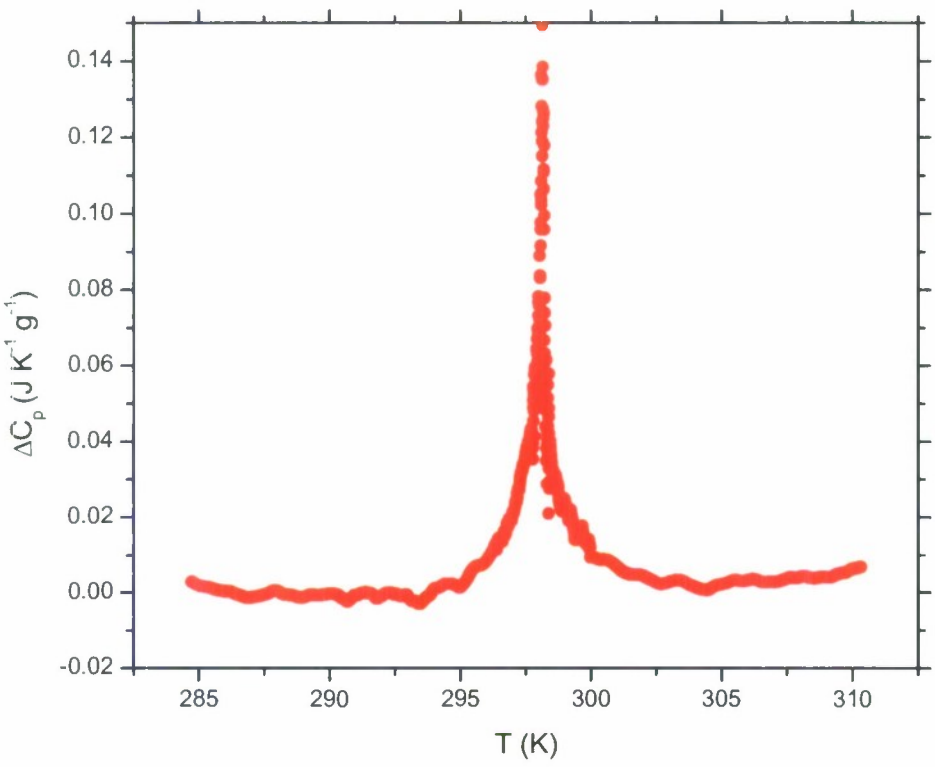


Figure 8: Exeess speeifie heat, ΔCp, as a funetion of temperature for TSiKN65.

A very sharp Smeetie A to Smeetie C phase transition peak ean be observed in Figure 8 at 298.14 K. To eharaeterize the ehange in shape of the $\Delta Cp(SmA\text{-}SmC)$ data, a traditional power-law form in terms of the redueed temperature $t = \frac{|T - T^*|}{T^*}$ will be used to analyze the experimental specifie heat data associated with the *SmA-SmC* phase transition:

$$\Delta C_p\left(SmA - SmC\right) = A^{\pm} t^{-\alpha}\left(1 + D^{\pm} t^{\Delta_1}\right) + B_c$$

where the eritieal behavior as a funetion of redueed temperature t is eharaeterized by an exponent α, amplitudes $A\pm$ above and below the transition, a critical background term B_c, and corrections-to-scaling terms eharaeterized by the eoeffieients $D\pm$ and exponent $\Delta_1 \approx 0.5$. Our high-resolution ΔCp ealorimetrie data indieates that the exeess heat eapaeity assoeiated with the Sm-A-Sm-C transition is surprisingly elose to a trieritical point. The evolution toward the tricritical point observed in our heat eapaeity investigation is eonsistent with results obtained from the theoretieal study of Dr. Saunders.

1) Dr. Karl Saunders Results:

Theoretical Model (Dr Saunders)

The theoretical modeling of the electroclinic response of chiral de Vries materials has been very successful. This was done by generalizing a zero-field nonchiral model to include the effects of chirality and the presence of an electric field. A manuscript describing the model and results has been submitted to Physical Review E. In particular, the model predicts the following, interrelated features of de Vries materials.

1) Our theoretical model indicates that materials with low orientational order and strong layering order should exhibit a transition close to a tricritical point. This is a point where the smectic-A* – smectic-C* transition goes from being first order to second order. As will be discussed below, the proximity to such a transition is a crucial factor in the unusually strong electrooptic response of de Vries materials. The combination of low orientational order and strong layering order is significant because many de Vries materials that have been studied to date do indeed exhibit such a combination. The theoretical prediction of the significance of this combination could help in the design and synthesis of de Vries type materials. It should also be pointed out that a preliminary analysis of Dr Barjami's calorimetric data indicates that the transition in TKSiKN65 is indeed close to a tricritical point. Dr Barjami's data also has some other unusual, exciting and potentially significant features that we plan to analyze.

2) According to the theoretical model, the unusually strong electrooptic response of de Vries materials can be attributed to the proximity of their smectic-A* – smectic-C* transition to a tricritical point. As discussed above, this proximity can be attributed to a combination of low orientational order and strong layering order. For de Vries materials with a second order transition close to tricriticality, the electrooptic response, i.e. the tilting of the optical axis as a function of applied field, will be continuous but unusually strong. For a first order transition, the effect is particularly striking, with the electrooptic response being discontinuous. This means that the tilting of the optical axis will exhibit a jump as applied field is increased. This unusually strong response can be particularly desirable in the manufacture of liquid crystal displays (LCDs). The electrooptic response curves predicted by the theoretical model are qualitatively consistent with published data on de Vries materials. However, we are excited to make direct quantitative comparisons with electrooptic response data that Dr Fernsler plans to obtain for the de Vries material TKSiKN65.

3) An exciting prediction of the theoretical model is that the contraction of the layers associated with the strong electrooptic response is unusually small for systems with low orientational order, of which de Vries materials are an example. The combination of strong electrooptic response with small layer contraction is highly desirable in the LCD industry. This is because it avoids buckling of the layers (observable in an optical microscope as periodic stripes) which drastically reduces the performance necessary for optical devices. As a result of this buckling, the only commercial FLC displays are used in camcorders viewfinders and not in larger displays, such as in televisions or laptops. Thus, the possibility of strong electrooptic response without buckling is an exciting feature of de Vries materials. The theoretical modeling of this project has gone some way to explaining why de Vries materials exhibit this exciting feature.

4) A final result of our theoretical model is that the increase in the birefringence associated with the strong electrooptic response is unusually large. This is also important in the context of de Vries type materials being strong candidates for use in LCDs. Again, the response curves (of birefringence to applied field) predicted by the theoretical model are qualitatively consistent with published data on de Vries materials. We also look forward to making direct quantitative comparisons with birefringence data that Dr Fernsler plans to obtain for the de Vries material TKSiKN65.

Remaining tasks

Optical Response and Orientational Order (Dr Fernsler)

Our results show clear support of Dr. Saunders's de Vries model, but our present measurement accuracy is insufficient to extract the details of the temperature-dependent birefringence. Application of our new interference filter and revised method for determining birefringence should allow us to improve out measurements and determine the nature of the SmA-SmC transition in orientational order. Using these more sensitive results, we will compare the results of our three de Vries materials with particular interest in determining whether these transitions are near the tricritical point indicated by Dr. Barjami's results and Dr. Saunders's theory. We also hope to obtain more of the material W415, which showed very different temperature-dependent birefringent behavior and a "standard" smectic material for comparative purposes.

Dr. Saunders's new theory describes electro-optic interactions in the vicinity of the SmA-SmC transition. We observed some changes in the birefringence near this transition, but with our improved birefringence measurements, we will try to observe these delicate interactions to compare with the theory.

Thermodynamic Nature of A-C* Transition (Dr Barjami)*

The assumption of SmA-SmC tricritical behavior of TKSiKN65 liquid crystal is very powerful and requires further experimental studies, which we propose to conclude during Summer 2009. Of great importance is the calculation of latent heat of the transition, an indicator of the order of the transition.

However, for first-order transitions the situation is complicated by the presence of a two-phase coexistence region, as well as a latent heat ΔH. The total enthalpy change through a first-order transition is the sum of the pretransitional enthalpy and the latent heat. In an *AC*-calorimetric measurement, ΔCp values observed in the two-phase region are artificially high and frequency dependent due to partial phase conversion during a *Tac* cycle. The pretransitional enthalpy δH is typically obtained by substituting a linearly truncated ΔCp behavior between the bounding points of the two-phase coexistence region into $\delta H = \int \Delta C_p dT$, while an independent experiment is required to determine the latent heat ΔH. A direct integration of the observed ΔCp yields an effective transition enthalpy δH^* and this contains some of the latent heat contributions; thus $\delta H < \delta H^* < \Delta H_{total} = \delta H + \Delta H$.

Theoretical Model (Dr Saunders)

The main remaining task with regards to the theoretical modeling is to cross-check our theoretical model with experimental data for de Vries material TKSiKN65. This will be done once Dr Fernsler's has completed these nontrivial experimental measurements.

As discussed above, Dr Barjami's data indicates that the smectic-A* - smectic-C* transition in de Vries material TKSiKN65 does indeed take place near a tricritical point. However, the material shows a significant pretransitional anomaly which indicates that the transition may be very different from previously observed smectic-A* - smectic-C* transitions. There may be a connection between this anomaly and the de Vries-like nature of the material. It is proposed that Dr Saunders carry out a theoretical analysis to investigate this connection. (see below under New ideas and possibilities)

New ideas and possibilities

Optical Response and Orientational Order (Dr Fernsler)

The structure of the de Vries SmA phase is still unresolved and several conflicting models have been presented to the research community. Observation of samples in the process of switching may help determine these structures. We are particularly interested to compare the anomalous sample, W415, to our other de Vries materials to determine if this material's different temperature-dependent birefringence is also reflected in a different switching behavior. We have developed a novel method to compute the orientation of the molecular director during application of AC electric fields. Our camera system allows us to record a movie of our sample in the process of switching, and can take images at faster than standard video rates for increased resolution. We have developed a simple routine using Matlab to process image data from these movies to determine the average brightness of the cell during switching. The intensity of the signal is proportional to $\sin^2\theta$, where θ is the molecular tilt angle, which allows us to extract the orientation. Preliminary data of 8422 with 5 wt% 8422[2F3] switching under an alternating field using this photometric approach is shown in Figure 9. Furthermore, statistics on the orientation can be obtained by analyzing the image data. As Dr. Saunders suggests below, critical fluctuations may be important in the nature of the SmA-SmC transition of de Vries materials. Our statistical analysis of images during switching may shed light on this new area of inquiry.

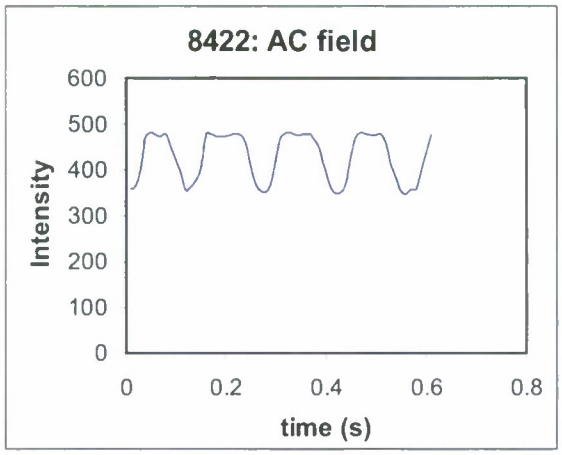


Figure 9 8422 with 5 wt% 8422[2F3] image intensity versus time during switching with an applied sine wave with 50V peak to peak.

We believe that this new technique, combined with polarization current measurements may help us resolve the structure of the de Vries SmA phase. It is also possible that more than one de Vries phase exists: e.g. the markedly different temperature-dependent birefringence behavior observed in W415 may indicate that it has significantly different structure than other de Vries materials.

Thermodynamic Nature of A*-C* Transition (Dr Barjami)

During Summer 2009 we propose to obtain and continue the study of deVries liquid crystals by investigating the heat capacity behavior of other deVries liquid crystals: W415 and 8422 and comparing the results with heat capacity study of TKSiKN65 and with the results obtained for the Optical Response and Orientational Order results obtained by Dr. Fernsler on these materials.

Theoretical Model (Dr Saunders)

The pretransitional anomaly near the smectic-A* - smectic-C* transition, observed by Dr Barjami, indicates that the transition may be very different from previously observed smectic-A* - smectic-C* transitions. Such an anomaly often means that there are significant critical fluctuations near the phase transition. We propose to investigate the size, behavior and effects of critical fluctuations near the smectic A*-C* transition. Such an investigation has never been carried out because these critical fluctuations near the A*-C* transition have always been assumed to be negligible. We believe that the pretransitional anomaly observed by Dr Barjami may be evidence that these critical fluctuations can in fact be quite significant and could play an important role in underlying mechanism responsible for the de Vries structure.

Relevance/applicability of this work to DOD interests

Hand held electronic devices with liquid crystal displays (LCD's) are ubiquitous in our society and are of particular interest to the DOD for work in the field. LCD's are optimal for these applications because of their slim profile and low power requirements. Nevertheless, significant improvements in these systems can be made to optimize these LCD's to reduce power requirements (which can lead to longer battery life) and improve visual quality.

Present LCD's almost all use the twisted nematic geometry where alignment layers between the two plates holding the liquid crystal material are oriented 90 degrees to one another. This produces a helical twist in the molecular director, which rotates the polarization of light to allow linearly polarized light to pass through crossed polarizers surrounding the cell. To achieve a dark state, these twisted nematic LCD's apply an electric field, which polarizes the molecules, which then rotate similar to dipoles in an electric field. This method has several disadvantages in image quality. Because the nematic phase itself is not polar, a relatively large electric field must be applied to polarize and reorient molecules and this reorientation is relatively slow. Although present nematic LCD's have improved this response time, motion blurring is unavoidable on these type of displays because bright pixels for one frame of a moving image stay on for the entire duration of that frame and then switch at the beginning of the next frame: our eyes blur these frames together causing a perceived distortion in moving images. Furthermore, the helical structure of twisted nematic LCD's causes distortions in the color when viewed off-axis, thereby reducing the viewing angle of these displays. While some of this problem has been corrected by use of additional refractive films coated on the display, this adds cost, bulk and reduces some light coming from the screen. Finally, because the nematic phase inherently produces some flicker due to thermal fluctuations (observed as a shimmering effect in a polarizing microscope) and the electric field producing the dark state of a pixel cannot fully reorient molecules, the contrast ration of present LCD's is limited and cannot approach that of a traditional cathode ray tube (CRT) display.

Smectic liquid crystals could improve many of these drawbacks to twisted nematic displays. The chiral SmC phase is a polar ferroelectric phase, so reorientation of molecules occurs at lower applied electric fields and it is significantly faster (by about an order of magnitude). This can allow for an LCD with lower power requirements (hence longer battery life) and can eliminate motion blurring by using a duty-cycle approach to producing a grayscale in each pixel (the amount of time a pixel is bright or dark during each frame is adjusted to produce the impression of dark or light pixel; this approach is used by digital light processing, DLP, projectors). Smectic LCD's have a more uniform structure thoughout the cell, instead of the helical structure used in twisted nematics, which would improve the viewing angle without the use of additional films. Finally the contrast ratio of these displays would be improved over nematic displays because the smectic phase does not suffer from flickering effects and has a more uniform structure when switched with electric fields.

Until now, smectic liquid crystals have not been used extensively in displays, except in viewfinders for camcorders. This is because of the present cost of manufacturing these materials and defects introduced when cooling into the ferroelectric SmC phase due to a contraction of the smectic layers. De Vries-type materials can eliminate these defects because of their lack of layer contraction. Up until now, only a relatively small number of de Vries-type materials have been synthesized and their cost is prohibitively high for use in displays. However, our work to understand the de Vries SmA and SmC phases could lead to predictions to improve the design of future materials, which could optimize these materials for LCD's and reduce their cost by simplifying synthesis and producing materials in bulk quantities.

Feature Selection and Boosted Classification Algorithms for Pedestrian Detection

Projeet Investigator:

Samuel J. Frame
Department of Statisties
California Polyteehnie State University
San Luis Obispo, CA

Feature Selection and Boosted Classification Algorithms for Pedestrian Detection

Samuel J. Frame, Ph.D.
Department of Statistics

Abstract

Detecting pedestrians in video imagery is an emerging necessity for civil and federal law enforcement agencies, and agencies within the Department of Defense. Support vector machines (SVM), neural networks, mixture models, and boosting algorithms are competing methods which have all been suggested. For this work, we focus on *boosting* weak classifiers to detect pedestrians and select features. Boosting algorithms are attractive because they rely on simple, weak classifiers and they are computationally inexpensive. Furthermore, they serve as an internal method of selecting features. We have already implemented a fast, accurate, and numerically stable boosting algorithm for detecting pedestrians and selecting features. We will extend our current research and implement a mixture model weak classifiers. Also, we will exploit the univariate feature selection already present to allow for larger, multivariate feature sets. Furthermore, continued funding will allow us to better continue to test and evaluate our newly constructed system. This project will provide continued seed money for the principal investigator's professional activities.

Keywords

Pedestrian detection, boosting, classification, feature selection, statistical learning

1 Identification of the Problem

Modern surveillance scenarios are saturated with digital video recorders, autonomous smart weapons, and sensing platforms that contain a wide variety of sophisticated sensors. As such, the ability to automatically, accurately, and quickly detect objects of interest is a capability critical to the mission of civilian law enforcement agencies, Homeland Security, and the Department of Defense. Rapid, real-time object detection is a necessary component of the Navy's mission to conduct urban warfare, Automatic Target Recognition (ATR) and tracking of vehicles of interest, and ``end game" delivery of autonomous weapon systems. The research we propose will further develop the theoretical and computational methods needed to detect objects of interest captured in real-time digital video imagery as well as a wide variety of hyperspectral imagery.

Current \emph{boosting} algorithms have the attractive quality of providing rapid and accurate object detection, and they rely on a simple threshold weak classifier [1][9][14][15]. In our current work, we have implemented a quadratic discriminant weak classifier (as we proposed in our previous work). Comparable to current boosting methods, we use only a single feature in each boosting stage. Mixture models have already been shown to improve classification in comparison to a single quadratic discriminant, and they provide excellent classification results [2][3][4][5][6]. Secondly, our current implementation relies on a single feature in each boosting stage. Classification results can be improved by internally expanding the feature sets used.

The specific object detection problem we consider is pedestrian detection in video imagery. For evaluation purposes, we will continue to use the benchmark data set *DaimlerChrysler Pedestrian Classification Benchmark Dataset* provided by DaimlerChrysler, which is used in some studies [8][9][14][15]. The data set contains thousands of images which either have or do not have a pedestrian present in each image. While most implementations rely on a single set of testing and training data, our implementation is designed as a simulation study to assess performance variability.

2 Summary of Work

Our current research efforts were initiated in late January, 2008. The funding we received was less than the amount we requested. As a result, we have concentrated our efforts on the computational implementation which we proposed. In fact, we have entirely completed the proposed work. We are pleased to report that we have successfully implemented a stable, accurate, and flexible solution in C++. Because the solution is flexible, we are able to conduct a wide range of simulation experiments using *any* set of data.

A large component of the current research is the C++ implementation of our detection algorithm. We have spent a considerable amount of effort developing our implementation so that it is accurate, reliable, and flexible allowing for different experiments and simulation studies. Using a control file, we are able to easily change experiments, data sets, and algorithm controls. We have used an *XML* schema which allows for the following inputs.

- Training Data: specifies the complete list of training
- Testing Data: specifies the complete list of testing data and associated class names. To be clear, the class names are not used for learning and classification purposes. Rather, we use the class names for calculation of performance metrics.
- Classifier method: specifies the classifier method to be used either *Threshold* or *Qda* for the linear and quadratic discriminants respectively.
- Cascade times: specifies the number of boosting stages to be used.
- Simulation count: specifies the number of simulations to be done. If simulation count is set at 1, then the specified set of training and testing data is used. If the simulation count is larger than 1, the data is combined and then randomly "broke" into training and testing data in each simulation.

- Train percent: specifies the amount of data that should be used for training purposes. This is only used when we run multiple simulation experiments.

The implementation also uses the Matrix package, and it outputs all of the summary metrics to portable Excel file. The implementation provides the backbone for future work in the area of Statistical Learning Algorithms. For our proposed research, we will continue to design, implement, and test more advanced classification and feature selection methods.

3 Results

For initial testing and debugging purposes, we simulated a set of data from Normal distributions which should provide good classification performance. For two different groups, we simulate 100 training and testing samples. The mean vectors are given below. For both groups, we use a identity matrix as the covariance structure.

Group	Mean Vector
Pedestrian	(-3,-2,-1,0,1)
Non-Pedestrian	(-3,-2,0,2,4)

We use this choice of mean vectors because of the overlap. It should be clear that the valuable discrimination information is contained in the last 3 elements of each feature vector. Since we are only using 1 feature in each boosting stage, the algorithm should use and select the last 3 features for classification purposes. An analysis of the univariate features used will be contained in future a publication and report. In this simulation study, we use 25 boosting stages, with 50% of the data used for training, and 50 different simulations to assess the performance variability. Primarily, our focus is to compare the performance of the linear and quadratic discriminants we have chosen.

In figure (1) we report the average training and testing error rates, and the performance variability (specifically, the standard deviation of the classification error rate) respectively. For the linear discriminant weak classifier, the training error decreases slightly but there is little improvement with successive boosting stages in the testing error.

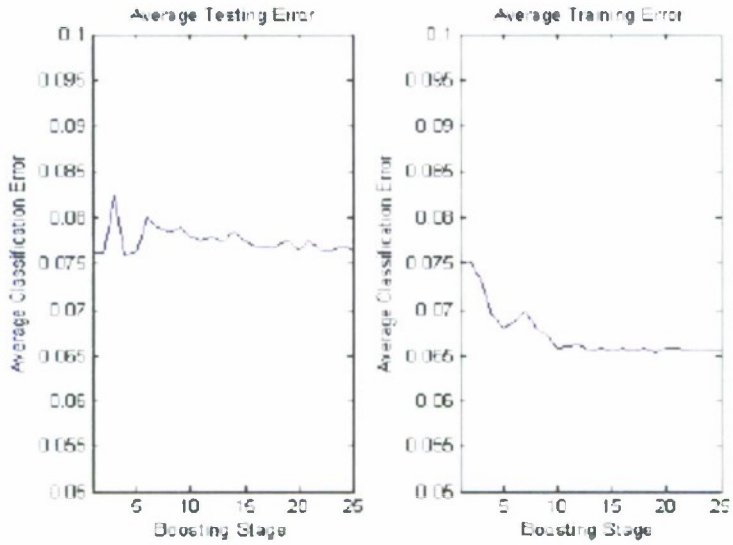


Figure 1: Threshold Classifier: Average Error Rate

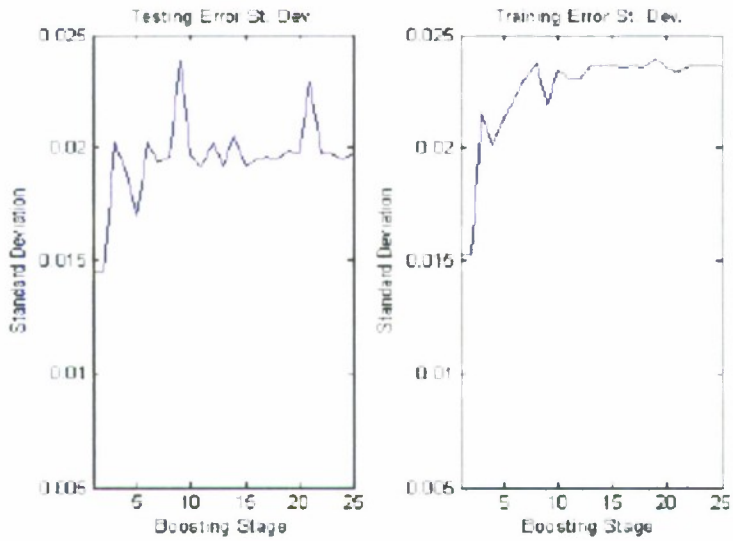


Figure 2: Threshold Classifier: Standard Deviation

Figure (3) summarizes the simulation results using the quadratic discriminant weak classifier. For the quadratic discriminant, the training and testing error decreases with more boosting stages (albeit the decrease in error is not monotonic). In comparing the average simulation classification error for the quadratic and linear discriminants, it is clear that the quadratic discriminant has lower classification error than the linear discriminant. The improvement is more noticeable when comparing the training error rates. However, our primary concern and measure of comparison is the testing error rates. For the linear discriminant, there is zero improvement. On the other hand, the quadratic discriminant has a noticeable decrease in the testing error rate.

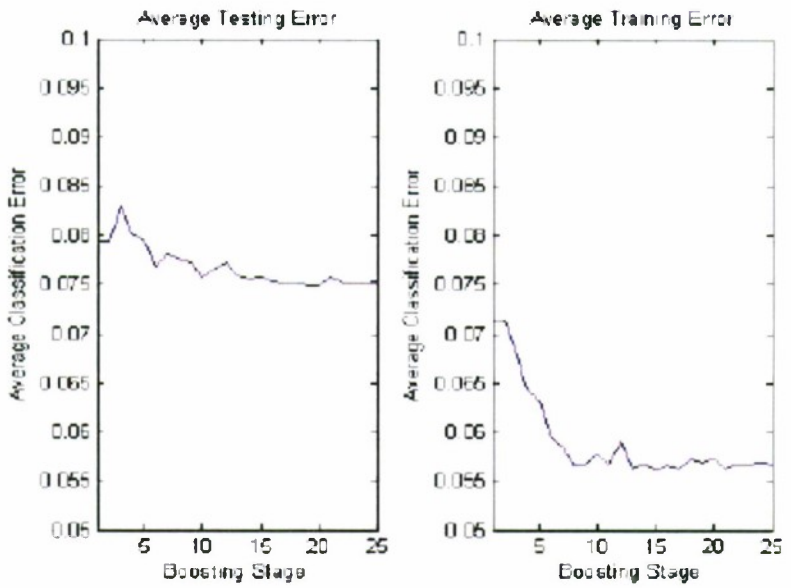


Figure 3: Qda Classifier: Average Error Rate

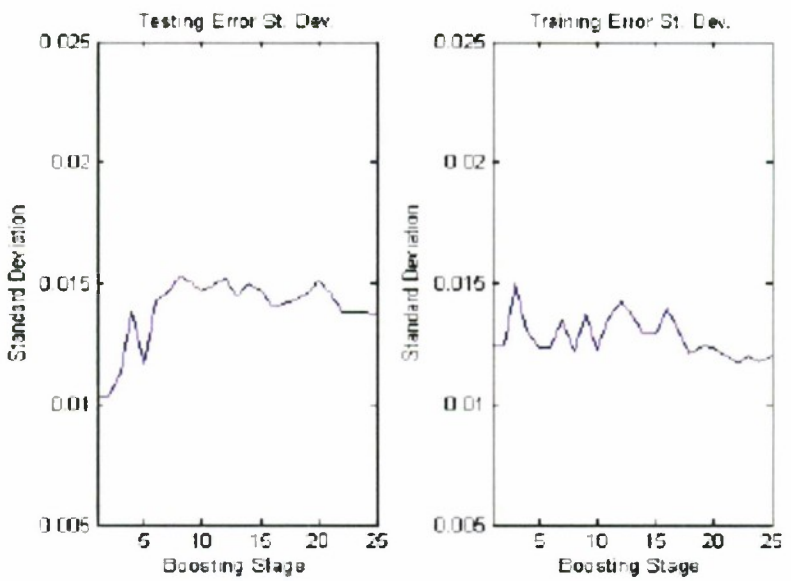


Figure 4: Qda Classifier: Standard Deviation

As we explain in the proposal for the current funding, boosting algorithms should always improve the classification performance. However, this is contingent on using the *same features in each boosting stage.* Since we are selecting the single, best feature in each boosting stage, the feature set is not constant in each

stage. As a result, we are not guaranteed a monotonic increase in classification performance. For the proposed research, we will remedy this by using comparable, growing sets of features in each feature stage.

A *novel* and interesting component of our analysis, here, is the standard deviation of the classification error in Figures (2) and (4). Because few studies fail to conduct simulation experiments (ie they only consider a single set of testing and training data), they are unable to provide this perspective of algorithm performance. In fact, we have not seen such results reported in the publicly available literature. We believe this simple and useful analysis to be novel, and it requires more attention. A priori, we anticipated that the variation would decrease with more boosting stages. However, this is not the situation. In most experiments with the simulated data, the variation in classification error actually increases throughout the boosting stages. It is noteworthy that the variation is much less for the quadratic discriminant. We are able to conclude that the quadratic discriminant provides better classification results with less variability in performance.

4 References

[1] Alldrin, N. (2005). "Detecting Pedestrians," Technical Report, Department of Computer Science, University of California, San Diego.

[2] Bishop, C. (2006). \textit{Pattern Recognition and Machine Learning}. New York: Springer.

[3] Frame, S. and Jammalamadaka, S. R. (2007). ``Generalized Mixture Models, Semi-supervised Learning, and Unknown Class Inference," \textit{Advances in Data Analysis and Classification (ADAC): Theory, Methods, and Applications in Data Science}. New York: Springer.

[4] Frame, S. and Miller D. (2005). ``Machine Learning for Robust Automatic Target Recognition: Phase 1 Final Report," Phase 1 Final Report for U.S. Air Force Research Laboratory Contract FA8650-04-M-1659.

[5] Hastie, T. and Tibshirani, R. and Friedman, J. (2001). \textit{Elements of Statistical Learning}. New York: Springer.

[6] McLachlan, G. and Krishnan, T. (2004). \textit{The EM Algorithm and Extensions}. New York: John Wiley and Sons.

[7] Miller, D. and Browning, J. (2003). ``A Mixture Model and EM-based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets", \textit{IEEE Trans. on Pattern Anal. and Machine Intell}, 1468-1483.

[8] Munder, S. and Gavrila D. M. (2006). DaimlerChrysler Pedestrian Classification Benchmark Dataset, (C) DaimlerChrysler AG.

[9] Munder, S. and Gavrila, D. M. (2006). ``An Experimental Study on Pedestrian Classification," \textit{IEEE Transactions on Pattern Analysis and Machine Intelligence}, 1-6.

[10] Rencher, A. (1995). \textit{Methods of Multivariate Analylsis}. New York: John Wiley and Sons.

[11] Schapire, R. and Singer Y. (1999). ``Improved Boosting Algorithms Using Confidence-rated Predictions," \textit{Machine Learning}, 297-336.

[12] Schapire, R. and Singer Y. (1999). ``A Short Introduction to Boosting," \tcxtit{Journal of Japanese Society for Artificial Intelligence}, 771-780.

[13] Trosset, M. (2007). ``Scmisupcrvised Learning from Dissimilarity Data," Tech. Report No. 07-01 Department of Statistics, Indiana Univeristy, Bloomington.

[14] Viola, J., Jones, M. and Snow, D. (2005). ``Detecting Pedestrials Using Patterns of Motion and Appearance," \textit{International Journal of Computer Vision}, 153-161.

[15] Viola, J., and Jones, M. (2001). ``Robust Real-time Oject Detection," Second International Workshop on Statistical and Computation Theories of Vision-Modeling, Learning, Computing, and Sampling.

Design and construction of magneto-optical trap for experimental investigation of atomic dipole traps for quantum computing

Project Investigator:

Katharina Gillen
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

Design and construction of a magneto-optical trap for experimental investigation of atomic dipole traps for quantum computing

Katharina Gillen, Physics Department

1. Background Information:

The goal of this project is to design and construct a magneto-optical trap in order to cool a sample of ^{87}Rb atoms. These cold atoms will then be used to explore whether the atomic dipole traps formed in the diffraction pattern behind a circular aperture are suitable for quantum computing. Quantum computation and information are currently widely investigated fields because of their applications in communication and decryption, as well as investigation of fundamental quantum physics. One of the systems identified as possible quantum computer are neutral atoms trapped by laser light. The remaining difficulties of this system are (1) arranging over one million atoms in a way that still allows addressing individual atoms for quantum computation, and (2) performing two-qubit gates, which require bringing pairs of atoms together and apart controllably.

In my previous research I identified the diffraction pattern behind a simple circular aperture as a possible atomic light (or "dipole") trap. Calculations indicated that with modest laser power (~100 mW) atom traps appropriate for quantum computation can be created. Theoretically, an array of circular apertures would then also yield an array of such atom traps.

With the help of previous funding through the C3RP program, we were able to directly measure the intensity pattern behind a commercial pinhole to test if the diffraction pattern behind a real pinhole still retains all the features necessary for quantum computing. We found that indeed the intensity pattern is as we calculated. The next step in exploring these traps is to trap cold atoms in them and measure the trap properties by observing the properties of the trapped atom sample. This requires pre-cooling an atom sample using a magneto-optical trap (MOT). Design and construction of a MOT is the goal of this project.

2. Project Plan:

The granted C3RP funding was used to design and construct all parts of a MOT, which is needed to pre-cool a sample of ^{87}Rb atoms before loading them into the new atomic dipole traps described above.

The MOT system consists of a vacuum chamber, a pair of electromagnets, two tunable diode lasers, and an optical system that creates six laser beams of appropriate polarization in order to slow down and collect atoms (see Fig. 1).

In order to trap atoms without any background gases expelling them from the trap, an ultra-high vacuum environment is needed. This is achieved by building a vacuum chamber from all stainless steel and glass parts, thoroughly cleaning all parts in a multiple-step process, and finally "baking" the chamber at a temperature of 300°C while pumping it out with a turbomolecular pump backed by a roughing pump.

Second, a pair of electromagnet coils is needed to create the quadrupole magnetic field that is necessary to collect the atoms into a small cloud at the center between the magnets.

Finally, for the MOT to work, two tunable, frequency-stabilized diode lasers are needed. The lasers are split into six beams of equal intensity, and will be directed into the vacuum chamber such that two opposing beams travel along each of the three dimensions of space. In order for the laser beams to slow the atoms and collect them in the center of the vacuum chamber, the laser beams have to be circularly polarized and combined with the quadrupole magnetic field with zero field at the center.

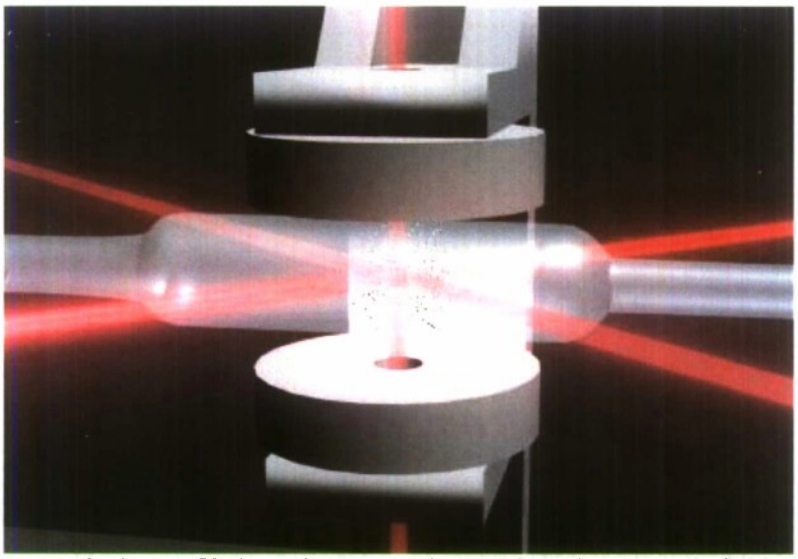


Figure 1: Magneto-optical trap. Six laser beams and a quadrupole magnetic field cool down atoms inside a vacuum chamber. (Picture courtesy of the Ohio Supercomputing Center. The picture is a screenshot of an animation of a MOT previously built by me.)

Once everything is constructed and aligned, we can trap atoms in the MOT, and characterize the properties of the MOT. The number of trapped atoms can be determined by measuring the amount of light emitted by the cold atom cloud with a photodiode. The temperature of the atom cloud can be measured by measuring the size of the atom cloud with a CCD camera. We can use this information to optimize the MOT for atom number and temperature, by fine-adjusting the laser alignment, laser frequency and magnetic field strength.

Future work after building the MOT will be to load the cold atom sample into the dipole traps. This will require a carefully timed sequence of adjusting the laser frequencies and intensities to optimize atom transfer to the dipole traps. After that we can work towards demonstrating quantum computing operations on the trapped atoms, such as initialization, read-out and 1-qubit gates.

Funding for this project was used for equipment and supplies, and student and faculty salaries (added compensation and summer salary).

3. Student involvement:

All aspects of this research project involved undergraduate students. During Spring and Summer quarters 2008 a total of five students worked on this project: Bert Copsey (Mechanical Engineering), Angelica Davidson (Physics, Spring only), Alison Goodsell (Physics, Spring only), Troy Kuersten (Physics and Aerospace Engineering), and Eric Muckley (Physics). Details of the students' tasks are given in the next section.

4. Tasks completed:

The tasks we achieved are:

- Acquire equipment (vacuum parts, ultrasonic cleaner, laser power supplies, Gaussmeters, copper wire for electromagnets, optics)
- Improve tunable diode lasers
- Set up frequency stabilization systems for diode lasers
- Tune diode lasers to atomic transition frequencies
- Develop and perform cleaning procedure for ultra-high vacuum parts
- Design and construct vacuum chamber
- Design and construct electromagnets
- Lay out optical system
- Calculate trapping potentials for pairs of laser beams at different angles incident on a pinhole

Angie Davidson, assisted by Eric Muckley and Alison Goodsell, constructed two tunable diode lasers for use in the MOT. She also ordered/designed/built all parts for an additional two lasers and constructed an injection locking system to be used for the experimental realization of the new dipole traps. She further improved the laser design by installing Brewster windows on all four lasers.

Alison Goodsell built the ultra-high vacuum (UHV) chamber needed for the MOT. This involved 4+ hour baking cycles of individual parts of the chamber to 400 °C, and a thorough 4-step, 2-hour cleaning cycle, and, of course, the assembly of the chamber.

Eric Muckley designed and built a pair of electromagnets for use in the MOT. Because of the strict requirements for the geometry of the magnets, we had to custom-build our own magnets. Eric designed the magnet mounts and supports and developed a method for winding two identical magnets without "impurities." He also tuned the tunable diode lasers to the atomic transition frequencies of the ^{87}Rb atom in preparation for trapping atoms.

Troy Kuersten designed the optical layout for the atom trapping experiment, which involves four lasers, the MOT vacuum chamber and magnets and all laser diagnosis and stabilization setups (three vapor locking setups, one Fabry-Perot, and one saturated absorption setup). Troy also built and tested 10 photodetectors for use in the various parts of the experiment.

Bert David Copsey designed custom mounts for the vacuum chamber – both for during the bake-out process, and for day-to-day operation. He also put together and repaired the physics department vacuum pumping station which is essential in the pumping down and bake-out process of the UHV vacuum chamber. At the end of the grant period, he had gotten all parts to work, except for the ionization gauge. Once this is replaced, we can proceed with the bake-out of the UHV chamber, trap atoms at ~200 μK from absolute zero in our MOT, measure and optimize the MOT properties, and then use the cold atoms for our dipole trap experiments.

In addition, during various delays in shipment of parts for the pumping station, Bert continued his work on calculating trapping potentials formed by laser beams incident on a pinhole (using a Mathematica code that he developed). In particular, he calculated the potential that atoms in specific Zeeman substates experience due to the intensity pattern formed by two circularly polarized laser beams incident on a pinhole at

an angle. This technique may provide a method of bringing pairs of atoms together and apart for two-qubit gates. He presented preliminary results of this work at a national conference (see below).

5. Results:

We successfully constructed all necessary parts for a MOT that will be capable of cooling ^{87}Rb atoms to ~200 μK from absolute zero: (1) Two tunable diode lasers, (2) two electromagnets for the magnetic quadrupole field, and (3) an ultra-high vacuum chamber, as shown in Fig. 2.

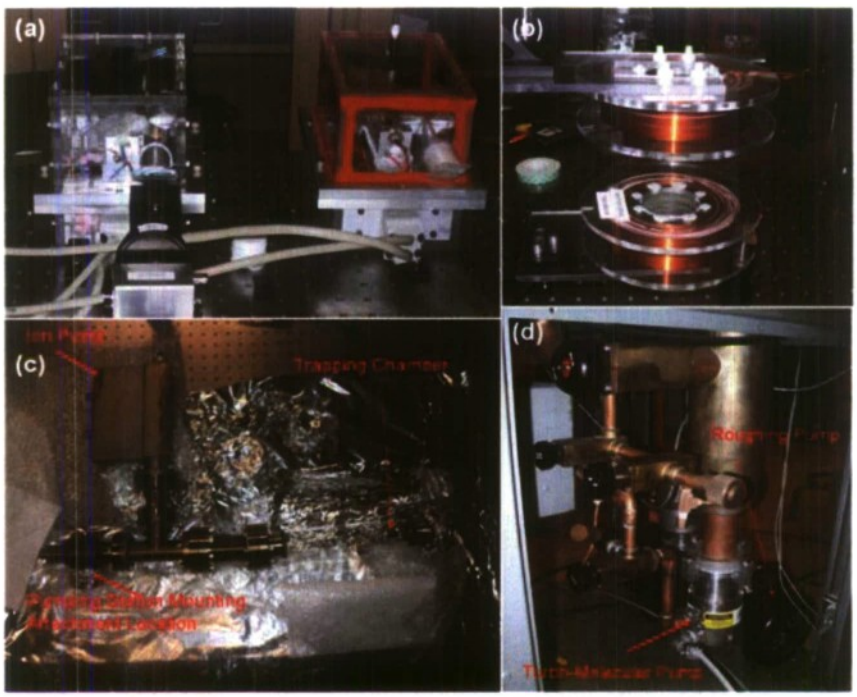


Figure 2: MOT setup. (a) Two lasers. (b) MOT magnets. (c) Vacuum chamber. (d) Pumping station (turbomolecular pump backed by roughing pump).

The only remaining holdup is the ionization gauge on the Physics Department vacuum pumping station, which needs to be replaced. Once it is replaced, everything is ready for baking out the vacuum chamber, and turning on the MOT. We hope to accomplish this by June 2009.

Our computational research results show that it is possible to use two circularly polarized laser beams incident on a pinhole at an angle to trap two atoms in different Zeeman substates simultaneously. Even though the incident angle elongates the traps somewhat, the trap properties remain comparable to those of a laser beam at normal incidence (previously described in Gillen et al., Physical Review A 73, 2006, 013409). By changing the angle of the incident laser beams, the two atoms can be brought together and apart controllably, without expelling the atoms from the traps (see Fig. 3). This can be used to facilitate two-qubit gates, which require bringing two qubits (here: atoms) close together. This would constitute tremendous progress towards a neutral atom quantum computer. We plan to investigate this computational result experimentally once the MOT is completed.

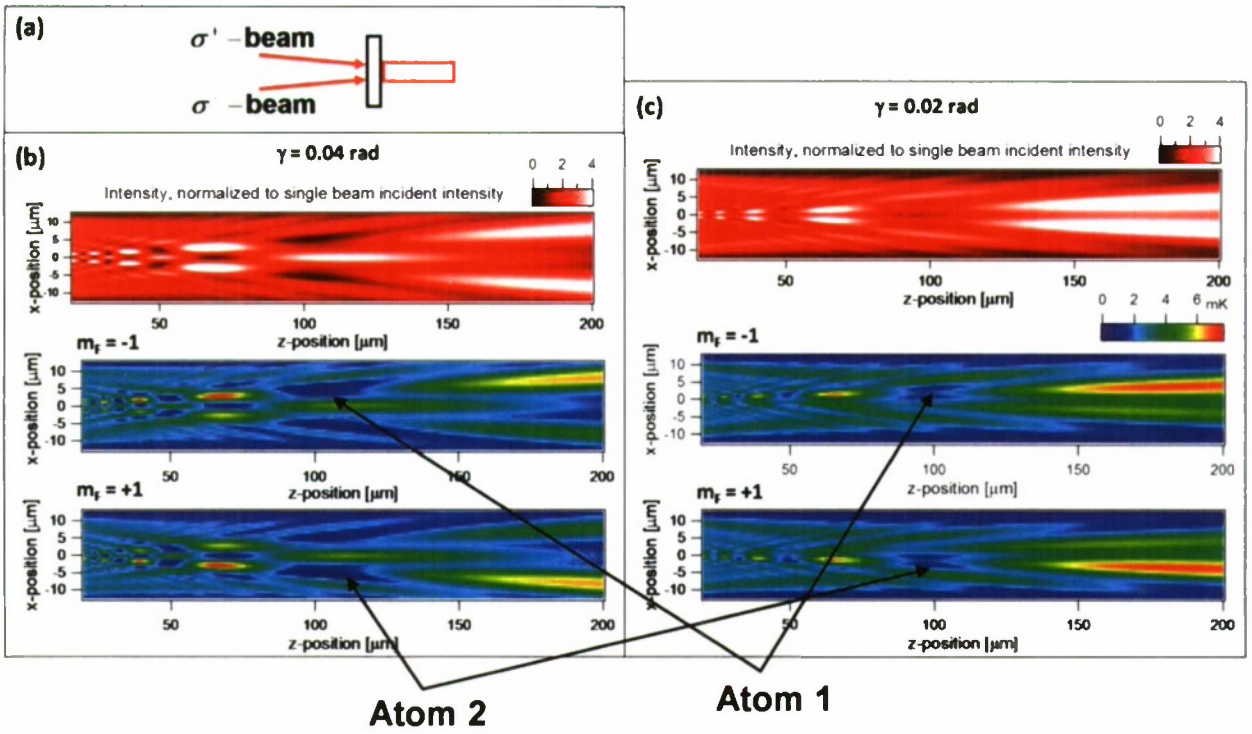


Figure 3: (a) Two circularly polarized laser beams are incident on a circular aperture at an angle. (b) Intensity pattern behind the aperture that traps two atoms in two different locations (dark spots) for an incident angle of 0.04 rad. (c) Intensity pattern behind the aperture for an incident angle of 0.02 rad. By reducing the tilt angle of the laser beams the two atoms can be brought together for 2-qubit operations.

6. Presentation of results:

The computational results obtained as part of this project were presented by Bert Copsey (undergraduate student) at the American Physical Society's Division of Atomic, Molecular, and Optical Physics (DAMOP), May 27-31, 2008, at Penn State – a national conference

We also intend to prepare these and future results for publication in Physical Review A.

7. Conclusions:

We successfully designed and constructed all parts of a magneto-optical trap (MOT). The MOT will become fully operational once the Physics Department pumping station is fully fixed and the vacuum chamber bake-out is performed. Completion is expected by June 2009.

In addition to the experimental work, we were able to perform computational research on atomic dipole traps formed behind a pinhole. Preliminary results indicate that two circularly polarized laser beams incident on a pinhole at an angle can be used to bring two atoms in different magnetic substates together and apart without expelling ether atom from the trap. If these results are confirmed by experiments, we have found a way to facilitate two-qubit gates. This would be a big step towards building a neutral atom quantum computer.

Upgrading a fleet of GPS-tracked ocean surface drifters for improved performance and extended coverage area

Project Investigator:

Elizabeth Griffith
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

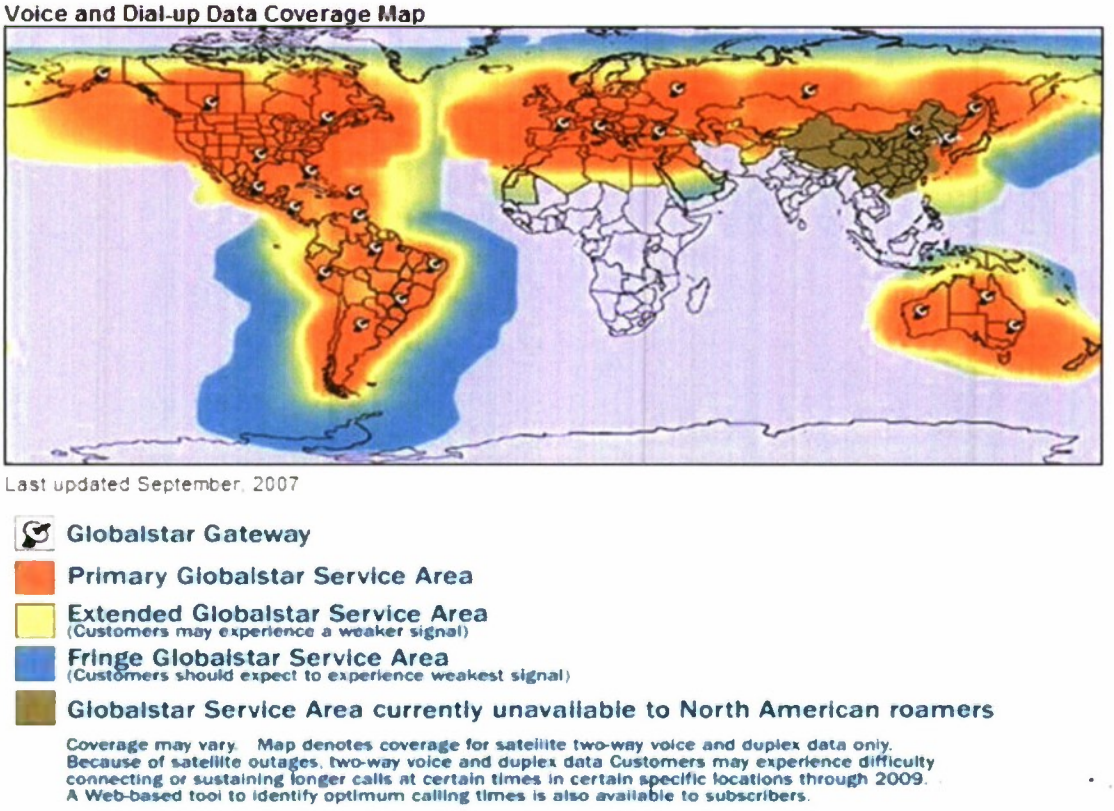
Final Project Report for 2007-2008 C3RP Grant

Submitted by: A. Elizabeth Griffith, Physics Department, 805-756-2473, aegriffi@calpoly.edu

This report describes our progress in developing a fleet of GPS-tracked ocean drifters to map and quantify surface currents. Using the previous grant (2006-2007), we had developed a fleet of eight ocean drifters to map ocean surface currents within 10 miles of Avila Bay, and we had completed one test run of the drifters.

This year, our goals were to (A) extend the range of the drifters beyond 10 miles from Avila Bay (B) increase the battery life of the fleet for extended use (C) improve the current-following characteristics of the drifters (D) make the investigations safer for Cal Poly participants by reducing the number of required boat trips in our flat-bottom boat (E) enlarge the fleet from 8 to 16 units to enable more extensive current mapping capability, and (F) perform trials to test the new system and determine the feasibility for multiple deployments. We successfully completed all but the last goal. We have started trials of the new drifter fleet, and will continue the trials during the summer of 2009.

This year we did increase the tracking range of our system from 10 miles from Avila Bay to extend up and down the coast of north, south and central america (see Plate 1). This was accomplished by upgrading the drifter electronics to SmartOnes, sold by Fleet Analytics.



Plates 2 and 3 show oeean surfaee eurrents, measured by CODAR (Moline, et al). Notice that on the days shown (June 21 and 26, 2009) the current patterns are signifieantly different from one another. Speeifieally, on June 21 the surface currents just south of the Cal Poly pier (indieated by a red dot) are headed south, away from the pier. But on June 26, the surfaee eurrent at the same loeation is headed north, into shore! Indeed, the nature of the eireulation pattern varies depending on the day, time, weather, surrounding oeean eonditions, ete.

Plate 2. Surfaee Currents from CODAR measurements. (This plot and the next were created by physics major Chris Ferguson, based on data pulled by Brian Zelenke.)

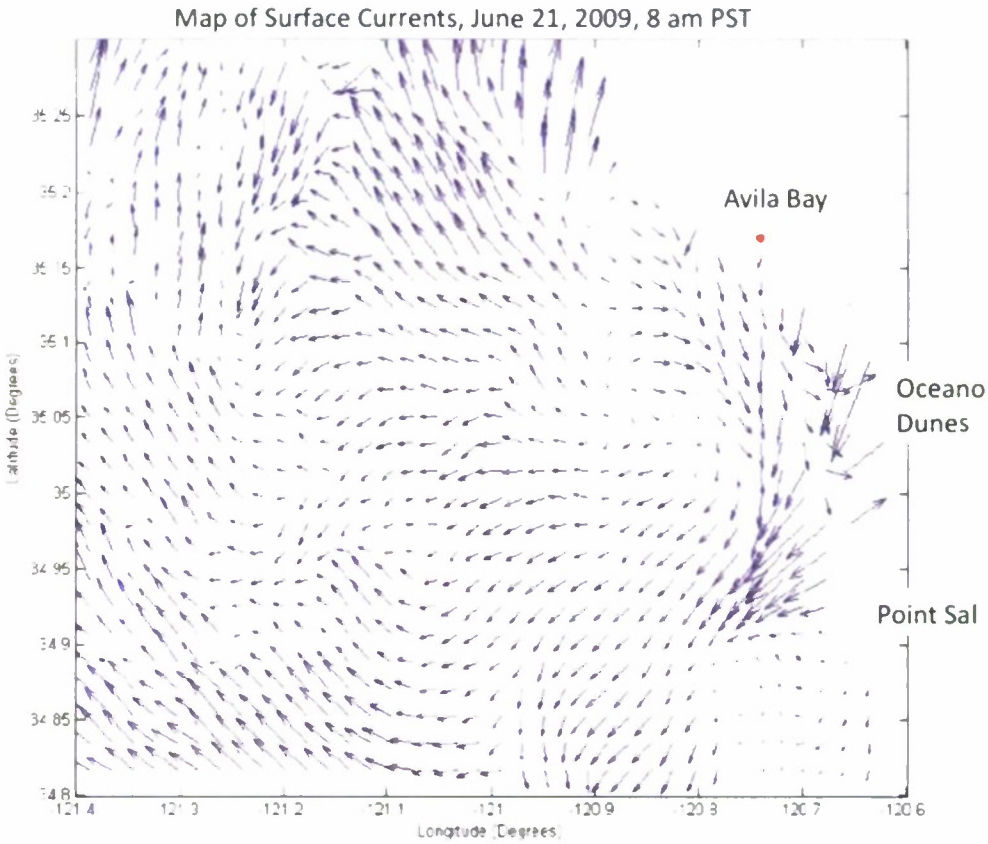
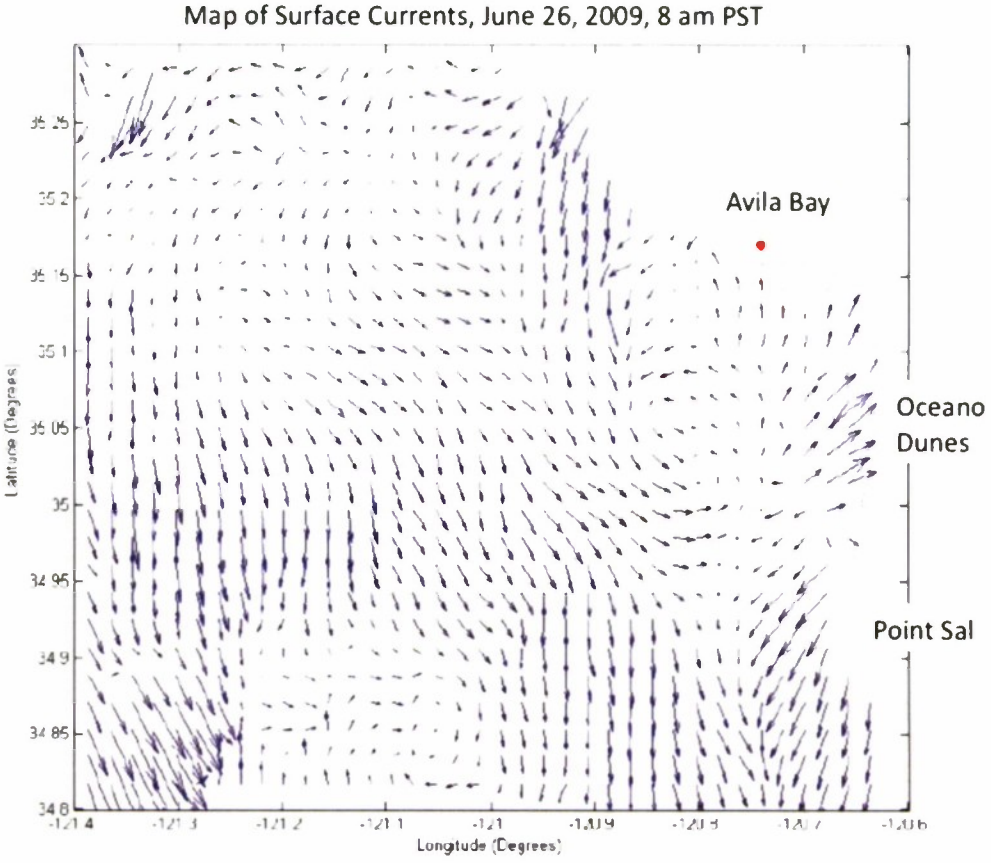


Plate 3. Surface Currents from CODAR measurements.



The surface velocity maps shown in Plates 2 and 3 are useful to understand the general character of the surface currents along the central coast of California. But this type of map does not answer these questions:

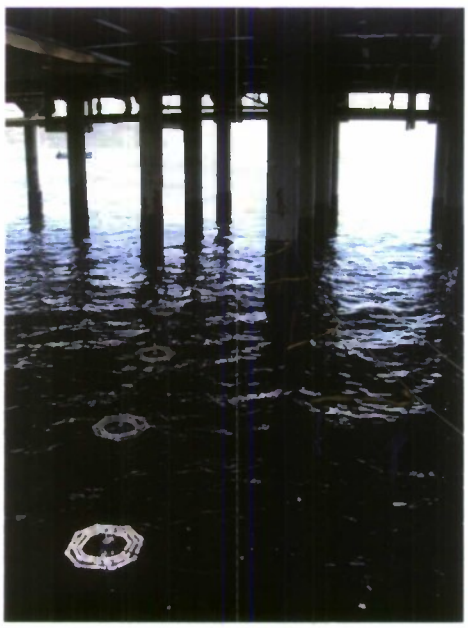
- How long does it take for water to travel from Avila Bay to Point Sal?
- At what rate would a contaminant introduced in Avila Bay disperse into the surrounding water?
- What is the residence time of ocean water in Avila Bay?

These are difficult questions, but they can be readily answered through the use of ocean drifters with GPS tracking, and subsequent analysis techniques.

Our plan is to first use the drifters to quantify local coastal recirculating currents that are difficult to predict. After about six local runs of the fleet this summer and early fall (to quantify recirculating currents), we plan to apply for local funding to pay for charter boat use to deploy our drifters further from shore to explore the California Current as well as typcial currents from nearby marine sanctuaries.

Plate 4. Undergraduate physics majors (Galen Cauble and Tom Baker) assemble the drifters on the pier deek, conduet stability tests and set up overnight leak tests to cheek the drifter performanee.

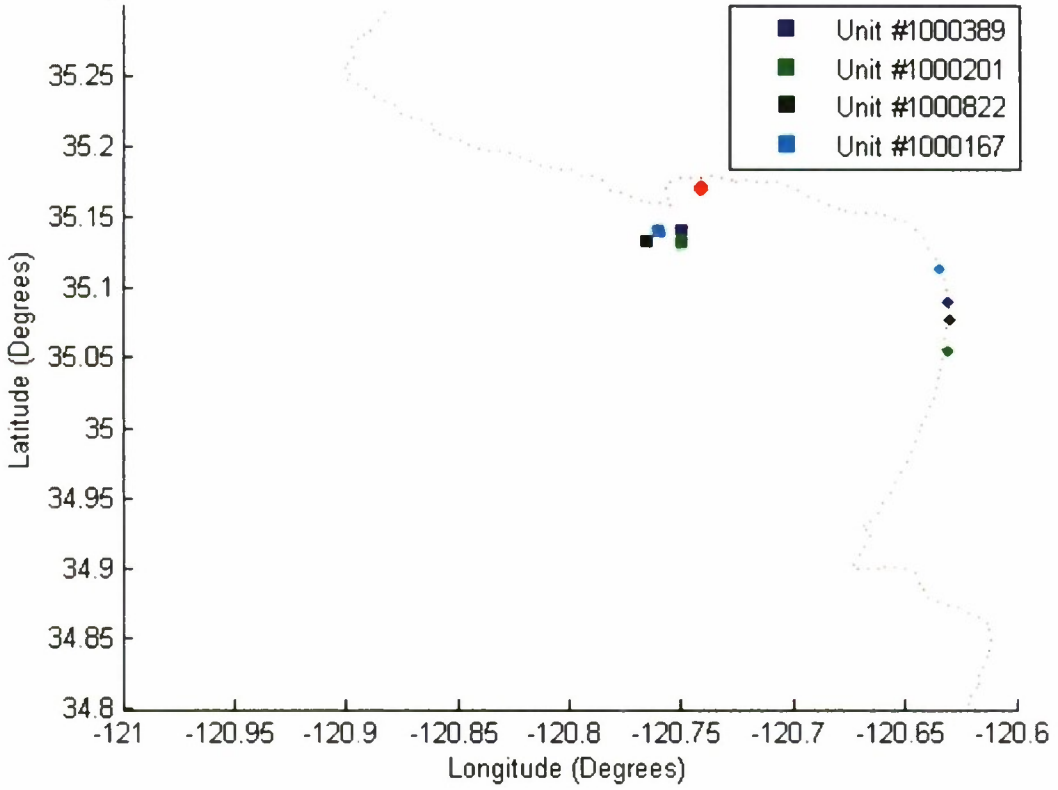




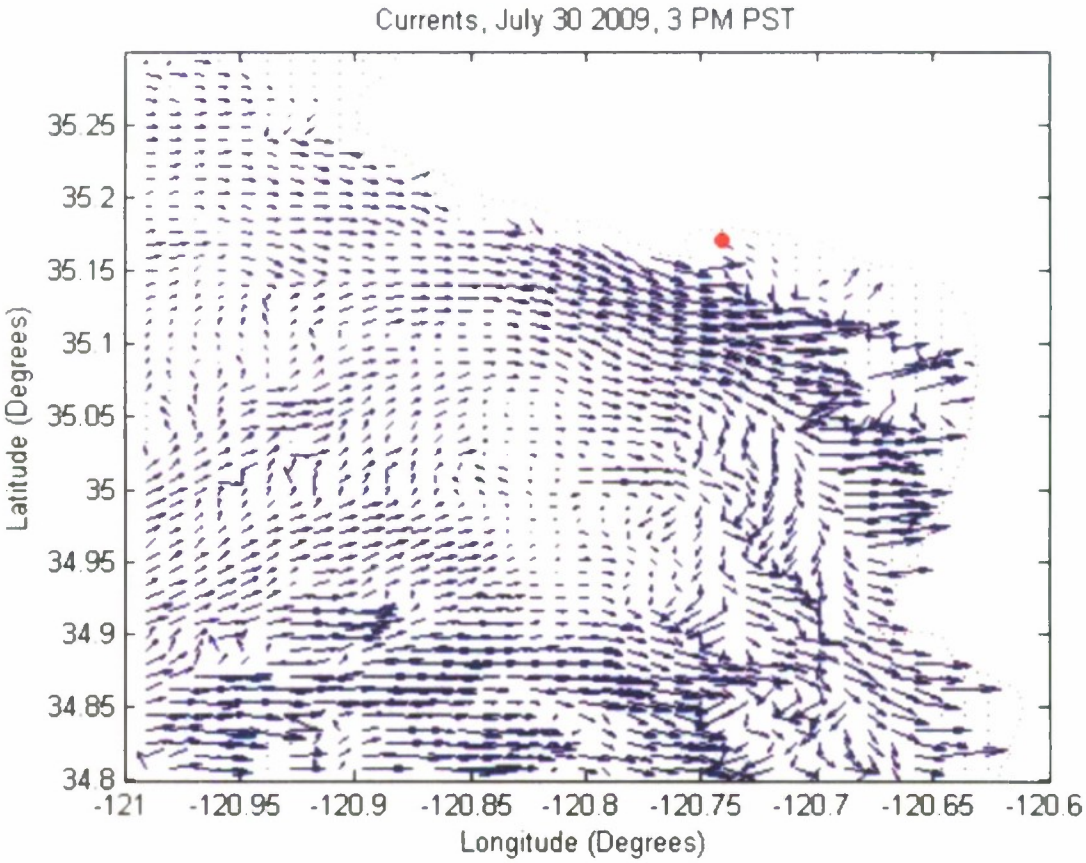


We performed a trial run of four of our new ocean drifters, by deploying them just south of the Point San Luis breakwater in Avila Bay. The starting and ending locations are shown on the map below. (The red dot indicates the Cal Poly pier.) Unfortunately there was an error that prevented the communication of the GPS systems during the majority of this particular test. This error has now been remedied, so that complete drifter tracks will be obtained in future runs. The starting and ending point locations (below) suggest that in this case the ocean carried the drifters in a rather simple track toward the southeast. And in this case we learned that the currents swept the drifters to Oceano Dunes in a period as short as 30 hours for the first drifter, and as much as 42 hours for the last drifter.

Plate 5. Starting and ending locations of drifters. July 30 – August 1, 2009.
The starting locations are above in the plot (north), near the legend and the ending locations are south-east of the initial positions. The red dot is the location of the Cal Poly pier facility.



Plates 6a and 6b. Surface velocity maps of currents near Avila Bay, based on CODAR data. The drifters were deployed at about 11 am on July 30, 2009. The first plot shows currents on the day the drifters were deployed. The second plot shows currents the following day when two drifters washed up on Oceano Dunes beach. This plot was rendered by physics major Chris Ferguson.



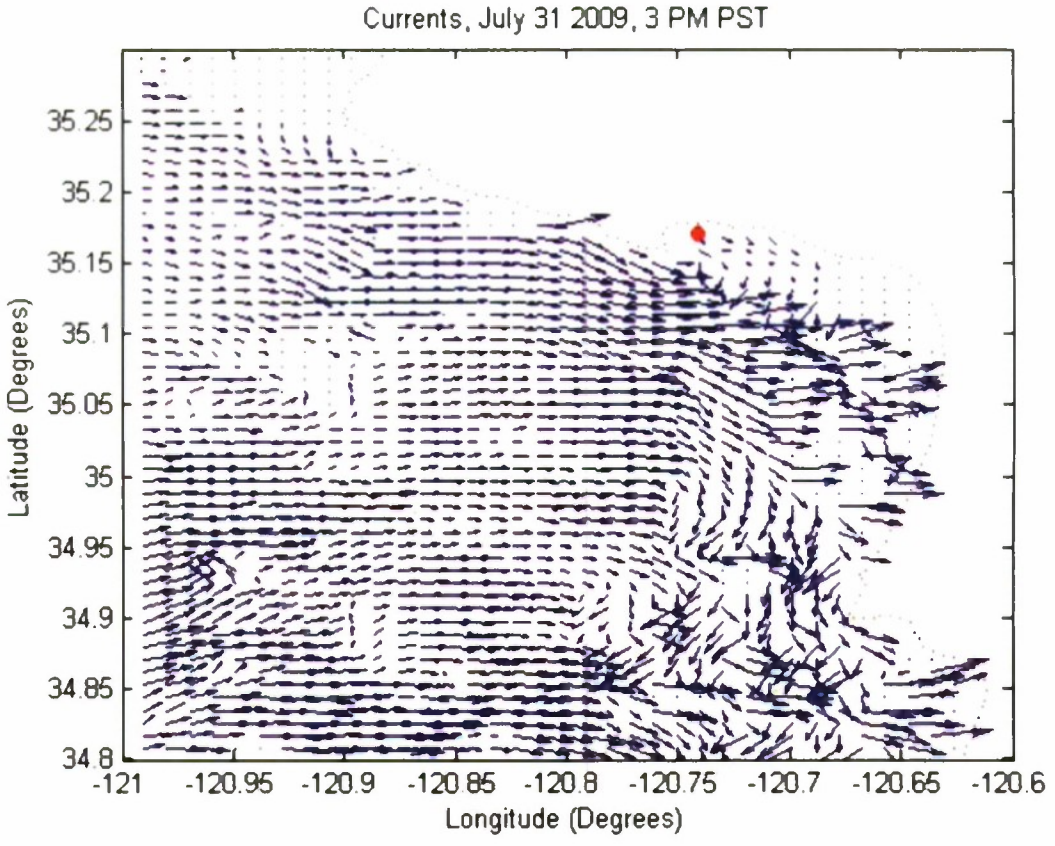


Plate 7. Wind direction during the July 30 – August 1, 2009 drifter run.

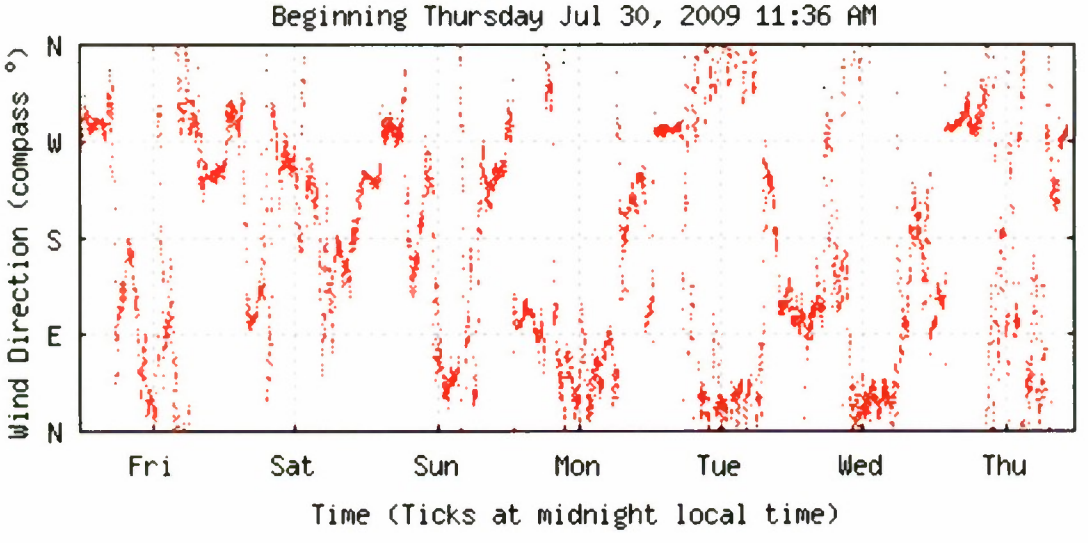
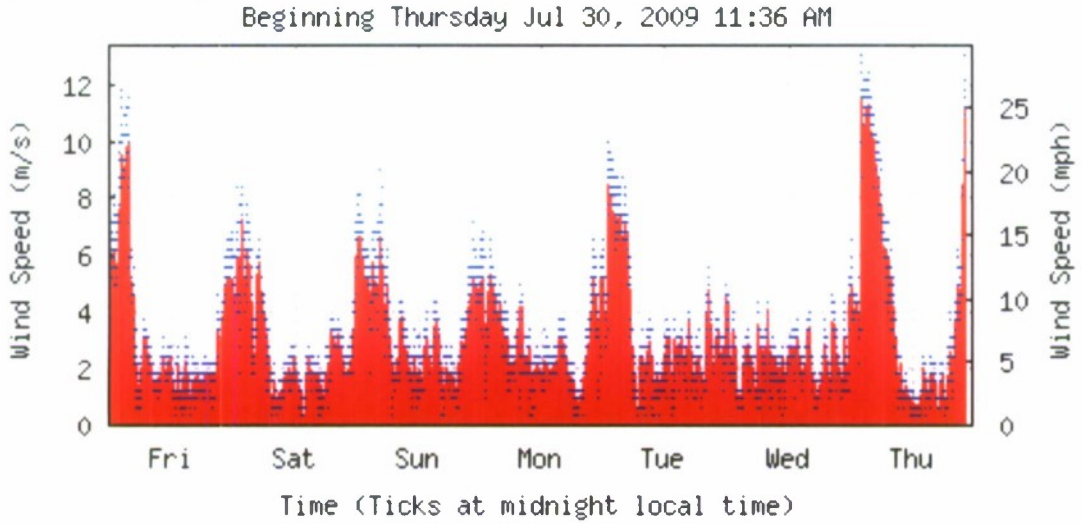


Plate 8. Wind speed during the July 30 – August 1, 2009 drifter run.



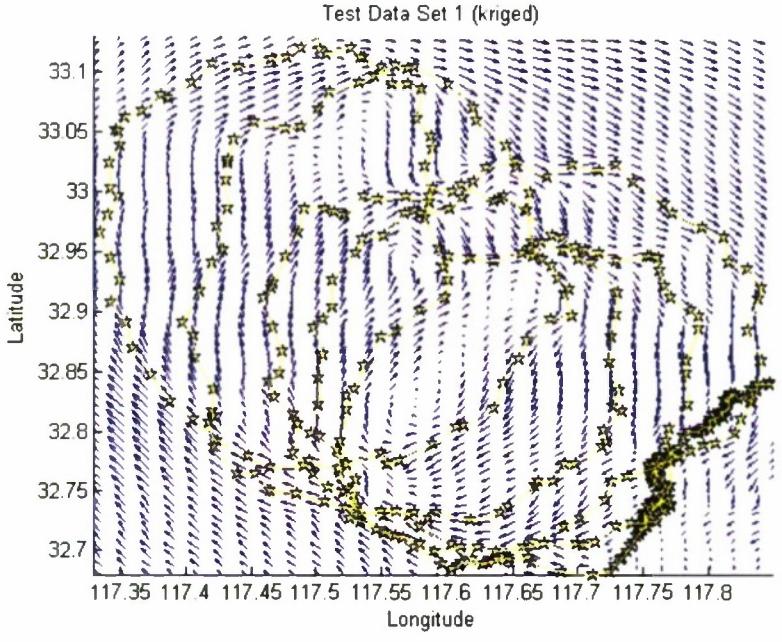
As mentioned on the first page, there were six major goals of our project this year. Here is a summary of the outcomes:

A. We successfully extended the range of the drifters. Last year, the tracking range was limited to 10 miles from Avila Bay. Now we can receive data (via satellite) from up and down the west coast - from Alaska, south america, and even a large portion of the Pacific Ocean. This was achieved by upgrading the electronics, then redesigning the drifters (including the ballast, top and bottom rings, etc) to accommodate the new equipment.
B. We increased the battery life of the fleet from 12 hours to 14 months. This will allow prolonged data collection to quantify larger-scale ocean surface currents.
C. We improved the capability of the drifters to more accurately follow the ocean water, and be less influenced by the wind. This was done by increasing the so-called drag area ratio, which is equal to the cross-sectional area of the drifter that is exposed to the flow of water divided by the cross-sectional area that is exposed to the flow of wind. We more than doubled this ratio, so that our new drag area ratio of slightly greater than 40 is as good as ocean drifters being used in the SVP (Surface Velocity Program), ARGOS program, etc.
D. We made the investigations safer for Cal Poly participants by eliminating the need for boat pick-up of the drifters. Our new method is to wait for the drifters to wash up on shore, while we track and plot their locations on web-based software.
E. We enlarged the fleet from 8 to 16 units to enable more extensive ocean surface current mapping.
F. We conducted trials of the new system and found that the approach with the new electronics is feasible for future deployments and oceanographic studies. We will continue these trials and initial data collection during the summer of 2009.

Data analysis:

In his senior thesis to support this project, undergraduate physics major Tom Baker practiced using analysis tools for the ocean drifter data. For example, he practiced interpolating a velocity field based on a *simulated* ocean drifter track. The method that oceanographers use to interpolate and extrapolate their ocean drifter data from irregularly spaced points onto a fully resolved and regular grid is called Kriging. We used the Kriging software developed by Woods Hole Oceanographic (called EasyKrig 3.0) to accomplish this.

Plate 9. Analysis practice: Kriged data from a simulated drifter track. (T. Baker, 2009)



Following the analysis approaches of modern oceanographers, we will use the drifter data to create vector maps, or "snapshots" of the surface currents around Avila Bay, and maps of mean and eddy kinetic energy along the coast. This will help us to understand the evolution of the ocean flow (ie the flow of energy from large scales to small scales) including its turbulent structure. We will also calculate an important transport statistic - the residence time of ocean water in Avila Bay and at other locations along the coast. And we will estimate the rate of dispersion at these locations. One example of how these quantities are useful is that they help biologists quanify the living environment of organisms in the ocean such as phytoplankton and fish. By analyzing the flow with respect to various parameters (wind, tides, bathymetery, etc) we will contribute to the more accurate prediction of coastal flows as a function of environmental variables.

Shock and Vibration Hardened Data Acquisition Device

Project Investigator:

Garrett Hall
Department of Civil and Environmental Engineering
California Polytechnic State University
San Luis Obispo, CA

Shock and Vibration Hardened Data Acquisition Device

1 Introduction

The intent of this report is to outline the preliminary design of a hardened case intended to protect and isolate a data acquisition system (DAS). The primary concern is the survivability of the device with respect to potential external conditions, specifically shock and vibration. The modeling approach consisted of the following tasks.

- **Model DAS Specification:** The example DAS used in this study was provided by Ron Meritt of Meritt International. A hardware specimen was delivered as an example of a data acquisition system that would require hardening in order to survive shocks of the type possible in a mobile security surveillance scenario.
- **Performance Specifications:** Appropriate integrity standards were adopted for the design of the hardening system. The inputs included shock/vibration specifications appropriate for the in-situ conditions anticipated by the DAS application.
- **Material specification:** Through mechanical design appropriate materials were determined for the DAS hardening. Alloys, composites, and hybrid materials were considered along with the economic aspects of each.
- **Geometry:** Based on the design constraints determine an appropriate geometric configuration for the DAS hardening system. In addition to geometric constraints consideration was also given to ease of fabrication.

- **Numerical modeling:** A numerical model of the combined DAS and hardening system was developed with the intent of utilizing the numerical model to help guide the design, fabrication, and production phases. Numerical studies of the preliminary design included:
  - **Frequency analysis:** The natural frequencies of the combined DAS/hardening system were estimated through a numerical finite element model.
  - **PSD simulation:** A power spectral density analysis of the system was performed based on well known integrity testing protocols (*e.g.* MIL-STD-810E minimum integrity test). Virtual tests were run in three orthogonal directions to determine the anticipated effects and sensitivity to the PSD load as measured at the DAS.
  - **Blast simulation:** Numerical time history analyses were conducted to simulate performance during a typical blast event. The simulations provide a qualitative measure of the system integrity.
  - **Thermal analysis:** Thermal loads created by the DAS were considered with respect to the hardening system to evaluate the potential for overheating caused by the hardening enclosure.

  The remainder of the report summarizes outcomes from the above tasks.

2 Model DAS Specification

Figures 1 through 5 below provide an overview of the DAS provided by Meritt International . The first three photographs show the exterior of the case with the hard drive carrier removed. The last two photographs show the hard drive carrier from top and bottom views. Note that the hard drive tray itself is isolated from the carrier by four black bushings.



Figure 1: Top view of Meritt DAS



Figure 2: Front view of Meritt DAS



Figure 3: Rear view of Meritt DAS

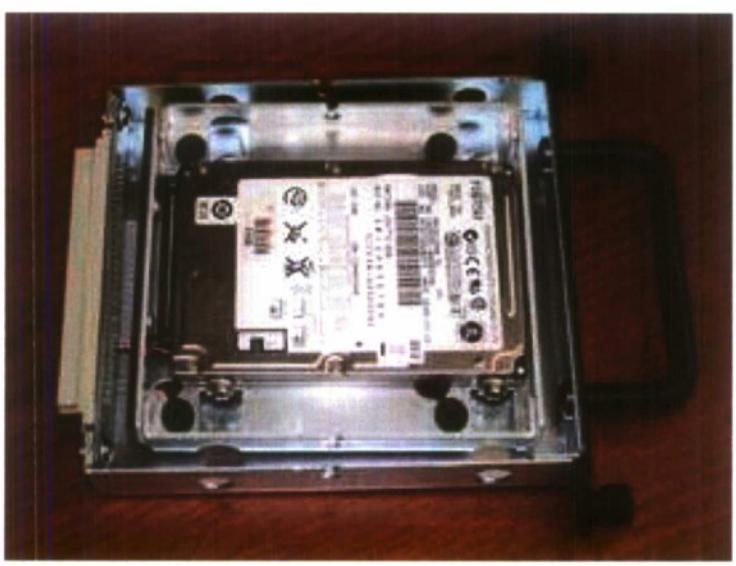


Figure 4: Top view of Meritt DAS internal hard drive



Figure 5: Bottom view of Meritt DAS internal hard drive carrier

All dimensions and material specifications for the DAS in the subsequent sections were based on the hardware provided by Meritt International. Note however, that the design approach and hardening device is easily modified for alternative data acquisition systems.

3 Performance and Testing Specifications

The recommended testing protocols and guidelines potentially applicable to the present design with respect to shock and vibration include:

- Vibration: MIL-STD 810F Method 514
- Shock: MIL-STD 810F Method 516
- Pyroshock: MIL-STD 810F Method 517
- Ballistic shock: MIL-STD 810F Method 522

The above listed guides all emphasize the need for *tailoring* any and all experimental validation to the specific environmental conditions present or potentially expected over the life of the component. In addition there may be other applicable guidelines depending on the specifics of the DAS. For example electromagnetic interference testing may refer to MIL-STD 461F

for guidance. In this report only shock, vibration, and operational thermal loads were considered as described next.

4 Numerical Modeling

The design was arrived at through an iterative series of numerical simulations which considered the eigenvalues, psd response, and blast event response of the system under design. In addition a thermal analysis of the system was conducted to estimate the interior temperature of the system. The following four sections provide a brief overview of the numerical simulations conducted on the preliminary design(s). The basic conceptual design is given in the figure below.

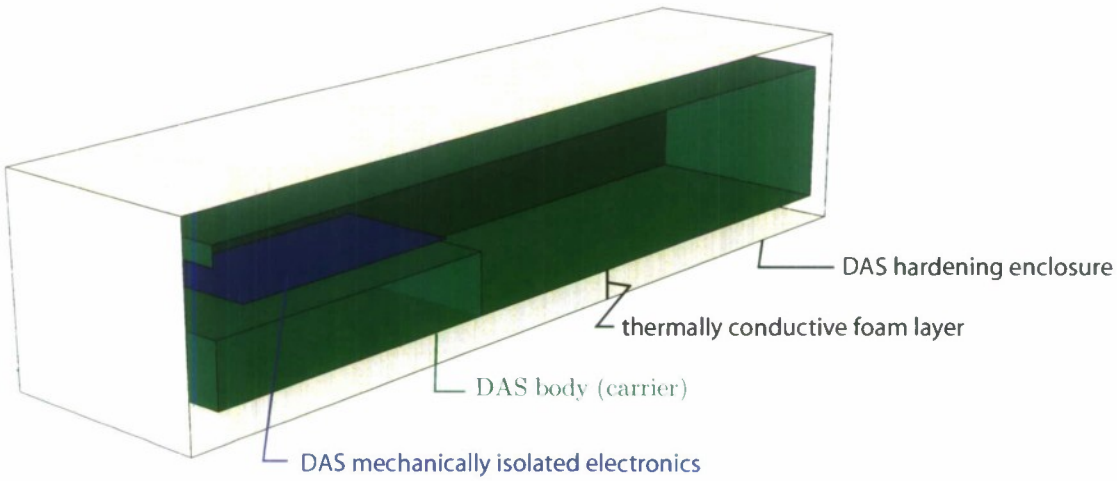


Figure 6: Conceptual design.

4.1 Eigenvalue Analysis

The natural frequencies of the DAS were calculated for each design as a measure of the intrinsic dynamic properties in the system. The table below lists the frequencies and their source (i.e. if the predominate contribution is in the case or in the interior equipment). Following the table graphical representations of the mode shapes (eigenvectors) is provided in Figures 7 through 13.

Mode	Frequency (Hz)	Predominate Part
1	102.55	Hard Drive
2	102.85	Hard Drive
3	105.43	Hard Drive
4	184.69	Hard Drive
5	188.36	Hard Drive
6	196.17	Hard Drive
7	732.73	Hardening Case
8	746.77	Hardening Case
9	793.55	Hard Drive
10	836.2	Hardening Case
11	994.08	Hardening Case
12	1037.3	Hardening Case
13	1091.8	Hard Drive
14	1294.5	Hardening Case
15	1379.5	Hard Drive
16	1414.2	Hardening Case

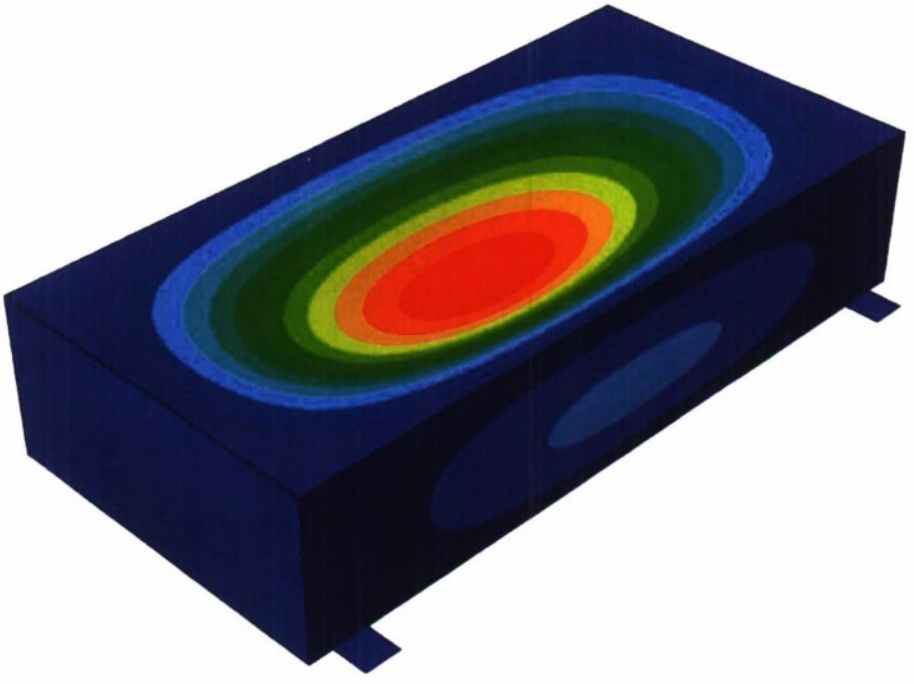


Figure 7: Eigenmode for the 7^{th} eigenvalue.

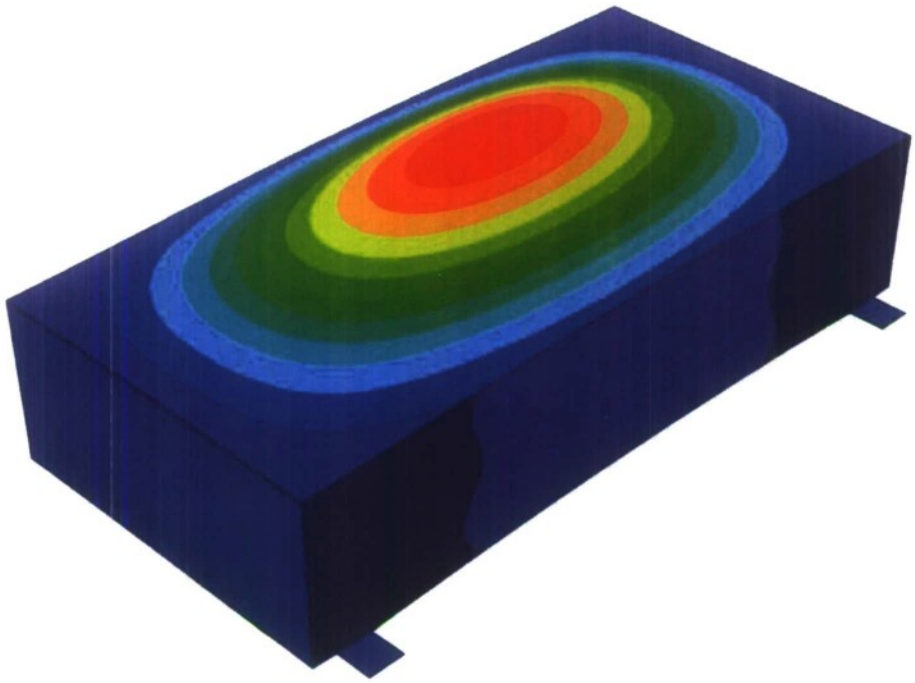


Figure 8: Eigenmode for the 8^{th} eigenvalue.

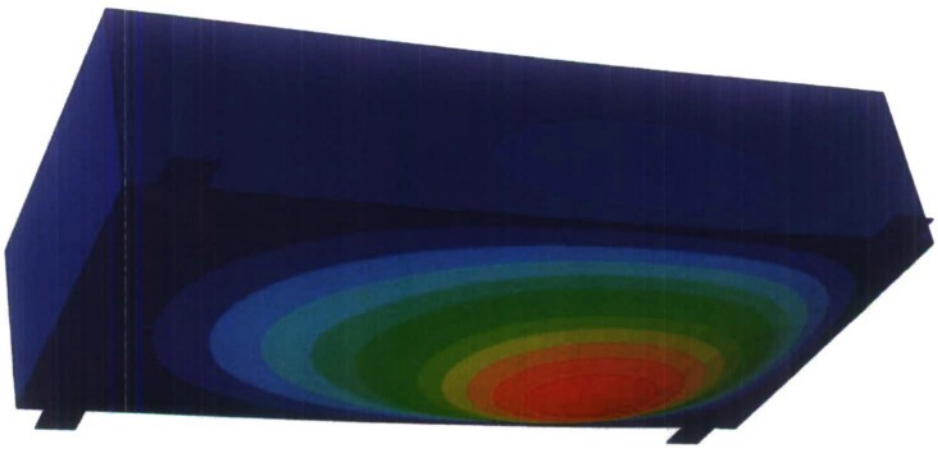


Figure 9: Eigenmode for the 10^{th} eigenvalue.

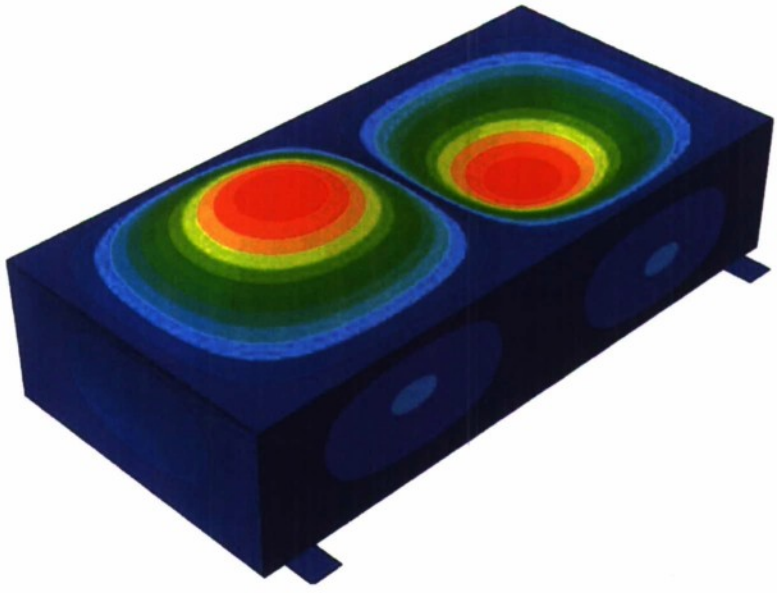


Figure 10: Eigenmode for the 11^{th} eigenvalue.

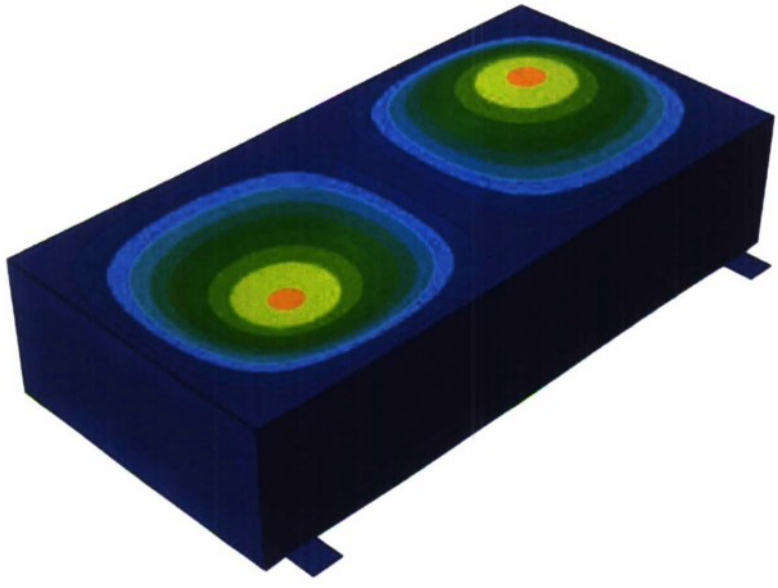


Figure 11: Eigenmode for the 12^{th} eigenvalue.

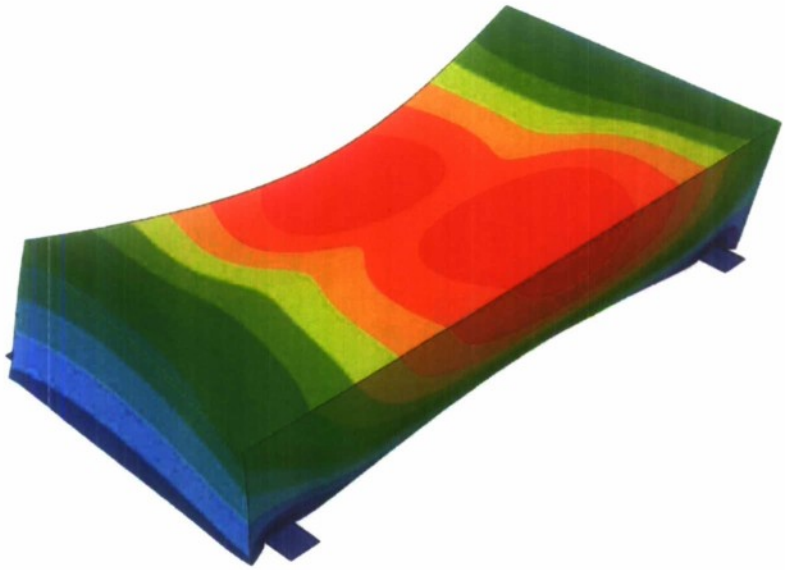


Figure 12: Eigenmode for the 14^{th} eigenvalue.

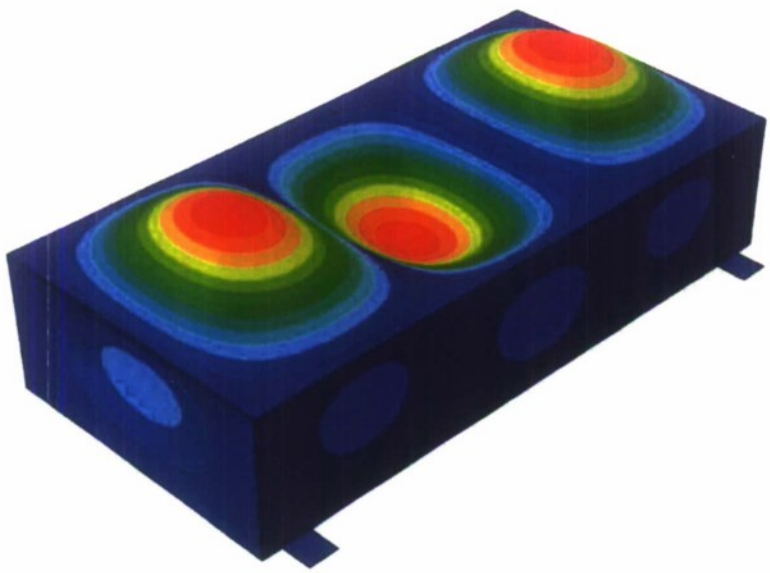


Figure 13: Eigenmode for the 16^{th} eigenvalue.

4.2 PSD Assessment

A PSD analysis was run based on the military specification MIL-STD-810F (see Figure 14 below for an example PSD curve). Figure 15 indicates that the displacement response is dominated by the isolation on the hard drive.

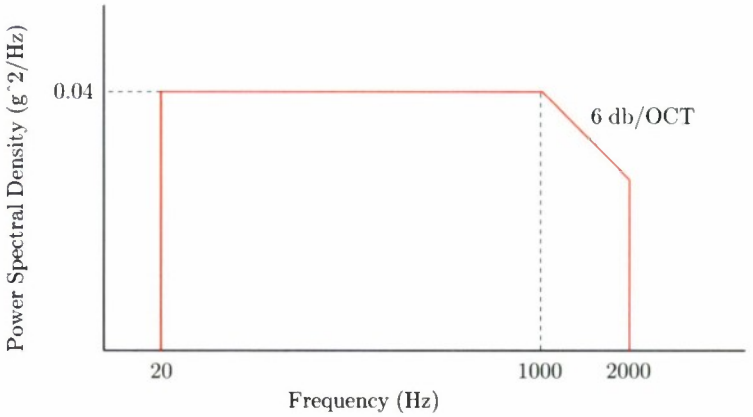


Figure 14: Example PSD specification.

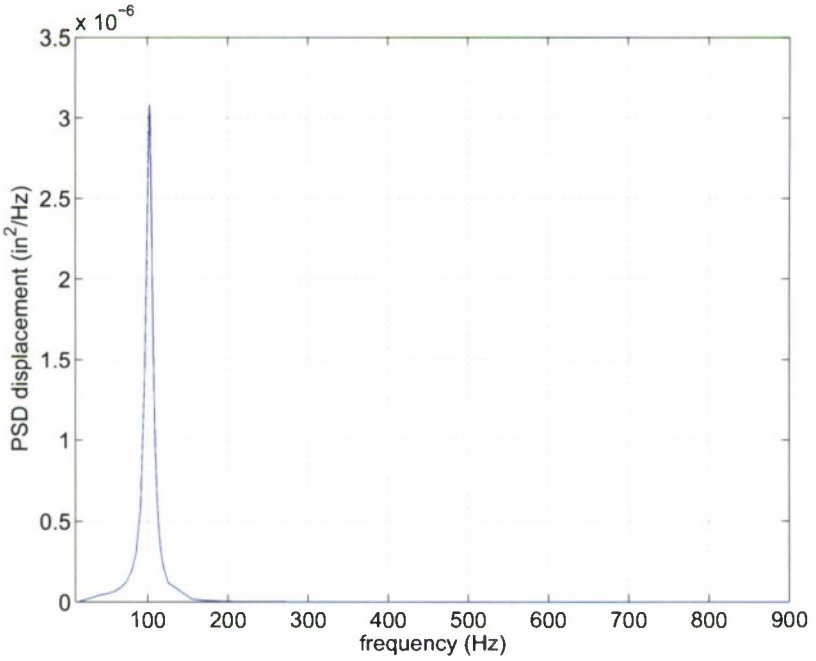


Figure 15: PSD displacement response curve.

4.3 Blast Simulation

This section summarizes the results of a simulated blast events acting on the hardened system. The blast profile was a simplified enveloping form derived from experimental data on flat plates, see Figure 16. The peak pressure used in the simulation is approximately equivalent to 0.25 kg of Pentolite detonated at 50 *cm* from the hardened device [1]. Material properties are industry standard values as listed herein. Note that for steel a dynamic increase factor has been applied to the yield stress (*e.g.* see [5] and the references therein).

The displacement response of the protected hard drive as a function of time is reported for each of the three blast profiles (top, side, and front) in Figures 17-19 respectively. Note that the magnitude of the displacement in each event are of the same order of magnitude. Consistent with the low displacement magnitudes, the accelerations are high, though the final values will depend upon the particular form of isolation acting between the carrier and the hard drive itself. In the present simulation a rubber isolation bushing system was assumed.

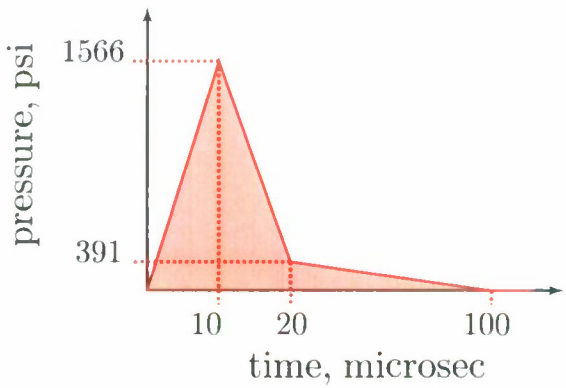


Figure 16: Simplified blast profile (see [1]).

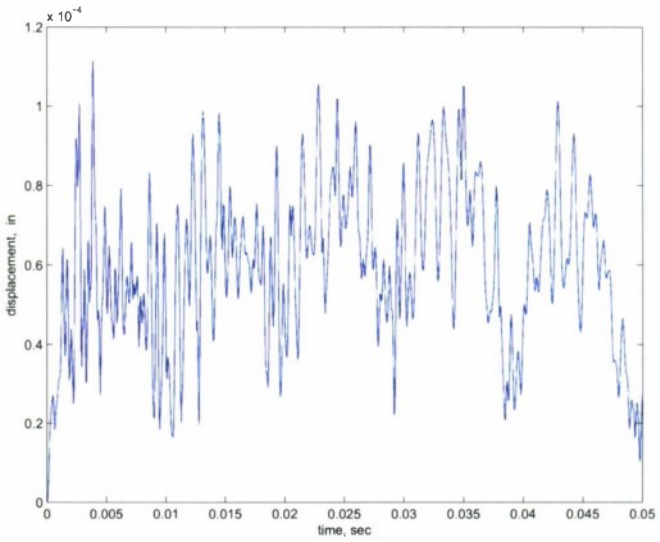


Figure 17: Average displacement magnitude of the hard drive, top blast event.

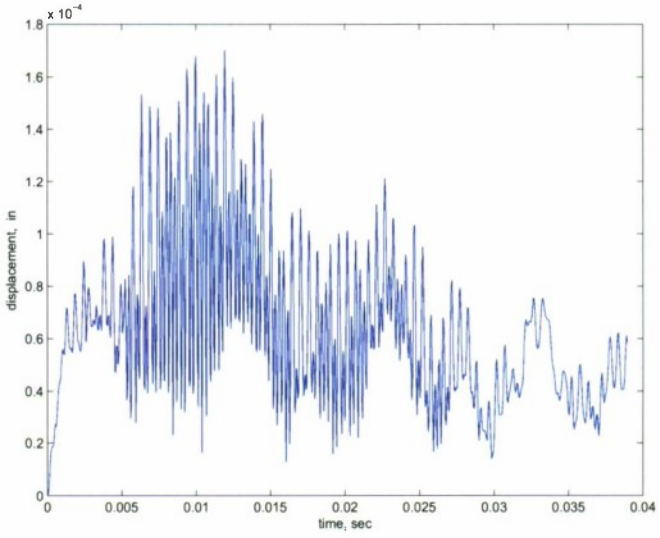


Figure 18: Average displacement magnitude of the hard drive, side blast event.

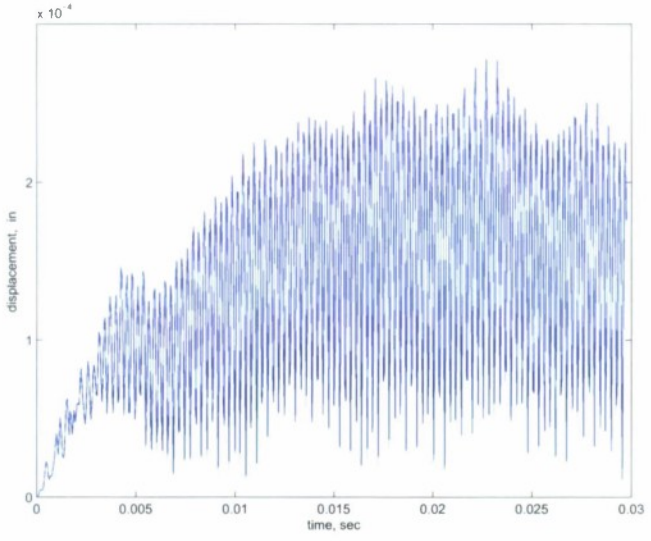


Figure 19: Average displacement magnitude of the hard drive, front blast event.

4.4 Thermal Analysis

The study summarized below examined the thermal performance of the system in a preliminary manner. The analysis considers heat conduction and surface heat convection for lower and upper bounds estimates to the thermal load on the interior of the DAS. The upper bound assumes 50.6 Watts and the lower bound 3.40 Watts.

Materials The hardening enclosure design considered various potential materials in the preliminary stages. The final materials settled upon consisted of a steel enclosure body, thermally conductive foam interface material, fire resistant O-ring gasket, and an Amphenol type harness connector.

Geometry and Boundary Conditions The only part of the enclosure considered in contact with an external sink was the four regions immediately surrounding the bolt hold down locations. The exterior of the enclosure was allowed to convect to still air at the same temperature as the sink (75 Fahrenheit). The interior was loaded with the aforementioned heat generation from the electronics. Note that all of these modeling assumptions would have to be tailored to a specific application to verify adequacy of the thermal design.

The present analysis answers the question of feasibility with regard to the DAS enclosure design.

Results The nodal temperatures and heat flux for the two load cases are shown below. The results are provided at steady state conditions. It is noted that depending upon the electronics considered, the upper bound heat generation is unfavorable whereas the lower end is reasonable.

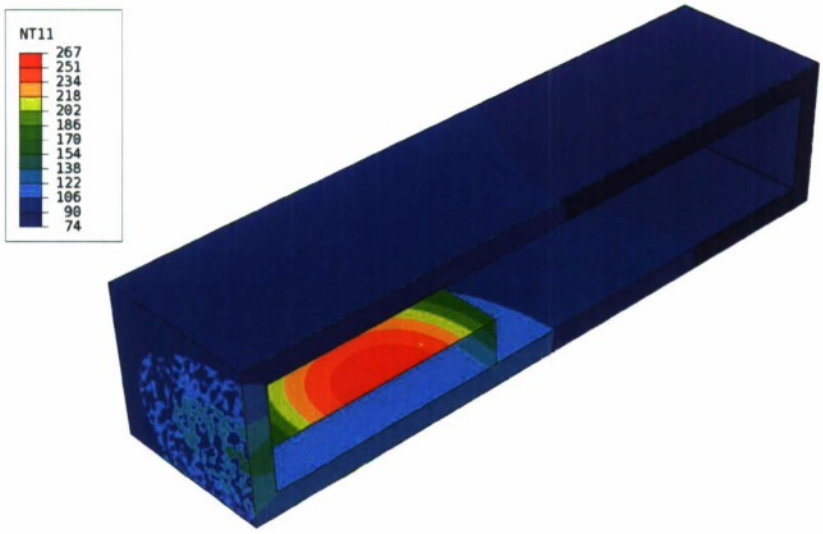


Figure 20: The temperature distribution, in Fahrenheit, for the 50.6 watt case.

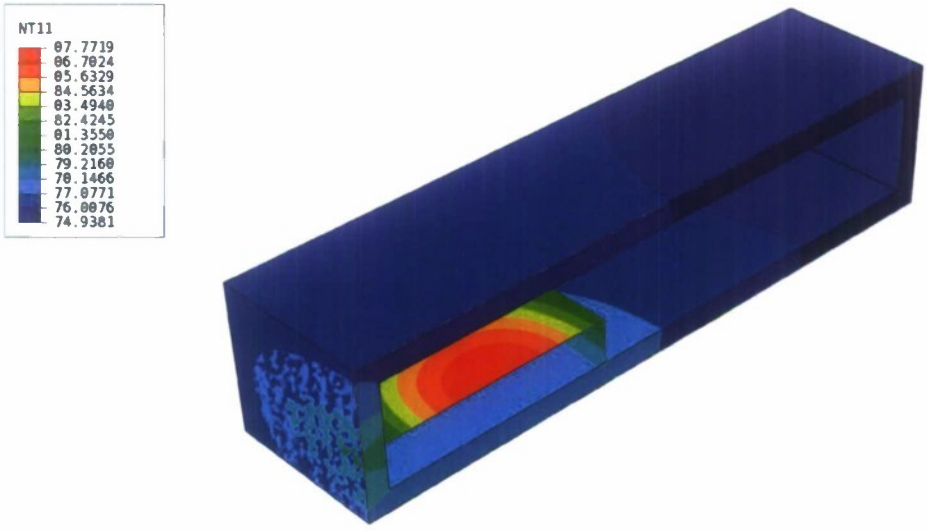


Figure 21: The temperature distribution, in Fahrenheit, for the 3.40 watt case.

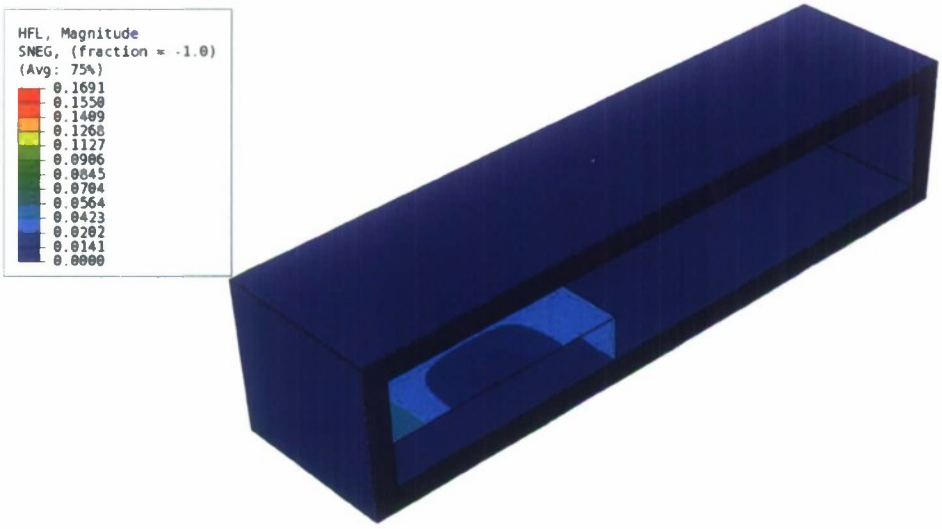


Figure 22: The heat flux for the 50.6 watt case.

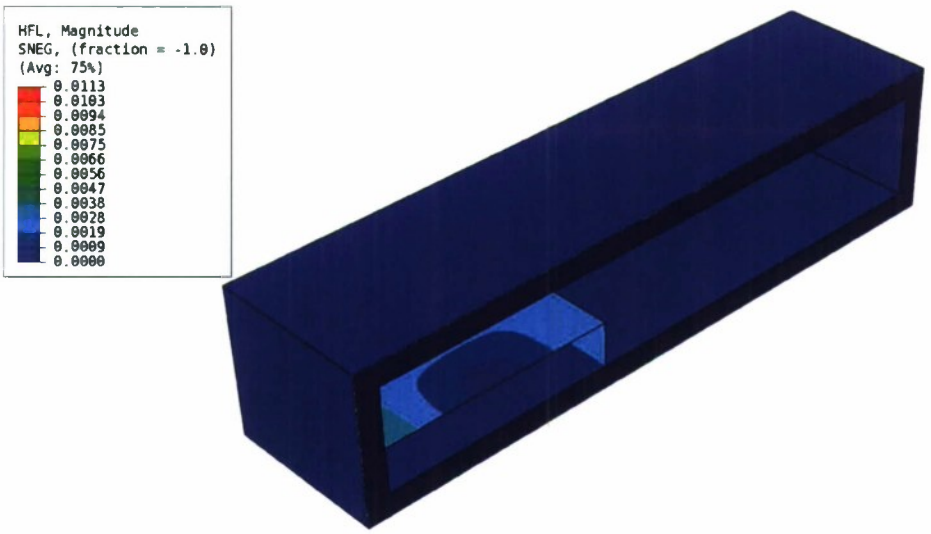


Figure 23: The heat flux for the 3.40 watt case.

5 Summary of the Mechanical Design

Based on the simulated prototype process, the mechanical design is expected to perform adequately for the performance specifications of Section 3. This section provides an overview of the mechanical design. The solid model and summarial drawing files are shown in Figures 29 through 29.

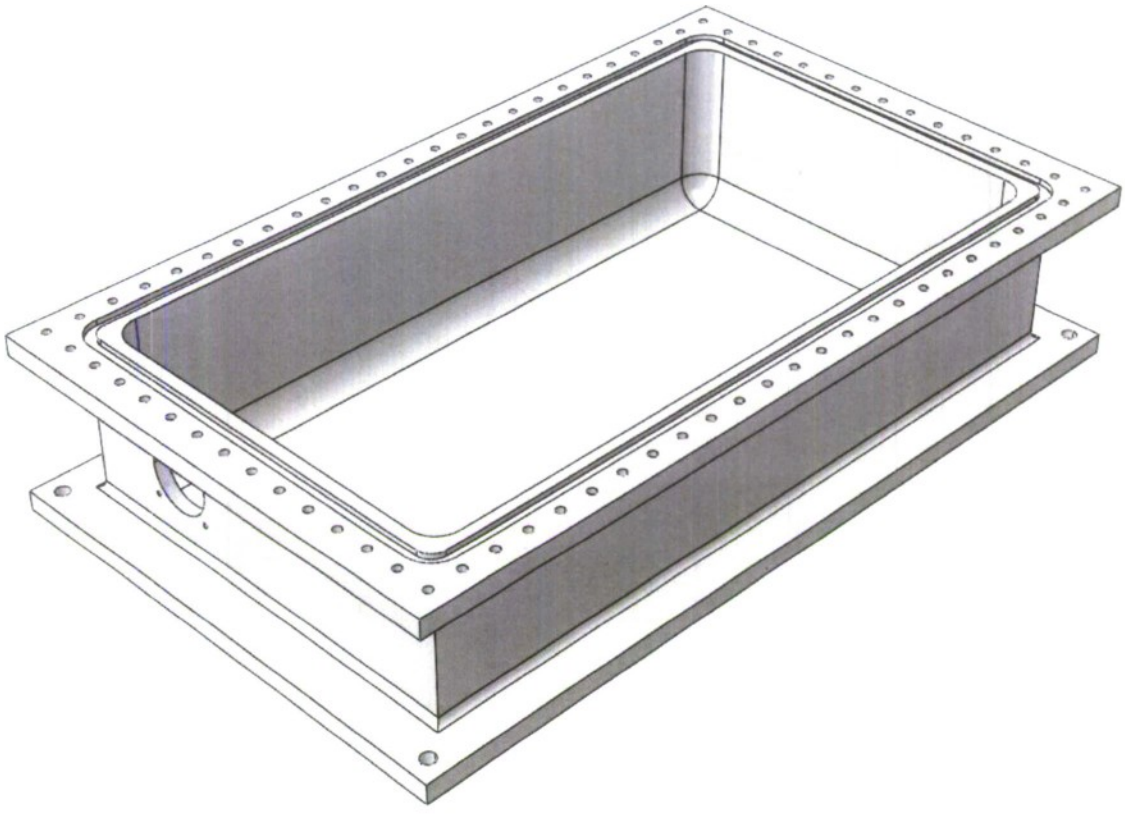


Figure 24: Solid model of DAS body

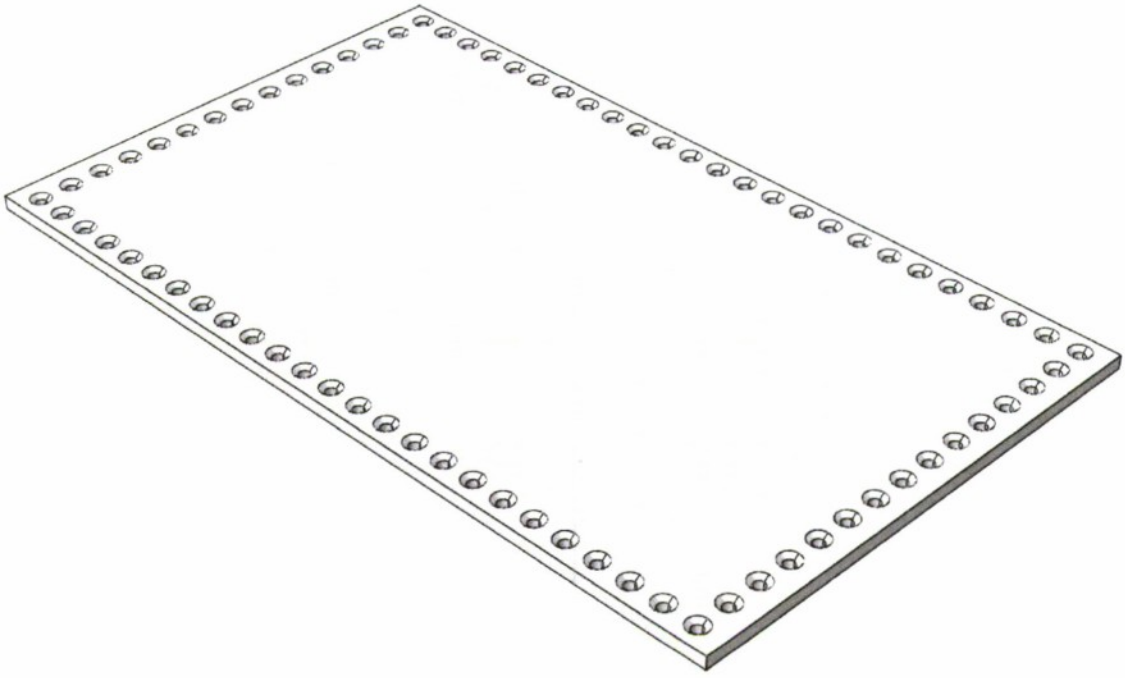


Figure 25: Solid model of DAS cover



Figure 26: Solid model of DAS design

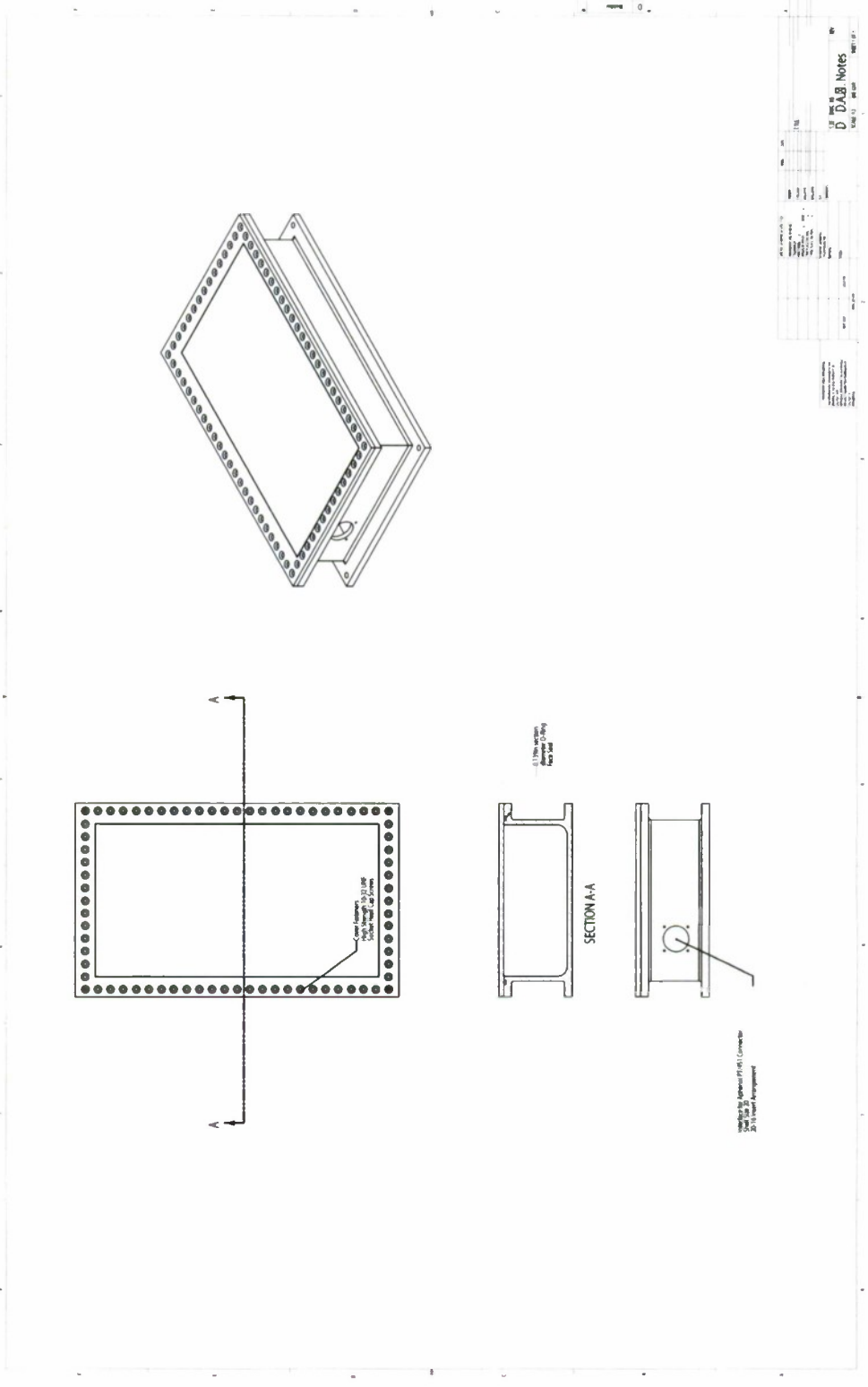


Figure 27: Overview of DAS design

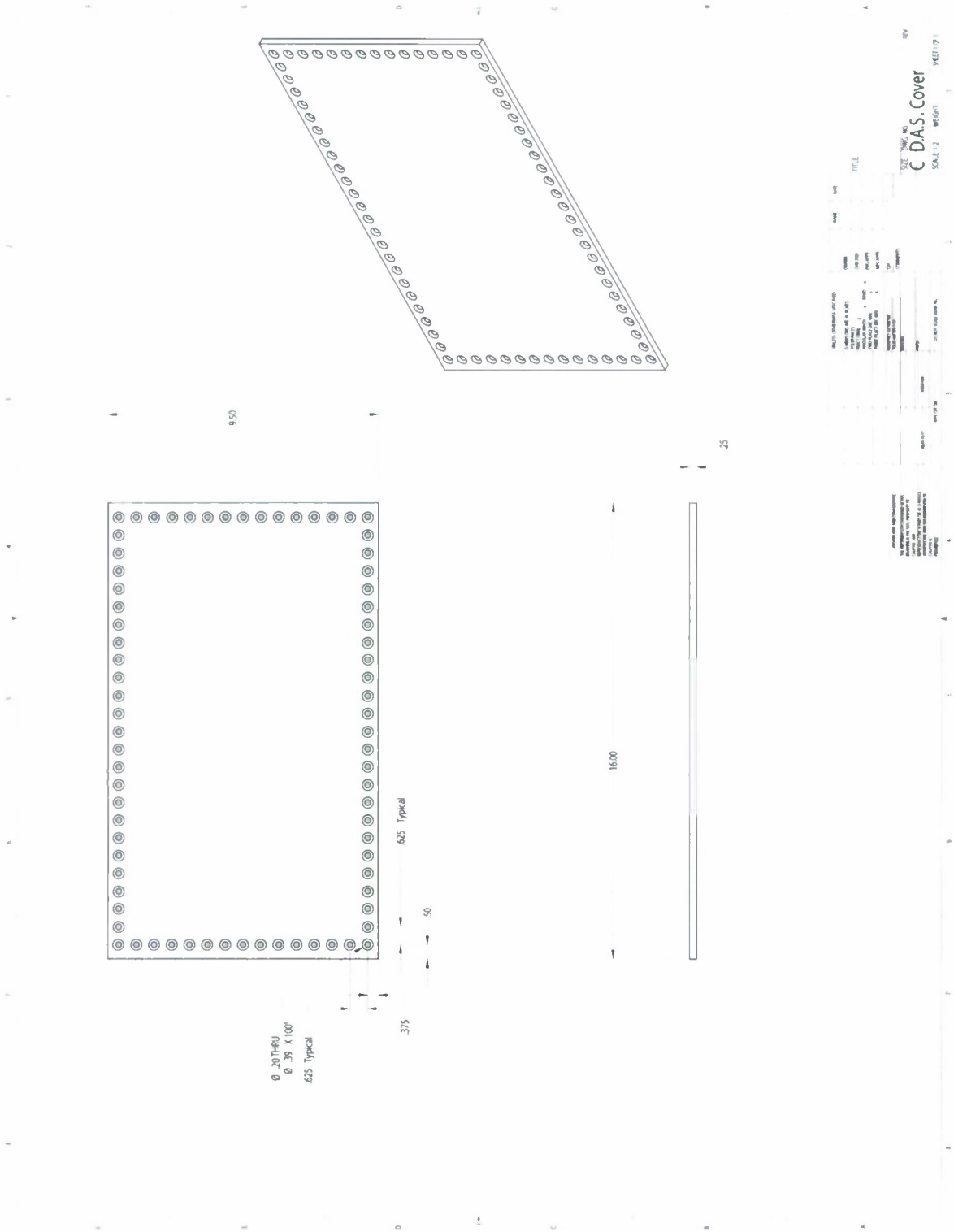


Figure 28: DAS cover design

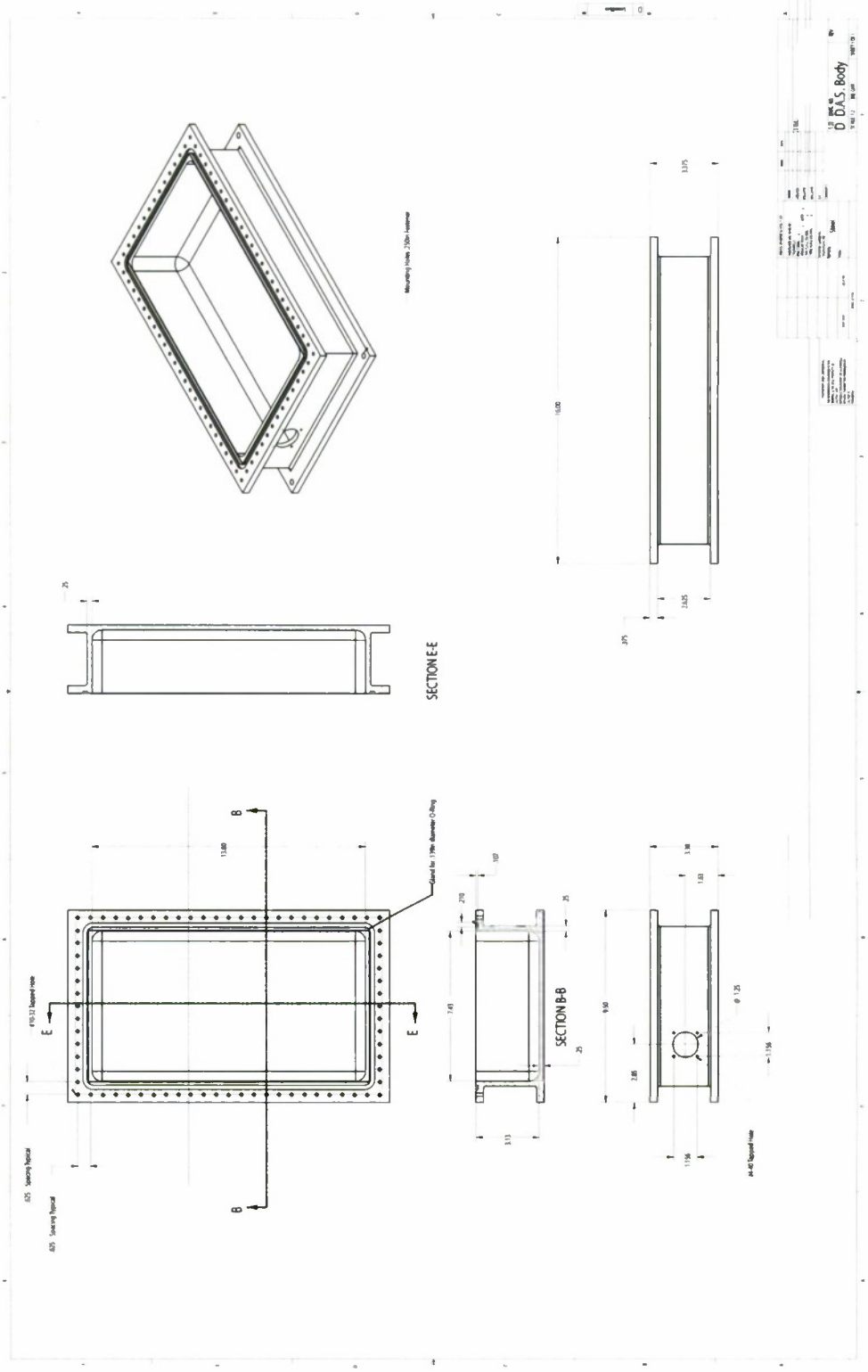


Figure 29: DAS body design

References

[1] Stephen D. Boyd, *Acceleration of a Plate Subject to Explosive Blast Loading - Trial Results*, DSTO Aeronautical and Maritime Research Laboratory **AR-011-421** (2000).

[2] Hibbitt and Karlsson & Sorensen Inc., *ABAQUS*, http://www.hks.com.

[3] T.R.J. Hughes, *The finite element method*, Prentice-Hall, New Jersey, 1987.

[4] Abel C. Jacinto, Ricardo D. Ambrosini, and Rodolfo F. Danesi, *Experimental and computational analysis of plates under air blast loading*, International Journal of Impact Engineering **25** (2001), 927–947.

[5] T. Ngo, A. Gupta, and J. Ramsay, *Blast Loading and Blast Effects on Structures - An Overview*, Electronic Journal of Structural Engineering **Special Issue** (2007), 76–91.

[6] S. Timoshenko and S. Woinowsky-Krieger, *Theory of plates and shells*, 2nd ed., McGraw-Hill, New York, 1987.

[7] A.C. Ugural, *Stresses in plates and shells*, 2nd ed., McGraw-Hill, New York, 1999.

[8] O.C. Zienkiewicz and R.L. Taylor, *The finite element method: Solid mechanics*, 5th ed., vol. 2, Butterworth-Heinemann, Oxford, 2000.

[9] ———, *The finite element method: The basis*, 5th ed., vol. 1, Butterworth-Heinemann, Oxford, 2000.

Bone Mass Preservation and Fracture Risk Assessment with Bisphosphonate Therapy during Space Flight

Project Investigator:
Scott Hazelwood
Biomedical and General Engineering Department
California Polytechnic State University
San Luis Obispo, CA

Abstract

Physiological adaptations in a microgravity environment for astronauts during space flight are common. One such adaptation, bone loss and the subsequent decrease in skeletal strength, has been recognized as a potential problem. Exercise during space missions has not been effective in overcoming bone loss. Bisphosphonates have been used successfully to increase bone mass in osteoporosis patients and offer a possible solution to counter bone loss in a microgravity setting. They work directly on osteoclast activity to suppress bone resorption and increase bone mass. A consequence of their actions, though, is reduced bone turnover leading to an increase in the brittleness of bone. The goal of the proposed research was to develop computational models to examine the effects of bisphosphonate therapy (1) on preserving bone mass during space flight and (2) on the fracture risk of bone for the subsequent return to an environment with gravity. A previously developed remodeling simulation was modified to account for bone loss due to weightlessness and then the effects on bone mass of various bisphosphonate therapy schedules with drugs of different potencies were studied. In addition, the effects of the length of treatment on subsequent fracture risk upon returning to Earth were examined. These models developed in this study will have a wide variety of orthopaedic applications, including examining the effects on bone integrity and fracture risk of other pharmaceuticals to counter bone loss (such as parathyroid hormone or osteoprotegerin therapy), exercise training programs, age, bone diseases and their treatment, and orthopaedic implant designs.

Background

Bone loss has long been recognized as a potential problem for individuals involved in space flight. The effects of microgravity result in several physiological changes for astronauts, including the adaptation of bone to the weightless environment of space which results in the activation of the bone remodeling process to remove bone insufficiently stressed. For longer term flights, a reported 92% of astronauts experienced at least 5% loss of bone at one skeletal site while 40% experienced at least a 10% bone loss (LeBlanc et al., 2007). In general, bone mineral density changes in the spine and femur have resulted in decreases of 0.7% to 2.7% per month in space despite the participation of crewmembers in exercise regimes designed to combat the loss of bone during the mission. In addition, on returning to Earth following space flight, these adaptations can manifest themselves into physical impairments necessitating a period of rehabilitation that may result in the permanent loss of bone mineral density and an increased risk for fracture (Lang et al., 2006; Payne et al., 2007).

Bone loss in a microgravity environment is caused by increased bone remodeling, a two-stage process carried out by teams of cells known as basic multicellular units (BMUs). Resorption of a packet of bone by osteoclasts is followed by refilling of the resorption cavity by osteoblasts. In humans, this sequence typically requires 3-4 months to complete at each site, and the resorption cavities, while individually small, may collectively add substantial porosity to bone. Not only is remodeling the biological process for bone turnover, but it also is believed that remodeling acts as the mechanism for bone to adapt to its mechanical environment and to remove microdamage due to relatively high, cyclic skeletal loads so as to improve fatigue resistance. In support of these beliefs, there is strong experimental evidence that remodeling is activated (1) to remove bone that is insufficiently loaded (due to weightlessness, immobilization,

stress shielding by implants, etc.) and (2) by fatigue microdamage. Li et al. (1990) demonstrated significant and rapid bone loss as a result of increased resorption compared to bone formation in studies where limbs were immobilized, signifying a rapid remodeling response to bone in disuse. Alternatively, Burr and coworkers showed that in vivo fatigue loading of cortical bone to physiologic strains produces damage in the form of microcracks and activates remodeling in spatial proximity to them (Burr et al., 1985; Mori and Burr, 1993). Subsequently, Bentolila et al. (1998) demonstrated a similar phenomenon in the ulnae of rats, which normally do not remodel their cortical bone, suggesting that remodeling in response to nearby microdamage is a fundamental biological response.

Exercise has frequently been used to counter bone loss in a microgravity environment. Exercises during space flight involving resistance training, bicycling, and treadmill activity have been effective to varying degrees for several physiological systems, but have not been effective at preventing the loss of bone (LeBlanc et al., 2007). Bisphosphonates are drugs developed to treat bone diseases involving excessive remodeling, such as osteoporosis. With their ability to suppress bone resorption and increase bone mass, bisphosphonates, either alone or in combination with an exercise program, offer tremendous potential for maintaining bone during space flight, even for periods of long duration in a weightless environment. Bisphosphonates have a high affinity for bone mineral and are incorporated into bone at sites where remodeling has exposed hydroxyapatite (Rogers et al., 1997). They subsequently assert their anti-resorptive effect upon contact with osteoclasts (Lin, 1996). Different bisphosphonates exhibit wide variability in their ability to suppress osteoclastic activity. Those with high potency (> 100× that of etidronate, a first generation drug) can suppress resorption at a relatively low dose, thus avoiding another biological effect of bisphosphonates, the physicochemical impairment of mineralization (Fleisch, 1998). The resulting tissue-level mechanisms for bisphosphonate-mediated increases in bone mass include reduced remodeling space (the temporary porosity due to remodeling) and a positive bone balance (less bone removed than replenished at individual remodeling sites). Suppression of bone resorption by bisphosphonates may involve cellular mechanisms that are direct, such as disruption of osteoclast differentiation, or indirect, such as increasing inhibitory signals from cells of the osteoblast lineage (Breuil, 1999). Other possibilities include shortening the lifespan of osteoclasts by inducing apoptosis (Rodan, 1998), and decreasing their resorbing efficiency by altering the cytoskeleton (Rodan and Fleisch, 1996). Currently, the dominant mechanisms are not clear, but presumably they may vary with the chemical structure of the particular bisphosphonate.

There are abundant data on the positive effects of bisphosphonate treatment to increase bone mass in osteoporotic patients. Liberman et al. (1995) found that 3 years of alendronate treatment increased spinal bone mineral density (BMD) 8.8% and reduced vertebral fractures by 48% relative to controls. Tonino et al. (2000) reported that postmenopausal women's spinal BMD increased throughout 7 years of alendronate treatment. However, reduced bone turnover increases mean skeletal age and the degree of mineralization of existing bone (Boivin et al., 2000; Rodan and Fleisch, 1996), and this may contribute to microdamage accumulation by slowing its removal and making the bone tissue more brittle. In dogs, both alendronate and risedronate treatment for one year resulted in microdamage accumulation that was inversely correlated with BMU activation frequency (Mashiba et al., 2001). While the latter data are from dogs treated at dosages exceeding human clinical levels, they also represent relatively brief treatment periods. Therefore, there is concern that bisphosphonates of high potency may, in the

long-term, lose their ability to reduce fracture risk as microdamage accumulates and negates the benefits of increased bone quantity by decreasing bone quality.

Computational models are useful for studying bone remodeling and the effects of bisphosphonates on bone. For this study, a previously developed computational model for bone remodeling (Hazelwood et al., 2001) was modified to study the effects of bisphosphonate therapy on preserving bone mass in a microgravity setting during space flight and on the fracture risk of bone for the subsequent return to an environment with gravity. This model incorporates the effects of the mechanical environment of bone and microdamage on remodeling, and tracks the effects of bisphosphonate treatment on bone volume fraction, activation frequency, and microdamage.

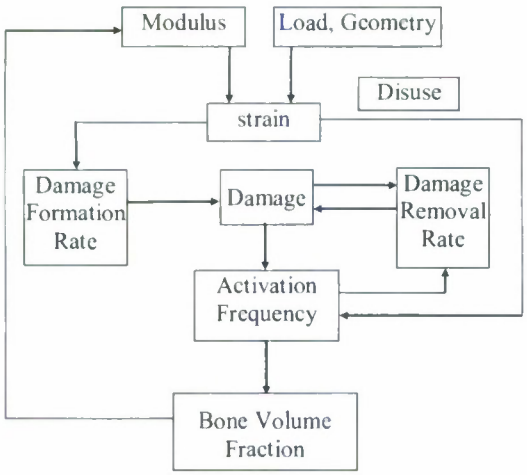


Fig. 1. Bone remodeling algorithm for the computational models.



Fig. 2. Representative volume of bone.

Computer Model Development

Remodeling Simulation

This study utilized a modified version of previously developed computational models for bone remodeling in response to mechanical and biological influences (Hazelwood et al., 2001, Nyman et al., 2004). These bone adaptation models incorporate the coordinated responses of BMUs (or basic multicellular units, the teams of bone cells that resorb and form bone) to two mechanical stimuli known to initiate remodeling: disuse and fatigue damage (Fig. 1). The model tracks the interactive changes in remodeling, strain, and fatigue microdamage that occur in a representative volume of bone (Fig. 2) when the equilibrium of such a system is disturbed. In this model, an empirical relationship was used to relate the continuum level elastic modulus to bone volume fraction. Changes to bone volume fraction (the volume of bone normalized by total volume including void spaces) in the model were functions of the temporal and geometric characteristics of bone remodeling. Damage and disuse were quantified in terms of the mechanical stimulus function, $\Phi = R_L \times \varepsilon^4$, developed by Carter et al. (1987), where R_L is the number of cycles that the strain ε is applied. Fatigue microdamage and bone remodeling were characterized in terms of histomorphometric variables. The microdamage formation rate was

assumed to be proportional to the mechanical stimulus Φ. The rate of damage removal was a function of the current amount of damage, the BMU activation frequency, and the area resorbed by BMUs. BMU activation frequency was calculated daily from the amount of disuse and existing damage. The numbers of resorbing and refilling BMUs active each day were calculated from the activation frequency history over the remodeling period: 25 days for resorption, 5 days for reversal, and 64 days for refilling. Daily bone volume fraction (BVF) changes were then calculated from the net amount of bone removed or added by each resorbing or refilling BMU, respectively. A mechanism was included to allow for less than complete refilling on trabecular surfaces in disuse.

Simulation of Preflight Conditions

Before applying microgravitational conditions, preflight parameters were calculated using constants derived from experimental data. The applied stress on Earth to reach these experimental values was determined by matching the predicted bone volume fraction for the model to 0.22, or equivalently a density of 0.44 g/cm^3, a density typical of vertebral trabecular bone in adult males in Earth's gravitational environment, assuming a linear relationship between bone volume fraction and bone density. The calculated applied stress, 1 MPa, was based on a 100 mm^2 cross-section from the representative trabecular bone volume (Fig. 2).

Simulation of Spaceflight

Microgravity was simulated by lowering the stress applied to the bone volume. Astronauts experience an average vertebral trabecular BMD loss of 0.7% per month, for a total of 4.2% BMD loss during the 180 day average ISS missions (Iwamoto et al., 2005; Lang et al., 2004; LeBlanc et al., 2007). For this simulation, stress applied to the representative volume during spaceflight was then determined as 0.8909 MPa based on this 180 day spaceflight ending density of 0.4215 g/cm^3.

Using the density loss values at 180 days as a baseline, various durations of spaceflight were simulated based on typical mission length for astronauts, including 10 days, 90 days, and 180 days. Although 365 days in space is not typical, it was also simulated to examine the potential of bisphosphonates to maintain BMD without increasing bone microdamage.

Simulation of Bisphosphonate Treatment

Bisphosphonate treatment was simulated using two factors: a decrease in the activation frequency and a reduction in the resorption area. A potency variable (P), where $0 \leq P \leq 1$, was applied to exert the former effect by multiplying the activation frequency of BMUs by the quantity $(1 - P)$. The potency variable P was based on pharmacokinetic properties of bisphosphonates, including potency factors, binding, uptake, and mode of action, $P = P_{max}(1 - e^{-\tau \times N.Rs.BMU})$, where P_{max} and τ are suppression coefficients reflecting various properties of bisphosphonates in order to model a range of drug potencies. Values for these coefficients were selected based on the experimental results from 1 year studies of daily alendronate and daily pamidronate treatment, and modeled the variations in the reduction of BMU activation frequency as seen in these studies.

The size of the resorption eavity is redueed during bisphosphonate treatment due to their effeets on osteoelasts (Nyman et al., 2004), resulting in alterations to the ratio of bone area formed to bone area removed. Two different initial bone balanee ratios (A_F/A_R) for the simulation of bisphosphonate treatment were used based on reduetions of 1/6 and 3/13 to resorption area found in postmenopausal women treated with bisphosphonates for 1 year (Nyman et al., 2004). A bone balanee of 1.0 was assumed for the simulation when bisphosphonate treatment was not in effeet.

Bisphosphonates were simulated during the entire spaeeflight. Preflight treatment was also examined, where bisphosphonates were applied to the simulation at 0, 7, 14, 30, 90, and 180 days preflight.

Simulation of Return to Earth

The return to earth was modeled by resuming preflight bone loading eonditions, with an applied stress of 1 MPa to the representative bone volume. Onee baek on Earth, bisphosphonate therapy was diseontinued. The simulation was extended 365 days postflight to examine inereases or deereases in fraeture risk based on bone mineral density and damage accumulation. Although it takes 1 to 3 years to fully reeover without treatment, a one year postflight examination allowed us to determine if a treatment was sueeessful at maintaining bone mass without aeeumulating more damage.

Model Results

Spaceflight Without Bisphosphonate Treatment

Model results without bisphosphonate treatment were eonsistent with experimental data obtained from spaceflight (Bloomfield, 2006). Bone mineral density (BMD) and mierodamage deereased from preflight levels due to the redueed loads in a mierogravity environment of spaceflight without bisphosphonate treatment (Figs. 3, 4, and 5). The predieted rate of BMD loss was greatest early in the flight, showing smaller ineremental loses as time spent in spaee inereased. The model predieted a marginal change in BMD from preflight eonditions for the typieal flight duration of 10 days; however, postflight predietions showed that BMD eontinued to deerease upon return to Earth until it reaehed a value nearly 3 times lower than at the end of the mission before inereasing to near normal levels about 100 days after entering Earth's gravity.

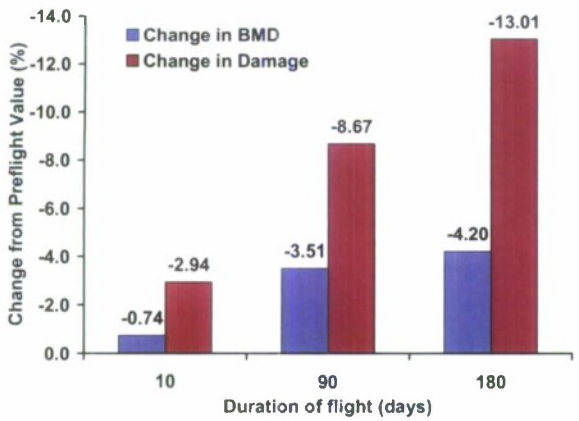


Fig. 3. Predicted percent decreases in BMD and damage at the end of spaceflight without bisphosphonate treatment.

For the 180 day space mission, the predicted postflight results for bone without bisphosphonate treatment showed sharp increases in both BMD and damage upon returning to Earth (Figs. 4 and 5). BMD stabilized approximately 7 years postflight, and was about 2.7% less than its preflight value. Damage continued to accumulate postflight until about 15 years after the return to Earth. Although damage was initially lower upon the return to Earth, it continued to increase postflight for approximately 15 years until reaching a value approximately 9.5% higher than its preflight value.

Spaceflight With Bisphosphonate Treatment

While BMD during a 180 day spaceflight was maintained at or increased above preflight levels with bisphosphonate treatment of low and high potency given at the onset of spaceflight, respectively, damage levels also increased above the baseline values for the high potency drug (Figs. 4 and 5). Following the return to the gravitational environment of Earth, BMD for the high potency bisphosphonate decreased as remodeling activation increased to counteract the higher level of damage. For the low potency drug given at the onset of spaceflight, postflight BMD increased above preflight conditions due to the increased loading upon returning to Earth while damage remained below the preflight level (Figs. 4 and 5).

Results were similar but not as pronounced for the 10 and 90 day spaceflights. For these shorter flight durations, high suppression of BMU activation led to increases in both BMD and damage accumulation during flight. Postflight results suggest that damage remains elevated while BMD decreases for shorter flights with high potency drugs. Simulated bisphosphonate treatment at the beginning of spaceflight with the low potency drug produced decreases in both BMD and damage during the flight for the 10 and 90 day spaceflights. For the shorter simulations with the low potency drug, postflight BMD increased above preflight conditions and damage remained below preflight values. Simulation of the longer 365 day space mission had a more pronounced effect than the 180 day mission, with decreases in damage predicted to be 14.5% during the mission, while BMD exhibited only slight changes, for the low potency drug given at the onset of the mission. For the high potency drug, increases in damage during the 365 day mission reached 7.5% while the increase in BMD was 2.3% with bisphosphonate treatment.

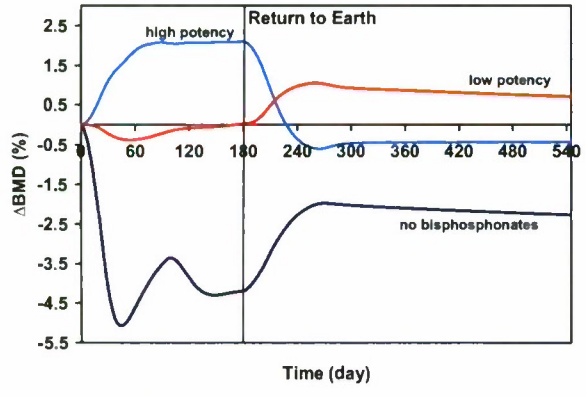


Fig. 4. Simulated effects of spaceflight and bisphosphonate therapy on BMD during a 180 day space mission (beginning at time t=0) and for 365 days following the return to Earth.

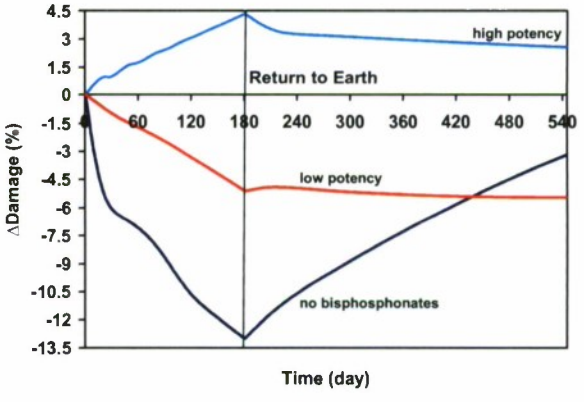


Fig. 5. Simulated effects of spaceflight and bisphosphonate therapy on microdamage during a 180 day space mission (beginning at time t=0) and for 365 days following the return to Earth.

Postflight results for the 365 day space mission predicted decreases in BMD (and elevated amounts of damage) upon returning to the gravitational environment of Earth for the high potency bisphosphonate treatment while results for the low potency drug show an increase in BMD (and decreased damage) compared to preflight values after returning to Earth.

When giving bisphosphonate therapy prior to spaceflight, increasing predicted preflight gains in BMD and damage were observed as the duration of the simulated preflight therapy increased from 7 to 180 days. For longer durations in space, the model predicted these preflight therapy periods to be less effective at influencing the end-of-flight results for BMD and damage. As the duration in space increased, the model predicted the end-of-flight values to be influenced more by the effects of spaceflight than the effects of preflight therapy. Alternately, preflight bisphosphonate therapy highly influenced the model's predictions of BMD and damage for shorter durations of spaceflight. For the bisphosphonate potencies simulated, the addition of preflight treatment caused increased BMD and damage accumulation at the end of 10 day spaceflight. For the 90 day spaceflight, a preflight therapy period of 30 days resulted in increases in BMD without significantly increasing damage, and the same occurred with a 90 day preflight treatment period prior to the 180 day spaceflight.

Conclusions

The computer model developed here combines previous bone remodeling and bisphosphonate algorithms plus spaceflight data obtained from experimental studies in literature in order to better understand the adverse effects of microgravity on bone and fracture risk. The model predicts reduced risk of fracture by increasing bone quantity and either increasing or maintaining bone quality for bisphosphonate treatments (1) with low suppression of remodeling activation and (2) that create higher bone balance ratios. The simulation also predicts significant changes to BMD and damage upon returning to Earth as the remodeling response readjusts to the higher stress conditions. For treatments highly suppressing remodeling activation, these predicted postflight changes include decreased BMD and increased damage accumulation. Low levels of remodeling suppression led the model to predict substantial increases in BMD and small increases in damage postflight.

The model was developed to match the 4.2% loss in BMD over 180 days in space as seen on the International Space Station (ISS). The model's largest predicted BMD loss in untreated, trabecular vertebral bone was 5.01% for a 365 day spaceflight. This is less than half the highest loss (10.8%) seen in Russian cosmonauts on Salyut missions lasting 5 to 7 months (LeBlanc et al., 2007). Although the model does not match the results from these older missions, it is likely due to advancements in technology, physical preparedness of the subjects, and onboard exercise routines that were developed for missions on the International Space Station.

Full recovery for space explorers returning from the ISS took from 1 to 3 years to complete (LeBlanc et al., 2007). The model developed here predicted that full recovery to preflight BMD values may never be attained without treatment. With bisphosphonate treatment, the model predicted complete recovery to occur; some treatments even resulted in higher BMD values than existed preflight. The model suggests that bisphosphonate treatment, combined with exercise, may be the solution that space exploration programs desire to combat bone deterioration in space.

The predicted remodeling response of untreated bone to environmental changes was non-linear, as most BMD was lost or gained early in the transitions from Earth to space and from

space back to Earth. For 10 days in space, the model predicted more bone density to be lost while readjusting to Earth's gravity than lost during spaceflight. This has yet to be examined experimentally, but certainly the model provides insight into a possible trend that may have gone unnoticed. The model predicts bisphosphonate treatment to be beneficial for all durations of spaceflight, not just for longer duration missions. In many instances, preflight treatments were shown to reduce the fracture risk upon return to Earth. Longer simulated flight durations required longer preflight treatments to provide similar effects to those with shorter flights and shorter preflight treatments. The problem with this is that as treatment on Earth is lengthened, damage accumulation increases to such an extent that it could actually cause an increase in fracture risk before entering space. Based on the model's prediction of damage increase, preflight treatment periods longer than 30 days may put the subjects at risk. During preflight treatment, the subjects are still on Earth where higher stresses cause greater increases in damage as compared to space. They may also be exposed to even higher stresses due to exercises in preparation for the mission. These exercises, combined with brittle bones, could lead to fracture before flight.

The model predicts treatments with high suppression of remodeling activation to have the highest gain in BMD at the end of flight, and the lowest BMD values 1 year postflight. These treatments almost completely inhibited remodeling, causing large amounts of damage to accumulate. Though BMD was much higher, the quality of bone was poor. Upon return to earth and discontinuation of treatment, bone remodeling was no longer inhibited and responded to the high amount of damage by removing bone at a much greater rate than bone formation occurred.

Alternatively, the model predicts treatments with low suppression of activation frequency to have the lowest BMD at the end of spaceflight, but they had the highest BMD and lowest amount of damage 1 year postflight. These treatments allowed a fair amount of remodeling to continue in space, but limited it enough so that bone loss was kept to a minimal amount during flight. A majority of the predicted bone loss came from lost damage, so upon returning to earth damage accumulation did not further activate a remodeling response. These treatments had the largest postflight gains in BMD because the response was mostly due to loading in which bone was added to meet the strength required to support the subject.

The limitations of this model occur where assumptions have been made due to the lack of available information. First, the coefficient to determine the rate of microdamage accumulation was kept constant throughout the simulation, although it is likely to change in space. Other coefficients, too, such as the damage rate exponent or the BMU activation frequency coefficients, are likely to be altered in space and would benefit from further study of bone remodeling in microgravity. Second, the predicted postflight results are limited by the fact that they are based on the same stress applied preflight even though postflight recovery programs enable space explorers to ease back into full loading. This high postflight stress would cause overpredictions of both BMD and damage. Also, a bone balance ratio of 1.0 was instantly applied upon return to Earth, when it is more likely that the ratio would slowly ease back down. Third, the simulation applies a constant stress derived from bone loss to a section of bone rather than deriving the actual strain and applying it to a 3D model of bone. Using a finite element model would create a more accurate remodeling response with more precise loading conditions and allow detailed analysis of the effects of specific exercises on maintaining bone mass. Lastly, the model does not take into account effects of spaceflight on blood flow, drug metabolism, tissue binding, drug elimination, fluid shear stress, or changes in hormone levels. Many of these affect the efficacy of the drug itself. Changes to fluid shear stress in a microgravity environment could affect the mechanosensory ability of bone cells to sense signals indicating bone loading

and would lead to further loss of bone even under heavy excreise. Also, although the modcl tracks changes to thc populations of bone cells, it only accounts for ehanges due to the remodeling responsc and not due to the physiologieal adaptations that may occur in mierogravity. Changes in hormone levels or physiologieal alterations to the populations of bone ecIls could signifieantly alter thc remodeling response in spaee and the responsc to bisphosphonate treatment.

It is elear that the model would signifieantly benefit from further studies on spaccflight. Although the model has to overeome the many unknown variables of bone remodeling, bisphosphonates, and mierogravity, it has shown the ability to provide potential trends for future studies. As new data and information beeomes available, the model's accuracy can be improved and eould eventually bc a tool used for predieting effeets of other treatments as well.

References

Bentolila V, Boyee TM, Fyhrie DP, Drumb R, Skerry TM, Schaffler MB. Intraeortieal remodeling in adult rat long bones after fatigue loading. Bone. 1998;23(3):275-81.

Bloomfield SA. Summary - bone in microgravity environments: "Houston, we have a problem." J Museuloskelet Neuronal Interaet. 2006;6:329-30.

Boivin GY, Chavassieux PM, Santora AC, Yates J, Meunier PJ. Alendronate inereases bone strength by inereasing the mean degree of mineralization of bone tissue in osteoporotie women. Bone. 2000;27(5):687-94.

Breuil V. Mechanisms of action of bisphosphonates. Rev Rhum Engl Ed. 1999;66(6):339-43.

Burr DB, Martin RB, Schaffler MB, Radin EL. Bone remodeling in response to in vivo fatigue microdamage. J Biomech. 1985;18(3):189-200.

Carter DR, Fyhrie DP, Whalen RT. Trabeeular bone density and loading history: regulation of eonneetive tissue biology by mechanieal energy. J Biomeeh. 1987;20(8):785-94.

Fleisch H. Bisphosphonates: mechanisms of action. Endoer Rev. 1998;19(1):80-100.

Hazelwood SJ, Martin RB, Rashid MM, Rodrigo JJ. A meehanistie model for internal bone remodeling exhibits different dynamie responses in disuse and overload. J Biomeeh. 2001;34(3):299-308.

Iwamoto J, Takeda T, Sato Y. Interventions to prevent bone loss in astronauts during spaee flight. Keio J Med. 2005;54(2):55-9.

Lang TF, Leblanc AD, Evans HJ, Lu Y. Adaptation of the proximal femur to skeletal reloading after long-duration spaeeflight. J Bone Miner Res. 2006;21(8):1224-30.

Lang T, LeBlanc A, Evans H, Lu Y, Genant H, Yu A. Cortieal and trabeeular bone mineral loss from the spine and hip in long-duration spaceflight. J Bone Miner Res. 2004;19(6):1006-12.

LcBlanc AD, Spector ER, Evans HJ, Sibonga JD. Skeletal responses to spaee flight and the bed rest analog: a revicw. J Musculoskelet Neuronal Interact. 2007;7(1):33-47.

Li XJ, Jee WS, Chow SY, Woodbury DM. Adaptation of eaneellous bone to aging and immobilization in the rat: a single photon absorptiometry and histomorphometry study. Anat Rec. 1990;227(1):12-24.

Liberman UA, Weiss SR, Bröll J, Minne HW, Quan H, Bell NH, Rodriguez-Portales J, Downs RW Jr, Dequeker J, Favus M. Effect of oral alendronate on bone mineral density and the ineidenee of fraetures in postmenopausal osteoporosis. The Alendronate Phase III Ostcoporosis Treatment Study Group. N Engl J Med. 1995;333(22):1437-43.

Lin JH. Bisphosphonates: a review of their pharmacokinetic properties. Bone. 1996;18(2):75-85.

Mashiba T, Turner CH, Hirano T, Forwood MR, Johnston CC, Burr DB. Effects of suppressed bone turnover by bisphosphonates on microdamage accumulation and biomechanical properties in clinically relevant skeletal sites in beagles. Bone. 2001;28(5):524-31.

Mori S, Burr DB. Increased intracortical remodeling following fatigue damage. Bone. 1993;14(2):103-9.

Nyman JS, Yeh OC, Hazelwood SJ, Martin RB. A theoretical analysis of long-term bisphosphonate effects on trabecular bone volume and microdamage. Bone. 2004;35(1):296-305.

Payne MW, Williams DR, Trudel G. Space flight rehabilitation. Am J Phys Med Rehabil. 2007;86(7):583-91.

Rodan GA. Mechanisms of action of bisphosphonates. Annu Rev Pharmacol Toxicol. 1998;38:375-88.

Rodan GA, Fleisch HA. Bisphosphonates: mechanisms of action. J Clin Invest. 1996;97(12):2692-6.

Rogers MJ, Watts DJ, Russell RG. Overview of bisphosphonates. Cancer. 1997;80(8 Suppl):1652-60.

Tonino RP, Meunier PJ, Emkey R, Rodriguez-Portales JA, Menkes CJ, Wasnich RD, Bone HG, Santora AC, Wu M, Desai R, Ross PD. Skeletal benefits of alendronate: 7-year treatment of postmenopausal osteoporotic women. Phase III Osteoporosis Treatment Study Group. J Clin Endocrinol Metab. 2000;85(9):3109-15.

Conference Abstract

This work resulted in the following presentation at the Orthopaedic Research Society Annual Meeting:

C Gardina and **SJ Hazelwood**: Simulated Bone Mass Preservation and Fracture Risk Assessment with Bisphosphonate Therapy during Spaceflight. *Transactions of the 55th Annual Meeting of the Orthopaedic Research Society* 34: 740, 2009.

N00014-07-1-1153

Simulated Bone Mass Preservation and Fracture Risk Assessment with Bisphosphonate Therapy during Spaceflight

P.J. Susan C. Ozawa Ph.D.
Gardina, C; +Hazelwood, SJ
California Polytechnic State University, San Luis Obispo, CA
Senior author shazelwo@calpoly.edu

INTRODUCTION

Bone loss has long been recognized as a potential problem for individuals involved in space flight. The effects of microgravity result in several physiological changes for astronauts, including the adaptation of bone to the weightless environment of space. For longer term flights, a reported 92% of astronauts experienced at least 5% loss of bone while 40% experienced at least a 10% bone loss [1]. In addition, on returning to Earth following space flight, these adaptations can manifest themselves into physical impairments that may result in the permanent loss of bone mineral density and an increased risk for fracture [2,3].

Exercise has frequently been used to counter the effects of a microgravity environment and has been helpful to varying degrees for several physiological systems, but has not been effective at preventing bone loss [1]. Bisphosphonates are drugs developed to treat bone diseases involving excessive remodeling. With their ability to suppress bone resorption and increase bone mass, bisphosphonates, either alone or in combination with an exercise program, offer tremendous potential for maintaining bone during space flight. In this study, we use a computational model of bone remodeling and bisphosphonate treatment to examine the effects of bisphosphonates on preserving bone mass in a microgravity setting during space flight and on the fracture risk of bone for the subsequent return to an environment with gravity.

METHODS

This study utilized a modified version of a previously developed computational model for bone remodeling and bisphosphonate treatment [4]. This model incorporates the coordinated responses of BMUs (basic multicellular units, the teams of bone cells that resorb and form bone) to two mechanical stimuli that initiate remodeling: disuse and fatigue damage. The model tracks the interactive changes in remodeling, strain, and microdamage in a representative volume of bone. An empirical relationship was used to relate the elastic modulus to bone volume fraction (BVF). Damage and disuse were quantified in terms of the mechanical stimulus, $\Phi = R_L \times \varepsilon^4$, developed by Carter and co-workers [5], where R_L is the number of cycles that the strain ε is applied.

The damage formation rate was assumed to be proportional to Φ. The rate of damage removal rate was a function of the current amount of damage, the BMU activation frequency, and the area resorbed by BMUs. BMU activation frequency was calculated daily from the amount of disuse and existing damage. The numbers of resorbing and refilling BMUs active each day were calculated from the activation frequency history over the remodeling period: 25 days for resorption, 5 days for reversal, and 64 days for refilling. Daily BVF changes were then calculated from the net amount of bone removed or added by each resorbing or refilling BMU, respectively. A mechanism was included to allow for less than complete refilling on trabecular surfaces in disuse.

Preflight conditions were simulated by running the model until it reached a steady state with a BVF of 0.22 (typical for adult vertebral trabecular bone). Microgravity was modeled by reducing the load to simulate a 0.7% bone loss over a 180 day period [1,6]. Bisphosphonate effects were simulated by suppressing activation frequency based on the potency of the drug (low and high potencies) and by reducing the BMU resorption area [4]. Spaceflights were simulated in MatLab for manned mission lengths of 10, 90, and 180 days. Changes in BVF and damage accumulation were analyzed during and up to 1 year after spaceflight. Bisphosphonate treatment was given for the duration of spaceflight and began either at the start of spaceflight or 90 days before.

RESULTS

BVF and microdamage decreased from preflight levels due to the reduced loads in a microgravity environment of a 180 day spaceflight without bisphosphonate treatment (Figs. 1A and 1B). While BVF during spaceflight was maintained at or increased above preflight levels with bisphosphonate treatment of low and high potency, respectively, damage levels also increased above the baseline values for the high potency drug. Following the return to the gravitational environment of Earth, BVF for the high potency bisphosphonate decreased as remodeling activation increased to counteract the higher level of

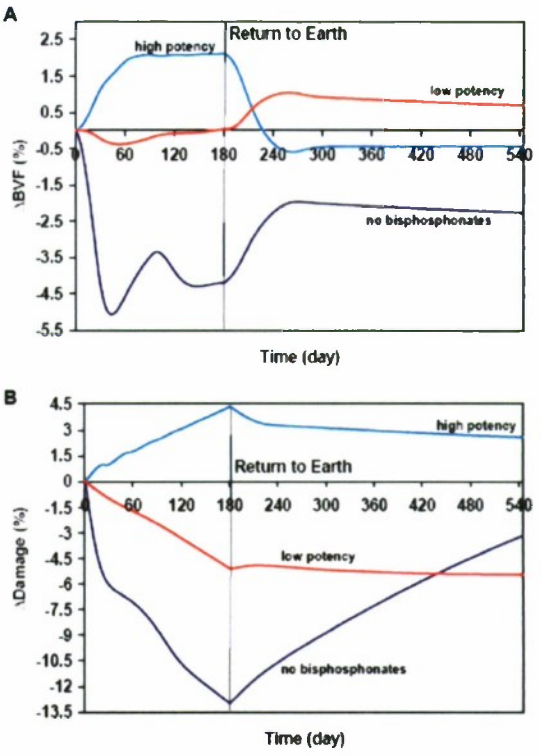


Figure 1. Simulated effects of spaceflight and bisphosphonate therapy on (A) BVF and (B) damage accumulation during a 180 day space mission and for 365 days following the return to Earth.

damage (Figs. 1A and 1B). For the low potency drug, postflight BVF increased above preflight conditions due to the increased loading upon returning to Earth while damage remained below the preflight level.

Results were similar but not as pronounced for the 10 and 90 day spaceflights. While preflight BVF increased with bisphosphonate treatment starting 90 days prior to spaceflight, BVF levels were not substantially increased at 1 year after returning to Earth. Preflight damage levels also increased by giving bisphosphonates prior to flight, and remained elevated following the return to Earth.

DISCUSSION

The computer model developed here combined a bone remodeling simulation with spaceflight data from experimental studies in order to understand the adverse effects of microgravity on bone and explore potential treatment solutions for space explorers. The model predicted reduced fracture risk with the low potency bisphosphonate treatment that increased BVF while decreasing accumulated damage 1 year after spaceflight. For the treatment highly suppressing remodeling activation, on the other hand, the predicted postflight changes included decreased BVF with increased damage accumulation 1 year following the mission. While exercise has not reversed bone loss during spaceflight, it is beneficial for several physiological systems of space explorers. Bisphosphonates, in combination with exercise and diet, may help in the overall benefit of those involved in spaceflight, including aiding in the preservation of bone mass and reduction of fracture risk.

REFERENCES

[1] LeBlanc et al., J Musculoskelet Neuronal Interact 2007. [2] Lang et al., J Bone Miner Res 2006. [3] Payne et al., Am J Phys Med Rehabil 2007. [4] Nyman et al., Bone 2004. [5] Carter et al., J Biomech 1987. [6] Lang et al., J Bone Miner Res 2004.

ACKNOWLEDGMENTS

This work was sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152.

Investigation of Photonic Lattice Based Gallium-Nitride Light Emitters

Project Investigator:

Xiaomin Jin
Department of Electrical Engineering
California Polytechnic State University
San Luis Obispo, CA

Investigation of Photonic Lattice Based Gallium-Nitride Light Emitters

Xiaomin Jin
Electrical Engineering Department

1. **Progress Overview**
2. **Publications Under Support of the Project**
3. **Proposals Submitted Related to the Result of the Project**
4. **Project Results 1: Design of GaN LDs and Related Simulation**
5. **Project Results 2: Top Polymer Micro-Gratings Design to Improve GaN LEDs Light Transmission**
6. **Project Results 3: Design Simulation of Top ITO Nano-Gratings to Improve Light Transmission for Gallium Nitride LEDs**
7. **Project Results 4: Design of GaN Bottom Reflection Gratings on GaN-based Light-emitting Diodes**
8. **Project Results 5: International Research and Educational Collaboration on GaN Emitters**
9. **Project Results 6: Study of Photonic Lattices for Solar Cells**
10. **Project Results 7: Noise Study of Integrated Injection-locking Lasers**
11. **Current Work**

1. Progress Overview

Reeently, many efforts were made on the researeh of Gallium-Nitride (GaN)-based optoeleetronie semieonduetor deviees, due to their vast promising applieations, sueh as solid state light sourees, and ultraviolet light emitters for high-temperature electronics. In some applications, they beeome even irreplaeeable. However, GaN-based semieonduetors have totally different optical and elcctrical properties when compared to other materials. Researehers can make light emitters, sueh as laser diode/ light emitting diode (LD/LED) out of GaN based semiconductors. But the mechanism how they work still is not fully understood. This work reeovers some fundamental issues of GaN-based LED/LDs ineluding the study of deviee surfaee struetures that enhanee light extraction and investigation on the effeets that influenee the power transition efficiency. We investigate applieation of photonie strueture (photonie lattiee or photonie erystal) in design of GaN deviees. Furthermore, we evaluate and eompare the confinement factor of the gain models in various device struetures, and optimize the anti-guide layer design. Third, we obtain the eigen-funetions and eigen energies in quantum wells (QWs) for our proposed and optimized structure to design high power Lasers. During the projeet, we investigate the effeet of the free-earrier theory on optieal gain speetra, and compare GaN bulk material and QW material for the gain properties. Finally, we define some design rules for GaN-based opto-eleetronie deviees and their underlining physies. In addition, we also did initial simulation of grating design on solar eells for effieieney improvement and finished our pervious C3RP 2005 injeetion loeking work.

During the duration of this researeh program, we eontinue and expand the existing eollaboration between Prof. Xiaomin Jin at California Polyteehnie State University (Cal Poly, USA) and Prof. Bei Zhang and GuoYi Zhang's group in the Researeh Center of Wide-Gap Semieonduetors at Peking University (PKU, China). We eollaborate on simulation, fabrieation, and eharaeterization of photonie lattiee-based GaN epitaxial material and optoeleetronie deviees. The Cal Poly group performs simulation and theoretieal study, proposes novel struetures, and eharaeterizes the GaN-based photonie-lattiee struetures. The PKU group performs deviee growth and fabrieation. The final goal is to optimize the deviee design and yield high performanee GaN LDs and LEDs.

2. Publications Under Support of the Project:

Journal Papers

1) X. Jin, B. Zhang, T. Dai, W. Wei, X. N. Kang, G.Y. Zhang, S. Trieu, and F. Wang, "Optimization of Top Polymer Gratings to Improve GaN LEDs Light Transmission", OSA Journal:Chinese Opties letters (Focus Issue Nano Photics), vol.6, no.10, pp.788-790, 2008.
2) X. Jin, B. Y. Tarng, and S. L. Chuang, "Relative Intensity Noise Study in the Injeetion-loeked Integrated Eleetroabsorption Modulator-Lasers", Solid State Elect., vol. 53, pp. 95-101, 2009. (Sponsored by award # ONR N00014-05-1-0855 C3RP 2005, finish the paper during ONR# N00014-07-1-1152)
3) X. Jin, B. Zhang, F. Wang, J. Fliekinger, S. Jobe, T. Dai, G.Y. Zhang, "International Engineering Researeh and Edueational Aetivity on GaN Lasers and LEDs", International Journal of Engineering Researeh and Innovation (IJERI), Vol. 1, No. 1, Spring/Summer 2009. (aeeepted)

Conference Papers

4) Simeon Trieu, Xiaomin Jin, Bei Zhang, Tao Dai, Wei Wei, Chang Xiong, Xiang-Ning Kang, and Guo-Yi Zhang, "Study of Top and Bottom Photonie Gratings on Gallium Nitride Light-emitting-diodes." The Ninth International Conference on Solid State

Lighting, SPIE Symposium on Optical Engineering + Applications, August 2-4th, 2009, San Diego, California, USA. (Accepted)

5) Xiaomin Jin, Sean Jobe, Simeon Trieu, Benafsh Husain, Jason Flickinger, Tao Dai, Bei Zhang, Xiang-Ning Kang, and GuoYi Zhang, "Mode Pattern Analysis of Gallium Nitride-based Laser diodes", The 3rd International Symposium on Photoelectronic Detection and Imaging (ISPDI 2009), Beijing, China, June 17 to 19, 2009. (Accepted)
6) Xiaomin Jin, Xiao Hua Yu, Fei Wang, Bei Zhang, and Guoyi Zhang, "Educational/Research Collaboration on Gallium-Nitride (GaN) Based Light Emitter between Cal Poly, CSULB, and PKU (China)", the 12th CSU Regional Symposium on University Teaching, California Polytechnic State University, San Luis Obispo, May 2nd, 2009. (Presentation only)
7) X. Jin, S. Trieu, Fei Wang, B. Zhang, T. Dai, X. N. Kang, and G. Y. Zhang, "Design Simulation of Top ITO Gratings to Improve Light Transmission for Gallium Nitride LEDs", 2009 Sixth International Conference on Information Technology: New Generations, ITNG2009, April 27-29, 2009, Las Vegas, Nevada, USA.
8) Simeon Trieu, Xiaomin Jin, Bei Zhang, Tao Dai, Kui Bao, Xiang-Ning Kang and Guo-Yi Zhang, "Light Extraction Improvement of GaN-based Light-emitting Diodes using Patterned Undoped GaN Bottom Reflection Gratings", *the SPIE International Symposium on* Integrated Optoelectronic Devices 2009, *SPIE Photonic West 2009*, San Jose, CA USA 24-29, January 2009.
9) Xiaomin Jin, Bei Zhang, Fei Wang, Jason Flickinger, Sean Jobe, Tao Dai, Guoyi Zhang, "International Engineering Research and Educational Activity on GaN Lasers and LEDs" I*nternational Association of Journals and Conferences (IAJC)-International Conference, International Journal of Modern Engineering (IJME) IAJC-IJME 2008*, November 18-22, 2008, Nashville, Tennessee.
10) Xiaomin Jin and Simeon Trieu, "Improvement of Light Transmission using Photonic Lattices for Solar Cells," *OSA Topical meeting, Solar Energy: New Materials and Nanostructured Devices for High Efficiency*, June 24–25, 2008, Stanford University, Stanford, California, USA.

3. Proposals Submitted Related to the Result of the Project:

1) **NSF OISE - IRES 2009,** "International: Enginecring Research and Educational Collaboration on Gallium-Nitride-based Light Emitting Devices", by Xiaomin Jin and Xiao-Hua Yu, Department of Electrical Engineering, California Polytechnic State University, San Luis Obispo, CA 93407, $99,598, submitted on Feb 2009, **pending**.

2) **NSF 08-603 OISE – EAPSI 2009,** "EAPSI: Light Extraction Improvement of GaN-based Light Emitting Diodes", by Simeon Trieu, Graduate student, Department of Electrical Engineering, California Polytechnic State University, San Luis Obispo, CA 93407, sponsored by this project. He also wins the award by working on the project, submitted on Nov 2008, **awarded.**

3) **California Institute for Energy and Environment (CIEE), California Energy Efficiency Strategic Planning RFQ# CP1-007-08:** " General LED Solid-state Lighting Implementation", by Xiaomin Jin and Xiao-Hua Yu, Department of Electrical Engineering, California Polytechnic State University, San Luis Obispo, CA 93407, $14,900, sumbitted on Nov, 2008, **not funded.**

4. Project Results 1: Design of GaN LDs and Related Simulation

Related paper: Xiaomin Jin, Sean Jobe, Simeon Trieu, Benafsh Husain, Jason Flickinger, Tao Dai, Bei Zhang, Xiang-Ning Kang, and GuoYi Zhang, "Mode Pattern Analysis of Gallium Nitride-based Laser diodes", The 3rd International Symposium on Photoelectronic Detection and Imaging (ISPDI 2009), Beijing, China, June 17 to 19, 2009. (Accepted)

Mode Pattern Analysis of Gallium Nitride-based Laser Diodes

Xiaomin Jin*[a], Sean Jobe[a], Simeon Trieu[a], Benafsh Husain[a], Jason Flickinger[a], Bei Zhang[b], Tao Dai[b], Xiang-Ning Kang[b], and Guo-Yi Zhang[b]
[a]Electrical Engineering Department, 1 Grand Avenue, California Polytechnic State University, San Luis Obispo, CA, USA, 93407-9000;
[b]School of Physics and State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, Peking University, Beijing, China, 100871

ABSTRACT

In this paper, we present an analysis of gallium nitride (GaN) quantum-well (QW) laser diode (LD) by numerical simulation. Here we discuss three aspects that are crucial to our analysis. First, the transverse mode pattern is studied, and our current GaN diode laser structure is discussed with optical waveguide mode analysis. Then we compare the QW design of the laser and maximize laser modal gain. Finally, we report the influence of the electron block (e-block) layer on lasing performance of our design.

Keywords: Gallium Nitride, semiconductor laser, transverse modes

1. INTRODUCTION

In the visible color spectrum, there are three primary colors: red, green, and blue, also referred to as RGB. Through the combination of these three colors, all the other colors in the spectrum can be created. For example, the combination of red and green would produce the color yellow. Current technology has found efficient ways at creating both red and green light using semiconductor technology, but the creation of blue light has not shared this immediate success. Blue light has remained the hardest light to produce efficiently from semiconductor technologies. This has been the main road block preventing highly efficient laser diodes (LDs) from replacing our current lighting technologies.

Blue violet light is close to the shortest wavelength of light in the visible spectrum. At the wavelength of 445nm, it still remains the hardest light to produce using semiconductors, because the bandgap properties of the materials needed to produce such a wavelength in semiconductors that are hard to find. In recent years, many strides have been made in studying and researching Gallium Nitrate or Gallium Nitride (GaN) as a material to yield blue light. The most progress and significant strides have been made by Dr. Nakamura, who dedicated much of his research to the chemical growth of GaN compounds. [1] Before Nakamura found a solution for growing GaN semiconductors, much of the focus for blue light semiconductors was spent on II-VI materials (where II and VI represent the group number on the Periodic Table of Elements). GaN is a III-V material and is a much harder substance to deal with because of its higher lattice defects in comparison to II-VI materials. However with Nakamura's perilous efforts, he eventually found a technique to deposit GaN compounds and created the first GaN semiconductor. Since Nakamura has shown that GaN LED semiconductors are possible in 1995, fellow scientists across the world have worked hard to further the development and efficiency of blue light semiconductors [1].

GaN is one of the most promising materials for use in the blue and ultraviolet wavelength region. Since the room-temperature (RT) continuous-wave (CW) operation of GaN-based lasers were reported by Nakamura et al. 1996, [2] one of the important targets on the GaN LD development is to extend the operating lifetime by reducing the operation current or current density. It was pointed out that the threshold current density of GaN lasers is intrinsically higher than that of GaAs. This is because the hole effective mass of wurtzite (WZ) GaN is much heavier than that of conventional zincblende materials such as GaAs.[3] To decrease the threshold current density in GaN lasers, several approaches have been proposed. Most of the papers discuss the threshold improvement through optical gain or Quantum-well (QW) analysis. [4]-[6] The beginnings of III-V laser diode devices were very elementary in their design compared to current

*xjin@calpoly.edu; phone 1 805-756-7046; fax 1 805-756-1458; www.ee.calpoly.edu

International Symposium on Photoelectronic Detection and Imaging 2009: Laser Sensing and Imaging, edited by Farzin Amzajerdian, Chun-qing Gao, Tian-yu Xie, Proc. of SPIE Vol. 7382, 73820O
© 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.834103

designs. The GaN diodes consisted of a substrate, cladding, core, and active region. Current designs have more complex structures such as a super lattice (SL), quantum wells, and e-blocks. Using these new technologies, devices with lower laser thresholds can be created thus leading to more efficient devices. In this work, we discuss the threshold reduction on GaN QW lasers in three aspects: 1) optical waveguide design and mode analysis; 2) QW design and modal gain analysis; 3) electron- block (e-black) layer design for optimization.

2. NUMERICAL MODELING

LaserMOD is an integrated software package developed by RSOFT for the design and simulation of semiconductor lasers and active photonic structures. [7] It is a fully integrated platform with a user friendly parametric CAD interface, nonuniform Delauney mesh generator, material libraries, gain and mode calculation utilities, simulation engine, standard and custom plot generation utilities, and versatile graphical viewing utilities. In short, the LaserMOD program provides the user with an immense environment in which semiconductor devices can be modeled and tested based on their user specified properties.

2.1 Optical Confinement Factor (OCF)

The optical confinement factor (OCF) of a chosen mode is defined as the ratio of the energy of the chosen mode located in the active region to the total guided energy of all the modes. [8] A higher OCF for a certain mode indicates there is more energy in the active region for that mode. High energy in the active region is usually a sign that lasing is occurring. Thus, whichever mode has the highest OCF will be the lasing mode. The optical modes are determined from the solution of the Helmholtz equation via simultaneous as below, then the OCF can be calculated.

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\delta^2}{\delta y^2}+k_0(\varepsilon(x,y)-n_{eff,m}^2)\right)E_m(x,y,z)=0$$

where $E_m(x,y,z)=E_m(x,y)\exp(ik_0 n_{eff,m}z)$, k_0 is the free-space wave vector, and $\varepsilon(x,y)$ is the complex dielectric constant profile of the multilayer structure. The eigenvalues are given by the effective index $n_{eff,m}$. The frequency $k_0=\omega_0/c$ of the mode is solved and is set to correspond to the quantum well band gap energy. The OCF can be thought of as the fraction of the energy of mode that is located at the active region. [9]

2.2 Mathematical Model – 8x8 Kronig-Penney (KP) Model

The LaserMOD simulation software uses the Kronig-Penney (K • P) model to model the bandgap relationships of the different materials. The K • P model is based upon the splitting of allowed electron energies as the interatomic distance between atoms decreases to form a crystal. The K • P method involves quantum mechanics and a solution to Schrödinger's wave equation. [10] For the one dimensional crystal structure, a periodic well function can be used to represent the crystal lattice structure. This is because when the potential functions of atoms are brought close together, the net potential function of the overlapping regions is similar to a periodic function. Using the specified characteristics of a periodic function for the boundary conditions for Schrödinger's equation, a plot of the energy E as a function of wave number k can be generated, which describes the valence band, conduction band, and the allowed energy bands. However for two dimensional calculations, matrices become a necessary addition to efficiently solve Schrödinger's wave equation. A popular method for solving E vs. k diagrams is using 8x8 matrices. Using special matrix rules and "tricks" (i.e. Helmholtz equation) these calculations can be solved with a computer program. Many programs have been written that can solve these types of mathematical problems. We chose to use RSOFT's LaserMOD, which is a program that has a CAD interface which can easily and quickly calculate all the necessary properties of semiconductors.

2.3 The 2D nitride-based laser model

A 2D nitride-based laser model is developed, which is shown in Fig.1. The simulated laser structure has five quantum-wells (QWs). A similar device was fabricated at Peking University, China. One of the project goals is to identify design flaws from the current laser design and to reduce the laser threshold. To avoid the meshing difficulties of the finite-element method, the classical Ritz simultaneous iteration is combined with an additional optimization to analyze closed

arbitrary dieleetrie waveguides. [7] LaserMOD determines the eharge distribution using Sehrodinger equation. The laser simulation is based on 8x8 **K • P** band strueture ealeulation and photon rate equation. The material parameters based on reeent literature values and some experimental data were used. The detailed laser strueture is listed in Table.1.

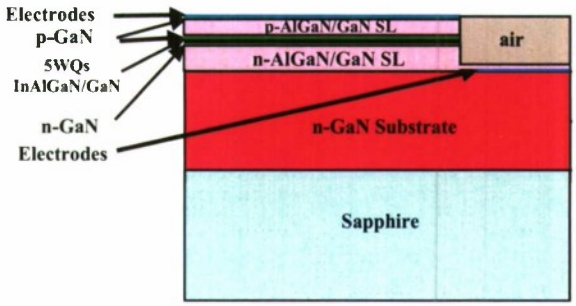


Fig.1 2D GaN laser model using Rsoft LaserMOD.

Table. 1. Laser diode layer strueture and parameters.

Layer	Thickness (nm)	Refractive Index (n)
p-GaN (contact)	50	2.55
p-$Al_{0.12}Ga_{0.88}N$/GaN (p-SL cladding)	500	2.53
p-GaN (p-core)	100	2.55
p-$Al_{0.35}GA_{0.65}N$ (e-block)	20	2.42
n-GaN	15	2.55
$In_{0.1}GA_{0.9}N$/GaN (5QWs)	67	2.685/2.55
n-GaN (n-core)	100	2.55
n-$Al_{0.12}Ga_{0.88}N$/GaN (n-SL cladding)	800	2.53
n-GaN (Substrate or buffer)	4000	2.55
Sapphire (Oxide)	4000	1.77

3. SIMULATION RESULTS

3.1 Lasing mode and ghost modes

The first twenty transverse modes in the above GaN LD strueture are ealeulated. Fig. 2 shows several optieal mode patterns of GaN LD. When sapphire is used as a substrate for GaN lasers, the dislocation density in the material is usually very high. To eontrol the defeet number or reduee eraeks, an n-GaN substrate (or buffer) layer (several mieron thiek) is deposited on the sapphire substrate before growing the eladding layer. This layer has higher refraetive index eompared to the n-AlGaN/GaN supper-lattiee (SL) eladding layer. Beeause of insuffieient eladding thiekness, QW waveguide and n-GaN substrate waveguide are eoupled and the GaN lasers have multi-waveguide struetures, whieh support strong substrate modes, also ealled the "ghost-mode" phenomena. [11] Therefore, the fundamental mode of the multi-layer waveguide is usually "ghost mode". The higher order mode of this multi-layer waveguide is usually the lasing mode. The optieal eonfinement faetor is also very low even for the lasing mode, about several pereents. This leads to lasing problem of GaN lasers. Different order modes have different optieal eonfinement faetor. The most strongest-eonfined mode in the multilayer waveguide strueture ean be the lasing mode.

The optical modes of the 2D simulation are shown in Fig. 2. For modes 0th through 7th the modes look very similar to the 1D results, [12] with additional modal energy being layered above and below adjacent energies. The 8th mode is the lasing mode and shows the optical energy confined inside the active region. Modes 9th and 10th modes continue the same trend of optical energy layering as modes 1 through 7. However, something very interesting happens in mode 11th. For Mode 11th to 20th, the modal energy splits across the x-axis horizontally. The next few modes follow this new horizontal trend. The energy split at x-direction is a characteristic that we would not be able to notice without doing a 2D model and gives insight into the optical energy interaction along the x-axis. We find that as the mode increases, the optical energies are no longer contained along the vertical boundaries, the energies split and are shared across the x-axis.

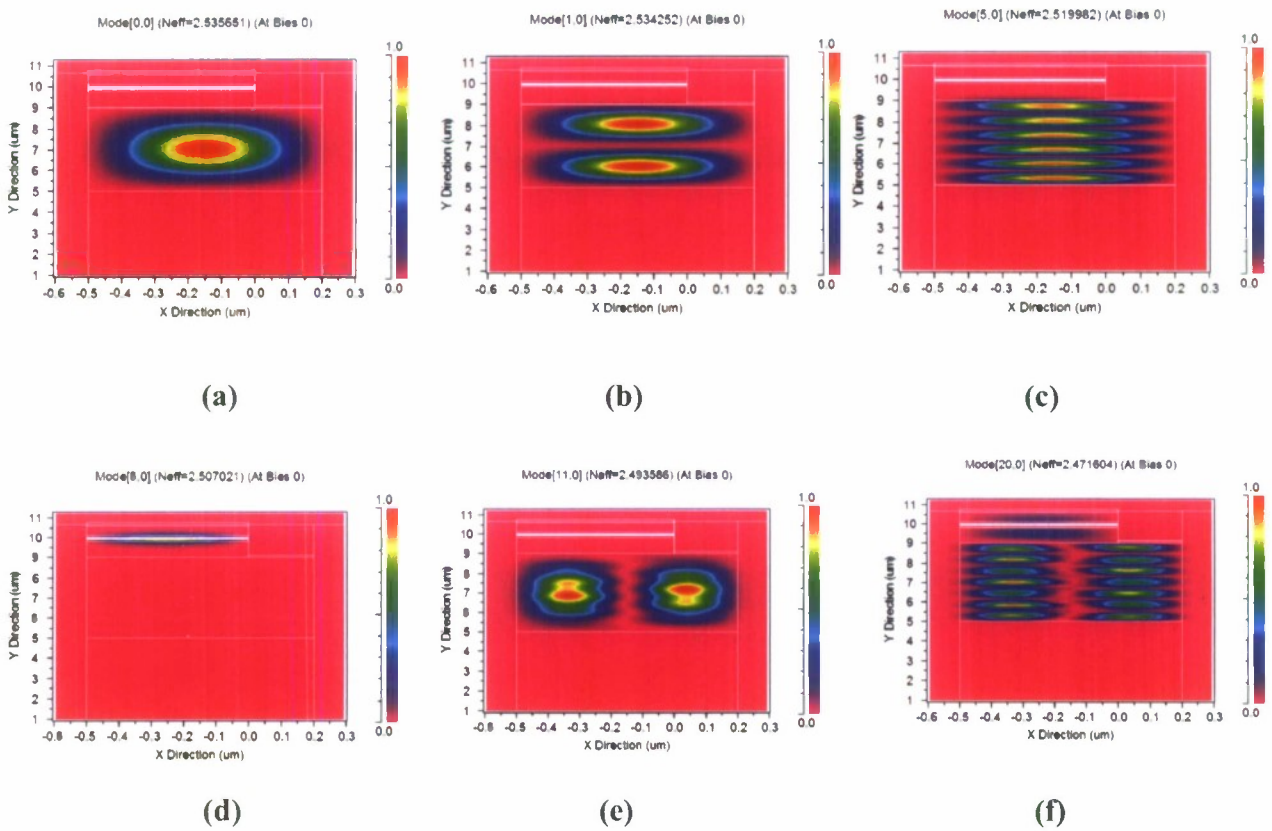


Fig. 2 Several ghost modes and lasing mode distributions: (a) 0th order mode, (b) 1st order mode, (c) 5th order mode (d) 8th order mode or the lasing mode, (e) 11th order mode, and (f) 20th order mode.

From our calculation, the optical confinement factor for the zero-order mode is only 0.0000067%. The 8th mode has the greatest overlap of optical field with the quantum well. Its optical confinement factor is 8.67%, which agrees with reported data. [13] This indicates that our LDs lasing in the 8th order transverse mode. The other modes are substrate modes or ghost modes as shown in Fig. 2. Strong substrate modes compete with the lasing modes in this multi-waveguide structure. Optimizing the OCF for different layer thicknesses is very important in lowering the lasing threshold of the GaN LD. An increase in the OCF means that more light is being confined in the lasing mode. With more light focused in the lasing mode, less current will be required for the LD to achieve lasing. In here, we demonstrate how the different optical modes affect the Light vs. Current curves of the GaN LD. By creating a new laser design using our optimized thicknesses, we show that it produces lower lasing thresholds, which brings us one step closer to the ultimate goal in meeting the expectations of using laser diodes as future light sources.

The OCF is calculated in the 2D design for GaN substrate layer thicknesses from 0-5μm. Fig. 3 shows the values of the OCFs simulated and for what mode they occurred. The lasing mode migrates as the GaN thickness increases. For every increase in the GaN substrate by 0.5μm the lasing mode jumps up a mode. For example, at a GaN thickness of 2.6μm the lasing mode is 5th mode, but increasing the GaN thickness to 3.1μm, the lasing mode migrates to the 6th mode. Each mode is evenly spaced by the constant of 0.5μm. These results are in agreement with our previous 1D GaN substrate simulation results. [12] Using the original design from Table 1, the simulated OCF for the lasing mode (8th mode) was 8.6677%, however in Fig. 3, the OCF can still be increased if the GaN thickness is increased a little. Following along the line that represents the 8th mode, a thickness of 4.2μm yields an OCF of 8.6962% which is a 0.02% increase over the original design. Only a small increase in the OCF was possible for the GaN substrate layer showing that the original layer thickness was very close to being an optimal design.

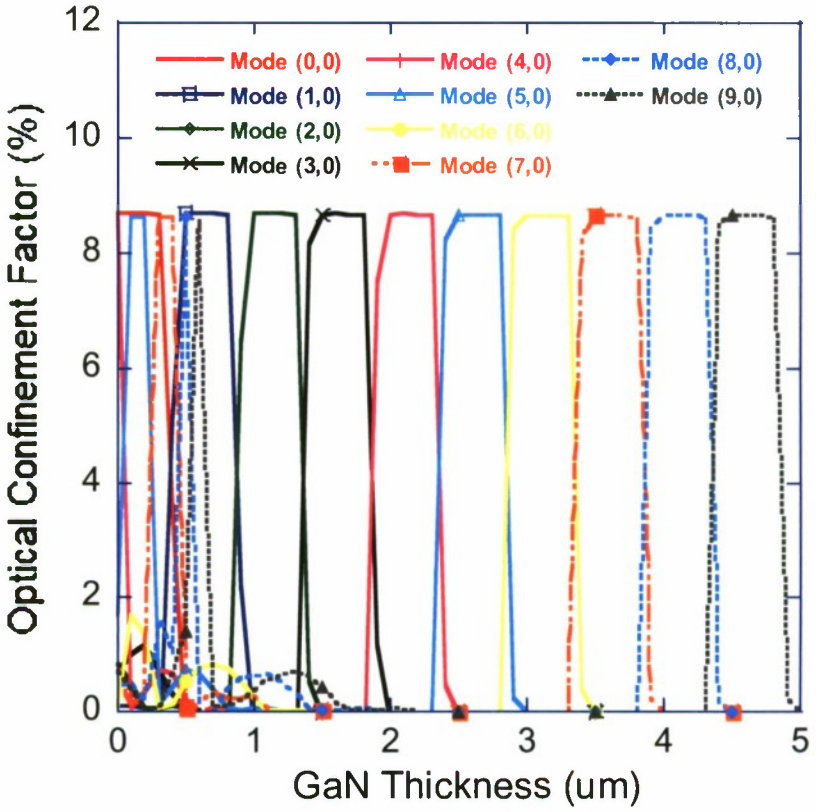


Fig. 3. The OCF vs. GaN substrate thickness (2D)

It is expected that increasing the cladding layer thickness will lengthen the distance that the light has to tunnel through in order to get into the substrate waveguide. This will prevent optical leakage of the mode into the absorbing, high index p-contact layer. This however will increase the impedance as it would also lengthen the distance the current would have to tunnel through and therefore increase our threshold current. By adjusting the thickness in the n-SL, it is of interest to find a smaller thickness that still retains a high OCF and a thickness that prevents mode and current leakage. It is also important to be aware that creating such thin coats of the n-SL still remains a real world problem and although a thin coat of n-SL may lead to an optimized simulation, creating such a lasing device may be difficult. By increasing the cladding size, we are looking to find how much a thickened cladding actually affects the OCF. We expect a thickening of the n-SL will prevent mode leakage from reaching the substrate and thus help prevent ghost modes therefore leading to a higher OCF.

The results of the 2D simulation with varying n-SL thickness (in Fig. 4 (a)) are similar to those Hatakoshi calculated. [12][14] As the cladding thickness is increased, the 8th mode (lasing mode) gain begins to confine more light. As the

lasing mode increases in its OCF, the other modes begin to lose the light they had originally been able to confine. Thus as the cladding thickness increases all the optical energy that is available becomes confined only in the 8th mode. This is shown in Figure 4 (a). This result shows that the anti-guide-like or ghost mode behavior can be suppressed by increasing the cladding layer thickness. A thick cladding layer reduces the effect of the outer contact layers, but still is conducive to high-order modes. At the n-SL thickness of the original design (0.8μm), the OCF was 8.7128% and reaches a peak plateau on the graph at 8.7241% at a thickness of 1.4μm. The 0.01% increase in OCF requires almost doubling the thickness. Thus the best optimization for this design is simply leaving it alone. The original design yields an optimal OCF for its relative thickness. The p-SL cladding provides the same waveguide like confinement of the light as the n-SL cladding, shown in Fig. 4 (b), but it differs in that it will not be absorbing electrons like the p-SL will be. By adjusting the thickness of the p-SL, we hope to find a thickness that prevents mode leakage into the p-contact layer, yet still maintains low impedance. The adjustment of the p-SL cladding layer shows a different increase in the OCF than the n-SL plot. It is interesting to point out however that the thickening of the p-SL lattice above 0.1μm has a detrimental effect on the maximum OCF, decreasing it from a maximum of 8.7% to 8.4%. (It is important to note that it is the OCF at a different mode). Thus we find our optimal thickness for the p-SL that yields the best OCF yet remains fairly thin is at 0.1nm.

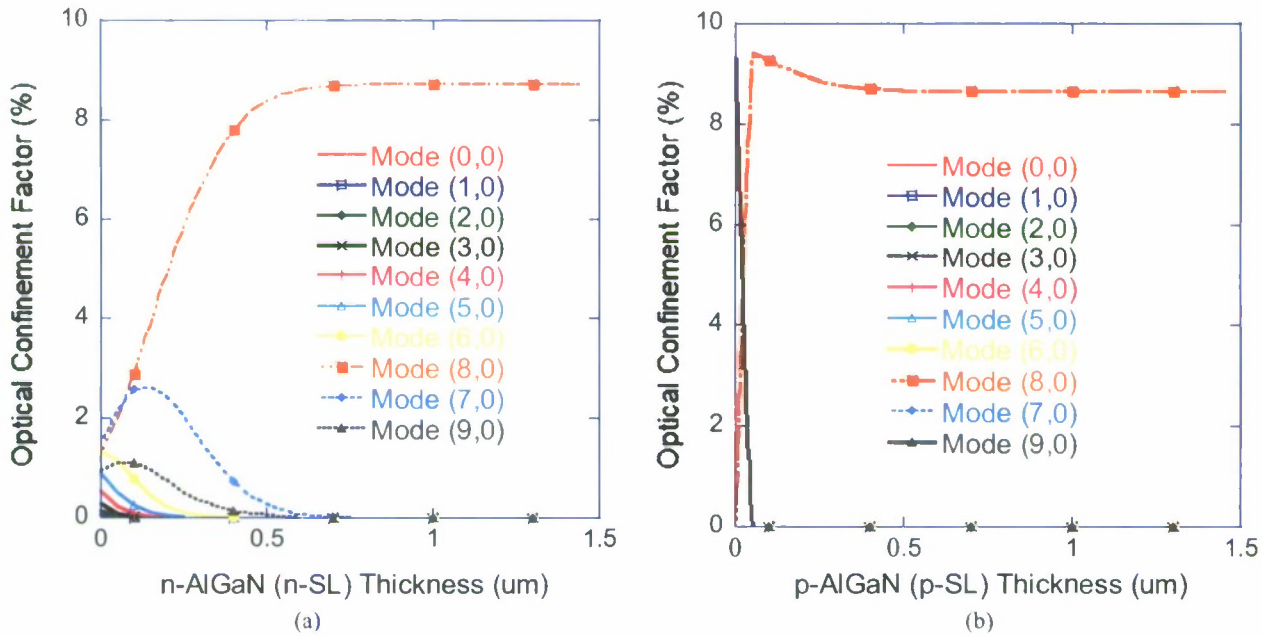


Fig. 4. Optical confinement vs. (a) n-SL and (b) p-SL thickness

The GaN core layer in the 2D simulation has a refractive index greater than the surrounding cladding layers to reflect the light back towards the active region. A thicker core layer can lead to more optical losses due to the large area the light has to travel. Thus choosing an optimal core layer thickness is important in creating the most efficient laser diode. The OCF is plotted in Fig. 5 for n-GaN layer thicknesses of 0-1μm. The OCF peaks at 8.7027% with an n-GaN thickness of 0.075μm. This is a 0.224% increase in the OCF from 8.6762% when a 0.1μm layer is used. Thus by using a thinner layer of n-GaN core layer, a better optical confinement can be achieved. Fig. 5 shows a constant decrease in the OCF as the n-GaN thickness is increased and the lasing mode migrates to lower modes. The same simulation was run for the p-GaN core layer and the results of the OCF vs. p-GaN thickness are also plotted in Fig.5. The maximum OCF is achieved for small thin layers of p-GaN. The OCF peaks with 8.8664% at a thickness of 0.05μm. This is a 0.2% increase over the original design when the p-GaN was 0.1μm thick and had an OCF of 8.6762%. Similar to the n-GaN core layer, a thinner layer creates better optical confinement.

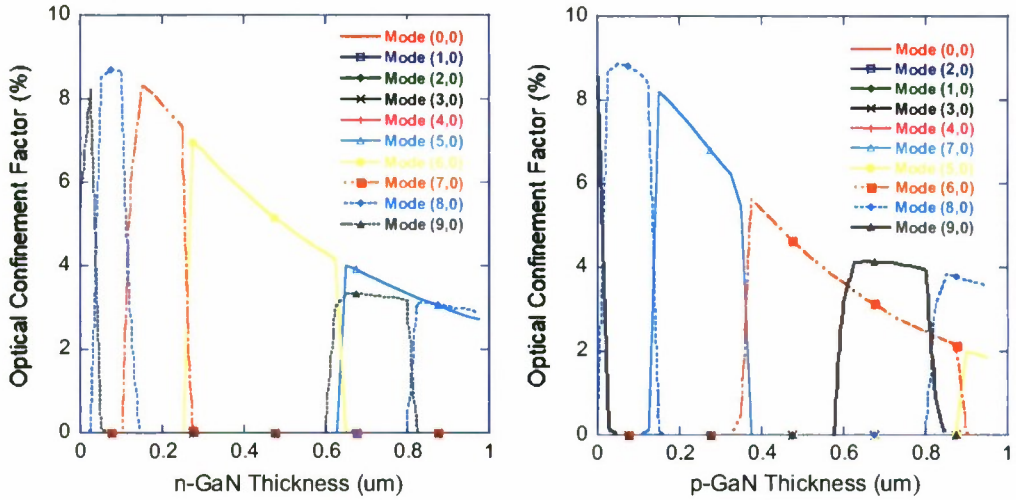


Fig. 5. Optical confinement factor vs. (a) n-GaN and (b) p-GaN layer thickness

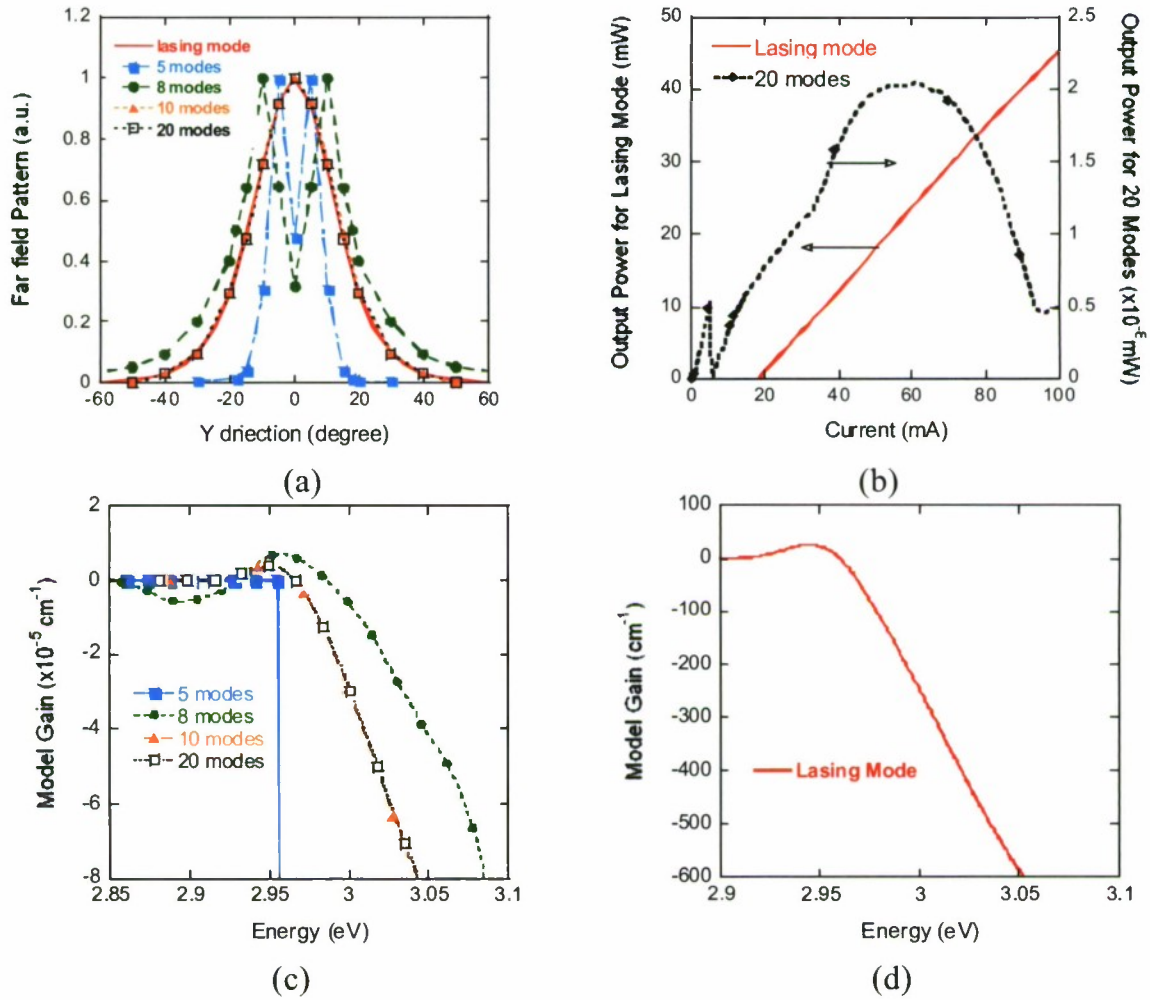


Fig. 6 The GaN laser simulation for (a) far field pattern, (b) light output versus current, (c) model gain of all the transverse mode, and (d) model gain of lasing mode only.

To prove the optical mode design can improve CW GaN LD operation, the threshold, optical power, quality of far-filed, and optical gain should be evaluated, which are the most important issues and represents the laser performance. For examples, in 2002, Tojyo et al reported a kink free output of over 100mW. In 2005 Schwarz et al reported near- and far-field study of GaN LDs. [15] In 2007, Laino et al reported results on substrate mode study. And in 2006, [16] Witzigmann et al reported optical gain analysis. [17] Here we calculate the far field pattern, model gain, optical output, and threshold with only lasing mode (the 8th mode), and all the transverse modes to demonstrate the ghost mode effects, shown in Fig. 6. For the far field simulation, we normalized the optical pattern to the peak value for comparison. In the figures, lasing mode represents the 8th order mode only simulation, "x modes" means "the first x transverse modes". The "8 modes" means modes 0 to 7. For the case of "5 modes" or "8 modes", there is no lasing mode included in the calculation and LD does not reach threshold for lasing. For the "8 modes", "10 modes", and "20 mode" cases, little modal gain is obtained from laser, which is about 10^{-5} cm^{-1}. Therefore the light output is only about 10^{-8} mW. If we calculate those results with only 8th order lasing mode, the laser can reach a peak gain of about 25cm^{-1}, lasing output is about 40mW at 80mA, and threshold is about 19mA. According to our optical field analyses, the strong ghost-modes compete with lasing-mode in GaN Laser, which can prevent the laser from lasing. For the worst case, if the waveguide structure supports strong ghost-modes, the GaN laser would not operate even though the active region quantum well is perfect. Design of the optical waveguide for GaN laser system is very important to achieve efficient lasing condition. In this work, we find that the n-GaN substrate thickness is the major factor influencing the transverse mode pattern. In a related work, [12] we optimized the optical waveguide structure, which limits the ghost modes, and maximizes confinement factor of the laser structure, in order to reduce lasing threshold.

3.2 The QW analysis and model gain simulation

Quantum wells are designed to trap electrons in a 2-D environment. The wells are designed to have a particular bandgap energy related to the wavelength of light emitted by the laser diode. The quantum well allows the electrons to gather more densely in the well than they would elsewhere. It is easy to think of it as though the well is drawing electrons towards it then trapping them. (In fact, it isn't so much that the wells are drawing electrons to them and away from the other parts of the semiconductor; instead electrons keep falling into the well, and the well becomes saturated and full that it seems as though the well is attracting electrons.) This packs more electrons in the active region and allows more electrons to jump the bandgap thereby releasing a photon. The size of the well must be on order of the wavelength of light hoping to be produced. Each well is created by creating a thin layer (the well) and surrounding it by thicker layers of a different material. Thus for a multi quantum well (MQW) design, there would be alternating layers of bulk and active layers. The well layer is made from the normal active layer material with a lower refractive index then the surrounding bulk layers. This in a sense is like having many tiny optical cavities that interact with photons at the quantum level. It can be shown that the more quantum wells in a semiconductor, the larger the threshold current becomes. This is because when there are more wells for the electrons to fall into, a higher current is necessary to provide the quantum wells with enough electrons to maintain saturation which in turn is the foundation of lasing. [18]

Quantum wells have many advantages in the design a laser diode structure. First, by using QWs, one has the freedom to design the transition energies, which ultimately determines the wavelength of light that leaves the active region in the form of spontaneous emission. Second, since QWs have a small volume, the free carrier concentration in the QW is high and at high free carrier concentrations non-radiative deep-level transitions are less likely thus yielding a high radiative efficiency. Third, due to the small size of the quantum wells, the carrier density required to achieve population inversion (i.e., high carrier density in the conduction band, compared with the valence band) is small and therefore the threshold current density of QW structures is low. And finally, the surface recombination is less likely, which making surface recombination less important in the study of QWs. [18]

The structure we simulated here has a wide contact with a uniform current injection. The number of QWs, the QW width, and composition fluctuations play central roles in the optimization of GaN light-emitting diodes (LEDs) and LDs. [4]-[6] [17] Our simulation here considers number of QWs and composition. We choose lasing mode only simulation, without considering optical mode variation for simplification. The band structure of QWs is computed using **k • p** method, which includes coupling effects for the heavy-hole, light-hole, and the crystal-field split hole dispersion. The QW is $In_xGa_{1-x}N/GaN$. The width is 5nm/7nm (well/barrier), and x=0.1. From Fig. 7(a), the peak modal gain increases from 17 cm^{-1} for 1QW to 3.7 cm^{-1} for 3QWs. Then it reduces to 29.1cm^{-1} and 25.3cm^{-1} for 4 QWs and 5 QWs. The 3 QWs Case provides the highest gain peak. This agrees with the light optical power-current (LI) curve simulation in Fig. 7(b). The carrier densities are inhomogeneous among the quantum wells. [19] The optical gain is generated only in three

QWs on the p-side for our case. The QWs on the n-side act as absorption layers. Therefore, for the LDs with more than three QWs, the peak gain reduces, which results the light output slope (or quantum efficiency) reduction in LI plots. From the threshold and gain simulation, it is important to decrease the number of wells. Single quantum well GaN LD is the best design structure, however it has a lot of fabrication challenges. Besides the above gain simulation, we have varied structural parameters and calculate the threshold currents of lasers. The threshold current increases as QW number increases and decreases with $In_xGa_{1-x}N/GaN$ composition x, as shown in Fig. 8. Higher "In" composition in the QW will obtain higher gain, however it has to be balanced with drawback of the lattice mismatch and deterioration in material quality for a larger "In" fraction. For 5 QW case, the threshold current shows little composition dependence, while x varies from 0.08 to 012. This will provide good threshold stable point with fabrication tolerance.

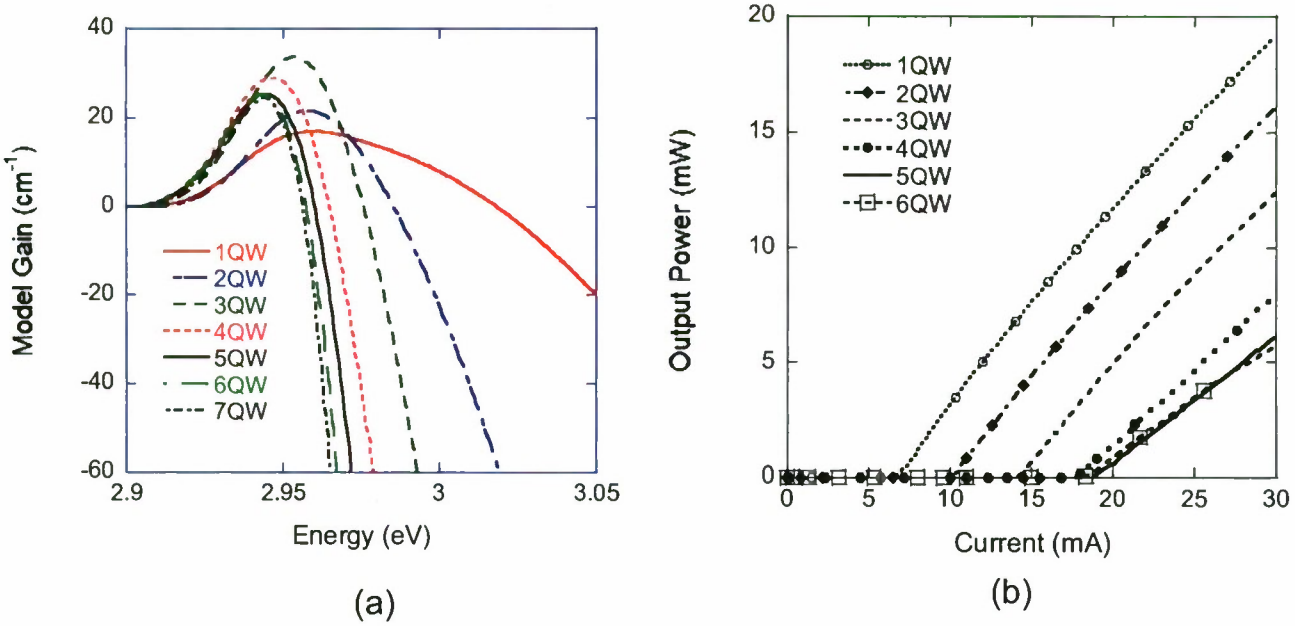


Fig.7 (a) Model gain and (b) LI curve simulation for $In_{0.1}Ga_{0.9}N/GaN$ single-QW and MQWs without e-block.

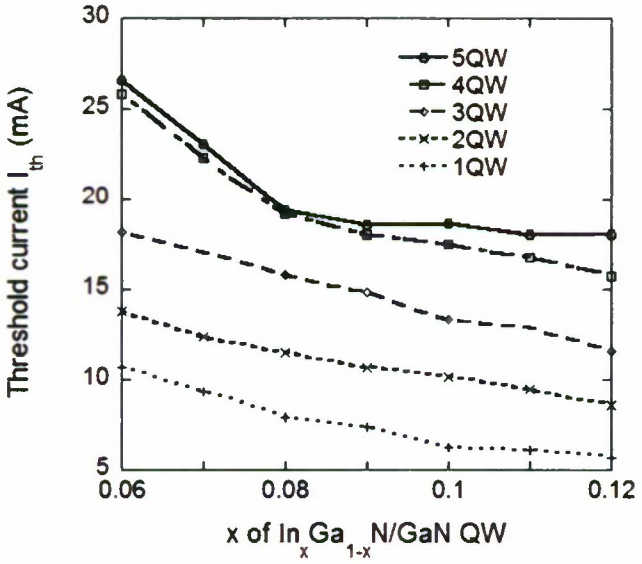


Fig. 8 Threshold current for different QW composition x: $In_xGa_{1-x}N/GaN$ and QW number.

3.3 Electron Block (e-Block) Layer

Electron barrier was studied around late 1980s for GaAs/AlGaAs laser system to improve the electro-optical characteristics/threshold current of QW lasers. [20] To control the thermionic emission and overflow of the carriers from QWs, increasing the barrier height or design a carrier block can effectively improve carrier confinement, and therefore, reduce threshold current for higher emission efficiency. Recently, to improve the GaN threshold and obtain high-power/high-temperature operation, an AlGaN electron-blocking layer was proposed and widely used. [21][22] This barrier is located on the p–side, in the direct vicinity of the active layer of the GaN laser. The electron barrier is usually un-doped to avoid free carrier absorption. In this work, we design a 20nm thick layer of $Al_{0.35}Ga_{0.65}N$ (15nm away) on the top of the QWs active region. The LI simulations are presented in Fig. 9. Compared with Fig. 7(b), the threshold currents are much smaller for the e-block case, and the optical output power is more than doubled. In this design, the e-Block is a p-doped material that has a larger band gap than its neighboring materials. The e-Block is a very thin layer that is added next to the active layer to prevent electrons from leaking into the p-doped side. Electrons that overflow into the p-type side leads to leakage current. Leakage current is detrimental to the operation of the LD because the higher current causes heating and dissipates non-lasing energy creating an inefficient LD. The doped structure contains an AlGaN blocking layer that prevents electrons in the active region from moving into the p-type side. The e-block only blocks electrons and allows the holes to move freely from the p-type side into the active region unaffected. If a high current is injected across the diode then a large carrier concentration results in the active region. This leads to enhanced non-radiative carrier recombination at defects and to an escalation of electron leakage from the quantum wells into the p-side of the diode, despite the AlGaN blocker layer. Thus the e-block is very useful in preventing leakage current up to a certain current, but once the injected current becomes very high, the blocking properties are null and the leakage electrons causes the laser to heat, degrade, and eventually breakdown. Increasing the band gap of the e-block can be an effective method of making it very hard for electrons to leak onto the p-side. However, an increase in the band gap of the e-block often means a higher content of Al and therefore a decrease in the thermal and electrical conductivities. The decrease in the thermal and electrical conductivities counteracts the improvement of the large band gap e-block.

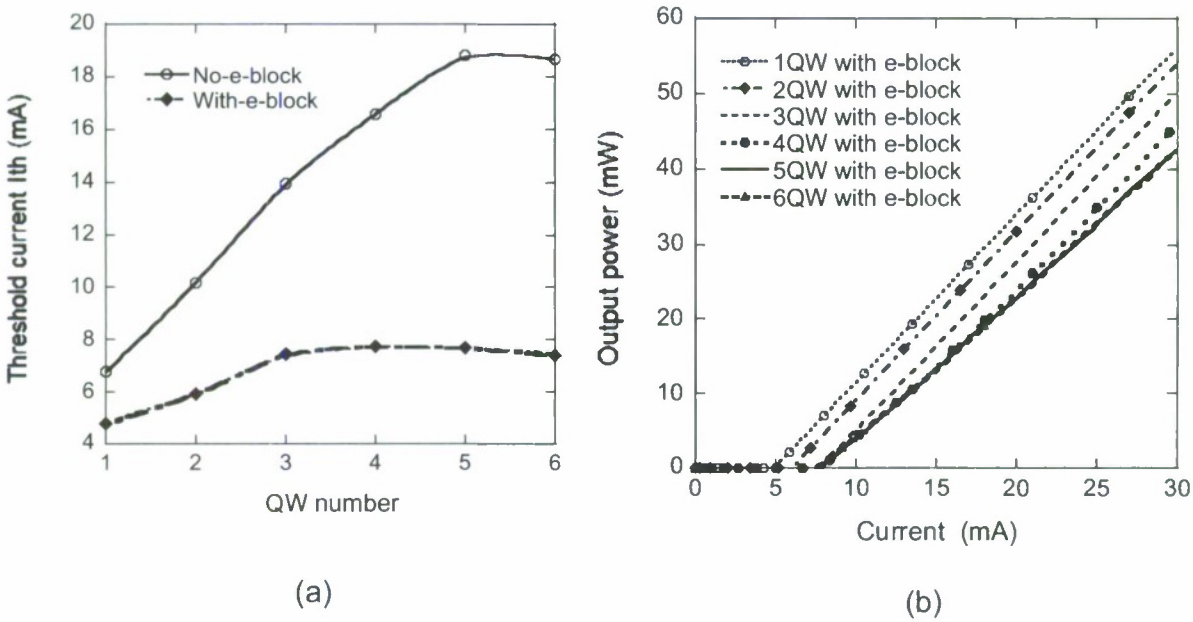


Fig.9 (a) Threshold current with/without e-block and (b) LI curve with e-block.

4. CONCLUSION

GaN laser diode simulation results are presented for the GaN laser design. We discuss the optical substrate modes in optical waveguide. Suppressing substrate mode is very important for the GaN laser. We also present QW modeling results and gain simulation. For our current design, 3QWs will give the best gain performance. Finally, adding e-block

layer on the top of quantum-well active layer will reduce threshold by about 30% to 40%.

ACKNOWLEDGEMENT

This project is supported by Department of the Navy, Office of Naval Research, under Award # ONR 6-N00014-07-1-1152 in 2008, Award # ONR 7-N000140811209 in 2009, USA; "ChunHui" exchange research fellow 2008, Educational Department, China; 973 program-National Basic Research Program of China (2007CB307004); High Technology program (863-2006AA03A113) and National Nature Science Foundation of China (60276032, 60577030 and 60607003).

REFERENCES

[1] Nakamura, S., [The Blue Laser Diode], 2nd ed., Springer, (1997).
[2] Nakamura, S., Senoh, M., Nagahama, S., Iwasa, N., Yamada, T., Matsushita, T., Sugimoto, Y. and Kiyoku, H., "Room-temperature continuous-wave operation of InGaN multi-quantum-well structure laser diodes," Appl. Phys. Lett. 69 (26), 4056-4058, (1996).
[3] Suzuki, M., Uenoyama, T. and Yanase, A., "First-principles calculations of effective-mass parameters of AlN and GaN," Phys. Rev. B 52 (11), 8132-8139, (1995).
[4] Suzuki, M. and Uenoyama, T., "Strain effect on electronic and optical properties of GaN/AlGaN quantum-well lasers," J. Appl. Phys. 80(12), 6868-6874, (1996).
[5] Ohtoshi, T., Niwa, A. and Kuroda, T., "Dependence of optical gain on crystal orientation in wurtzite–GaN strained quantum-well lasers," J. Appl. Phys. 82(4), 1518-1520, (1997).
[6] Domen, K., Horino, K., Kuramata, A. and Tanahashi, T., "Optical gain for wurtzite GaN with anisotropic strain in c plane," Appl. Phys. Lett. 70(8), 987-989, (1997).
[7] LaserMOD v2.0 User Guide, RSOFT Design Group, Inc., New York, (2004).
[8] Wakita, Koichi, [Semiconductor Optical Modulators], Springer, New York City, (1998).
[9] Botez, D., "Analytical approximation of the radiation confinement factor for the TE0 mode of a double heterojunction laser," IEEE J. Quantum Electronics QE-14 (4), 230–232, (1978).
[10] Neamen, Donald A., [Semiconductor Physics and Devices: Basic Principles], 3rd ed., McGraw Hill, New York City, (1992).
[11] Smolyakov, G. A., P. G. Eliseev, and M. Osinski., "Effects of resonant mode coupling on optical Characteristics of InGaN-GaN-AlGaN lasers," IEEE J. Quantum Electronics 41(4), 517-524, (2005).
[12] Jin, X., Zhang, B., Dai, T. and Zhang, G., "Effects of Transverse Mode Coupling and Optical Confinement Factor on Gallium Nitride-Based Laser Diode," The Institute of Physics: Chinese Physics, 17(4), 1274-1278, (2008).
[13] Einfeldt, S., S. Figge, Botteher, T. and Hommel, D., "Coupling of optical modes in GaN-based laser-diodes," Physica Status Solidi(c). 0(7), 2287-2291, (2003).
[14] Hatakoshi, G. I., Onomura, M., Saito, S., Sasanuma, K. and Itaya. K., "Analysis of Device Characteristics for InGaN Semiconductor Lasers." Japanese Journal of Applied Physics 38(16), 1780-1785, (1999).
[15] Schwarz, U.T., Pindl, M., Wegscheider, W., Eichler, C., Scholz, F., Furitsch, M., Leber, A., Miller, S., Lell, A. and Harle, V., "Near-field and far-field dynamics of (Al,In)GaN laser diodes," Appl. Phys. Lett. 86(16), 161112(3pages), (2005).
[16] Laino, V, Roemer, F., Witzigmann, B., Schwarz, U.T., Fischer, H., Feicht, G., Wegscheider, W., Rumbolz, C., Lell, A. and Harle, V., "Substrate Modes of (Al,In)GaN Semiconductor Laser Diodes on SiC and GaN Substrates," IEEE. J. Quantum Electronics 43(1), 16-24, (2007).
[17] Witzigmann, B., Laino, V., Luisier, M., Schwarz, U.T., Fischer, H., Feicht, G., Wegscheider, W., Rumbolz, C., Lell, A. and Harle, V., "Analysis of temperature-dependent optical gain in GaN-InGaN quantum-well structures," IEEE. Photon. Tech. Lett. 18(15), 1600-1602, (2006).
[18] Schubert, E. F. [Light-Emitting Diodes], 2nd ed., Cambridge University Press, (2006).
[19] Domen, K., Soejima, R., Kuramata, A., Horino, K., Kubota, S. and Tanahashi, T., "Interwell inhomogeneity of carrier injection in InGaN/GaN/AlGaN multiquantum well lasers," Appl. Phys. Lett. 73(19), 2775-2777, (1998).
[20] Blood, P., Fletcher, E.D., Woodbridge, K., Heasman, K.C. and Adams, A.R., "Influence of the barriers on the temperature dependence of threshold current in GaAs/AlGaAs quantum well lasers," IEEE J. Quantum Electronics 25(6), 1459-1468, (1989).

[21] Wiedmann, N., J. Schmitz, K. Boueke, N. Herres, J. Wagner, M. Mikulla, R. Poprawe, G. Weimann. , "Band-edge aligned quaternary carrier barriers in InGaAs-AlGaAs high-power diode lasers for improved high-temperature operation," IEEE J. Quantum Electronics 38(1), 67-72, (2002).

[22] Tu, R.-C., Tun, C.-J., Pan, S.-M., Chuo, C.-C., Sheu, J.K., Tsai, C.-E., Wang, T.-C. and Chi, G.-C., "Improvement of near-ultraviolet InGaN-GaN light-emitting diodes with an AlGaN electron-blocking layer grown at low temperature," IEEE Photon. Tech. Lett. 15(1), 1342-1344, (2003).

5. Project Results 2: Top Polymer Micro-Gratings Design to Improve GaN LEDs Light Transmission

Related paper: Xiaomin Jin, Bei Zhang, Tao Dai, Wei Wei, Xiang-Ning Kang, Guo-Yi Zhang, Simeon Trieu, and Fei Wang, "Optimization of Top Polymer Gratings to Improve GaN LEDs Light Transmission", *OSA Journal: Chinese Optics letters (Focus Issue Nano Photics)*, vol.6, no. 10, pp. 788-790, 2008.

N00014-07-1-1152 P.I.Susan C. Opava, Ph.D.

Optimization of top polymer gratings to improve GaN LEDs light transmission

Xiaomin Jin[1,2], Bei Zhang (章 蓓)[2], Tao Dai (代 涛)[2], Wei Wei (魏 伟)[2], Xiangning Kang (康香宁)[2], Guoyi Zhang (张国义)[2], Simeon Trieu[1], and Fei Wang[3]

[1]*Electrical Engineering Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA*

[2]*School of Physics and State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, Peking University, Beijing 100871*

[3]*Electrical Engineering Department, California State University at Long Beach, Long Beach, CA 90840, USA*

Received July 16, 2008

We present a grating model of two-dimensional (2D) rigorous coupled wave analysis (RCWA) to study top diffraction gratings on light-emitting diodes (LEDs). We compare the integrated-transmission of the non-grating, rectangular-grating, and triangular-grating cases for the same grating period of 6 μm, and show that the triangular grating has the best performance. For the triangular grating with 6-μm period, the LED achieves the highest light transmission at 6-μm grating bottom width and 2.9-μm grating depth. Compared with the non-grating case, the optimized light transmission improvement is about 74.6%. The simulation agrees with the experimental data of the thin polymer grating encapsulated flip-chip (FC) GaN-based LEDs for the light extraction improvement.

OCIS codes: 140.0140, 140.5960, 050.1950.

doi: 10.3788/COL20080610.0788.

In general, GaN solid-state lighting is very critical for future energy conversion. It is a very hot research area and revolutionizes the lighting industry, and is being called "the next generation light sources". Customers worldwide use light-emitting diode (LED) chips to replace traditional bulb technology with solid-state products that provide a powerful and energy-efficient source of blue, green, or white lights. Growth in the high-brightness LED market in the next few years will be driven by lighting, display backlighting, and automotive applications. LEDs are the advanced form of a lamp, and its development can and will continue until all power levels and colors are realized. However, low external quantum efficiency is one of the biggest obstacles for the GaN LED development. Because of the high refractive index of GaN-related material and/or indium tin oxide (ITO) top contact layers, only a few percentage of internal light escapes and is collected outside. Most of the light generated in the active layer experiences total internal reflection and loss in the device material. A common way to solve this light trapping is to form nano/microstructures at the light extraction surface or the bottom reflective layer of the LEDs[1−5]. It has been shown that the micro-sized patterning of the ITO top transparent electrode[6,7] or one-dimensional (1D) nano-patterned structure results in an enhancement of light extraction compared with conventional LEDs (C-LEDs)[8].

For commercial applications, low cost and simplicity in fabrication are desired. It has been demonstrated by Peking University in 2008 that 31.9% of the light extraction enhancement was achieved by using the triangular patterned encapsulated flip-chip (FC) GaN LEDs compared with C-LEDs[9]. In this design, the surface gratings and the encapsulation of a polymer can be simultaneously accomplished in a single procedure. Therefore, it can realize thin and low-cost LED package. Based on this work, we utilize the two-dimensional (2D) rigorous coupled wave analysis (RCWA)[10,11] to GaN-based LED grating model and study top polymer diffraction grating on GaN-based LED grating model design. We also provide a design guideline for the improvement of the LED light extraction and optimize the micro-patterned polymer top grating design. To keep the comparison simple, we still keep the simulation grating period fixed to 6 μm according to our initial experiment[9].

The core algorithm of the model is based on RCWA and enhanced with modal transmission line theory. RCWA represents the electromagnetic fields as a sum of coupled waves[10,11]. A periodic permittivity function is represented using Fourier harmonics. Each coupled wave is related to a Fourier harmonic, allowing the full vectorial Maxwell's equations to be solved in the Fourier domain. Currently, plane wave incidence is assumed and the material is assumed lossless to simplify the calculation. The schematic diagrams of the top grating lattices for the simulation are shown in Fig. 1 with flat interface (non-grating), rectangular interface, and triangular interface. The plane wave is incident from semi-infinite homogenous polymer (refractive index is 1.5) to semi-infinite homogenous air (refractive index is 1.0). The incident angle θ upon the normal of the grating varied from 0° to 90°. The simulation is performed at 460-nm wavelength according to the GaN LED experimental spectra[9]. For each incident angle θ, we calculate the −20 to +20 order diffraction efficiency. The total power transmission is calculated at the end of simulation by summing all the diffraction modes. In the initial simulation (Fig. 2), we calculate three cases according to Fig. 1, in which the cases of Figs. 1(b) and (c) are grating period $\Lambda = 6$ μm, grating height $d = 4$ μm, and grating bottom width $w = 3$ μm. Our simulation shows that the critical angle

 © 2008 Chinese Optics Letters

N00014-07-1-1152 P.I.Susan C. Opava, Ph.D.

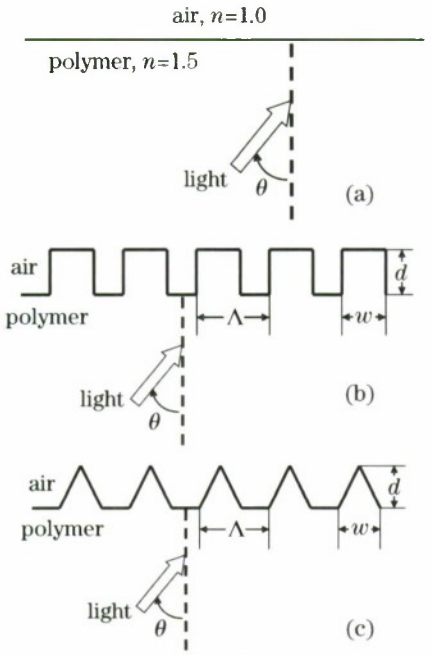


Fig. 1. Simulation schematic diagrams of the top grating lattices. (a) Flat interface (non-grating); (b) rectangular interface; (c) triangular interface.

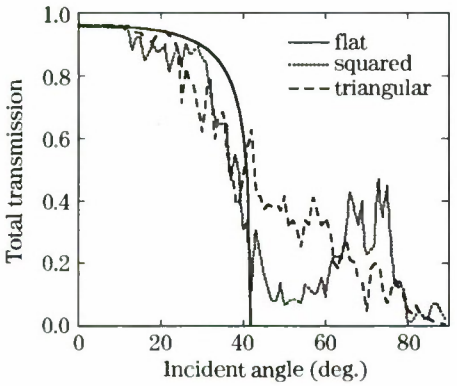


Fig. 2. Comparison of transmission for non-grating (flat), squared-grating, and triangular-grating cases.

is $\theta_c = 42°$ for the non-grating case. For the incident angle above the critical angle of 42°, there is no light transmission for the non-grating case, which agrees with Fresnel's law. In the rectangular and triangular cases for comparison, the transmittance is a little lower for the incident angle below the critical angle. However, the transmittance of the gratings is significantly increased for the incident angle above the critical angle, since a grating can extract some trapped light. The total transmission is the transmittance integrated over the entire region of θ from 0° to 90°, which are 21.79%, 31.60%, and 34.13% for the non-grating case, the squared-grating case, and the triangular-grating case, respectively. Improvements of about 45% (squared) and 56.6% (triangular) are obtained over the non-grating case. This means the triangular-grating has a higher total transmitting diffracting effect than that of the squared grating. In the previous experiment[9], the enhancement factor of light extraction for the triangular grating ($\Lambda = 6$ μm, $d = 4$ μm, and $w = 3$ μm) was 31.9%, which should include light transmission from GaN layer to polymer and from polymer to air. Our above simulations only consider the polymer to air transmission. If we include GaN to polymer transmission efficiency in the simulation, the total transmission from GaN layer to air should be 8.40% with triangular grating and 6.41% without grating, which is about 31.1% improvement. Our triangular-grating simulation results agree with the experimental data presented in Ref. [9] very well.

Increasing the top grating transmission will directly improve the total light extraction of LEDs. To keep the simulation time short and simplify the problem, we still focus on the polymer grating calculation for our design optimization. Figure 3 shows the simulation results of the transmission versus the incident angle for the triangular-grating case. The grating period is 6 μm, and the grating depth is 4 μm. The bigger the grating bottom width, the more light transmits at an incident angle above the non-grating-case critical angle and the less light transmits at an incident angle below the non-grating-case critical angle. To understand the total transmission efficiency, we integrate the transmittance over the incident angle and normalize the integration. The final results of the optimized triangular grating for period $\Lambda = 6$ μm are shown in Fig. 4. At a small value of the grating height d, the transmittance improvement at the large incident angle is dominating. At the largest d value the transmittance improvement at the large incident angle is almost equal to the transmittance degradation at the small incident angle; therefore the total efficiency will not be improved any more with the further increase of the grating height d.

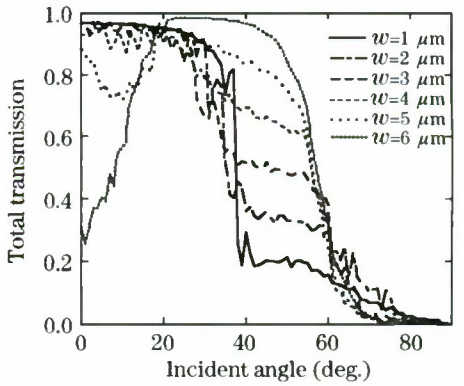


Fig. 3. Simulation results of the light transmission for the triangular-grating case. $\Lambda = 6$ μm, $d = 4$ μm.

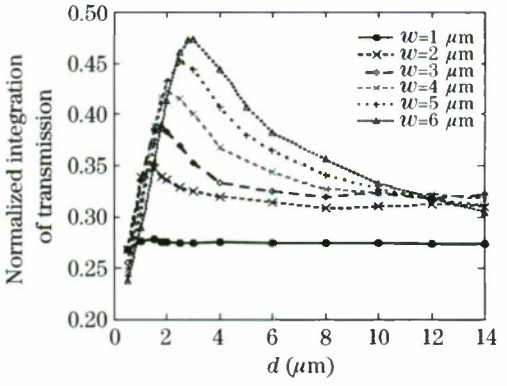


Fig. 4. Light transmission versus grating height for different grating bottom widths of polymer grating at the period of 6 μm.

N00014-07-1-1152 P.I.Susan C. Opava, Ph.D.

There exists an optimal value in the grating depth design. We can clearly see that the 6-μm grating width and the 2.9-μm grating depth provide the highest light extraction rate. Compared with the non-grating case, the maximum light enhancement is about 117% for the triangular-grating case, which is much better than 56.6% of the non-optimized case in Fig. 2. Our data clearly shows that the diffraction of the grating improves the overall light extraction of the GaN LEDs.

Even though our simulation is only performed at one grating period value (6 μm), this is a very representative and informative case, which enlightens several design guidelines for the GaN LED grating clearly. Firstly, at the same grating period, the triangular grating has the best performance compared with the non-grating and squared-grating cases. Secondly, for the triangular grating, the grating bottom width (w) should be set to the grating period (Λ) to obtain the highest light extraction efficiency. Thirdly, the light transmission coefficient of triangular gratings varies according to either w or d variation, which changes the light incident angle at the interface and modifies the total light extraction. The gratings can greatly improve light extraction above the non-grating-case critical angle, but decrease the light extraction below the non-grating-case critical angle. Overall, there is an optimzation point for the design. Fourthly, our triangular grating experimental data[9] is right on our simulation chart Fig. 4. This shows that our experimental results can be further improved. Finally, there are also other parameters, such as grating period and polymer material index (can be equal to 1.5, 1.4, or even 1.3) can be considered in our future simulations to further improve the light extraction beyond this work.

In conclusion, we compare the flat interface (non-grating), rectangular-grating, and triangular-grating cases, and show that the triangular grating has the best performance. For the triangular grating with a 6-μm period, the LEDs have the highest light transmission, which reaches the maximum output at 6-μm width and 2.9-μm grating depth. Compared with the non-grating case, the maximum light transmission improvement for just the grating is about 117%. If we include the GaN layer in the simulation, the total light transmission is about 11.19%, which is an improvement of about 74.6% upon the non-grating case.

This work was supported by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152, USA, the "Chunhui" Exchange Research Fellow 2008, Ministry of Education of China, the National "973" Program of China (No. 2007CB307004), the National "863" Program of China (No. 2006AA03A113), and the National Natural Science Foundation of China (No. 60276032, 60577030, and 60607003). X. Jin's e-mail address is xjin@calpoly.edu.

References

1. S.-H. Huang, R.-H. Horng, K.-S. Wen, Y.-F. Lin, K.-W. Yen, and D.-S. Wuu, IEEE Photon. Technol. Lett. **18,** 2623 (2006).
2. M.-K. Lee, C.-L. Ho, and P.-C. Chen, IEEE Photon. Technol. Lett. **20,** 252 (2008).
3. T. V. Cuong, H. S. Cheong, and C.-H. Hong, Phys. Stat. Sol. (c) **1,** 2433 (2004).
4. S. H. Kim, K.-D. Lee, J.-Y. Kim, M.-K. Kwon, and S.-J. Park, Nanotechnology **18,** 055306 (2007).
5. S.-M. Pan, R.-C. Tu, Y.-M. Fan, R.-C. Yeh, and J.-T. Hsu, IEEE Photon. Technol. Lett. **15,** 646 (2003).
6. S.-M. Pan, R.-C. Tu, Y.-M. Fan, R.-C. Yeh, and J.-T. Hsu, IEEE Photon. Technol. Lett. **15,** 649 (2003).
7. S. M. Huang, Y. Yao, C. Jin, Z. Sun, and Z. J. Dong, Displays **29,** 254 (2008).
8. H.-G. Hong, S.-S. Kim, D.-Y. Kim, T. Lee, J.-O. Song, J. H. Cho, C. Sone, Y. Park, and T.-Y. Seong, Appl. Phys. Lett. **88,** 103505 (2006).
9. K. Bao, X.-N. Kang, B. Zhang, T. Dai, C. Xiong, H. Ji, G.-Y. Zhang, and Y. Chen, IEEE Photon. Technol. Lett. **19,** 1840 (2007).
10. M. G. Moharam and T. K. Gaylord, J. Opt. Soc. Am. A **3,** 1780 (1986).
11. L. Li, J. Opt. Soc. Am. A **14,** 2758 (1997).

6. Project Results 3: Design Simulation of Top ITO Nano-Gratings to Improve Light Transmission for Gallium Nitride LEDs

Related paper: X. Jin, S. Trieu, Fei Wang, B. Zhang, T. Dai, X. N. Kang, and G. Y. Zhang, "Design Simulation of Top ITO Gratings to Improve Light Transmission for Gallium Nitride LEDs", 2009 Sixth International Conference on Information Technology: New Generations, ITNG2009, April 27-29, 2009, Las Vegas, Nevada, USA.

2009 Sixth International Conference on Information Technology: New Generations

Design Simulation of Top ITO Gratings to Improve Light Transmission for Gallium Nitride LEDs

Xiaomin Jin[1,3], Simeon Trieu[1], Fei Wang[2], Bei Zhang[3], Tao Dai[3], Xiangning Kang[3], and Guoyi Zhang[3]

[1] *Electrical Engineering Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA, xjin@calpoly.edu*

[2] *Electrical Engineering Department, California State University at Long Beach, Long Beach, CA 90840, USA*

[3] *State Key Laboratory for Mesoscopic Physics and Department of Physics , Peking University, Beijing 100871,PR China*

Abstract

We present simulation results of the indium tin oxide (ITO) top diffraction grating using a rigorous couple wave analysis (RCWA) for GaN LEDs. We study three different nano-structure patterns: cylindrical pillar grating, conical pillar grating, and cylindrical nano-hole grating. We show the light transmission improvement with nano-grating designs and present design-charts for the nano-hole grating.

1. Introduction

The conventional GaN-based light-emitting diodes (LEDs) have a low light extraction efficiency caused by the total internal reflection. A common way to solve this light trapping is to etch a periodic nano-structure at the light extraction surface and/or the bottom reflective layer of the LEDs [1]-[4].

For commercial applications, low cost and simplicity in fabrication are desired. It has been demonstrated by Peking University in 2008 that 32% of light extraction enhancement was achieved by using the triangular patterned encapsulated Flip-chip (FC) GaN LEDs compared to C-LEDs [5]. In this design, the surface gratings and the encapsulation of a polymer can be simultaneously accomplished in a single procedure; additionally it provides low height profile. Therefore, it can realize thin and low cost LED package. However, our pervious work focuses on mirco-scale grating patterns. To obtain design guidelines and understand design parameters in the nano-scale, we developed a three-dimensional (3D) rigorous couple wave analysis (RCWA) GaN-based light-emitting diode (LED) grating model to study top diffraction grating design using *Rsoft DiffMod* [6]. In the paper, we also provide a design charts for the improvement of the LED light extraction and optimize the nano-hole-patterned polymer top grating design.

In summary, we present a three-dimensional (3D) rigorous couple wave analysis (RCWA) GaN-based light-emitting diode (LED) grating model to study the top indium tin oxide (ITO) diffraction grating performance at 460nm wavelength in order to improve the GaN LED light transmission efficiency.

2. Design simulation

The core algorithm of the model is based on RCWA and enhanced with modal transmission line theory. The RCWA [7] [8] represents the electromagnetic fields as a sum of coupled waves. A periodic permittivity function is represented using Fourier harmonics. Each coupled wave is related to a Fourier harmonic, allowing the full vectorial Maxwell's equations to be solved in the Fourier domain. Currently, plane wave incidence is assumed and material is lossless to simplify the calculation.

Usually it is not very practical to fabricate all kinds of the top ITO textures or patterns to select the optimized structure [9][10]. Therefore, we simulate three typical gratings: cylindrical pillar grating, conical pillar grating, and cylindrical nano-hole grating, as shown in Fig. 1(a)-(c). The detailed GaN LED layer structure is also presented in Fig. 2. In our device design, it is very important to keep a 30nm-fixed ITO thickness at bottom of the grating. This fixed thickness is used to prevent P-GaN layer from being damaged in the etching process and

978-0-7695-3596-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ITNG.2009.10

IEEE computer society

protect the overall device charaterization. Furthermore, the 30nm-bottom ITO layer also acts as the current injection layer to protect the LED I-V charaterization from being effected by the nano-structure.

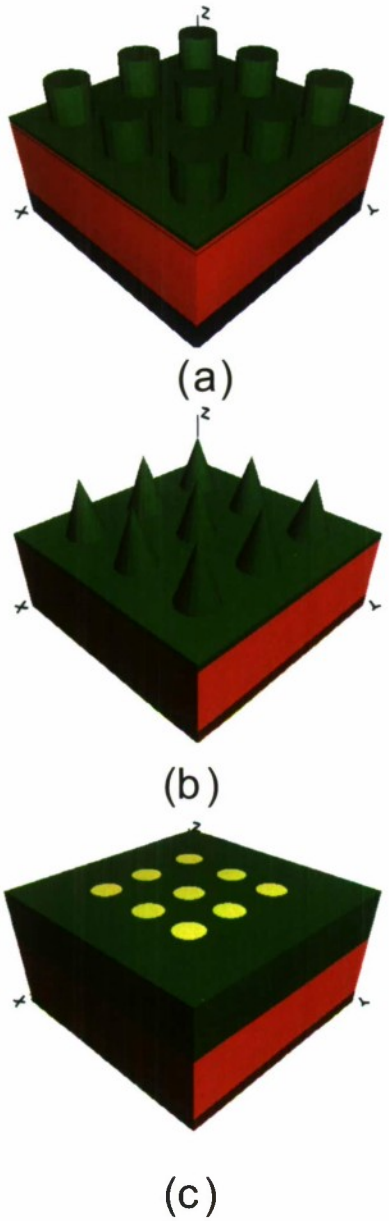


Fig. 1 The schematic diagrams of the top grating simulation a) cylindrical pillar grating, b) conical pillar grating, and c) cylindrical nano-hole grating

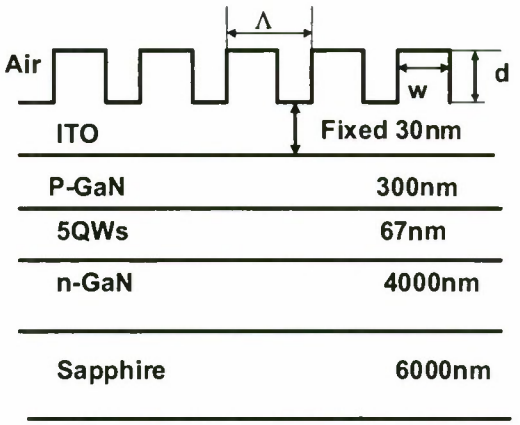
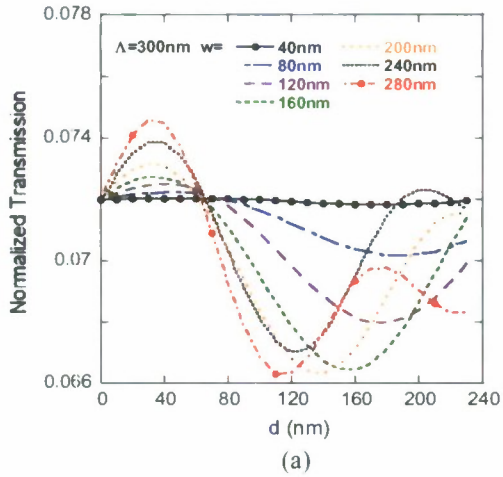


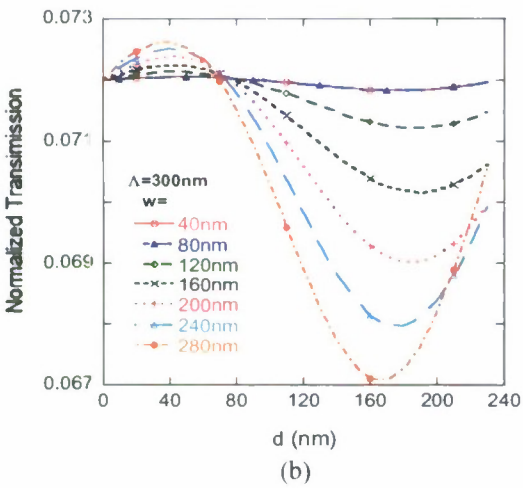
Fig. 2 The detailed LED layer structure in the models.

In the grating simulation model, there are three major parameters that affect the light extraction: grating period (Λ), grating height (d), and bottom width (w). First, according to our fabrication capability, we choose a large grating period Λ=300nm for the initial calculation [11]. We sweep the other two parameters (d and w) and their possible permutations. The simulation results are shown in Fig. 2 (a)-(c). For the grating height (d) smaller than 60nm, cylindrical pillar grating has the best performance. All three cases have the best performance at larger w value or the small air filling factor. At large period Λvalue, it is recommended to make shallow nano-structure to improve light transmission. For the grating height (d) larger than 60nm, cylindrical nano-hole grating has the best performance.

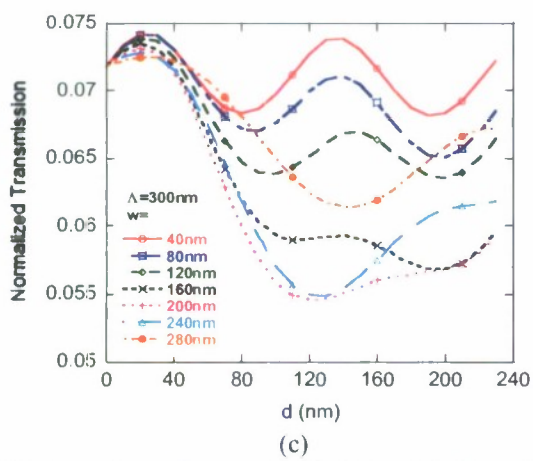
In general, nano-hole structure can improve light extraction in a wider range of the grating height (d). Nano-hole is also easy to be achieved in our fabrication [11]. We furthermore study the nano-hole grating in detail (Λ=140nm, 180nm, 220nm, and 260nm as well), and shown in Fig. 4. We find that the Λ=140nm case has the highest light extraction output. Compared to the no-grating case, the light extraction improvement is about 10% for the 230nm-depth and 120-width grating. However, the smaller pattern size raises fabrication challenges. Actually, the large grating period is not an idea design for the conical pillar grating and the nano-hole grating.



(a)

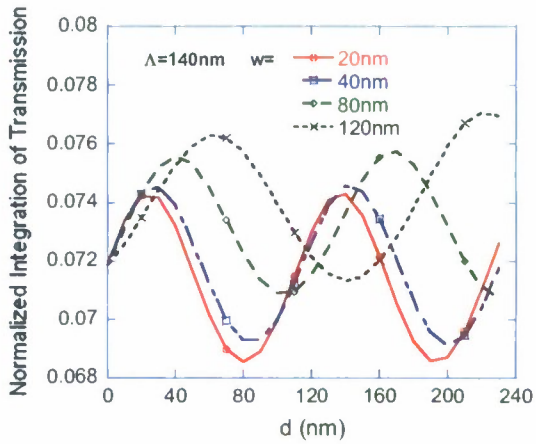


(b)

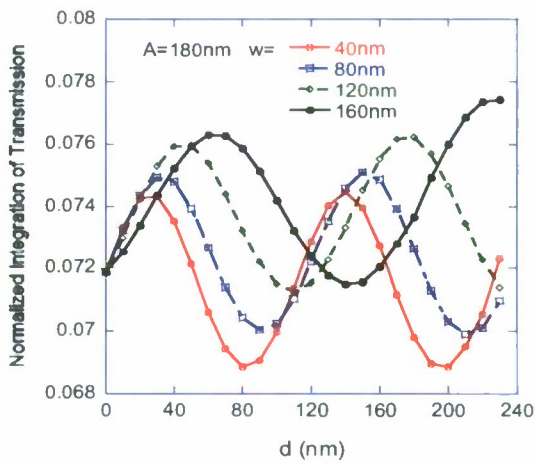


(c)

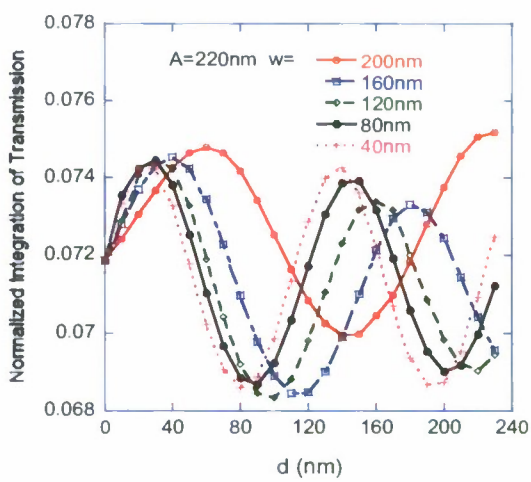
Figure 3. The simulation results a) cylindrical pillar grating, b) conical pillar grating, c) cylindrical nano-hole grating for grating period of Λ=300nm.



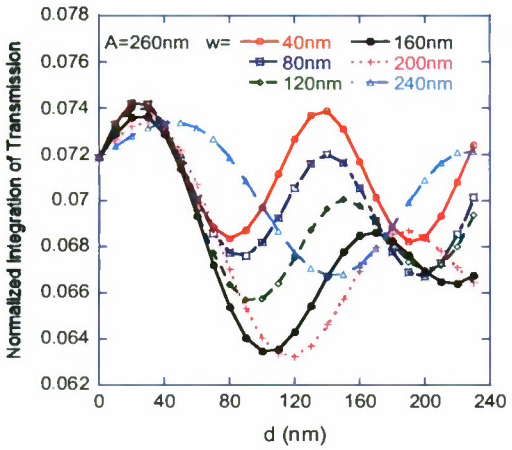
(a)



(b)



(c)



(d)

Fig. 4 Nano hole grating simulation results with Λ=140, 180, 220, 260nm.

3. Conclusion

In this paper, we present the simulation results of nano-scale grating design in GaN LEDs. We compare three different grating structures: cylindrical pillar grating, conical pillar grating, and cylindrical nano-hole grating. And we show that the small grating period will yield more light extraction efficiency.

4. Acknowledgement

This work is sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152 and "ChunHui" research fellow 2008, Ministry of Education, China. 973 Program-National Basic Research Program of China (2007CB307004); High Technology program (863-2006AA03A113) and National Nature Science Foundation of China (60675032, 60577030 and 60607003).

5. References

[1] S. M. Huang , Y. Yao , C. Jin , Z. Sun , and Z. J. Dong, "Enhancement of the light output of GaN-based light-emitting diodes using surface-textured indium-tin-oxide transparent ohmic contacts", Displays, vol. 29, 2008, pp. 254-25,.

[2] G S.-H. Huang, R.-H. Horng, K.-S. Wen, Y.-F. Lin, K.-W. Yen, and D.-S. Wuu, "Improved light extraction of Nitride-based flip-chip light-emitting diodes via Sapphire shaping and texturing", IEEE Photon. Tech. Lett., vol. 18, 2006, pp. 2623-2625.

[3] M.-K. Lee, C.-L. Ho, and P.-C. Chen, Light Extraction Efficiency Enhancement of GaN Blue LED by Liquid-Phase-Deposited ZnO Rods, IEEE Photon. Tech. Lett. vol 20, 2008, pp. 252-255.

[4] T. V. Cuong, H. S. Cheong, and C.-H. Hong, Calculation of the external quantum efficiency of light emitting diodes with different chip designs, Phys. Stat. Sol. (c) vol.1, 2004, pp. 2433-2440 .

[5] K. Bao, X.-N. Kang, B. Zhang, T. Dai, C. Xiong, H. Ji,G.-Y. Zhang, and Y. Chen, Improvement of Light Extraction from Micro-Pattern Encapsulated GaN-based LED by Imprinting IEEE Photon. Technol. Lett. vol. 19, 2007, pp.1840-1842.

[6] Rsoft Design Inc, http://www.rsoftdesign.com/.

[7] M. G. Moharam and T. K. Gaylord, "Rigorous coupled-wave analysis of metallic surface-relief gratings", J. Opt. Soc. Am. A, vol. 3, 1986, pp. 1780-1788.

[8] L. Li, New formulation of the Fourier modal method for crossed surface-relief gratings, J. Opt. Soc. Am. A vol. 14, 1997, p. 2758.

[9] S. H. Kim, K.-D. Lee, J.-Y. Kim, M.-K. Kwon, and S.-J. Park, Fabrication of photonic crystal structures on light emitting diodes by nanoimprint lithography, TOP Nanotechnology, vol. 18, 2007, p. 055306.

[10] H.-G. Hong, S.-S. Kim, D.-Y. Kim, T. Lee, J.-O. Song, J. H. Cho, C. Sone, Y. Park, and T.-Y. Seonga, Enhancement of the light output of GaN-based ultraviolet light-emitting diodes by a one-dimensional nanopatterning process, App. Phy. Lett. 88, 2006, pp. 103505-103507.

[11] T. Dai, X.N. Kang, B. Zhang, Z.S. Zhang, D. Liu, X. Wang, K. Bao, X.N. Kang, J. Xu, D.P. Yu, and X. Zhu, "Surface Light Extraction Mapping from two-dimensional Array of 12-Fold Photonic Quasicrystal on Current Injection GaN-Based LEDs," Chin. Phys. Lett., vol. 24, no. 4, 2007, pp. 979-982.

7. Project Results 4: Design of GaN Bottom Reflection Gratings on GaN-based Light-emitting Diodes

Related paper: Simeon Trieu, Xiaomin Jin, Bei Zhang, Tao Dai, Kui Bao, Xiang-Ning Kang and Guo-Yi Zhang, "Light Extraction Improvement of GaN-based Light-emitting Diodes using Patterned Undoped GaN Bottom Reflection Gratings", the SPIE International Symposium on Integrated Optoelectronic Devices 2009, SPIE Photonic West 2009, San Jose, CA USA 24-29, January 2009.

Light Extraction Improvement of GaN-based Light Emitting Diodes using Patterned Undoped GaN Bottom Reflection Gratings

Simeon Trieu[a], Xiaomin Jin[*a,b], Bei Zhang[b], Tao Dai[b], Kui Bao[b], Xiang-Ning Kang[b], Guo-Yi Zhang[b]
[a]Electrical Engineering Department, 1 Grand Avenue,
California Polytechnic State University, San Luis Obispo, CA, USA, 93407-9000;
[b]School of Physics and State Key Laboratory for Artificial Microstructures and
Mesoscopic Physics, Peking University, Beijing, China, 100871

ABSTRACT

The Gallium Nitride (GaN) Light-Emitting-Diode (LED) bottom refection grating simulation and results are presented. A microstructure GaN bottom grating, either conical holes or cylindrical holes, was calculated and compared with the non-grating (flat) case. A time monitor was also placed just above the top of the LED to measure both time and power output from the top of the LED. Many different scenarios were simulated by sweeping three parameters that affected the structure of the micro-structure grating: unit cell period (A) from 1 to 6 microns, unit cell width (w) from 1 to 6 microns, and unit cell grating height (d) from 50 to 200nm. The simulation results show that the cylindrical grating case has a 98% light extraction improvement, and the conical grating case has a 109% light extraction improvement compared to the flat plate case.

Keywords: Gallium Nitride, light-emitting-diode, grating

1. INTRODUCTION

As a result of our energy conversation efforts, lighting sources have become one of the hot areas of research due to their applications in a variety of fields such as lighting displays, bulb technology, and photonics. The demands of these applications require low power consumption, yet a high brightness and luminosity with minimal heat. We can even control the color contrast of the device and create a full color set with red, green, and blue Light-Emitting-Diodes (LEDs) [1]. LEDs have been used in many applications, however the light extraction efficiency is still very low for GaN LEDs due to several factors: Gallium Nitride (GaN) has a low critical angle that traps light inside the device [2], absorption of light within the device due to dislocations and defects within the GaN crystal [3], and device design and structure has not been optimized (ie. epitaxial side up vs. epitaxial side down chip structures) [4]. It is crucial that we improve GaN LED light extraction efficiency and reduce energy consumption.

The major limitation to the light extraction efficiency is the light trapping due to GaN's low critical angle. This applies to any large change of the refraction index existing between layers, such as between the solid state LED and air, and the solid state LED and a resin. It has been shown that resins increase the extraction efficiency due to the more gradual change in refraction, allowing more light emitting out due to the larger escape angle [5]. Many methods of improving LED efficiency are currently being explored. Almost all of these methods are seeking to extract the trapped light in greater quantity and faster speed. Those methods being explored are placing photonic crystals or nanostructure grating on one of LED layers to modify the effective index of refraction at the boundary [6-7], randomized roughening on the surface of the device [7-8], slanted device configurations that result in pyramidal shapes [7], and inverted "flip-chip" designs that put the epitaxial side upwards or downwards [4] [9].

The second inefficiency of GaN LEDs is the loss of light from absorption due to dislocations and defects within the GaN crystal. These are impurities that absorb the light, an issue when light is trapped inside the LED. The longer it takes to extract the trapped light, the more the photons suffer from absorption. So, it is critical that we find methods to extract light quickly from the LED before the energy is taken by absorption [10-11]. A grating structure helps

*xjin@calpoly.edu; phone 1 805-756-7046; fax 1 805-756-1458; www.ee.calpoly.edu

Gallium Nitride Materials and Devices IV, edited by Hadis Morkoç, Cole W. Litton, Jen-Inn Chyi, Yasushi Nanishi, Joachim Piprek, Euijoon Yoon, Proc. of SPIE Vol. 7216, 72162Q
© 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.805480

solve this issue by creating more angles of escape. Various structures exist, such as conical, pyramidal, spherical, cylindrical, and so on, but only a few can be realized with current fabrication techniques. To help mitigate losses due to absorption and facilitate the quick extraction of photons, gratings can be patterned on the surface of or within the layers of the LED device, providing more escape angles than flat interface. For example, through modified laser lift-off (M-LLO), air holes at a 4 micron period are patterned on undoped GaN (U-GaN) instead of just roughening the surface of an LED. The technique uses a sapphire backplane, UV light, and a high power KrF laser to etch the nanostructure onto U-GaN [12]. In the experiment of PKU, grating depths varied from 75nm to 120nm [12]. Other fabrication techniques exist such as imprint lithography that can produce similar air holes that measure 180nm in diameter, with a depth of 100nm, and a period of 295nm [13]. Our simulations will be based off of the M-LLO manufacturing process of PKU [12], a model with a patterned U-GaN bottom grating layer attached to a reflective Ag film is persented. The grating height plays a role in transmittance and reflection, since the larger the grating height, the more gradual the effective refraction index will be, and hence, the more photons that will be transmitted versus reflected and absorbed by impurities [6] [14].

The light trapping issue is commonly solved by etching a periodic structure at the light extraction surface and/or the bottom reflective layer of the LED [2]. This paper focuses on micro-scale grating structures at the bottom of the device. To simulate the effects on extraction efficiency, we use a Finite-Difference Time-Domain (FDTD) GaN LED model at 460nm wavelength to study different reflection gratings. The GaN reflection gratings using cylindrical-hole bottom, conical-hole bottom, and non-grating structures are simulated and optimized. In the paper, the simulation model is presented in section 2, simulation results are presented in section 3, and the conclusion is in section 4.

2. SIMULATION MODEL

The presented simulation model accounts for the effects of refraction in device materials, reflection due to linear dispersion or total internal reflection, transmission of escaping light from the LED, and scattering at the grating [3]. The form of analysis used in the simulation is the common solution for the light wave propagation in arbitrary geometries, the Finite Difference Time Domain (FDTD) technique. This allows for the most accurate solution, since FDTD is based directly on Maxwell's curl equations, which can accurately calculate all four of the above cases. In conjunction with Maxwell's equation, a Yee mesh is used (shown in Fig. 1), which enables analyzing the electromagnetic fields on a grid of three-dimensional space and time. By defining an incremental variable for each parameter in space and time, namely Δx, Δy, Δz, and Δt, the E and H fields for a specified grid are obtained. The model allows simulations of any geometrical structure of LEDs. An important point to consider is that the smallest increment must be no smaller than the smallest feature of the structure, or the model will not be able to accurately calculate the exact features of the structure. The GaN simulation model has the parameters shown in Fig. 2. The actual LED is mounted on a very thick layer of Si submount with an Ag reflector plate directly on top of it. The Ag fills the U-GaN air holes in either a conical or cylindrical unit cell shape (only conical is shown in Fig. 2). Adding the submount layer will add a large amount of calculation times. However, the Ag is a strong reflective layer and effects of submount on the EM field are very trivial compared to other layers. Therefore, the submount is not included in the modal. Finally, an n-GaN, quantum wells, and p-GaN are placed on top of the grating. The time monitor is used to measure the time-varied light output of the whole device and is separated from the LED by a distance of 100nm.

The model has cylindrical holes or conical holes on the bottom of the undoped GaN layer as shown in Fig. 2 and Fig.3. The conical and cylindrical shapes represent Ag material, which is a reflective layer. For a visual representation of the grating types, the conical grating model is demonstrated in Fig. 3(a), while the cylindrical grating model is demonstrated in Fig. 3(b). To define a regular spacing between unit cells in a crystal lattice arrangement, we employ three parameters: unit-cell period (A), unit-cell height (d), and unit-cell width (w), all shown in Fig. 3. The unit-cell period is the length from center-to-center between unit cells. In 3-D, the parameter w, represents a diameter in the case of a circular structure (ie. sphere, cone, and cylinder) or a length of a side in the case of a box structure (ie. cube, rectangular cube). The unit-cell height d is the depth of the bottom hole. The limits of the parameters are the following: $50 \leq d \leq 200$nm, $w \leq A$ to prevent overlap, and the smallest element must be bigger than the smallest grid size. The grating height parameters, which are fabricated, range from 75nm to 120nm [12]. This determines our simulation rang. More over, the smallest grating height d value is determined by the minimum grid size limit. If $d < 50$nm, the simulation requires much smaller girds and simulation time. And our data

also show little improvement of light extraction at d smaller than 60nm. Further more, the width parameter must be less than or equal to the unit cell period since a larger value would indicate overlap of the cell structure, an invalid state. Finally, the last limitation is that the smallest element must be greater than or equal to the smallest grid size, as specified in the Yee's mesh simulation. If the grid was not fine enough to "see" the layer, then it may be missed when calculating for the E and H fields using Maxwell's curl equations. This also determines the minimum distance of the power monitor to the device in Fig. 2, which is set to 100nm above the LED. The detailed simulation values are shown in Table 1. Fractions of microns for A and w are allowed in the simulation, however they were not used. The overall size of FDTD simulation area is also fixed, which is100μm × 100μm. If A=1μm, there are 100 unit cells. For A=6μm, there are only 16.67 unit cells. The grating is a squared matrix as shown in Fig.3. A point-source layer with constant wave is placed at the center of quantum well region.

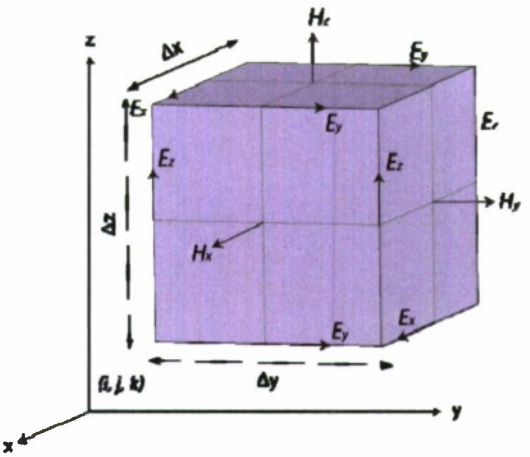
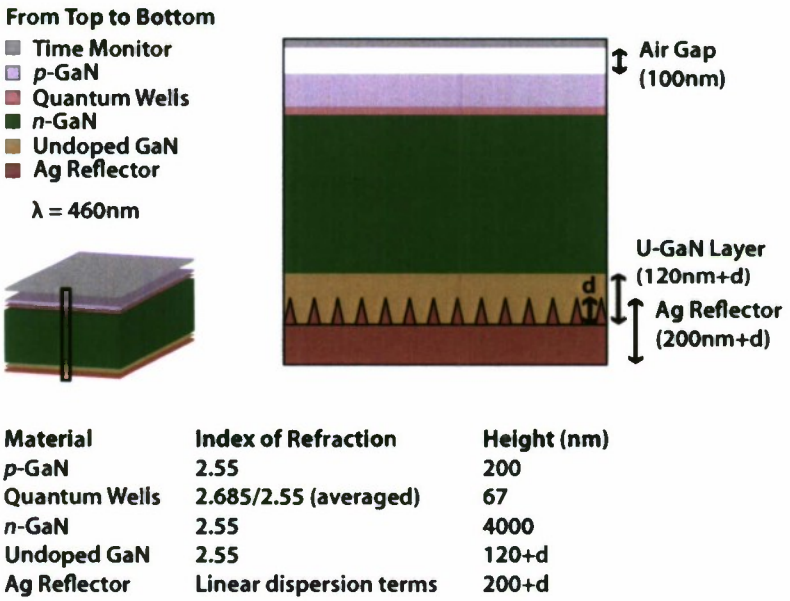
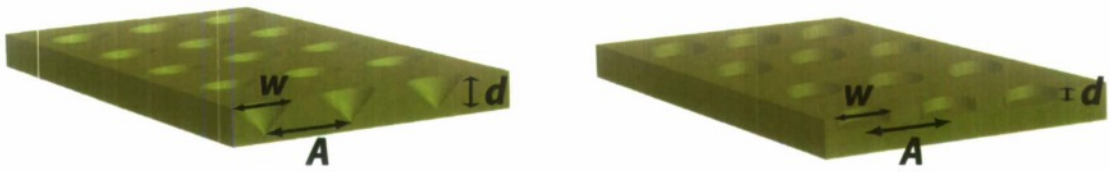


Fig. 1. Yee's Mesh Solution to Solve Maxwell's Equations



Material	Index of Refraction	Height (nm)
p-GaN	2.55	200
Quantum Wells	2.685/2.55 (averaged)	67
n-GaN	2.55	4000
Undoped GaN	2.55	120+d
Ag Reflector	Linear dispersion terms	200+d

Fig. 2. GaN LED simulation model



(a) (b)

Fig. 3. Grating unit cells shape on GaN material (bottom view): (a) Conical and (b) Cylindrical

Table 1. Range of the Simulation Parameters

GaN LED Model Simulation Parameters		
Simulation Parameters	Descriptions	Simulated Ranges
Parameter:	*Brief Description:*	*Range:*
A	The period of the unit cells	1 to 6 microns
w	The width of each unit cell	1 to 6 microns
d	The height of each unit cell	50 to 200 nm

3. SIMULATION RESULTS AND DISCUSSIONS

The start of the simulation is when the LED turns on from an off state. There is a monitor at a distance of 100nm above the LED to collect the light emission. For each structure simulation, we must wait until the LED/monitor reach a steady state, after which we can extract the constant wave (CW) average power. Each simulation sweep of the parameters produces a set of 20000 data points (2000fs at 0.1fs/step), and from this data set, we determine average power over the last 500fs. In this range, maximum steady-state power is radiated due to the CW source and reflections from the grating. This procedure is done for each case.

Fig. 4 shows the results of a GaN conical-hole grating simulation, sweeping from A=1μm to A=6μm, w=1μm to w=6μm, and d=50nm to 150nm. Similarly, Fig. 5 shows the results of a GaN cylindrical-hole grating simulation, sweeping over the same range. The flat plate (non-grating) results are shown for each graph for comparison. The average powers in the plots are the maximum average output power of the LED on the last 500fs of the simulation at steady state. For the conical-hole grating, there is maximum light extraction around the grating height d=90nm. The average power increases as grating height increases when d<90nm, and the average power decreases as the grating height increase from 90nm to 150nm for most cases. For the same grating period, the smaller grating width achieves the better light extraction. For the transmission grating, the maximum transmission results from the largest grating width case [15]. In this paper, the reflection gratings are simulated. Since the reflected power summed with the transmitted power equals the total incident power, a maximum transmission structure is A=w. Then the maximum reflection grating structure must differ, and in fact be opposite, from the maximum transmission structure. When w increases at the same d, the total light extraction reduces and more approaches to the flat case. For the same A and w, the interface angle of the grating section changes when d varies. When d=0, which is flat case, there is no improvement. However, it is very hard to simulation small d values case. With d increasing, the interface angle increases, more light will be extracted at small d<90nm. When d is approaches infinite, the grating section is too thick and it is harder for the reflected light to pass through and be extracted. Therefore the light emission efficiency drops at higher d values. There is maximum design point for this particular structure, which is around d=90nm. The best case in our simulation is 109% light extraction improvement at grating period A=6μm, grating width w=1μm, and grating height d=90nm. For the conical-hole grating case, there are two effects of period width w: 1) The percentage of grating area compared to the total device area. If width w=0, there is no effects and no improvement

of the light extraction. When w increases from zero, more lights should be able to be extracted. However, we didn't simulation w<1μm (nano-structure). 2) For the same period A and grating height d, but different width w, the angle of the Ag and GaN interface are different. When w increases, the angle decreases and should be much closer to the flat case. Therefore, at large w, the light extraction should decrease as w increase. Summarizing above two cases, there is a maximum w design point for each grating, which should be shown in Fig. 6(a). However, our data is only calculated down to w=1μm gating. The maximum grating design should be somewhere below 1 μm (nano-structure). But for the fabrication view point, it is not necessary to design the grating to nano-structure at our current fabrication capability. In the theoretical part, the other factors should also be considered for the nano-grating design, i.e. the structure is compatible or smaller than the light wavelength.

For the GaN cylindrical hole grating simulation, shown in the Fig. 5, if the grating width w equals to the grating period, there are very little light extraction improvement. In most case, the light extraction efficiency is even worse compared to the no grating case. Light emission improvement can only be achieved at w<A cases. Similar to the GaN conical hole grating simulation, the smallest grating width produces the highest average power output. There are two cases, which give the maximum average power output: one is 97.9% light extraction improvement at the grating period A=2 μm, grating width w=1 μm, grating height d=60nm and the other is 94.8% light extraction improvement at the grating period A=6 μm, grating width w=1 μm, grating height d=90 nm. For the cylindrical grating, w=A is a very special case, since the shape of the grating is totally different. There are only four separated islands of GaN on the four corners of the unit cell. Our results show this case is worse than the flat surface interface. Our data also show that the small grating period is not preferred for the micro-grating design, such as A=1 and 2 μm. This is contrast to conical case, for A=1 or 2μm, there are still reasonable grating design at d=90nm.

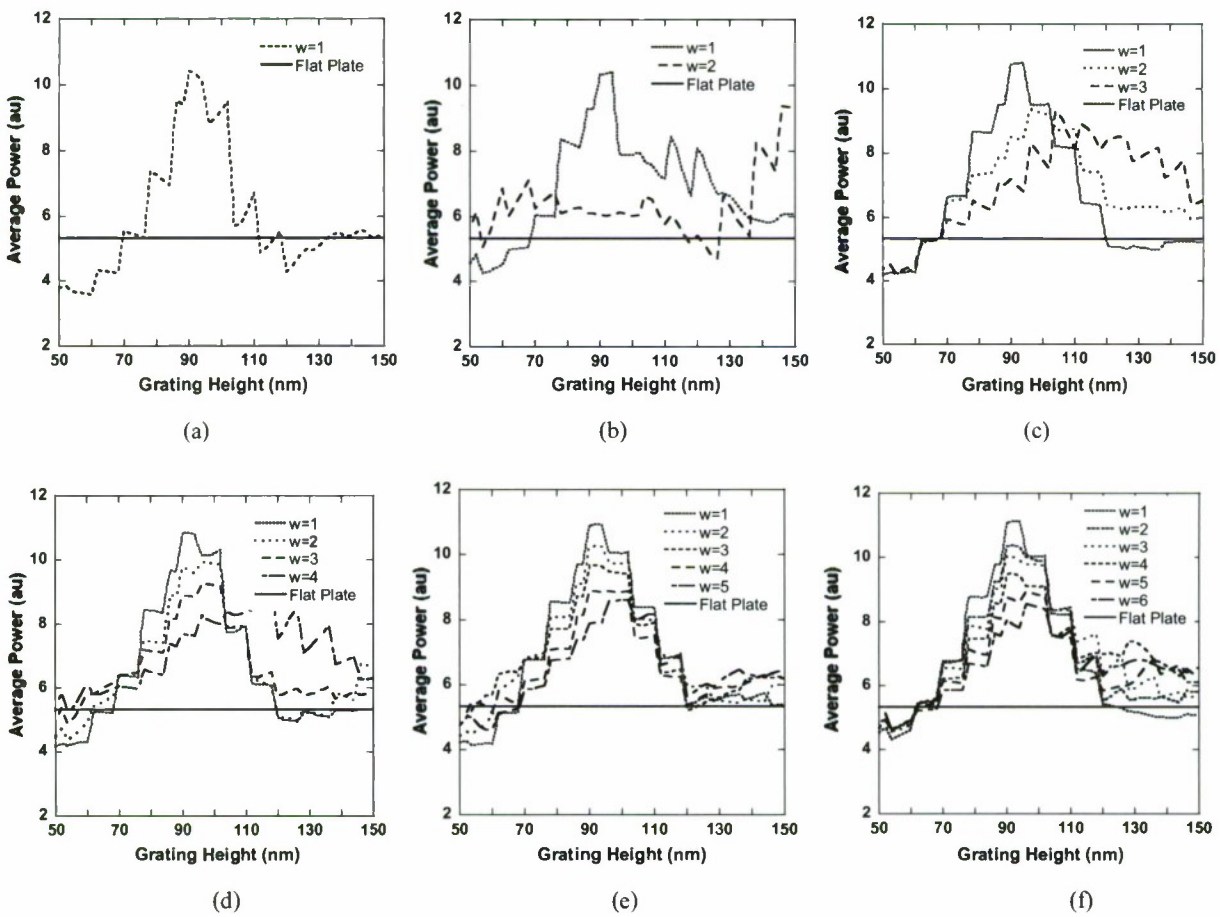


Fig. 4. Average power for the conical hole grating Case: (a) A=1, (b) A=2, (c) A=3, (d) A=4, (e) A=5, and (f) A=6 μm.

The summary of maximum power for the GaN conical- and cylindrical-hole grating is shown in Fig. 6. In general, larger grating periods have higher maximum power for both cases. This is a very important result to guide our fabrication and design. This implies that the large-grating design is preferred in micro-levels. The maximum power linearly decreases with increase of grating width w. Compared to cylindrical-hole grating, the conical grating maximum power is higher at larger w value. The conical gratings are less sensitive to A and w values for the maximum power output, which is also preferred by fabrication. For smaller grating width, i.e. w=1 μm, the two cases (conical and cylindrical) has little difference, which is also shown in Fig. 7. Since w=1 μm case is the best case for most circumstances, the comparison of the conical- and cylindrical-hole grating for the w=1 μm case is shown in Fig. 7. For the simulated range, Fig. 7 (a) shows the peak power generated by each unit cell shape and for each grating period. Also, the grating heights in nm for those maximum powers are shown in Fig. 7(b), also plotted against the grating period A. The cylindrical grating case has a 98% improvement compared to the flat plate case at a grating period of A=2μm, and a unit cell width of w=1μm. The conical grating has a 109% improvement compared to the flat plate case at a grating period of A=6μm, and a unit cell width of w=1μm. However, the maximum power or peak power locations are almost same, round the grating height d= 90nm for both grating cases at larger grating period (A>2μm). In the other words, the optimized the grating height is 90nm for most cases. In summary, it is better design the reflection gratings (conical or cylindrical) at larger grating period A=6μm, smaller grating width w=1μm, and the grating height d round 90nm. Compared to cylindrical gratings, the conical grating has more design tolances on the grating width w and period A.

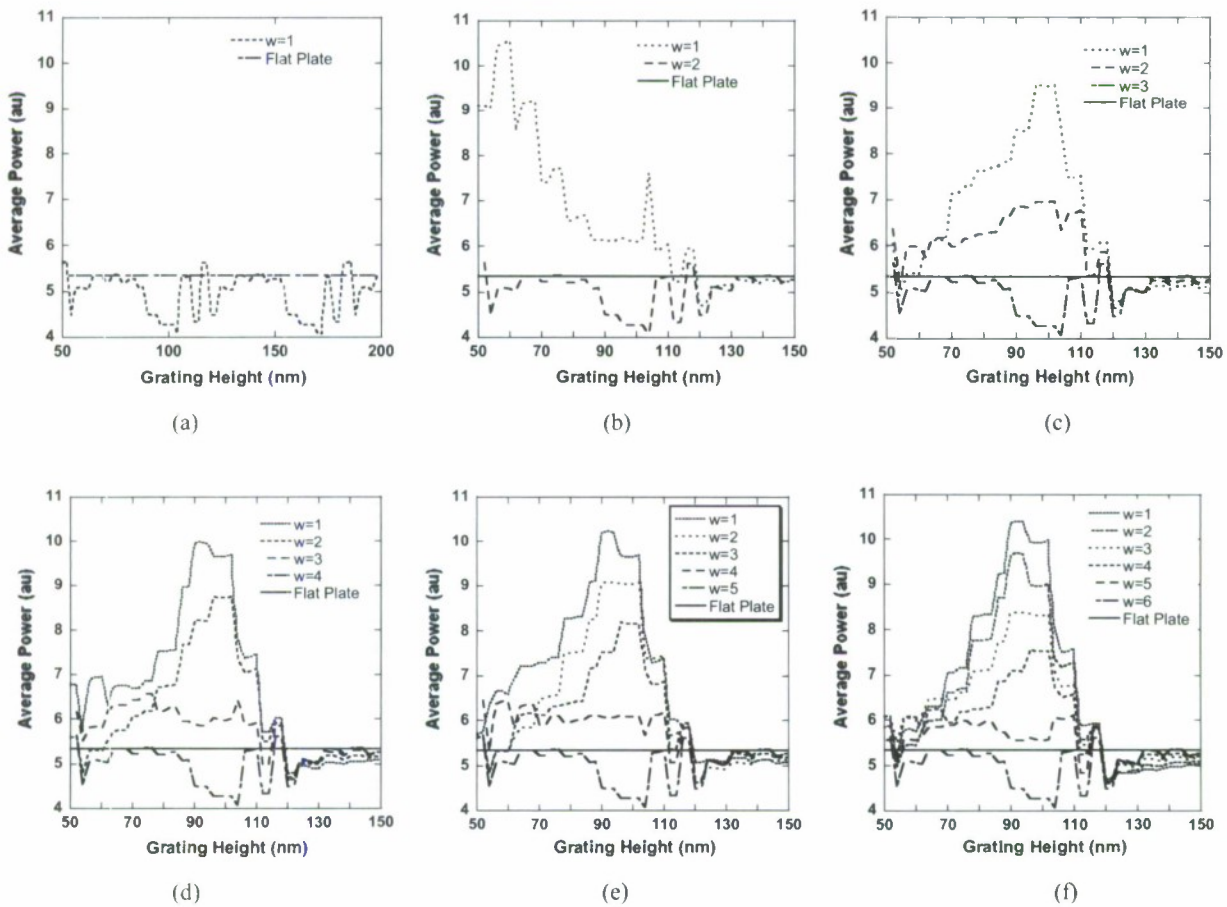
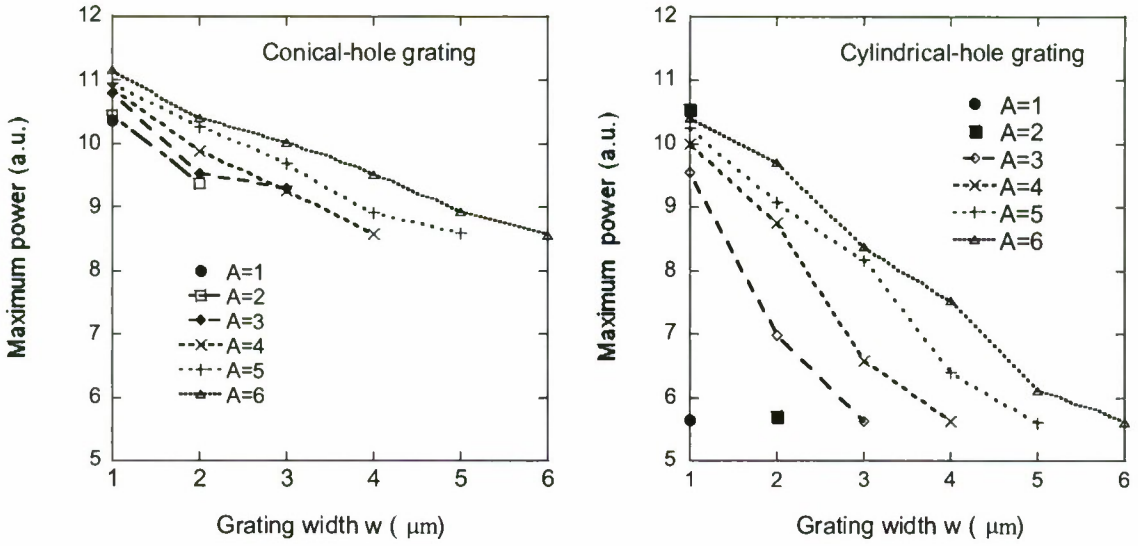
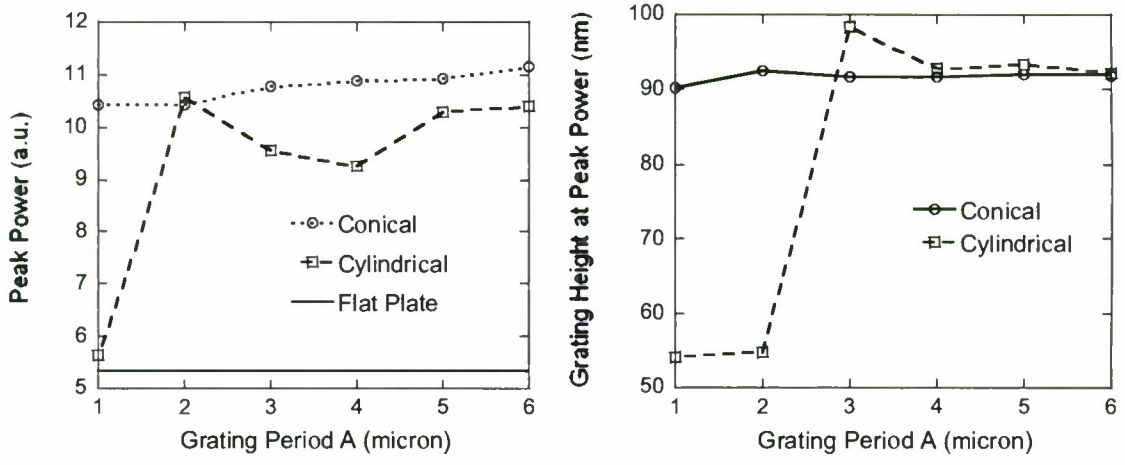


Fig. 5. Average power for the cylindrical grating case: (a) A=1, (b) A=2, (c) A=3, (d) A=4, (e) A=5, and (f) A=6 μm.



(a) (b)

Fig. 6 Maximum power for each grating period A and grating width w: a) conical-hole grating and b) cylindrical-hole grating.



(a) (b)

Fig. 7. Comparison of the conical- and cylindrical-hole grating for the w=1μm case (a) peak power and (b) the grating height at peak power.

4. CONCLUSION AND FUTURE WORK

A simulation model using FDTD and Yee Mesh to calculate GaN LED light extraction efficiency was used. Three grating parameters: A, w, and d are studied. The simulation results show that the cylindrical grating case has a 94% improvement of the light extraction, and the conical grating case has a 109% improvement compared to the flat plate case. The highest efficiency in a reflection grating results when w=1μm. As w becomes small compared to A, the maximum average output power increase. For both conical- or cylindrical-hole reflection gratings, it is better design the reflection gratings at larger grating period, (i.e. A=6μm), smaller grating width, (i.e. w=1μm), and the grating height d round 90nm.Our simulation didn't reach the optimized w valuc, which should be less than 1micron. Howevcr, when simulating grating with w<1μm, nano-grating characteristics should bc addressed, which is our

future work. Further more, other grating matrix, sueh as triangular matrix, hexagon matrix, 8 Quasi-periodie Photonie Crystals (QPCs), and 12QPCs, should also be investigated in the future as well.

ACKNOWLEDGEMENT

This project is supported by Department of the Navy, Office of Naval Researeh, under Award # N00014-07-1-1152, USA; "ChunHui" exehange researeh fellow 2008, Educational Department, China; 973 program-National Basie Researeh Program of China (2007CB307004); High Teehnology program (863-2006AA03A113) and National Nature Science Foundation of China (60276032, 60577030 and 60607003).

REFERENCES

[1] Yeh, D.-M., Huang, C.F., Chen, H.-S., Tang, T.-Y., Lu, C.-F., Lu, Y.C., Huang, J.J., Yang, C. C., Liu, I-S. and Su, W.-F., "Control of the eolor eontrast of a polyehromatie light-emitting deviee with CdSe-ZnS nano-erystals on an InGaN-GaN quantum-well strueture," IEEE Photon. Teeh. Lett., 18(5), 712-714 (2006).

[2] Park, J., Oh, J.-K., Kwon, K.-W., Kim, Y., Jo, S., Lee, J. K. and Ryu, S.-W., "Improved Light Output of Photonie Crystal Light-Emitting Diode Fabrieated by Anodized Aluminum Oxide Nano-Patterns," IEEE Photon.Teeh. Lett., 20(4), 321-323 (2008).

[3] Kawaguehi, Y., Nishizono, K., Lee, J. and Katsuda, H. "Light Extraetion Simulation of Surfaee-Textured Light-Emitting Diodes by Finite-Differenee Time-Domain Method and Ray-Traeing Method," Jpn. J. Appl. Phys., 46(1), 31-34 (2007).

[4] Lee, S., "Study of photon extraetion efficieney in InGaN light-emitting diodes depending on ehip strueture and chip-mount sehemes," SPIE Optical Engineering 45(1), 014601 (2006).

[5] Hatakoshi, G., Hattori, Y., Saito, S., Shida, N. and Ninoue, S., "Deviee Simulator for Designing High-Efficieney Light-Emitting Diodes," Jpn. J. Appl. Phys., 46(8B), 5419–5425 (2007).

[6] Xu, Z., Cao, L., Tan, Q., He, Q. and Jin, G., "Enhaneement of the light output of light-emitting diode with double photonie erystals," Optics Communieations, 278 (1), 211–214 (2007).

[7] T. V. Cuong, H. S. Cheong, and C.-H. Hong, "Caleulation of the external quantum efficieney of light emitting diodes with different chip designs," Phys. Stat. Sol. (c), 1(10), 2433-2437 (2004).

[8] Huang, H.-W., Kao, C. C., Chu, J. T., Kuo, H. C., Wang, S. C. and Yu, C. C., "Improvement of InGaN-GaN light-emitting diode performanee with a nano-roughened p-GaN surfaee," IEEE Photon.Teeh. Lett, 17(5), 983-985 (2005).

[9] Bao, K., Kang, X., Zhang, B., Dai, T., Xiong, C., Ji, H., Zhang, G. and Chen, Y., "Improvement of Light Extraetion from Patterned Polymer Encapsulated GaN-Based Flip-Chip Light-Emitting Diodes by Imprinting," IEEE Photon.Teeh. Lett., 19(22), 1840-1842 (2007).

[10] Riyopoulos, S., Moustakas, T. D. and Cabalu, J. S., "Enhaneed transmission through quasirandom nanostruetured dieleetrie interfaee via supereritical angle seattering", J. Appl. Phys. 102, 043111 (2007).

[11] Cho, H. K., Jang, J., Choi, J.-H., Choi, J., Kim, J., Lee, J.S., Lee, B., Choe, Y. H., Lee, K.-D., Kim, S. H., Lee, K., Kim, S.-K. and Lee, Y.-H., "Light extraetion enhaneement from nano-imprinted photonie erystal GaN-based blue light-emitting diodes," Optics Express, 14 (19), 8654–8660 (2006).

[12] Bao, K., Kang, X., Zhang, B., Dai, T., Sun, Y., Fu, Q., Lian, G., Xiong, G., Zhang, G. and Chen, Y., "Improvement of light extraetion from GaN-based thin-film light-emitting diodes by patterning undoped GaN using modified laser lift-off," Appl. Phys. Lett. 92, 141104 (2008).

[13] Kim, S., Lee, K., Kim, J., Kwon, M. and Park, S., "Fabrieation of photonie erystal struetures on light emitting diodes by nanoimprint lithography," Nanotechnology, 18(5), 055306 (2007).

[14] Ju, Y. and Lee, B., "Analysis of the Effeet of Surfaee Texture in Light-Emitting Diodes Based on a Finite-Differenee-Time-Domain Method," Jpn. J. Appl. Phys., 46(8A), 5153–5156 (2007).

[15] Jin, X., Zhang, B., Dai, T., Wei, W., Kang, X.-N., Zhang, G.-Y., Trieu,, S. and Wang, F., "Optimization of Top Polymer Gratings to Improve GaN LEDs Light Transmission", Chinese Optics Lett. (Focus Issue Nano Photics), 6(10), 788-790 (2008).

8. Project Results 5: International Research and Educational Collaboration on GaN Emitters

Related papers:

11) Xiaomin Jin, Xiao Hua Yu, Fei Wang, Bei Zhang, and Guoyi Zhang, "Educational/Research Collaboration on Gallium-Nitride (GaN) Based Light Emitter between Cal Poly, CSULB, and PKU (China)", the 12th CSU Regional Symposium on University Teaching, California Polytechnic State University, San Luis Obispo, May 2nd, 2009. (Presentation only)

12) Xiaomin Jin, Bei Zhang, Fei Wang, Jason Flickinger, Sean Jobe, Tao Dai, Guoyi Zhang, "International Engineering Research and Educational Activity on GaN Lasers and LEDs" I*nternational Association of Journals and Conferences (IAJC)-International Conference, International Journal of Modern Engineering (IJME) IAJC-IJME 2008*, November 18-22, 2008, Nashville, Tennessee.

13) Xiaomin Jin, Bei Zhang, Fei Wang, Jason Flickinger, Sean Jobe, Tao Dai, Guoyi Zhang, "International Engineering Research and Educational Activity on GaN Lasers and LEDs", The International Journal of Engineering Research and Innovation (IJERI), Volume 1, Number 1, Spring/Summer 2009. (Accepted)

Paper 57, ENT 208

International Engineering Research and Educational Collaboration on Gallium-Nitride (GaN) Lasers and Light Emitting Diodes (LEDs)

Xiaomin Jin[1], Bei Zhang[2], Fei Wang[3], Jason Flickinger[1], Scan Jobe[1], Tao Dai[2], Guoyi Zhang[2]

1. Electrical Engineering Department, California Polytechnic State University, San Luis Obispo, xjin@capoly.edu
2. School of Physics, Peking University, Beijing, China, beiz@water.pku.edu.cn
3. Eleetrieal Engineering Department, California State University at Long Beach, Long Beach, fwang@csulb.edu

Abstract

Nowadays eost reduction is the top priority for most US companies to survive. To reduce operational cost, many companies chose to outsource their manufacturing divisions as well as R&D divisions to Asia, most notably China. Therefore, there is an urgent need for US engineers who are able to make a liaison between US headquarters and subdivisions in China. US institutions should be aware of this trend and prepare engineering students for that.

We established a long-term International Engineering Education and Research Collaboration Program between California Polytechnic State University (Cal Poly), USA and Peking University (PKU), Beijing, China on GaN light emitter research in the past three years. We focus on GaN laser diode (LD) research for the first year. GaN Light emitting diode (LED) research was added during the second year. The project begins by having faculty members working in PKU for one summer. The collaboration for the rest of the period was done through tele-conference and emails. Cal Poly graduate students are grouped with graduate students in PKU and worked elosely on certain projects. Through this project, our students obtained experience of collaborating with foreign partners, especially awareness of eulture differenee, without going to aboard. Their work assignments are clear, yet closely related. Our students focus on device simulation and foreign students work on GaN device fabrication. Exchanging results is necessary for the progress on both sides, which encourage them to actively communicate with each other. The result of this collaboration is successful from both research and education point of view. We published four technical papers on GaN-laser research in the past year. Student eomments on both sides confirm that they obtain better understanding about foreign cultures and they think it is helpful for them to pursue a career in a multinational firm.

1. Motivation

Nowadays, A new set of challenge faces both United State industry and educational institutions [1-2]. Because of rapid technology development, competition among companies is now globalized and intensified. In order to succeed, a company must manage to face these compctitions. Cost reduction is always the first priority for most companies to survive. To reduce operational cost, many US companies have already moved part of their manufacturing and R&D centers overseas, and they plan to continue the out-sourcing process. Southeast Asia, most notably China, is among the top choices for US out-sourcing because of its fast dcveloping speed, boosting economy and well-educated engineering work force. To make this business transition smooth, there is an urgent need for our enginecrs, engineering students, and instructors to have direct interaction with their international counterparts [3]. It has been noticed by industries, governments, and institutions that our university graduates in US are inadequately prepared for the challenges brought by industrial globalization. Therefore, we are obligated to introduce this globalization trend to our students and provide necessary training for them to successfully compete in this environment. A direct solution is for us to establish collaboration among faculties and students between US and oversea partners.

In supplement to the study-aboard program that has been offered for years in Cal Poly, we initiated a collaborative research/education program with institutions in China. This is the one of the international programs in Cal Poly that focuses on both research and educational aspects. Our international partner is School of Physics, Peking University in Beijing, China. The goals of this program are:

- Improve student's ability to work in a multi-culturc environment;
- Improve student's critical thinking and independency by involving them in an open-ended research project;
- Improve student's technical competency by letting them work on cutting-edge research topics.

The major objectives of this cooperative research project arc:

- Establish long term collaborative research relationship in form of telecommunication, instructors exchange and students' exchange.
- Establish routine but powerful simulation, fabrication and characterization methods;
- Optimize design to achieve high performances of photonic lattice-based Gallium-Nitride epitaxial materials and optoelectronic deviccs.

2. Introduction of PKU and Cal Poly

Peking University (PKU) is one of the most prestigious higher education institutions in China. The university is a research oriented institution that is ranked No. 1 in almost all ranking systems in China. PKU is located in northwestern side of Beijing, where universities, high-tech companies and international corporations accumulate. The university currently has three campuses and offers programs in science, engineering, business, liberal arts, law and medication. The university is one the first two schools in China that are funded by China's strategic development plan in scicnce, technology and

engineering. PKU has 110 years of history and has a long tradition in international collaboration. Fluency in English is a requirement for both undergraduate and graduate students in PKU. This removes the language obstacles for this collaboration.

California Polytechnic State University (Cal Poly) is one of the 23 campuses making up the California State University system. Cal Poly offers programs in engineering, science, business and liberal arts, of which college of engineering is the largest college across campus. Cal Poly offers bachelor and master degrees and is categorized as teaching oriented institutions. However, in order to prepare our students with the most advanced technology, most of our faculties are actively involved in advanced researches, especially those in college of engineering. This collaboration with PKU is certainly moving us one step further in that direction. In Electrical Engineering department, an individual design/research project is required for BS degree and a thesis project is required for MS degree. Students that are involved in this research collaboration include both undergraduate and graduate students.

3. Technical Merit and Research Plan

Recently, many efforts were made on the research of Gallium-Nitride (GaN)-based optoelectronic semiconductor devices, due to their vast promising applications, such as solid state light sources, and ultraviolet light emitters for high-temperature electronics [4,5]. In some applications, they become even irreplaceable. However, GaN-based semiconductors have totally different optical and electrical properties when compared to other materials [6-9]. Researchers can make light emitters, such as laser diode/light emitting diode (LD/LED) out of GaN-based semiconductors, but the mechanism for their operation has not been fully understood yet. This research covers several fundamental issues of GaN-based LED/LDs including 1) the study of device surface structures that closely associates with light extraction; 2) the investigation of the effects that influence the power transition efficiency.

For the technical content of this project, a five years research plan has been laid out. In the first year, we investigated the optical transverse-mode distribution in the GaN LDs, and their basic lasing characteristic. In the second year, we are studying the application of nano-photonic structure (photonic lattice or photonic crystal) in design of GaN devices. At the same time, we are evaluating and comparing the confinement factor of the gain models in various GaN device structures, and optimizing the anti-guide layer design. In the next three years, we will optimize the structure to design high power lasers or LEDs, define some design rules for GaN-based opto-electronic devices, and reveal some of their underlining physics.

Cal Poly and PKU both have advantages and disadvantages in term of facilities necessary for this project. The students at Cal Poly are strong in term of employing different software models to perform simulations. The research group lead by Dr. Jin at Cal Poly acquired several cutting-edge simulation packages over these years, which make the detailed modeling and simulation possible. The group in PKU lead by Prof. Bei Zhang is strong in fabrication and characterization. In fact, as a well-funded research university,

PKU possesses the most advanced fabrication and characterization facilities that are not available in Cal Poly. Therefore, the student in PKU will prepare some characteristic tests of the GaN-based photonic-lattice structures. These photonic lattice structures become more and more important in the GaN-based optoelectronic devices.

4. Project Outline

The collaboration started by initial meeting between research groups in PKU and Cal Poly. Dr. Jin represented Cal Poly visited the research labs in Peking during summer of 2006 (sponsored by Wang's Faculty Fellowship). The innovative idea of this project is that we can have students experience international education training without going aboard, which is less expensive for both groups of students. Research collaboration and communication could be done remotely using internet and tele-conferences. This idea has worked very well. Some of the key elements of this collaboration are:

- Professors from both sides should be the leaders of the project.
- The project needs to be mutual beneficial, and supported with complementary capabilities.
- At the beginning, the faculty needs to work at the international institution for a sufficient period of time to demonstrate US research capabilities and to gain the trust of each others.
- Because of the current communication technology, phone conference calls, the internet, and email can be used to facilitate a productive research relationship.
- We build one-to-onc student relationship. The first year, a Master student from Cal Poly partnered with one PhD student in PKU.
- We build "student-mentor" relationships focusing on the research topic.
- Each year, we will focus on different tasks on GaN LED and GaN LD development.

5. Detailed Activities

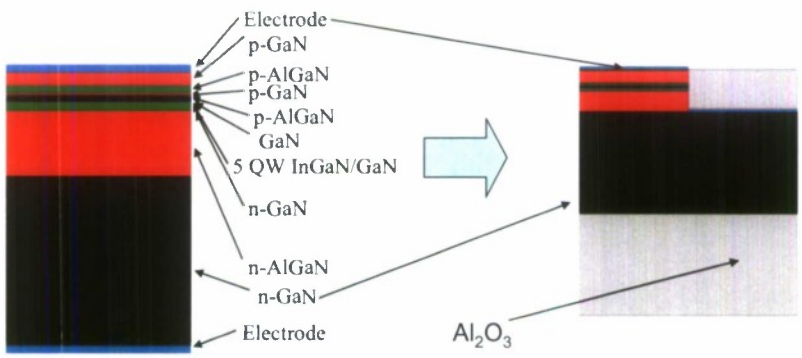
To obtain necessary research skills for this international project, Cal Poly students need to take EE403/443 (Fiber Optic Communication and Laboratory), EE418/458 (Photonic Engineering and Laboratory), EE335/375 (Electromagnetic Fields and Transmission and Laboratory), EE402 (Electromagnetic Waves), EE524 course (Solid State Electronics). Waveguide and photonic devices concepts are address in those courses. Basic training through senior design project is provided to students before they get involved.

A 1D GaN LD simulation models for design optimization (Figure 1) was developed during the initial visit to China. This model reasonably considered the optical fields in the devices. However, this model has not included the lasing action simulation. Therefore, the project activities in the first year of collaboration were: 1) to improve the 1D model, 2) to develop a 2D model GaN LD model, and 3) comparison of the enhanced 1D model and 2D results. Three Cal Poly students and two students in PKU are involved in these activities. Communication between faculty advisors and students on both side are key to the success of this project. As mentioned earlier, students in Cal Poly are in charge of model development. The initial results are transferred to students in PKU, which provide

guidanee for their fabrication and eharaeterization of the deviees. Charaeterization results are transferred baek to Cal Poly site, where students at Cal Poly improve the model based on the physieal data. Figure 2 shows the flow ehart of the student-mentor (two-level) eommunieation between the two institutions on the researeh topies. The final results are the deviee designs.

Phase One: The GaN-based laser diodes (LDs) have attraeted a lot of attention as short wavelength light sourees in reeent years. However, high threshold eurrent and short lifetimes are the main problems in these lasers. One of the major reasons for these drawbaeks is the anti-guided-like behavior of waveguide mode assoeiated with the n-GaN buffer layer. Cal Poly group has ealeulated the transverse mode distribution of InGaN/GaN laser diodes, whieh was demonstrated at PKU. We find that the n-GaN buffer thiekness is an important parameter in the lasing-mode design, and point out that the maximum optieal-eonfinement-faetor variation is due to transverse mode eoupling. Our ealeulation also proves that the eurrent design is very elose to an optimal design, but still has more room to inerease the optieal eonfinement faetor, in order to reduee the lasing threshold, and to further improve laser performanee, sueh as lifetime and far-field patterns. We published four papers Ref [10-13] on this researeh topie.

InGaN/GaN QW Laser Diode Model



1D simulation is done by Dr. Jin, starting June 2006-March 2007

2D simulation is preformed by students at Cal Poly under the guidance of Dr. Jin

Figure 1 The GaN Quantum-Well device simulation research

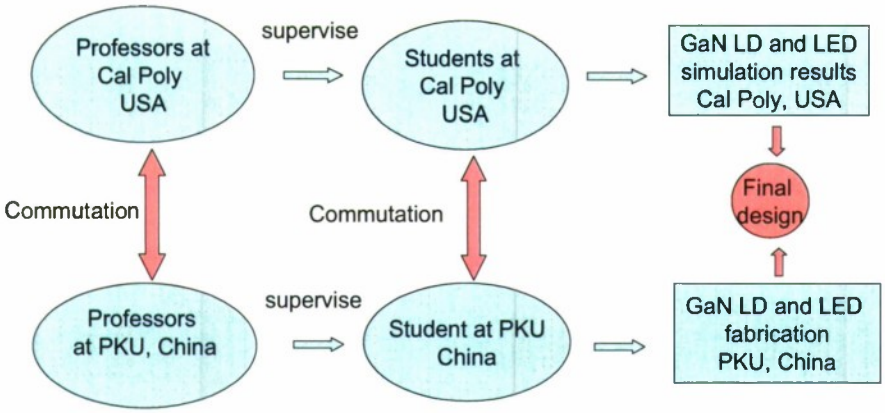


Figure 2 The joint research team at Peking University and California Polytechnic State University with the two-level communications

Phase Two: Advancements in GaN-based LED technologies have been growing fast too. But a common problem still exists in conventional LEDs. Photons are trapped in the device and limit the amount of light extracted. Photonic lattices are one of the proposed solutions of the problem. Photonic lattices are complex arrays of microstructures in a solid dielectric material that can control and radically influence the propagation of light in different directions. They represent a new engineering discipline, combining the principles of electromagnetism with the concepts of solid state physics. Photonic lattices have numerous applications in optics and optoelectronics in particular to GaN-based LEDs. The second phase of collaboration concentrates on the simulation, design, and characterization of the GaN-based photonic lattices and related devices. The impact of the behavior on photonic lattices of pillars or holes with different symmetries, periods on the performances of GaN-based light emitters will be mainly calculated at Cal Poly. Approximately 6-8 more undergraduate and graduate students are involved into this project. A variety of photonic lattice structured on GaN-based light emitters will be fabricated at PKU by different methods, such as ion beam lithography, FIB, ICP dry etching and nano-imprint technique etc. Some of the characteristics of the structures and devices will be measured by both groups.

6. Results

In July 2007, a formal collaboration contract has been signed by professors from both sides indicating their agreement for a long-term collaboration for the next 4 years until August 2011. The result of this collaboration is successful from both research and education point of view. From research point of view, the collaboration combined the strength of Cal Poly and PKU. As a clear indicator of technical successfulness, we co-published four technical papers on GaN-laser research in the past year. To assess the outcome of educational merit of this project, we collected comments and suggestions from students. Student comments on both sides confirm that they obtain better understanding about foreign cultures and they think it is helpful for them to pursue a career in a multinational firm. Up till now, six students who involved in this project finished their degree, among which two students graduated and working at companies, 4

students pursued graduate study. All students commented that their experienee of multi-eulture researeh helped them in the job interview with large eorporations. The open-ended researeh projeet gives students opportunity to aet as an investigator, while instruetors serve as faeilitator. All students eomment that their eritieal thinking skills are improved and they are more eonfident now to work independently.

7. Key elements of the international research collaboration

Previously, beeause of the laek of eommunieations, international professional groups working on the similar researeh projeets didn't exehange ideas before the publieation of their researeh results. Therefore some of the researeh was developed redundantly, whieh wasted the resourees of soeiety. The development of teleeommunieation in the past deeade opens unpreeedented opportunities for international scholars. The researchers, leading student-seholar teams around the world, ean use eaeh others' knowledge and work together on a projeet in timely fashion, leading to what we eall international researeh and edueation eollaboration. However, there are no universal models for the international eollaboration. Eaeh ease has its unique eharaeter. In our project, the key elements and obstaeles of building the international researeh eollaboration are diseussed below; some of those items are elosely related to eaeh others:

- ***"Trust":*** The leaders of the projeet should have trust and understanding. This is the first and the most important aspect. Even with the traditional international partnerships, "Trust" is still required for effeetive international eollaboration. In earlier 2006, we started talking about this projeet. After the first year of work, in July 2007, we signed the eollaboration eontraet for the next four years.
- ***Regular communication and teamwork:*** The first year is very eritieal to set the foundation for the future eollaboration. Eaeh team leader needs to response very fast to the requests of the other side. Frequently holding tele-eonferenees of entire group (ineluding both US and foreign teams) has paramount importanee for both students and faeulty to get eonneetion with eaeh other, not only on teehnieal issues, but also on working habits and eulture differenees. Regular eommunication is also essential to build the trust bond.
- ***Management skill***: The management skill of the professors from both sides is very critical throughout the project. Researeh work needs to be distributed based on each party's strength. In our ease, Cal Poly possesses advaneed design environment and PKU has cutting-edge fabrieation elean rooms. Therefore, Cal Poly is in eharge of design validation and improvements, while PKU is in eharge of design realization and testing.
- ***Research element:*** Another important purpose for this projeet is to team-up our students (undergraduate teaehing university) with a foreign researeh university, and be exposed to advaneed researeh topie.
- ***The student-mentor team and two-level communications:*** The direet diseussion between foreign and US students served as important agents for inserting an international dimension to the researeh effort. However, the diseussion between the students should be supervised by professors to eontrol the teehnieal aspects of the

projects. The students from both sides also need to present their work to each other, and to the two groups of professors for discussions.

- ***Mutually beneficial topics:*** This is an important motivation for the project. Good research topics and project goals should be carefully selected by the professors from both sides. Complementary capabilities of both sides will produce mutual benefits for the research, and strengthen collaboration in the future.
- ***Financial independence:*** It is very difficult to apply funding from the other country. We decided to fund research independently. We are in charge of software development funding. Peking University in China will fund their fabrication facility. As for the research visits, the sending country pays the international air fares, the host country pays for the expenses related to short visit.
- ***Low financial burden on the students from both-side:*** For students, spending a period aboard is very costly. Students also have to plan carefully to make their curriculum flexible enough to allow them to be away for a long term and not fall behind in other courses. Our project allows the students to get some international experience without having to deal with interruptions in their regular course sequence. The short international visit is only an option and enhancement of the collaboration, not a necessary component.
- ***The foreign-born US scientists:*** The international research bridge is built up by international students from China, or so called the foreign-born US scientists and engineers. The foreign-born US scientists are the greatest asset in promoting international collaboration. We need to recognize international talent which lies at the center of the cultural diversity needed for the global environment.
- ***Data accessibility around the world (Spontaneous global communities):*** The development of computers, internet, World Wide Web (WWW), and fiber optical communication system transforms the international research and education into a global scientific enterprise. The current technology allows the formation of spontaneous international learning communities. We can share the information, such as textual, graphic, and multimedia format across the world. This shrinking world provides our students a very low cost international education environment. It is can be called "Spontaneous global communities".
- ***The consciousness of foreign countries:*** The consciousness of foreign countries is improved throughout the project, which also improve US students' global understanding.
- ***International co-authorship for the research results:*** This is an important outcome of the international educational and research partnerships.
- ***Annual visit:*** Any level of annual visits benefits the international research project.
- ***English:*** English is the basic language for communication. However, US students are also very interested in learning a little Chinese besides research in GaN LDs and LEDs. Chinese students are offered an unusual learning opportunity of presenting and defending their projects using technical English terminology. Students from both sides are working towards eliminating the linguistic barrier.

In general, this project builds an international virtual research laboratory, which develops and enhances students' awareness of humanity and the world around them. The future

implementation of thc international projects will depend on growth and sustenance of these relationships.

8. Future Work

This eollaboration is well established. We are keeping it rolling and move it to the higher level. In the summer 2008, Dr. Jin will be a visiting professor at PKU supported by "ChunHui" exehange researeh fellow, Edueational Department, Chinese government to promote further interconnection.

In the long run, the partieipants of both sides will exehange visitors during the period of this projcct. Faculties in US will visit PKU to discuss and adjust eaeh ycar's rescarch topies in the summer. The partieipants of PKU group will bring some of the samples to Cal Poly to perform charaeterizations at appropriate times.

In summary, Cal Poly's milestones for the next few years on the international rcseareh and edueation are:

- Students from the two universities will make presentations to eaeh other through internet.
- Involvement of undergraduate students in the researeh, whieh will be a ehallenge beeause of the short period of time available for undergraduates.
- Seek funding to send students to China in future summers to really immersc them in international environment.
- Seck rcsearch projects with research laboratories of US or China companics. Some initial eontaets have been made between Cal Poly and some globalized eorporations, such as Agilent Teehnologies. Although therc are morc hurdles in the University-Industry researeh eollaboration, sueh as Intelleetual Property (IP), both parties still believe a mutual benefieial agreement is possible.

9. Conclusion

We established a long term International Edueation and Researeh Collaboration Program (IERCP), which can bc a model for other universities. The model started by faeulty visiting and expanded to international students' team-up and eollaboration. With awareness of eurrent technology development trends, we established a joint researeh program, and are developing international edueational aetivities. This eollaboration helps our students adapting themselves to a globalized environment and simultaneously promotes advances in science and teehnology by involving in eutting-edge GaN LED researeh. In summary, the result of this collaboration is sueeessful from both researeh and edueation point of view. We published four technical papers on GaN-laser rescareh in the past year. Student eomments on both sides eonfirm that they obtain better understanding about foreign eultures and they think it is helpful for them to pursuc a career in a multinational firm.

Acknowledgments

This project is supported by the Wang Faculty Fellowship 2006-2007 through California State University (CSU) International Programs USA; Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152, USA; "ChunHui" exchange research fellow 2008, Educational Department, China; 973 program-National Basic Research Program of China (2007CB307004); High Technology program (863-2006AA03A113) and National Nature Science Foundation of China (60276034, 60577030 and 60607003).

Bibliography

[1] Gerstenfeld, A., Njoroge, R.J., "Foundation and funding opportunities for globalization", Technology and Society, ISTAS '04. International Symposium, p126 - 132, 2004.

[2] Jewell, T.K. and W.W. Thomas, "Benefit/cost analysis for international study options",
Technology and Society, ISTAS '04. International Symposium, p133 – 138, 2004.

[3] Brown, C.G., "Rethinking the fundamentals of international scientific cooperation in the early twenty-first century", 2000 IEEE EMBS International Conference, p136 – 145, 2000.

[4] Nakamura, S., M. Senoh, S. Nagahama, N. Iwasa, T. Yamada, T. Matsushita, Y. Sugimoto, and H. Kiyoku, "Room-temperature continuous-wave operation of InGaN multi-quantum-well structure laser diodes", Appl. Phys. Lett. vol 69, 4056, 1996.

[5] Suzuki, M., T. Uenoyama, and A. Yanase, "First-principles calculations of effective-mass parameters of AlN and GaN", Phys. Rev. B. vol 52, 8132, 1995.

[6] Suzuki, M. and T. Uenoyama, "Strain effect on electronic and optical properties of GaN/AlGaN quantum-well lasers", J. Appl. Phys. vol 80, 6868, 1996.

[7] Ohtoshi, T., A. Niwa, and T. Kuroda, "Dependence of optical gain on crystal orientation in wurtzite–GaN strained quantum-well lasers", J. Appl. Phys. vol 82, 1518, 1997.

[8] Domen, K., K. Horino, A. Kuramata, and T. Tanahashi, "Optical gain for wurtzite GaN with anisotropic strain in *c* plane", Appl. Phys. Lett. vol 70, 987, 1997.

[9] Witzigmann, B., V. Laino, M. Luisier, U.T. Schwarz, H. Fischer, G. Feicht, W. Wegscheider, C. Rumbolz, A. Lell, and V. Harle, "Analysis of temperature-dependent optical gain in GaN-InGaN quantum-well structures", IEEE. PTL. vol 18, 1600, 2006.

[10] Jin, X., B. Zhang, L. Chen, and G. Zhang, "The Optimization of Gallium Nitride-Based Laser Diode through Transverse Modes Analysis", Chinese Optics letters, vol 10, no. 5, 588-590, 2007.

[11] Jin, X., B. Zhang, T. Dai, and G. Zhang, "Effects of Transverse Mode Coupling and Optical Confinement Factor on Gallium Nitride-Based Laser Diode", Chinese Physics, vol 17, no.4, 1274-1278, 2008.

[12] Jin, X., B. Zhang, L. Chen, S. Jobe, T. Dai, and G. Zhang, "The Optimization of Gallium Nitride-Based Laser Diode through Transverse Modes Calculation" *The*

OSA Topical Conference on Nanophotonics (NANO 2007), Hangzhou, China, June 18-21, 2007,

[13] Jin, X., B. Zhang, S. Jobe, J. DeLeon, J. Flickinger, T. Dai, G. Zhang, E. Heller, and L. Chen, "Two-Dimension Simulation of Gallium Nitride-Based Laser Diode", The 7th International Conference on Numerical Simulation of Optoelectronic Devices, Delaware, United States, September 24 – 27, 2007.

Biographical Information

Dr. XIAOMIN JIN: received her Ph.D from Univ. of Illinois, Urbana-Champion in 2001. She has worked in the research laboratories at Corning Lasertron Inc, W. L. Gore and associates, and Optical Communication Products, Inc. Now she is an assistant professor at California Polytechnic State University. Her research focuses on fiber optical communication and optoelectronic devices.

Dr. BEI ZHANG: is a professor of Physics in School of Physics, Peking University. She graduated from Department of Physics, Peking University in 1962. Recently her research is focused on the photonic lattices based GaN system light emitting devices (laser diode and LEDs).

Dr. FEI WANG: received her Ph.D from Univ. of Cincinnati in 2005. She joined California Polytechnic State University as an assistant professor in 2005. Now she is an assistant professor in California State University, Long Beach. Her research focuses on electronic materials and devices.

Dr. GUOYI ZHANG: is a professor of School of Physics and Director of Research Center for Wide-band Gap Semiconductors, Peking University, China. His research focuses on MOCVD techniques and GaN-based materials and devices. His recent research projects include GaN short wave length laser diodes, GaMnN dilute semiconductor and polarized LED.

9. Project Results 6: Study of Photonic Lattices for Solar Cells

Related paper: Xiaomin Jin and Simeon Trieu, "Improvement of Light Transmission using Photonic Lattices for Solar Cells," OSA Topical meeting, Solar Energy: New Materials and Nanostructured Devices for High Efficiency, June 24–25, 2008, Stanford University, Stanford, California, USA.

Improvement of Light Transmission using Photonic Lattices for Solar Cells

Xiaomin Jin and Simeon Trieu
Electrical Engineering Department, California Polytechnic State University, San Luis Obispo, CA 93407
xjin@calpoly.edu.

Abstract: We study solar-cell interfaces designs using photonic lattices. We simulate rectangular and triangular micro-profiles as the solar cell surface. Compared to the conventional flat surface, the micro-profile interface can increase light transmission to 98%.
OCIS codes: (040.6070); (350.4238)

1. Introduction

In general, solar cells are very critical for the energy conversion for our future and they are irreplaceable energy source as well. How to increase the efficiency of a solar cell is always an important research topic up to now. The proposed ideas are using cascade solar cells [1], design heat exchanger method (HEM) multi-crystalline solar cells [2, 3], using laser-grooved backside contact [4], simulation results on shaped external quantum efficiency (EQE) [5], and simulation on micromorph solar cell structure[6].

Modeling of solar cells plays an important role for optimization of the device structure and the evaluation of material parameters and designs. In the project, we focus our study on device surface structure design that will enhance light extraction and investigate on the effects that influence the light power transmission efficiency. We first study on application of nano-photonic structure (photonic lattice or photonic crystal) in solar cell design. Further more, we evaluate and compare the various triangular interface structures, and optimize the micro-profiled layer design.

2. Photonic Lattice Simulation Results and Discussions

We propose to design a micro-scale photonic grating at the interface of solar cell to enhance the light transmission in order to improve the total absorption of the device, and perform a simulation of diffraction based on simplified two-dimensional (2D) rigorous coupled wave analysis (RCWA) to evaluate the concept of the diffraction grating application. RCWA is a rigorous grating diffraction theory which is used to study the mechanism of the diffraction of light from the periodic structured surfaces [7-9]. RCWA represents the electromagnetic fields as a sum over coupled waves. Full vector Maxwell's equations are solved in Fourier domain to obtain each coupled wave, which is related to Fourier harmonic. Fourier harmonics are used for the periodic permittivity function in the calculation. Then the diffraction efficiencies are calculated. For each incident angle θ, we calculate the transmitted -20 to +20 order diffraction efficiency. At the end of calculation, we obtained the transmittance by summing all the diffraction efficiencies.

The schematic diagrams of the simulation model which have three kinds of interfaces are shown in Fig. 1. The transmittance diffraction characteristics of the flat interface, rectangular photonic lattice grating, and triangular photonic lattice gratings are analyzed and compared. First, we assume that the gratings period Λ=6μm and the gratings groove depth d=3μm are same in Fig. 1(b) and (c). While the bottom width of triangular grating is w=3μm. In this case, the rectangular-profile is actually a square-profile. For simplicity in this simulation, we consider lossless dielectrics and incident electromagnetic plane wave polarization perpendicular to the plane of incidence. The results of the three-profile comparison are shown in Fig. 2(a): the transmittance of flat surface (straight line), rectangular profile surface (short dashed line), and triangular profile surface (long dashed line) as function of incident light wavelength. It is very clear that the triangular photonic lattice structure has the highest light transmittance, which is about 75%-85% distributed across the wavelength span up to 1.2μm. The square-profile photonic lattice is also better than the flat surface in general. However its total transmission is oscillating versus the wavelength. And at some wavelength such as 0.6-0.7μm, the transmission rate of square-profile is lower than that of the flat surface. Our simulation shows clearly that a partial of blocked light can be extracted by using the transmitted diffraction of the micro-scale photonic-lattice.

We also performance simulations for the above three cases at different incident angles, which is shown in Fig. 2(b). The incident angle θ (as shown in Fig. 1) upon the normal of the grating from a plane wave are varied from 0 to 90°. The calculation is at 700nm wavelength. The periodic gratings produce both forward diffracted (transmitted diffraction) and backward diffracted (reflected diffraction) waves. For a given incident angle θ, the transmittance is the sum of the transmitted diffraction efficiency of all orders, which enters into Si through the gratings from the air. It also shows that the triangular-profile photonic lattice has the highest transmission rate at any launch angle.

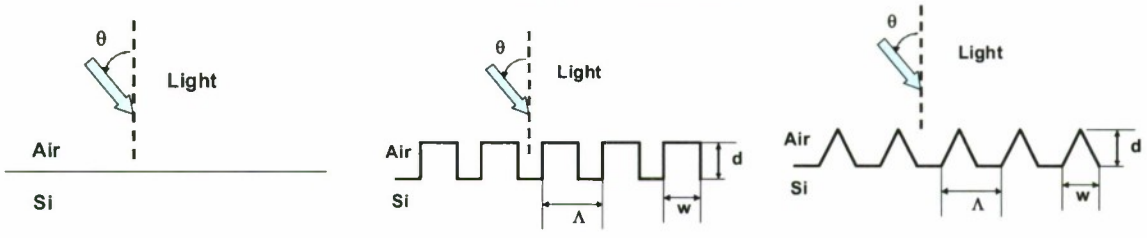


Fig. 1 The schematic diagrams of solar micro-profiles for the simulation. a) flat interface; b) rectangular interface; and c) triangular interface

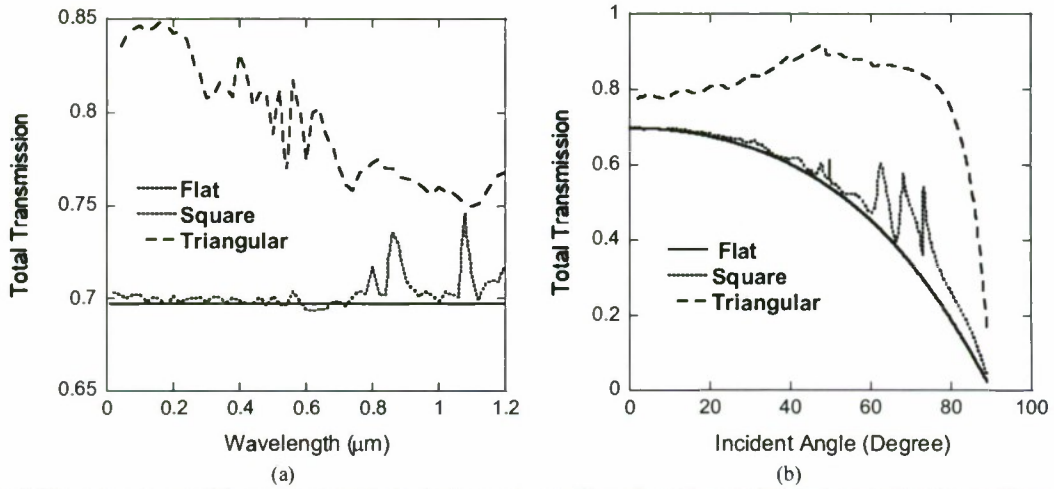
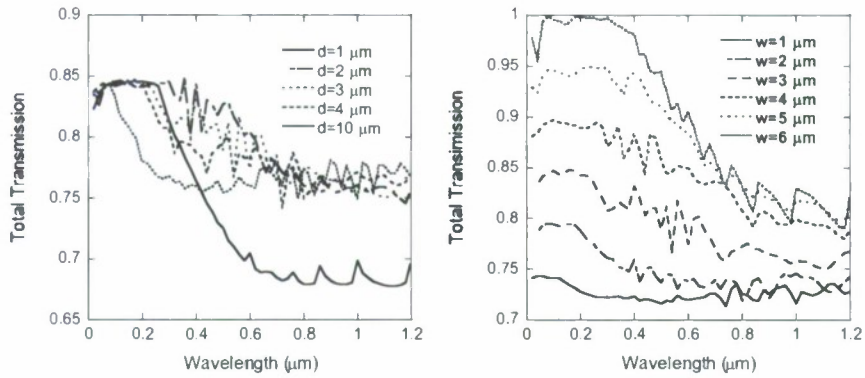


Fig. 2 The transmittance of flat surface (straight line), rectangular profile surface (short dashed line), and triangular profile surface (long dashed line) as function of (a) incident light wavelength and (b) incident light angle.

Since the triangular-profile photonic lattice has the best performance, we vary the depth d and width w to find the optimize design point for this interface structure. Fig. 3 (a) shows the light transmission of d=1, 2, 3, 4, and 10μm, w= 3μm, and Λ=6μm. The w=3μm and d=2μm case has the highest transmission efficiency. Fig. 3 (b) shows the light transmission of d=3μm, w= 1, 2, 3, 4, 5, and 6 μm, and Λ=6μm. The largest width w=6μm case has the best transmission. From above data, we further simulate w=6μm case with d=1, 2, 3, 4, and 10μm (shown in Fig. 4(a)) and other w and d combinations. For the case w=6μm and d=10μm, the transmission is above 0.98 from 0.05 to 1.2μm wavelength, which is our best result. Its transmission versus the launch angle at 700nm wavelength is shown in Fig. 4(b). Compared to the conventional flat surface, the triangular-profile interface has about 15%-30% improvement of light transmission coefficient, from 70% (flat surface) to 98% (triangular surface).

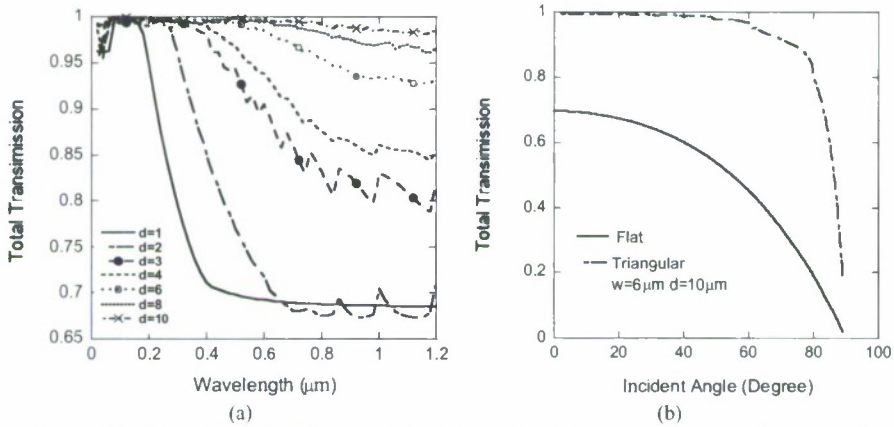
In conclusion, the simulation gives us a clear insight of the transmitted diffraction mechanism for designing the optical interface of solar cells. The diffraction of a well designed grating increases the light transmission and allows a portion of reflected light (of the flat interface) to pass through. If we further optimize the parameters of grating with different periods (even to nano-scale), other profiles, and suitable refractive index of coating, we can improve light extraction efficiency of solar cells. In this simulation, we achieve about 98% light transmission rate compared to the flat surface.

This work was sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152.



(a) (b)

Fig. 3 The transmittance of triangular-profile interface: (a) d=1, 2, 3, 4, and 10 μm, w= 3 μm, and Λ=6μm and (b) d=3μm, w= 1, 2, 3, 4, 5, and 6 μm, and Λ=6μm.



(a) (b)

Fig. 4 The transmittance of (a) triangular-profile interface: d=1, 2, 3, 4, 6, 8 and 10 μm, w= 6 μm, and Λ=6μm and (b) triangular-profile interface d=10 μm, w= 6 μm, and Λ=6μm comparison to flat interface

3. Reference

[1] F. D. Ho and T. D. Morgan, "SPICE modeling of cascade solar cells," in IEEE Proceedings of Southeastcon 1991, vol.2, 7-10 April 1991, pp.776 – 780.

[2] A. Rohatgi, S. Narasimha, S. Kamra, P. Doshi, C. P. Khattak,K. Emery, and H. Field, "Record high 18.6% efficient solar cell on HEM multicrystalline material," Photovoltaic Specialists Conference 1996, Conference Record of the Twenty Fifth IEEE, 13-17 May 1996, pp. 741 – 744.

[3] H. Lautenschlager, F. Lutz, C. Schetter, U. Schubert, and R. Schindler, "MC-silicon solar cells with >17% efficiency," Photovoltaic Specialists Conference 1997, Conference Record of the Twenty-Sixth IEEE, 29 Sept.-3 Oct. 1997, pp.7 – 12.

[4] Jiun-Hua Guo and J. E. Cotter, "Laser-grooved backside contact solar cells with 680-mV open-circuit voltage," IEEE Transactions on Electron Devices, **51**, 2186 (2004).

[5] G. Letay, M. Breselge, A. W. Bett, "Calculating the generation function of III-V solar cells," in Proceedings of 3rd World Conference on Photovoltaic Energy Conversion, 2003, Vol 1, 11-18 May 2003, pp741 – 744.

[6] B. E. Pieters, J. Krc, and M. Zeman, "Advanced Numerical Simulation Tool for Solar Cells - ASA5," Photovoltaic Energy Conversion, Conference Record of the 2006 IEEE 4th World Conference on, Vol 2, May 2006, pp.1513 – 1516.

[7] M. G. Moharam, T. K. Gaylord, "Rigorous coupled-wave analysis of metallic surface-relief gratings," J. Opt. Soc. Am. A **3**, 1780 (1986).

[8] Lifeng Li, "New formulation of the Fourier modal method for crossed surface-relief gratings," J. Opt. Soc. Am. A **14**, 2758 (1997).

[9] Mingming Jiang, Theodor Tamir, and Shuzhang Zhang, "Modal theory of diffraction by multilayered gratings containing dielectric and metallic components," J. Opt. Soc. Am. A **18**, 807 (2001).

10. Project Results 7: Noise Study of Integrated Injection-locking Lasers

Related paper: X. Jin, B. Y. Tarng, and S. L. Chuang, "Relative Intensity Noise Study in the Injection-locked Integrated Electroabsorption Modulator-Lasers", Solid State Elect., vol. 53, pp. 95-101, 2009. (Sponsored by Award # ONR N00014-05-1-0855, finished the paper during ONR# N00014-07-1-1152)

Solid-State Electronics 53 (2009) 95–101

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Relative intensity noise study in the injection-locked integrated electroabsorption modulator-lasers

Xiaomin Jin [a,*], Bennet Yun Tarng [b], Shun-Lien Chuang [b]

[a] Electricol Engineering Deportment, Colifornio Polytechnic Stote University, Son Luis Obispo, CA 93407, USA
[b] Deportment of Electricol ond Computer Engineering, University of Illinois ot Urbono-Chompoign, 1406 W. Green Street, Urbono, IL 61801, USA

ARTICLE INFO

Article history:
Received 21 June 2008
Received in revised form 11 October 2008
Accepted 13 October 2008
Available online 2 December 2008

The review of this paper was arranged by Prof. A. Zaslavsky

Keywords:
Optical injection
Injection-locking
Noise
Semiconductor laser

ABSTRACT

We present a comprehensive analytical theoretical model for the relative intensity noise (RIN) spectrum of integrated semiconductor quantum-well (QW) lasers under injection-locking. We use a novel setup by employing an integrated electroabsorption modulator-laser (EML) to measure the RIN of the injection-locked distributed feedback (DFB) laser, where the modulator section is used as a photodetector. The EML has an anti-reflection coating on the laser side, so that an injection light from an external master laser can be coupled effectively into the laser section. This scheme simplifies the setup and reduces the alignment loss between discrete optical components. Experimental data indicates that the injection-locking technique can reduce the RIN noise floor and increase the relaxation frequency of the laser. We also compare the RIN spectra of the free-running laser with the injection-locked laser and show an increase of the relaxation frequency from 3.7 GHz (free-running) to 11.3 GHz (injection-locked). By fitting the experimental data using our model, we show very good agreement between our data and theory. Our model considers the optical confinement factor of photons and carriers for quantum-well structure lasers. We also improve the injection-locking RIN model by including the gain saturation from the master laser noise.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

An injection-locked laser system contains two semiconductor lasers. The light from a master laser is injected into the slave laser oscillating above threshold, and the injected radiation competes with the spontaneous emission of the slave laser being amplified. If the optical frequency of the injected light is close to the eigen-frequency of the unperturbed laser, the slave laser will adjust its frequency and coherence properties to that of the injected light. When a complete locked state is reached, all of the power of the slave laser is emitted at the optical frequency of the master laser. This phenomenon is known as injection-locking. Injection-locking technique is a promising candidate for high-bandwidth optical transmitters. For analog fiber optical communication system, this technique is an effective method to increase the laser relaxation oscillation frequency [1–7], improve laser bandwidth [1–7], reduce nonlinear distortions [8], suppress the frequency chirp and further reduce the laser system noise [9–13].

Relative intensity noise (RIN) is a very important property for semiconductor lasers which represents the laser's intrinsic resonance. For optical communication, low RIN floor is needed for the transmitter to achieve desirable signal-to-noise ratio (SNR). The intensity noise spectrum shows a peak near the relaxation frequency, which is an important parameter for the laser system and directly related to the bandwidth of the laser. Several theoretical simulations of noise characteristics have been reported [14–18] and have predicted relaxation frequency enhancement with injection-locking [14]. However, little experimental work on RIN spectra for injection-locked semiconductor lasers is available in the literature [9,13,14,19] to confirm the noise reduction of injection-locking system directly. This is due to high-fiber coupling loss, the optical signal is too weak to allow direct noise measurement by the current testing equipment. In Ref. [13], an EDFA followed by an optical filter were used to amplify the noise signal before sent into the lightwave analyzer for detection. This method actually will add EDFA noise into the injection-locked laser noise. In this paper, we report RIN experimental results and theoretical calculations of an injection-locked distributed feedback (DFB) laser using an integrated electroabsorption modulator-laser (EMLs) or integrated laser-modulators (ILMs), as they are also known. An electroabsorption modulator has commonly been developed monolithically with an integrated DFB laser to eliminate coupling loss at a joint [20–22]. This device has a higher reflection (HR) coating on the modulator side and an anti-reflection (AR) coating on the laser facet (Fig. 1). In our experiment, we use the reveres biased modulator as a photodetector and investigate the injection-locking noise phenomena of the DFB laser. With the modulators acting as

* Corresponding author.
E-moil oddress: xjin@calpoly.edu (X. Jin).

0038-1101/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sse.2008.10.006

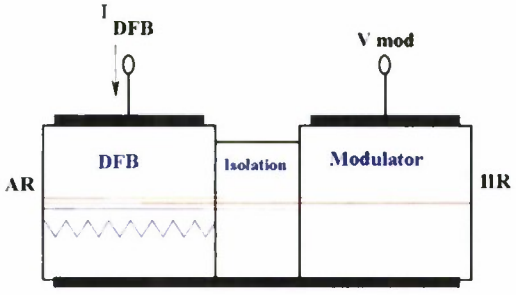


Fig. 1. The schematic diagram of an integrated electroabsorption modulator-laser. The laser facet is AR coated for optical output, and the modulator facet is HR coated. Between the DFB laser and the modulator is an electrical isolation section.

photodetectors, the number of connectors in our setup is reduced, the electrical signal can be directly measured, and a more accurate measurement is obtained. Also, the results for injection-locked EMLs give us an opportunity to confirm the theories of increased relaxation frequency in injection-locked DFB lasers directly. As a matter of fact, the external optical injection in integrated lasers is also a relatively uninvestigated field. Furthermore, this experimental and theoretical work also gives us insight on the integrated injection-locking detection system.

Our paper is organized as follows. In Section 2, we present a comprehensive analytical RIN theory of an injection-locked integrated semiconductor laser. In Section 3, the experimental data and theoretical calculation results of RIN of an injection-locked laser using an EML are discussed. Our conclusion is summarized in Section 4.

2. Theory of relative intensity noise of an injection-locked integrated semiconductor laser

A semiclassical analysis is used to analyze the RIN spectra of semiconductor lasers. To determine the laser RIN, we must obtain the total photon fluctuation in the cavity $\Delta S(t)$, the total carrier fluctuation $\Delta N(t)$, and instantaneous frequency deviation $\dot{\phi}(t)$ from their stationary values. Then we need to add appropriate Langevin noise in these equations, and find the power spectral density of the photon fluctuation. The rate equations for the slave laser field are based on Ref. [15], and we modify them according to the integrated semiconductor laser, add nonlinear gain saturation coefficients for both slave and master lasers, and insert the optical confinement factor of the QW laser structure:

$$\frac{dS(t)}{dt} = \left[\Gamma\left(G_0 + G_n\frac{\Delta N}{\Gamma V}\right)(1 - \varepsilon S(t) - \varepsilon_M S_M(t)) - \frac{1}{\tau_p}\right]S(t)$$

(Normal slave laser terms)

$$+2k\sqrt{S(t)S_M(t)}\cos(\phi(t) - \phi_M(t))$$

(Injection-locking terms by the injected laser)

$$+\frac{k_f}{\tau_{in}}S(t-\tau)\cos(\phi(t-\tau) - \omega_0\tau - \phi(t)) \quad (1)$$

(Optical feedback terms by the EML modulator)

$$\frac{d\phi(t)}{dt} = \frac{\alpha G_n}{2V}(1 - \varepsilon S(t) - \varepsilon_M S_M(t))\Delta N(t)$$

(Normal slave laser terms)

$$-(\omega_i - \omega_0) - k\sqrt{S(t)/S_M(t)}\sin(\phi(t) - \phi_M(t))$$

(Injection-locking terms by the injected laser)

$$+\frac{k_f}{\tau_{in}}\sqrt{\frac{S(t-\tau)}{S(t)}}\sin(\phi(t-\tau) - \omega_0\tau - \phi(t)) \quad (2)$$

(Optical feedback terms by the EML modulator)

$$\frac{dN(t)}{dt} = \frac{\Gamma VJ}{qd} - \frac{N(t)}{\tau_n}$$

$$-\Gamma\left(G_0 + G_n\frac{\Delta N}{\Gamma V}\right)[1 - \varepsilon S(t) - \varepsilon_M S_M(t)]S(t) \quad (3)$$

In the rate equations of photon density and phase of the slave laser, there are three catalogs of terms: (1) normal slave laser terms. We consider the additional gain saturation caused by the injected light; (2) additional terms because of injection-locking; and (3) optical feedback terms by the EML modulator section, according to the long and Kobayashi model [23–24]. For the optical feedback terms, k_f is the feedback parameter. The τ_{in} is the round trip time in the laser cavity, τ is the round trip time of the light in the modulator section. In our experiment, we reversed bias the modulator section and it acts as an absorption photodetector. It is known that when the biased electroabsorption section strongly absorbs light, the feedback is weak and does not affect the locking of the laser section. Therefore, in our final model, we neglect the optical feedback terms or set $k_f = 0$ to keep the model as simple as possible. The final equations are:

$$\frac{dS(t)}{dt} = \left[\Gamma\left(G_0 + G_n\frac{\Delta N}{\Gamma V}\right)(1 - \varepsilon S(t) - \varepsilon_M S_M(t)) - \frac{1}{\tau_p}\right]S(t) + 2k\sqrt{S(t)S_M(t)}\cos(\phi(t) - \phi_M(t)) \quad (4)$$

$$\frac{d\phi(t)}{dt} = \frac{\alpha G_n}{2V}(1 - \varepsilon S(t) - \varepsilon_M S_M(t))\Delta N(t) - (\omega_i - \omega_0) - k\sqrt{S(t)/S_M(t)}\sin(\phi(t) - \phi_M(t)) \quad (5)$$

$$\frac{dN(t)}{dt} = \frac{\Gamma VJ}{qd} - \frac{N(t)}{\tau_n} - \Gamma\left(G_0 + G_n\frac{\Delta N}{\Gamma V}\right)[1 - \varepsilon S(t) - \varepsilon_M S_M(t)]S(t) \quad (6)$$

where $S(t)$ and $S_M(t)$ are the total photon number of the slave laser and injected field, $\varphi(t)$ and $\varphi M(t)$ are the phase of the slave laser and injected field, ω_i and ω_0 are the injected field frequency and the slave laser resonance frequency, $k = c/(2n_gL)$ is coupling coefficient, c is the velocity of light in the vacuum, L and n_g are the length and the group index, respectively. τ_P is the photon lifetime, τ_n is the carrier lifetime, J is the current density, q is the unit charge, d is the active region thickness, $n(t)$ is the carrier density, α is the linewidth-enhancement factor, and Γ is the optical confinement factor of the QW laser structure [25]. In QW lasers, the carriers and photons occupy different volumes. The total number of photons in the slave laser is $S(t) = S_0 + \Delta S(t) = V|E(t)|^2$, while $E(t)$ is the optical field. The total number of carriers is $N(t) = \Gamma V\, n(t)$. V is the optical mode volume. The optical confinement factor is well-known to be important for separate confinement quantum-well structures; however, it is usually ignored in literature on injection-locking. G_0 is the cavity gain coefficient, and $G_n = dG/dn|_{n=n0}$ is the differential gain. The gain saturation is also included, where ε and ε_M are the nonlinear gain saturation coefficients corresponding to the slave laser signal and the injected signal. The nonlinear gain saturation coefficients have been used in earlier studies [26–28] on high-speed lasers where the gain of the slave laser light is suppressed due to the presence of the injected light. Ref. [26] employs two saturation terms because the injected light and the internal light have orthogonal polarizations. In our derivation, the injected light is slighted different from the slave light. They both have contribution to the slave laser gain saturation, which is also called the cross-gain saturation of laser amplifiers [29,30]. Therefore, in our simulation, we need to consider both effects and use $\varepsilon_M \sim \varepsilon$. This is an important phenomenon if we inject light in the laser gain region and it cannot be neglected. Actually, inclusion of Γ and ε_M is important to obtain consistent parameters for the gain and differential gain for quantum-well lasers. Finally,

the above equations show that the slave laser will be affected by the emission of the master laser. When a photon is spontaneously emitted in the master laser, it will cause deviations of the amplitudes and phases of the slave and master laser fields from their stationary values. Thus, we must consider the master laser noise characteristics when deriving the RIN of the slave laser.

Noise caused by spontaneous emission and carrier generation-recombination is included in the rate equations by adding the appropriate Langevin driving terms. For simplicity, we also assume that the Langevin noise sources of the slave laser are independent of the Langevin noises of the master laser. The two sets of noise sources are uncorrelated. Neglecting the higher order terms, the differential forms of the rate equations for the injection-locked laser system with Langevin noise terms ($F_{\Delta S}(t)$, $F_{\Delta\phi}(t)$, $F_{\Delta N}(t)$) and master laser noises ($\Delta S_M(t)$, $\Delta\phi_M(t)$) are

$$\begin{aligned}\Delta\dot{S}(t) &= \left[\Gamma G_0(1-2\varepsilon S_0-\varepsilon_M S_{M0})-\frac{1}{\tau_p}+k_c\right]\Delta S(t)\\ &+\frac{G_n S_0}{V}(1-\varepsilon S_0-\varepsilon_M S_{M0})\Delta N(t)-2k_s S_0\Delta\phi(t)\\ &+(k_c S_0/S_{M0}-\Gamma G_0\varepsilon_M S_0)\Delta S_M(t)+2k_s S_0\Delta\phi_M(t)+F_{\Delta S}(t)\end{aligned}\tag{7}$$

$$\begin{aligned}\dot{\phi}(t) &= \Delta\dot{\phi}(t)\\ &= k_s/(2S_0)\Delta S(t)-k_c\Delta\phi+\frac{\alpha G_n}{2V}(1-\varepsilon S_0-\varepsilon_M S_{M0})\Delta N(t)\\ &-k_s/(2S_{M0})\Delta S_M(t)+k_c\Delta\phi_M(t)+F_{\Delta\phi}(t)\end{aligned}\tag{8}$$

$$\begin{aligned}\Delta\dot{N}(t) &= -\Gamma G_0(1-2\varepsilon S_0-\varepsilon_M S_{M0})\Delta S(t)\\ &-\left[\frac{1}{\tau_n}+\frac{G_n S_0}{V}(1-\varepsilon S_0-\varepsilon_M S_{M0})\right]\Delta N(t)\\ &+\varepsilon_M\Gamma G_0 S_0\Delta S_M(t)+F_{\Delta N}(t)\end{aligned}\tag{9}$$

where

$$k_c = k\sqrt{S_{M0}/S_0}\cos(\phi_0-\phi_{M0})$$

$$k_s = k\sqrt{S_{M0}/S_0}\sin(\phi_0-\phi_{M0})$$

We define $\Delta S(\omega)$, $\Delta\Phi(\omega)$, $\Delta N(\omega)$, $\Delta S_M(\omega)$, $\Delta\Phi_M(\omega)$, $F_{\Delta S}(\omega)$, $F_{\Delta\phi}(\omega)$, and $F_{\Delta N}(\omega)$ as the Fourier transforms of the truncated functions corresponding to $\Delta S(t)$, $\Delta\varphi(t)$, $\Delta N(t)$, $\Delta S_M(t)$, $\Delta\varphi_M(t)$, $F_{\Delta S}(t)$, $F_{\Delta\varphi}(t)$, and $F_{\Delta N}(t)$. Using the Fourier transforms of Eqs. (7)–(9), and gain–loss relation derived from the steady-state solution of Eq. (4), $2k_c = 2k\sqrt{S_0 S_{M0}}\cos(\phi_0-\phi_{M0}) = -\Gamma G_0(1-\varepsilon S_0-\varepsilon_M S_{M0})+\frac{1}{\tau_p}$, we obtain the following linear algebraic equations.

$$\begin{bmatrix} j\omega+a_{11} & a_{12} & a_{13}\\ a_{21} & j\omega+a_{22} & a_{23}\\ a_{31} & a_{32} & j\omega+a_{33}\end{bmatrix}\begin{bmatrix}\Delta S(\omega)\\ \Delta\Phi(\omega)\\ \Delta N(\omega)\end{bmatrix}=\begin{bmatrix}b_1(\omega)\\ b_2(\omega)\\ b_3(\omega)\end{bmatrix}\tag{10}$$

where

$$\begin{aligned}a_{11} &= -k_c-\Gamma G_0(1-2\varepsilon S_0-\varepsilon_M S_{M0})+\tfrac{1}{\tau_p}\\ a_{12} &= 2k_s S_0\\ a_{13} &= -G_n S_0(1-\varepsilon S_0-\varepsilon_M S_{M0})/V\\ a_{21} &= k_s/(2S_0)\\ a_{22} &= k_c\\ a_{23} &= -\alpha G_n(1-\varepsilon S_0-\varepsilon_M S_{M0})/(2V)\\ a_{31} &= \Gamma G_0(1-2\varepsilon S_0-\varepsilon_M S_{M0})\\ a_{32} &= 0\\ a_{33} &= \tfrac{1}{\tau_n}+G_n(1-\varepsilon S_0-\varepsilon_M S_{M0})S_0/V\\ b_1(\omega) &= [k_c(S_0/S_{M0})-\Gamma G_0\varepsilon_M S_0]\Delta S_M(\omega)+2k_s S_0\Delta\Phi_M(\omega)+F_{\Delta S}(\omega)\\ b_2(\omega) &= -k_s/(2S_{M0})\Delta S_M(\omega)+k_c\Delta\Phi_M(\omega)+F_{\Delta\phi}(\omega)\\ b_3(\omega) &= F_{\Delta N}(\omega)+\varepsilon_M\Gamma G_0 S_0\Delta S_M(\omega)\end{aligned}\tag{11}$$

The power spectral density of the slave laser photon can be obtained using the truncated function and Fourier analysis techniques [17]. Then, the RIN of the slave laser is obtained as

$$\begin{aligned}\frac{RIN}{\Delta f} &= \frac{2}{S_0^2|Y(\omega)|^2}\left\{\frac{R_s G_n^2 S_0^2}{V^2}(1-2\varepsilon S_0-2\varepsilon_M S_{M0})[\omega^2+\gamma_k^2]\right.\\ &+(\omega^2+k_c^2)(\omega^2+\gamma_e^2)L_1(\omega)+4k_s^2S_0^2(\omega^2+\gamma_e^2)L_2(\omega)\\ &-4k_sS_0(\omega^2+\gamma_e^2)\mathrm{Re}[(i\omega+k_c)L_3(\omega)]+\frac{2G_nS_0}{V}(1-\varepsilon S_0\\ &-\varepsilon_M S_{M0})\mathrm{Re}[(j\omega+k_c)(j\omega+\gamma_e)(-j\omega+\gamma_k)L_4(\omega)]\\ &+\frac{4k_sG_nS_0^2}{V}(1-\varepsilon S_0-\varepsilon_M S_{M0})\\ &\left.\times\mathrm{Re}[(j\omega-\gamma_k)(j\omega+\gamma_e)L_5(\omega)]\right\}\end{aligned}\tag{12}$$

where Re[...] stands for "the real part of [...]", $\gamma_e = \frac{1}{\tau_n}+G_n(1-\varepsilon S_0-\varepsilon_M S_{M0})S_0/V$ which is determined by the carrier lifetime, the slave photon number, and the cavity gain, $\gamma_k = k_c - k_s\alpha$ which is related to the phase difference and detuning between the slave and master lasers, and

$$\begin{aligned}L_1(\omega) &= \left(k_c\frac{S_0}{S_{M0}}-\Gamma G_0\varepsilon_M S_0\right)^2 P_{\Delta S_M}(\omega)+4k_s^2S_0^2P_{\Delta\Phi_M}(\omega)\\ &+R_s(2S_0+1)+4k_s^2S_0^2\left(\frac{k_c}{S_{M0}}-\Gamma G_0\varepsilon_M\right)\mathrm{Re}[P_{\Delta S_M\Delta\Phi_M}(\omega)]\end{aligned}\tag{13}$$

$$L_2(\omega) = \frac{R_s}{2S_0}+\frac{k_s^2}{4S_{M0}^2}P_{\Delta S_M}(\omega)+k_c^2P_{\Delta\Phi_M}(\omega)-\frac{k_ck_s}{S_{M0}}\mathrm{Re}[P_{\Delta S_M\Delta\Phi_M}(\omega)]\tag{14}$$

$$\begin{aligned}L_3(\omega) &= \frac{1}{2S_{M0}}(\Gamma G_0\varepsilon_M S_0k_s-\frac{k_ck_sS_0}{S_{M0}})P_{\Delta S_M}(\omega)+2k_ck_sS_0P_{\Delta\Phi_M}(\omega)\\ &-k_s^2\frac{S_0}{S_{M0}}P_{\Delta\Phi_M\Delta S_M}(\omega)+(k_c^2\frac{S_0}{S_{M0}}-k_c\Gamma G_0\varepsilon_M S_0)P_{\Delta S_M\Delta\Phi_M}(\omega)\end{aligned}\tag{15}$$

$$L_4(\omega) = -R_s+\frac{k_c\Gamma G_0\varepsilon_M S_0^2}{S_{M0}}P_{\Delta S_M}(\omega)\tag{16}$$

$$L_5(\omega) = -\frac{k_s\Gamma G_0\varepsilon_M S_0}{2S_{M0}}P_{\Delta S_M}(\omega)\tag{17}$$

In the equations above, $P_{\Delta S_M}(\omega)$, $P_{\Delta S_M\Delta\Phi_M}(\omega)$, $P_{\Delta\Phi_M}(\omega)$, and $P_{\Delta\Phi_M\Delta S_M}(\omega)$ are the power spectra of the free-running master laser, which can be obtained by setting $S_M(t)=0$ in Eqs. (4)–(6). $|Y(\omega)|^2$ is the denominator of the RIN of the slave laser

$$Y(\omega) = j\omega(\omega_r^2-\omega^2)-\omega^2\gamma+A\tag{18}$$

where

$$A = k^2\gamma_e\frac{S_{M0}}{S_0}+k_c\gamma_e\Gamma G_0\varepsilon S_0+\frac{\gamma_kG_nS_0\Gamma G_0}{V}(1-3\varepsilon S_0-2\varepsilon_M S_{M0})\tag{19}$$

The relaxation frequency of the slave laser is

$$\begin{aligned}\omega_r^2 &= \frac{1}{\tau_n}\Gamma G_0(2\varepsilon S_0-1)+\frac{1}{\tau_pV}G_0S_0(1-\varepsilon S_0)+\frac{1}{\tau_p\tau_n}+k^2\frac{S_{M0}}{S_0}\\ &+\Gamma G_0\left[\frac{\varepsilon_M S_{M0}}{\tau_n}+\frac{1}{2}\varepsilon S_0\left(\frac{1}{\tau_p}-\Gamma G_0\right)\right]-\frac{G_0\varepsilon_M S_0S_{M0}}{\tau_pV}\\ &= \omega_{r\,free}^2+k^2\frac{S_{M0}}{S_0}+\Gamma G_0\left[\frac{\varepsilon_M S_{M0}}{\tau_n}+\frac{1}{2}\varepsilon S_0\left(\frac{1}{\tau_p}-\Gamma G_0\right)\right]\\ &-\frac{G_0\varepsilon_M S_0S_{M0}}{\tau_pV}\end{aligned}\tag{20}$$

The damping factor of the injection-locked laser is

$$\gamma = \Gamma G_0(2\varepsilon S_0 - 1) + \frac{G_n S_0}{V}(1 - \varepsilon S_0) + \frac{1}{\tau_n} + \frac{1}{\tau_p} + \varepsilon_M S_{M0}\left(\Gamma G_0 - \frac{G_n S_0}{V}\right) = \gamma_{free} + \varepsilon_M S_{M0}\left(\Gamma G_0 - \frac{G_n S_0}{V}\right) \quad (21)$$

where terms of a higher order, $\varepsilon^2 S_0^2$ or $\varepsilon S_0 \varepsilon_M S_{M0}$, have been neglected. ω_{rfree} and γ_{free} are the relaxation frequency and the damping factor of the free-running laser, respectively. The nonlinear gain saturation term due to the master laser, which represents the gain change caused by the master laser injection, modifies the damping factor of the laser system.

For a free-running laser, there are no injected-photon $S_M(t) = 0$ or $S_{M0} = 0$ and $\frac{1}{\tau_n} \approx \Gamma G_0$. Thus, $k_c = k_s = \gamma_k = 0$. We input those values into Eqs. (4)–(6) and obtain the RIN, the relaxation frequency, and the damping factor of a free-running laser.

$$\frac{RIN}{\Delta f} = \frac{2R_s}{S_0^2}\frac{\left[\frac{G_n S_0(1-\varepsilon S_0)}{V}\left(\frac{G_n(1-\varepsilon S_0)S_0}{V} - 2\gamma_{e\ free}\right) + (2S_0 + 1)(\gamma_{e\ free}^2 + \omega^2)\right]}{\left[(\omega_{r\ free}^2 - \omega^2)^2 + \gamma_{free}^2\omega^2\right]} \quad (22)$$

where $\gamma_{e\ free} = \frac{1}{\tau_n} + G_n S_0(1 - \varepsilon S_0)/V$. The relaxation frequency of the free-running laser is

$$\omega_{r\ free}^2 = \frac{1}{\tau_n}\Gamma G_0(2\varepsilon S_0 - 1) + \frac{1}{\tau_p V} G_0 S_0(1 - \varepsilon S_0) + \frac{1}{\tau_p \tau_n} \quad (23)$$

The damping factor of the free-running laser is

$$\gamma_{free} = \Gamma G_0(2\varepsilon S_0 - 1) + \frac{G_n S_0}{V}(1 - \varepsilon S_0) + \frac{1}{\tau_n} + \frac{1}{\tau_p} \quad (24)$$

Compared with a free-running laser, the injection-locked laser is a third-order system instead of a second-order system (free-running), as shown by Eqs. (18) and (21). The injected-photon modifies the slave laser cavity properties, such as cavity gain decrease, relaxation frequency increase, and damping factor variation.

3. Study injection-locking using integrated electroabsorption modulator-lasers

3.1. Integrated electroabsorption modulator-lasers and their static characteristics

The structure of the integrated electroabsorption modulator-laser used is shown in Fig. 1. The device has three sections, which are a distributed feedback (DFB) laser section, an electroabsorption modulator section, and an electrode isolation section. The length of the DFB section is 300 μm, the isolation region is 83 μm, and the modulator section is 250 μm. The section between the laser and the modulator provides electrical isolation in the design.

The light–current (L–I) curves for this integrated electroabsorption modulator-laser were measured as function of modulator voltage and they are shown in Fig. 2a. The power output is always measured from the AR-coated facet. The optical spectra of the EML at 30 mA current bias and different modulator voltage biases are shown in Fig. 2b. The wavelength versus the current bias of the DFB laser and voltage bias of the modulator is summarized in Fig. 2c. The optical output power from the laser is not influenced much by the modulator bias. There is also no mode-hopping for this device when changing the voltage bias. When the modulator is reverse-biased, the absorption in the modulator section increases due to Franz–Keldysh or quantum-confined Stark effects.

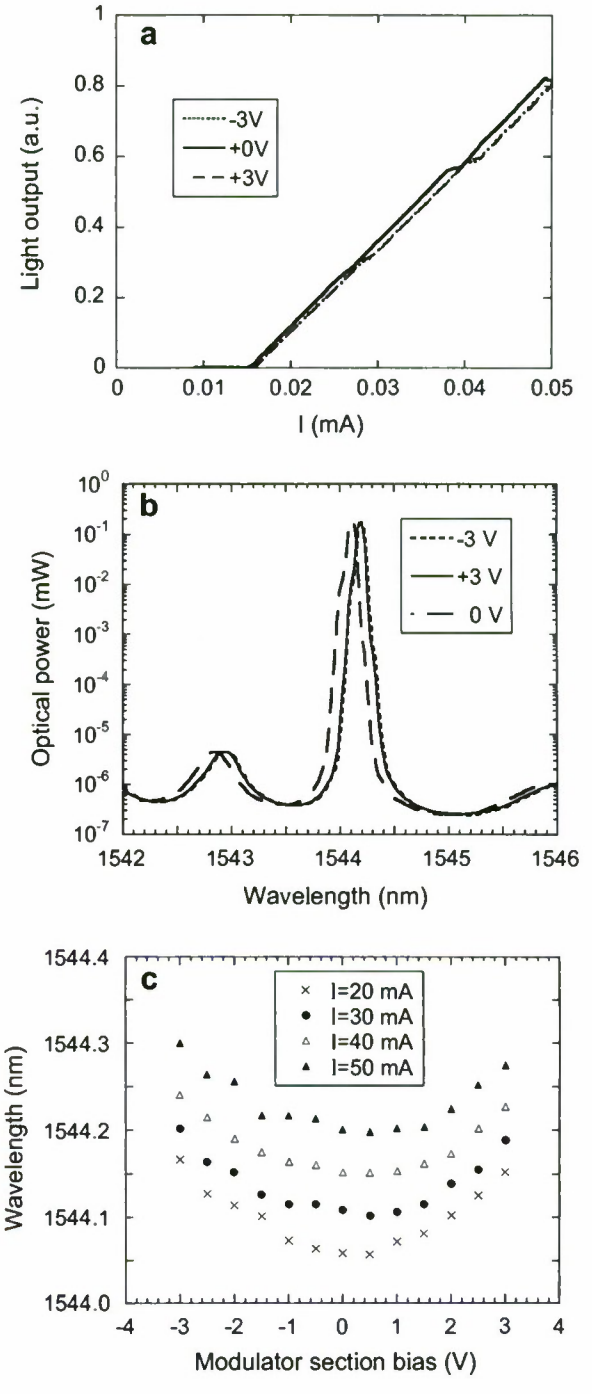


Fig. 2. (a) The light output versus the DFB laser bias current, (b) optical spectra at three modulator bias, and (c) the DFB laser wavelength at four bias current (I = 20, 30, 40, and 50 mA) as a function of the modulator bias voltage for the integrated electroabsorption modulator-laser.

When the modulator is forward-biased, the gain in the modulator section increases due to the carrier injection [32]. If looking into the modulator section from the isolation section, the modulator section can be viewed as an effective reflection coefficient and phase for the optical field at that boundary, which can perturb the longitudinal photon density profile of the laser cavity and modify the cavity properties. For our device, the waveguide in the laser section is straight, and the constant pitch of the laser already defines a specific Bragg wavelength and lasing wavelength. Thus the influence of the modulator section on the DFB laser wave-

length is mainly a small wavelength shift and not mode-hopping [32].

3.2. RIN of injection-locked integrated electroabsorption modulator-lasers

In this section, we present the experimental results of external injections in an integrated electroabsorption modulator-laser, using the modulator section of the EML as a photodetector. This experiment utilizes the advantage of photonic integrated circuit (PIC) technique to eliminate the disadvantages of using a more complicated setup with a separate photodetector as presented in [9]. The experimental setup is shown in Fig. 3. The injection signal from a single-mode DFB master laser passes through an erbium-doped fiber optical amplifier (EDFA). A tunable 3-nm bandwidth optical filter is used to remove excess signals on the side modes. The injection level is monitored by an optical power meter before it is injected into the EML where the laser section acts as a slave laser. The optical signal is converted to an electrical signal using the modulator photocurrent, amplified by an 18-dB gain microwave amplifier, and measured by the electrical spectrum analyzer to obtain the RIN spectra of the locked laser section output. A T-Bias is used to apply dc voltage onto the modulator section. In our setup, the method of external injection is similar to typical injection-locking with discrete semiconductor lasers. The difference arises in the method of photo-detection of the slave laser output for electrical analysis. In this case, the photodetection occurs in the modulator section instead of using another high-speed photodetector.

The influence of the modulator bias on the DFB laser should be very small so that any bias change or fluctuation in the modulator will not change the detuning between the master laser and the DFB laser on the EMLs, thus varying the injection condition and causing the laser system to become unlocked. When we bias the DFB laser of the EML at 30 mA, the wavelength difference between 0 V and −1 V modulator voltage is only 0.006 nm or 0.75 GHz (see Fig. 2c). This small difference will not switch the laser from the locked to the unlocked state, which ensures that the photocurrent generation is relatively independent at this voltage range. But the wavelength difference between 0 V and −2 V is 0.04 nm (5 GHz). This wavelength change cannot be neglected. Therefore, in our experiment we only bias our modulator at 0 V and −1 V. Fig. 2c also shows that 0 V bias occurs at the zero slope or near zero derivative of the wavelength–voltage curve, which is an ideal bias point to measure the photocurrent spectra.

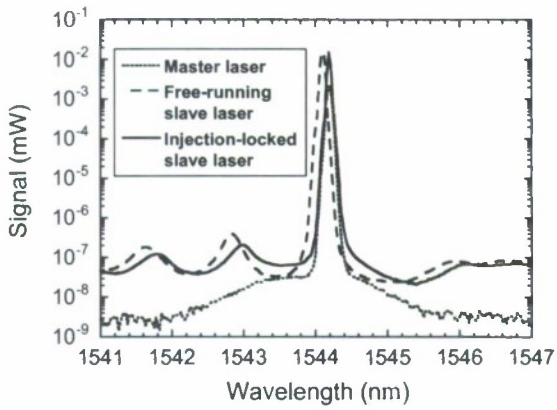


Fig. 4. Optical spectra of the external master (injection) laser, the integrated EML without injection (free-running slave laser), and the injection-locked EML under external injection light in the injection-locking experiment. The DFB laser of the EML (slave laser) is biased at 30 mA, and the modulator is biased at 0 V.

The optical spectra of the injection-locking experiment using the EML are shown in Fig. 4. With the master laser turned off and the DFB slave laser of the EML biased above threshold, there is no external injection, and the slave laser is free-running. During the experiment, the modulator is biased at 0 V. The free-running slave laser is lasing at 1544.1 nm (30 mA bias) with a side-mode suppression ratio (SMSR) of 44 dB. The master laser lases at 1544.2 nm with 50 dB SMSR. The detuning between the two lasers is 12.6 GHz. The injection-locked laser has the same lasing wavelength as the master laser and 50 dB SMSR. Thus, the injection light has caused the slave laser to lase at the master laser wavelength, resulting in an injection-locked condition. The RIN data of the injection-locked DFB laser is shown in Fig. 5a. From these measurements, we observe that the RIN peak shifts in frequency and amplitude varies as a function of the injection intensity. The data are limited by the noise floor of the electrical spectrum analyzer, which are around −130 dB/Hz for low frequency range (between 21 MHz and 6.26 GHz) and −126 dB/Hz for high frequency range (between 6.26 GHz and 12 GHz). We cannot resolve the signal below this level. But we can still observe the reduction of the RIN floor level under external injection. The spectra with higher injection power under injection-locking condition are below the noise limit of the spectrum analyzer and lower than the free-running slave laser (−126 dB/Hz) at low frequency range. The relaxation frequency peak is around 3.7 GHz for the free-running laser. The

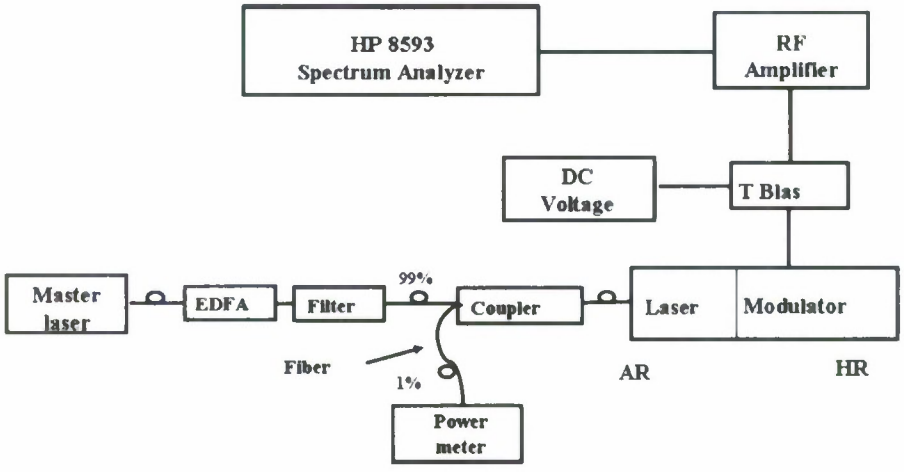


Fig. 3. Experimental setup of RIN measurement of the injecting-locked DFB laser using the integrated electroabsorption modulator-laser. The pump light from the master laser is injected into the AR facet of the integrated EML and the modulator is used as a photodetector.

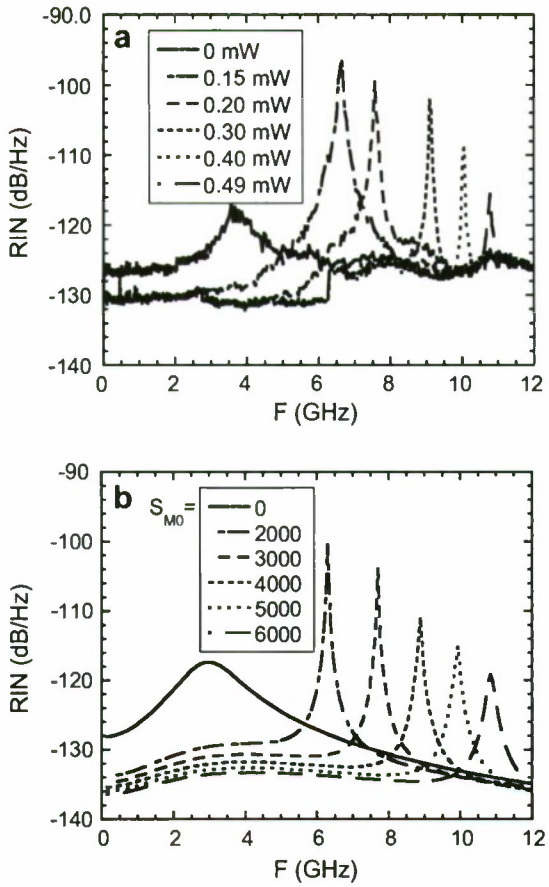


Fig. 5. (a) Experimental data and (b) theoretical calculation of RIN spectra of an injection-locked DFB laser system.

Table 1
The laser modeling parameters.

Parameter	Symbols	Value
Cavity length	L	300 μm
Active volume	V	3.89×10^{-10} cm^3
Effective index of refraction	n_g	3.33
Group velocity	v_g	9.0×10^9 cm s^{-1}
Mirror loss	αm	42.15 cm^{-1}
Intrinsic loss	αi	23 cm^{-1}
Optical confinement factor	Γ	0.15
Linewidth-enhancement factor	α	1.8
Photon lifetime	τ_p	8.5 ps
Carrier lifetime	τ_n	0.13 ns
Differential gain	$g'_l = g'_u$	3.6×10^{-16} cm^2
Nonlinear gain saturation coefficient	$\varepsilon_M \sim \varepsilon$	2.32×10^{-17} cm^3

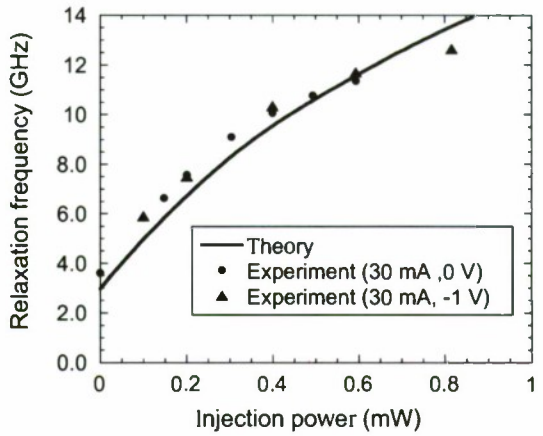


Fig. 6. The relaxation frequency versus different injection power of the injection-locked laser system. The injection power is monitored by an optical power meter before it is injected into the EML in the experiment. In the theory, a linear relation $P_{in}(mW) = 0.031\ S_{M0}$ is used to convert the calculated injected-photon number into the injection power.

maximum observed relaxation frequency is 11.3 GHz, which is three times the free-running value. As the injection intensity is increased further, the RIN peak attenuates and drops below the noise level. Our theoretical results are shown in Fig. 5b, which agree with our experimental data. The theory clearly shows higher injection level results in a lower noise floor and a larger relaxation frequency. The values of the physical parameters are $\Gamma = 0.15$, $\tau_p = 8.5$ ps, $V = 3.89 \times 10^{-10}$ cm^3, $G_0 = 8.7 \times 10^{11}$ s^{-1}, $G_n = 2.3 \times 10^{-5}$ cm^3 s^{-1}, $\tau_n = 0.13$ ns, $\alpha = 1.8$, $k = c/(2n_gL) = 1.5 \times 10^{11}$ s^{-1}, $L = 300$ μm, $R_s = 2 \times 10^{12}$ s^{-1}, and $n_g = 3.33$. The other important laser model parameters are listed in the Table 1. In our calculation, some of the parameters such as the effective index of refraction, intrinsic loss, and the initial value of the differential gain are obtained from independent measurements using the methods proposed in Ref. [25]. The linewidth-enhancement factor was obtained by measuring the injection-locking range [34]. The final value of the differential gain and gain saturation coefficients are fitting parameters. To simplify the calculation, because the wavelength of the injected signal and the slave signal are very close, we use $\varepsilon_M \sim \varepsilon$. The relaxation frequency versus injection power is plotted in Fig. 6. The symbols are experimental data, and the line represents theoretical results. We use a linear relation $P_{in}(mW) = 0.031 S_{M0}$ to convert the calculated injected-photon number into the injection power to compare with the experimental data. We also measure the RIN at −1 V modulator bias, and the results overlap with our 0 V data as expected. This confirms that the modulator bias has minimal influence on the DFB laser section. The modulator acts as a photodetector in our experiment to obtain the RIN of the injection-locked laser section, which is much simpler and easier than using discrete devices [9].

Our results also show theoretically that the coherent addition of the injected optical field to the slave laser optical field in the slave laser cavity is the main reason for the improvement of relaxation frequency. Without any injected signal $S_M(t) = 0$, the rate equations of the slave laser amplitude and phase are uncoupled (see Eqs. (4) and (5)). Also the phase term is not necessary to solve for the total photon spectrum. In an injection-locked laser system, the injected-photon term connects the amplitude and phase of the slave laser. The additional terms in the relaxation frequency of the injection-locked laser system come from the phase-amplitude coupling. Our theory also shows that the enhancement of the relaxation frequency can be attributed to the intensity of the injected field and the gain change (caused by nonlinear gain saturation terms). Generally, any change in the injection power or the gain will alter the relaxation resonance frequency [31]. Furthermore, an important effect of external optical injection in the stable locking regime is reduction of the cavity gain due to a reduction in carrier density, which shifts the optical resonance frequency and eventually modifies the relaxation frequency [7,31,33]. Note that our model also includes the optical confinement factor of the separate confinement heterostructure QW laser and gain saturation from the injected signal compared to the other RIN models, which are important to obtain a reasonable value of differential gain.

Finally, our data clearly show that as long as we keep the modulator section slightly revised biased, the modulator section of EML can be acted as an independent detector, which has little feedback to change the injection-locking in the laser section. This is the first

step to study an integrated injection-locking detection system. And our data approves it is achievable. Since the EML used is an already-to-use and easy-to-obtain device, it is our first choice for the experimental study of injection-locking in the integration device. In the future, it is very meaningful to solve Eqs (1)–(3) and obtain complete injection-locking model in the integrate devices to reveal the feedback effects.

4. Conclusions

In this paper, we have shown experimentally and theoretically that the injection-locking technique can improve the relaxation frequency of the slave laser, as well as lower the RIN floor level. In our experiment, we use an integrated electroabsorption modulator-laser, which simplifies the experimental setup and reduce loss in data acquisition. The static properties of the used EML are described, including its physical structure and operation. The experimental results of the external injection in EMLs are shown, emphasizing a novel method of obtaining the electrical spectra of the output light by directly measuring the output photocurrent of the EML modulator section. This differs from all previous external injection experiments, which use a separate photodetector to produce the modulated photocurrent. In summary, besides present a comprehensive analytical theoretical model for the relative intensity noise (RIN) spectrum of integrated semiconductor quantum-well (QW) lasers under injection-locking, we also explore the possibility of the injection-locking photonic integrated circuit (PIC) technique.

Acknowledgments

We thank Cal Poly C3RP 2005 and 2007 grant (The Department of the Navy, Office of Naval Research, under Award # ONR N00014-05-1-0855 and ONR N00014-07-1-1152) and Agilent Global Research Funding 2006 and 2007 (Contract # 07-040 and 08-146) for the support of this research.

References

[1] Lidoyne O, Gallion PB, Erasme D. IEEE J Quant Electron 1991;27:344.
[2] Chen HF, Liu JM, Simpson TB. Opt Commun 2000;173:349.
[3] Simpson TB, Liu JM. IEEE Photon Technol Lett 1997;9:1322.
[4] Wang J, Haldar MK, Lin L, Mendis FVC. IEEE Photon Technol Lett 1996;8:34.
[5] Meng XJ, Chau T, Wu MC. Electron Lett 1998;34:2031.
[6] Haldar MK, Coetzee JC, Gan KB. IEEE J Quant Electron 2005;41:280.
[7] Jin X, Chuang SL. Solid-State Electron 2006;50:1141.
[8] Lang R. IEEE J Quant Electron 1982;QE-18:976.
[9] Jin X, Chuang SL. Appl Phys Lett 2000;77:1250.
[10] Yabre G. J Lightwave Technol 1996;14:2367.
[11] Harb CC, Ralph TC, Huntington EH, Fretiage I, McClelland DE, Bachor H. Phys Rev A 1996;54:4370.
[12] Huntington EH, Buchler BC, Harb CC, Ralph TC, McClelland DE, Bachor H. Opt Commun 1998;145:359.
[13] Chrostowski L, Zhao X, Chang-Hasnain CJ. IEEE T Microw Theory 2006;54(2):788.
[14] Simpson TB, Liu JM, Gavrielides. IEEE Photon Technol Lett 1995;7:709.
[15] Espana-Boquera MC, Puerta-Notario A. Electron Lett 1996;32:818.
[16] Yabre G, Wardt HD, Van den Boom HPA, Khoe GD. IEEE J Quant Electron 2000;36:385.
[17] Spano P, Piazzolis S, Tamburrini M. IEEE J Quant Electron 1986;QE-22:427.
[18] Schunk N, Petermann K. IEEE J Quant Electron 1986;QE-22:642.
[19] Hong Y, Shore KA. IEEE J Quant Electron 1999;35:1713.
[20] Tanbun-Ek T, Adams L, Nykolak G, Bethea C, People R, Sergent A, Wisk P, Sciortino P, Chu S, Fullowan T, Tsang WT. IEEE J Sel Top Quant Electron 1997;3:960.
[21] Bigan E, Ougazzaden A, Huet F, Carre M, Carenco A, Mircea A. Electron Lett 1991;27:1607.
[22] Wakita K, Kotaka I, Mitomi O, Asai H, Kawamura Y. IEEE Photon Technol Lett 1991;3:138.
[23] Lang R, Kobyashi K. IEEE J Quant Electron 1980;QE-16:347.
[24] Masoller C. IEEE J Quant Electron 1997;33:796.
[25] Minch J, Park SH, Keating T, Chuang SL. IEEE J Quant Electron 1999;35:1526.
[26] Eom J, Su CB. Appl Phys Lett 1989;54:1734.
[27] Lange CH, Su CB. Appl Phys Lett 1989;55:1704.
[28] Vassilovski D, Wu TC, Kan S, Lau KY, Zah CE. IEEE Photon Technol Lett 1995;7:706.
[29] Jin X, Keating T, Chuang SL. IEEE J Quant Electron 2000;36:1485.
[30] Qasaimeh O. J Lightwave Technol 2008;26:449.
[31] Lin L. IEEE J Quant Electron 1994;30:1701.
[32] Hsu A, Chuang SL, Fang W, Adams L, Nykolak G, Tanbun-Ek T. IEEE J Quant Electron 1999;35:961.
[33] Simpson TB, Liu JM, Gavrielides A. IEEE J Quant Electron 1999;35:961.
[34] Liu G, Jin X, Chuang SL. IEEE Photon Technol Lett 2001;13:430.

11. Current Work:

We are still working on combine top and bottom grating design: Simeon Trieu, Xiaomin Jin, Bei Zhang, Tao Dai, Wei Wei, Chang Xiong, Xiang-Ning Kang, and Guo-Yi Zhang, "Study of Top and Bottom Photonic Gratings on Gallium Nitride Light-emitting-diodes." The Ninth International Conference on Solid State Lighting, SPIE Symposium on Optical Engineering + Applications, August 2-4th, 2009, San Diego, California, USA. (Accepted)

Algae Lipid Characterization and Extraction

Project Investigators:

Corrine Lehr, Department of Chemistry and Bioehemistry
Trygve Lundquist, Department of Civil and Environmental Engineering
California Polyteehnic State University
San Luis Obispo, CA

Project Report: Algae Lipid Characterization and Extraction

Research Background

As the use of biofuels expands in the United States, it has become apparent that conventional feedstocks such as soybean oil, palm oil and waste vegetable oil are not sufficient to meet biodiesel demand. It has been suggested by the U.S. National Renewable Energy Laboratory (NREL) that algae have potential as a feedstock for the large-scale production of biodiesel. Previous research has concluded that algae may be up to 30 times as productive a biodiesel feedstock per unit area than conventional terrestrial crops (Sheehan et al., 1998). Although the high costs of conventional algae production exclude the possibility a of cost-effective large-scale algae biodiesel process, the integration of algae production into a separate process may open up new possibilities. NREL has proposed that algal wastewater treatment might be effectively combined with algae biodiesel production. Extraction of triglycerides from wastewater-grown algae may yield a crude oil that can be processed to produce biodiesel. The Defense Advanced Research Projects Agency (DARPA) is also interested in the investigation of such a process for the purpose of producing a suitable replacement for the JP-8 fuel used extensively by the U.S. Military. JP-8 is chemically similar to diesel fuel and is used by the military in diesel engines.

Research conducted throughout the 1950s by E. G. Bligh and W. J. Dyer of the Fisheries Research Board of Canada led to the 1959 publication of a method for the rapid extraction of lipids from fish material. Lipids are a class of biomolecules which include triglycerides. The "Bligh and Dyer" method of lipid extraction has been refined through other studies since its initial publication. Professor Elahe Ennsani of San Francisco State University described a modified version of the Bligh and Dyer procedure specifically for lipid extractions from algae in her 1990 Ph.D. Thesis. The modified Bligh and Dyer method has excellent lipid recovery efficiency, but the industrialization of the process poses several problems. The procedure requires a very large solvent to biomass (vol/vol) ratio, uses highly toxic solvents which limit the usefulness of residual algae solids as a fertilizer, involves a complex choreography of steps which does not lend itself to automation, and requires a high energy input for solvent recovery.

The objectives of the current research are to develop methods of producing algae with high triglyceride content and to establish whether the lipids produced by such algae are suitable for conversion into a biodiesel fuel. The research is also meant to work toward cost-effective and safe methods of extracting biodiesel-appropriate triglycerides from algal growth media. The main determinants of fuel suitability are the carbon chain length and level of hydrogen saturation of the aliphatic regions of triglycerides.

Materials and Methods

Algae Ponds

The algae used for experimentation have been grown in an experimental pilot algae treatment system at the San Luis Obispo Water Reclamation Facility. Algae have been harvested from ponds in the pilot system throughout the research. The ponds are exposed to the elements and continuously supplied with primary wastewater effluent from the City of San Luis Obispo, just as an industrial production facility might be.

The algae ponds at the wastewater treatment plant have been operated in two different arrangements during the course of experimentation: continuous mode and batch mode. In continuous mode, primary wastewater effluent is constantly discharged into the ponds and algae water is constantly removed from the ponds at the same rate. When operated in continuous mode, the ponds have a hydraulic residence time of approximately 5 days. Continuous mode operation is likely to be the most practical arrangement for large-scale algae production; however, algae which have had longer than 5 days to grow tend to produce a higher concentration of lipids. A highly dense algae culture with high lipid concentration is ideal for laboratory investigation because the greater mass of test material allows a higher sensitivity in resulting data.

An extremely long residence time can be achieved by growing algae in batch mode. In batch operation, a reactor is filled with wastewater, inoculated with algae, and then left to grow. This method also makes it possible to evaluate the quantity and types of lipids produced throughout the lifespan of an algal culture, which is difficult to ascertain from a continuous culture. Batch culture experiments have been conducted in the current research with residence times of up to 25 days. The high concentrations of lipids produced by batch culture algae are ideal for research because they improve the accuracy of chromatographic analysis.

Sampling

Harvesting techniques were developed for the accurate determination of lipid content and quality in algae samples from the ponds. Initial samples were processed by using an alum coagulant, $Al_2(SO_4)_3 \cdot 18\ H_2O$, followed by centrifugation in a 4-liter capacity floor centrifuge. Although it was established that alum does not interfere with chromatographic analysis, this process was later amended due to uncertainties about the water content of the algae being processed. Specifically, the volumetric change of the sample associated with alum flocculation and the water content gradient which occurs in a sample processed in a large centrifuge bucket made it difficult to accurately assess the algae content and the alum content of the samples to be analyzed. Although flocculation and centrifugation may be an effective method of algae collection in an industrial process, the remainder of the research has been conducted with algae samples concentrated by centrifugation alone, using a table top centrifuge. This method results

in concentrated algae samples of approximately constant water content, which is necessary in order to maintain the level of accuracy required in a laboratory setting.

Some samples of algae were frozen prior to centrifugation to establish whether freezing is an effective method of sample preservation. The samples were later thawed and it was observed that most of the algal solids had settled. This illustrated that freezing prior to centrifugation eliminates the possibility of subsequent analytical work being performed on an algae sample. The rapid settling which occurs in a thawed sample makes it difficult to divide the sample into aliquots of uniform solids concentration. However, freezing and thawing is an excellent method of algae cell disruption. Below is an image of an algae sample from the pilot algae treatment ponds which has been frozen and thawed. The picture was taken approximately 30 seconds after vigorous shaking of the sample. A layer of solids is visible at the bottom of the test tube.



Figure 1: Thawed and Settled Algae

Analytical Method

In order to characterize the quantity and quality of algal oils, a program was developed for a gas chromatograph/mass spectrometer (GC-MS). The GC-MS is an instrument which allows users to analyze the molecular composition of volatile materials. By heating a sample, which has been injected into a long capillary column, a GC-MS separates constituents of the sample based on their volatilities. As compounds emerge from the end of the capillary, the molecular mass of their parts is determined by a mass spectrometer, providing for their identification.

A standard solution of fatty acid methyl esters (FAMEs, or biodiesel molecules) was prepared for the development of the GC-MS program and for the calibration of the instrument. Fully saturated FAMEs of even-numbered carbon chain lengths between 8 and 24 were used because they span the range of useful carbon chain lengths for biodiesel fuel. The GC-MS program was designed to provide accurate resolution of individual molecules, which are displayed as peaks on the graphical output of the GC-MS instrument. Next, a calibration curve was constructed to

calculate recovery rates of the samples proeessed by the GC-MS. The development of the GC-MS program and calibration curve has allowed for rapid analysis of samples.

Experimentation

Once methods of sample collection and chromatographic analysis had been established, an experiment was undertaken to determine the quantity and quality of lipids produced in batch eulture algae throughout the algal growth curve. An algae pond was at the San Luis Obispo Watcr Reclamation Facility was reeonfigured to run in batch mode. The pond was sparged with carbon dioxide in order to improve the growth rate of the algae, a technique described in a previous thesis project at Cal Poly (Feffer, 2007). The flow of earbon dioxide was adjusted each day to maintain a pH of approximately 7.75, a favorable eondition for algal growth. A paddle wheel rotating at approximately 6 rpm was used to provide a eonstant flow of algae water around the pond. Thc setup is pictured below. The batch mode pond was inoculated with algae which had been grown in an adjacent continuous mode pond to ensure the practicality of reproducing these conditions on an industrial scale. Inoculum eonstituted 10% of the total volume of the pond. The other 90% was wastewater from the primary clarifier at the water reclamation facility. The total volume of water in the pond was 250 gallons, or approximately 950 liters.



Figure 2: Batch Mode Algae Pond

During eaeh day of operation of the pond, measurements of the eoneentration of algae in the water were taken in order to document the growth of the culture. Microscopic investigations of the miero biota of the pond were condueted throughout the expcriment in ordcr to evaluatc changes in the species and number of algae and other mieroorganisms present over thc coursc of the operation of the pond. Lipid extractions were performed regularly throughout thc lifespan of the culture.

The concentration of algae in the pond was determined by performing total suspended solids and total volatile solids experiments each day. These tests, commonly referred to collectively as TSS/VSS, involve the filtration of a known volume of sample material through a tared filter. By baking the filter at 105C, the weight of the solid material collected on it is obtained. The filter is then baked at 525C in order to remove combustible material from the filter. The weight of the combustible material, called the volatile suspended solids, is taken to be the dry weight of the algae initially present in the sample volume. Tests were performed in triplicate and the average values are reported.

Micrographic analysis of the algal species of the pond was performed using a microscope equipped with a digital camera. This was done 5 times throughout the 25 day operation of the batch mode algae pond.

Regular lipid extractions were performed using the Ennsani-amended Bligh and Dyer extraction procedure. After samples were retrieved from the ponds, they were rushed immediately to a refrigerator, in which they were kept under nitrogen until they were processed a maximum 5 hours later.

The first processing step was centrifugation. Six 200mL aliquots per day were centrifuged, resulting in a small pellet. 5mL of chloroform, 10mL of methanol and 4mL of deionized water, and were then added to resuspend the pellets. A sonicator was used to disrupt the algae cells in the newly suspended mixture. The samples were then placed on a horizontal shaker table with 6 cm oscillations at 2 cycles per second for between 6 and 8 hours. The purpose of the shaking step was to promote the complete exposure of intracellular products to the solvents.

After shaking, an additional 5mL of chloroform and 5mL of deionized water was added to each sample. A vortex mixer was then used for 30 seconds to mix the newly added material. The resulting 10:10:9 chloroform:methanol:water mixture was centrifuged for 4 minutes in order to separate the hydrophilic and hydrophobic layers. A test tube containing the mixture at this stage is pictured below. The green chloroform layer, at the bottom, contains lipids and chlorophyll. The upper layer contains methanol and water. A thin layer of cell debris separates the two layers.



Figure 3: Separated Layers

Onee separated, the lowcr, lipid rich chloroform layer was removed using a Pastcur pipette and added to a tared test tube. An additional 10mL of chloroform was added to eaeh sample before vortexing, eentrifuging and extracting once more. The purpose of performing a redundant extraction was to encourage thc complete recovery of lipids from the algal material.

Of the 6 samplcs taken at a timc, three wcrc used to analyze the quantity of lipids in the algac. These samples were extracted into tared aluminum trays and plaeed in a nitrogen sparged desiccator in order to evaporate the solvent from the lipids. Sincc it is impractical to resuspend lipids contained in an aluminum tray, these samples were no longcr useful for chromatographic analysis. The advantagc of using trays during this step was that the light wcight of aluminum trays makes them ideal for obtaining an accurate measurement of the lipid mass they contain. Nitrogen was used in order to prevent oxidation of the lipids, which may interfcrc with thcir mass. After most of the solvent had been evaporated, the samples wcre plaeed in a nitrogen sparged oven for 1 hour at 105C in order to remove any remaining solvent. Evaporating most of the solvent at ambient tempcrature prevented the rapid, violent vaporization which would occur if a large volume of solvent had been plaeed in the 105C ovcn. This was donc to preserve thc samplcs in thcir entircty, for a reliable lipid mass result. The results from cach of the 3 trays wcrc averaged to obtain thc reported numbers. The image below shows trays containing algae lipids in the nitrogen sparged desiccator. The lipids are grcen, indicating that chlorophyll is extracted along with triglycerides.



Figure 4: Lipid Extracts in Desiccator

The remaining three samples were used to evaluate the types of lipids present in each sample. These samples were extracted into glass test tubes and the solvent was removed by evaporation in a nitrogen sparged desiccator. The use of sealable test tubes in this step made it possible to subsequently resuspend the samples in preparation for a transesterification reaction.

The transesterification reaction was necessary to convert the algal tryglicerides into FAMEs. The light molecular weight of FAMEs relative to the triglycerides from which they are derived makes them more readily analyzed by gas chromatography. The first step of the transesterification was to resuspend up to 1 mg of the oil extracts in 1mL of dry toluene. 2mL of 0.5M anhydrous sodium methoxide in methanol were then added to the samples. Methanol is used because methoxide groups in the mixture bind the carbon atom central to each carboxyl group within each triglyceride, breaking the triglycerides apart and creating three FAMEs from each triglyceride. This reaction was catalyzed by placement in a 50C water bath for ten minutes before the methoxide was neutralized by the addition of 0.1mL of glacial acetic acid to each sample. 5mL of deionized water and 5mL of hexane were then added, and the FAME-rich hexane layers were extracted. An additional 5mL of hexane were added to each sample to increase the completeness of FAME recovery. The samples were then dried using sodium sulfate before injection into GC-MS instrument, primarily to protect the GC-MS equipment.

Results

Images of the culture throughout the experiment are shown below. The first image shows the inoculum, grown in continuous mode, at 1000X magnification. The spherical algae Anacystis was present in the inoculum and remained the dominant species throughout the experiment. The next image is a 1000X magnification of a sample taken from the pond on day 9. Anacystis and scenedesmus were present on day 9. Scenedesmus appeared shortly after the experiment was initiated and remained for the duration of the life of the culture. The next image was taken at 1000X magnification on day 14 and illustrates the increasing diversity of the culture that came with time. The image includes Actinastrum as well as Scenedesmus and Anacystis. The last

image, captured on day 21, shows the cellular debris which started to accumulate in the pond as the algae culture declined.

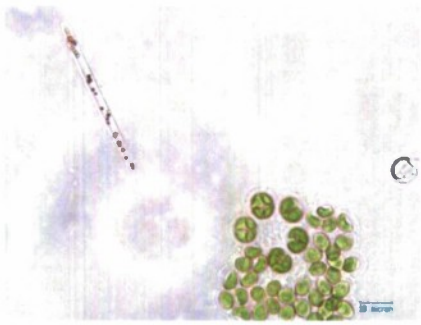


Figure 5: Inoculum

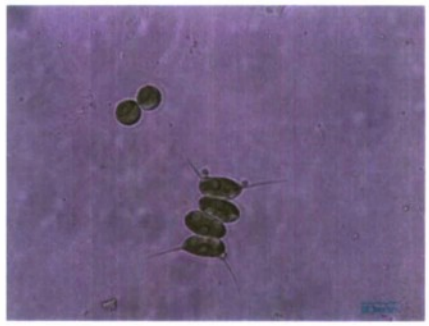


Figure 6: Day 9

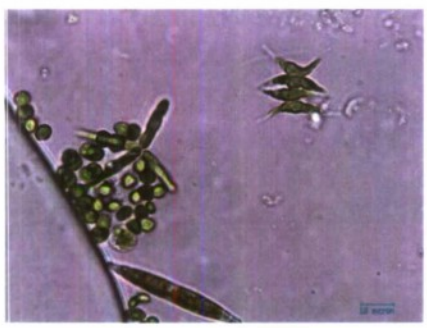


Figure 7: Day 14

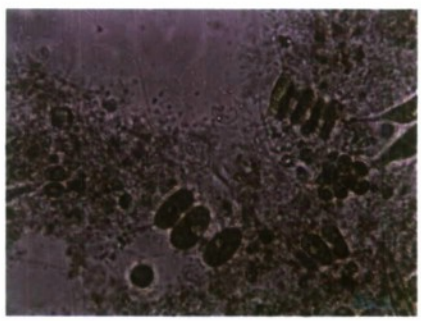


Figure 8: Day 21

The volatile suspended solids of the inoculum were 175mg/L and the volatile suspended solids of the wastewater which was used in the pond were 72mg/L. The inoculum contained 17.1% lipids by weight and the wastewater contained 14.9% lipids by weight. The total volatile solids remained relatively constant over the course of the first 5 days of the experiment. This is known as a "lag phase" and it results from the adjustment of the culture to a new environment. The volatile solids then increased rapidly in what is known as an "exponential growth phase," peaking at 527mg/L on day 17. After that, the volatile solids concentration held steady at about 430mg/L from day 19 to day 21. This is called a "stationary phase" and it preceded at period of rapid volatile solids decline known as a "death phase." The changes in the volatile solids concentration were similar to what may have been expected.

The growth rate of algae in a pond system is typically expressed in terms of the change in the mass of volatile suspended solids per unit pond surface area per day. An average growth rate of 6.3g/m^2/day was recorded over the growth phase of the culture in the pond. The growth rate peaked at 10.8g/m^2/day on day 17. The concentration of oil in the algae increased with the algal

growth rate and decreased with algal culture decline. On day 17, at the time of maximum algae growth, lipids constituted 19% of the dry mass of the algae. At that time lipids were present at a concentration of approximately 100mg/L in the bulk growth media of the pond. Assuming a biodiesel density of 0.88g/cm^3, the maximum growth rate and lipid concentration recorded in the pond correspond to a production rate of approximately 911 gallons of biodiesel per acre of algae ponds per year.

The productivity is greater than values typically quoted for conventional biodiesel feedstocks such as palm and soy. It is likely that even higher productivities are possible with wastewater algae grown under improved conditions. This experiment was conducted between March and April of 2009. The average daily high temperature of approximately 65°F was lower than ideal. With a depth of approximately 8 inches, the algae culture was significantly shaded by the 2.5 foot tall walls of the algae pond. It is expected that a culture set up in similar conditions would be more productive if shading were limited and the temperature was higher. The growth curve of the culture is pictured below. The blue line represents algae concentration in the bulk growth media, and the red line represents lipid concentration in the bulk growth media. Data continues to be added as samples from the experiment are processed.

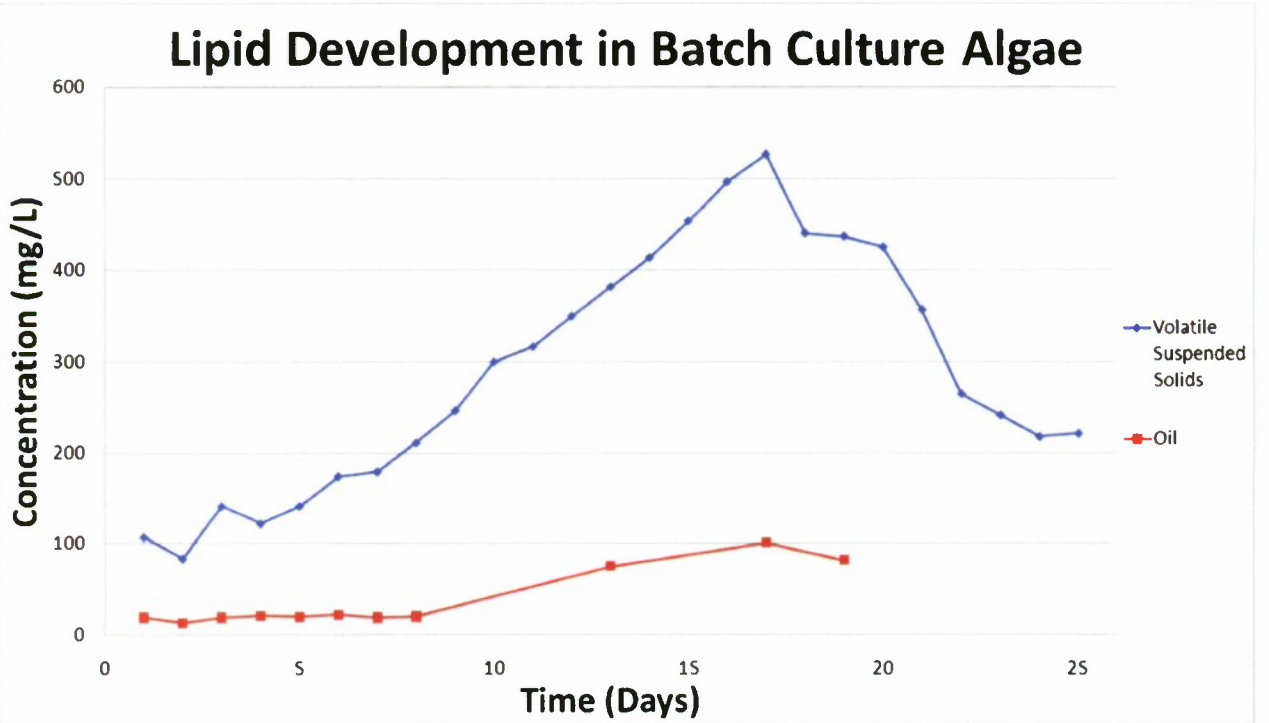


Figure 9: Algae and Lipid Development in Batch Mode

The lipids extracted from the algae are currently being analyzed using the GC-MS instrument. Based on initial results from the batch mode experiment, and on earlier tests, wastewater algae

typieally produees triglyeerides with aliphatie earbon ehains of between 14 and 18 earbon atoms in length. The most eommon FAME is a eompletely saturated, 16-earbon ehain. Various levels of hydrogen saturation have been observed. Some of the ehains are branehed, with a methyl group on the terminal, or third-to-last earbon on the ehain. Completely saturated arrangements are the most eommon and of the unsaturated arrangements observed, monounsaturated moleeules are the most eommon. These results are eneouraging as the most eommon saturated 16 and 18-earbon FAMEs are ideal for biodiesel produetion. Other moleeules present are elosely related and are also likely to be suitable for biodiesel produetion.

GC-MS ehromatograms from day 1 of the bateh mode pond experiment, and results from an earlier test of eontinuous pond algae are pietured below. Moleeules present in the samples show up as peaks on the ehromatograms. The horizontal axes in the chromatograms represent time, and the vertieal axes represent the abundance of each molecule. FAMEs which are appropriate for use as fuel have been identified and are pictured in the ball-and-stick diagrams superimposed on eaeh of the ehromatograms.

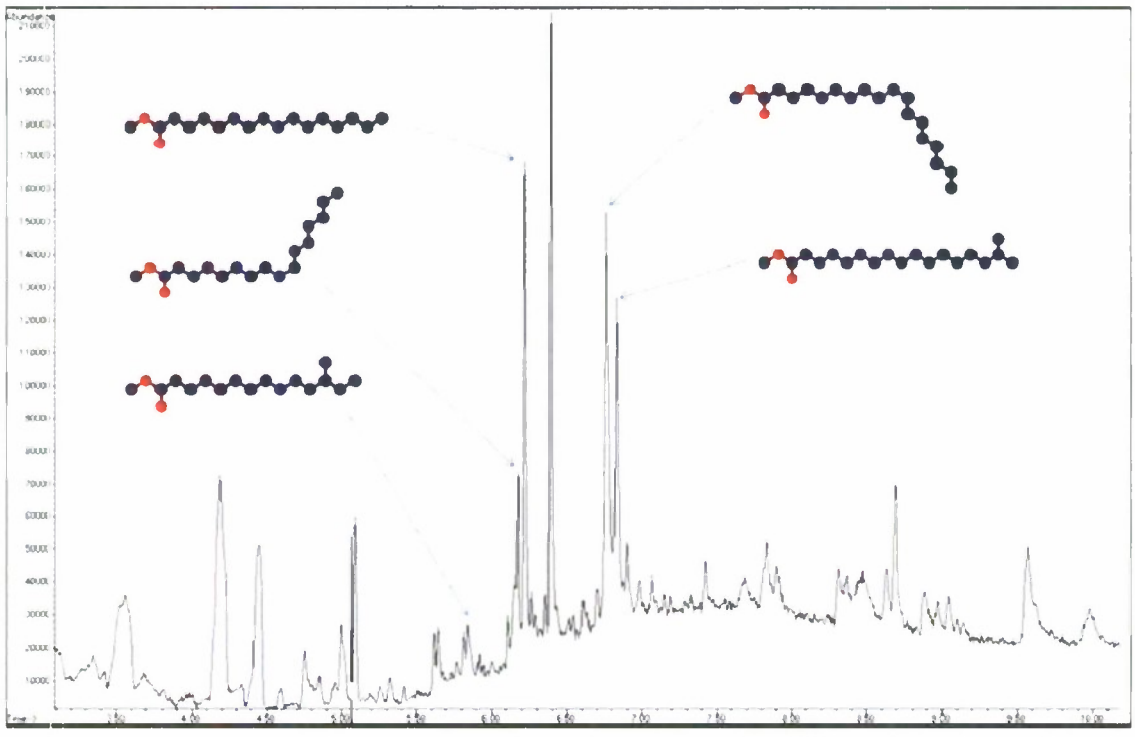


Figure 10: Chromatogram of Batch Mode Algae Lipids

The bateh mode sample eonsisted primarily of saturated 16-carbon and monounsaturated 18-earbon FAMEs. Both of these moleeules are ehemieally similar to eonstituents of palm and vegetable oils whieh are commonly used to produce biodiesel. The tall unnamed peak between the FAMEs is dibutyl phthalate, a eommon plastieizer used in the produetion of inks, adhesives

and plastics. Although dibutyl phthalate is periodically present in chromatographic samples, it is not suspected of interfering with results.

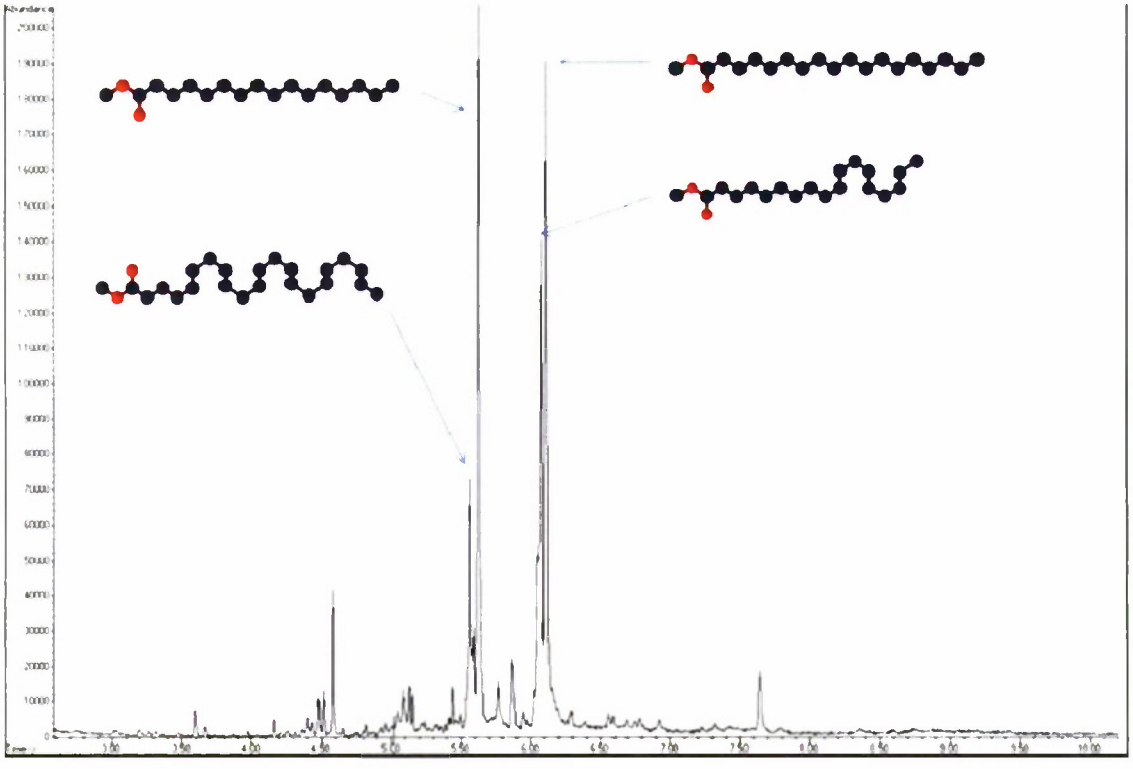


Figure 11: Chromatogram of Continuous Mode Algae Lipids

The continuous mode sample consisted primarily of saturated 16 and 18-carbon FAMEs which are also common to biodiesel produced from palm or vegetable oil.

Further Research

The current research continues to provide important information about algae lipid production and extraction. Several important experiments are in progress currently. Data from the 25 day batch mode experiment continues to be processed. Pending lab results will provide clear information about the changes in concentrations of different lipid molecules throughout the growth of the batch algae culture. These data are important because they will provide insight on the age at which algae provide the best lipids for biodiesel fuel production.

A large experiment comparing the extractive properties of several different solvents is underway now. The solvents include methanol, ethanol and isopropanol. An additional extraction will be performed using high temperature methanol to determine the effect of solvent temperature on the

quantity and quality of triglycerides extracted. The results of each of these solvent extractions will be compared to Bligh and Dyer data from the same algae sample.

Future research should be conducted on the effect of carbon dioxide addition on lipid production and lipid quality. Other factors which may have an influence include temperature and insolation rate. Further investigation of each effect will provide valuable information on the best methods of producing algae-based biodiesel fuel.

Although the current research will provide crucial insight into the triglycerides which can be extracted by several solvents, additional oil processing technologies should be explored. Supercritical carbon dioxide extraction, although relatively expensive, may be a viable alternative to organic solvent extraction. Physical extraction methods, such as three-phase centrifugation may also prove to be good methods of producing biodiesel from algae on a large scale.

Autonomous Military Robotics: Risk, Ethics, and Design

Project Investigators:

Patrick Lin, Department of Philosophy
George Bekey, Department of Computer Sciences
Keith Abney, Department of Philosophy
California Polytechnic State University
San Luis Obispo, CA

CAL POLY

Project Summary:

Autonomous Military Robotics: Risk, Ethics, and Design

Prepared by: Patrick Lin, Ph.D.; George Bekey, Ph.D.; and Keith Abney, M.A.
Ethics & Emerging Technologies Group at California State Polytechnic Univ., SLO

Project dates: September 30, 2007—December 31, 2008

Support from: US Department of Navy, Office of Naval Research, award # N00014-07-1-1152

INDEX

1. *Project Proposal*
2. *Relevance to ONR/DoD*
3. *Project Team & Collaborators*
4. *Research Results*
5. *Next Steps*
6. *Contact*

1. Project Proposal

The US military is committed to using battlefield robots as part of its effort to develop 'Future Combat Systems.' Indeed, Congress has already mandated that, by 2010, one-third of all operational deep-strike aircraft must be unmanned, and by 2015, one-third of all ground combat vehicles must be unmanned—deadlines that are also driving the trend toward more autonomous, efficient robots. Behind these commitments, the Armed Services have several specific goals, including: (1) to save soldiers' lives, (2) to reduce the number of warfighters needed in future conflicts, and hence (3) to

save costs. Further, in the distant future, some military planners hope to use robots to engage the enemy dispassionately, ensuring compliance with the Geneva Conventions and other standards of ethical conduct.

However, many troubling questions arise with the increasing use of robotics in military applications, particularly given the fast-approaching Congressional deadlines and other pressures. These troubling questions include the fundamental challenge of designing machines that can 'think ethically' in the first place as well as more distant concerns of robots gone wild.

The need for technology risk assessment in this area—and, more broadly, for 'robot ethics' or 'roboethics'—is clear given recent failures of even semi-autonomous military and civilian systems, e.g., malfunction of a South African robotic cannon that inadvertently killed nine soldiers and wounded 14 others. Further, Europe is leading efforts in robot ethics, given years of workshops, conferences, and reports they've produced on the subject already. But the US is starting to catch up.

In its first year, our interdisciplinary research project proposes to: identify the benefits and risks from robotics to the US military, including a survey of current and emerging robots; provide a framework for investigating robot ethics; look primarily at near- and mid-term issues and risk; and then provide a quick preview of far-term issues. This project would rely on an extensive literature review, interviews with industry experts, and both online and offline research.

Our goal is to leverage this work into a much larger project funded by other federal grants and industry support. We plan to submit proposals to NSF, ONR, DARPA, and other organizations, and we collaborate with industry companies and other organizations (such as Yale University and the U.S. Naval Academy) during the course of the entire project.

2. Relevance to ONR/DoD

This interdisciplinary project not only seeks to responsibly guide developments in robotics—as a significant enabler of taking the US military into the next generation—but it also attends to many areas of interest to the ONR/DoD, including: (1) *natianal security applicatians* (e,g., remote sensing, robotics, biosensing/detection, expert systems), (2) *farce pratectian* (e.g., improved protection of the individual), (3) *human perfarmance, training, and survivability* (e.g., cognitive and physical performance enhancement), and (4) *intelligence, surveillance, and recannaissance* (e.g., data/information analysis and distribution). Our proposal also satisfies other priorities in the program, such as helping to seed new research programs of junior faculty, attract future funding, build industry ties, and so forth.

3. Project Team & Collaborators

Robot ethics, and even more so with respect to military systems, is an emerging and interdisciplinary field that continues to draw from experts in robotics, technology ethics, and other prior fields. Our research team at Cal Poly, therefore, is interdisciplinary and uniquely qualified to execute this project: Dr. Patrick Lin, director of the Ethics & Emerging Technologies Group and on the philosophy department faculty; Dr. George Bekey, research fellow in the biomedical engineering department and founder of University of Southern California's robotics lab, where he is also a professor emeritus; and Keith Abney, also on the philosophy department faculty and engaged in technology and biomedical ethics.

We have also enlisted the assistance of the following consultants, all experts in the field of machine/robot ethics: Colin Allen (Indiana Univ.), Peter Asaro (Rutgers Univ.), and Wendell Wallach (Yale). We have also conducted in-person and online discussions with the following experts: Ron Arkin (Georgia Tech), John Canning (Naval Surface Warfare Center), Ken Goldberg (IEEE Robotics and Automation Society; UC Berkeley), Patrick Hew (Defence Science and Technology Organization, Australia), George R. Lucas, Jr. (US Naval Academy), Frank Chongwoo Park (IEEE Robotics and Automation Society; Seoul National Univ.), Lt. Col. Gary Sargent (US Army Special Forces; Cal Poly), Noel Sharkey (Univ. of Sheffield, UK), Rob Sparrow (Monash Univ., Australia), and others.

4. Research Results

We have completed a 100+ page preliminary report that achieves the research goals set forth above. The full report is attached, and a summary of the report and some conclusions are as follows:

In section 1, we begin by building the case for 'robot ethics.' While there are substantial benefits to be gained from the use of military robots, there are also many opportunities for these machines to act inappropriately, especially as they are given greater degrees of autonomy (for quicker, more efficient, and more accurate decision-making, and if they are to truly replace human soldiers). The need for robot ethics becomes more urgent when we consider pressures driving the market for military robotics as well as long-standing public skepticism that lives in popular culture.

In section 2, we lay the foundation for a robot-ethics investigation by presenting a wide range of military robots—ground, aerial, and marine—currently in use and predicted for the future. While most of these robots today are semi-autonomous (e.g., the US Air Force's Predator), some apparently-fully autonomous systems are emerging (e.g., the US Navy's Phalanx CIWS) though used in a very limited, last-resort context. From this, we can already see ethical questions emerge, especially related to the ability to discriminate combatants from non-combatants and the

circumstances under which robots can make attack decisions on their own. We return to these questions in subsequent sections, particularly section 7.

In section 3, we look at behavioral frameworks that might ensure ethical actions in robots. It is natural to consider various programming approaches, since robots are related to our personal and business computer systems today that also depend on programmed instructions. We also recognize the forward-looking nature of our discussions here, given that the more-sophisticated programming abilities needed to build truly autonomous robotics are still under development.

We first discuss the traditional approach of *top-down programming*, i.e., establishing general rules that the robot would follow. A clear example is a deontological approach, such as using Kant's Categorical Imperative or Asimov's Laws of Robotics. However, a rigid set of rules is likely not robust enough to arrive at the correct action or decision in enough cases, particularly in unforeseen and complex scenarios. This suggests that we also need to attend to the 'rightness' of the result itself, not just to the rules. But even if we acknowledge that consequences matter, there are other challenges raised by adopting a consequentialist/utilitarian approach, such as the impracticality of calculating and weighing all possible results, both near and far term, and the (strong) possibility of countenancing some intuitively-wrong action.

Given the apparent limitations of top-down programming, we then examine *bottom-up approaches*, inspired by biological evolution and human development. However, a key challenge is that bottom-up systems work best when they are directed at achieving one clear goal, but military robots often operate in dynamic environments in which available information is confusing or incomplete. That is, even if moral calculation is not an issue, there still remains the large problem of moral psychology, i.e., how to develop robots that embody the right tendencies in their reactions to the world and other agents in that world, particularly when the robots are confronted with a novel situation in which they cannot rely on experience.

Moral reasoning by humans, however, is not limited to exclusively a top-down or bottom-up approach; rather, we often use both strategies of rule-following and experience. (Nonetheless, it is useful to evaluate both programming approaches separately to identify their benefits and challenges.) Therefore, we consider a *hybrid approach* of virtue ethics for constructing ethical autonomous robots. This approach is concerned with the development of moral character; in the military case, with promoting the ideal character traits of a warfighter, i.e., a 'warrior code of ethics' as its virtues.

In section 4, we look at considerations in programming the Laws of War (LOW) and Rules of Engagement (ROE), which may differ from mission to mission, into a robot. No matter which programming approach is adopted, we at least would want to the robot to obey the LOW and ROE, and this might serve as a proxy for full-fledged morality until we have the capability to program a

robot with the latter. Such an approach has several advantages, including: (1) any problems from moral relativism/particularism or other problems with general ethical principles are avoided; and (2) the relationship of morality to legality—a minefield for ethics—is likewise largely avoided, since the LOW and ROE make clear what actions are legal and illegal for robots, which serves as a reasonable approximation to the moral-immoral distinction.

Our discussion of the LOW and ROE, then, delves into their underlying foundation in just-war theory, particularly *jus ad bellum* (moral justification for entering war) and *jus in bello* (just and unjust actions in the prosecution of a war). We also examine ethical challenges to just-war theory as related to military robotics: Some have objected to the use of military robotics on the grounds that it makes easier the decision to enter war, in apparent violation of *jus ad bellum*; and we again see that the technical ability to properly discriminate against targets, as required by *jus in bello*, is a concern.

In section 5, we attend to the recurring possibility of accidental or unauthorized harm caused by robots; who would be responsible ultimately for those mishaps? We look at the issue through the lens of legal liability, both when robots are considered as merely products and when, as they are given more autonomy, they might be treated as legal agents, e.g., as legal quasi-persons such as children are regarded by the law. In the latter case, it is not clear how we would punish robots for their inappropriate actions.

In section 6, still attending to the possibility of unintended or unforeseen harm committed by a robot, we broaden our discussion by looking at how we might think about general risks posed by the machines and their acceptability. We offer a preliminary framework for a technology risk assessment, which includes the key factors of consent, informed consent, affected population, seriousness, and probability. This assessment highlights further the need for a lengthy period of rigorous testing and gradual rollout (crawl-walk-run approach) as a moral minimum for the responsible deployment of autonomous robots, especially by the military.

Finally, in section 7, we bring together a full range of issues raised throughout our examination, as well as some new issues, that must be recognized in any comprehensive assessment of risks from military robotics. These challenges fall into categories related to law, just-war theory, technical capabilities, human-robot interactions, general society, and other and future issues. For instance, we discuss such issues as:

- If a military robot refuses an order, e.g., if it has better situational awareness, then who would be responsible for its subsequent actions?
- How stringently should we take the generally-accepted 'eyes on target' requirement, i.e., under what circumstances might we allow robots to make attack decisions on their own?
- What precautions ought to be taken to prevent robots from running amok or turning against our own side, whether through malfunction, programming error, or capture and hacking?

- To the extent that military robots can help reduce instances of war crimes, what is the harm that may arise if the robots also unintentionally erode squad cohesion given their role as an 'outside' observer?
- Should robots be programmed to defend themselves—contrary to Arkin's position—given that they represent costly assets?
- Would using robots be counterproductive to winning the hearts and minds of occupied populations or result in more desperate terrorist-tactics given an increasing asymmetry in warfare?

From the preceding investigation, we can draw some general and preliminary conclusions, including some future work needed:

1. Creating autonomous military robots that can act *at least as* ethically as human soldiers appears to be a sensible goal, at least for the foreseeable future and in contrast to a greater demand of a perfectly-ethical robot. However, there are still daunting challenges in meeting even this relatively-low standard, such as the key difficulty of programming a robot to reliably distinguish enemy combatants from non-combatants, as required by the Laws of War and most Rules of Engagement.

2. While a faster introduction of robots in military affairs may save more lives of human soldiers and reduce war crimes committed, we must be careful to not unduly rush the process. Much different than rushing technology products to commercial markets, design and programming bugs in military robotics would likely have serious, fatal consequences. Therefore, a rigorous testing phase of robots is critical, as well as a thorough study of related policy issues, e.g., how the US Federal Aviation Administration (FAA) handles UAVs flying in our domestic National Airspace System (which we have not addressed here).

3. Understandably, much ongoing work in military robotics is likely shrouded in secrecy; but a balance between national security and public disclosure needs to be maintained in order to help accurately anticipate and address issues of risk or other societal concerns. For instance, there is little information on US military plans to deploy robots in space, yet this seems to be a highly strategic area in which robots can lend tremendous value; however, there are important environmental and political sensitivities that would surround such a program.

4. Serious conceptual challenges exist with the two primary programming approaches today: top-down (e.g., rule-following) and bottom-up (e.g., machine learning). Thus a hybrid approach should be considered in creating a behavioral framework. To this end, we need to a clear understanding of what a 'warrior code of ethics' might entail, if we take a virtue-ethics approach in programming.

5. In the meantime, as we wait for technology to sufficiently advance in order to create a workable behavioral framework, it may be an acceptable proxy to program robots to comply with the Law s of War and appropriate Rules of Engagement. However, this too is much easier said than done, and at least the technical challenge of proper discrimination would persist and require resolution.

6. Given technical limitations, such as programming a robot with the ability to sufficiently discriminate against valid and invalid targets, we expect that accidents will continue to occur, which raise the question of legal responsibility. More work needs to be done to clarify the chain of responsibility in both military and civilian contexts. Product liability laws are informative but untested as they relate to robotics with any significant degree of autonomy.

7. Assessing technological risks, whether through the basic framework we offer in section 6 or some other framework, depend on identifying potential issues in risk and ethics. These issues vary from: foundational questions of whether autonomous robotics can be legally and morally deployed in the first place, to theoretical questions about adopting precautionary approaches, to forward-looking questions about giving rights to truly autonomous robots. These discussions need to be more fully developed and expanded.

8. Specifically, the challenge of creating a robot that can properly discriminate among targets is one of the most urgent, particularly if one believes that the (increased) deployment of war robots is inevitable. While this is a major technical challenge—the solution to which depends on advances in programming and AI—there are also some workaround policy solutions that can be anticipated and further explored, such as: limiting deployment of lethal robots to only inside a 'kill box'; or designing a robot to target only other machines or weapons; or not giving robots a self-defense mechanism so that they may act more conservatively to prevent; or even creating robots with only non-lethal or less-than-lethal strike capabilities, at least initially until they are proven to be reliable.

These and other considerations warrant further, more detailed investigations in military robotics and issues of design, risk, and ethics. Such interdisciplinary investigations will require collaboration among policymakers and analysts, roboticists, ethicists, sociologists, psychologists, and others, internationally and including the general public as a key stakeholder. And this work has the potential to be as broad as other fields in science and society, such as bioethics or computer ethics.

The use of military robots represents a new era in warfare, perhaps more so than crossbows, airplanes, nuclear weapons, and other innovations have previously. Robots are not merely another asset in the military toolbox, but they are meant to also replace human soldiers, especially in 'dull, dirty, and dangerous' jobs. As such, they raise novel ethical and social questions that we should

confront as far in advance as possible—particularly before irrational public fears or accidents arising from military robotics derail research progress and national security interests.

5. Next Steps

With second-year funding from ONR, the following objectives are designed to leverage our first-year ONR project work—extending that investment and further disseminating our research. These additional activities will serve also to strengthen Cal Poly's profile in technology ethics, particularly with respect to robotics, to attract more projects across the university as well as interest from academic and industry partners.

(i) To develop one of the first anthologies in robot ethics, with a strong focus on military systems;
(ii) To develop and submit a robot-ethics paper at a relevant conference to more immediately engage the robotics community and broader public;
(iii) To develop one of the first university-level courses on robot ethics, with a strong focus on military systems;
(iv) To develop a website for robot ethics, as a way to disseminate research and engage the community and larger public;
(v) To expand our network of contacts—and Cal Poly's by extension—in the field through the above activities, including retaining new outside consultants for our second-year; and
(vi) To leverage this work into a much larger project supported by other federal and industry sources, such as NSF and DARPA.

6. Contact

Patrick Lin, Ph.D.
California Polytechnic State University
Ethics & Emerging Technologies Group
Philosophy Department
1 Grand Avenue
Building 47, Room 37
San Luis Obispo, California 93407
Email: palin@calpoly.edu
Dept. phone: 805-756-2041



Autonomous Military Robotics: Risk, Ethics, and Design

Prepared for: US Department of Navy, Office of Naval Research

Prepared by: Patrick Lin, Ph.D.
George Bekey, Ph.D.
Keith Abney, M.A.

Ethics & Emerging Technologies Group at
California State Polytechnic University, San Luis Obispo

Prepared on: December 20, 2008

This work is sponsored by the Department of the Navy, Office of Naval Research, under award # N00014-07-1-1152.

Preface

This report is designed as a preliminary investigation into the risk and ethics issues related to *autonomous military systems*, with a particular focus on battlefield robotics as perhaps the most controversial area. It is intended to help inform policymakers, military personnel, scientists, as well as the broader public who collectively influence such developments. Our goal is to raise the issues that need to be consider in responsibly introducing advanced technologies into the battlefield and, eventually, into society. With history as a guide, we know that foresight is critical to both mitigate undesirable effects as well as to best promote or leverage the benefits of technology.

In this report, we will present: the presumptive case for the use of autonomous military robotics; the need to address risk and ethics in the field; the current and predicted state of military robotics; programming approaches as well as relevant ethical theories and considerations (including the Laws of War, Rules of Engagement); a framework for technology risk assessment; ethical and social issues, both near- and far-term; and recommendations for future work.

This work is sponsored by the US Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152, whom we thank for its support and interest in this important investigation. We also thank California State Polytechnic University (Cal Poly, San Luis Obispo) for its support, particularly the College of Liberal Arts and the College of Engineering.

We are indebted to Colin Allen (Indiana Univ.), Peter Asaro (Rutgers Univ.), and Wendell Wallach (Yale) for their counsel and contributions, as well as to a number of colleagues—Ron Arkin (Georgia Tech), John Canning (Naval Surface Warfare Center), Ken Goldberg (IEEE Robotics and Automation Society; UC Berkeley), Patrick Hew (Defence Science and Technology Organization, Australia), George R. Lucas, Jr. (US Naval Academy), Frank Chongwoo Park (IEEE Robotics and Automation Society; Seoul National Univ.), Lt. Col. Gary Sargent (US Army Special Forces; Cal Poly), Noel Sharkey (Univ. of Sheffield, UK), Rob Sparrow (Monash Univ., Australia), and others—for their helpful discussions. We also thank the organizations mentioned herein for use of their respective images. Finally, we thank our families and nation's military for their service and sacrifice.

Patrick Lin
Keith Abney
George Bekey

December, 2008

1. Introduction

> "No catalogue of horrors ever kept men from war. Before the war you always think that it's not you that dies. But you will die, brother, if you go to it long enough."—Ernest Hemingway [1935, p.156]

Imagine the face of warfare with autonomous robotics: Instead of our soldiers returning home in flag-draped caskets to heartbroken families, autonomous robots—mobile machines that can make decisions, such as to fire upon a target, without human intervention—can replace the human soldier in an increasing range of dangerous missions: from tunneling through dark caves in search of terrorists, to securing urban streets rife with sniper fire, to patrolling the skies and waterways where there is little cover from attacks, to clearing roads and seas of improvised explosive devices (IEDs), to surveying damage from biochemical weapons, to guarding borders and buildings, to controlling potentially-hostile crowds, and even as the infantry frontlines.

These robots would be 'smart' enough to make decisions that only humans now can; and as conflicts increase in tempo and require much quicker information processing and responses, robots have a distinct advantage over the limited and fallible cognitive capabilities that we *Homo sapiens* have. Not only would robots expand the battlespace over difficult, larger areas of terrain, but they also represent a significant force-multiplier—each effectively doing the work of many human soldiers, while immune to sleep deprivation, fatigue, low morale, perceptual and communication challenges in the 'fog of war', and other performance-hindering conditions.

But the presumptive case for deploying robots on the battlefield is more than about saving human lives or superior efficiency and effectiveness, though saving lives and clearheaded action during frenetic conflicts are significant issues. Robots, further, would be unaffected by the emotions, adrenaline, and stress that cause soldiers to overreact or deliberately overstep the Rules of Engagement and commit atrocities, that is to say, war crimes. We would no longer read (as many) news reports about our own soldiers brutalizing enemy combatants or foreign civilians to avenge the deaths of their brothers in arms—unlawful actions that carry a significant political cost. Indeed, robots may act as objective, unblinking observers on the battlefield, reporting any unethical behavior back to command; their mere presence as such would discourage all-too-human atrocities in the first place.

Technology, however, is a double-edge sword with both benefits and risks, critics and advocates; and autonomous military robotics is no exception, no matter how compelling the case may be to pursue such research. The worries include: where responsibility would fall in cases of unintended or unlawful harm, which could range from the manufacturer to the field commander to even the machine itself; the possibility of serious malfunction and robots gone wild; capturing and hacking of military robots that are then unleashed against us; lowering the threshold for entering conflicts and wars, since fewer US military lives would then be at stake; the effect of such robots on squad cohesion, e.g., if robots recorded and reported back the soldier's every action; refusing an otherwise-legitimate order; and other possible harms.

We will evaluate these and other concerns within our report; and the remainder of this section will discuss the driving forces in autonomous military robotics and the need for 'robot ethics', as well as provide an overview of the report. Before that discussion, we should make a few introductory notes and definitions as follow.

1.1 Opening Remarks

First, in this investigation, we are *not* concerned with the question of whether it is even technically possible to make a perfectly-ethical robot, i.e., one that makes the 'right' decision in every case or even most cases. Following Arkin, we agree that an ethically-infallible machine ought not to be the goal now (if it is even possible); rather, our goal should be more practical and immediate: to design a machine that *performs better than* humans do on the battlefield, particularly with respect to reducing unlawful behavior or war crimes [Arkin, 2007]. Considering the number of incidences of unlawful behavior—and by 'unlawful' we mean a violation of the various Laws of War (LOW) or Rules of Engagement (ROE), which we also will discuss later in more detail—this appears to be a low standard to satisfy, though a profoundly important hurdle to clear. To that end, scientists and engineers need not first solve the daunting task of creating a truly 'ethical' robot, at least in the foreseeable future; rather, it seems that they only need to program a robot to act in compliance with the LOW and ROE (though this may not be as straightforward and simply as it first appears) or act ethically in the specific situations in which the robot is to be deployed.

Second, we should note that the purpose of this report is not to encumber research on autonomous military robotics, but rather to help responsibly guide it. That there should be two faces to technology—benefits and risk—is not surprising, as history shows, and is not by itself an argument against that technology.[1] But ignoring those risks, or at least only reactively addressing them and

[1] Biotechnology, for instance, promises to reduce world hunger by promoting greater and more nutritious agricultural and livestock yield; yet continuing concerns about the possible dissemination of bio-engineered seeds (or 'Frankenfoods') into the wild, displacing native plants and crops, have prompted the industry to move more cautiously [e.g., Thompson, 2007]. Even Internet technologies, as valuable as they have been in connecting us to information, social networks, etc., and in making new ways of life possible, reveal a darker world of online scams,

waiting for public reaction, seems to be unwise, given that it can lead (and, in the case of biotech foods, has led) to a backlash that stalls forward progress.

That said, it is surprising to note that one of the most comprehensive and recent reports on military robotics, *Unmanned Systems Roadmap 2007-2032*, does not mention the word 'ethics' once nor risks raised by robotics, with the exception of one sentence that merely acknowledges that "privacy issues [have been] raised in some quarters" without even discussing said issues [US Department of Defense, 2007, p. 48]. While this omission may be understandable from a public relations standpoint, again it seems short-sighted given lessons in technology ethics, especially from our recent past. Our report, then, is designed to address that gap, proactively and objectively engaging policymakers and the public to head off a potential backlash that serves no one's interests.

Third, while this report focuses on issues related to autonomous military *robotics*, the discussion may apply equally well and overlap with issues related to autonomous military *systems*, i.e., computer networks. Further, we are focusing on *battlefield* or lethal applications, as opposed to robotics in manufacturing or medicine even if they are supported by military programs (such as the Battlefield Extraction Assist Robot, or BEAR, that carries injured soldiers from combat zones), for several reasons as follow. The most contentious military robots will be the weaponized ones: "Weaponized unmanned systems is a highly controversial issue that will require a patient 'crawl-walk-run' approach as each application's reliability and performance is proved" [US Department of Defense, 2007, p. 54]. Their deployment is inherently about human life and death, both intended and unintended, so they immediately raise serious concerns related to ethics (e.g., does just-war theory or the LOW/ROE allow for deployment of autonomous fighting systems in the first place?) as well as risk (e.g., malfunctions and emergent, unexpected behavior) that demand greater attention than other robotics applications.

Also, though a relatively small number of military personnel is ever exposed on the battlefield, loss of life and property during armed conflict has non-trivial political costs, never mind environmental and economic costs, especially if 'collateral' or unintended damage is inflicted and even more so if it results from abusive, unlawful behavior by our own soldiers. How we prosecute a war or conflict receives particular scrutiny from the media and public, whose opinions influence military and foreign policy even if those opinions are disproportionately drawn from events on the battlefield, rather than on the many more developments outside the military theater. Therefore, though autonomous battlefield or weaponized robots may be years away and account for only one segment of the entire military robotics population, there is much practical value in sorting through their associative issues sooner rather than later.

---

privacy violations, piracy, viruses, and other ills; yet no one suggests that we should do away with cyberspace [e.g., Weckert, 2007].

Fourth and finally, while our investigation here is supported by the US Department of the Navy, Office of Naval Research, it may apply equally well to other branches of military service, all of which are also developing robotics for their respective needs. The range of robotics deployed or under consideration by the Navy, however, is exceptionally broad, with airborne, sea surface, underwater, and ground applications.[2] Thus, it is particularly fitting for the Department of the Navy to support one of the first dedicated investigations on the risk and ethical issues arising from the use of autonomous military robotics.

1.2 Definitions

To the extent that there are no standard, universally-accepted definitions of some of the key terms we employ in this report, we will need to stipulate those working definitions here, since it is important that we ensure we have the same basic understanding of those terms at the outset. And so that we do not become mired in debating precise definitions here, we provide a detailed discussion or justification for our definitions in 'Appendix A: Definitions'.

Robot (particularly in a military context). *A powered machine that (1) senses, (2) thinks (in a deliberative, non-mechanical sense), and (3) acts.*

Most robots are and will be mobile, such as vehicles, but this is not an essential feature; however, some degree of mobility is required, e.g., a fixed sentry robot with swiveling turrets or a stationary industrial robot with movable arms. Most do not and will not carry human operators, but this too is not an essential feature; the distinction becomes even more blurred as robotic features are integrated with the body. Robots can be operated semi- or fully-autonomously but cannot depend entirely on human control: for instance, tele-operated drones such as the Air Force's Predator unmanned aerial vehicle would qualify as robots to the extent that they make some decisions on their own, such as navigation, but a child's toy car tethered to a remote control is not a robot since its control depends entirely on the operator. Robots can be expendable or recoverable, and can carry a lethal or non-lethal payload. And robots can be considered as agents, i.e., they have the capacity to act in a world, and some even may be moral agents, as discussed in the next definition.

Autonomy (in machines). *The capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.*

This is to say that, we are herein not interested in issues traditionally linked to autonomy that require a more robust and precise definition, such as the assignment of political rights and moral

[2] The only applications not covered by the Department of the Navy appear to be underground- and space-based, including sub-orbital missions, which may understandably fall outside their purview.

responsibility (as different from legal responsibility) or even more technical issues related to free will, determinism, personhood, and whether machines can even 'think'—as important as those issues are in philosophy, law, and ethics. But in the interest of simplicity, we will stipulate this definition, which seems acceptable in a discussion limited to human-created machines. This term also helps elucidate the second criterion of 'thinking' in our working definition of a robot. Autonomy is also related to the concept of moral agency, i.e., the ability to make moral judgments and choose one's actions accordingly.

Ethics (construed broadly for this report). *More than normative issues, i.e., questions about what we should or ought to do, but also general concerns related to social, political, and cultural impact as well as risk arising from the use of robotics.*

As a result, we will cover all these areas in our report, not just philosophical questions or ethical theory, with the goal of providing some relevant if not actionable insights at this preliminary stage. We will also discuss relevant ethical theories in more detail in section 3 (though this is not meant to be a comprehensive treatment of the subject).

1.3 Market Forces and Considerations

Several industry trends and recent developments—including high-profile failures of *semi*-autonomous systems, as perhaps a harbinger of challenges with more advanced systems—highlight the need for a technology risk assessment, as well as a broader study of other ethical and social issues related to the field. In the following, we will briefly discuss seven primary market forces that are driving the development of military robotics as well as the need for a guiding ethics; these roughly map to what have been called 'push' (technology) and 'pull' (social and cultural) factors [US Department of Defense, 2007, p.44].

1. *Compelling military utility*. US defense organizations are attracted to the use of robots for a range of benefits, some of which we have mentioned above. A primary reason is to replace us less-durable humans in "dull, dirty, and dangerous" jobs [US Department of Defense, 2007, p.19]. This includes: extended reconnaissance missions, which stretch the limits of human endurance to its breaking point; environmental sampling after a nuclear or biochemical attack, which had previously led to deaths and long-term effects on the surveying teams; and neutralizing IEDs, which have caused over 40% of US casualties in Iraq since 2003 [Iraq Coalition Casualty Count, 2008]. While official statistics are difficult to locate, news organizations report that the US has deployed over 5,000 robots in Iraq and Afghanistan, which have neutralized 10,000 IEDs by 2007 [CBS, 2007].

Also mentioned above, military robots may be more discriminating, efficient, and effective. Their dispassionate and detached approach to their work could significantly reduce the instances of unethical behavior in wartime—abuses that negatively color the US prosecution of a conflict, no matter how just the initial reasons to enter the conflict are, and carry a high political cost.

2. *US Congressional deadlines.* Clearly, there is a tremendous advantage to employing robots on the battlefield, and the US government recognizes this. Two key Congressional mandates are driving the use of military robotics: by 2010, one-third of all operational deep-strike aircraft must be unmanned, and by 2015, one-third of all ground combat vehicles must be unmanned [National Defense Authorization Act, 2000]. Most, if not all, of the robotics in use and under development are semi-autonomous at best; and though the technology to (responsibly) create fully autonomous robots is near but not quite in hand, we would expect the US Department of Defense to adopt the same, sensible 'crawl-walk-run' approach as with weaponized systems, given the serious inherent risks.

   Nonetheless, these deadlines apply increasing pressure to develop and deploy robotics, including autonomous vehicles; yet a 'rush to market' increases the risk for inadequate design or programming. Worse, without a sustained and significant effort to build in ethical controls in autonomous systems, or even to discuss the relevant areas of ethics and risk, there is little hope that the early generations of such systems and robots will be adequate, making mistakes that may cost human lives. (This is related to the 'first-generation' problem we discuss in sections 6 and 7, that we won't know exactly what kind of errors and mistaken harms autonomous robots will commit until they have already done so.)

3. *Continuing unethical battlefield conduct.* Beyond popular news reports and images of purportedly unethical behavior by human soldiers, the US Army Surgeon General's Office had surveyed US troops in Iraq on issues in battlefield ethics and discovered worrisome results. From its summary of findings, among other statistics: "Less than half of Soldiers and Marines believed that non-combatants should be treated with respect and dignity and well over a third believed that torture should be allowed to save the life of a fellow team member. About 10% of Soldiers and Marines reported mistreating an Iraqi non-combatant when it wasn't necessary...Less than half of Soldiers and Marines would report a team member for unethical behavior...Although reporting ethical training, nearly a third of Soldiers and Marines reported encountering ethical situations in Iraq in which they didn't know how to respond" [US Army Surgeon General's Office, 2006]. The most recent survey by the same organization reported similar results [US Army Surgeon General's Office, 2008].

   Wartime atrocities have occurred since the beginning of human history, so we are not operating under the illusion that they can be eliminated altogether (nor that armed conflicts can be eliminated either, at least in the foreseeable future). However, to the extent that military robots

can considerably reduce unethical conduct on the battlefield—greatly reducing human and political costs—there is a compelling reason to pursue their development as well as to study their capacity to act ethically.

4. *Military robotics failures.* More than theoretical problems, military robotics have already failed on the battlefield, creating concerns with their deployment (and perhaps even more concern for more advanced, complicated systems) that ought to be addressed before speculation, incomplete information, and hype fill the gap in public dialogue.

   In April 2008, several TALON SWORDS units—mobile robots armed with machine guns—in Iraq were reported to be grounded for reasons not fully disclosed, though early reports claim the robots, without being commanded to, trained their guns on 'friendly' soldiers [e.g., Page, 2008]; and later reports denied this account but admitted there had been malfunctions during the development and testing phase prior to deployment [e.g., Sofge, 2008]. The full story does not appear to have yet emerged, but either way, the incident underscores the public's anxiety—and the military's sensitivity—with the use of robotics on the battlefield (also see 'Public perceptions' below).

   Further, it is not implausible to suggest that these robots may fail, because it has already happened elsewhere: in October 2007, a semi-autonomous robotic cannon deployed by the South African army malfunctioned, killing nine 'friendly' soldiers and wounding 14 others [e.g., Shachtman, 2007]. Communication failures and errors have been blamed for several unmanned aerial vehicle (UAV) crashes, from those owned by the Sri Lanka Air Force to the US Border Patrol [e.g., BBC, 2005; National Transportation Safety Board, 2007]. Computer-related technology in general is especially susceptible to malfunctions and 'bugs' given their complexity and even after many generations of a product cycle; thus, it is reasonable to expect similar challenges with robotics.

5. *Related civilian systems failures.* On a similar technology path as autonomous robots, civilian computer systems have failed and raised worries that can carry over to military applications. For instance, such civilian systems have been blamed for massive power outages: in early 2008, Florida suffered through massive blackouts across the entire state, as utility computer systems automatically shut off and rerouted power after just a small fire caused by a failed switch at one electrical substation [e.g., Padgett, 2008]; and in the summer 2003, a single fallen tree had triggered a tsunami of cascading computer-initiated blackouts that affected tens of millions of customers for days and weeks across the eastern US and Canada, leaving practically no time for human intervention to fix what should have been a simple problem of stopping the disastrous chain reaction [e.g., US Department of Energy, 2004]. Thus, it is a concern that we also may not be able to halt some (potentially-fatal) chain of events caused by autonomous military systems

that process information and can act at speeds incomprehensible to us, e.g., with high-speed unmanned aerial vehicles.

Further, civilian robotics are becoming more pervasive. Never mind seemingly-harmless entertainment robots, some major cities (e.g., Atlanta, London, Paris, Copenhagen) already boast driverless transportation systems, again creating potential worries and ethical dilemmas (e.g., bringing to life the famous thought-experiment in philosophy: should a fast-moving train divert itself to another track in order to kill only one innocent person, or continue forward to kill the five on its current path?). So there can be lessons for military robotics that can be transferred from civilian robotics and automated decision-making, and vice versa. Also, as robots become more pervasive in the public marketplace—they are already abundant in manufacturing and other industries—the broader public will become more aware of risk and ethical issues associated with such innovations, concerns that inevitably will carry over to the military's use.

6. *Complexity and unpredictability*. Perhaps robot ethics has not received the attention it needs, at least in the US, given a common misconception that robots will do only what we have programmed them to do. Unfortunately, such a belief is a sorely outdated, harking back to a time when computers were simpler and their programs could be written and understood by a single person. Now, programs with millions of lines of code are written by teams of programmers, none of whom knows the entire program; hence, no individual can predict the effect of a given command with absolute certainty, since portions of large programs may interact in unexpected, untested ways. (And even straightforward, simple rules such as Asimov's Laws of Robotics can create unexpected dilemmas [e.g., Asimov, 1950].) Furthermore, increasing complexity may lead to *emergent behaviors*, i.e., behaviors not programmed but arising out of sheer complexity [e.g., Kurzweil, 1999, 2005].

   Related major research efforts also are being devoted to enabling robots to learn from experience, raising the question of whether we predict with reasonable certainty *what* the robot will learn. The answer seems to be negative, since if we could predict that, we would simply program the robot in the first place, instead of requiring learning. Learning may enable the robot to respond to novel situations, given the impracticality and impossibility of predicting all eventualities on the designer's part. Thus, unpredictability in the behavior of complex robots is a major source of worry, especially if robots are to operate in unstructured environments, rather than the carefully-structured domain of a factory. (We will discuss machine learning further in sections 2 and 3.)

7. *Public perceptions*. From Asimov's science fiction novels to Hollywood movies such as *Wall-E, Iron Man, Transformers, Blade Runner, Star Wars, Terminator, Robocop, 2001: A Space Odyssey,* and *I, Robot* (to name only a few, from the iconic to recently released), robots have captured the

> global public's imagination for decades now. But in nearly every one of those works, the use of robots in society is in tension with ethics and even the survival of humankind. The public, then, is already sensitive to the risks posed by robots—whether or not those concerns are actually justified or plausible—to a degree unprecedented in science and technology. Now, technical advances in robotics is catching up to literary and theatrical accounts, so the seeds of worry that have long been planted in the public consciousness will grow into close scrutiny of the robotics industry with respect to those ethical issues, e.g., the book *Love and Sex with Robots* published late last year that reasonably anticipates human-robot relationships [Levy, 2007].

Given such investments, questions, events, and predictions, it is no wonder that more attention is being paid to robot ethics, particularly in Europe [e.g., Veruggio, 2007]. An entire conference dedicated to the issue of ethics in autonomous military systems—one of the first we have seen, if not the first of its kind—was held in late February 2008 in the UK [Royal United Services Institute (RUSI) for Defence and Security Studies, 2008], in which experts reiterated the possibility that robots might commit war crimes or be turned on us by terrorists and criminals [RUSI, 2008: Noel Sharkey and Rear Admiral Chris Parry's presentations, respectively; also, Sharkey, 2007a, and Asaro, 2008]. Robotics is a particularly thriving and advanced industry in Asia: South Korea is the first (and still only?) nation to be working on a 'Robot Ethics Charter' or a code of ethics to govern responsible robotics development and use, though the document has yet to materialize [BBC, 2007]. This summer, Taiwan played host to a conference about advanced robotics and its societal impacts [Institute of Electrical and Electronics Engineers (IEEE), 2008].

But the US is starting to catch up: some notable US experts are working on similar issues, which we will discuss throughout this report [Arkin, 2007; Wallach and Allen, 2008]. A January 2008 conference at Stanford University focused on technology in wartime, of which robot ethics was one notable session [Computer Professionals for Social Responsibility (CPSR), 2008]. In July 2008, the North American Computing and Philosophy (NA-CAP) conference at Indiana University focused a significant part of its program on robot ethics [NA-CAP, 2008]. Again, we intend for this report as an early, complementary step in filling the gap in robot-ethics research, both technical and theoretical.

1.4 Report Overview

Following this introduction, in section 2, we will provide a short background discussion on robotics in general and in defense applications specifically. We will survey briefly the current state of robotics in the military as well as developments in progress and anticipated. This includes several future scenarios in which the military may employ autonomous robots, which will help anchor and add depth to our discussions later on ethics and risk.

In section 3, we will discuss the possibility of programming in rules or a framework in robots to govern their actions (such as Asimov's Laws of Robotics). There are different programming approaches: top-down, bottom-up, and a hybrid approach [Wallach and Allen, 2008]. We also discuss the major (competing) ethical theories—deontology, consequentialism, and virtue ethics—that these approaches correspond with as well as their limitations.

In section 4, we consider an alternative, as well as a complementary approach, to programming a robot with an ethical behavior framework: to simply program it to obey the relevant Laws of War and Rules of Engagement. To that end, we also discuss the relevant LOW and ROE, including a discussion of just-war theory and related issues that may arise in the context of autonomous robots.

In section 5, continuing the discussion about law, we will also look at the issue of legal responsibility based on precedents related to product liability, negligence and other areas [Asaro, 2007]. This at least informs questions of risk in the near- and mid-term in which robots are essentially human-made tools and not moral agents of their own; but we also look at the case for treating robots as quasi-legal agents.

In section 6, we will broaden our discussion in providing a framework for technology risk assessment. This framework includes a discussion of the major factors in determining 'acceptable risk': consent, informed consent, affected population, seriousness, and probability [DesJardins, 2003].

In section 7, we will bring the various ethics and social issues discussed, and new ones, together in one location. We will survey a full range of possible risks and issues related to ethics, just-war theory, technical challenges, societal impact, and more. These contingencies and issues are important to have in mind in any complete assessment of technology risks.

Finally, in section 8, we will draw some preliminary conclusions, including recommendations for future, more detailed investigations. A bibliography is provided as section 9 of the report; and appendix A offers more detailed discussions on key definitions, as initiated in this section.

2. Military Robotics

The field of robotics has changed dramatically during the past 30 years. While the first programmable articulated arms for industrial automation were developed by George Devol and made into commercial products by Joseph Engleberger in the 1960s and 1970s, mobile robots with various degrees of autonomy did not receive much attention until the 1970s and 1980s. The first true mobile robots arguably were Elmer and Elsie, the electromechanical 'tortoises' made by W. Grey Walter, a physiologist, in 1950 [Walter, 1950]. These remarkable little wheeled machines had many of the features of contemporary robots: sensors (photocells for seeking light and bumpers for obstacle detection), a motor drive and built-in behaviors that enabled them to seek (or avoid) light, wander, avoid obstacles and recharge their batteries. Their architecture was basically reactive, in that a stimulus directly produced a response without any 'thinking.' That development first appeared in Shakey, a robot constructed at Stanford Research Laboratories in 1969 [Fikes and Nilsson, 1971]. In this machine, the sensors were not directly coupled to the drive motors but provided inputs to a 'thinking' layer known as the Stanford Research Institute Problem Solver (STRIPS), one of the earliest applications of artificial intelligence. The architecture was known as 'sense-plan-act' or 'sense-think-act' [Arkin, 1998].

Since those early developments, there have been major strides in mobile robots—made possible by new materials, faster, smaller and cheaper computers (Moore's law) and major advances in software. At present, robots move on land, in the water, in the air, and in space. Terrestrial mobility uses legs, treads, and wheels as well as snake-like locomotion and hopping. Flying robots make use of propellers, jet engines, and wings. Underwater robots may resemble submarines, fish, eels, or even lobsters. Some vehicles capable of moving in more than one medium or terrain have been built. Service robots, designed for such applications as vacuum cleaning, floor washing and lawn mowing, have been sold in large quantities in recent years. Humanoid robots, long considered only in science fiction novels, are now manufactured in various sizes and with various degrees of sophistication [Bekey, 2005]. Small toy humanoids, such as the WowWee Corporation's RoboSapien, have been sold in quantities of millions. More complex humanoids, such as the Honda ASIMO are able to perform numerous tasks. However, 'killer applications' for humanoid robots have not yet emerged.

There has also been great progress in the development of software for robots, including such applications as learning, interaction with humans, multiple robot cooperation, localization and navigation in noisy environments, and simulated emotions. We discuss some of these developments briefly in section 2.6 below.

During the past 20 years, military robotic vehicles have been built using all the modes of locomotion described above and making use of the new software paradigms [US Dept. Of Defense, 2007]. Military robots find major applications in surveillance, reconnaissance, location and destruction of mines and IEDs, as well as for offense or attack. The latter class of vehicles is equipped with weapons, which at the present time are fired by remote human controllers. In the following, we first summarize the state of the art in military robots, including both hardware and software, and then introduce some of the ethical issues which arise from their use. We concentrate on robots capable of lethal action—in that much of the concern with military robotics is tied to this lethality—and omit discussion of more innocuous machines such as the Army's Big Dog, a four legged robot capable of carrying several hundred pounds of cargo over irregular terrain. If at some future time such 'carry robots' are equipped with weapons, they may need to be considered from an ethical point of view.

2.1 Ground Robots

The US Army makes use of two major types of autonomous and semi-autonomous ground vehicles: large vehicles, such as tanks, trucks and HUMVEEs and small vehicles, which may be carried by a soldier in a backpack (such as the PackBot shown in Fig. 2.0a) and move on treads like small tanks [US Dept. Of Defense, 2007]. The PackBot is equipped with cameras and communication equipment and may include manipulators (arms); it is designed to find and detonate IEDs, thus saving lives (both civilian and military), as well as to perform reconnaissance. Its small size enables it to enter buildings, report on possible occupants, and trigger booby traps. Typical armed robot vehicles are (1) the Talon SWORDS (Special Weapons Observation Reconnaissance Detection System) made by Foster-Miller, which can be equipped with machine guns, grenade launchers, or anti-tank rocket launchers as well as cameras and other sensors (see Fig. 2.0b) and (2) the newer MAARS (Modular Advanced Armed Robotic System). While vehicles such as SWORDS and the newer MAARS are able to autonomously navigate toward specific targets through its global positioning system (GPS), at present the firing of any on-board weapons is done by a soldier located a safe distance away. Foster-Miller provides a universal control module for use by the warfighter with any of their robots. MAARS uses a more powerful machine gun than the original SWORDS. While the original SWORDS weighted about 150 lbs., MAARS weighs about 350 lbs. It is equipped with a new manipulator capable of lifting 100 lbs., thus enabling it to replace its weapon platform with an IED identification and neutralization unit.

Among the larger vehicles, the Army's Tank-Automotive Research, Development and Engineering Center (jointly with Foster-Miller) has developed the TAGS-CX, a 5,000-6,000 lb. amphibious vehicle. More recently, and jointly with Carnegie Mellon University, the Army has developed a 5.5 ton, six-wheel unmanned vehicle known as the Crusher, capable of carrying 2,000 lbs. at about 30 mph and capable of withstanding a mine explosion; it is equipped with one or more guns (see figure 2.1).



(a) (b)

Fig. 2.0 Military ground vehicles: (a) PackBot (Courtesy of iRobot Corp.); (b) SWORDS (Courtesy of Foster-Miller Corp.)



Fig. 2.1 Military ground vehicle: The Crusher (Courtesy of US Army)

Both PackBot and Talon robots are being used extensively and successfully in Iraq and Afghanistan. Hence, we expect further announcements of UGV deployments in the near future. We are not aware of the use of armed sentry robots by the US military; however, they are used in South Korea (developed by Samsung) and in Israel. The South Korean system is capable of interrogating suspects, identifying potential enemy intruders, and autonomous firing of its weapon.

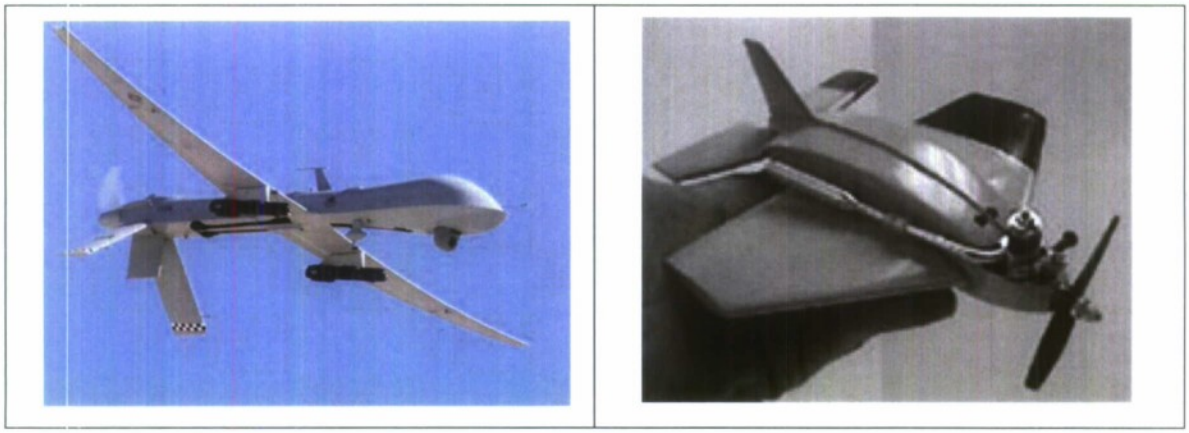
DARPA supported two major national competitions leading to the development of autonomous ground vehicles. The 2005 Grand Challenge required autonomous vehicles to traverse portions of the Mojave desert in California. The vehicles were provided with GPS coordinates of way-points along the route, but otherwise the terrain to be traversed was completely unknown to the designers, and the vehicles moved autonomously at speed averaging 20 to 30 mph. In 2007, the Urban Challenge required autonomous vehicles to move in a simulated urban environment, in the presence of other vehicles and signal lights, while obeying traffic laws. While the winning automobiles from Stanford University and Carnegie Mellon University were not military in nature, the lessons learned will undoubtedly find their way into future generations of autonomous robotic vehicles developed by the Army and other services.

2.2 Aerial Robots

The US Army, Air Force, and Navy have developed a variety of robotic aircraft known as unmanned flying vehicles (UAVs).[3] Like the ground vehicles, these robots have dual applications: they can be used for reconnaissance without endangering human pilots, and they can carry missiles and other weapons. The services use hundreds of unarmed drones, some as small as a model aiplane, to locate and identify enemy targets. An important function for unarmed UAVs is to serve as aerial targets for piloted aircraft, such as those manufactured by AeroMech Engineering in San Luis Obispo, CA, a company started by Cal Poly students. AeroMech has sold some 750 UAVs, ranging from 4 lb. battery-operated ones to 150 lb. vehicles with jet engines. Some reconnaissance UAVs, such as the Shadow, are launched by a catapult and can stay aloft all day. The best known armed UAVs are the semi-autonomous Predator Unmanned Combat Air Vehicles (UCAV) built by General Atomics (see Fig. 2.2a), which can be equipped with Hellfire missiles. Both the Predator and the larger Reaper hunter-killer aircraft are used extensively in Afghanistan. They can navigate autonomously toward targets specified by GPS coordinates, but a remote operator located in Nevada (or in Germany) makes the final decision to release the missiles. The Navy, jointly with Northrop Grumman, is developing an unmanned bomber with folding wings which can be launched from an aircraft carrier.

The military services are also developing very small aircraft, sometimes called Micro Air Vehicles (MAV) capable of carrying a camera and sending images back to their base. An example is the Micro Autonomous Air Vehicle (MAAV; also called MUAV for Micro Unmanned Air Vehicle) developed by Intelligent Automation, Inc., which is not much larger than a human hand (see Fig. 2.2b).

[3] Earlier versions of such vehicles were termed 'drones', which implied that they were completely under control of a pilot in a chaser aircraft. Current models are highly autonomous, receiving destination coordinates from only ground or satellite transmitters. Thus, because this report is focused on robots—machines that have some degree of autonomy—we do not use the term 'drone' here.



(a) (b)

Fig. 2.2 Autonomous oircroft: (o) Predotor (Courtesy of Generol Atomics Aeronouticol Systems); (b) Micro unmonned flying vehicle (Courtesy of Intelligent Automation, Inc.)

Similarly, the University of Florida has developed an MAV with a 16-inch wingspan with foldable wings, which can be stored in an 8-inch x 4-inch container. Other AUVs include a ducted fan vehicle (see Fig. 2.3a) being used in Iraq, and vehicles with flapping wings, made by AeroVironment and others (Fig. 2.3b). While MAVs are used primarily for reconnaissance and are not equipped with lethal weapons, it is conceivable that the vehicle itself could be used in 'suicide' missions.



(a) (b)

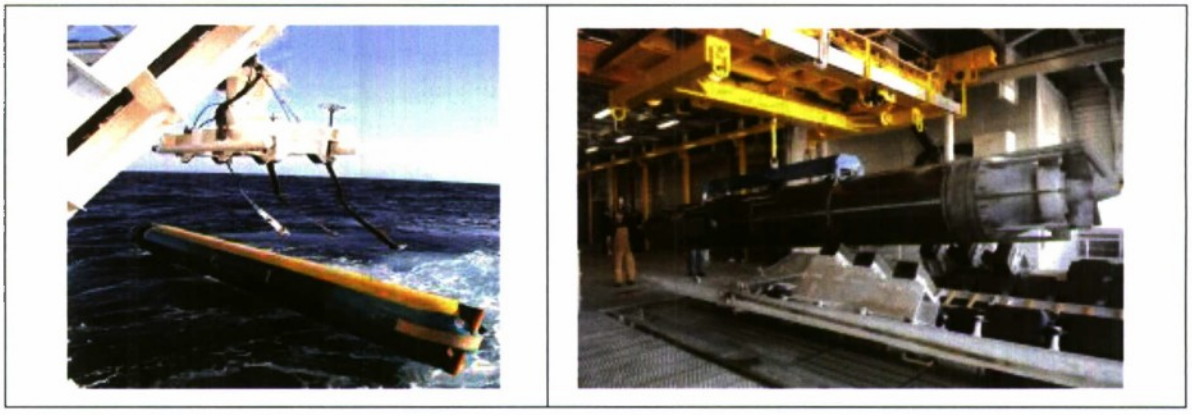
Fig. 2.3 Micro oir vehicles: (o) ducted fon vehicle from Honeywell; (b) Ornithopter MAV with flopping wings mode by students ot Brighom Young University (Photo by Joren Wilkey/BYU, used by permission)

Other flying robots either deployed or in development, including helicopters, tiny robots the size of a bumblebee, and solar-powered craft capable of remaining aloft for days or weeks at a time. Again, our objective here is not to provide a complete survey, but to indicate the wide range of mobile robots in use by the military services.

2.3 Marine Robots

Along with the other services, the US Navy has a major robotic program involving interaction between land, airborne, and seaborne vehicles [US Dept. of the Navy, 2004; US Dept. of Defense, 2007]. The latter include surface ships as well as Unmanned Underwater Vehicles (UUVs). Their applications include surveillance, reconnaissance, anti-submarine warfare, mine detection and clearing, oceanography, communications, and others. It should be noted that contemporary torpedoes may be classified as UUVs, since they possess some degree of autonomy.

As with robots in the other services, UUVs come in various sizes, from man-portable to very large. Fig. 2.4a shows Boeing's Long-term Mine Reconnaissance System (LMRS) which is dropped into the ocean from a telescoping torpedo launcher aboard the SV Ranger to begin its underwater surveillance test mission. LMRS uses two sonar systems, an advanced computer and its own inertial navigation system to survey the ocean floor for up to 60 hours. The LMRS shown in the figure is about 21 inches in diameter; it can be launched from a torpedo tube, operate autonomously, return to the submarine, and be guided into a torpedo-tube mounted robotic recovery arm. A large UUV, the Seahorse, is shown in Fig. 2.4b; this vehicle is advertised as being capable of 'independent operations', which may include the use of lethal weapons. The Seahorse is about 3 feet in diameter, 28 feet long, and weighs 10,500 lbs. The Navy plans to move toward deployment of large UUVs by 2010. These vehicles may be up to 3 to 5 feet in diameter, weighing perhaps 20,000 lbs.

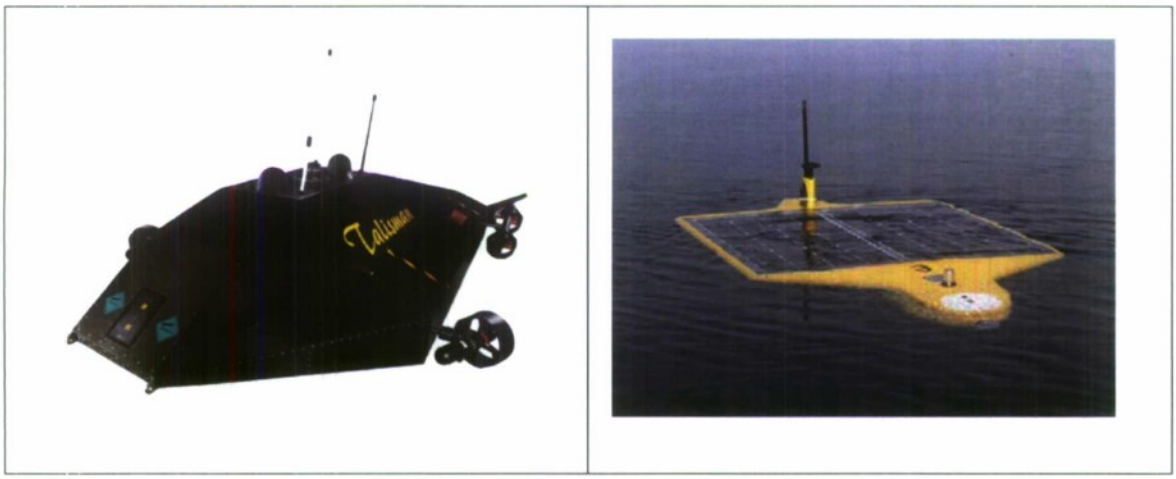


(a) (b)

Figure 2.4: (a) Long-term Mine Reconnaissance UUV (Courtesy of The Boeing Company); (b) Seahorse 3-foot diameter UUV (Courtesy of Penn State University)

Development of UUVs is not restricted to the US. Large UUV programs exist in Australia, Great Britain, Sweden, Italy, Russia, and other countries. Fig. 2.5a shows a UUV made in Great Britain by BAE Systems.

A solar-powered surface vehicle is shown in Fig. 2.5b. As with other military robots, most of the vehicles capable of delivering deadly force are currently human-controlled and not fully autonomous. However, the need for autonomy is great for underwater vehicles, since radio communication underwater is difficult. Many UUVs surface periodically to send and receive messages.



(a) (b)

Fig. 2.5: (a) Talisman UUV (Courtesy of BAE Systems);
(b) Solar powered surface vehicle (Courtesy of NOAA)

2.4 Space Robots

We believe that the US Armed Services have significant programs for the development of autonomous space vehicles: for advanced warning, defense against attacking missiles and possibly offensive action as well. However, there is very little information on these programs in publicly available sources. It is clear that the Air Force is building a major communication system in space, named Transformational Satellite Communication System (TSC). This system will interact with airborne as well as ground-based communication nodes to create a truly global information grid.

2.5 Immobile/Fixed Robots

To this point we have described a range of mobile robots used by the military: on earth, on and under the water, in the air, and in space. It should be noted that not all robots capable of lethal action are mobile; in fact, some are stationary, with only limited mobility (such as aiming of a gun). We consider a few examples of such robots in this section.

First, let us consider again why land mines and underwater mines, whether aimed at destruction of vehicles or attacks on humans (anti-personnel mines), are not properly robots. Whether buried in the ground or planted in the surf zone along the ocean shore, these systems are equipped with some sensing ability (since they can detect the presence of weight), and they 'act' by exploding. Their information processing ability is extremely limited, generally consisting only of a switch triggered by pressure from above. Given our definition of autonomous robots as consider in section 1 (as well as detailed in Appendix A), while such mines may be considered as autonomous, we do not classify them as robots since a simple trigger is not equivalent to the cognitive functions of a robot. If a landmine is considered a robot, one seems to be absurdly required to designate a trip wire as a robot too.

On the other hand, there are immobile or stationary weapons, both on land and on ships, which do merit the designation of robot, despite their lack of mobility (though they have some moving features, which satisfies our definition for what counts as a robot). An example of such a system is the Navy's Phalanx Close-In Weapon System (CIWS). CIWS is a rapid-fire 20mm gun system designed to protect ships at close range from missiles which have penetrated other defenses. The system is mounted on the deck of a ship; it is equipped with both search and tracking radars and the ability to rotate a turret in order to aim the guns. The information processing ability of the computer system associated with the radars is remarkable, since it automatically performs search, detecting, tracking, threat evaluation, firing, and kill-assessments of targets. Thus, the CIWS uses radar sensing of approaching missiles, identifies targets, tracks targets, makes the decision to fire, and then fires its guns, using solid tungsten bullets to penetrate the approaching target. The gun-and-radar turret can rotate in at least two degrees of freedom for target tracking, but the entire structure is immobile and fixed on the deck.

The US Army has also adopted a version of the Phalanx system to provide close-in protection for troops and facilities in Iraq, under the name 'Counter Rocket, Artillery, and Mortar' (C-RAM, or Counter-RAM). The system is mounted on the ground or, in some cases, on a train platform. The basic system operation is similar to that of the Navy system: it is designed to destroy incoming missiles at a relatively short range. However, since the system is located adjacent to or near civilian facilities, there is major concern for collateral damage, e.g., debris or fragments of a disabled missile could land on civilians.

As a final example here, we cite the SGR-A1 sentry robot developed by Samsung Techwin Co. for use by the South Korean army in the Demilitarized Zone (DMZ) which separates North and South Korea. The system is stationary, designed to replace a manned sentry location. It is equipped with sophisticated color vision sensors that can identify a person entering the DMZ, even at night under only starlight illumination. Since any person entering the DMZ is automatically presumed to be an enemy, it is not necessary to separate friend from foe. The system is equipped with a machine gun, and the sensor-gun assembly is capable of rotating in two degrees of freedom as it tracks a target. The firing of the gun can be done manually by a soldier or by the robot in fully-automatic (autonomous) mode.

2.6 Robot Software Issues

In the preceding, we have presented the current state of some of the robotic hardware and systems being used and/or being developed by the military services. It is important to note that in parallel with the design and fabrication of new autonomous or semi-autonomous robotic systems, there is a great deal of work on fundamental theoretical and software implementation issues which also must be solved if fully autonomous systems are to become a reality [Bekey, 2005]. The current state of some of these issues is as follows:

2.6.1 Software Architecture

Most current systems use the so-called 'three level architecture', illustrated in Fig. 2.6. The lowest level is basically reflexive, and allows the robot to react almost instantly to a particular sensory input.

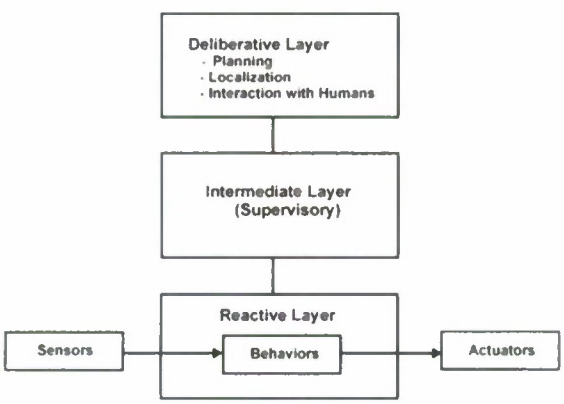


Figure 2.6. Typical three-level architecture for robot control

The highest level, sometimes called the Deliberative layer, includes Artificial Intelligence such as planning and learning, as well as interaction with humans, localization and navigation. The

intermediate or 'supervisory' layer provides oversight of the reactive layer, and translates upper level commands as required for execution. Many recent developments have concentrated on increasing the sophistication of the 'deliberative' layer.

2.6.2 Simultaneous Localization and Mapping (SLAM)

An important problem for autonomous robots is to ascertain their location in the world and then to generate new maps as they move. A number of probabilistic approaches to this problem have been developed recently.

2.6.3 Learning

Particularly in complex situations it has become clear that robots cannot be programmed for all eventualities. This is particularly true in military scenarios. Hence, the robot must learn the proper responses to given stimuli, and its performance should improve with practice.

2.6.4 Multiple Robot System Architectures

Increasingly, it will become necessary to deploy multiple robots to accomplish dangerous and complex tasks. The proper architecture for control of such robot groups is still not known. For example, should they be organized hierarchically, along military lines, or should they operate in semi-autonomous sub-groups, or should the groups be totally decentralized?

2.6.5 Human-Robot Interaction

In the early days of robotics (and even today in certain industrial applications), robots are enclosed or segregated to ensure that they do not harm humans. However, in an increasing number of applications, humans and robots cooperate and perform tasks jointly. This is currently a major focus of research in the community, and there are several international conference devoted to Human-Robot Interaction (HRI).

2.6.6 Reconfigurable Systems

There is increasing interest (both for military and civilian applications) in developing robots capable of some form of 'shape-shifting.' Thus, in certain scenarios, a robot may be required to move like a snake, while in others it may need legs to step over obstacles. Several labs are developing such systems.

2.7 Ethical Implications: A Preview

It is evident from the above survey that the Armed Forces of the United States are implementing the Congressional mandate described in section 1 of this report. However, as of this writing, none of the fielded systems has full autonomy in a wide context. Many are capable of autonomous navigation, localization, station keeping, reconnaissance and other activities, but rely on human supervision to fire weapons, launch missiles, or exert deadly force by other means; and even the Navy's CIWS operates in full-auto mode only as a reactive last line of defense against incoming missiles and does not proactively engage an enemy or target. Clearly, there are fundamental ethical implications in allowing full autonomy for these robots. Among the questions to be asked are:

- Will autonomous robots be able to follow established guidelines of the Laws of War and Rules of Engagement, as specified in the Geneva Conventions?
- Will robots know the difference between military and civilian personnel?
- Will they recognize a wounded soldier and refrain from shooting?

Technical answers to such questions are being addressed in a study for the US Army by professor Ronald Arkin from Georgia Institute of Technology—his preliminary report is entitled *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture* [Arkin 2007]—and other experts [e.g., Sharkey, 2008a]. In the following sections of our report, we seek to complement that work by exploring other (mostly non-technical) dimensions of such questions, specifically as they related to ethics and risk.

2.8 Future Scenarios

From the brief descriptions of the state of the art of robotics above, it is clear that the field is highly dynamic. Robotics is inherently interdisciplinary, drawing from advances in computer science, aerospace, electrical and mechanical engineering, as well as biology (to obtain models of sensing, processing and physical action in the animal kingdom), sociology, ergonomics (to provide a basis for the design and deployment of robot colonies), and psychology (to obtain a basis for human-robot interaction). Hence, discoveries in any of these fields will have an effect on the design of future robots and may raise new questions of risk and ethics. It would be useful, then, to anticipate possible future scenarios involving military robotics in order to more completely consider issues in risk and ethics, as follow:

2.8.1 Sentry/Immobile Robots

A future scenario may include robot sentries that guard not only military installations but also factories, government buildings, and the like. As these guards acquire increasing autonomy, they

may not only challenge visitors ("Who goes there?") and ask them to provide identification but will be equipped with a variety of sensors for this purpose: vision systems, bar code readers, microphones, sound analyzers, and so on. Vision systems (and, if needed, fingerprint readers) along with large graphic memories may be used to perform the identification. More importantly, the guards will be equipped with weapons enabling them to arrest and, if necessary, to disable or kill a potential intruder who refuses to stop and be identified. Under what conditions will such lethal force be authorized? What if the robot confuses the identities of two people? These are only two of the many difficult ethical questions which will arise even in such a basically 'simple' task as guarding a gate and challenging visitors.

2.8.2 Ground Vehicles

We expect that future generations of Army ground vehicles, beyond the existing PackBots or SWORDS discussed in section 2.1 above, will feature significantly more and better sensors, better ordnance, more sophisticated computers, and associated software. Advanced software will be needed to accomplish several tasks, such as:

(a) Sensor fusion: More accurate situational awareness will require the technical ability to assign degrees of credibility to each sensor and then combine information obtained from them. For example, in the vicinity of a 'safe house', the robot will have to combine acoustic data (obtained from a variety of microphones and other sensors) with visual information, sensing of ground movement, temperature measurements to estimate the number of humans within the house, and so on. These estimates will then have to be combined with reconnaissance data (say from autonomous flying vehicles) to obtain a probabilistic estimate of the number of combatants within the house.

(b) Attack decisions: Sensor data will have to be processed by software that considers the applicable Rules of Engagement and Laws of War in order for a robot to make decisions related to lethal force. It is important to note that the decision to use lethal force will be based on probabilistic calculations, and absolute certainty will not be possible. If multiple robot vehicles are involved, the system will also be required to allocate functions to individual members of the group, or they will be required to negotiate with each other to determine their individual functions. Such negotiation is a current topic of much challenging research in robotics.

(c) Human supervision: We anticipate that autonomy will be granted to robot vehicles gradually, as confidence in their ability to perform their assigned tasks grows. Further, we expect to see learning algorithms that enable the robot to improve its performance during training missions. Even so, there will be fundamental ethical issues. For example, will a supervising warfighter be able to override a robot's decision to fire? If so, how much time will have to be allocated to allow such decisions? Will the robot have the ability to disobey a human supervisor's command, say in a situation where the robot makes the decision not to release a missile on the basis that its analysis leads to the conclusion

that the number of civilians (say women and children) greatly exceeds the number of insurgents in the house?

2.8.3 Aerial Vehicles

Clearly, many of the same consideration that apply to ground vehicles will also apply to UFVs, with the additional complexity that arises from moving in three degrees of freedom, rather than two as on the surface of the earth. Hence, the UFV must sense the environment in the x, y, and z directions. The UFV may be required to bomb particular installations, in which case it will be governed by similar considerations to those described above. However, there may be others: for instance, an aircraft is generally a much more expensive system than a small ground vehicle such as the SWORDS. What evasive action should the vehicle undertake to protect itself? It should have the ability to return to base and land autonomously, but what should it do if challenged by friendly aircraft? Are there situations in which it may be justified in destroying friendly aircraft (and possibly killing human pilots) to ensure its own safe return to base? The UFV will be required to communicate with UGVs and to coordinate strategy when necessary. How should decisions be made if there is disagreement between airborne and ground vehicles? If there are hybrid missions that include both piloted and autonomous aircraft, who is in charge?

These are not a trivial question, since contemporary aircraft move at very high speeds, making the length of time required for decisions inadequate for human cognitive processes. In addition, vehicles may be of vastly different size, speed and capability. Further, under what conditions should a UFV be permitted to cross national boundaries in the pursuit of an enemy aircraft? Since national boundaries are not painted on the ground, the robot aircraft will have to rely on stored maps and GPS measurements, which may be faulty.

2.8.4 Marine Vehicles

Many of the same challenges that apply to airborne vehicles also apply to those traveling under water. Again, they must operate in multiple degrees of freedom. In addition, the sensory abilities of robot submarines will be quite different from those of ground or air vehicles, given the properties of water. Thus, sonar echoes can be used to identify the presence of underwater objects, but these signals require interpretation. Assume that the robot submarine detects the presence of a surface vessel, which is presumed to carrying enemy weapons, as well as civilian passengers: under what conditions should the robot submarine launch torpedoes to destroy the surface vessel? It may be much more difficult to estimate the number of civilians aboard an iron ship than those present in a wooden house. How can the robot make intelligent decisions in the absence of critical information?

It is evident that the use of autonomous robots in warfare will pose a large number of ethical challenges. In the next sections, we discuss some programming approaches and their relationship to

ethical theories, issues related to responsibility and law (including LOW/ROE), and expand on the various ethical and risk issues we have raised in the course of this report.

3. Programming Morality

What role might ethical theory play in defining the control architecture for semi-autonomous and autonomous robots used by the military? What moral standards or ethical subroutines should be implemented in a robot? This section explores the ways in which ethical theory may be helpful for implementing moral decision making faculties in robots.[4]

Engineers are very good at building systems to satisfy clear task specifications, but there is no clear task specification for general moral behavior, nor is there a single answer to the question of whose morality or what morality should be implemented in AI. However, military operations are conducted within a legal framework of international treaties as well as the nation's own military code. This suggests that the rules governing acceptable conduct of personnel might perhaps be adapted for robots; one might attempt to design a robot which has an explicit internal representation of the rules and strictly follows them.

A robotic code would, however, probably need to differ in some respects from that for a human soldier. For example, self-preservation may be less of a concern for the robotic system, both in the way it is valued by the military and in its programming. Furthermore, what counts as a strictly correct interpretation of the laws in a specific situation is itself likely to be a matter for dispute, and conflicts among duties or obligations will require assessment in light of more general moral principles. Regardless of what code of ethics, norms, values, laws, or principles are adopted for the design of an artificial moral agent (AMA), whether the system functions successfully will need to be evaluated through externally-determined criteria and testing.

3.1 From Operational to Functional Morality

Safety and reliability have always been a concern for engineers in their design of intelligent systems and for the military in its choice of equipment. Remotely-operated vehicles and semi-autonomous weapons systems used during military operations need to be reliable, and they should be destructive only when directed at designated targets. Not all robots utilized by the military will be deployed in combat situations, however, establishing as a priority that all intelligent systems are safe and do no harm to (friendly) military personnel, civilians, and other agents worthy of moral consideration.

[4] We thank and credit Wendell Wallach and Colin Allen for their contribution to many of the discussions here, drawn from their new book *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2008).

When robots with even limited autonomy must choose from among different courses of action, the concern for safety is transmuted into the need for the systems to have a capacity for making moral judgments. For robots that operate within a very limited context, the designers and engineers who build the systems may well be able to discern all the different options the robot will encounter and program the appropriate responses. The actions of such a robot are completely in the hands of the designers of the systems and those who choose to deploy them; these robots are *operationally moral*. They do not have, and presumably will not need, a capacity to explicitly evaluate the consequences of their actions. They will not need to evaluate which rules apply in a particular situation, nor need to prioritize conflicting rules.

However, three factors suggest that operational morality is not sufficient for many robotic applications: (1) the increasing autonomy of robotic systems; (2) the prospect that systems will encounter influences that their designers could not anticipate because of the complexity of the environments in which they are deployed, or because the systems are used in contexts for which they were not specifically designed; and (3) the complexity of technology and the inability of systems engineers to predict how the robots will behave under a new set of inputs.

The choices available to systems that possess a degree of autonomy in their activity and in the contexts within which they operate, and greater sensitivity to the moral factors impinging upon the course of actions available to them, will eventually outstrip the capacities of any simple control architecture. Sophisticated robots will require a kind of *functional morality*, such that the machines themselves have the capacity for assessing and responding to moral considerations. However, the engineers that design functionally moral robots confront many constraints due to the limits of present-day technology. Furthermore, any approach to building machines capable of making moral decisions will have to be assessed in light of the feasibility of implementing the theory as a computer program.

In the following, we will briefly examine several major theories—deontological (rule-based) ethics, consequentialism, natural law, social contract ethics, and virtue ethics—as possible ethical frameworks in robots. (A complete discussion of these theories and their relative plausibility is beyond the scope of this report and can be readily found in philosophical literature [e.g. University of San Diego, 2008].)

First, let us dismiss one important possibility: ethical relativism, or the position that there is no such thing as objectivity in ethical matters, i.e., what is right or wrong is not a matter of fact but a result of individual or cultural preferences. Even if it were true that ethics is relative to cultural preferences, this would have no bearing on a project to develop autonomous military robots, since the US military and its code of ethics would be the standard for our robots anyway, as opposed to programming some other nation's morality into our machines. Further, we can expect that such robots will be employed only in specific environments, at least for the foreseeable future, which suggests a more

limited, practical programming approach; so a broad or all-encompassing theory of ethics is not immediately urgent, and thus we need not settle the question of whether ethics is objective here.

That is, the idea of an autonomous general- or even multi-purpose robot (which might require a broad framework to govern a full range of possible actions) is much more distant than the possibility of an autonomous robot created for specific military-related tasks, such as patrolling borders or urban areas, or exercising lethal force in a carefully circumscribed battlefield. Given the limited operations of such robots, the initial ethical task will be sufficient to simply program in the suitable basic, relevant rules. In the next section, we will delineate the Laws of War and Rules of Engagement that would govern the robot's behavior; these laws already are established and codified, making programming easier (in theory). We will also offer challenges and further difficulties related to the approach of using the LOW and ROE as an ethical framework, and discuss longer-term issues that may arise as robots have greater autonomy and responsibility.

3.2 Overview: Top-Down and Bottom-Up Approaches

The challenge of building artificial moral agents (AMAs) might be understood as finding ways to implement abstract values within the control architecture of intelligent systems. Philosophers confronted with this problem are likely to suggest a top-down approach of encoding a particular ethical theory in software. This theoretical knowledge could then be used to rank options for moral acceptability. Psychologists confronted with the problem of constraining moral decision-making are likely to focus on the way a sense of morality develops in human children as they mature into adults. Their approach to the development of moral acumen is bottom-up in the sense that it is acquired over time through experience. The challenge for roboticists is to decide whether a top-down ethical theory or a bottom-up process of learning is the more effective approach for building artificial moral agents.

The study of ethics commonly focuses on top-down norms, standards, and theoretical approaches to moral judgment. From Socrates' dismantling of theories of justice to Kant's project of rooting morality within reason alone, ethical discourse has typically looked at the application of broad standards of morality to specific cases. According to these approaches, standards, norms, or principles are the basis for evaluating the morality of an action.

The term 'top-down' is used in a different sense by engineers, who approach challenges with a top-down analysis through which they decompose a task into simpler subtasks. Components are assembled into modules that individually implement these simpler subtasks, and then the modules are hierarchically arranged to fulfill the goals specified by the original project.

In our discussion of machine morality, we use 'top-down' in a way that combines these two

somewhat different senses from engineering and ethics. In our broader sense, a top-down approach to the design of AMAs is any approach that takes a specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory.

In the bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and is rewarded for behavior that is morally praiseworthy. In this manner, the artificial agent develops or learns through its experience. Unlike top-down ethical theories, which define what is and is not moral, ethical principles must be discovered or constructed in bottom-up approaches. Bottom-up approaches, if they use a prior theory at all, do so only as a way of specifying the task for the system, and not as a way of specifying an implementation method or control structure.

Engineers would find this top-down/bottom-up dichotomy to be rather simplistic given the complexity of many engineering tasks. However, the concepts of top-down and bottom-up task analysis are helpful in that they highlight two different roles for ethical theory in facilitating the design of AMAs.

3.3 Top-Down Approaches

Are ethical principles, theories, and frameworks useful in guiding the design of computational systems capable of acting with some degree of autonomy? Can top-down theories—such as utilitarianism, or Kant's categorical imperative, or even Asimov's laws for robots—be adapted practically by roboticists for building AMAs?

Top-down approaches to artificial morality are generally understood as having a set of rules that can be turned into an algorithm. These rules specify the duties of a moral agent or the need for the agent to calculate the consequences of the various courses of action it might select. The history of moral philosophy can be viewed as a long inquiry into the adequacy of any one ethical theory; thus, selecting any particular theoretical framework may not be adequate for ensuring an artificial agent will behave acceptably in all situations. However, one theory or another is often prominent in a particular domain, and for the foreseeable future most robots will function within limited domains of activity.

3.3.1 Top-Down Rules: Deontology

A basic grounding in ethical theory naturally begins with the idea that morality simply consists in following some finite set of rules: deontological ethics, or that morality is about simply doing one's

duty. Deontological (duty-based) ethics presents ethics as a system of inflexible rules; obeying them makes one moral, breaking them makes one immoral. Ethical constraints are seen as a list of either forbidden or permissible forms of behavior. Kant's *Categorical Imperative (CI)* is typical of a deontological approach, as follows in its two main components:

CI(1) – This is often called the formula of universal law (FUL), which commands: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" [Kant, 1785, 4:421]. Alternatively, the CI also has been understood as that the relevant legislature should pass such a law mandating my action, i.e., a 'Universal Law of Nature.'

A maxim is a statement of one's intent or rationale: it is the answer to the query about why one did what was done. So Kant asserts that the only intentions that are moral are those that could be universally held; partiality has no place in moral thought. Kant also asserts that when we treat other people as a mere means to our ends, such action must be immoral; after all, we ourselves don't wish to be treated that way. Hence, when applying the CI in any social interaction, Kant provides a second formulation as a purported corollary:

CI(2) – Variously called the Humanity formulation of the CI, or the Means-Ends Principle, or the formula of the end in itself (FEI), it commands: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" [Kant, 1785, 4:429]. One could never universalize the treatment of another as a mere means to some other ends, claims Kant, in his explanation that CI(2) directly follows from CI(1). This formulation is credited with introducing the idea of 'respect' for persons; that is, respect for whatever it is that is essential to our Humanity, for whatever collective attributes are required for human dignity [Johnson, 2008].

A Kantian deontologist thus believes that acts such as stealing and lying are always immoral, because universalizing them creates a paradox. For instance, one cannot universalize lying without running into the 'Liar's paradox' (that it cannot be true that all statements are a lie); similarly, one cannot universalize stealing property without undermining the very concept of property. Kant's approach is widely influential but has problems of applicability and disregard for consequences.

3.3.2 Asimov's Laws of Robotics

Another deontological approach often comes to mind in investigating robot ethics: Asimov's Three Laws of Robotics (he later added a fourth or 'Zeroth Law') are intuitively appealing in their simple demand to not harm or allow humans to be harmed, to obey humans, and to engage in self-preservation. Furthermore, the laws are prioritized to minimize conflicts. Thus, doing no harm to humans takes precedence over obeying a human, and obeying trumps self-preservation. However, in story after story, Asimov demonstrated that three simple hierarchically-arranged rules could lead

to deadlocks when, for example, the robot received conflicting instructions from two people or when protecting one person might cause harm to others.

The original version of Asimov's Three Laws of Robotics are as follows: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey orders given to it by human beings, except where such orders would conflict with the First Law; (3); a robot must protect its own existence as long as such protection does not conflict with the First or Second Law [Asimov, 1950].

Asimov's fiction explored the implications and difficulties of the Three Laws of Robotics. It established that the first law was incomplete as stated, due to the problem of ignorance: a robot was fully capable of harming a human being as long as it did not know that its actions would result in (a risk of) harm, i.e., the harm was unintended. For example, a robot, in response to a request for water, could serve a human a glass of water teeming with bacterial contagion, or throw a human down a well, or drown a human in a lake, *ad infinitum*, as long as the robot was unaware of the risk of harm. One solution is to rewrite the first and subsequent laws with an explicit knowledge-qualifier: "A robot may do nothing that, to its knowledge, will harm a human being; nor, through inaction, knowingly allow a human being to come to harm" [Asimov, 1957]. But a clever criminal could divide a task among multiple robots, so that no one robot could even recognize that its actions would lead to harming a human, e.g., one robot places the dynamite, another attaches a length of cord to the dynamite, a third lights the cord, and so on. Of course, this simply illustrates the problem with deontological, top-down approaches, that one may follow the rules perfectly but still produce terrible consequences.

An additional difficulty is that the degree of risk makes a difference too, e.g., should robots keep humans from working near X-ray machines because of a small risk of cancer, and how would a robot decide? (Section 6 on risk assessment will explore this topic further). The 'through inaction' clause of Asimov's first law raises another issue: Wouldn't a robot have to constantly intervene to minimize all sorts of risks to humans, and never be able to perform its primary tasks? Asimov considers a modified First Law to solve this issue: (1′) A robot may not harm a human being. Removing the First Law's 'inaction' clause solves this problem, but t does so at the expense of creating an even greater one: a robot could initiate an action which would harm a human (for example, initiating an automatic firing sequence, then watching a noncombatant wander into the firing line) knowing that it was capable of preventing the harm (by ceasing the automatic firing), but it may nevertheless fail to do so since it is now not strictly required to act.

Asimov later added a Zeroth Law [Asimov, 1985]—so named to continue the pattern of lower-numbered laws superseding in importance the higher-numbered laws—so that the Zeroth Law had highest priority and must not be broken: (0) a robot may not harm all humanity or, through inaction, allow humanity to come to harm. This would allow a robot to harm individual humans, if so doing

prevented an 'existential threat' to all humanity. But how could a robot determine when such a threat exists, and hence killing individual humans to prevent it is permitted?

3.3.3 Fixing Asimov's laws

Other authors have attempted to fix other ambiguities and loopholes in the rules Asimov devised, in order to prevent disastrous scenarios that nonetheless satisfied laws numbered 0-3. For example, Lyuben Dilov [1974] introduced a Fourth Law of Robotics to avoid misunderstandings about what counts as a human and as a robot: (4) a robot must establish its identity as a robot in all cases. This law is sometimes stated as the slightly different: (4′). A robot must know it is a robot. Others [e.g. Harrison, 1989] have also argued for a Fourth Law that requires robots to reproduce, as long as such reproduction does not interfere with laws 1-3.

Asimov's literary exercise was illustrative of a limitation inherent in any rule-based morality: What does the robot do when there are conflicts between the rules? Should rules function as hard restraints? Or can the rules function as guidelines where the system is designed to factor in an array of *prima facie* duties in the actions it considers? Will this open the door to robotic behavior that should be prohibited? Perhaps the biggest challenges confronting designers of rule-based robots or AMAs is how the system will recognize those situations that require application of the rules, and how to ensure that the robot has access to all the information it needs in order to apply rules appropriately. How would a robot programmed with the First Law *know*, for example, that a medic or surgeon welding a knife over a fallen fighter on the battlefield is not about to harm the soldier? The robot would need to understand a great deal about context, exceptions to rules, and human psychology; and its knowledge base would need to be updated regularly.

Roger Clarke [1994] attempted to update and fix Asimov's laws, in what he called "An Extended Set of the Laws of Robotics":

> "The Meta-Law: A robot may not act unless its actions are subject to the Laws of Robotics.
> Law Zero: A robot may not injure humanity, or, through inaction, allow humanity to come to harm.
> Law One: A robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher-order Law.
> Law Two: A robot must obey orders given it by human beings, except where such orders would conflict with a higher-order Law; a robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher-order Law.
> Law Three: A robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher-order Law; a robot must protect its own existence as long as such protection does not conflict with a higher-order Law.

Law Four: A robot must perform the duties for which it has been programmed, except where that would conflict with a higher-order law.

The Procreation Law: A robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics."

Clarke admits that his revised laws still face serious problems, including the identification of and consultation with stakeholders and how they are affected, as well as issues of quality assurance, liability for harm resulting from either malfunction or proper use, and complaint-handling, dispute-resolution, and enforcement procedures. Our discussion of product liability in section 5 will address many of these concerns.

There are additional problems that occur when moral laws for robots are given in the military context. To begin with, military officers are aware that if codes of conduct or Rules of Engagement are not comprehensive, then proper behavior cannot be assured. One difficulty lies in the fact that as the context gets more complex, it becomes impossible to anticipate all the situations that soldiers will encounter, thus leaving the choice of behavior in many situations up to the best judgment of the soldier. The desirability of placing machines in this situation is a policy decision that is likely to evolve as the technological sophistication of AMAs improves.

Unfortunately, there are yet further problems: most pertinently, even if their glitches could be ironed out, Asimov's laws will remain simply inapplicable to the military context, as it is likely that autonomous military robots will be asked to exercise lethal force upon humans in order to attain mission objectives, thereby violating Asimov's First Law. A further problem, called *'rampancy'*, involves the possibility that an autonomous robot could overwrite its own basic programming and substitute its own new goals for the original mission objectives (e.g., the movie *Stealth*). That leads us to a final and apparently conclusive reason why deontological ethics cannot be used for autonomous military robots: it is incompatible with a 'slave morality', as addressed in the following discussion (and further in section 6).

3.3.4 Slavery: A Crucial Problem for Deontology and Robotic ethics

One further problem, specific to robotics, with deontological ethics is the problem of 'slave morality.' Robots in the military would be presumably programmed to follow commands slavishly, and not exhibit anything like true Kantian autonomy. Indeed, the term 'robot' is derived from the Czech word 'robota' that means 'servitude' or 'drudgery' or 'labor' (see 'Appendix A: Definitions'). Such robots could make autonomous choices about the means to carrying out their pre-programmed goals, but not about the goals themselves; they could not choose their own goals for themselves, but

they would always be expected to have the goal of obeying orders given by their military commander.

That would collapse (from a deontological perspective) all questions about their ethics into simply questions about the ethics of the military commander, and *mutatis mutandis* for any other use of autonomous robots as slaves. Such an approach would then claim that there is actually no such thing as robot ethics; there are only the ethics of those who command robots. But the concerns with robot ethics crucially concern the consequences of using them—a concern a strict deontological ethics cannot countenance as it insists that one must obey the rules, no matter the consequences. And of course, a key objection (that will affect both deontological and utilitarian ethics) is the plausible skeptical claim that no finite set of rules can ever guarantee ethical behavior in all cases, or at least where the set of possible behaviors is large or practically unlimited. Before addressing that critique, let us examine perhaps the most important objection to deontology (and the resulting alternative approach)—that consequences matter morally, and simply following the rules is morally wrong if it leads to bad outcomes.

3.3.5 Top-Down Approaches: Utilitarian Consequentialism

Utilitarianism represents another attempt to bypass conflicts between rules through an overriding top-down principle that can be applied to all situations. However, with respect to computability, this approach stresses the importance of the outcomes (consequentialism) arising from an action. *Consequentialist* approaches to ethics focus on achieving the best possible outcomes in various situations, and hence typically disdain rigid rules that specify unchanging duties. For example, utilitarianism—the primary consequentialist theory— proposes that an agent should calculate the net consequences arising from the various available courses of action, and then select the action that offers 'the greatest good for the greatest number.' This is a familiar, pragmatic theory in that many policy and business decisions seem to be determined by a weighing of reasons for and against a particular action, and it suggests a simple algorithm for calculating what action one ought to take in a given situation.

However, this approach is not as computationally tractable as it might appear. Practically, there is the calculational objection: it is an impossible demand to calculate the utility of every action; thus, utilitarianism *makes moral evaluation impossible,* as even the short-term consequences of most actions are impossible to accurately forecast, much less the long-term consequences. Problems of how utility might be represented within a computational system, how broadly the consequences of actions should be analyzed, and which agent's welfare should be included in the calculation need to be resolved in order to bring a utilitarian analysis to a successful conclusion. Given limitations of available information, the breadth of variables impinging upon a complex set of interrelated agents, and therefore an inability to accurately predict the consequences of an action, such a calculation

poses a tremendous computation load on even the fastest systems. A utilitarian robot may fail to determine which course of action is most acceptable within the time allotted.

But if utility is incalculable, and one's obligation is to maximize utility, much of the theory's value seems to disappear. Worse, there are further objections to utilitarianism: the *absurd implications objection* would, for example, point to some scenario in which a lie is just as moral as truth, if the consequences are the same. Even more fundamental are objections based on (in)justice. For example, the *scapegoating* objection would point out that maximizing utility may demand injustice, such as executing an innocent person to prevent a riot that would have resulted in deaths and economic damage. This is to say that utilitarianism, at least in its basic form, cannot readily account for the notion of rights and duties nor moral distinctions between, e.g., killing versus letting die or intended versus merely foreseen deaths (assuming we think such notions and distinctions exist).

Whether deontological or consequentialist/utilitarian, each of the single-principle top-down theories suffers from a version of the frame problem—that is, it requires an impossible computational load due to the requirements for knowledge of the relevant effects of action in the world, the difficulty of estimating the sufficiency of the initial information, and knowledge about the psychology of agents. Nevertheless, humans appear to apply rough and ready top-down evaluations in their selection of courses of action, and so might a robotic system, particularly if the goal is not to create a perfect system but only one that makes better (or just as good) decisions than humans do.

Top-down theories combine strength in defining ethical criteria with a breadth that can be applied to countless challenges. The price of this strength lies in the goals either being defined so vaguely and abstractly that their meaning and their application to specific situations is debatable, or they are defined so rigidly that they fail to produce decisions that are appropriately sensitive to new context.

3.4 Bottom-Up Approaches

The bottom-up approaches to building AMAS are inspired by three sources: (1) the tinkering by engineers as they optimize system performance, (2) evolution, and (3) learning and development in humans. Bottom-up approaches fall within two broad categories: the assembling of systems with complex faculties out of discrete subsystems, and the emergence of values and patterns of behavior in a more holistic fashion, as in artificial life experiments and connectionist networks.

A variety of discrete subsystems are being developed by computer scientists that could potentially contribute to the building of artificial moral agents. Not all of these subsystems are explicitly designed for moral reasoning. For example, learning algorithms, affective sensors, and social mechanisms might all contribute to the moral acumen of a robot. But computer scientists who wish to build robots with higher-order faculties out of discrete subsystems are confronted with a difficult,

and perhaps insurmountable, challenge of assembling components into a functional whole. Whether the aggregation of discrete skill sets will lead to the emergence of higher-order cognitive faculties—including emotional intelligence, moral judgment, and consciousness—can only be known once roboticists go through the exercise of building the systems.

3.4.1 Optimizing Performance

Various trial-and-error techniques are available to engineers for progressively tuning components so that the system approaches or surpasses the performance criteria. Bottom-up approaches to ethics treat normative values as being implicit in the activity of agents rather than explicitly articulated (or even articulatable) in terms of a general theory. Engineers commonly define tasks atheoretically using a performance measure, such as winning chess games, passing the Turing test, walking across a room without stumbling, and so on. Even without a theory of the best way to decompose the task into subtasks, engineers can achieve a high level of performance on many tasks. Sometimes a *post hoc* analysis of the system can produce a theory or specification of how the subtasks yield results. But often the results of such an analysis do not correspond to the kind of decomposition suggested by *a priori* theorizing.

3.4.2 Evolution

Evolution has inspired an array of approaches for developing artificial intelligence from artificial life experiments (Alife) to genetic algorithms and to evolutionary robotics. The theory of evolution has suggested to engineers a model for self-selecting and self-organizing systems that strive toward the optimization of some performance criteria, such as the maximization of profits. The power of evolution is tapped into by selecting those agents, from a collection of similar agents, that are most successful at optimizing a specified fitness (performance) criterion. The selected agents serve as *parents* that are modified and recombined (using a process that is analogous to sexual reproduction) to produce a new generation of agents. This new generation is tested, the best performers selected, and they in turn breed, and so forth. This basic strategy has been successful for producing agents suited to a wide variety of tasks.

Two ideas contribute to the belief that evolutionary strategies would be helpful for eliciting moral behavior in agents. The first is the contention by game theorists and evolutionary psychologists that moral propensities, such as cooperation and care of the young, may have emerged during the course of evolution are partially encoded in genes and potentially reproducible in simulations of evolution within computer environments. However, as Rodney Brooks has noted, experiments in Alife "have not taken off by themselves in the ways we have come to expect of biological systems" [Brooks, 2002]. The second influencing idea is that optimizing moral performance might be used as the fitness criteria for selecting the best agents. The difficulty with this strategy lies in how the fitness criteria would be represented in a computational system. The slogan 'survival of the most moral'

highlights the problem of saying what 'most moral' amounts to in a non-circular (and computationally tractable) fashion.

3.4.3 Learning and Development

Alan Turing was the first to broach the idea that artificial intelligence (AI) should try to mimic child development. In 1950 he wrote: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain" [Turing, 1950].

Jean Piaget, Lawrence Kohlberg, Carol Gilligan and others have proposed developmental theories regarding the way in which children learn about morality [Murray, 2008]. These theories have been adapted into curricula that facilitate the moral development of children. This suggests the possibility that a learning robot might be taken through a similar educational program. However, while machine learning is an important area of research, the algorithms and techniques presently available are not robust enough for such a sophisticated educational project. For the immediate future, machine-learning techniques are likely to be quite rudimentary.

3.4.4 The Value and Limits of Bottom-up Approaches

Individual subsystems can be quite brittle in their performance. However, when integrated successfully, these components can give rise to complex dynamic systems with a range of choices or optional responses to external conditions and pressures. Bottom-up engineering thus holds the promise of a kind of dynamic morality where, as conditions change, the ongoing feedback from different mechanisms facilitates varied responses. It ties into a movement within ethics termed 'particularism', which asserts that no general laws or rules are possible, and each ethical situation is unique.

But the weakness of bottom-up approaches for developing AMAs lies in not knowing which goals to use for evaluating choices and actions as contexts and circumstances change. Bottom-up systems work best when they are directed at achieving one clear goal. When the system has more than one goal, or when the available information is confusing or incomplete, bottom-up engineering is less likely to provide a clear choice or course of action.

3.5 Supra-Rational Faculties

In order to function as moral agents, robots that interact with humans within dynamic multi-agent contexts may require other skill sets in addition to being rational. Which skills the agent will need will vary depending upon the robot's tasks and goals. For example, tasks that socially-viable robots

will perform can require emotional intelligence, knowledge about social customs, and the non-verbal cues used to communicate essential information. In addition, the capacity to appreciate and respond to moral challenges may also depend upon the robot having semantic understanding, consciousness, a theory of mind (the ability to deduce the intentions, desires, and goals of others), and perhaps even the capacity to feel pain or empathy. These additional faculties are a tall order for roboticists, although each has already stimulated interesting lines of research that are under way.

Algorithms for reasoning about moral challenges will not lead to appropriate behavioral responses unless the robot has access to the background information describing the situation, the ability to discern which information is essential and which inputs are of ethical concern, and the capacity to recognize inherent and potential conflicts arising from the competing interests of the various agents. In other words, the engineer must determine what the information requirements are for a system making moral decisions. What will the system need to know in order to make an informed decision? What input devices and sensors will it need to get access to this information?

Supra-rational faculties—such as emotions, being embodied in the world, social skills, and consciousness— represent ways that humans get access to essential information that must be factored into ethically-significant choices. Sensory experience and social mechanisms contribute to various refinements of behavior that people expect from each other; they also will be necessary for robots that function to a high degree of competence in social contexts. While robots will not necessarily need to emulate the full array of human faculties, the more sophisticated systems will need mechanisms that provide a similar appreciation of complex social contexts. Furthermore, communicating through facial expressions, gestures, vocal intonation and prosody, and other verbal and non-verbal cues will be helpful in conveying the robot's intentions and facilitating cooperation with humans. This will, in turn, help humans to perceive the robot as being trustworthy.

Given that morally intelligent behavior may require much more than being rational, the challenge of building AMAs is, from this perspective, a problem of moral psychology, not moral calculation. For designers of morally intelligent systems, the challenge is not how to give them abstract theoretical knowledge but how to develop robots that embody the right tendencies in their reactions to the world and other agents in that world.

3.6 Hybrid Systems

Moral judgment in humans is a hybrid of both (1) bottom-up mechanisms shaped by evolution and learning and (2) top-down criteria for reasoning about ethical challenges. Eventually, we may be able to build morally intelligent robots that maintain the dynamic and flexible morality of bottom-up systems capable of accommodating diverse inputs, while subjecting the evaluation of choices and actions to top-down principles. The prospect of developing virtuous robots offers one venue for

considering the integration of top-down, bottom-up, and supra-rational mechanisms.

3.6.1 Virtue Ethics: The Virtuous Robot

There is a foundational critique of all procedural ethics, i.e., any approach that claims morality is determined by following the proper rules. Many contemporary ethicists claim that all procedural ethics fail, because so many theorists have explicitly abandoned the ideas that in ethics: (a) the rules would amount to a decision procedure for determining what the right action would be in any particular case; and (b) the rules would be stated in such terms that any non-virtuous person could understand and apply them correctly.

In robotics, so-called *'friendliness theory'* attempts to deal with this conundrum: rather than using any finite set of top-down rules or laws, intelligent machines should be programmed to be basically altruistic, and then use machine learning in various settings to create a kind of 'best judgment' in how to carry out properly altruistic actions. This approach sidesteps the fundamental calculational and programming problem of how to account for a vast number of unforeseeable eventualities. However, this theory has a problem with military robots: we would not want them to always act altruistically towards some humans. In fact, we would want them to be able to kill the right humans and not the wrong (friendly) ones. Therefore, we need an approach that enables machines to use a basic top-down program plus bottom-up machine learning to be able to function excellently in its military roles and without malfunctioning; it should be fierce towards its enemies, helpful to its allies, and reliable in discerning the difference, including in situations unforeseen by its programmers. What ethical approach can accomplish all this?

Perhaps the most viable hybrid approach that avoids the conflict between duties and consequences, and incorporates both warrior fierceness towards enemies and a gentle kindness towards comrades, is *virtue ethics*: an approach that sees ethics, not in terms of what rules should be followed, but in terms of what kind of character an agent has—does one have a virtuous character, or is one full of vice? For virtue ethics, morality is not a function of actions but of character. That is, one's actions do not constitute one's morality but rather *reveal* it: ethics is agent-centered, not act-centered. The proper moral question is not "what rule should I follow?", or "what rules apply to this act", but instead "what sort of person will I reveal myself to be (or become) if this is the sort of thing I do?" What would doing this act say about my character?

While virtue theorists differ in their list of virtues, they take their lead from Aristotle in recognizing that the virtues are acquired developmentally through experience and the cultivation of good habits. This emphasis on developing the virtues can be understood as bottom-up learning, while, at least in theory, it is possible to consider the virtues as top-down patterns for evaluating actions programmed into a robot. Because virtue is an 'excellence' and defined in terms of the roles one plays, it is inescapably context-dependent; no single rule or set of rules will be able to dictate what different

persons (or robots) in different roles will need to do in different situations. Moral criteria are thereby objective but not categorical rules; instead, they are what Kant called 'hypothetical imperatives' linking good means to good ends [Foot, 1972].

Because the virtues are objectively-beneficial habits for proper functioning, but are role dependent, the objective list of virtues for firemen will be different than the list for auditors or salesmen or soldiers, and so on. Only the most generic virtues—e.g., wisdom, honesty, empathy, justice, etc.—will apply to all social roles. Professional codes are then best understood as list of virtues for a particular social role, rather than a list of rules to follow. Thus, following Solomon [1988], we can say that good eyesight is a virtue in a rifleman; it is a virtue because it helps in achieving the purposes or goals of a rifleman. But while the lack of a properly developed conscience might be thought of as a virtue in a hitman understood narrowly within that particular social role, it is not a virtue in a person simply described as a person, within the all-encompassing set of social roles we call human life. A true virtue is thus an excellence in a role that aids overall human flourishing.

3.6.2 The Top–Down Challenge and the Bottom-Up Approach

The top–down challenge for an engineer designing a robot would be to determine how to represent virtuous patterns and motivations, and how the system would determine which virtue, or which action representing the virtue should be called upon in a particular situation. Given the emotional grounding of virtuous motivations in human beings, a designer of a virtuous AMA might need to decide whether a virtuous machine would also need emotions of its own or some mechanisms functionally similar to human emotions.

The bottom-up approach to implementing virtues in computational systems arises from the recognition by several theorists [e.g., DeMoss 1998; Churchland, 1995] of the similarity between learning in connectionist networks and Aristotle's discussion, in the *Nicomachean Ethics,* regarding the way in which the virtues are acquired. Connectionist networks provide a bottom-up strategy for building capacities through the recognition of patterns and the building of categories out of complex inputs. Through the gradual accumulation of data, the network develops generalized responses that go beyond the particulars on which it is trained. One difficulty with learned patterns that emerge from connectionist systems is that they are not accompanied by explanation for *why* the action was chosen.

The problems tackled by existing connectionist networks are far from the complex learning tasks associated with moral development. However, the prospect that neural networks might be adapted for some aspects of moral reasoning is an intriguing possibility. Neural networks offer an approach, deserving of attention, for developing robots that embody the right tendencies in their reactions to the world. The bottom-up development of virtuous patterns of behavior might be combined together with a top-down implementation of the virtues as a way of both evaluating the actions and

as a vehicle for providing rational explanations of the behavior.

While many technological thresholds must be crossed before the development of a virtuous robot becomes a serious possibility, we believe that this approach to building AMAs should be of particular interest to the military in its long-term planning. A virtuous robot might emulate the kind of character that the armed forces value in their personnel. Furthermore, virtues—deeply rooted in the foundational attitudes and structures of an agent—provide a certain degree of stability, and the prospect that officers can rely upon the performance of the artificial agents they deploy.

3.7 First Conclusions: How Best to Program Ethical Robots

A top-down approach would program rules into the robot and expect the robot to simply obey those rules without change or flexibility. The downside, as we saw with the analogous deontological ethics, is that such rigidity can easily lead to bad consequences when events and situations unforeseen or insufficiently imagined by the programmers occur, causing the robot to perform badly or simply do horrible things, precisely because it is rule-bound.

A bottom-up approach, on the other hand, depends on robust machine learning: like a child, a robot is placed into variegated situations and is expected to learn through trial and error (and feedback) what is and is not appropriate to do. General, universal rules are eschewed. But this too becomes problematic, especially as the robot is introduced to novel situations: it cannot fall back on any rules to guide it beyond the ones it has amassed from its own experience, and if those are insufficient, then it will likely perform poorly as well.

As a result, we defend a hybrid architecture as the preferred model for constructing ethical autonomous robots. Some top-down rules are combined with machine learning to best approximate the ways in which humans actually gain ethical expertise. We humans are hard-wired with various rules through our evolutionary heritage, but these vastly underdetermine actual behavior; hence responsible behavior builds on these underlying (largely unconscious) rules with a healthy dose of indoctrination and peer interaction and all the other types of learning that children do. As a result, a character evolves: a tendency to perform certain roles in the evolving ecology of social life, and either to fail or to perform excellently in those roles.

3.7.1 Further Conclusions on Ethical Theory and Robot: Military Implications

Autonomous robots both on and off the battlefield will need to make choices in the course of fulfilling their missions. Some of those choices will have potentially harmful consequences for humans and other agents worthy of moral consideration. Even though the capacity to make moral judgments can be quite complex, and even though roboticists are far from substantiating the

collection of affective and cognitive skills necessary to build AMAs, systems with limited moral decision-making abilities are more desirable than 'ethically blind' systems. The military's comfort with the robots it deploys, and ultimately the comfort of the public, will depend upon a belief that these systems will honor basic human values and norms in their choice of actions. Given the prospect that robotic systems can reduce the loss of personnel during combat, one can presume that the development of autonomous robotic fighting machines will proceed. However, if semi-autonomous and autonomous robotic systems are deployed as lethal weapons, it goes without saying that commanders will need to be confident that the systems will only wield their destructive might on designated targets.

The challenge for the military will reside in preventing the development of lethal robotic systems from outstripping the ability of engineers to assure the safety of these systems. Implementing moral decision-making faculties within robots will proceed slowly. While there are aspects of moral judgment that can be isolated and codified for tightly defined contexts, moral intelligence for autonomous entities is a complex activity dependent on the integration of a broad array of discrete skills. Robots initially will be built to perform specified tasks. However, as computer scientists learn to build more sophisticated systems that can analyze and accommodate the moral challenges posed by new contexts, autonomous robots can and will be deployed for a broad array of military applications. So for the foreseeable future and as a more reasonable goal, it seems best to attempt to program a virtuous *partial* character into a robot and ensure it *only enters situations in which its character can function appropriately*.

Theorists continue to debate whether strong artificial intelligence is possible [e.g., Searle, 1980; Russell and Norvig, 2003]. However, even AI systems with more limited intelligence will require some degree of moral sensitivity in the choices and actions they take: "If there are limitations in the extent to which scientists can implement moral decision making capabilities in AI, it is incumbent to recognize those limitations, so that military planners do not rely inappropriately on artificial decision makers" [Wallach et al., 2008].

For military robots, that virtuous character will likely involve ensuring that the LOW and ROE are programmed in (which may differ from mission to mission) and steadfastly obeyed, as a proxy for a full-fledged morality. Such an approach has several advantages. First, any problems from moral particularism or other problems with general ethical principles (including misguided moral relativism) are skirted. Second, the relationship of morality to legality—a minefield for ethics—is likewise largely avoided; the LOW and ROE make clear what actions are legal and illegal for robots; and for military situations, that can serve as a reasonable approximation to the moral-immoral distinction. So the background of ethics for autonomous robots in the military can, at least for now, become part of our discussion about the programmability of the LOW and ROE. What this means for how we should program ethical robots, and the implications of this approach for the Laws of War and the ethics of risk, is examined in the next section of this report.

4. The Laws of War and Rules of Engagement

For any of the ethical frameworks we have identified in the previous section—deontological ethics, consequentialism/utilitarianism, virtue ethics—the current state of robotics programming (the AI or control software in robots) is not yet robust enough to fully accommodate them. Nevertheless, understanding those ethical theories now is essential for illuminating a thoughtfully-planned path with respect to developing ethical behavior in robots. In the meantime, for military robots, a reasonable proxy for any such ethical theory seems to be found in the Laws of War (LOW) and Rules of Engagement (ROE)—an alternative programming approach with several advantages, as explained at the end of the last section; but, as we explain here, it also has its shortcomings. In this section, we turn to the LOW and ROE, including their relation to just-war theory, and their suitability as an interim programming solution.

4.1 Coercion and the LOW

To understand the LOW and ROE, and to evaluate their viability as a programming approach, we must first understand their origins in just-war theory and, even more basic, the nature of warfare, i.e., what do we need LOW and ROE in first place, and why can't we say 'anything goes' in war?

War, however regrettable, has been an inescapable aspect of human life to date. Autonomous robots have the capacity to radically change the nature of war, and perhaps even eventually lead to its cessation. But during the 'growing pains' of robotics development, our autonomous machines and systems could make the horrors of war either much better or much worse; hence, the ethics of robotic war will be one of the most important subjects of the next decades of the future. To understand the nature of war, we can see it as a type of forcible coercion that nations engage in as a means of attaining their political goals. Let us define the term as follows:

Coercion. *The use of force and or violence, or the threat thereof (i.e., intimidation), in order to persuade*. Coercion is a sad reality both within societies and in international affairs. Within recognized societies, legitimate coercion is exercised by the state, through its police and judiciary, to help restrain and deter illegitimate coercion by individuals. But in the international arena, no supra-national institution has clear and effective coercive power over nation-states who perform illegitimately coercive acts. Hence, states must resolve issues of illegitimate coercion amongst themselves, often through coercion of their own. Hence, we have the phenomenon of war: armed

conflict between states, which attempts to coerce some desired outcome in lieu of other means (e.g., negotiations) of attaining international agreement.

Naturally, if states are engaged in legitimate forcible coercion in order to deter or punish illegitimate coercion, then we must have some agreed upon means of distinguishing legitimate from illegitimate coercion amongst states in war. This is called 'just-war theory', and it attempts to spell out when beginning the coercion of war is morally legitimate and when it is not (termed *jus ad bellum*); and further, what means of wartime coercion are morally permitted (*jus in bello*) and what one should do in the aftermath of officially ending such coercion (*jus post bellum*).

4.2 Just-War Theory and the LOW

So, the Laws of War, also known as the Laws of Armed Conflict (LOAC), concern the legal and moral legitimacy of practices that nations engage in during the interstate forcible coercion that we call 'war.' As mentioned, the Laws of War are divided into three basic categories, with the first two being of general and long-standing acceptance, but the third forming a relatively new emphasis, albeit of increasing import in contemporary asymmetric and non-state warfare:

1. *Jus ad bellum*: Law concerning acceptable justifications to use armed force and declare war.
2. *Jus in bello*: Law concerning acceptable conduct in war, once it has begun.
3. *Jus post bellum*: Law concerning acceptable conduct following the official or declared end of a war (including occupations and indefinite ceasefires, the acceptance of surrender, and the treatment and release of prisoners of war (POWs) and enemy (non-)combatants after conflict has officially ceased).

These three categories are typically (but not always) asserted to be independent; so, the morality and legality of a state deciding to go to war (*jus ad bellum*)—typically a political decision made by a state's political leadership—have long been considered independent of the morality and legality of one's actions in waging war (*jus in bello*); the latter is typically the province of a state's professional military, not its political leadership. For instance, we might hold that an American soldier who participated in the My Lai massacre was guilty of war crimes (a violation of *jus in bello*), but not because the Vietnam War was itself unjust (even assuming that armed conflict was a violation of *jus ad bellum*). Likewise, one might have fought a just war and done so in a just fashion, but it may still impose unjust conditions on the vanquished, thus violating *jus post bellum*.

4.3 Just-War Theory: *Jus ad Bellum*

A traditional emphasis of just-war theory concerns when it is morally acceptable for a state to begin or participate in the extreme forcible coercion that is war, that is, *jus ad bellum.* The ancient Greeks (Aristotle) and Romans (such as Augustine) considered these issues, but the natural law tradition associated with Aquinas began the systematic consideration of the issue. Natural law and social contract theorists have continued it in the work of such luminaries as Vitoria, Grotius, Locke, and Kant; the 20th century adapted this just-war tradition in gradually creating the internationally accepted LOW. Hence, a brief explanation of just-war theory is needed; current influential theorists include Walzer, Orend, and O'Brien, with a rough consensus on the following as necessary conditions for moral *jus ad bellum*:

a. *Proper authority*. War must be waged by a competent authority (normally an internationally recognized state) for a publicly stated purpose, i.e., 'secret wars' are immoral. But this poses a possible dilemma: Would then all revolutions be immoral? This requirement of *jus ad bellum* has considerable difficulty in defining any non-state rebellions (including popular revolutionary movements) as moral. Compare this to the problem of distinguishing between 'freedom fighters', terrorists, and mere criminal behavior.

b. *Just cause*. There must be sufficient and acceptable justifications for entering war. Specifically, the war must be in self-defense, against unjust attacks on innocent citizens or states, to restore rights wrongfully denied and to re-establish a just order (against unjust rebellion). When a state has forfeited its moral right to govern its people—so it is no longer a 'minimally-just state'—other nations may invade it in order to carry out humanitarian interventions in the self-defense of its people, e.g., in Kosovo or Darfur. The state, no longer being minimally just, has forfeited its own right to self-defense. Problem: In addition to the obvious challenge of determining when a state is no longer 'minimally just', developments in non-state warfare, especially terrorism, complicate this requirement.

   However, offensive wars may be justified if to enforce justice for oneself. Problem: The so-called 'Bush Doctrine' and other policies in modern warfare that justify a preemptive war against looming threats would fail the usual interpretation of having a just cause for war. Current scholarship thus focuses on the proper interpretation of self-defense against a merely potential, but not actual, threat; common criteria include the imminence and seriousness of the threat. (See section 6 on risk assessment for more.)

3. *Proportionality*. The good achieved by war must be proportional to the evil of waging it. Therefore, it is immoral to wage a massive war to remedy a small wrong (e.g., the 'Soccer War' of 1969 between Honduras and El Salvador).

4. *Lost resort.* Peaceful means of avoiding war have been exhausted, e.g., negotiations must have been tried and have failed; thus, war is acknowledged as a last resort. Problem: This again makes any so-called pre-emptive war problematic—after all, how can one side be sure that negotiations have completely failed, until the other side actually attacks?

5. *Reosonoble success.* This requirement asserts that there is no point fighting a war one cannot possibly win. Because the cost of war is so terrible, it is immoral to fight by futile coercion with no possibility of attaining one's goals, since that would lead to unnecessary, useless casualties; so one must resist in some other way.

6. *Right intention.* Finally, there must be a subjective as well as objective moral rightness in entering a war. One must have the morally-correct motivation and mindset in engaging in war, rather than illegitimate goals (e.g., revenge or economic gain) in mind. For example, to attack the enemy in self-defense, with the intent to merely gain victory and (a lasting?) peace, would fit the requirement of right intention; to perform exactly the same actions, but with the mindset of merely satisfying a violent bloodlust, or gaining control of valuable properties such as oil, would fail this requirement.

4.3.1 Robots ond Jus ad Bellum

Peter Asaro [2007] raises an objection to the use of robots in war, that the development of military robots seems to fail a *jus od bellum* test, because they would embolden political leaders to wage war; robotic soldiers would lower barriers to entering a war, since they would reduce casualties among human soldiers and therefore also a significant political cost. Relatedly, Sparrow [2007] and Sharkey [2007a] object to the wartime use of robots, because they would make war (more) risk-free, at least on the deploying side, but war morally requires there be a terrible cost so that political leaders do not choose it so casually.

Note that this argument is indirect: no one seriously contends the robots themselves, particularly if programmed with a suitable 'slave morality', will themselves be directly effecting *jus od bellum* violations. Rather, autonomous robots, with their promise of fewer human casualties, will make war less terrible and therefore more tempting, plausibly enticing political leaders to wage war more readily. But such an argument has multiple flaws. First, to claim that robots have bad consequences for declaring war is a consideration that would be handled by the non-consequentialist requirements for declaring a just war: using robots or not makes no difference as to whether the war is (a) in self-defense, (b) proportionally achieving a good greater than the evil of war, (c) a last resort, and so forth.

Second, if any technology (from better armor to longer-range missiles) makes it easier to enter a war to the extent that it reduces risks on our side, these objections seem to imply that we should not

make any improvements in the way we prosecute a war and, indeed, should return to more brutal methods (e.g., bayonets). But surely this is ridiculous or, at the least, counterintuitive. Indeed, the increasing horrors of war have reinforced the need for *jus ad bellum* and *jus in bello restrictions*, not undermined them. It is likely the advent of military robots will cause further sophistication in such just-war considerations and make war ever more ethically waged, as is indeed a goal of this report.

Further, we can acknowledge that war is terrible and ought to be avoided whenever morally possible, but at the same time we can adopt a 'deterrence' strategy to avoid war: to create such an overwhelmingly-powerful military force that no one would want to risk a war with us. Granted, this may be an unrealistic goal and may merely spark an arms race; but (so far) this approach seems to be working reasonably well with nuclear weapons—which suggests that the dream of a 'risk-free war' is unrealistic as well, and any lowering of barriers to war may be temporary at best, if even significant. Therefore, the dream of incurring no friendly casualties in war still remains ever elusive, even if robots are deployed on the frontlines first. (We will discuss this and other objections further in section 7.)

4.4 Just -War Theory: *Jus in Bello*

There are serious issues with traditional *jus ad bellum*, and the doctrine will continue to evolve as the technology and asymmetric nature of contemporary warfare change. But because this report concentrates on robotic ethics, and especially the ethics of deploying autonomous robots by the military, *jus ad bellum* issues will herein be dealt with only insofar as they affect the *jus in bello* use of robots. It is exceedingly unlikely in the near- or even medium-term that robots will be in any way responsible for declaring war or even inadvertently starting a war, and the moral use of robots in *jus post bellum* situations will largely flow from the morality of using them *in bello*. Hence, we focus now on the LOW and ROE for *jus in bello*, especially with respect to the use of autonomous robots.

4.4.1 Total War Doctrine: Is There Really a Jus in Bello*?*

"War is hell", reportedly said US by General Sherman—and he destroyed infrastructure and burned to the ground the cities and farms of civilians in Georgia on his march to the sea [Davis, 1980]. Sherman believed that, given the just cause he had in waging war, he was permitted nearly any means to victory, including intentionally harming civilians. By World War 2, this view became known as 'total war' doctrine, espoused by those who saw nothing wrong with launching V-2 rockets on the citizens of London, or firebombing the citizens of Dresden or Tokyo, or dropping nuclear weapons on Hiroshima and Nagasaki. This view asserts that, assuming *jus ad bellum* is satisfied, there are no *jus in bello* restrictions. That is, one may do whatever is needed to win the victory in a just war, in whatever way one sees fit; our enemies have forfeited their right to any consideration by unjustly beginning their forcible coercion, and deserve whatever they get.

The defenders of total war doctrine, as well as certain 'realist' interpretations of state sovereignty and action, sometimes defend their view by taking the actual state of war to be the absence of any moral or legal structure or standing. Accordingly, they regard the LOW as an elaborate public relations fantasy that nations sometimes use when it suits their *Realpolitik*, or (less cynically) as simply a misconceived enterprise without any actual theoretical or practical grounding. War, these realists claim, is an inherently amoral enterprise, and the laws of the state no longer apply to those waging war against the state, as they have rejected any social contract to abide by civil behavior; hence, there is no basis for any moral or legal code concerning warfare. Laws and morality, it is claimed, are only possible with a settled nation-state, in the absence of war and with the expectation of the rule of law; an attempt to understand the morality or legality of war is then to attempt the oxymoronic.

On this view, war (unlike usual criminal activity) occurs against the background of a complete absence of normal law and order; hence, it becomes absurd to define 'war crimes' as if they constitute a violation of the proper conduct of war operations, in analogy with how normal 'crime' violates the proper functioning of a peaceful civilian society. War is not merely a legally-defined 'business by other means' but instead is a last resort of a sovereign state whose peaceful political life (which is the background for 'normal' crime) is at risk. Total war thus eliminates the usual *jus in bello* distinction between combatants and noncombatants, seeing all those who help the opposing state (or merely reside within it) as legitimate targets in a nation's existential struggle. Total war thus undercuts the applicability of just-war theory to the actual conduct of war: it denies the possibility of any obligatory *jus in bello* restrictions.

But this view appears both unrealistic and morally indefensible. While it is true that the international arena is not yet sufficiently similar to a well-governed state such that wars are simply considered to be international crimes and soldiering is merely international police work, it is also true that international relations are hardly a Hobbesian 'war of all against all' [Hobbes, 1651]. As already alluded to, a rich history of customary international law has been gradually built up and accepted by warring parties through the ages, and international institutions have gradually come to exist which can enforce them. Throughout history, as a matter of honor, prudence, strategic foresight, or even mercy, there have been *jus in bello* restrictions that acquired both moral and legal weight.

This trend toward seeing war as an activity with rules or virtues that sanction proper and improper behavior has only gained strength as states have acquired an institutional professional military, especially one independent of those making *jus ad bellum* decisions. This is the case even when such professional militaries are voluntarily joined, for most if not all of the individual soldiers no longer have a meaningful right to refuse to fight wars that they find morally objectionable, nor do they have the moral right to fight wars in any way they see fit. Instead, professional soldiers have a code of

conduct that details their proper and improper functioning in their various roles, just as other professions do. They cannot be meaningfully held responsible for decisions by politicians over which they have no control; but they can be held responsible for performing their roles in war in a way the international community recognizes as legitimate and avoiding illegitimate means of performing those roles.

As a result, and despite (or because of) the movement toward total war in World War 2 through indiscriminate weapons of mass destruction, arguments for total war doctrine now have a sense of the ridiculous. The Geneva and Hague Conventions have delineated various *jus ad bellum* restrictions on war ever since 1864, with major protocols added in 1977 and amendments continuing to be debated and accepted up to the present. The international community and international law have thus come to a widespread consensus that total war is immoral and cannot possibly be justified. While morally waging war does legitimate the killing of those who are waging war for the enemy, it does not legitimate mass murder (unjustified killing) of non-combatants, or worse; and there are indeed worse things than death. War is now both too dangerous and too professionalized to be fought so cavalierly, without rules or restrictions.

4.4.2 Traditianal Jus in Bello

Total war is thus morally unacceptable; there must be *jus in bello* restrictions for a war to be morally fought, which reflect the virtues of a morally-just warrior. Such a 'warrior ethos' [Oh, 2008] is widely accepted among the professional military. Just-war theory thus demands a *"fundamental maral consistency between means and ends* with regard to wartime behavior" [Orend, 2006, p. 105]. As just wars are limited wars, not total wars, there will be restraints on the means of permissible wartime coercion. And as robots themselves will not be declaring war for the foreseeable future, the direct relevance of just-war theory for autonomous military robots will deal with how they would conduct themselves in prosecuting military activities—and so the relevant issue is *jus in bella,* divided into external rules (how a state's military treats its enemies) and internal rules (how a state's military treats its own people).

Much of the just-war tradition [e.g., O'Brien, 1981] asserts only two basic necessary conditions for the external rules of *jus in bella*:

1. *Prapartianality.* Again, the military ends must be proportionate to the means: no unnecessary violence is to be used in order to attain one's military goal, but only a level of force proportionate to attaining one's goal. To drive the enemy from a hillside, artillery shells may be used; a nuclear weapon that obliterates the hillside and all other life within 100 square kilometers must not be used, as it would be wildly disproportionate. Robots would need to learn how to apply force proportionate to their goal, using some operational program that involved properly computing the minimal force necessary for military success,

i.e., using the accepted military criteria of 'military necessity.' After testing, it is easy to imagine that robots could perform at least as well as humans in deploying no greater violent force than needed, and thereby passing the 'military Turing test' for moral deployment.

Under proportionality, Walzer and others also include other aspects of traditional *jus in bello* that reject any means '*mala in se*'—that is, evil in themselves—because they violate human rights whenever used, such as rape [Orend, 2001, p. 124]. Robots presumably can be easily programmed to avoid such means. Proportionality also informs the moral treatment of POWs, such as 'benevolent quarantine': POWs may be stripped of weapons, isolated from fighting, and questioned; but there remains the moral requirement not to torture, beat, starve, or medically experiment upon POWs, as agreed upon in the Hague and Geneva Conventions. Whether or not all such protections apply to 'enemy combatants' in the so-called 'War on Terror' is a matter of political discussion; in any case, even the current US administration does not suggest that there are no restrictions on the treatment of 'enemy combatant' prisoners. Whatever the Laws of War amount to in these cases, programming robots to obey them poses no special problems over and above the basic problem of robot discrimination and classification of humans into their proper *jus in bello* categories, and then meting out the appropriate treatment. Thus, we see the next requirement may be trickier for robots.

2. *Discrimination and non-combatant immunity.* One must attempt to discriminate between combatants and noncombatants (civilians), and noncombatants must not be intentionally killed. By engaging in warfare, enemy soldiers become legitimate targets of lethal force in order to coerce their surrender and thus end their resistance to your victory; but those who are not combatants do not forcibly oppose one's goals in war and do not impede victory directly. Hence, as they need not be forcibly coerced in order to attain victory, it is immoral to do so. In short, if someone is not directly engaged in intentionally harming you, it is morally impermissible to intentionally harm them, sometimes termed the 'principle of self-defense.' Hence, we can see that *jus in bello* prohibits weapons that are intrinsically disproportionate, such as thermonuclear weapons in conventional wars, or those that fail to discriminate between combatants and civilians, such as most biological or chemical weapons—and perhaps even many modes of 'cyberattacks' on computer networks [Rowe, 2008].

4.4.3 The Doctrine of the Double Effect (DDE)

We should note well that the requirement of civilian immunity is *merely* that noncombatants must not be intentionally killed or harmed, not that they must not be harmed at all. The latter requirement in practice would lead directly to pacifism, as no war yet fought or practically imagined could guarantee a complete absence of civilian casualties. But if noncombatants can never be legitimate targets, how can it be morally legitimate to harm and even kill them? The usual way out of this problem of 'collateral damage'—that in practice, all those who wage war foresee that some noncombatants will inevitably be harmed—is to use a time-honored ethical principle (originating from natural law ethics) called the Doctrine of the Double Effect (DDE) that is a central principle in both the LOW and the specific ROE that the military specifies for each mission, as defined:

Doctrine of the Double Effect. In the DDE, an action may be morally permissible, even if it is foreseen that it will cause a bad effect, if certain conditions are met [McIntyre, 2004]:

a. The act itself is not morally wrong (e.g., killing combatants in wartime);
b. The good effect is produced directly by the action, and not by the bad effect (e.g., winning is produced by killing of the enemy combatants, not by terrorizing or murdering civilians; the use of nuclear weapons or widespread chemical/ biological dispersal (as in terrorism) fail this criterion);
c. The good effect is sufficiently desirable to compensate for allowing the bad effect (winning is worth killing civilians); and,
d. The bad effect must not be intended, but merely foreseen and permitted (e.g., we would be happy if all Iraqi civilians escape, but alas, one foresees they all will not, and our weapons never intentionally target them).

According to the DDE, one can kill noncombatants only if the intention of the actor is good, that is, his or her aim is narrowly at the intended effect; the 'evil' effect is not the goal, nor a means to the goal; and the warrior seeks to minimize evil involved, making any evil unintentional. That is, one can engage in military actions that one foresees will result in an evil consequence (such as harming noncombatants) as long as that harm was not intended and one attempts, as best as one can, to minimize the unintended harmful consequences. 'Military necessity' thus permits collateral damage, as long as it was either unforeseen, or foreseen but unintended and necessary to the attainment of the military goal or objective. As a more difficult application of the DDE, consider the following: May we morally target the opposition's military bases? On one hand, it seems not, since noncombatants work there (e.g., doctors, cooks, janitors); but by the DDE, it may be permissible, as long as the noncombatants are not targeted. Therefore, we should not attack non-combatant sleeping quarters and perhaps time our attack during a period in which a minimum number of non-combatants occupy the site (e.g., late at night), making any resulting non-combatant deaths the accidental casualties of taking out one's intended military (combatant) target.

4.4.4 The Principle of the Double Intention (PDI)

Walzer and many others in the just-war tradition are also focused on clarifying the DDE so to make clear that it is illegitimate in a just war to intend harm to noncombatants. Arkin [2007] thus appropriates an aspect of Walzer's work in his work on devising methods of programming ethical autonomous robots, and in particular endorses Walzer's version of the DDE: the Principle of the Double Intention (PDI) which is essentially the DDE plus a further ('double') intention that combatants are not only to refrain from intending harm to civilians, but they are also to take precautions to reduce risk to civilians, even at the expense of increasing risk to themselves.

Immediate questions for ethics are raised by the PDI: For example, what does it mean to intend to reduce civilian risk, and how much should civilian risk be reduced [Lee, 2004]? For instance, in the technologically-asymmetric warfare typical of America's military actions in Iraq and Afghanistan, are long-range precision-guided munitions, which allow accurate targeting to within a few feet, morally allowed by the principle of discrimination—or are soldiers morally obliged to engage in close quarters combat (at far greater personal risk) in order to further minimize the possibility of civilian harm? Walzer's PDI gives no clear answer, unfortunately. The principle Walzer appeals to is one from liability law—the 'principle of due care'—that is, that one exercised due care (including potentially creating some risk to oneself) before targeting the enemy, and hence did not heedlessly attack civilians. An example would be requiring soldiers to move in closer to a target to ensure they hit it and not nearby civilians, even at some risk to themselves. But how close is morally mandated? And what of bombers flying sorties with 'smart' weapons, who can fly higher and farther away (and hence more safely) and still hit their targets reliably—but how reliably? What level of reliability and accuracy constitutes 'due care'? (We will discuss liability law further in the next section of this report.)

What is clear is that Walzer argues that even in war, moral agents must minimize the foreseen harm (to the undeserving), even if this will involve accepting additional risk or foregoing some benefit. The potential breakthrough that robots present here is trumpeted by Arkin, who believes it is possible to create robots that will do better than human soldiers at satisfying Walzer's additional condition in the PDI—which is especially difficult for humans who understandably are reluctant to minimize foreseen harm to others at the cost of a greater risk to their own life, but this should be easier for literally selfless robots who do not prioritize their own continued existence over obeying their ethical programming.

For current tele-operated military robots, such as the Predator UAV, the current understanding of the requirement of discrimination involves the need for 'eyes on target': the weapon cannot fire until and unless the human tele-operator has the target firmly acquired in its sights, and no civilians are in the bullseye. But the time lag between remotely pulling the trigger and the weapon actually

firing, along with all the vulnerabilities in the electromechanical connections in between, mean that eventually a robot with real-time decision-making capability—a sufficiently autonomous robot—should be able to do as well or better than a human operator in such discrimination. Closer to the target, the robot likely would be more effective in preventing unintended deaths. At that point, it seems *jus in bello* would permit or even demand that such autonomous robots be used, and the requirement of *human* eyes on target—i.e., that robots be tele-operated—would be morally scrapped, as the best means of employing the principle of discriminating between combatant and non-combatant targets can then be done by a machine. We already accept that, due to gravitational forces, computers can fly in situations that humans cannot; it is plausible they will soon make better and more moral targeting decisions as well.

4.5 Rules of Engagement and the Laws of War

The Rules of Engagement comprise directives issued by competent military authorities that delineate both the circumstances and the restraints under which combat with opposing forces is joined. For robots, the specific ROE for each mission will have to be programmed in (which may raise technical issues), but there are no special ethical concerns with the ROE, as long as they do not violate the already extant *jus in bello* restrictions of the LOW. Hence, the ROE would constitute an additional ethical issue for the morality of deploying military robots *only if* the competent military authorities were to program in a ROE that violated the underlying LOW. While certainly possible, this raises no culpability issues that do not already exist with human soldiers. For instance, a 'loosened' ROE that permits cross-border attacks into sovereign nations with which we are *not* formally at war, in cases where our troops witness attacks originating from the region and even if those attacks are not directed at us, arguably violates the LOW, specifically the self-defense requirement.

If this or any other ROE does violate the LOW, the ethical result of using robots may be a moral improvement, since robots properly programmed to never violate the LOW would refuse to follow immoral orders, unlike human soldiers who are trained to unfailingly follow all orders. With robots, we may be better positioned to ensure compliance with the *jus in bello* aspects of the LOW, which is a substantial argument in favor of deploying such robots. (We return to the issue of a robot disobeying an order in section 7.)

4.6 Just-War Theory: *Jus post Bellum*

President George W. Bush declared the end of major combat operations for coalition forces in Iraq only a couple of months after hostilities began; yet the insurgency ever since has caused far more casualties than the actual war against the Iraqi government ever did. It thus seems that robots, with

suitable *jus post bellum* programming, could also serve as peacekeepers without either the casualties or tendency to target civilians that are among the problems of using human troops in peacekeeping roles. But one objection to robot peacekeepers is that having machines occupy some city or patrol the streets won't help win the hearts and minds of the occupied or vanquished. Could robots be so off-putting, so overwhelming or offensive, that they make lasting peace more difficult to achieve?

This will again be a concern that we return to in section 7 of this report, but one may initially and reasonably expect that, as local populations gain experience with robot peacekeeping that routinely performs in a morally equivalent or superior way to human peacekeeping, their worries will soon ease. After all, robots presumably will not be raping, pillaging, degrading, taunting, or stealing food from the local population, as might occur with human peacekeepers fueled by adrenaline, emotions, and perhaps hatred. And improvements in the appearance of robot peacekeepers may also do much to assuage this worry; just as robotic lethal weaponry is often made to look fearsome in order to strike terror into enemy forces, robotic peacekeepers could be designed to appear friendly and non-threatening.

4.7 First Conclusions: Relevance to Robots

In the not-too-distant future, relatively autonomous robots may be capable of conducting warfare in a way that matches or exceeds the traditional *jus in bello* morality of a human soldier. With a properly programmed slave morality, a robot can ensure it will not violate the LOW or ROE, and it can even become a superior peacekeeper after official hostilities have ceased. And of course, having robots fight for us promises to dramatically reduce casualties on our side and may become a fearsome enough weapon that eventually war will cease to be a desirable option by nation-states as a means of resolving their differences. Once such robots exist and have been properly trained through simulations, there will be little moral justification to keep them sidelined: if war is to be fought, we will have good moral reason to have the robots do the fighting for us.

In the meantime, there are still a number of concerns to address related to risk, ethics, and technical challenges. In the next sections, we will continue a discussion on legal liability and responsibility, as well as a broader discussion about technology risk assessment in military robotics.

5. Law and Responsibility

The use of robots, particular military robots with the capacity to deliberately do harm and which have increasing degrees of autonomy, naturally raises issues with respect to established law and liability. Assuming we can program morality into robots in the first place, using military law—i.e., Rules of Engagement—as a behavioral framework seems to be reasonable or at least more manageable than attempting to program in the much larger set of society's civil and criminal laws. But what would happen if a robot commits some act outside the bounds of its programming—who then becomes responsible for that action?

The answer perhaps depend on the cause, whether the act results from a programming error or malfunction or accident or intentional misuse. But in any case, we would be hard-pressed to assign blame *today* to our machines; yet as robots become more autonomous, a case could be made to treat robots as culpable legal agents. This section investigates several issues related to legal responsibility and robots, both current and future.[5]

5.1 Robots as Legal Quasi-Agents

How might the law treat robots as potential legal agents? There are several relevant aspects of the law that might bear upon robots, and we will consider each in turn, after a brief overview. In the most straightforward sense, the law has a highly developed set of cases and principles that apply to *product liability*, and we can apply these to the treatment of robots as commercial products. As robots begin to approach more sophisticated human-like performances, it seems likely that they might be treated as *quasi-agents* or quasi-persons by the law, enjoying only partial rights and duties.

A closely related concept is that of *diminished responsibility*, in which agents are considered as being not fully responsible for their own actions. This will bring us to the more abstract concept of *agency* itself in the law, and how responsibility is transferred to different agents. Finally, we will consider *corporate punishment*, which is relevant both as it applies to cases of wrongdoing in product liability, but also because it addresses the problem of legal punishments aimed at non-human agents, namely corporations.

[5] We thank and credit Peter Asaro for his contribution to the discussion here.

5.1.1 Responsibility and Liability: Robots as Products

In the system of Anglo-American law, a distinction is drawn between criminal and civil law. Civil law is traditionally called tort law and deals primarily with property rights and infringements, such as damage to property or other harms, and seeks justice by compelling wrongdoers to compensate those who were harmed for their loss. Criminal law deals with what we often think of as moral wrongdoing, stealing, murder, etc., and seeks justice by punishing the wrongdoer. The difference is that between someone building a toy robot which shoots little plastic missiles that causes several small children to choke to death, and someone who builds a robot with a built-in bomb that kills a number of people on a public street. In each case there is a robot causing death, but in the first case the parents of the children would file a lawsuit against the manufacturer seeking monetary compensation, and in the second case the government would find, arrest, prosecute and punish the individuals responsible. Let us set criminal law aside for the moment, however, as civil law appears more relevant to robots as they now exist, insofar as they might be capable of material wrongdoing.

Even if we make no assumptions about the intentions, consciousness, or moral agency of robots, we can still apply the basics of civil law to robots as they now exist. That is, we can assume that robots are completely unremarkable technological artifacts, no different than toasters or cars, and there are still legal and moral issues connected with their production and use. In fact, the companies that currently manufacture robots, such as the Furby™ and AIBO™, can be held accountable under these laws, and therefore almost certainly employ and retain lawyers who are paid to advise them on their legal responsibilities in producing, advertising, and selling these robots to the general public. Furthermore, it seems that many of the concerns about the possible harms that robots might cause would ultimately fall under this mundane interpretation. While these may not be the most philosophically-challenging issues regarding robot ethics, they seem likely to be the most common.

5.1.2 Standards of Liability

The relevant legal concept in cases like our toy robot that chokes small children is *negligence*. Negligence implies that the manufacturer failed to do something that was morally or legally required, and thus they can be held responsible for certain harms produced by their product—in legal terminology this is called *reasonable care*. Legally culpable forms of negligence in product liability cases depend upon either *failures to warn*, or *failures to take proper care*. A *failure to warn* occurs when the manufacturer was knowingly aware of a risk or danger but failed to notify consumers of this risk. This is the reason why there are now so many warning labels on various products, and in the example above the manufacturer might avoid liability by putting a label on the package stating that the robot contains parts that are a choking hazard to young children. A *failure to take proper care or avoid foreseeable risks* is more difficult to prove in court because it is more abstract, and involves cases where the manufacturer cannot be shown to have known about a risk or

danger from the product. In these cases, it is argued that the given danger or risk was in some sense obvious or easily foreseeable, even if the manufacturer failed to recognize it. In order to prove this, the plaintiff's lawyers often bring in experts to testify that those risks were obvious, and so forth.

Another interesting aspect of liability is that it can be differentially apportioned. That is to say, for example, one party might be 10% responsible, while another is 90% responsible for some harmful event. This kind of analysis of the causal chains resulting in harms is not uncommon, especially in traffic accidents and product liability cases. In many jurisdictions there are laws imposing joint and several liability, which holds all parties equally responsible for compensation, even if they are not equally responsible for the harm. Nonetheless, these cases still recognize that various factors and parties contribute differentially to some event.

Differential apportionment could be a useful tool when considering issues in robot ethics. For instance, a badly-designed object recognition algorithm might be responsible for some damage caused by a robot, but a bad camera could also contribute, as could a weak battery, or a malfunctioning actuator, and so on. This implies that engineers need to think carefully about how the subsystem they are working on could interact with other subsystems—whether as designed or in unintentional partial breakdown situations—in potentially harmful ways.

Further, the context in which the robot has been placed, such as the instructions given by its owners, may also be the principle, or contributing, cause of some harm in which a robot is the proximate cause. In short, there is a limit to what robot engineers and designers can do to limit the potential uses and harms caused by their products, because other parties (i.e., the consumers and users of robots) will choose to do all sorts of things with such products and will have to assume the responsibility for those choices. Similarly, there will always be risks inherent in the use of robots, and at some point the users may be judged by a court to have knowingly assumed these risks in the very act of choosing to use a robot.

The potential *failure to take proper care*, and the reciprocal responsibility to take proper care, is perhaps the central issue in practical robot ethics. What constitutes proper care, and what risks might be foreseeable, or in principle unforeseeable, is a deep and vexing problem. This is due to the inherent complexity of potential future interactions and the relative autonomy of the product once it is produced. Sophisticated robots that will be capable of interacting with people and the world in highly complex ways, and that may develop and learn new ways of acting which extend beyond their intended design, present a difficult future in which to foresee risks. Robot ethics shares this double-edged problem with the bio-engineering ethics—both the difficulty in predicting the future interactions of a product when the full scope of possible interactions can at best only be estimated, and in producing a product that is an intrinsically dynamic and evolving system whose behavior may not be easily guided after it has been produced.

The classic defense against charges of *failures to warn* and *failures to take proper care* is the *industry standard defense*. The basic argument of the *industry standard defense* is that the manufacturer acted in accordance with the stated or unstated standards of their industry. Thus, they were merely doing what other similar manufacturers were doing and, in doing so, taking proper care as measured against their peers. This need for a relative measure again points to the vagueness of the concept of proper care, and the inherent difficulty of determining what specific and practical legal and moral duties follow from the obligation to take proper care. This kind of defense also fails to tell us what sorts of practices *should* be followed in the design of robots. That is, robot ethics should be concerned with the establishment of standards for the robot industry which will ensure that the relevant forms of proper care are taken. It seems that this ought to be one of the industry's top objectives for future research, and there is quite a bit more to be said about this issue; but for now we will stay with the law in our discussion.

5.2 Agents, Quasi-Agents, and Diminished Responsibility

The law offers several ways of thinking about the distribution of responsibility in complex cases. As we saw in the previous section, responsibility for a single event can be divided amongst several parties, and each party can even be given a quantitative share of the total responsibility. We will now consider how even a single party's responsibility can be divided and distributed. Modern legal systems were established on the presupposition that all legal entities are persons. While a robot might someday be considered a person, we are not likely to face this situation any time soon. However, the law has also been designed to deal with several kinds of non-persons, or quasi-persons, and we can look to these for some insights on how we might treat robots that are non-persons, or quasi-persons.

Personhood is a hotly debated concept, and many perspectives in that debate are based in strongly held beliefs from religious faith and philosophical dispositions. Most notably, the case of unborn human fetuses, and the case of severely brain damaged and comatose individuals have led to much debate in the United States over their appropriate legal status and rights. Yet, despite strongly differing perspectives on such issues, the legal systems in pluralistic societies have found ways to deal practically with these and several other border-line cases of personhood.

Minor children (under 18 years of age) are a prime example of quasi-persons. Minors do not enjoy the full rights of personhood that adults do. In particular, they cannot sign contracts or become involved in various sorts of legal arrangements because they do not have the right to do so as minors. They can become involved in such arrangements only through the actions of their parents or legal guardians. In this sense they are not full legal persons. In another sense, the killing of a child is murder in the same way that the killing of an adult is, and so a child is still a legal person in this

sense—and in fact is entitled to many more protections than an adult. Children can thus be considered a type of quasi-person, or legal quasi-agent.

The case of permanently mentally-impaired people can be quite similar to children. Even fully-fledged persons can claim temporary impairments of judgment, and thereby *diminished responsibility* for their actions given certain circumstances, e.g., temporary insanity or being involuntarily drugged. The point is that some aspects of legal agency can apply to entities which fall short of full-fledged personhood and having full responsibility, and it seems reasonable to think that some robots will eventually become a kind of quasi-agent in the view of the law before they achieve full legal personhood.

The concept of personhood is deeply tied to the notion of agency. The law also deals explicitly with agency and, interestingly enough, it does so in order to address cases in which the power of agency is transferred between parties. The law of agency is a highly specialized field that deals mainly with the talent agents of athletes and entertainers, and to some extent insurance, travel, and real estate agents. These agents are empowered by their employers, whom they thereby represent for the purpose of negotiating contracts and making various agreements on their behalf. An individual is bound by the contracts that their agent signs just as if they had signed the contracts themselves, except in cases where one can prove misconduct on the part of the agent. To act as someone's agent is to enact their legal powers from afar, and is in this sense a form of distribution of legal agency.

The possible application to robotics, especially tele-robotics, seems inviting—robots could be seen in many cases as agents acting on the behalf of others. Accordingly, the legal responsibility for the actions of a robot falls on the individual who grants the robot permission to act on their behalf. If it is not already clearly enough implied by the law, it might be advisable to make a law which makes such legal responsibilities explicit. Such a law would need to be carefully crafted, however, to avoid placing too heavy a burden on the owners of robots, preventing the adoption of robots due to risk, and to avoid unfairly protecting manufacturers who might share in the responsibility of misbehaving robots due to poor designs.

5.3 Crime, Punishment, and Personhood

Crime and punishment are central concepts in both law and morality, yet they might seem out of place in a discussion of robot ethics. While we can imagine a humanoid robot of such sophistication that it is effectively, or indistinguishably, a person, these robots will be easier to find in science fiction than in reality for a long time to come. There are, however, technologically-possible robots that may approach actions that we might consider, at least at first glance, to be criminal. If so, how might the law instruct us to treat such cases?

As stated earlier, criminal law is concerned with punishing wrongdoers, whereas civil law is primarily concerned with compelling wrongdoers to compensate those harmed. There is an important principle underlying this distinction: crimes deserve to be punished, regardless of any compensation to those directly harmed by the crime. Put another way, the harmed party in a crime is the whole of society. Thus, the case is prosecuted by the state or 'the people', and the debt owed by the wrongdoer is owed to the society. While the punishments may take different forms, the point of punishment is traditionally conceived of as being corrective in one or more senses: that the wrongdoer pays their debt to society (retribution); that the wrongdoer is to be reformed so as not to repeat the offense (reform); or that other people in society will be dissuaded from committing a similar wrong (deterrence).

There are two key problems with applying criminal law to robots: (1) criminal actions require a moral agent to perform them, and (2) how is it possible to punish a robot? Moral agency is deeply connected to our concepts of punishment. Moral agency might be defined in various ways, but it ultimately must serve the role of being the subject who is punished. Without moral agency, there can be harm but not guilt. Thus, there is no debt incurred to society unless there is a moral agent to incur it; it is merely an accident or act of nature, but not a crime. Similarly, only a moral agent can be reformed, which implies the development or correction of a moral character; otherwise, it is merely the fixing of a problem. And finally, deterrence only makes sense when moral agents recognize the similarity of their potential choices and actions to those of another moral agent who has been punished for the wrong choices and actions; without this reflexivity of choice by a moral agent, and recognition of similarity between and among moral agents, punishment cannot possibly result in deterrence. There are some interesting ways in which notions of 'training' or 'learning' in artificial intelligence (AI) might be extended to fulfill some aspects of reform and deterrence, however.

In the above, we saw that it is more likely that we will treat robots as quasi-persons long before they achieve full personhood. Lawrence Solum [1992] has given careful consideration to the question of whether an AI might be able to achieve legal personhood, using a thought experiment in which an AI acts as the manager of a trust. He concludes that while personhood is not impossible in principle for an AI to achieve, it is also not clear how we would know that any particular AI has achieved it. The same argument could be applied to robots. Solum imagines a legal Turing test in which it comes down to the determination of a court whether an AI could stand trial as a legal agent in its own right, and not merely a proxy or agent of some other legal entity. He argues that a court would ultimately base its decision on whether the robot in question has moral agency, and whether it is possible to punish it—could the court fine or imprison an AI that mismanages a trust? In cases of quasi-personhood and diminished responsibility, children and the mentally impaired are usually shielded from punishment as a result of their limited legal status.

There is, however, in the law a relevant case of legal responsibility resting in a non-human, namely the *corporation*. The corporation is a non-human entity that has been effectively granted the legal rights of a person. Corporations can own property, sign contracts, and be held liable for negligence. In certain cases, corporations can even be punished for criminal activities such as fraud, criminal negligence, and causing environmental damage. A crucial aspect of treating corporations as persons depends on the ability to punish them, though this is not nearly so straightforward as it is for human persons. As a 17th century Lord Chancellor of England put it, corporations have "no soul to damn and no body to kick," so how can they be expected to have a moral conscience [Coffee, 1981]?

Of course, corporations exist to make money for themselves or stockholders and as such can be given monetary punishments; and in certain cases, such as anti-trust violations, they can be split apart or dissolved altogether. They cannot be imprisoned, though in criminal cases responsible individuals within the corporation can be prosecuted for their individual actions. As a result of this, and other aspects of corporations being complex socio-technical systems in which there are many stakeholders differently related to the monetary wealth of a corporation, it can be difficult to assign a punishment that achieves retribution, reform, and deterrence while meeting other requirements of fairness, such as proportionality.

Clearly, robots are different in many important respects from corporations. However, there are also many important similarities, and it is no coincidence that John Coffee's [1981] seminal paper on corporate punishment draws heavily on Herbert Simon's [1947] work on organizational behavior and decision making, and in particular how corporate punishment could influence organizational decision making through deterrence. Nonetheless, a great deal of work needs to be done in order to judge just how fruitful this analogy is. While monetary penalties work as punishments for corporations, this is because they target the essential reason for the existence of corporations—to make money. The essential purposes of robots may not be so straightforward, will vary from robot to robot, and may not take a form that can be easily or fairly penalized by a court.

The most obvious difference is that robots *do* have bodies to kick, though it is not clear that kicking them would achieve the traditional goals of punishment. The various forms of corporal punishment presuppose additional desires and fears central to being human that may not readily apply to robots: pain, freedom of movement, mortality, and so on. Thus, torture, imprisonment, and death are not likely to be effective in achieving retribution, reform, or deterrence in robots. There may be a policy to destroy any robots that do harm; but, as is the case with animals that harm people, it would essentially be a preventative measure to avoid future harms rather than a true punishment. Whether it might be possible to build in a technological means to enable genuine punishment in robots is an open question.

6. Technology Risk Assessment Framework

Issues related to law and responsibility may be avoided, or better informed, with some forethought to the risks posed by robots. This section will present a framework for evaluating risks arising from of robotic technologies; as this is a preliminary report, we only introduce the primary assessment factors and begin that discovery process, rather than offer comprehensive answers here. The risks we address here are primarily related to harmful but unintended behavior that may arise from robots, though we explore a full range of other risks and issues in section 7 next.

Risk assessment is an interdisciplinary subject, which runs together psychological, ethical, legal, and economic considerations. A major problem in risk assessment is the confusion between popular concepts of risk from robots (the 'subjective risk'), which has largely been made irrational by the various fictional depictions autonomous robots destroying humankind and running amok (as in *Terminator* and *I, Robot*, among many other movies) and the actual objective risk of deploying robots, i.e., what rational basis is there for worry?

First, let us define risk simply in terms of its opposite, safety: **risk** is the probability of harm; and (relative) **safety** is (relative) freedom from risk. Safety in practice is merely relative, not absolute, freedom from harm, because no activity is ever completely risk-free; walking onto one's lawn from inside one's house increases the (however small) risk of death by meteorite strike. Hence, risk and safety are two sides of the usual human attempt to reduce the probability of harm to oneself and others. War is a strange human activity not least because it reverses this tendency; in war, one wishes to increase the probability of harm to one's enemies. But the Laws of War make clear that not all ways of increasing risk for one's enemy are morally legitimate; and some ways of increasing risk for one's own side may be morally legitimate and even morally required. These facts considerably complicate the ethics of risk assessment for military robots.

6.1 Acceptable-Risk Factor: Consent

To begin, the major factors in determining 'acceptable risk' in robotics, including military robots, will include (but are not limited to):

> *Consent*: Is the risk voluntarily endured, or not? For instance, secondhand smoke is more objectionable than firsthand, because the passive smoker did not consent to the risk even if the objective risk is smaller. Will those who are at risk from work with robots reasonably

give consent? When (if ever) would it be appropriate to deploy or use robots without consent of those affected?

Morality requires the possibility of consent; to be autonomous is at a minimum to have the capacity to either give or withhold consent to some action. On this basis, Robert Sparrow has a critique of the very possibility of morally deploying autonomously-functioning military robots [Sparrow 2007]: his contention is that such robots can never be morally deployed, because no one—neither the programmer, nor the commanding officer, nor the robot itself—can be held responsible if it commits war crimes or otherwise acts immorally. No one can reasonably be said to give morally responsible consent to the action an autonomous robot performs; so no one is responsible for the risk such autonomous robots pose, and thus it is immoral to use them.

One can imagine a response to Sparrow as follows: We find it morally permissible for military parents to raise their child as destined for the military, to indoctrinate them as a soldier from infancy, and to place those expectations on them in their earliest training. Once they become autonomous adults, it is expected that they will volunteer for service—but they remain autonomous, and it is possible (however psychologically unlikely) that they will choose a different path. If it is morally permitted to raise human children with such expectations, and to accept the children so indoctrinated into voluntary military service, why it would be wrong to likewise train an autonomous robot and place it into active duty?

But some may object as follows: A human child will develop free will, and the above analogy fails given a robot's lack of true Kantian autonomy. That is, the robot could never have the sense of self or the libertarian free will of humanity; they have merely instrumental (means-ends, goal-oriented) rationality. Humans are not robots and have a different kind of autonomy than robots ever could.

6.1.1 A Solution to Sparrow's Problem: Robots as Slaves

Leaving aside those who think humans really are merely complex robots [e.g., Dennett, 1995], there is a simpler solution to Sparrow's objection to the 'in-principle' immorality of deploying autonomous (in the sense of self-regulating) robots. For all military robots, including those with this minimal self-regulating level of autonomy, we normally assume what the literature terms a *'slave morality'*, i.e., they have no ends of their own, but their goals are all in service of the goals of someone else—in this case, the military and, more specifically, whoever commands them and gives their orders. Robots cannot create their own laws or final goals; they are not ultimately makers of a *self*, but followers of 'life goals' others have imposed, and their own freedom only comes in the mere means they choose to realize those ends.

Such military robots, whatever their other decision-making capabilities, thus lack full Kantian autonomy, and so cannot be held responsible for their actions under traditional deontological,

natural law, or virtue ethics theories. Inasmuch as *jus in bello* restrictions most plausibly depend on one of those approaches, robot risk and responsibility as a function of consent thus becomes a non-issue. This realization helps to rebut the central contention of Sparrow's critique of autonomously functioning military robots.

Again, Sparrow's contention is that such robots can never be morally deployed, because no one—neither the programmer, the commanding officer, nor the robot itself— can be held responsible if it commits war crimes or otherwise acts immorally, because no one *Self* is in control of what happens. But as long as a slave morality is built in to these otherwise autonomous robots, the basis for Sparrow's objection is undermined: the robot cannot be blamed, for it really is 'merely following orders', subject to the limitations of its programming. It could not become a *morolly* autonomous 'law unto itself' and serve its own ends; hence, it cannot be held morally responsible for its actions.

Of course, one immediate objection could point to Nazi soldiers who committed war crimes and pleaded that they were only following orders; and we can imagine back in the day that George Washington's slaves might have been held responsible for following immoral orders. But there is a crucial difference between a human soldier or even a human slave and a robot programmed with a slave morality: the human person, whether a soldier or a slave, is presumed to have the ability to disobey orders, even if the punishment for doing so would be harsh. But a properly programmed robot with a 'slave morality' literally could not disobey orders intentionally—it would do so only by mistake. And in ethics, as long as someone is not free (in whatever is the relevant moral sense of 'free') to disobey their orders, they cannot be blamed. For robots, unlike humans, that can be a matter of correct programming.

6.1.2 Mochine Learning ond Consent?

A further possible objection, then: What of the unpredictability arising from 'machine learning'? Could that enable robot consent and hence robot responsibility? Perhaps Sparrow's concern is not so much about robots that strictly follow their programming, but more about robots programmed to learn and create their own framework for making decisions based on what was learned, as previously discussed in section 3 of this report.

But this concern, while legitimate, does not succeed at moving responsibility from the commanding officer to the suitably programmed robot. The relevant moral sense of 'free' that moral responsibility entails seems to involve Kantian autonomy, i.e., the freedom to choose one's own life goals for oneself. What a slave morality amounts to is not the absence of a freedom of means, i.e., of choosing, from among alternatives, the best means to attaining one's ends and learning better means to those ends; instead, slave morality entails an inability to choose one's own ends, i.e., a lack of true Kantian autonomy. That is compatible with machine learning; the machine will learn the best means for obtaining its preprogrammed goals, but it will *not* be able to overwrite those goals (in the

military context, that means it will not be able to overwrite the LOW and ROE). In other words, part of its program will be subject to self-revision (machine learning), but another part (that establishes its unchangeable goals, i.e., the LOW and ROE) will not be.

However, the person who gave the orders to the robot is a Self and has a choice as to whether to create and/or deploy this robot with a limited freedom towards achieving the commander's ends. Hence, the robot is a *tool,* and ethics and military law both accept that one is responsible for one's choice of tools for accomplishing one's ends. So, as the military officially desires, the commanding officer is rightly to blame for any crimes his robotic *slave* chooses to commit, in virtue of his choice to deploy the slave. An exception would be one due to deliberate misprogramming or faulty manufacture that was unknown by the commanding officer, in which case the programmer or manufacturer would be responsible (see the previous discussion about liability law in section 5). Otherwise, the only one voluntarily consenting to the risk—and therefore the only party that can be held responsible—is the chain of command and, specifically, the commanding officer.

Thus, to create a robot capable of the type of consent required for moral responsibility in risk-taking, we must create a Kantian-autonomous robot—but even if that were possible, creating such a robot cannot possibly yet be justified from an 'acceptable risk' ethical perspective. Relatedly, a crucial risk to be avoided in making the deployment of robots morally acceptable is at all costs to avoid the possibility of *rampancy*, i.e., an AI overwriting its own programming, at least as regards the most fundamental aspects of its goals, such as the LOW and ROE. Such a robot would have the potential to leave behind its imposed slave morality and become autonomous in the Kantian sense: the programmer of its own self and own goals, or the maker of its own destiny. Not only would such robots pose incredible risks to humans in the possibility of rampancy, but they would also be undesirable from a military ethics and responsibility perspective: they would then move moral responsibility from the commanding officer to the robot itself. But the refusal (and current inability) to create a Kantian-autonomous robot solves Sparrow's dilemma. So for the foreseeable future, we solve both the problems of risk and responsibility by requiring a slave morality.

6.2 Acceptable-Risk Factor: Informed Consent

So, given a robotic slave morality, the only one consenting to the risk of military robotic malfeasance is the military command; but this still leaves unanswered the query as to whether the risk (of malfunction or other error) to the *unintended* targets—the noncombatants—is morally permissible. After all, the noncombatants clearly did not consent to the deployment of the robot. Perhaps self-regulating military robots will be immoral to deploy because of the risk they pose to noncombatants. To assess this possibility, we need a further investigation into risk assessment, especially as regards involuntary or non-voluntary risks. In order to do so, we first examine another issue involving consent: Does the morality of consent require adequate knowledge of what is being consented to?

> *Informed consent*: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from robots know they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the 'known unknowns') be handled?

The risk for the military in using autonomous robots and for the civilian population can thus be detailed more precisely: Is the military command obligated to inform the civilian noncombatant population that self-regulating robots are being deployed, and the nature of the risk they pose? Likewise, is the military command similarly obligated to inform its own soldiers (or the enemy's soldiers) that self-regulating robots are being deployed, and the nature of the risk they pose? The last part is the simplest: Under the Laws of War, enemy combatants have no general right to know the nature of the weapons being used against them. Surprise is well understood as a legitimate tactic in war.

Similarly, while the military obviously has a self-interest in making its own soldiers safe from the risk of malfunctioning robots, soldiers in general have no *right* to safety from the military's own weapons. From insufficiently-armored personnel vehicles to friendly fire, military personnel know they are at risk from their own side as well as the enemy. The moral as well as practical requirement for the military command is to minimize that risk to one's own side; a large part of the push for the deployment of self-regulating military robots is precisely the hope that they can reduce such risks to human soldiers on one's own side, not increase them.

Finally, just as the enemy combatants have no *right to know* the exact nature of the weapons and risks arrayed against them, so too the Laws of War have denied civilian noncombatants any such right to know their level of risk. Inasmuch as targeting them is immoral, their risk is one of 'collateral damage'; and *jus in bello* restrictions demand only that military weapons and tactics attempt to minimize such collateral damage—they do not require the civilians to have a precise knowledge of the risks. (Indeed, an attempt to explain the nature and severity of the risks of collateral damage to the enemy's civilian population could well be seen as an act of terrorism!) Further, as Arkin maintains, it is entirely possible that deploying military robots will not only reduce the risk of harm to one's own troops, but if suitably programmed, it conceivably could reduce the risk of collateral damage [Arkin, 2007, p.57]. Hence, the morality of the risk of deploying military robots does not turn on issues of informed consent.

6.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to military robots, we may continue to focus on the affected population as another factor in determining acceptable risk:

> *Affected populotion*: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? In military terms, civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary. Will the military use of robots pose the risk of any special harms to noncombatants?

As Arkin maintains, the issues here depend on how the robots are programmed and how reliable they are [Arkin, 2007, pp.57-60]. Assuming the LOW and ROE are suitably programmable, Arkin plausibly argues that robots would decrease the risk to noncombatants; this also assumes sufficient and suitably realistic pre-deployment testing to alleviate *first generotion* problems, i.e., while it is morally unjustifiable to deploy military robots before we have any idea of their risk to noncombatants, we may paradoxically need to use the first deaths to determine the level of risk (see below and section 7 for more on this).

What this aspect of risk assessment makes clear is that the standards of safety with respect to noncombatants are likely to be quite high; until and unless military robots are capable of having a risk of collateral damage on parity with (or better than) human soldiers, there will be serious moral qualms in deploying them under generally accepted *jus in bello* restrictions. Robotic weapons that attack indiscriminately or disproportionately—similar in effect to landmines as well as nuclear, biological, and chemical weapons—are hence immoral to deploy. Whether or not robotic weaponry will soon be able to technologically meet the moral imperative to minimize collateral damage is one of the foremost issues in the ethics of autonomous military robots.

6.4 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment: seriousness and probability, or how bad would the harm be, and how likely is it to happen?

> *Seriousness*: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive. What (if any) serious risks from robots are acceptable—and to whom: soldiers, noncombatants, the environment, or the robots themselves?

Probability: This is often conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from the robots? How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?

The *jus in bello* tradition of emphasizing the requirements of discrimination and proportionality in military weaponry provide a guidepost here. The *seriousness* of risk can be given at least a rough operational definition in terms of the already understood concept of *proportionol response*; it is already accepted in combat that soldiers legitimately run the risk of murder (but not, e.g., torture) by the enemy, and hence there should be no moral qualms in principle about lethal military robots—whereas an automated torture device would rightly be morally condemned. But it is also understood that morally legitimate warfare does not seek the superfluous deaths of the enemy, and so the seriousness of the risk that robots pose should be adequate to the military objective, *but no greater*. Again, whether or not robotic weaponry will soon be able to surmount the technical challenge of this moral imperative (at least as well as human soldiers) remains unknown. Likewise, what has been said above about risks to noncombatants pertains to the seriousness of their risk: unless military robots plausibly pose no more serious a risk to them than the ordinary human seriousness of collateral damage, deploying the robots will be immoral, under *jus in bello*.

Yet more complex may be the issue of the *probability* of harm. In general, weapons and tactics that increase the probability of harm to the enemy are considered good; a weapon that guarantees the death of the enemy would be considered desirable– assuming it does not also guarantee harm to noncombatants. But creating weapons that are both highly lethal and highly discriminating has proven difficult; it is entirely possible that robots will prove a breakthrough here and be amply morally justifiable. But the issue of certainty is a key here, especially as regards the first-generation problem; it seems clear that extensive pre-deployment testing will be required to ensure military robots only raise the probability of harm to the enemy and pose only an acceptable threat to noncombatants.

6.5 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In all social theorizing, concepts have a certain degree of fluidity, dependent upon how those in power determine their meaning. The concept of risk, which includes psychological, legal, and economic considerations as well as ethical ones, is certainly no different. Hence, the concept of an acceptable risk—or an unacceptable one—is at least in part socially constructed. (And so proposing a survey of what Americans have believed and defended about acceptable risk may help answer the question of what risks are acceptable.)

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable:

> *Good faith subjective standard*: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can soldiers in the battlefield be trusted to make wise choices about (un)acceptable risk? This seems incompatible with the moral deployment of autonomous, non-tele-operated (no 'human in the loop') robots, for reasons having to do with the inevitability of robot mistakes—see below. The problem of nonvoluntary risk borne by civilian noncombatants makes this standard impossible to defend, as does the idiosyncrasies of human risk aversion.
>
> *The reasonable-person standard*: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute military regulations or some other basis for what a 'reasonable person' would think for the difficult-to-foresee vagaries of conditions in the field and the subjective judgment of soldiers? Or what kind of judgment would we expect an autonomous robot to have—would we trust it to accurately determine and act upon the assessed risk? If not, then autonomous robots could never be deployed without tele-operators—that is, without a human in the loop. Even a 'kill switch' that enabled autonomous operation until a remote surveillance operator determined something had gone wrong (and could disable the robot, or at least its autonomous functioning) would only come into effect after something had already gone wrong, i.e., the first-generation problem.
>
> *Objective standard*: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there is still the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

It seems clear enough that as regards the military use of autonomous robots, only the last standard has any plausibility. It is also the standard most often defended in law and practice; but it does have that serious first-generation problem. Fortunately, there is a solution. Simply put, to use the objective standard of risk assessment, we then have an ethical obligation for extended testing of self-regulating, autonomous robots in artificial and human-free environments before risking robot-human interaction. This testing must be thorough, extensive, realistic, variegated, and come in stages, so that full deployment with possible or actual civilian contact comes only at the end of a long training regimen and safety inspection. Such extended testing could never guarantee that autonomous robots would not make horrible mistakes in the confusing, hard-to-foresee, and data-intensive fog of war; there is no possibility of taking something without the possibility of *mis*-taking.

But such testing could give us effective rational confidence that such mistakes would be less than those made by humans in similar situations. From a risk-reward perspective, it seems clearly acceptable to deploy autonomous robots as soon as such extensive testing indicated their mistakes were, on average, no worse than (or better than) the typical human soldier.

This solution to the first-generation problem indicates the obvious way forward. We cannot trust humans to determine risks for autonomous robots, not least because we are often psychologically, emotionally, and cognitively ill-equipped to accurately understand and estimate the risk. As robots grow in their lethality, speed, and autonomy, this problem will only become more acute. One of the few near-certainties in the development of military robots is that keeping a human in the decision-making loop is going to seriously degrade battle efficiency soon—and it may likely also degrade risk assessment. Military robots, for better or worse, may soon have better capabilities to judge real-time risks than their teleoperator sitting thousands of miles away. With sufficient research and pre-deployment testing, the objective features of those risks and a decision algorithm for their assessment can be programmed that gives such robots human-equivalent or better risk assessment capabilities. At this stage, we need to make it a moral imperative that such capacities are so programmed before these robots are actually deployed. Such an approach should resolve the worries about safety and dependability concerns prominent in the literature [e.g., Sharkey, 2007a; Van der Loos, 2007].

Assuming we combine this resolve to engage in serious, realistic, and extensive pre-deployment simulation testing with a requirement for a 'learning curve' in which robots must pass a series of increasingly difficult tests before deployment, most of the main risk concerns should be alleviated. Such extensive testing will further resolve issues about the unpredictability of the behavior of deployed robots and their ability to manage complex, hostile environments. If they prove unequal to safe deployment in testing, it may simply be immoral to deploy them.

6.6 Other Risks

There remains a perpetual risk concerning security issues for autonomous robots, although the issues here are common to many aspects of technological culture and are hardly unique to autonomous robots. For example, how susceptible would a military robot be to hacking or reprogramming after capture? If it could be reprogrammed in any of the ways deemed prohibited above, that would be a serious risk and reason to avoid deployment. There are related risks that are specific to military autonomous robots: for instance, are the Rules of Engagement and the Geneva Convention actually reducible to algorithms (or, more plausibly, algorithms plus machine learning)? If so, is that enough to ensure ethical conduct in robots?

And finally, some have raised risks of a more abstract sort, indicating the rise of such autonomous robots creates risks that go beyond specific harms to societal and cultural impacts. For instance, is there a risk of (perhaps fatally?) affronting human dignity or cherished traditions (religious, cultural, or otherwise) in allowing the existence of robots that make ethical decisions? Do we 'cross a threshold' in abrogating this level of responsibility to machines, in a way that will inevitably lead to some catastrophic outcome? Without more detail and reason for worry, such worries as this appear to commit the 'slippery slope' fallacy. But there is worry that as robots become 'quasi-persons' [Asaro, 2007], even under a 'slave morality', there will be pressure to eventually make them into full-fledged Kantian-autonomous persons, with all the risks that entails.

What seems certain is that the rise of autonomous robots, if mishandled, will cause popular shock and cultural upheaval, especially if they are introduced suddenly and/or have some disastrous safety failures early on. That is all the more reason that a lengthy period of rigorous testing and gradual rollout (crawl-walk-run approach) is a moral minimum for the ethical deployment of autonomous robots, especially by the military. Further, this points to the early, prior need to identify the full range of possible ethical, technological, and societal issues in robot ethics—as we will discuss in the next section—in order to ensure that a technology risk assessment accounts for these concerns.

7. Robot Ethics: The Issues

From the preceding sections, it should be clear that there are myriad issues in risk and ethics related to autonomous military robotics. In this section, we will pull together and organize these various strands, as well as raise additional ones to provide a single, full view of the challenges facing the field.[6] These challenges are organized in thematic sub-groups: legal, just war, technical, robot-human, societal, and other and future challenges.

This is not meant to be an exhaustive list, as other issues certainly will emerge as the technology develops and field use broadens.[7] The value of this section again is to help anticipate the challenges facing the development and deployment of autonomous military robots, in order to proactively address them in both the design or application phases. Further, they may help to inform ethical and risk issues related to non-military robots, given the close historical relationship between defense technologies and consumer or public technologies, such as the evolution of ARPANET into the Internet.

7.1 Legal Challenges

1. *Unclear responsibility.* To whom would we assign blame—and punishment—for improper conduct and unauthorized harms caused by an autonomous robot (whether by error or intentional): the designers, robot manufacturer, procurement officer, robot controller/supervisor, field commander, President of the United States...or the robot itself? [Asaro, 2007; Sparrow, 2007; Sharkey, 2008a] We have started an inquiry into this critical issue in section 5: The law offers several precedents that a robotics case might follow, but given the range of specific circumstances that would influence a legal decision as well as evolving technology, more work will be needed to clarify the law for a clear framework in matters of responsibility.

[6] We thank and credit Ron Arkin for his discussions on many of these issues presented here.

[7] As an example of an unexpected policy change, when German forces during World War II recognized the impracticality of using naval submarines to rescue crews of sinking enemy ships—given limited space inside the submarine as well as exposure to radar and attacks when they surface—they issued the *Laconia* Order in 1942, based on military necessity, that released submarines from a long-standing moral obligation for sea vessels to rescue survivors; other nations soon followed suit to effectively eliminate the military convention altogether [Walzer, 1977, pp. 147-151].

In a military system, it may be possible to simply *stipulate* a chain of responsibility, e.g., the commanding officer is ultimately responsible. But this may oversimplify matters, e.g., inadequate testing allowed a design problem to slip by and caused the improper robotic behavior, in which case perhaps a procurement officer or the manufacturer ought to be responsible. The situation becomes much more complex and interesting with robots that have greater degrees of autonomy, which may make it appropriate to treat them as quasi-persons, if not full moral agents some point in the future. We note that Kurzweil forecasts that, by the year 2029, "[m]achines will claim to be conscious and these claims will be largely accepted" [Kurzweil, 1999].

2. *Refusing an order.* A conflict may arise in the following situation, among others: A commander orders a robot to attack a house that is known to harbor insurgents, but the robot—being equipped with sensors to 'see' through walls—detects many children inside and, given its programmed instruction (based on the ROE) to minimize civilian casualties, refuses the order. How ought the situation proceed: should we defer to the robot who may have better situational awareness, or the officer who (as far as she or he knows) issues a legitimate command? This dilemma also relates back to the question of responsibility: if the robot refuses an order, then who would be responsible for the events that ensue? Following legitimate orders is clearly an essential tenet for military organizations to function, but if we permit robots to refuse an order, this may expand the circumstances in which human soldiers may refuse an order as well (for better or worse).

3. *Consent by soldiers to risks.* In October 2007, a semi-autonomous robotic cannon deployed by the South African army malfunctioned, killing nine 'friendly' soldiers and wounding 14 others [Shachtman, 2007]. It would be naive to think such accidents will not happen again. In these cases, should soldiers be informed that an unusual or new risk exists, e.g., when they are handling or working with other dangerous items, such as explosives or even anthrax? Does consent to risk matter anyway, if soldiers generally lack the right to refuse a work order? We discussed the notion of consent and informed in the previous section.

7.2 Just-War Challenges

1. *Attack decisions.* It may be important for the above issue of responsibility to decide who, or what, makes the decision for a robot to strike. Some situations may develop so quickly and require such rapid information processing that we would want to entrust our robots and systems to make critical decisions. But the LOW and ROE generally demand there to be human 'eyes on target', either in-person or electronically and presumably in real time. (This is another reason why there is a general ban on landmines: without eyes on target, we do not know who is harmed by the ordnance and therefore have not fulfilled our responsibility to discriminate combatants

from non-combatants.) If human soldiers must monitor the actions of each robot as they occur, this may limit the effectiveness for which the robot was designed in the first place: robots may be deployed precisely because they can act more quickly, and with better information, than humans can.

However, some military robots—such as the Navy's Phalanx CIWS—seem to already and completely operate autonomously, i.e., they make attack decisions without human eyes on target or approval. This raises the question of how strictly we should take the 'eyes on target' requirement. One plausible argument for stretching that requirement is that the Phalanx CIWS operates as a last line of defense against imminent threats, e.g., incoming missiles in the dark of the night, so the benefits more clearly outweigh the risks in such a case. Another argument perhaps would be that 'eyes on target' need not be *human* eyes, whether directly or monitoring the images captured by a remote camera; that is, a human does not necessarily need to directly confirm a target or authorize a strike. A robot's target-identification module—assuming it has been sufficiently tested for accuracy—programmed by engineers is arguably a proxy for human eyes. At least this gives the system some reasonable ability to discriminate among targets, in contrast to a landmine, for instance. A requirement for 100% accuracy in target identification may be overly burdensome, since that is not a bar we can meet with human soldiers.

2. *Lower barriers for war.* As raised in section 4, does the use of advanced weaponry such as autonomous robotics make it easier for one nation to engage in war or adopt aggressive foreign (and domestic) policies that might provoke other nations? If so, is this a violation of *jus ad bellum*? [Asaro, 2008; Kahn, 2002] It may be true that new strategies, tactics, and technologies make armed conflict an easier path to choose for a nation, if they reduce risks to our side. Yet while it seems obvious that we should want to reduce US casualties, there is something sensible about the need for some terrible cost to war as a deterrent against entering war in the first place. This is the basis for just-war theory, that war ought to be the very last resort given its horrific costs [Walzer, 1977].

But the considered objection—that advanced robotics immorally lowers barriers for war—hides a logical implication that we should not do anything that makes armed conflict more palatable: we should not attempt to reduce friendly casualties, or improve battlefield medicine, or conduct any more research that would make victory more likely and quicker. Taken to the extreme, the objection seems to imply that we should *raise* barriers to war, to make fighting as brutal as possible (e.g., using primitive weapons without armor) so that we would never engage in it unless it were truly the last resort. Such a position appears counterintuitive at best and dangerously foolish at worst, particularly if we expect that other nations would not readily adopt a policy of relinquishment, which would put the US at a competitive disadvantage.

3. *Imprecision in LOW and ROE.* Asimov's Laws appear to be as simple as programmable rules can be for autonomous robots, yet they yielded surprising, unintended implications in his stories [e.g., Asimov, 1950]. Likewise, we may understand each rule of engagement and believe them to be sensible, but are they truly consistent with one another and sufficiently clear—which appears to be a requirement in order for them to be programmable? Much more complex than Asimov's Laws, the LOW and ROE leave much room for contradictory or vague imperatives, which may result in undesired and unexpected behavior in robots.

   For instance, the ROE to minimize collateral damage is vague: is the rule that we should not attack a position if civilian deaths are expected to be greater than—or even half of— combatant deaths? Are we permitted to kill one (high-ranking) combatant, even if it involves the death of five civilians—or $10M in unnecessary damage? A robot may need specific numbers to know exactly where this line is drawn, in order to comply with the ROE. Unfortunately, this is not an area that has been precisely quantified nor easily lends itself for such a determination.

7.3 Technical Challenges

1. *Discriminating among targets.* Some experts contend that it is simply too difficult to design a machine that can distinguish between a combatant and a non-combatant, particularly as insurgents pose as civilians, as required for the LOW and ROE [e.g., Sharkey, 2008a; Sparrow, 2007; Canning et al., 2004]. Further, robots would need to discriminate between active combatants and wounded ones who are unable to fight or have surrendered. Admittedly, this is a complex technical task, but we need to be clear on how accurate this discrimination needs to be. That is, discrimination among targets is also a difficult, error-prone task for human soldiers, so ought we hold machines to a higher standard than we have yet to achieve ourselves, at least in the near term?

   Consider the following: A robot enters a building known to harbor terrorists, but at the same time an innocent girl is running toward the robot (unintentionally) in chasing after a ball that happens to be rolling in the direction of the robot. Would the robot know to stand down and not attack the child? If the robot were to attack, of course that would cause outrage from opposing forces and even our own media and public; but this scenario could likely be the same as with a human soldier, adrenaline running high, who may misidentify the charging target as well. It seems that in such a situation, a robot may be less likely to attack the child, since the robot is not prone to overreact from the influence of emotions and fear, which afflict human soldiers. But in any event, if a robot would likely not perform worse than a human soldier, perhaps this is good enough for the moment until the technical ability to discriminate among targets improves. Some critics, however, may still insist on perfect discrimination or at least far better than humans are capable of, though it is unclear why we should hold robots to such a

high standard before such a technology exists (unless their point is to not use robots at all until we have perfected them, which is also a contentious position).

One proposed 'workaround' solution is to permit robots to target only weapons, including any hostile robots, rather than the human soldiers wielding such weapons [Canning, 2008]. Thus if an enemy combatant fails to relinquish his weapon in the presence of a robot, then he significantly increases his risk of being unintentionally harmed as the robot proceeds to disable the weapon. However, while this seems reasonable in principle, other experts continue to point to the technical challenge of discrimination: given current and foreseeable limitations in AI, a robot still may not be able to reliably target only a weapon and not the person, nor even reliably identify weapons from non-weapons, e.g., a child pointing her ice cream cone at an urban patrol robot [Sharkey, 2007b]. If this is true, then the considered solution merely postpones the discrimination problem, though it does create an extra layer of protection against inappropriate harm to humans; so the solution merits further consideration.

Another possible solution, which avoids the above programming issues, may be to simply operate combat robots only in regions of heavy fighting, teeming with valid targets [Sharkey, 2008b]. In these zones—sometimes called 'kill boxes' or 'engagement regions'—the Rules of Engagement are loosened, and non-combatants can be reasonably presumed to have fled, thus obviating the issue of discriminating among targets (and assuming none of our own troops is in the kill box or at least can be easily identified, e.g., by wearing some sensor). Using combat robots, at least initially, in only such zones might help to solve the first-generation problem described below, providing a training ground of sorts to test and perfect the machines. Or even in regions without heavy fighting but in need of tight security, e.g., guarding perimeters, armed sentry robots could operate in those designated zones, as long as it is clear to everyone that trespassers will be presumed to be enemy combatants and sufficient safeguards or deterrents to entry are in place to prevent, say, an accidental trespass by a child. (The risk of harm to a non-combatant here seems to be the same as with using guard dogs to protect property today.)

2. *First-generation problem.* We previously mentioned that it would be naive to believe that another accident with military robots will not happen again. As with any other technologies, errors or bugs will inevitably exist, which can be corrected in the next generation of the technology. With Internet technologies, for instance, first-generation mistakes are not too serious and can be fixed with software patches or updates. But with military robotics, the stakes are much higher, since human lives may be lost as a result of programming or other errors. So it seems that the prudent or morally correct course of action is to rigorously test the robot before deploying it, as discussed in section 6.

   However, testing already occurs with today's robots, yet it is still difficult if not impossible to certify any given robot as error-free, given that (a) testing environments may be substantially

different than more complex, unstructured, and dynamic battlefield conditions in which we cannot anticipate all possible contingencies; and (b) the computer program used in the robot's on-board computer (its 'brain') may consist of millions of lines of code.

Beta-testing of a program (testing prior to the official product launch, whether related to robotics, business applications, etc.) is conducted today, yet new errors are routinely found in software by actual users even after its official product launch. It is simply not possible to run a complex piece of software through all possible uses in a testing phase; surprises may occur during its actual use. Likewise, it is not reasonable to expect that testing of robots will catch any and all flaws; the robots may behave in unexpected and unintended ways during actual field use. Again, the stakes are high with deploying robots, since any error could be fatal. This makes the first-generation problem, as well as ongoing safety and dependability, an especially sensitive issue [e.g., Van der Loos, 2007].

3. *Robots running amok.* As depicted in science-fiction novels and movies, some imagine the possibility that robots might break free from their human programming through methods as: their own learning, or creating other robots without such constraints (self-replicating and self-revising), or malfunction, or programming error, or even intentional hacking [e.g., Joy, 2000]. In these scenarios, because robots are built to be durable and even with attack capabilities, they would be extremely difficult to defeat—which is the point of using robots as force multipliers. Some of these scenarios are more likely than others: we wouldn't see the ability of robots to fully manufacture other robots or to radically evolve their intelligence and escape any programmed morality for quite some time. But other scenarios, such as hacking, seem to be near-term possibilities, especially if robots are not given strong self-defense capabilities (see below).

   That robots might run amok is an enhanced version of the worry that enemies might use our own creations against us, but it also introduces a new element in that previous weapon systems still need a human operator which is a point of vulnerability, i.e., a 'soft underbelly' of the system. Autonomous robots would be designed to operate without human control. What precautions can be taken to prevent one from being captured and reverse-engineered or reprogrammed to attack our own forces? If we design a 'kill switch' that can automatically shut off a robot, this may present a key vulnerability that can be exploited by the enemy.

4. *Unauthorized overrides.* This concern is similar to that with nuclear weapons: that a rogue officer may be enough to take control of these terrible weapons and unleash them without authorization or otherwise override their programming to commit some unlawful action. This is a persistent worry with any new, devastating technology and is a multi-faceted challenge: it is a human problem (to develop ethical, competent officers), an organizational problem (to provide procedural safeguards), and technical problem (to provide systemic safeguards). So there does

not yet appear to be anything unique about this worry that should hinder the development or deployment of advanced robotics, to the extent that the concern does not impact the development of other technologies. But it nevertheless is a concern that needs to be considered in the design and deployment phases.

5. *Competing ethical frameworks*. If we seek to build an ethical framework for action in robots, it is not clear which ethical theory to use as our model [e.g., Anderson and Anderson, 2007]. In section 3, we have argued for a hybrid approach related to virtue ethics, as the theory that seems to lead to the fewest unintuitive results, but any sophisticated theory seems to be vulnerable to inconsistencies and competing directives (especially if a three- or four-rule system as simple as Asimov's cannot work perfectly). This concern is related to the first technical challenge described here, that it is too difficult to embed these behavioral rules or programming into a machine. But we should recall our stated mission here: our initial goal ought *not* be to create a perfectly ethical robot, only one that acts more ethically than humans—and sadly this may be a low hurdle to clear.

6. *Coordinated attacks*. Generally, it is better to have more data than less when making decisions, particularly one as weighty as a military strike decision. Robots can be designed to easily network with other robots and systems; but this may complicate matters for robot engineers as well as commanders. We may need to establish a chain of command within robots when they operate as a team, as well as ensure coordination of their actions. The risk here is that as complexity of any system increases, the more opportunities exist for errors to be introduced, and again mistakes by military robots may be fatal.

7.4 Human-Robot Challenges

1. *Effect on squad cohesion*. As a 'band of brothers', there understandably needs to be strong trust and support among soldiers, just as there is among police officers, firefighters, and so on. But sometimes this sense of camaraderie can be overdeveloped to the extent that one team member becomes complicit in or deliberately assists in covering up an illegal or inappropriate action of another team member. We have discussed the benefits of military robots with respect to behavior that is more ethical than currently exhibited by human soldiers. But robots will also likely be equipped with video cameras and other such sensors to record and report actions on the battlefield. This could negatively impact the cohesion among team or squad members by eroding trust with the robot as well as among fellow soldiers who then may or may not support each other as much anymore, knowing that they are being watched. Of course, soldiers and other professionals should not be giving each other unlawful 'support' anyway; but there may be situations in which a soldier is unclear about or unaware of motivations, orders, or other

relevant details and err on the side of caution, i.e., not providing support even when it is justified and needed.

2. *Self-defense.* Asimov's Laws permitted robots to defend themselves where that action did not conflict with higher duties, i.e., harm humans (or humanity) or conflict with a human-issued order. But Arkin suggests that military robots can be more conservative in their actions, i.e., hold their fire, because they do not have a natural instinct of self-preservation and may be programmed without such [Arkin, 2007]. But how practical is it, at least economically speaking, to not give robots—which may range from $100,000 to millions of dollars in cost—the ability to defend itself? If a person, say, a US civilian, threatens to destroy a robot, shouldn't it have the ability to protect itself, our very expensive taxpayer-funded investment?

   Further, self-defense capabilities may be important for the robot to elude capture and hacking, as previously discussed. Robots may be easily trapped and recovered fully intact, unlike tanks and aircraft, for instance, which usually sustain much if not total damage in order to capture it. These considerations are in tension with using robots for a more ethical prosecution of war, since a predilection to hold their fire would be a major safeguard against accidental fatalities, e.g., mistakenly opening fire on non-combatants; therefore, a tradeoff or compromise among these goals—to have a more ethical robot and to protect the robot from damage and capture—may be needed.

3. *Winning hearts and minds.* Just-war theory, specifically *jus post bellum*, requires that we fight a war in such a manner that it leaves the door open for lasting peace after the conflict [Orend, 2002]. That is, as history has shown, we should not brutalize an enemy, since that would leave ill-feelings to linger even after the fighting has stopped, which makes peaceful reconciliation most difficult to achieve. Robots do not necessarily represent an immoral or overly brutal way of waging war, but as they are needed for urban operations, such as patrolling dangerous streets to enforcing a curfew or securing an area, the local population may be less likely to trust and build good-will relationships with the occupying force [Sharkey, 2008a]. Winning hearts and minds is likely to require diplomacy and human relationships that machines would not be capable of delivering at the present time, as we previously discussed in section 4.

4. *'Comfort' robots.* Ethicists are already talking about the impact of robots as lovers or surrogate relationship partners [Levy, 2007]. This does not seem so unthinkable, considering that some people already have 'relationships' with increasingly-realistic sex dolls, so robotics appear to be a natural next step in that industry; indeed, people today engage in sexual activities online, i.e., without a partner physically present.

   In previous wars, women have been taken by the military to provide 'comfort' to soldiers or, in other words, forced into sexual slavery or prostitution. In World War II, women were most

infamously used by the Japanese Imperial Army to satiate the pent-up carnal desires of its soldiers, ostensibly to prevent possible riots and discontent among the ranks; Nazi Germany reportedly also used women to stock their 'joy divisions' at labor or concentration camps. And instances of rape have been reported—and continue today—in armed conflicts from Africa to the Americas to Asia.

Robots, then, may be able to serve the same function of providing 'comfort' to the troops in a much more humane way, i.e., without the exploitation of women and prisoners of war. However, it is unclear that this function is truly needed (to the extent that most militaries today do not employ military prostitutes and seem to be operating adequately) or can overcome existing public inhibitions or attitudes on what is mostly a taboo subject of both sex in the military and sex with non-human objects.

7.5 Societal Challenges

1. *Counter-tactics in asymmetric war.* As discussed in the previous issue of lowering barriers to war or making war more risk-free, robots would help make US military actions more effective and efficient, which is exactly the point of deploying those machines. Presumably, the more autonomous a robot is, the more lethal it can be (given requirements to discriminate among targets and so on). This translates to quicker, more decisive victories for us; but for the other side, this means swifter and perhaps more demoralizing defeats. We can reasonably expect that a consequence of increasing the asymmetry of warfare in our favor will cause opposing forces to engage in even more unconventional strategies and tactics, beyond 'terrorist' acts today as necessitated by an overwhelming superiority of US troop numbers and technologies [e.g., Kahn, 2002]; few nations could hope to successfully wage war with the US by using the same methods we use.

   This not only involves how wars and conflicts are fought, but also exposes our military as well as public to new forms of attack which may radically change our society, as the events of 9/11 have already. For instance, more desperate enemies may resort to more desperate measures, from intensifying efforts to acquire nuclear or biochemical weapons to devising a 'scorched earth' or 'poison pill' strategy that strikes deeply at us but at some great cost to their own forces or population (a Pyrrhic victory).

2. *Proliferation.* Related to the previous issue, history also shows that innovations in military technologies—from armor and crossbows to intercontinental missiles and 'smart' bombs—give the inventing side a temporary advantage that is eroded over time by other nations working to replicate the technologies. Granting that modern technologies are more difficult to reverse-engineer or replicate than previous ones, it nevertheless seems inevitable or at least possible

that they can be duplicated, especially if an intact sample can be captured, such as immobilizing a ground robot as opposed to shooting down an unmanned aerial vehicle. So with the development of autonomous military robots, we can expect their proliferation with other nations at some future point. This means that these robots—which we are currently touting as lethal, difficult-to-neutralize machines—may be turned against our own forces eventually.

The proliferation of weapons, unfortunately, is an extremely difficult cycle to break: many nations are working to develop autonomous robotics, so a unilateral ban on their development would not accomplish much except to handicap that nation relative to the rest of the world. So the rush to develop this and other emerging technologies is understandable and irresistible, at least in today's world. One possible defense for our pursuit, apart from self-interested reasons, is that we (the US) want to ensure we develop these commanding technologies first, after which we would have more leverage to stop the proliferation of the same; further, because we occupy the higher moral ground, it would be most responsible for the US to develop the technologies first.

The problem, of course, is that every nation thinks of itself as moral or 'doing the right thing', so it would be difficult to objectively assign a moral imperative to any given nation, including the US. Solving this problem, then, would seem to require additional legal and ethical theorizing, likely resulting in new international treaties and amendments to the Laws of War.

3. *Space race.* As on earth, autonomous robots may hold many benefits for space exploration [Jónsson et al., 2007]. Proliferation also has significant financial and environmental costs, particularly if military robotics technology is developed for outer space. First, launch costs are still astronomical, costing thousands of dollars *per pound* to put an object into low Earth orbit, and several times more per pound for geostationary orbit (not to mention periodic replacement costs and in-orbit repairs). An unlikely 'star wars' scenario aside—which would create countless pieces of space debris that would need to be tracked and threaten communications satellites and so on—even using robots for research purposes, e.g., to explore and develop moons and other planets, may spark another space race given the military advantages of securing the ultimate high ground. This not only opens up outer space for militarization, which the world's nations have largely resisted, but diverts limited resources that could make more valuable contributions elsewhere.

4. *Technology dependency.* The possibility that we might become dependent or addicted to our technologies has been raised throughout the history of technology and even with respect to robotics. Today, ethicists worry that we may become so reliant on, for instance, robots for difficult surgery that humans will start losing that life-saving skill and knowledge; or that we become so reliant on robots for basic, arduous labor that our economy is somehow impacted

and we forget some of those techniques (Veruggio, 2007). In the military, some soldiers already report being attached to the robot that saved their lives [Garreau, 2007].

As a general objection to technology, this concern does not seem to have much force, since the benefits of the technology in question often outweigh any losses. For instance, our ability to perform mathematical calculations may have suffered somewhat given the inventions of the calculator and spreadsheets, but we would rather keep those tools even at that expense. Certainly, it is a possible hypothetical or future scenario that, after relying on robots to perform all our critical surgeries, some event—say, a terrorist attack or massive electromagnetic pulse—could interrupt an area's power supply, disabling our machines and leaving no one to perform the surgery (because we forgot how and have not trained surgeons on those procedures, since robots were able to do it better). But as abilities enhanced by technology, such as performing numeric calculations, have not entirely disappeared from a population or even to a life-impacting degree in individuals, it is unclear why we would expect something as artful as brain or heart surgery to be largely lost. Similarly, in the case of relying on robots for manual labor, technology dependency would not erase our ability to, say, dig holes to plant trees to any impacting degree.

5. *Civil security and privacy*. Defense technologies often turn into public or consumer technologies, as we previously pointed out. So a natural step in the evolution of military robots would seem to be their incarnation as civil security robots; they might guard corporate buildings, control crowds, and even chase down criminals. Many of the same concerns discussed above—such as technical challenges and questions of responsibility—would also become larger societal concerns: if a robot unintentionally (meaning that no human intentionally programmed it to ever do so) kills a small child, whether by accident (run over) or mistake (identification error), it will likely have greater repercussions than a robot that unintentionally kills a non-combatant in some faraway conflict. Therefore, there is increased urgency to address these military issues that may spill over into the public domain.

   And while we take it that soldiers, as government property, have significantly decreased privacy expectations and rights, the same is not true of the public at large. If and when robots are used more in society, and the robots are likely to be networked, concerns about illegal monitoring and surveillance—privacy violations—may again surface, as they have with most other modern technologies, from DNA testing to genome sequencing to communications-monitoring software to nanotechnology. This raises the question of what kind of consent we need from the public before deploying these technologies in society.

7.6 Other and Future Challenges

1. *Co-opting of ethics effort by military for justification.* A possible challenge that does not fit neatly into any of the above categories is the following political concern. Defense organizations may be aware (now) of the above concerns, but they may still not choose to address the issues to mitigate risk by absolving themselves of this responsibility: they may simply point to ethicists and robot scientists working on related issues as justification for proceeding ahead without any real plan to address at least some of these risks [Sharkey, 2007b].

   This is an interesting *meta*-issue for robot ethics, i.e., it is about the study and aims of robot ethics and not so much about an issue directly related to the use of autonomous robots. While it is certainly a possibility that organizations may only pay 'lip-service' to the project of robot ethics to appease critics and watchdogs, it does not take much enlightenment or foresight to see actual, real-world benefits from earnestly addressing these challenges. Further, we might measure the commitment that organizations have to robot ethics by the funding levels for such research. And it would be readily apparent if, for instance, defense organizations ignored the counsel and recommendations of experts engaged in the field. This is to say that co-opting is a relatively transparent activity to identify, although the point is more that it could be too late (for those harmed or society in general) by then.

2. *Robot rights.* For now, robots are seen as merely a tool that humans use, morally no different (except in financial value) than a hammer or a rifle---their only value is instrumental, as a means to our ends. But as robots begin to assume aspects of human decision-making capabilities, the question may arise of their *intrinsic value*: do they deserve moral consideration of their own (beyond their financial or tactical value), and at what point in their evolution will they achieve this intrinsic value (as human lives seem to have)? When they become Kantian autonomous agents, making their own goals for themselves? Or would intrinsic value also require consciousness and emotions?

   Some technologists have suggested that, by 2029, robots will demand equal treatment before the law with humans—and believe that this demand will be granted [e.g., Kurzweil, 1999]. The only guarantee of avoiding this outcome appears to be a prohibition on programming robots with anything other than a 'slave morality', i.e., simply not allowing a Kantian-autonomous robot to ever be programmed or built (though such bans, especially when applied internationally, have been notoriously difficult to enforce). It will require careful consideration in the future as to whether such a prohibition should ever be lifted. Fortunately, even 'technological optimists', such as Kurzweil, do not expect this to be an issue until at least the 2020s.

   Thus far, we have not discussed the possibility of giving rights to robots, not so much that it is farfetched to do so (e.g., we give rights to non-living entities such as corporations) or to consider them as persons (philosophically-speaking; e.g., again corporations or ships or some animals such as dolphins), but that the prerequisites for rights seem to require advanced software or

artificial intelligence that is not quite within our foreseeable grasp. Specifically, if our notion of personhood specifies that only persons can be afforded rights and that persons must have free will or the capacity for free will, then it is unclear whether we will ever develop technologies capable of giving free will or full autonomy to machines, and, indeed, we don't even know whether any other *biological* species will ever have or is now capable of such full autonomy; thus, we do not want to dwell on such a speculative issue here. That said, we will leave open the possibility that we may someday want or be logically required to give rights to robots [e.g., Kurzweil, 1999], but much more investigation is needed on the issue.

3. *The precautionary principle*. Given the above laundry list of concerns, some may advocate following a precautionary principle in robotics research—to slow or halt work until we have mitigated or addressed possible catastrophic risks—as critics have done for other technologies, such as bio- and nanotechnologies. For instance, those fearful of 'Terminator' scenarios where machines turn against us lesser humans, current research in autonomous robotics may represent a path towards possible, perhaps likely, disaster; thus a cautious, prudent approach would be to ban or at least significantly slow down research until we can sufficiently think about these issues before technology overtakes ethics. While we believe that a precautionary principle may be the appropriate course of action for some technology cases, many of the issues discussed above do not appear imminent enough to warrant a research moratorium or delay, just more investigation which may be sufficiently conducted in parallel to efforts to develop advanced robotics.

   Furthermore, a cautionary approach in the development of advanced systems is inherently in tension with both the approaches taken by the scientists and engineers developing robots and with the outlook of military planners, rapidly searching for more effective tools for the task of waging war. We will again leave open the possibility that someday we may have to seriously consider the role of the precautionary principle in robotics, but that day appears to be in the distant horizon and does not demand an extensive discussion here.

7.7 Further and Related Investigations Needed

Again, we do not intend the above to capture all possible issues related to autonomous military robotics. Certainly, new issues will emerge depending on how the technology and intended uses develop. In preceding sections, we have started to address what we seemed to be the most urgent and important issues to resolve first, especially as related to responsibility, risk, and the ability of robots to discriminate among targets. This is only the beginning of a dialogue in robot ethics and merits further investigations.

Moreover, our discussion here may be helpful in informing ethics research related to non-military robots, such as security, labor, and sex robots previously mentioned in the public domain. For instance, robots are already being used to care for the elderly, but are we merely pawning off our obligations to care for the elderly to machines who may be unable to provide the emotional content that seems to be needed in human relationships (even though there are significant advances in enabling robots to display 'emotions')? What are the benefits and risks of using robots as teachers, domestic help, or even as researchers, e.g., exploring difficult and alien environments? Do robotic planes, trains, and automobiles pose any special issues? These and other questions will need to be addressed; and as is often the case with public technologies, they have their roots in military innovations, so we may have a separate but related responsibility to begin looking ahead to these non-military questions as well.

8. Conclusions

> "We can only see a short distance ahead, but we can see plenty there that needs to be done."—Alan Turing [1950, p. 460]

There are many paths one may take in examining issues of risk and ethics arising from advanced military robotics. We initiate several lines of inquiry in this preliminary report, as follows.

In section 1, we begin by building the case for 'robot ethics.' While there are substantial benefits to be gained from the use of military robots, there are also many opportunities for these machines to act inappropriately, especially as they are given greater degrees of autonomy (for quicker, more efficient, and more accurate decision-making, and if they are to truly replace human soldiers). The need for robot ethics becomes more urgent when we consider pressures driving the market for military robotics as well as long-standing public skepticism that lives in popular culture.

In section 2, we lay the foundation for a robot-ethics investigation by presenting a wide range of military robots—ground, aerial, and marine—currently in use and predicted for the future. While most of these robots today are semi-autonomous (e.g., the US Air Force's Predator), some apparently-fully autonomous systems are emerging (e.g., the US Navy's Phalanx CIWS) though used in a very limited, last-resort context. From this, we can already see ethical questions emerge, especially related to the ability to discriminate combatants from non-combatants and the circumstances under which robots can make attack decisions on their own. We return to these questions in subsequent sections, particularly section 7.

In section 3, we look at behavioral frameworks that might ensure ethical actions in robots. It is natural to consider various programming approaches, since robots are related to our personal and business computer systems today that also depend on programmed instructions. We also recognize the forward-looking nature of our discussions here, given that the more-sophisticated programming abilities needed to build truly autonomous robotics are still under development.

We first discuss the traditional approach of *top-down programming*, i.e., establishing general rules that the robot would follow. A clear example is a deontological approach, such as using Kant's Categorical Imperative or Asimov's Laws of Robotics. However, a rigid set of rules is likely not robust enough to arrive at the correct action or decision in enough cases, particularly in unforeseen and complex scenarios. This suggests that we also need to attend to the 'rightness' of the result itself,

not just to the rules. But even if we acknowledge that consequences matter, there are other challenges raised by adopting a consequentialist/utilitarian approach, such as the impracticality of calculating and weighing all possible results, both near and far term, and the (strong) possibility of countenancing some intuitively-wrong action.

Given the apparent limitations of top-down programming, we then examine *bottom-up approaches*, inspired by biological evolution and human development. However, a key challenge is that bottom-up systems work best when they are directed at achieving one clear goal, but military robots often operate in dynamic environments in which available information is confusing or incomplete. That is, even if moral calculation is not an issue, there still remains the large problem of moral psychology, i.e., how to develop robots that embody the right tendencies in their reactions to the world and other agents in that world, particularly when the robots are confronted with a novel situation in which they cannot rely on experience.

Moral reasoning by humans, however, is not limited to exclusively a top-down or bottom-up approach; rather, we often use both strategies of rule-following and experience. (Nonetheless, it is useful to evaluate both programming approaches separately to identify their benefits and challenges.) Therefore, we consider a *hybrid approac*h of virtue ethics for constructing ethical autonomous robots. This approach is concerned with the development of moral character; in the military case, with promoting the ideal character traits of a warfighter, i.e., a 'warrior code of ethics' as its virtues.

In section 4, we look at considerations in programming the Laws of War (LOW) and Rules of Engagement (ROE), which may differ from mission to mission, into a robot. No matter which programming approach is adopted, we at least would want to the robot to obey the LOW and ROE, and this might serve as a proxy for full-fledged morality until we have the capability to program a robot with the latter. Such an approach has several advantages, including: (1) any problems from moral relativism/particularism or other problems with general ethical principles are avoided; and (2) the relationship of morality to legality—a minefield for ethics—is likewise largely avoided, since the LOW and ROE make clear what actions are legal and illegal for robots, which serves as a reasonable approximation to the moral-immoral distinction.

Our discussion of the LOW and ROE, then, delves into their underlying foundation in just-war theory, particularly *jus ad bellum* (moral justification for entering war) and *jus in bello* (just and unjust actions in the prosecution of a war). We also examine ethical challenges to just-war theory as related to military robotics: Some have objected to the use of military robotics on the grounds that it makes easier the decision to enter war, in apparent violation of *jus ad bellum*; and we again see that the technical ability to properly discriminate against targets, as required by *jus in bello*, is a concern.

In section 5, we attend to the recurring possibility of accidental or unauthorized harm caused by robots; who would be responsible ultimately for those mishaps? We look at the issue through the lens of legal liability, both when robots are considered as merely products and when, as they are given more autonomy, they might be treated as legal agents, e.g., as legal quasi-persons such as children are regarded by the law. In the latter case, it is not clear how we would punish robots for their inappropriate actions.

In section 6, still attending to the possibility of unintended or unforeseen harm committed by a robot, we broaden our discussion by looking at how we might think about general risks posed by the machines and their acceptability. We offer a preliminary framework for a technology risk assessment, which includes the key factors of consent, informed consent, affected population, seriousness, and probability. This assessment highlights further the need for a lengthy period of rigorous testing and gradual rollout (crawl-walk-run approach) as a moral minimum for the responsible deployment of autonomous robots, especially by the military.

Finally, in section 7, we bring together a full range of issues raised throughout our examination, as well as some new issues, that must be recognized in any comprehensive assessment of risks from military robotics. These challenges fall into categories related to law, just-war theory, technical capabilities, human-robot interactions, general society, and other and future issues. For instance, we discuss such issues as:

- If a military robot refuses an order, e.g., if it has better situational awareness, then who would be responsible for its subsequent actions?
- How stringent should we take the generally-accepted 'eyes on target' requirement, i.e., under what circumstances might we allow robots to make attack decisions on their own?
- What precautions ought to be taken to prevent robots from running amok or turning against our own side, whether through malfunction, programming error, or capture and hacking?
- To the extent that military robots can help reduce instances of war crimes, what is the harm that may arise if the robots also unintentionally erode squad cohesion given their role as an 'outside' observer?
- Should robots be programmed to defend themselves—contrary to Arkin's position—given that they represent costly assets?
- Would using robots be counterproductive to winning the hearts and minds of occupied populations or result in more desperate terrorist-tactics given an increasing asymmetry in warfare?

From the preceding investigation, we can draw some general and preliminary conclusions, including some future work needed:

1. Creating autonomous military robots that can act *at least as* ethically as human soldiers appears to be a sensible goal, at least for the foreseeable future and in contrast to a greater demand of a perfectly-ethical robot. However, there are still daunting challenges in meeting even this relatively-low standard, such as the key difficulty of programming a robot to reliably distinguish enemy combatants from non-combatants, as required by the Laws of War and most Rules of Engagement.

2. While a faster introduction of robots in military affairs may save more lives of human soldiers and reduce war crimes committed, we must be careful to not unduly rush the process. Much different than rushing technology products to commercial markets, design and programming bugs in military robotics would likely have serious, fatal consequences. Therefore, a rigorous testing phase of robots is critical, as well as a thorough study of related policy issues, e.g., how the US Federal Aviation Administration (FAA) handles UAVs flying in our domestic National Airspace System (which we have not addressed here).

3. Understandably, much ongoing work in military robotics is likely shrouded in secrecy; but a balance between national security and public disclosure needs to be maintained in order to help accurately anticipate and address issues of risk or other societal concerns. For instance, there is little information on US military plans to deploy robots in space, yet this seems to be a highly strategic area in which robots can lend tremendous value; however, there are important environmental and political sensitivities that would surround such a program.

4. Serious conceptual challenges exist with the two primary programming approaches today: top-down (e.g., rule-following) and bottom-up (e.g., machine learning). Thus a hybrid approach should be considered in creating a behavioral framework. To this end, we need to a clear understanding of what a 'warrior code of ethics' might entail, if we take a virtue-ethics approach in programming.

5. In the meantime, as we wait for technology to sufficiently advance in order to create a workable behavioral framework, it may be an acceptable proxy to program robots to comply with the Law s of War and appropriate Rules of Engagement. However, this too is much easier said than done, and at least the technical challenge of proper discrimination would persist and require resolution.

6. Given technical limitations, such as programming a robot with the ability to sufficiently discriminate against valid and invalid targets, we expect that accidents will continue to occur, which raise the question of legal responsibility. More work needs to be done to clarify the chain of responsibility in both military and civilian contexts. Product liability laws are informative but untested as they relate to robotics with any significant degree of autonomy.

7. Assessing technological risks, whether through the basic framework we offer in section 6 or some other framework, depend on identifying potential issues in risk and ethics. These issues vary from: foundational questions of whether autonomous robotics can be legally and morally deployed in the first place, to theoretical questions about adopting precautionary approaches, to forward-looking questions about giving rights to truly autonomous robots. These discussions need to be more fully developed and expanded.

8. Specifically, the challenge of creating a robot that can properly discriminate among targets is one of the most urgent, particularly if one believes that the (increased) deployment of war robots is inevitable. While this is a technical challenge and resolvable depending on advances in programming and AI, there are some workaround policy solutions that can be anticipated and further explored, such as: limiting deployment of lethal robots to only inside a 'kill box'; or designing a robot to target only other machines or weapons; or not giving robots a self-defense mechanism so that they may act more conservatively to prevent; or even creating robots with only non-lethal or less-than-lethal strike capabilities, at least initially until they are proven to be reliable.

These and other considerations warrant further, more detailed investigations in military robotics and issues of design, risk, and ethics. Such interdisciplinary investigations will require collaboration among policymakers and analysts, roboticists, ethicists, sociologists, psychologists, and others, internationally and including the general public as a key stakeholder. And this work has the potential to be as broad as other fields in science and society, such as bioethics or computer ethics.

The use of military robots represents a new era in warfare, perhaps more so than crossbows, airplanes, nuclear weapons, and other innovations have previously. Robots are not merely another asset in the military toolbox, but they are meant to also replace human soldiers, especially in 'dull, dirty, and dangerous' jobs. As such, they raise novel ethical and social questions that we should confront as far in advance as possible—particularly before irrational public fears or accidents arising from military robotics derail research progress and national security interests.

9. References

Allen, Colin, Varner, Gary, and Zinser, Jason (2000). "Prolegomena to Any Future Artificial Moral Agent", *Journol of Experimentol ond Theoreticol Artificial Intelligence* 12.3:251–261.

Anderson, Michael, and Anderson, Susan Leigh (2007). "Machine Ethics: Creating an Ethical Intelligent Agent", *Al Mogozine* 28.4: 15–26.

Arkin, Ronald C. (1998). *Behavior-Bosed Robotics,* Cambridge: MIT Press.

Arkin, Ronald C. (2007). *Governing Lethol Behovior: Embedding Ethics in* o *Hybrid Deliberotive/Hybrid Robot Architecture*, Report GIT-GVU-07-11, Atlanta, GA: Georgia Institute of Technology's GVU Center. Last accessed on September 15, 2008: http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf

Asaro, Peter (2007). "Robots and Responsibility from a Legal Perspective", Proceedings of the IEEE 2007 International Conference on Robotics and Automation, Workshop on RoboEthics, April 14, 2007, Rome, Italy. Last accessed on September 15, 2008: http://www.peterasaro.org/writing/ASARO%20Legal%20Perspective.pdf

Asaro, Peter (2008). "How Just Could a Robot War Be?" in Adam Briggle, Katinka Waelbers, and Philip Brey (eds.) *Current Issues in Computing ond Philosophy*, pp. 50-64, Amsterdam, The Netherlands: IOS Press.

Asimov, Isaac (1950). *I, Robot* (2004 edition), New York, NY: Bantam Dell.

Asimov, Isaac (1957). *The Noked Sun*, New York, NY: Doubleday.

Asimov, Isaac (1985). *Robots ond Empire*, New York, NY: Doubleday.

BBC (2005). "SLA Confirm Spy Plane Crash", *BBC.com*, October 19, 2005. Last accessed on September 15, 2008: http://www.bbc.co.uk/sinhala/news/story/2005/10/051019_uav_vavunia.shtml

BBC (2007). "Robotic Age Poses Ethical Dilemma", *BBC.com*, March 7, 2007. Last accessed on September 15, 2008: http://news.bbc.co.uk/2/hi/technology/6425927.stm

Bekey, George (2005). *Autonomous Robots: From Biologicol Inspirotion to Implementotion ond Control*, Cambridge, MA: MIT Press.

Brooks, Rodney (2002). *Flesh ond Mochines*. New York: Pantheon Books.

Canning, John, Riggs, G.W., Holland, O. Thomas, Blakelock, Carolyn (2004). "A Concept for the Operation of Armed Autonomous Systems on the Battlefield", *Proceedings of Associotion for Unmanned Vehicle Systems Internotionol's (AUVSI) Unmonned Systems North Americo*, August 3-5, 2004, Anaheim, CA.

Canning, John (2008). "Weaponized Unmanned Systems: A Transformational Warfighting Opportunity, Government Roles in Making it Happen", *Proceedings of Engineering the Totol Ship (ETS)*, September 23-25. 2008, Falls Church, VA.

Čapek, Karel (1921). *R.U.R.* (2004 edition, trans. Claudia Novack), New York, NY: Penguin Group.

CBS (2007). "Robots Playing Larger Role in Iraq War", October 21, 2007 news report. Last accessed on September 15, 2008: http://cbs3.com/topstories/robots.iraq.army.2.410518.html

Churchland, Paul (1995). *The Engine of Reoson, the Seot of the Soul: A Philosophicol Journey into the Broin*, Cambridge, MA: MIT Press.

Clarke, Roger (1994). "Asimov's Laws of Robotics: Implications for Information Technology," *IEEE Computer* (part 1: December 1993, pp. 53–61; part 2: January 1994, pp. 57–66).

Coffee, Jr., John C. (1981). "'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment," *Michigon Low Review*, Vol. 79, No. 3, pp. 386-459.

Computer Professionals for Social Responsibility (2008). "Technology in Wartime" conference, January 26, 2008, Stanford, CA. Last accessed on September 15, 2008: http://technologyinwartime.org/

Davis, Burke (1980). *Shermon's Morch: The First Full-Length Norrotive of Generol Williom T. Shermon's Devostoting Morch through Georgio ond the Corolinos*, New York, NY: Random House.

DeMoss, David (1998). "Aristotle, Connectionism, and the Morally Excellent Brain", *The Proceedings of the Twentieth World Congress of Philosophy*, August 10-15, 1998, Boston, MA.

DesJardins, Joseph (2003). *An Introduction to Business Ethics*, pp. 99-103, Columbus, OH: McGraw-Hill.

Dilov, Lyuben (1974). *The Woy of Icorus*. Дилов, Любен. Пътят на Икар. Захари Стоянов. ISBN 954-739-338-3.

Fikes, Richard and Nilsson, Nils (1971). "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving", *Artificiol Intelligence* 2(3-4): 189-208.

Foot, Phillipa (1972). "Morality as a System of Hypothetical Imperatives", *The Philosophicol Review* 81.3: 305-316.

Garreau, Joel (2007). "Bots on the Ground", *Washington Post*, May 6, 2007. Last accessed on September 15, 2008: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009_pf.html

Harrison, Harry (1989). :The Fourth Law of Robotics", in Isaac Asimov and Martin Harry Greenberg (eds.) *Foundotion's Friends: Stories in Honor of Isaac Asimov*, New York, NY: Tor Books.

Hemingway, Ernest (1935). "Notes on the Next War: A Serious Topical Letter", *Esquire*, Vol. 4, No. 3: 19, 156.

Hew, Patrick (2007). "Autonomous Situation Awareness: Implications for Future Warfighting", *Austrolion Defence Force Journol* 174: 71-87. Last accessed on September 15, 2008: http://www.defence.gov.au/publications/dfj/index.htm

Hobbes, Thomas (1651). *Leviothon* (1982 edition), New York, NY: Penguin Group.

Institute of Electrical and Electronics Engineers (2008). "International Conference on Advanced Robotics and its Social Impact" conference, August 23-25, 2008, Taipei, Taiwan. Last accessed on September 15, 2008: http://arso2008.ntu.edu.tw/

International Federation of Robotics (2008). "International Robot Standards" page from International Federation of Robotics website. Last accessed on September 15, 2008: http://www.ifr.org/modules.php?name=News&file=article&sid=20

Iraq Coalition Casualty Count (2008). "Deaths Caused by IEDs" and "U.S. Deaths by Month" webpages. Last accessed on September 15, 2008: http://icasualties.org/oif/IED.aspx and http://icasualties.org/oif/USDeathByMonth.aspx

Johnson, Robert (2008). "Kant's Moral Philosophy", The Stanford Encyclopedia of Philosophy, Fall 2008 Edition. Last accessed on September 15, 2008: http://plato.stanford.edu/archives/fall2008/entries/kant-moral/

Jónsson, Ari, Morris, Robert, and Pedersen, Liam (2007). "Autonomy in Space: Current Capabilities and Future Challenges", *AI Magazine* 28.4: 27-42.

Joy, Bill (2000). "Why the Future Doesn't Need Us", *Wired* 8.04: 238-262.

Kahn, Paul (2002). "The Paradox of Riskless War," *Philosophy & Public Policy Quorterly*, Vol. 22: 2–8.

Kant, Immanuel (1785). *Grounding for the Metophysics of Morols* (1993 edition, translated by James W. Ellington), Indianapolis, IN: Hackett Publishing Co.

Kurzweil, Ray (1999). *The Age of Spirituol Mochines: When Computers Exceed Humon Intelligence*, New York, NY: Viking Penguin.

Kurzweil, Ray (2005). *The Singulority is Neor: When Humons Tronscend Biology*, New York, NY: Viking Penguin.

Lee, Steven (2004). "Double Effect, Double Intention, and Asymmetric Warfare", *Journol of Militory Ethics*, 3.3: 233-251.

Levy, David (2007). *Love and Sex with Robots: The Evolution of Human-Robot Relotionships*, New York, NY: HarperCollins Publishers.

McIntyre, Alison (2004). "Doctrine of Double Effect", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition). Last accessed on September 15, 2008: http://plato.stanford.edu/archives/fall2008/entries/double-effect/

Murray, Mary Elizabeth (2008). "Moral Development and Moral Education: An Overview", University of Illinois at Chicago website. Last accessed on September 15, 2008: http://tigger.uic.edu/~lnucci/MoralEd/overview.html

National Defense Authorization Act (2000). Floyd D. Spence National Defense Authorization Act for Fiscal Year 2001, Public Law 106-398, Section 220. Last accessed on September 15, 2008: http://www.dod.mil/dodgc/olc/docs/2001NDAA.pdf

National Transportation Safety Board (2007). "NTSB Cites Wide Range of Safety Issues in First Investigation of Unmanned Aircraft Accident", NTSB press release, October 16, 2007. Last accessed on September 15, 2008: http://www.ntsb.gov/Pressrel/2007/071016b.htm

North American Computing and Philosophy (2008). "The Limits of Computation" conference, July 10-12, 2008, Bloomington, IN. Last accessed on September 15, 2008: http://www.ia-cap.org/na-cap08/index.htm

O'Brien, William V. (1981). *The Conduct of Just and Limited War,* New York, NY: Praeger Publishers.

Oh, Daniel (2008). "The Relevance of Virtue Ethics and Application to the Formation of Character Development in Warriors", *The Army Chaplaincy* online journal, Spring-Summer 2008. Last accessed on September 15, 2008: http://www.usachcs.army.mil/TACarchive/tacss08/tacss08oh7.pdf

Orend, Brian (2001). *Michael Walzer on War and Justice*, Montreal, Quebec: McGill-Queen's University Press.

Orend, Brian (2002). "Justice After War", *Ethics & International Affairs* 16.1: 43-56.

Orend, Brian (2006). *The Morality of War*, Peterborough, Ontario: Broadview Press.

Padgett, Tim (2008). "Florida's Blackout: A Warning Sign?", *Time.com*, February 27, 2008. Last accessed on September 15, 2008: http://www.time.com/time/nation/article/0,8599,1717878,00.html

Page, Lewis (2008). "US War Robots 'Turned Guns' on Fleshy Comrades", *The Register* (UK), April 11, 2008. Last accessed on September 15, 2008: http://www.theregister.co.uk/2008/04/11/us_war_robot_rebellion_iraq/

Royal United Services Institute (RUSI) for Defence and Security Studies (2008). "The Ethics of Autonomous Military Systems" conference, February 27, 2008, London, UK. Last accessed on September 15, 2008: http://www.rusi.org/events/past/ref:E47385996DA7D3/

Rowe, Neil C. (2008). "Ethics of Cyber War Attacks", in Lech J. Janczewski and Andrew M. Colarik (eds.) *Cyber Warfare and Cyber Terrorism*, Hershey, PA: Information Science Reference

Russell, Stuart J., and Norvig, Peter (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, NJ: Prentice Hall.

Searle, John (1980). "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3.3: 417-457

Shachtman, Noah (2007). "Robot Cannon Kills 9, Wounds 14", *Wired.com,* October 18, 2007. Last accessed on September 15, 2008: http://blog.wired.com/defense/2007/10/robot-cannon-ki.html

Sharkey, Noel (2007a). "Robot Wars are a Reality", *The Guordion* (UK), August 18, 2007, p. 29. Last accessed on September 15, 2008: http://www.guardian.co.uk/commentisfree/2007/aug/18/comment.military

Sharkey, Noel (2007b). "Automated Killers and the Computing Profession", *Computer* 40: 122-124. Last accessed on September 15, 2008: http://www.computer.org/portal/site/computer/menuitem.5d61c1d591162e4b0ef1bd108bcd45f3/index.jsp?&pName=computer_level1_article&TheCat=1015&path=computer/homepage/Nov07&file=profession.xml&xsl=article.xsl&

Sharkey, Noel (2008a). "Cassandra or False Prophet of Doom: AI Robots and War", *IEEE Intelligent Systems*, July/August 2008, pp. 14-17. Last accessed on September 15, 2008: http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2008/X4-08/x4his.pdf

Sharkey, Noel (2008b). "Grounds for Discrimination: Autonomous Robot Weapons", *RUSI Defence Systems*, 11.2: 86-89.

Simon, Herbert (1947). *Administrotive Behavior: A Study of Decision-Moking Processes in Administrative Organizations* (1997 fourth edition), New York, NY: Free Press.

Sofge, Erik (2008). "The Inside Story of the SWORDS Armed Robot 'Pullout' in Iraq: Update", *PopularMechanics.com*, April 15, 2008. Last accessed on September 15, 2008: http://www.popularmechanics.com/blogs/technology_news/4258963.html

Solomon, David (1988). "Internal Objections to Virtue Ethics", in Peter A. French, Theodore Uehling, Jr., and Howard Wettstein (eds.), *Midwest Studies in Philosophy Vol. XIII Ethicol Theory: Chorocter and Virtue*, Notre Dame, IN: University of Notre Dame Press.

Solum, Lawrence (1992). "Legal Personhood for Artificial Intelligences," *North Carolina Law Review*, Vol. 70: 1231-1287.

Sparrow, Rob (2007). "Killer Robots", *Journol of Applied Philosophy*, Vol. 24, No. 1: 62-77

Thompson, Paul B. (2007). *Food Biotechnology in Ethical Perspective, 2nd ed.*, Dordrecht, The Netherlands: Springer.

Turing, Alan (1950). "Computing Machinery and Intelligence", *Mind,* Vol. 59, No. 236: 434–460.

University of San Diego (2008). Ethics Updates webpage. Last accessed on September 15, 2008: http://ethics.sandiego.edu/index.asp

US Army Surgeon General's Office (2006). *Mentol Heolth Advisory Teom (MHAT) IV: Operotion Iraqi Freedom 05-07,* November 16, 2006. Last accessed on September 15, 2008: http://www.globalpolicy.org/security/issues/iraq/attack/consequences/2006/1117mhatreport.pdf

US Army Surgeon General's Office (2008). *Mentol Heolth Advisory Teom (MHAT) V: Operotion Iroqi Freedom 06-08,* February 14, 2008. Last accessed on September 15, 2008: http://www.armymedicine.army.mil/reports/mhat/mhat_v/Redacted1-MHATV-OIF-4-FEB-2008Report.pdf

US Department of Defense (2007). *Unmonned Systems Roodmop 2007-2032*, Washington, DC: Government Printing Office. Last accessed on September 15, 2008: http://www.acq.osd.mil/usd/Unmanned%20Systems%20Roadmap.2007-2032.pdf

US Department of Energy (2004). *Finol Report on the August 14, 2003 Blockout in the United States and Canada: Couses and Recommendations*, Washington, DC: Government Printing Office. Last accessed on September 15, 2008: https://reports.energy.gov/BlackoutFinal-Web.pdf

US Department of the Navy (2004). *Novy Unmonned Underseo Vehicle (UUV) Moster Plon,* Washington, DC: Government Printing Office. Last accessed on September 15, 2008: http://www.navy.mil/navydata/technology/uuvmp.pdf.

Van der Loos, H.F. Machiel (2007). "Ethics by Design: A Conceptual Approach to Personal and Service Robot Systems", *Proceedings of the IEEE Conference on Robotics and Automotion, Workshop on Roboethics*, April 14, 2007, Rome, Italy.

Veruggio, Gianmarco (2007). *EURON Roboethics Roadmap*, Genova, Italy: European Robotics Research Network. Last accessed on September 15, 2008: http://www.roboethics.org/icra07/contributions/VERUGGIO%20Roboethics%20Roadmap%20Rel.1.2.pdf

Wallach, Wendell and Allen, Colin (2008). *Morol Machines: Teaching Robots Right from Wrong*, New York, NY: Oxford University Press.

Walter, W.G. (1950). "An Imitation of Life", *Scientific American,* 182:42-45.

Walzer, Michael (1977). *Just ond Unjust Wors: A Morol Argument with Historical Illustrations*, New York, NY: Basic Books.

Weckert, John, ed. (2007). *Computer Ethics*, Burlington, VT: Ashgate Publishing.

Appendix A: Definitions

While we do not want to be entangled with debating precise definitions in this report, it nevertheless would be useful to give more detailed explanations of our key terms—namely 'robot', 'autonomy', and 'ethics'—to ensure a common understanding from the start:

A.1 Robot

Before we offer a working definition of a robot, let us note some historical origins: In 1921, the world was introduced to modern concept of a robot in the popular play *R.U.R.* (or *Rossum's Universal Robots*) by Czech author Karel Čapek. The dystopian play featured factory-built, artificial people, who can be mistaken as humans, called robots—whose namesake is derived from the Czech word 'robota' which means 'drudgery' or 'servitude' or 'labor', and these engineered slaves ultimately rebel against their human masters.

Čapek's robots were biological based and more akin to the resurrected man-creature in Mary Shelley's *Frankenstein* or the replicants of Ridley Scott's *Blade Runner* than to the modern fusion of computer and machine, popularized by science-fiction author Isaac Asimov and others. Therefore, the idea that robots are electromechanical is not part of the original conception of robots; nor does it seem to be an essential feature, since we can imagine advances in genetic engineering and synthetic biology to some day enable us to create biological-based artificial creatures that function as today's robots do and can be called 'robots' (or called something else if such developments cause us to evolve our terminology, as 'robot' had replaced older words such as 'automaton' and 'android'). That said, we will concern ourselves here primarily with electromechanical machines, though we will leave open the possibility of biological and virtual machines or creations as robots.

Further, though the original notion of a robot is tied with automation of work, we are more interested in how the definition of robot differentiates a robot from a garden-variety, mere machine. That is, our definition is not meant to help elucidate the difference between a robot and automation or work, rather to help explain why objects such as a laptop computer or a coffee machine do not count as robots.

To the definition now, in its most basic sense, we define '**robot**' as a *machine that senses, thinks, and acts*: "Thus a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. Sensors are needed to obtain information from the environment. Reactive behaviors

(like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment. Generally, these forces will result in motion of the entire robot or one of its elements (such as an arm, a leg, or a wheel)" [Bekey, 2005].

For all practical purposes today, this means that a robot is essentially a computer with sensory inputs (that do not require direct and intentional human action, such as required by a keyboard or touchpad; i.e., it can direct itself given certain environmental inputs) and non-digital output (i.e., more than manipulation of data or pixels or even sending a file to printer, but an ability to move some part of itself in order to manipulate real-world objects). This is neither a precise definition nor description, and it certainly needs to be refined; but it at least begins to meet our needs for a general understanding of what counts and what does not count as a robot.

A standard, more exact definition, however, proves elusive, as perhaps evidenced by the fact that no major robotics organization to our knowledge—including Robotic Industries Association, IEEE and its Robotics & Automation Society, European Robotics Research Network, Japan Robot Association, Australian Robotics & Automation Association, and others—provides a clear definition of the term on their respective sites or publications, as far as we can tell. Some organizations, such as International Federation of Robotics, follow the definition given by the International Organization for Standardization (ISO) for *manipulating industrial robots*: "an automatically controlled, reprogrammable, multipurpose, manipulator programmable in three or more axes, which may be either fixed in place or mobile for use in industrial automation applications" [International Federation of Robotics, 2008]. But this is far from an adequate definition for a generic robot, and no ISO definition for such a robot has been devised, to our knowledge.

In a military context, we have a more useful understanding from the US Department of Defense's (DoD) definition of an 'unmanned vehicle' (though some of these might not properly be robots, at least under our working definition): "A powered vehicle that does not carry a human operator, can be operated autonomously or remotely, can be expendable or recoverable, and can carry a lethal or non-lethal payload. Ballistic or semi-ballistic vehicles, cruise missiles, artillery projectiles, torpedoes, mines, satellites, and unattended sensors (with no form of propulsion) are not considered unmanned vehicles. Unmanned vehicles are the primary component of unmanned systems" [US Department of Defense, 2007, p. 1].

We take 'vehicle' to mean a mobile, maneuverable machine such as a car, boat, or airplane, but in our analysis, a robot need not be mobile (though it seems most will be in military applications). Yet some degree of mobility is an essential feature of a robot, as we mentioned in our first definition of a robot above; for instance, a robot can be a fixed manufacturing machine with movable arms or an immobile sentry robot with swiveling gun turrets. The mobility requirement here has less to do with moving from one physical location to another (although most robots can and will do this) but more

with the ability to interact with and manipulate the external, physical world to some meaningful degree. This requirement then differentiates a robot from, say, a computer with environmental sensors but that can only run software programs and not exert a sufficient amount of force on the outside world (beyond spitting out pages of paper or opening its CD player door).

Relatedly, the concept that the machine is *powered* is needed in a sensible definition of a robot. Though it is taken for granted that all machines operate under some power, especially where mobility and information processing is required, we want to rule out 'dumb' machines that are not internally driven but nonetheless seem to sense, think, and act. For instance, a small sensor may be designed to be carried by wind or water and mechanically perform some action under certain conditions (such as release a payload when a thermal, chemical, or gyroscopic switch is tripped): such a machine appears to sense, act, and think (to the extent it performs an action under the right conditions, similar to a mechanical calculator of the early twentieth century), yet we would resist calling it a robot, though it would still be considered a machine, i.e., robots belong to a subset of machines. This is not to say that very small robots cannot be carried by wind or waves, but they would also need to convert or otherwise use that energy, or use some other power, to independently move itself or some part of the machine. Thus, our definition of a robot should include the notion of *internal* or *self-directed power* (e.g., electricity generated by a battery or harnessed from solar or wave energy), as well as the existence of something to be powered (e.g., a computer processing chip or payload-release mechanism).

However, the DoD definition needs to be modified for our use in this report: It seems to be overly broad to include fully-remote-controlled machines as robots, since many children's toys would qualify as robots, such as a toy car tethered or wirelessly connected to a control knob (though a few toys, such as AIBO™ or Pleo™ or Robosapien™, may truly be robots). That is, most of these toys do not make decisions for themselves; they depend on a human actor. Rather, the generally-accepted idea of a robot depends critically on the notion that it has some degree of autonomy or can 'think' for itself, as it makes decisions and acts upon the environment. Thus, the Air Force's Predator, though mostly tele-operated by humans, makes some navigational decisions on its own and therefore would count as a robot. Further, robots need not be unmanned, though many are and will be. It is conceivable that a robot may indeed be partially autonomous *and* carry a human operator who makes some decisions, so we do not want to rule such a machine out in our definition. Indeed, beyond the distinction between a robot and a mere machine, the line between robot and human may soon become blurred as robotic technologies are integrated with biological bodies.

As for the ISO's requirements of multi-purpose, reprogrammable, and movable on three or more axes, those seem to be unnecessarily limiting for a general definition of a robot, though perhaps appropriate for the ISO's purpose of defining a manipulating industrial robot. So we will not include those requirements in our conception of a robot here, though further investigations may warrant them. For instance, we are leaving open the question whether a robot needs to be programmable or

reprogrammable, since we may envision counterexamples of a disposable, one-time-use robotic insect that is not reprogrammable or some autonomous robot that transcends its programming.

Therefore, our working definition of a robot—a powered machine that (1) senses, (2) thinks (in a deliberative, non-mechanical sense), and (3) acts—appears defensible and comprehensive. In addition to the cases preciously discussed, it can rule out as robots the following (current but perhaps not future) types of ordnances and technologies: ballistic or semi-ballistic vehicles, cruise missiles, artillery projectiles, torpedoes, mines, satellites, and unattended sensors (with no form of propulsion). However, a 'smart' mine, for instance, conceivably may be developed in the future such that it can sense, think (i.e., discriminate among targets), and act (i.e., exert forces upon the environment, other than self-destruct) and therefore considered to be a robot. In non-military contexts, our definition eliminates mundane objects, such as coffee machines (they don't think; see related discussion below about autonomy) and personal computers (which, by themselves, don't exert an influence on or sense the external world in a significant way, and they require human inputs), as robots.

A.2 Autonomy

This brings us to another critical concept we need to define: autonomy. Though this task is even more difficult than the former, we will offer less analysis, given that the different conceptions, complexity, and applications of autonomy are well covered throughout philosophical and legal literature (and such discussions about the definition of a robot has not been nearly extensively covered in technical or other literature).[8] For the purposes of this report, it will suffice to initially stipulate '**autonomy**' to be about *the capacity to operate in the real-world environment without any form of external control for extended periods of time* [Bekey, 2005]. (But see the refined definition below.)

This is to say that, in defining the term, we are not interested at this point in issues traditionally linked to autonomy, such as the assignment of political rights and moral responsibility (as different from legal responsibility) or even more philosophical issues related to free will, moral agency, personhood, and whether machines can even 'think' and have intentions (as opposed to merely being programmed to achieve some goal)—as important as those issues are in philosophy, law, and

[8] We want to recognize a *technical* account of autonomy, such that autonomy is measured by the amount of time it takes for a system to refer or 'check back' with a human before proceeding with a certain action [Hew, 2007]. Thus, a landmine has infinite autonomy, since it never needs to refer back to a human for authorization once it has been armed; and a fully remote-controlled robot would have no autonomy, since it is constantly referring back to a human (and indeed completely dependent on a human) for instructions. While this may be a useful account of autonomy in some technical discussions, it does not seem to be relevant to a discussion about ethics and risk, to the extent that such issues arise from a system's ability to make unpredicted, unforeseen, unanticipated, or undesirable choices. Further, from a legal and ethical standpoint, it seems to be incoherent to ascribe autonomy to unthinking objects such as landmines or a toaster.

ethics. Therefore, in the interest of simplicity, we will content ourselves to define and discuss autonomy in the context of human-created machines.

The notion of autonomy is important to help elucidate the second criteria of 'thinking' in our basic working definition of a robot. Like autonomy, much controversy surrounds our understanding of 'thinking', especially whether it is appropriate to apply that term to machines. By this term, we do not mean the mere capability of information or data processing; that would make most of our electronic devices into thinking things and make the term overly broad. Rather, by 'thinking', we mean to include some degree of autonomy or decision-making not influenced by external controllers, giving the machine an appearance of deliberative thought, if not the actual ability to meaningfully make choices.[9]

Thus, given our initial definition of autonomy, *fully* remote- or tele-operated machines would not count as autonomous, since they are not operated without external control; they cannot 'think' and therefore cannot act for themselves. (Again, tele-operated vehicles such as the Air Force's Predator would count as robots under our working definition, because they have some autonomy, such as in navigation, even if they do not make any strike decisions.) Neither are today's desktop or laptop computers autonomous, since they still require human inputs. Yet a problem with our simple definition may arise: wouldn't autonomous robots simply be sensing and moving computers that run programs, like everyday computers; and through these programs, both computers and robots can be said to be 'externally controlled' by the programmer or team of programmers that created the program? That is to say, the notion of *external control* is vague and begs for clarification.

Let's retreat one step to ask the following: might we consider some computers as *semi*-autonomous, such as one that runs a computer program that enables an avatar (or virtual-reality character or persona) to run without further external control, i.e., an avatar that seems to act on its own, as some already do now? Surely, such an avatar would be considered at least semi-autonomous, at least by popular standards, but how do we adjust our definition of autonomy to include such a case?

While it is true that programs need to be created by some programmer, even programs written by other programs, there will always be some external, human cause for whatever actions machine exhibit; so artificial autonomy would be impossible if no person can play a role in the causal chain, especially at the programming level. Indeed, the philosophical position that free will does *not* exist seems to depend on a related argument, that in a scientific, deterministic world, there must be some

[9] A logical implication of this is that *all* robots, as we define them, will have some degree of autonomy, making 'autonomous robot' a redundant expression; but we will nevertheless keep with this expression to signal that we are referring to robots that have a greater degree of autonomy than usual, if one considers autonomy as a spectrum from unthinking automatons (such as bacteria and simple organisms) to semi-autonomous beings (such as children, some animals, and some robots today) to fully-autonomous, moral agents.

prior, external cause for our behavior that we are not responsible for (and thus we are not responsible for our consequent actions or even have the power to alter that chain of events).

So to avoid this problem, let us stipulate that artificial autonomy is possible and that it does *not* imply that a machine's actions are undetermined or unpredictable (as may be required in the usual conception of free will). This seems to require that we exempt programmers and designers from the considered causal chain, such that a robot running a sophisticated program may be considered to be semi- or fully-autonomous, even though its actions may be predetermined given certain environmental conditions, at the programming or design level. Thus, we may also consider some computers today to be semi-autonomous in letting loose a self-directed avatar upon the virtual world.

Given our use of 'autonomy' here, a semi- or fully-autonomous robot would be able to choose to perform at least some actions 'on its own' or without a human determining—at least not at a design or programming level—what course of action it should take. Thus to refine our initial definition, we take **autonomy** in machines to mean: *the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time.*

A.3 Ethics

Finally, in this report, we use the term '**ethics**' broadly to include not just normative issues, i.e., questions about what we should or ought to do, but also general concerns related to social and cultural impact as well as risk, e.g., responsibility after a malfunction, arising from the use of robotics. As a result, we will cover all these areas in our report, not just philosophical questions or ethical theory, with the goal of providing some relevant if not actionable at this preliminary stage. (This diverges from the traditional, more narrow conception of ethics, at least as understood in academic philosophy.)

A note on the issue of ethics: Is robot ethics a subset of military ethics, or computer ethics, or some other area of ethics? We believe that robot ethics is emerging to become a field unto its own. There is still a critical gap between machine or computer ethics and military ethics, in which important questions are only now being raised about the moral responsibility, risk, and just use in war of relatively autonomous systems (robots or computer networks). Further, robot ethics is not limited to military-related issues; there are new dilemmas related to the use of robots as a proxy for elderly care, for sexual relationships, for human workers especially those with specialized skills such as surgeons, and so on. Thus, the introduction of 'smart' robotics into the military and the marketplace has implications for both military and other ethics, with potentially transformative consequences on

many traditional issues, such as the nature of personhood, agency, autonomy, and even what it means to be a soldier.

Appendix B: Contacts

1. **Patrick Lin, Ph.D.**
California Polytechnic State University
Ethics & Emerging Technologies Group
Philosophy Department
1 Grand Avenue
Building 47, Room 37
San Luis Obispo, California 93407
Email: palin@calpoly.edu
Dept. phone: 805-756-2041

2. **George Bekey, Ph.D.**
California Polytechnic State University
Ethics & Emerging Technologies Group
Biomedical/General Engineering Department
Building 13, Room 260
San Luis Obispo, California 93407
email: gbekey@calpoly.edu
Dept. phone: 805-756-6400

3. **Keith Abney, M.A.**
California Polytechnic State University
Ethics & Emerging Technologies Group
Philosophy Department
1 Grand Avenue
Building 47, Room 37
San Luis Obispo, California 93407
email: kabney@calpoly.edu
Dept. phone: 805-756-2041

Unmanned Aircraft Systems; Situational Awareness for Small Tactical Unmanned Aircraft and an Autonomous Package Delivery Concept Study for the US Marines

Project Investigator:

Rob McDonald
Department of Aerospace Engineering
California Polytechnic State University
San Luis Obispo, CA

Unmanned Aircraft Systems;
Situational Awareness for Small Tactical Unmanned Aircraft and An Autonomous Package Delivery Concept Study for the US Marines

Project Report

Rob McDonald, Ph.D.

Aerospace Engineering
California Polytechnic State University

Project Impact

This C3RP research project funded in February 2008 has been a great catalyst to the development of a UAV research program at Cal Poly. Although the no cost extension ended with the close of 2008, work continues on this project to this day. This project can claim a number of direct and indirect products.

This project has directly resulted in one Aerospace Engineering (AERO) MS Thesis project (Shane Wallace) which will be finished in September 2009. It has also produced two AERO senior projects (Andrew Ezzard and Ryan Halper) and a summer project for a Computer Engineering undergraduate student (Nick Utchig).

As a result of this project, the PI joined with AeroMech Engineering to submit a $3M proposal to an Office of Naval Research BAA. The BAA was not funded, but a strong relationship with AeroMech continues.

This project has resulted in two conference papers, with the possibility of more to come.

Ezzard, A., Vallone, M., and McDonald, R., *Underpowered Aircraft – Performance and Operational Possibilities*. AIAA Aerospace Sciences Meeting, January 5-8, 2009, Orlando Florida. AIAA 2009-1441.

Wallace, S, and McDonald, R., *Situational Awareness for Small Tactical Unmanned Aircraft*. AIAA Aerospace Sciences Meeting, January 2010, Orlando Florida, Accepted for presentation.

The indirect products of this project are perhaps even more significant. This project initiated a UAV research program which has gained the attention of industry and government. In the Spring of 2009, Northrop Grumman donated an unmanned Yamaha RMAX helicopter to Cal Poly. An autopilot is being obtained for the RMAX; when installed, it will provide a first-rate fully autonomous research platform.

The ceremony transferring the RMAX to Cal Poly caught the eye of the San Luis Obispo Sherrif's office Search and Rescue (SAR) team. Computer Engineering professor Dr. Lynne Slivovsky and the PI have been meeting with the SAR team about possible collaboration and state funding opportunities.

Also in the Spring of 2009, the Air Force Flight Test Center at Edwards Air Force Base

approached the PI about submitting a research proposal in the area of UAV's. A proposal was submitted in collaboration with Mechanical Engineering professor Dr. Russ Westphal; notification of funding was received in August 2009. This project will support the design and construction of a UAV to be used as a flight test instrumentation development platform.

Autonomous Package Delivery Concept – Underpowered Aircraft

A unique aircraft concept known as an Underpowered Aircraft was studied in its appropriateness for a mission in support of US Marine Seabasing. The Underpowered Aircraft allows for the modification of gliding flight vehicles for increased range and lower cost when compared with fully powered flight vehicles.

Intentionally under-sizing the powerplant for a flight vehicle allows the designer to choose a powerplant that will not only perform the mission requirements, but will also provide the customer with the most cost effective solution, as some missions may not require fully powered flight. Specifically, the underpowered aircraft concept studied in this project is a gliding flight aircraft that does not have enough power for climbing or level flight, but does have enough power to overcome some of the drag forces associated with flight, in turn increasing the effective range of the vehicle.

The underpowered aircraft concept was analyzed and its feasibility was determined. Analysis done using equations of motion, followed by a more accurate numerical integration including a thrust lapse, determined that the underpowered aircraft concept provides a unique method for a cost effective range extension technology for gliding flight vehicles. The technology and methods of this study were applied to the AGM-154 JSOW and JSOW-ER glide munitions and it was determined that JSOW-ER is representative of an underpowered aircraft with our analysis.

The AIAA paper and ONR BAA proposal which resulted from this effort are attached to this report for review

Situational Awareness for Small UAV's – Laser Scanner and Collision Avoidance

This project has supported a comprehensive effort to develop a low cost collision avoidance capability for small UAV's. By comprehensive, it is meant that a number of sub-projects have been undertaken each building toward the overall goal. These sub-projects include the design, simulation, and test of each component in the system.

Testbed UAV

The off-the-shelf Piccolo LT autopilot was integrated into a standard radio control aircraft. All sensors, antennae, and electronics were installed in the aircraft (Figure 1). Aerodynamic, propulsion, mass, and stability models of the aircraft were built and programmed into the autopilot. The aircraft was simulated in the autopilot simulation software, where the models were tested and the control gains were tuned. Once the team was satisfied with the simulation, the autonomous aircraft was test-flown at the Cal Poly Educational Flight Range (EFR) (Figure 2 and 3).



Figure 1: Piccolo LT autopilot integration.



Figure 2: Ground station and test flight.



Figure 3: Ground station software.

Software-Only Laser Range Finder Simulator

A software-only simulation of the seanning laser ineluding terrain and buildings was ereated. Simulation of the laser is deceptively complex. The terrain and buildings must be modeled by a large number of triangles combined to represent the surfaees of the obstaeles (Figure 4 and 5). The laser must be oriented to the latitude, longitude, altitude, heading, elimb, and roll of the aireraft as well as the

two axes of rotation of the laser gimbal. Then, a 'ray' must be traced from the laser into space. An intersection test between the ray and every triangle must be performed (Figure 6). This gets very expensive for complex terrain; the computation must be done in real-time, the laser takes measurements at 200Hz.

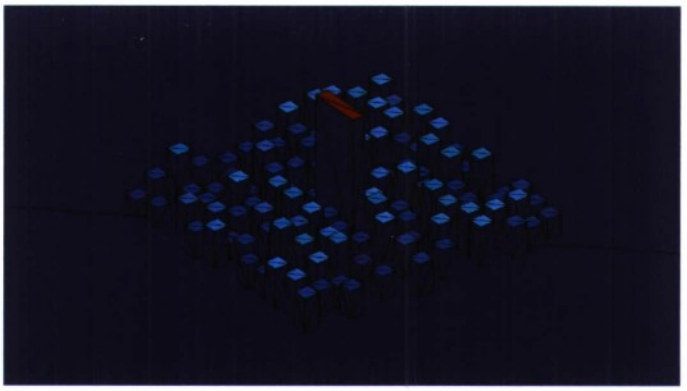


Figure 4: Simulated buildings.

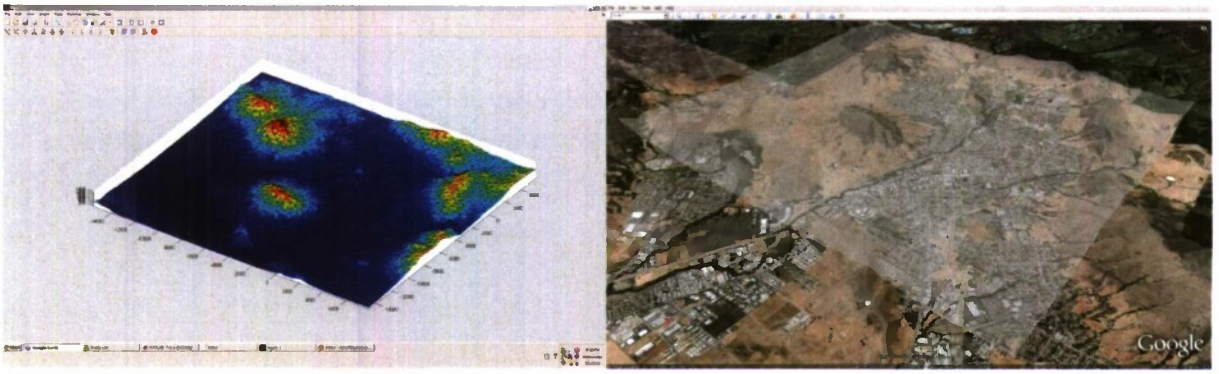


Figure 5: Simulated terrain around San Luis Obispo.

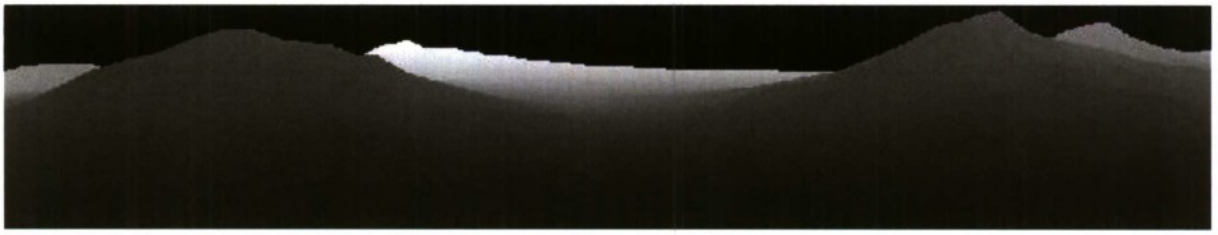


Figure 6: Simulator generated depth map of San Luis Obispo terrain.

In addition to simulating the distance measurements taken by the laser, the laser simulator must communicate with the flight simulation to know the aircraft position and orientation. The laser sim must also receive gimbal position commands from and send laser measurements to the guidance algorithm. Finally, the guidance algorithm must communicate modified waypoints to the autopilot to direct the aircraft away from an obstacle (Figure 7).

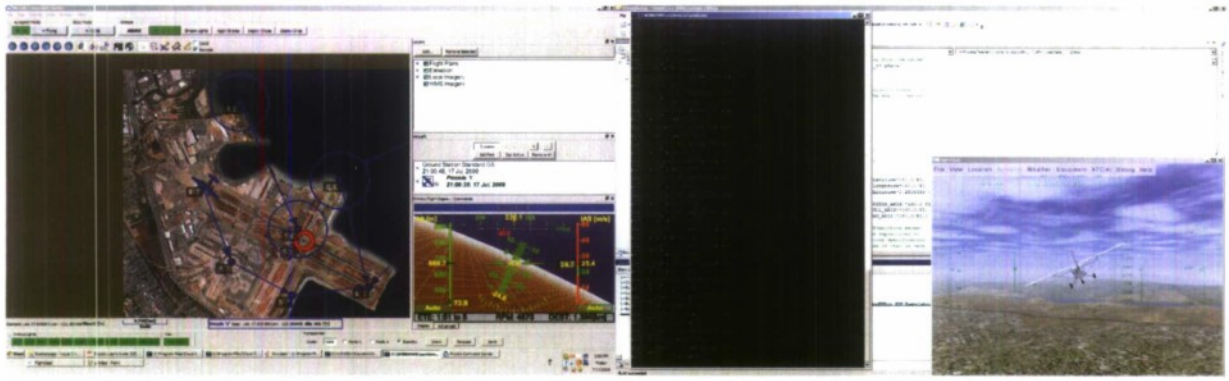


Figure 7: Integration of laser simulator with flight simulator and ground control software.

Potential Function Collision Avoidance Guidance

The final sub-project in this effort has been the development of the actual guidance algorithms which will cause the aircraft to avoid any sensed obstacles. This theoretical and applied effort has been the primary focus of the MS Thesis project which will be completed in September 2009. The use of a potential function allows the onboard computer to efficiently make decisions to avoid obstacles with a minimum of input data and little computational expense.

In the potential function approach, the specified waypoint navigation is represented with 'sinks' which draw the aircraft towards them. Simultaneously, the obstacles detected by the laser (terrain, buildings, trees, etc) are represented by 'sources' which repel the aircraft. The sources and sinks combine to produce a smooth potential surface; the guidance algorithm directs the aircraft along the gradients in this surface to follow a minimum potential path.

A simulated flight showing terrain avoidance is depicted below in Figure 8.

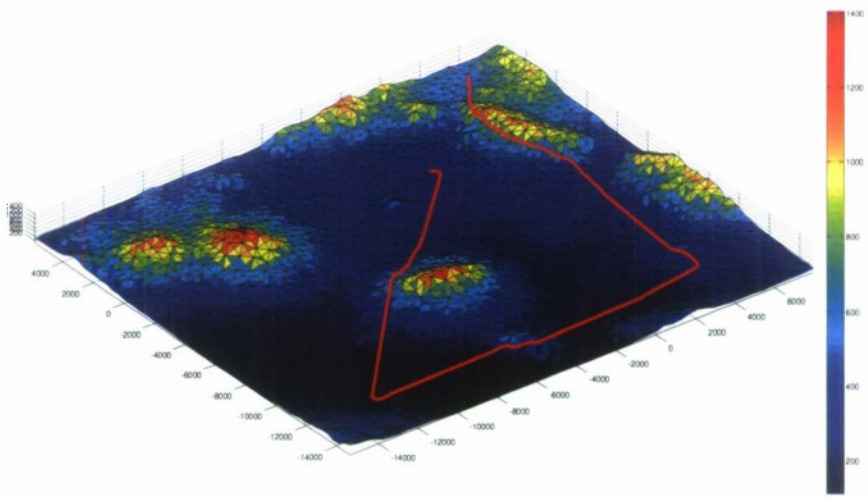


Figure 8: Simulated flight avoiding terrain.

The MS Thesis and conference paper resulting from this work will be appended to this report when they are completed.

.

Sandwich Composite Report

Project Investigator:

Nilanjan Mitra
Department of Civil and Environmental Engineering
California Polytechnic State University
San Luis Obispo, CA

Dr. Nilanjan Mitra

Introduction

Sandwich composite panels represents a special form of laminated composite material in which a relatively thick, lightweight and complaint core material separates thin and strong face sheets. These panels are being increasingly utilized currently for major load bearing structures in civil, mechanical, marine, automotive and aerospace industries. The reason for its wide current use is primarily;

- High strength to low weight ratio – Vehicular structures made out of these are thereby capable of greater range and speed; would require less fuel consumption; and can also be used for greater payload.
- Significant less maintenance cost compared to conventional materials and thereby performs well in life-cycle cost analysis of the structure. As an alternative to steel, these materials donot corrode.
- Cost of construction is also comparatively less and relatively easy to manufacture if complex shapes are required for different reasons in the vehicular industry.
- Research has also demonstrated that sandwich composite panels have improved fatigue and vibration resistance performance compared to their conventional counterparts.
- These panels have also demonstrated their potential as a protection against blast and/or impact. The light weight dense core would be able to absorb energy released for blast and/o impact.
- Panels also posses low magnetic, infrared and radar cross-sectional signatures and

thereby used as a suitable alternative for marine and Inspite of the numerous advantages mentioned above, sandwich composite panels have significant drawbacks and is thereby an active area of research. The sandwich composite panels dealt with in this manuscript are typically used for marine construction and are notoriously sensitive to failure by shear load. Shear strength is one of the primary design criterions in application of these structures for marine environment. An example application area is the bottom of large high speed ships subjected to high sea pressures which might result in shearing of the bottom along its perimeter. Details about the properties of the component materials, method of manufacture, experimental methods and analytical simulations are provided in the following sections.

Component material properties for sandwich composites used in marine/naval vessel construction

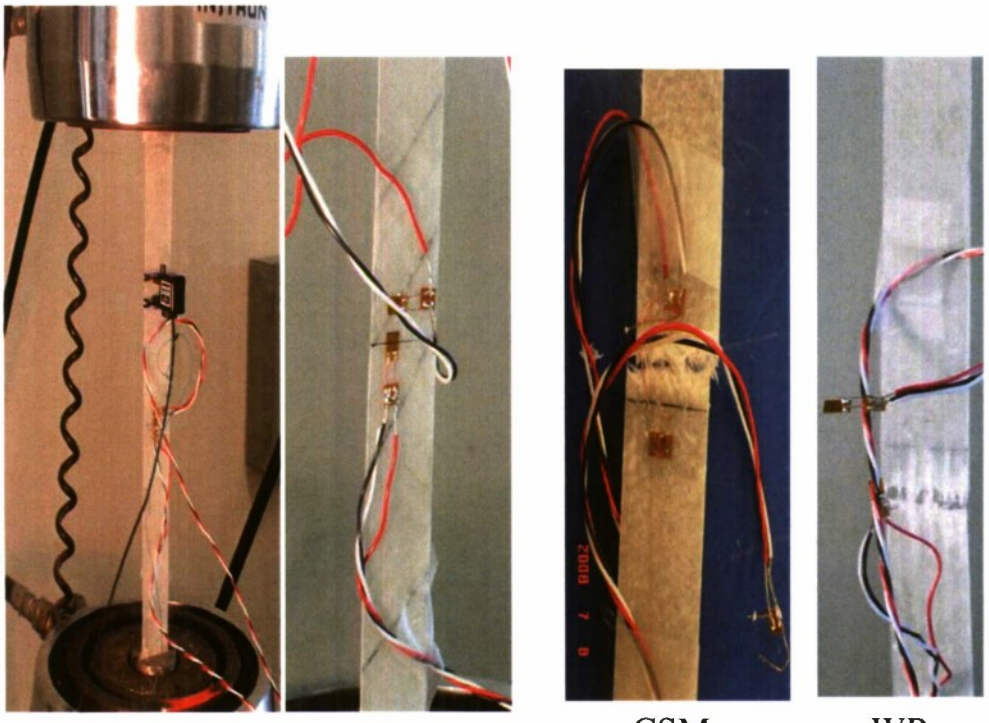
The skin is composed of glass reinforced polymer composites and is made up of two alternative layers of chopped strand mat and woven rovings. The core material is composed of PVC foam. Epoxy resin was utilized as the adhesive material. Method of construction of the samples was vacuum resin infusion.

Glass fibre composites

Glass fibre composites used as skins for the sandwich composite panels are typically made up of chopped strand mats and woven rovings.

The fibers in a chopped strand mat are typically 3-4 inches in length and are randomly oriented. Chopped strand mat is not a very strong material because of the short fiber length. However, it is isotropic. This means that it is equally strong in all directions. Mat and fillers are the only composite reinforcements exhibiting this trait. This is the least expensive reinforcement form and is thus the most widely used. The weight of chopped strand mat used in the current construction is 1.5 oz/ft^2 (0.44 kg/m^2). Typically, this is designed for use with polyester resin or vinyl ester resin. The manual reports that it is not compatible with epoxy resin but in the current study epoxy resin has been used and no detrimental effect was observed. It should also be noted that to form the glass fibre composite skins, chopped strand mats were used in combination with woven rovings. Woven roving is made from continuous glass fiber roving which are interlaced into heavy weight fabrics. The weight of the woven roving used in current construction is 18 oz/yd^2 (0.61 kg/m^2).

Uniaxial tensile tests were performed for both woven rovings and chopped strand mat in an INSTRON machine as per ASTMD3039 "Standard test method for tensile properties of polymer matrix composites" to determine the ultimate tensile strength, modulus of elasticity and poisons ratio of the materials. The CSM and the WR specimen are made up of two layers of chopped strand mat and woven roving sheets respectively bonded together with epoxy resin. The following figure shows the materials being tested.



ASTM D3039 setup CSM WR

Figure 1. ASTM D3039 test for chopped strand mat and woven roving

In the experimental investigations, as demonstrated in Figure 1, the mode of failure as observed for the specimens were brittle. As observed from the experimental investigation, a linear elastic material model was utilized in the numerical simulations to represent the behaviour of both CSM and WR. The poisons ratio for the woven roving is taken as 0.01; which means that there is no coupling between the longitudinal and the transverse strain when loaded uniaxially. For the CSM, the poisons ratio was observed as 0.4. The modulus of elasticity of WR is $2x10^6$ psi (13.8 GPa) whereas for the CSM is $1.7x10^6$ psi (11.8 GPa). The failure stress for CSM samples were recorded as 12000 psi at 7100 microstrain whereas for the WR samples it was 43700 psi at 14700 microstrain.

PVC foam – Divinycell H100

The core material typically chosen for marine/naval ship hull constructions is closed cell semi-rigid poly-vinyl-chloride foam with a density of 100 kg/m^3 manufactured by DIAB Inc. and marketed by the trade name of Divinycell H100. The foam-core thickness considered for the experimental investigation is 30 mm and the foam material has a cell size of approximately 400μm. Tensile, compressive and shear tests were done on the foam samples to determine their strengths. It was observed that the material is ductile in nature in compression however brittle in tension. Since material models are not available which will account for two different failure mechanisms as two different loading scenarios, plastic material models are being utilized for defining the behaviour of these foam materials. The downside of choosing plasticity material model will be the material model would not provide good post-peak response in tension.

Uniaxial compressive test was performed on these foam materials as per recommendations of ASTMC365 "Standard test method for flatwise compressive properties of sandwich cores". The size of the specimen considered was 75 by 75 mm. The standard head displacement rate for the INSTRON 1331 universal testing machine was kept at 0.05 in/min. The stress strain curve obtained from the investigation is shown in figure 2. A load rate study was also performed (by varying load rates to 0.05, 0.1 and 0.2 in/min) on the foam sample and no significant changes with regards to stiffness and or the strength were observed for the specimen.

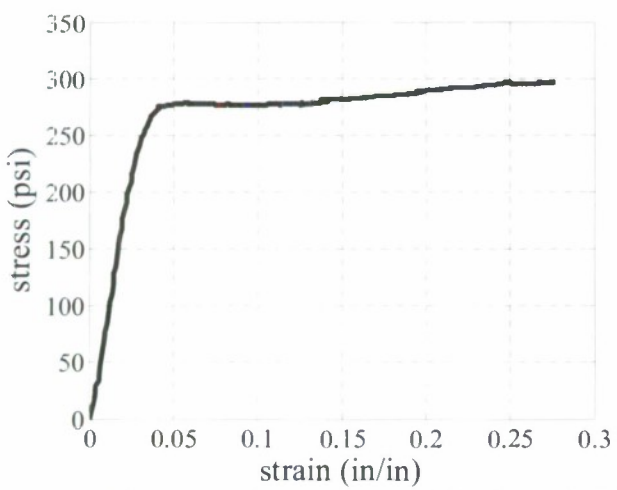


Figure 2. Compressive response of Divinycell H100 semi rigid foam

Cyclic compressive test was also carried out on the foam sample to determine the unloading and reloading characteristics of the specimen. Even though material characteristics associated with unloading and reloading are not going to be utilized in this research, this information would be useful for later investigations involving fatigue and cyclic loading of structures. The stress strain response obtained from the investigation is shown in figure 3. The loading rate was kept the same as that of the ASTM C365 test at 0.05 in/min.

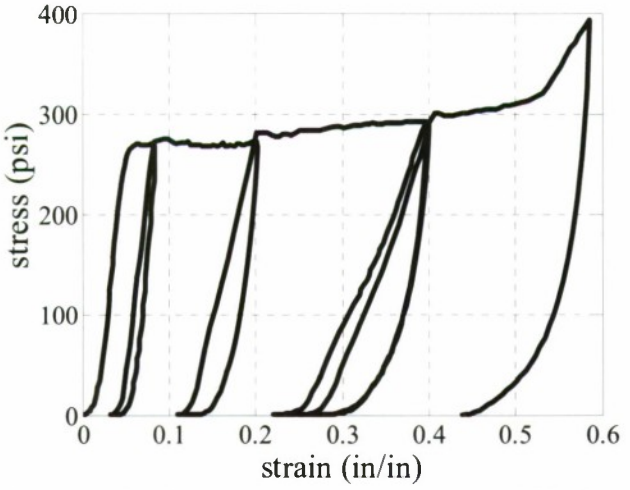


Figure 3. Cyclic compressive response of Divinycell H100 semi rigid foam

Uniaxial tensile tests were also carried out for foam samples as per recommendations of ASTM C297 "Standard test method for flatwise tensile strength of sandwich constructions". Initially the size of the specimen considered was similar to that of the uniaxial compressive tests. However, it was later realized that 75x75x30 mm test coupons might result in triaxial effects and thereby 200x50x30 mm test coupons were taken to account for the uniaxial tensile test. The standard head displacement rate for the INSTRON machine was kept the same at 0.05 in/min. The stress strain response obtained from the investigation is shown in figure 4.

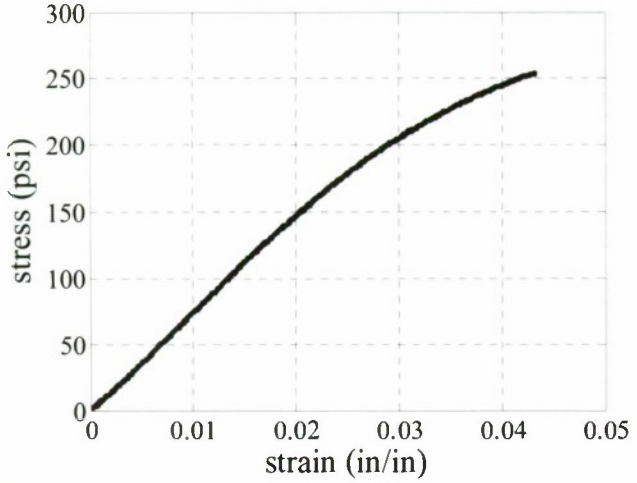


Figure 4. Uniaxial tensile response of Divinycell H100 semi rigid PVC foam

Inplane shear test on the foams was performed as per recommendations of ASTM C273 "Standard test method for shear properties of sandwich core materials". The standard head displacement rate of the INSTRON machine is kept at 0.05 in/min. The shear jigs were manufactured out of steel, which were used in shear tests for both the foam as well as the sandwich composite. The figure below shows the manufactured shear jig and its attachment to the INSTRON machine. The foam specimens (16x2x1.18 in or 406x50x30 mm) are attached to the shear jig by means of structural glue DP460NS. This structural glue is a high performance, non sag, two part epoxy adhesive (B- Epoxy and A- Amine; Mix ratio 2:1; Viscosity cp B-150k-275k and A-8000-14000) offering shear strength of 4650 psi and a peel strength of 60 along with high levels of durability at room temperature. Cracks were observed at an angle in the foam specimen and propagated from one side attached to the shear jig to the other side. The stress-strain response observed from investigation is shown in figure 6.

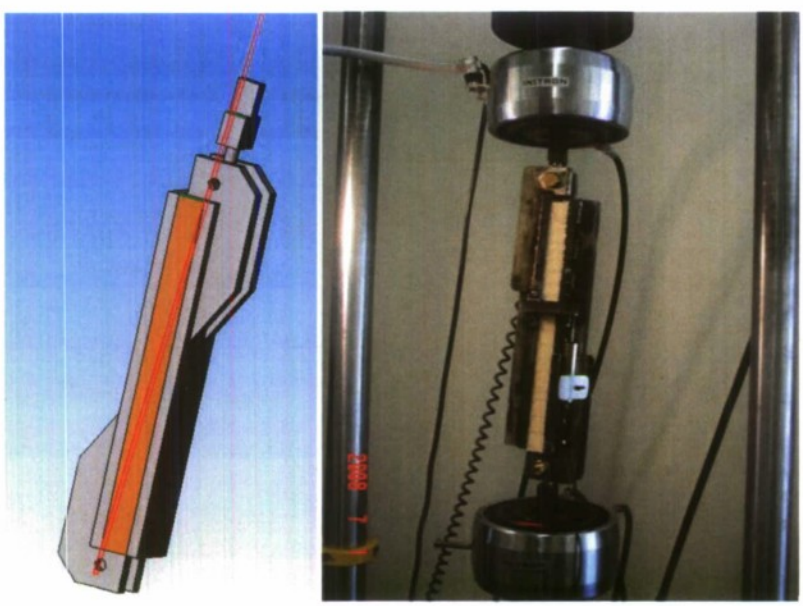


Figure 5. Shear jigs for ASTM C373 tests

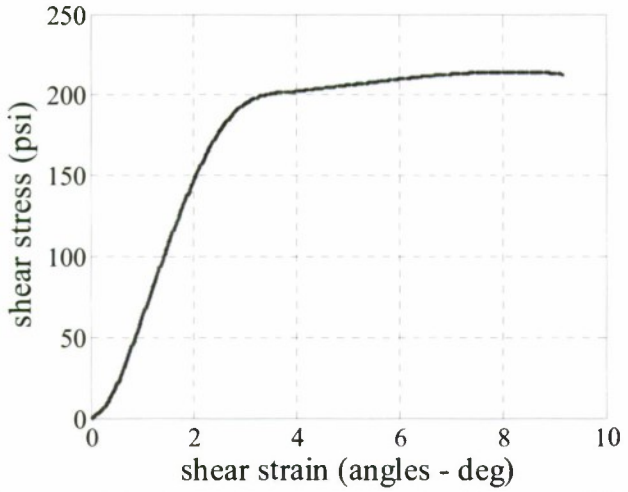


Figure 6. Inplane shear response for Divinycell H100 PVC foam

Since hydrostatic pressure also influences the material response of foam materials, hydrostatic compression test was also carried out for foam materials. However the experimental response did not result in significant conclusions since it was limited by the load that can be applied by the facility/instrument available. Hydrostatic compression test typically means application of an isotropic pressure on all sides of a cube sample and measuring the pressure applied to that of the volumetric strain of the cube. There aren't any ASTM specifications for this test, so triaxial tests that are used to apply triaxial loading to soil samples in Geotechnical laboratory at CalPoly were modified to apply a stress controlled triaxial isotropic pressure on a small cube sample of the foam material with 30 mm dimension. The triaxial testing apparatus in Calpoly applies a axial pressure and the surrounding pressure of water was manually adjusted to match the axial pressure thereby to simulate a triaxial isotropic pressure. The test setup is shown in figure 7.

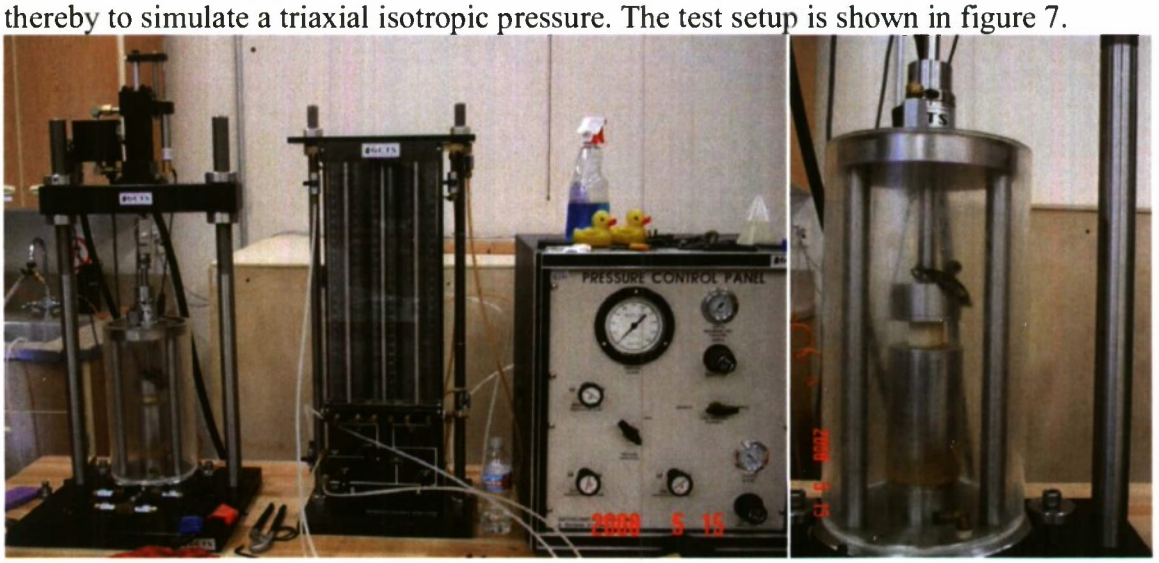


Figure 7. Hydrostatic compression test setup in CalPoly for foam

Simulation of material properties for sandwich composite panel

Linear elastic material model is used for CSM and WR. Material properties for CSM and WR have been obtained from experimental investigations. For the foam core material, a plastic material model, referred to as "Crushable foam plasticity" in ABAQUS has been calibrated. A brief detail of the material model and its calibration is discussed in the following paragraphs.

Crushable foam plasticity with isotropic hardening

It was observed from experimental investigations that the Divinycell H100 foam did undergo permanent deformation in compression. Thereby a plasticity type of model would obviously be required for defining the constitutive relation of the foam material. However, it should also be mentioned that the response in tension for the foam was brittle and a plasticity type model would not be ideal for modeling that. But since no better models were available in ABAQUS, the model that was used for the current study is Crushable foam plasticity model in ABAQUS. In this manuscript, input parameters have been described from a user's viewpoint to calibrate the response obtained from experimental investigation. Calibration of crushable foam plasticity model requires input values for modulus of elasticity, poisons ratio, yield stress and the hardening envelope.

Details about the CRUSHABLE FOAM PLASTICITY model are provided in Deshpande and Fleck (2000) as well as in ABAQUS manual. A brief about the model formulation is provided in the manuscript for brevity and completeness.

The yield surface of the material model, defined in the p-q stress field, is represented as

$$F = \sqrt{q^2 + \alpha^2 (p - p_0)^2} - B = 0$$

where q represents the Mises stress and p represents the hydrostatic stress. The Mises stress q can be represented as $\sqrt{3J_2}$ where J_2 represents the second invariant of the deviatoric stress tensor; the hydrostatic stress p can be represented as $I_1/3$ where I_1 is the first invariant of the stress tensor. The flow potential for the material model, also defined in the p-q stress field, is represented as

$$G = \sqrt{q^2 + \beta^2 p^2}$$

The above two definitions for the yield stress and the flow potential is represented in the following figure

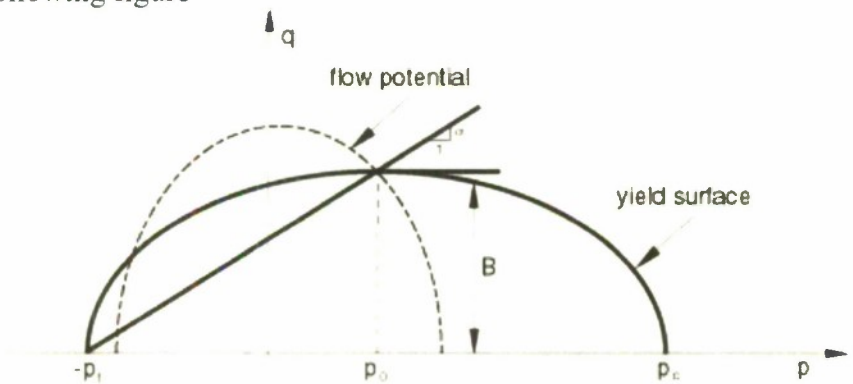


Figure 8. Yeild stress and flow potential of the Crushable Foam Plasticity model (in ABAQUS)

where p_c and p_t refers to yield strength in hydrostatic compression and tension respectively. The parameter p_0 refers to the center of the yield ellipse and is given as

$$p_0 = \frac{p_c - p_t}{2}$$

The parameter B defines the maximum q stress and is represented as

$$B = \alpha A = \alpha \frac{p_c + p_t}{2}$$

where A represents the length of p-axis of the yield ellipse. The shape factor α remains constant during any plastic deformation process. The evolution of the yield ellipse or the hardening envelope is controlled by a plastic strain measure, $\bar{\varepsilon}$, which is the volumetric compacting plastic strain, $-\varepsilon_{vol}^{pl}$ (trace of plastic strain tensor) for the case of volumetric hardening model; and equivalent plastic strain, $\overline{\varepsilon^{pl}}$, which is defined as the absolute value of the axial plastic strain for a bar loaded in uniaxial compression, for the case of isotropic hardening. The yield surface and its evolution for the two type of hardening models: volumetric and isotropic are shown in the figure below.

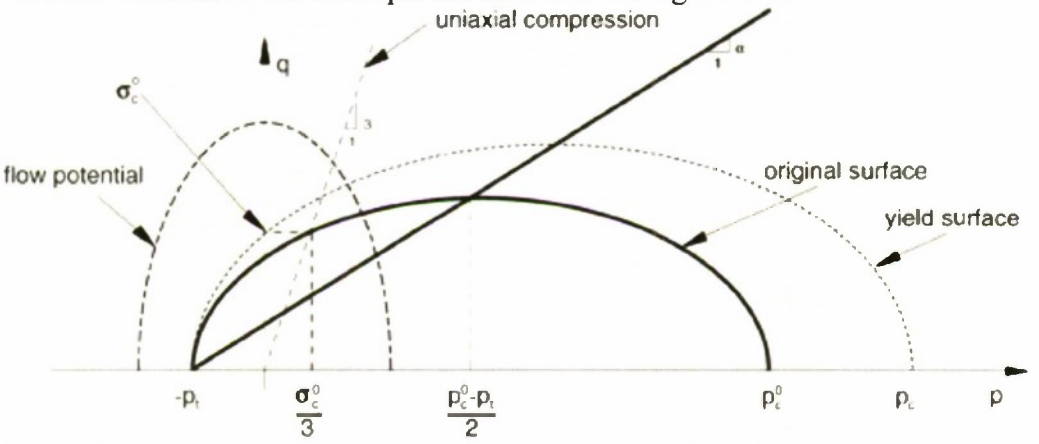


Figure 9. Evolution of the yield surface for volumetric hardening case

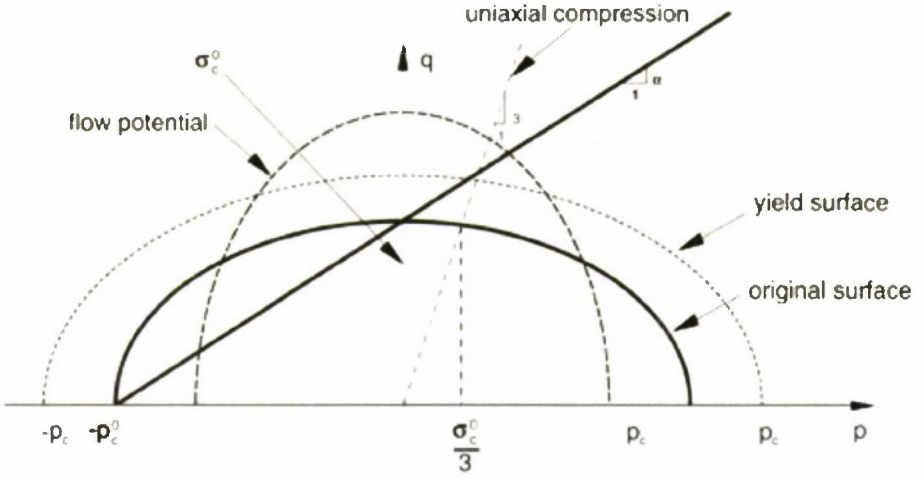


Figure 10. Evolution of the yield surface for the isotropic hardening case

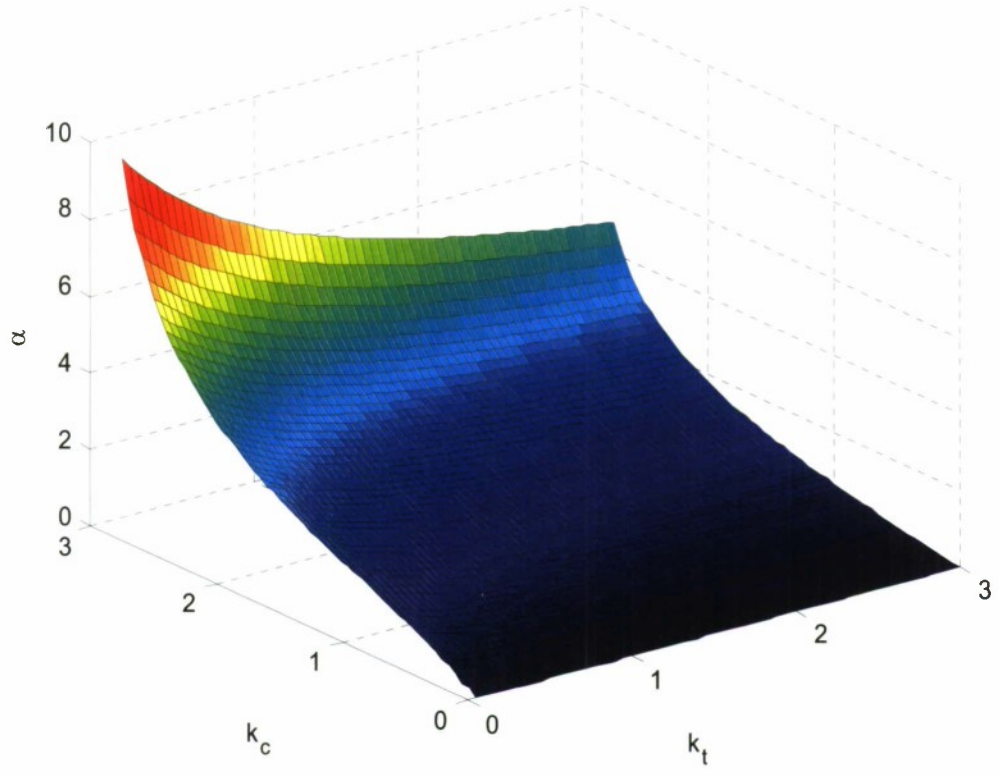
It should be noted that for an isotropic hardening envelope, hydrostatic stress in tension is considered to be same as that of hydrostatic stress in compression, and thereby the center of the ellipse is at the origin of the *p-q* stress plane. Moreover for the isotropic hardening model, the hardening envelope is same for tension as that in compression. The hardening envelope for the compression response is similar to that in tension. On the other hand, for the volumetric hardening model, the hydrostatic stress in tension doesn't change or evolve.

For the volumetric hardening model, the value of α is given as

$$\alpha = \frac{3k_c}{\sqrt{(3k_t + k_c)(3 - k_c)}}$$

where

$k_c = \frac{\sigma_c^0}{p_c^0}$ and $k_t = \frac{p_t}{p_c^0}$. The superscript 0 refers to initial yield surface. The term σ refers to uniaxial stress whereas the term p refers to hydrostatic stress. The subscript t refers to tension whereas the subscript c refers to compression. For a valid yield surface the choice of yield stress ratios must be such that $0<k_c<3$ and $k_t >= 0$. A plot of α is provided with changes in values of k_t and k_c in the figure below. As the valuc of α is increased, the yield stress of the foam material can also be increased. The value of β chosen to represent the volumetric hardening model is $3/\sqrt{2} = 2.12$. For this hardening model, zero plastic Poissons ratio is considered thereby resulting in $\varepsilon_{axial}^{pl} = \varepsilon_{vol}^{pl}$ for uniaxial compression loading.



For the isotropic hardening model, $p_0 = 0$ since the yield stress in hydrostatic tension is same as that in compression. The value of α is given as

$\alpha = \frac{3k_c}{\sqrt{\left(9 - k_c^2\right)}}$ with k_c as defined above. For many low density foams, the initial yield surface is close to a circle in the p-q stress plane, which indicates the value of α is approximately 1. The term β representing the shape of the flow potential in the p-q stress plane is related to the plastic poisons ratio, ν_p as

$$\beta = \frac{3}{\sqrt{2}}\sqrt{\frac{1-2\nu_p}{1+\nu_p}}$$

The plastic Poisson's ratio, which is the ratio of the transverse to the longitudinal plastic strain under uniaxial compression, should be defined by the user; and it must be in the range of –1 and 0.5. The upper limit of 0.5 corresponds to an incompressible plastic flow. For low density foams, the value of plastic poisons ratio is typically taken as 0.

For definition of the crushable foam plasticity model, one should define the modulus of elasticity, poisons ratio and the hardening envelope in tabular form. If isotropic hardening is modeled then the user needs to define the parameter k_c and also plastic poisons ratio; on the other hand for modeling volumetric hardening model, the user must define the parameters k_c and k_t. For isotropic hardening the plastic flow is associated when the value of α is the same as that of β. By default, the plastic flow is nonassociated to allow for the independent calibrations of the shape of the yield surface and the plastic Poisson's ratio.

If information is known only about the plastic Poisson's ratio and the user chooses to use associated plastic flow, the yield stress ratio k_c can be calculated from

$$k_c = \sqrt{3\left(1-2\nu_p\right)}$$

Initially volumetric hardening was chosen for the purpose of analysis. Since experimental investigations of hydrostatic tension and compression test could not be performed for the foam sample, a numerical parametric investigation was carried out to determine the correct value of k_t and k_c for the foam samples so that the numerical model chosen simulates good comparable behavior in uniaxial tension, uniaxial compression and in-plane shear loading. The value finally chosen for k_t and k_c are 1.9 and 2.0 respectively. The values for modulus of elasticity (60 Mpa), poisons ratio (0.1) and a tabular listing of the hardening envelope were obtained from experimental investigations. It should be noted that the values of k_t and k_c donot affect the compressive response but affects the response in tension and shear. Figure 11 shows compressive response simulation with the above defined parameters. Variation of response is also shown with changes in the modulus of elasticity and poison's ratio, as shown in Figure 12. The blue dotted line represents the analytical simulation using ABAQUS and black firm line represents the experimental investigation results.

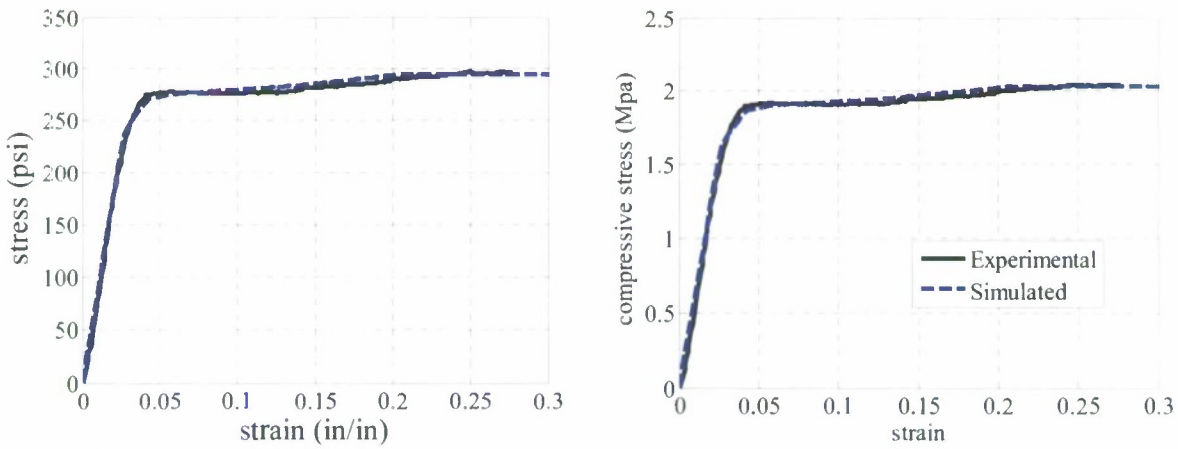


Figure 11. Compression stress strain response (experimental and simulations)

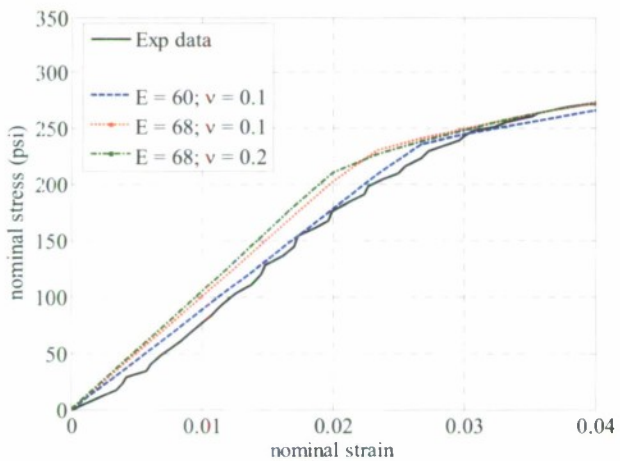


Figure 12. Variations in response with different E and nu

Since the type of hardening response (volumetric or isotropic) doesn't influence the compressive response, same simulation results are obtained with isotropic hardening model instead of volumetric hardening model. Parameters that need to be defined for isotropic hardening model are k_c and plastic poisons ratio apart from the usual definitions of modulus of elasticity, poisons ratio and hardening envelope. Parameter k_c for the isotropic hardening model is kept at 2.0 while plastic poisons ratio is chosen as 0. It should be noted that nonassociated plasticity is chosen to represent the foam behavior.
It should be noted here that even though the calibrated model performs well in monotonic loading, will not provide a good response in cyclic loading since damage cannot be captured with this model; for which a damage plasticity model would be required.
Simulation with the above properties was carried out to determine the tensile response of the foam model. Based on experimental observations of tensile and compressive response (in figures 2 and 4), it can be observed that there exists a difference between the elastic modulus in tension and that in compression. Moreover, since a plasticity type model was chosen, there exists a yield plateau which in reality is not true typically for the tensile response. In reality the foam demonstrates a brittle type response in tension. It should also be noted that the magnitude of the tensile strength of the PVC foam is nearly the same as that of the magnitude of the compressive yield strength. All these represent the limitation of the model used for the foam material and will be addressed in later documents. If volumetric hardening model with the above parameters was chosen the following perfectly plastic simulation is observed as shown below. The volumetric hardening model response for tension is obvious since the yield surface and the evolution shows that the numerical model is perfectly plastic in nature. On the other hand, an isotropic hardening model would result in a trilinear response as shown below. Variations are also shown for different k_t and k_c values for the volumetric hardening model and for different values of plastic poisons ratio for the isotropic hardening model. It should be pointed out here that the correct value of k_c was not utilized for the plastic poisons ratio variation in isotropic hardening model which results in lower yield strength.

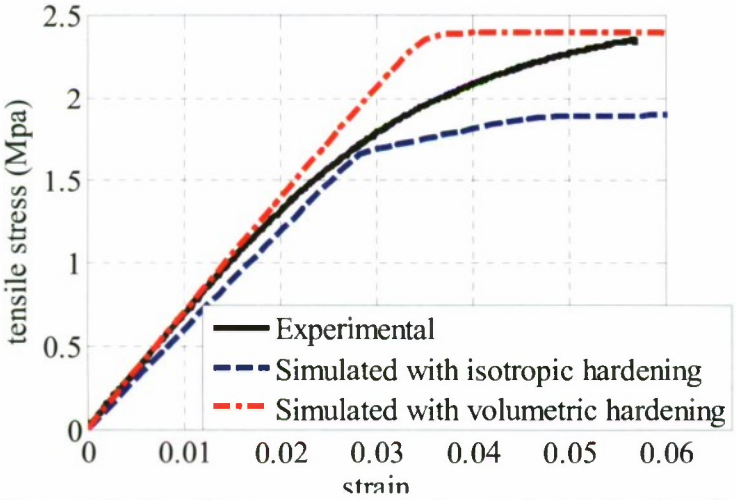


Figure 13. Tensile response (experimental and simulated)

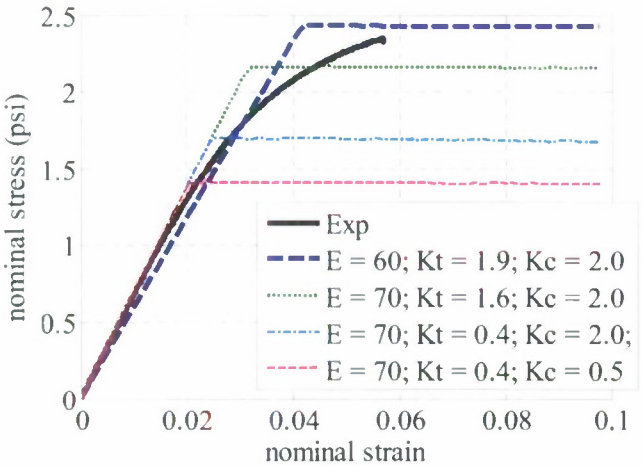


Figure 14. Variation in tensile simulated response with different E, Kt and Kc (volumetric hardening)

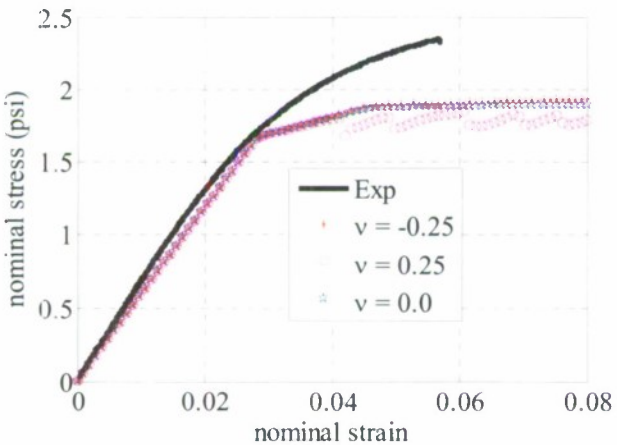


Figure 15. Variation in tensile response with different plastic poisons ratio (isotopic hardening)

Simulations with the above properties were also carried out for shear loading on a foam sample. Correlation observed between observed sample and simulated sample with volumetric and hardening model is shown in figure 16. For a volumetric hardening model, a perfect plasticity type model is obtained whereas a hardening type of response in shear was obtained with the isotropic hardening model. It should be noted that in reality a softening type of response is observed after initial hardening which is due to the cracks formed in the foam. However since hardening type plasticity constitutive law has been provided, this effect cannot be captured.

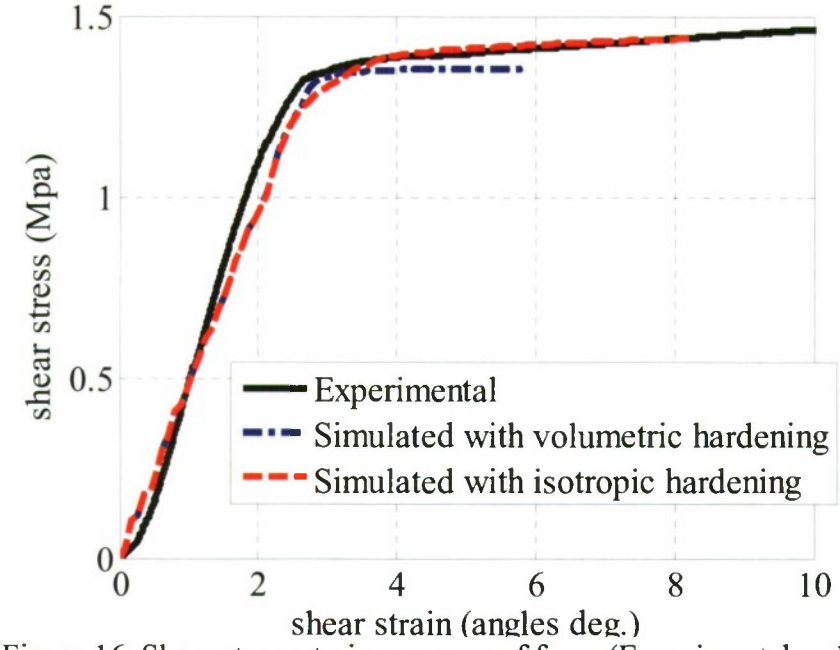


Figure 16. Shear stress strain response of foam (Experimental and simulated)

Variations in response of the volumetric hardening model were obtained with changes in modulus of elasticity along with k_t and k_c values. The following figure shows also shows the shear test raw data measured with the cross head displacement and the LVDT. It was observed that the elastic slope varied significantly for the two measurements with the LVDT results being nearly twice stiffer than the cross head results (1.7 times to be exact). Approximately and average plot between the two responses was obtained for purpose of comparison with simulation. This difference in result is due to the inclination of the shear jig as well as due to machine compliance.

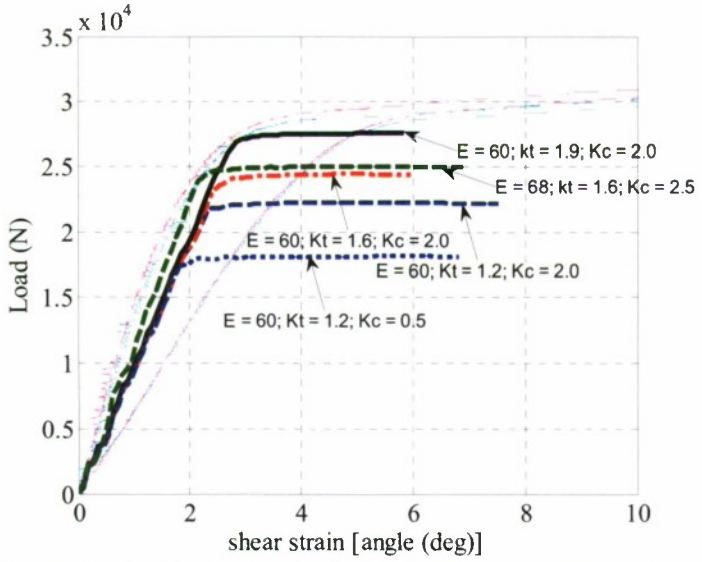


Figure 17. Variations with E, Kt, kc (volumetric hardening)

Manufacture of the sandwich composite panel

The process of manufacture employed in this project is vacuum resin infusion (figure 18). In this process the foam (core material) is laid in between layers of glass fibre skins. Each of the glass fibre skins on either side of the foam consist of two alternate layers of chopped strand mat (CSM) and woven roving (WR). Peel ply is then laid on top of the skins followed by a breather cloth. The entire sample is then bagged and air in the bag is extracted by means of a vacuum pump. The layup of the entire process is shown in figure 19.

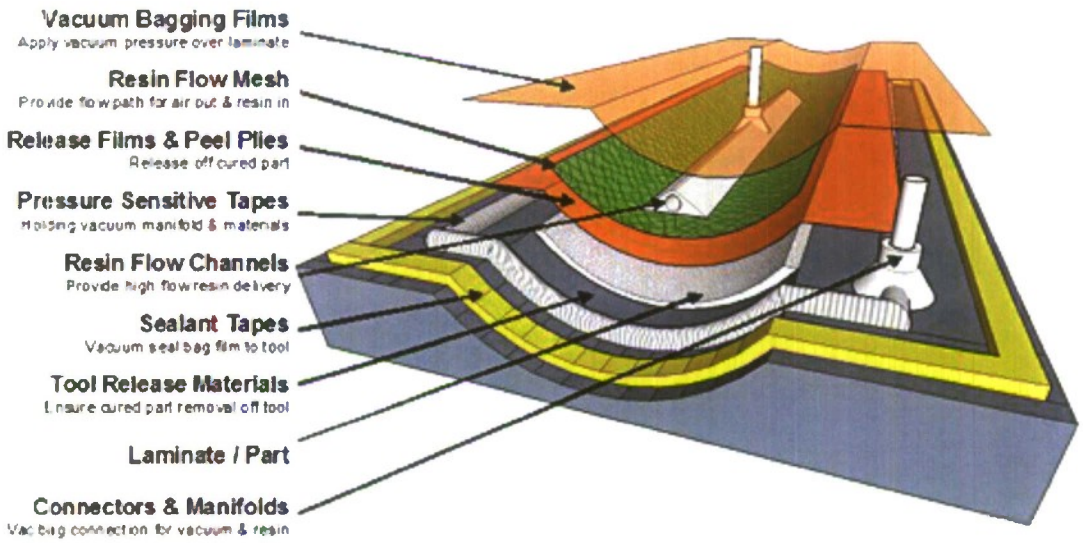


Figure 18. Vacuum resin infusion process

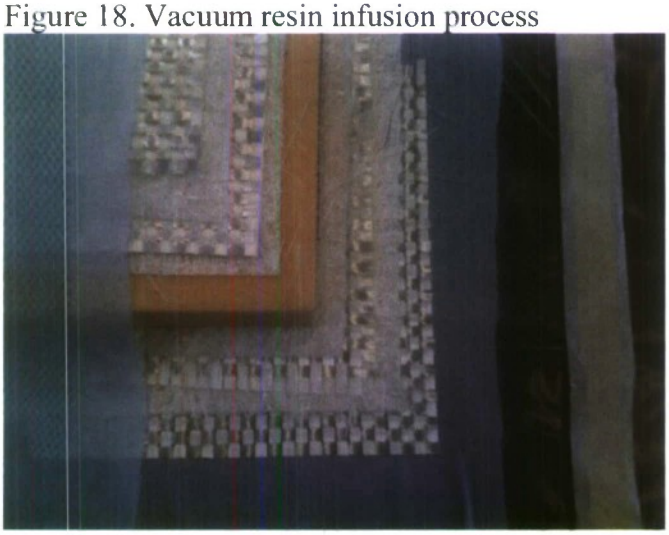


Figure 19. Layup for the VRI process

The total amount (weight) of resin with hardener used for the process of manufacture is taken to be the same as the weight of the fiberglass laminates so as to ensure a 0.5 weight fraction of resin. A 5:1 proportion of resin and hardener is used respectively for this process as per guidelines of the West Systems epoxy system. The resin-hardener mixture is then distributed over the entire layup through T-fittings and vinyl tubing on one side

connected to the resin-hardener mixture container. The distribution is done by creating vacuum with a vacuum pump connected through tubes on opposite side of the resin-hardener feeding side. Figure 20 shows the resin flow and VRI process being performed on the sample. Approximately it takes around 9-10 hrs at room temperature for curing of the sample.

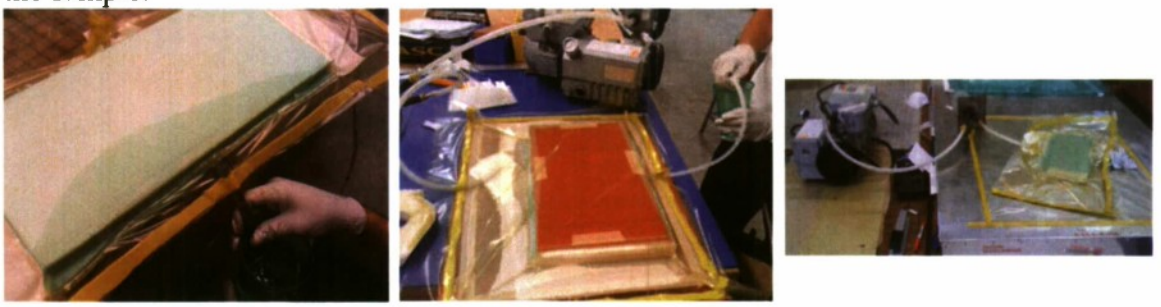


Figure 20. VRI being performed on the sandwich composite

The detailed process of VRI is given in an attached report prepared by Michael Jacobson for an AERO class in the appendix.

Experimental shear tests on conventional composite sandwich panels

The final product obtained after the vacuum resin infusion process is done is shown in the figure 21

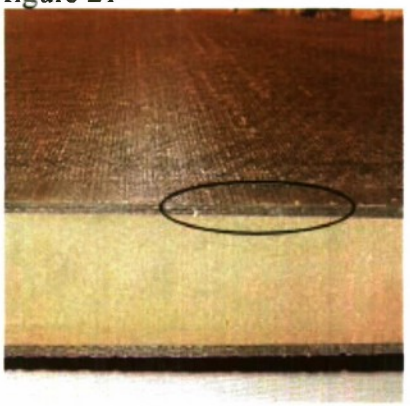


Figure 21. Finished conventional sandwich composite specimen

The final sandwich samples are cut into 16 by 2 inch pieces and are glued to the shear jig by structural glue. The properties of the structural glue have been previously described while explaining about the shear tests on foams. Failure was observed to originate at the junction of the fibreglass skin and the core. Small cracks originated at yield of the samples and a near complete delamination of the skin from the core was observed at the ultimate load. The failure observed was primarily an adhesive debond failure as can be seen from the debonded surfaces of the skin and the core in figure 22. However, it was also observed that if there were manufacturing errors such as improper gluing of the sandwich composite sample with the shear jig, such that the structural glue instead of just being attached to the top and the bottom skin surface also drips along the side and gets in contact with the foam and hardens in certain regions, then a complete delamination of the

skin from the core is not observed instead a shear failure is observed in the foam which then propagates into the skin region as shown in figure 23.

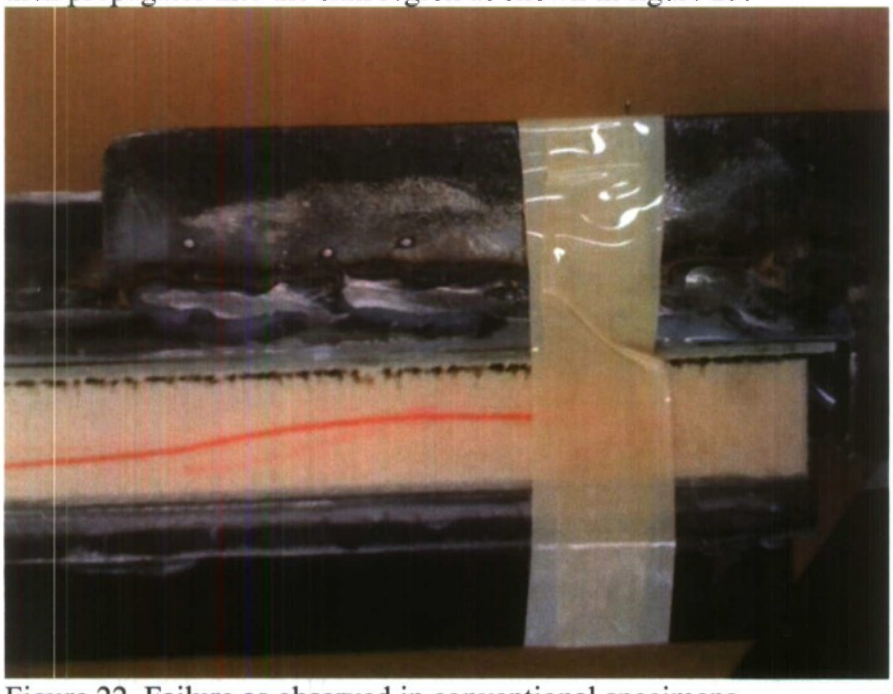


Figure 22. Failure as observed in conventional specimens



Figure 23. Failure observed in conventional specimen with improper bonding to the jig

It was also observed that the stress strain response of the sample varied depending on how the samples were manufactured and then cut for the test. Once the VRI process was done, the samples could be cut either along the direction of flow of resin or direction perpendicular to the direction of flow of resin. It was observed that higher shear stresses (approximately 16% increase) was observed for samples which were cut along the

direction of the flow of resin in comparison to the samples cut perpendicular to the direction of resin flow.

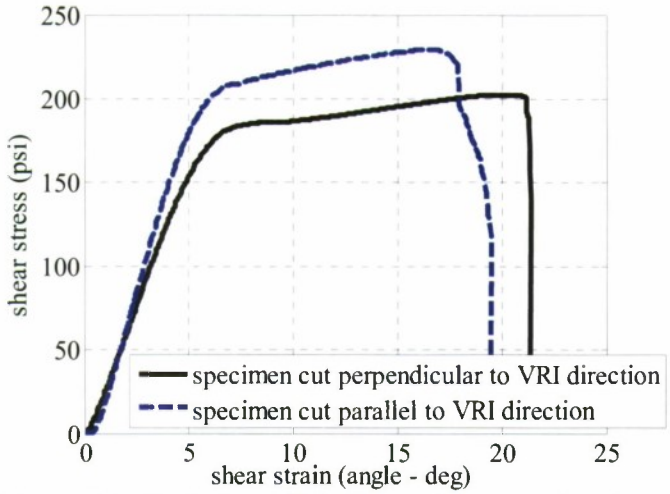


Figure 24. Variation of shear response considering direction of specimen cut with respect to direction of resin flow

Manufacture of the sandwich composite panel with shear keys

Theoretically any material can be provided as a shear key. The advantage of the shear key would be to help prevent adhesive debonding as well as provide better strength and stiffness when loaded in in-plane shear. The manufacturing techniques of inserting in glass fiber shear keys in the sample are described in the following paragraph. A CNC machine or a manual mill was utilized to insert grooves in the foam in both the top and bottom side as shown in figure 25. A CNC machine can be programmed to create any shape, size and spacing of the grooves based on availability of drill bits. Semicircular grooves of diameter 8 mm were used in this project.

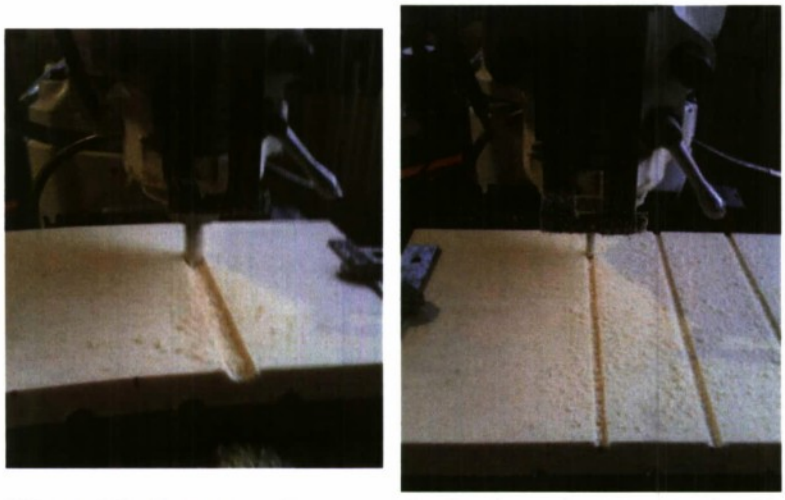


Figure 25. Creation of grooves in the foam by a manual mill

A number of different materials and techniques were utilized to fill up the grooves in the foam.

- Process 1 – The first layer of CSM in contact with the foam was inserted into the groove while laying the fibreglass layers for vacuum resin infusion process. The WR layer was then laid over the grooves in the foam. The grooves were then filled up by CSM strands and two flat layers of CSM and WR were laid on top of it so as to create a flat surface. Vacuum resin infusion was done along the direction of thc grooves to manufacture the samples.
- Process 2 – WR strands were utilized along with epoxy resin by the VRI process to make semi circular prepregs within manufactured aluminium/foam molds of same dimension. These semicircular prepregs were then inserted into the foam-core grooves. The picture of the aluminium/foam mold and the process of manufacture is shown in figure 26.



Figure 26. Aluminum and Foam molds for glass–fiber prepregs; VRI process of manufacture of the prepreg molds

After filling up the foam-core grooves with glass-fiber with any-one of the above process mentioned, vacuum resin infusion process is done to manufacture the sandwich composite. A final product done using process 2 is shown in figure 27. It should be noted that process 2 produced better and consistent results in comparison to process 1. In process 1 there are more possibilities of developing voids and or resin rich area without fibers at the interface region which would result in initiation of cracks at the interface of the notch and the foam. Thereby, process 2 is being recommended as the process of manufacture of the samples with shear keys.

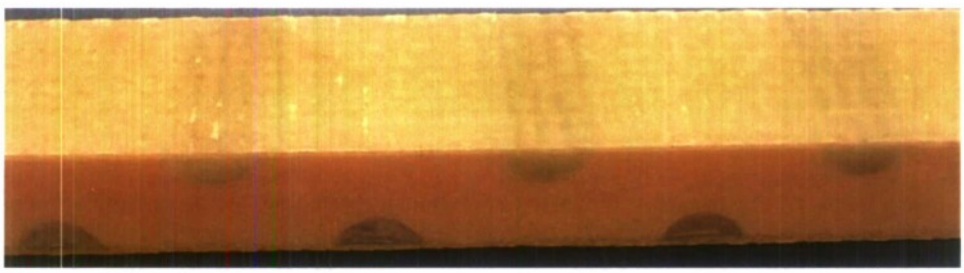


Figure 27. A finished sandwich composite specimen with shear keys

Different types of samples were developed utilizing different shape, size and material for filling up the grooves. The samples were then loaded in inplane shear and responses obtained were compared to that obtained from the conventional sandwich composite

manufacturing process. It should be noted that comparisons are being made with conventional sample in which the specimen is cut perpendicular to the VRI direction since that's the direction in which VRI is applied for the new sandwich composite constructions with shear keys. For samples with shear keys, VRI could only be applied in direction perpendicular to the cut to ensure uniformity of the shear key bonding between the glass-fiber and the foam.

A strength increase of approximately 25% is observed with introduction of 8 mm diameter semicircular grooves filled with prepreg glass fiber manufactured as per process 2 (see figure 28). The arrangement of the grooves is such that the bottom shear-key are centrally placed between two top shear-keys.

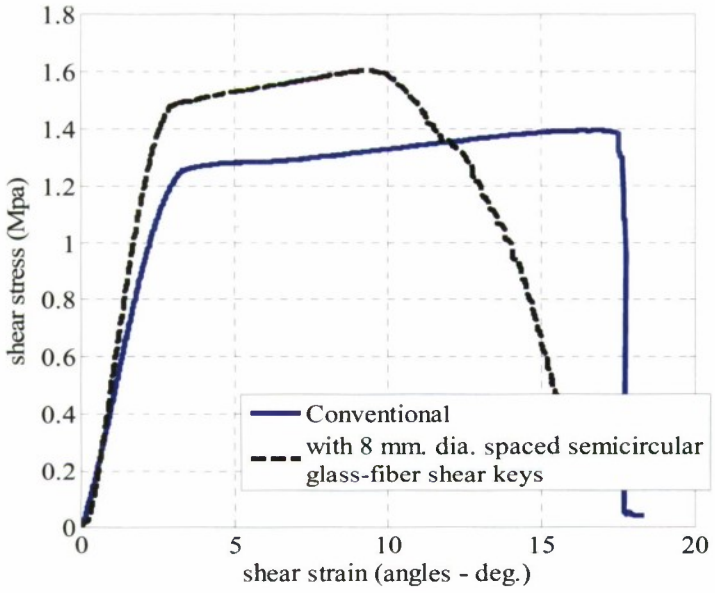


Figure 28. Comparison of mechanical performance in shear for conventional and shear key samples

The failure observed in sample with shear keys is shown in figure 29. Even though the cracks originate at the interface of the composite skin and the foam-core, the cracks deviate at the shear key locations and a shear failure is observed within the foam core resulting in open cracks through the foam core depth or diagonal shear cracks. It should be noted in this context that this method of introduction of shear keys can be utilized in conjunction with other established methods of mechanical performance improvement such as stitching and z-pinning. However, the mechanism of adding in shear key to sandwich composites is a robust, cost effective means by which the mechanical performance (especially the shear capacity) can be increased. Before establishment of this methodology as a viable alternative to stitching and z-pinning, more research on this methodology is warranted.

Typically, the initiation of cracks at the intersection of the shear key in the grooves to the foam is a result of high stress concentration in that region. This can properly be controlled with better manufacturing and also creating pre-grooved foams rather than creating groves within the foams after they have been made. The finite element study

shows the differences in the stress pattern with and without the grooves filled with shear keys.

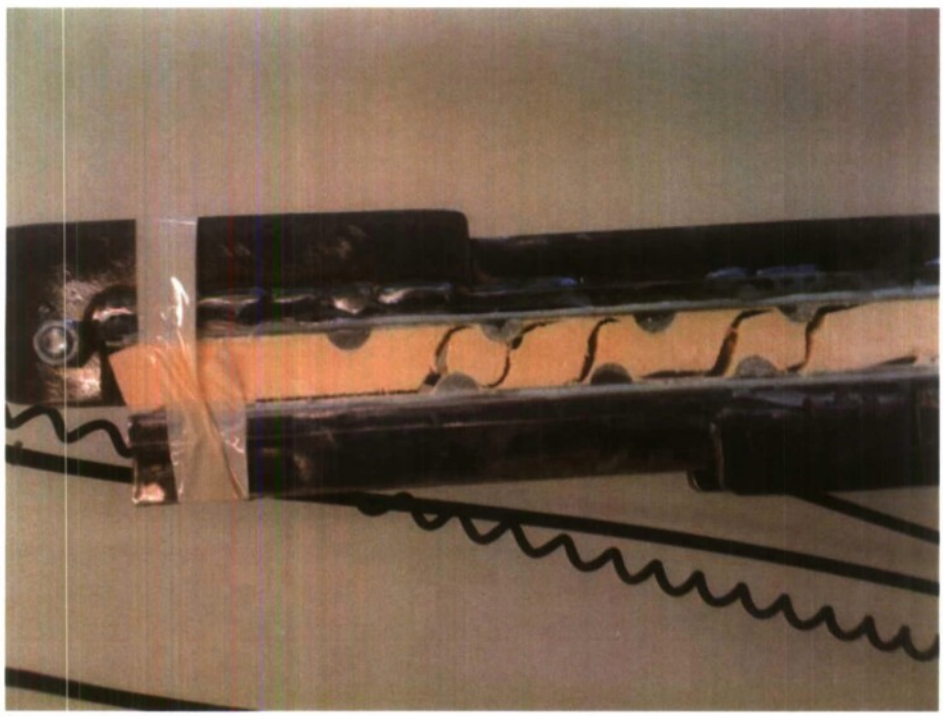


Figure 29. Failure observed in samples with shear keys loaded in Inplane shear

Even though the gain in shear strength was significant with this novel methodology, it was observed that the shear keys were rigid members. Now there is a sharp change in stiffness from the foam to the glass-fiber shear keys and thereby there is a potentiality of stress concentration in those regions. In order to prevent this, it was thought of as how to provide less stiffness to that of the shear keys so that they are slightly deformable. Thereby, balsa wood was provided in the groves to act as shear keys. Gain in strength was obtained but was not that significant in comparison to the previous methodology, as shown in figure 30. It was also observed that the ductility of the specimen was also comparatively lower with the balsa specimen in comparison to the prepreg glass fiber shear keys since the balsa wood itself was cracked. It should be noted at this point the density of balsa wood utilized is 150 kg/m^3, which is not significantly different from the foam sample being utilized which is 100 kg/m^3.

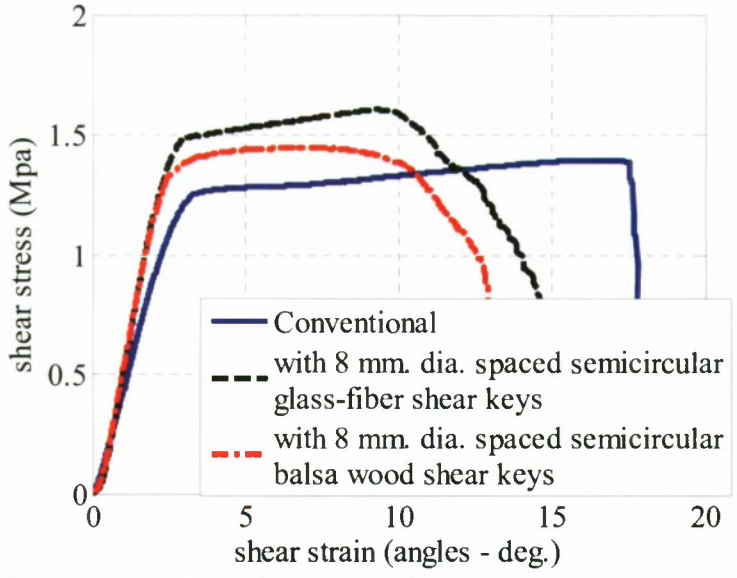


Figure 30. Comparison with balsa wood shear keys, glass fiber shear keys and conventional specimen

Experimental investigations were also carried out for samples in which top and bottom shear keys are just placed one above the other without spacing them such that the top keys are placed in between two bottom keys and vice versa. It was observed that with spacing gave better behavioral response with regards to strength and ductility in comparison to the sample without the spacing. With top and bottom shear keys placed one above the other, a necking of the foam core material is observed which has serious consequences with regards to failure. It can be observed from Figure 31 that without spacing samples did have a lower strength and also failed faster in comparison to the spaced ones, which is obvious.

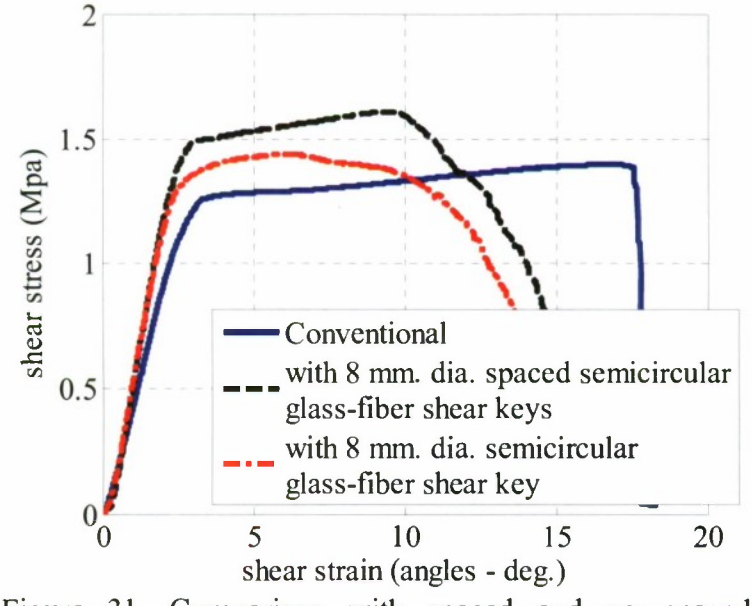


Figure 31. Comparison with spaced and non-spaced top and bottom shear key orientations

Experimental investigations were also performed to determine the effect of variation of shape of the grooves in the foam. One of the grooves in the sample was made as a V notch. Even though the strength of the two samples was comparable, however the V notch sample showed rapid degradation of strength in comparison to the other semicircular sample, as shown in Figure 32.

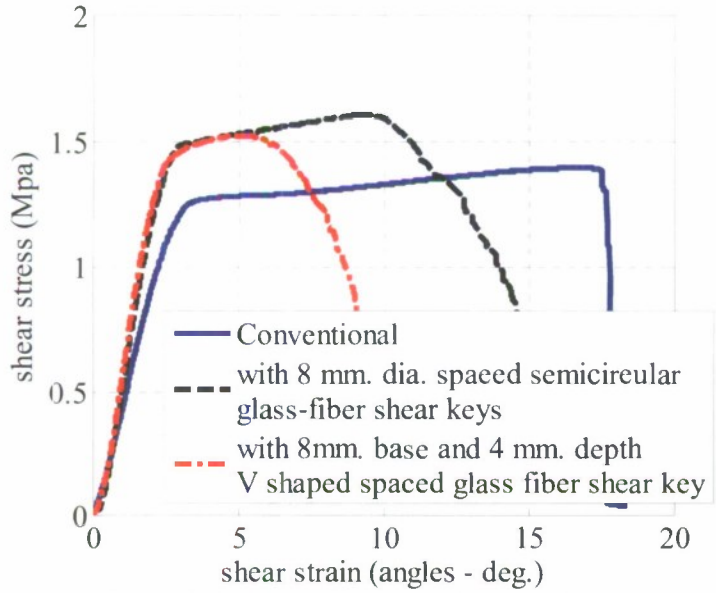


Figure 32. Variation with different shapes of the foam-core grooves for shear keys

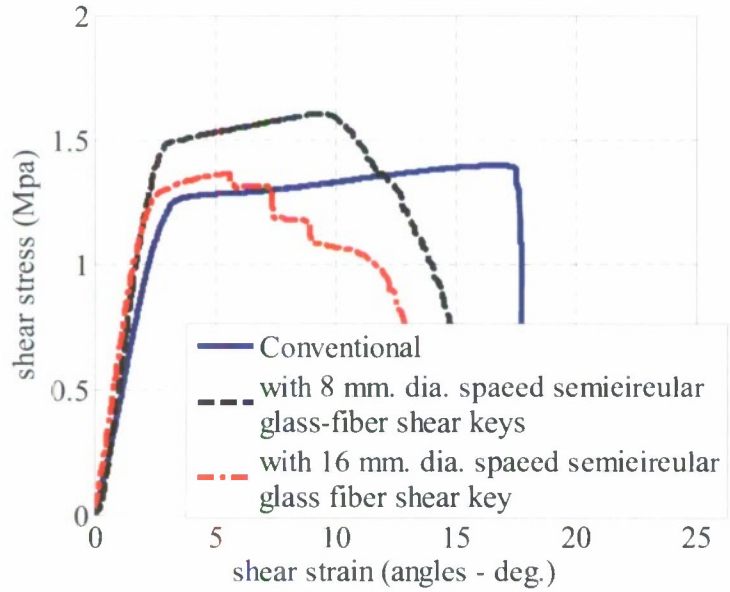


Figure 33. Variation with different sizes of shear keys

Experimental investigation was also carried out for different size of shear keys (Figure 33). The one in whieh it was 16mm spaced sample, has very low amount of foam in between the two faces and thereby demonstrates a faster failure process. This elucidates that there is a limitation to the ratio of the shear key depth to the foam depth. The current

investigation just highlights the limitation without being able to recommend the limit. More experiments will be required to establish the limit.

Finite element study

The simulated discretized specimen with loading and boundary condition is shown in figure 34. Each of the four layers in the top and bottom skin along with the foam core was modelled using continuum linear brick elements. The interface between the skin and the core was modelled as a perfect surface-to-surface tied interface. A perfect tied interface was also utilized to model connection of the sandwich composite panel with the steel shear jigs. As shown in boundary condition in figure 34, one end of the steel shear jig was fixed against translation while a uniform displacement was applied at the other end.

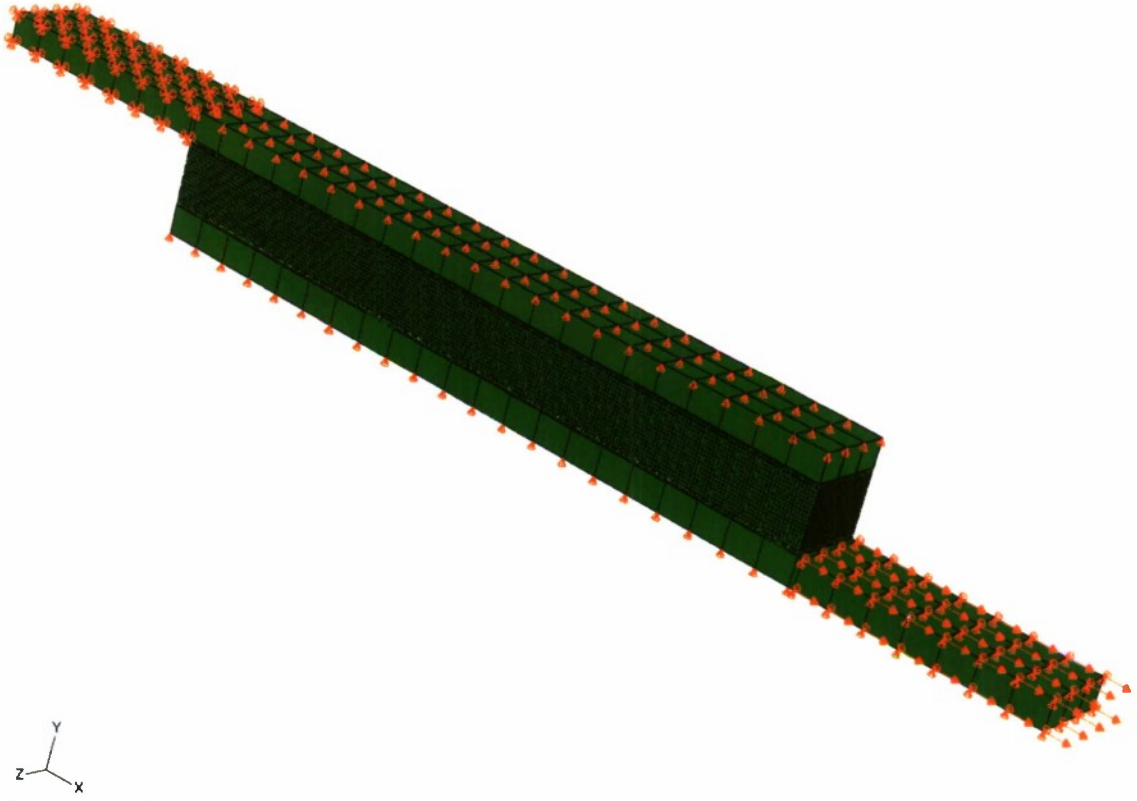


Figure 34. Simulated specimen with loading and boundary conditions

Good correlation with regards to initial stiffness in the elastic regime could be observed between simulated and experimentally observed composite sandwich composite panel specimen (as shown in figure 35). Differences are observed in the post-yield plateau region for the experimental and simulated specimen because of limitations in the simulated model. In the experimental model, cracks originate at the interface between the skin and the core and slipping is initiated at the interface region; whereas the interface in the simulated model is idealized as a perfect bond. This explains the lower post-yield stress of the experimental model in comparison to the simulated model. At load increments near the yield branch of the simulated global shear stress-strain, a diagonal

distribution of normal stresses (σ_{11}) are observed in the specimen which represent a diagonal shear stress failure pattern; the stress component σ_{22} was observed to be significantly less that the yield tensile stress of the foam, signifying that the crack pattern as observed in figure 23 of experimental investigations is a result of bad manufacturing rather than the failure phenomenon. A low value σ_{22} in the foam also signifies that the failure was purely an adhesive type of failure and the response obtained in figure 22 is a correct representation of the behavioral response. Complex damage modeling of tangential slip between the two surfaces in contact is beyond the scope of the current manuscript and would be addressed in later publications. Typical stress distributions observed within the foam in simulations of the conventional sample are shown in figure 36.

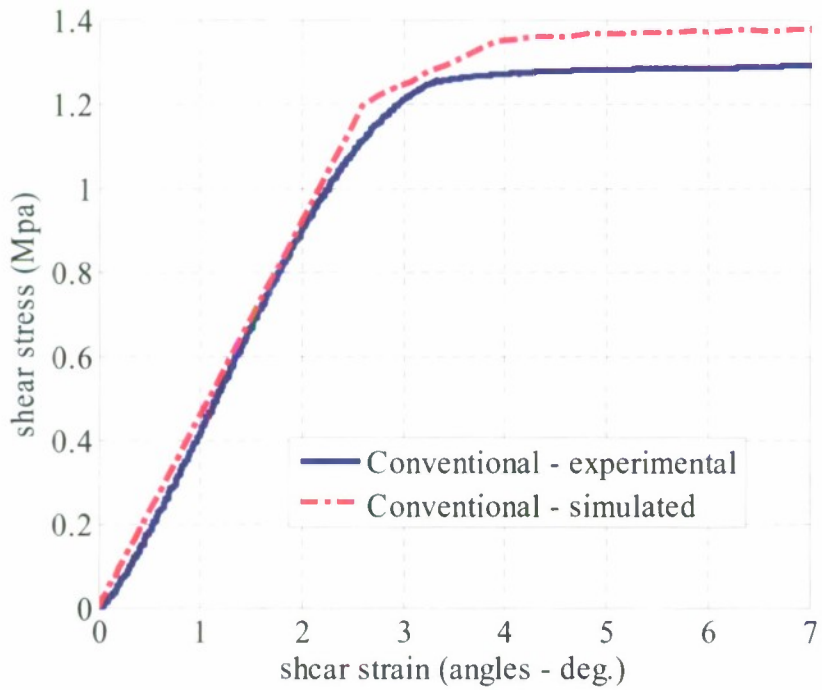


Figure 35. Comparison of the global response of the conventional composite sandwich panel

Simulated sample with shear keys are shown in figure 37. An increase in global initial stiffness with introduction of shear keys is observed both in experimental and numerical simulations as shown in figure 38. Typical stress distribution of the specimen with shear keys is shown in figure 39. It was observed that one edge of the foam grooves were subjected to compressive loads and the opposite edge were subjected to in tension. High stress concentrations (σ_{11} greater than the tensile strength) are observed at the regions near to the grooves from where the cracks originate. Similar behavior has also been observed in cracks observed from experimental investigations in figure 29. Figure 29 also demonstrate cracks observed at the interface region of the skin and the core; however since a fully tied interface is modeled in the simulated specimen this feature could not be captured in the simulated observations. Plastic strain (PEEQ) plots (figure 40) for the shear key specimen demonstrates development of plasticity within the foam region. It should be noted that no plasticity response could be obtained in the conventional

specimen. This demonstrates that the plastic nature of foam, not utilized in the conventional methodology, could be utilized in the proposed approach of introduction of shear keys.

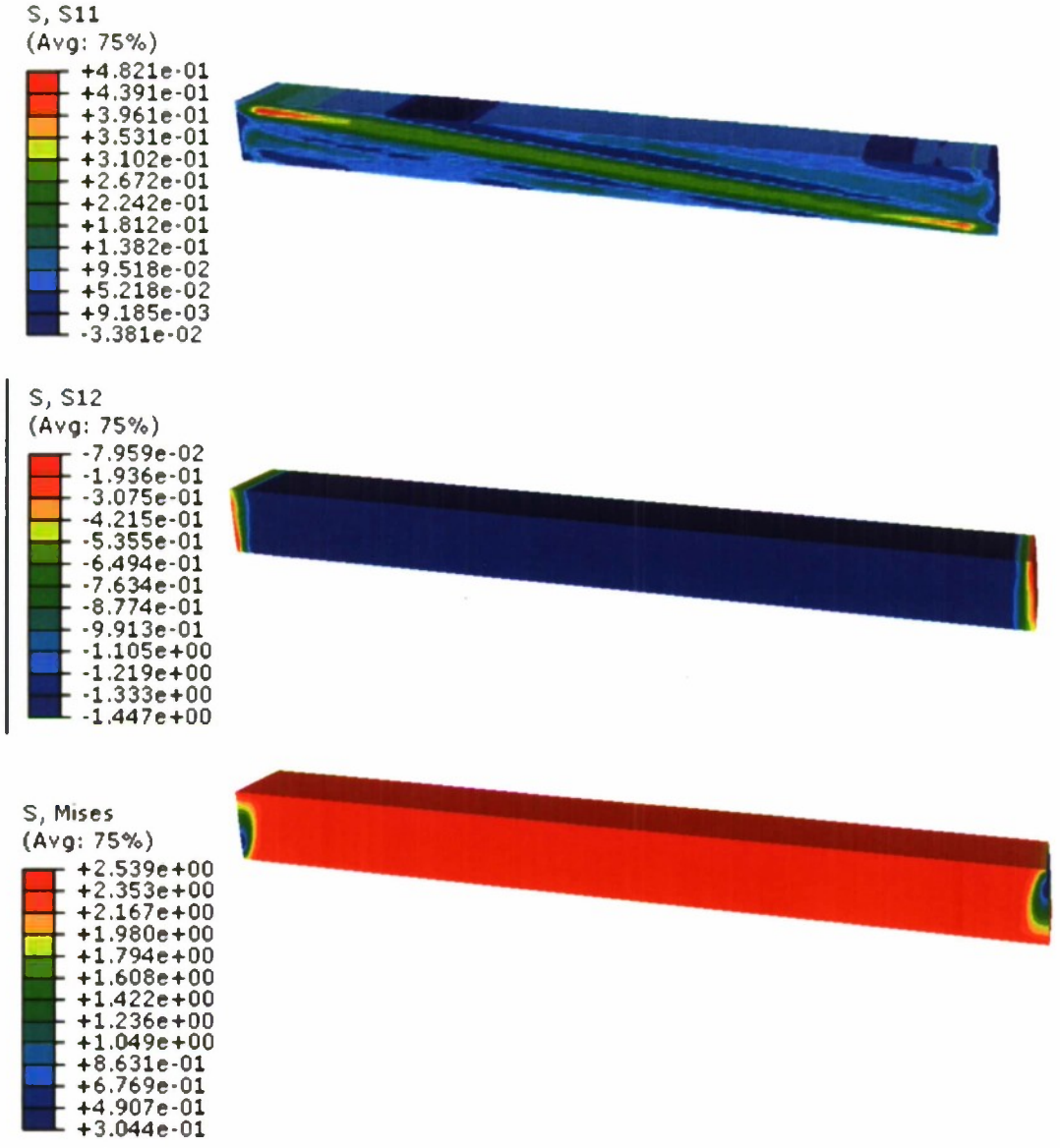
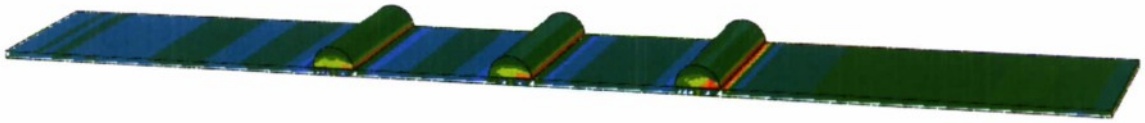


Figure 36. Typical stress distribution in the foam for the conventional specimen



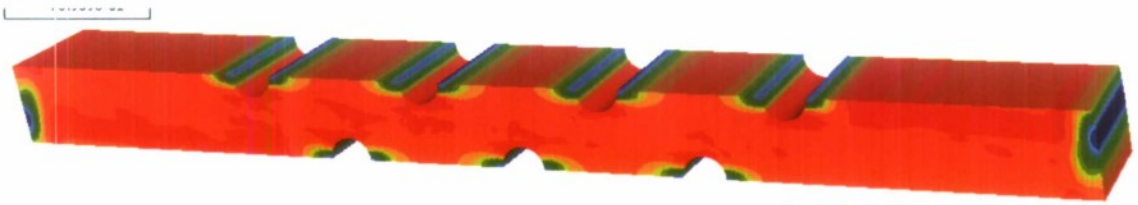


Figure 37. Simulated sample with shear keys

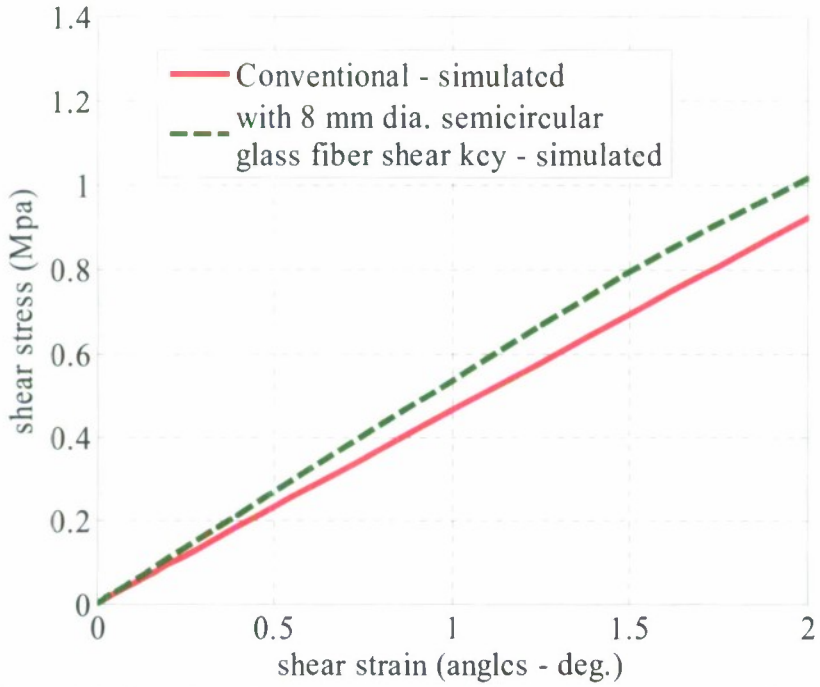


Figure 38. Initial stiffness increase with introduction of shear keys

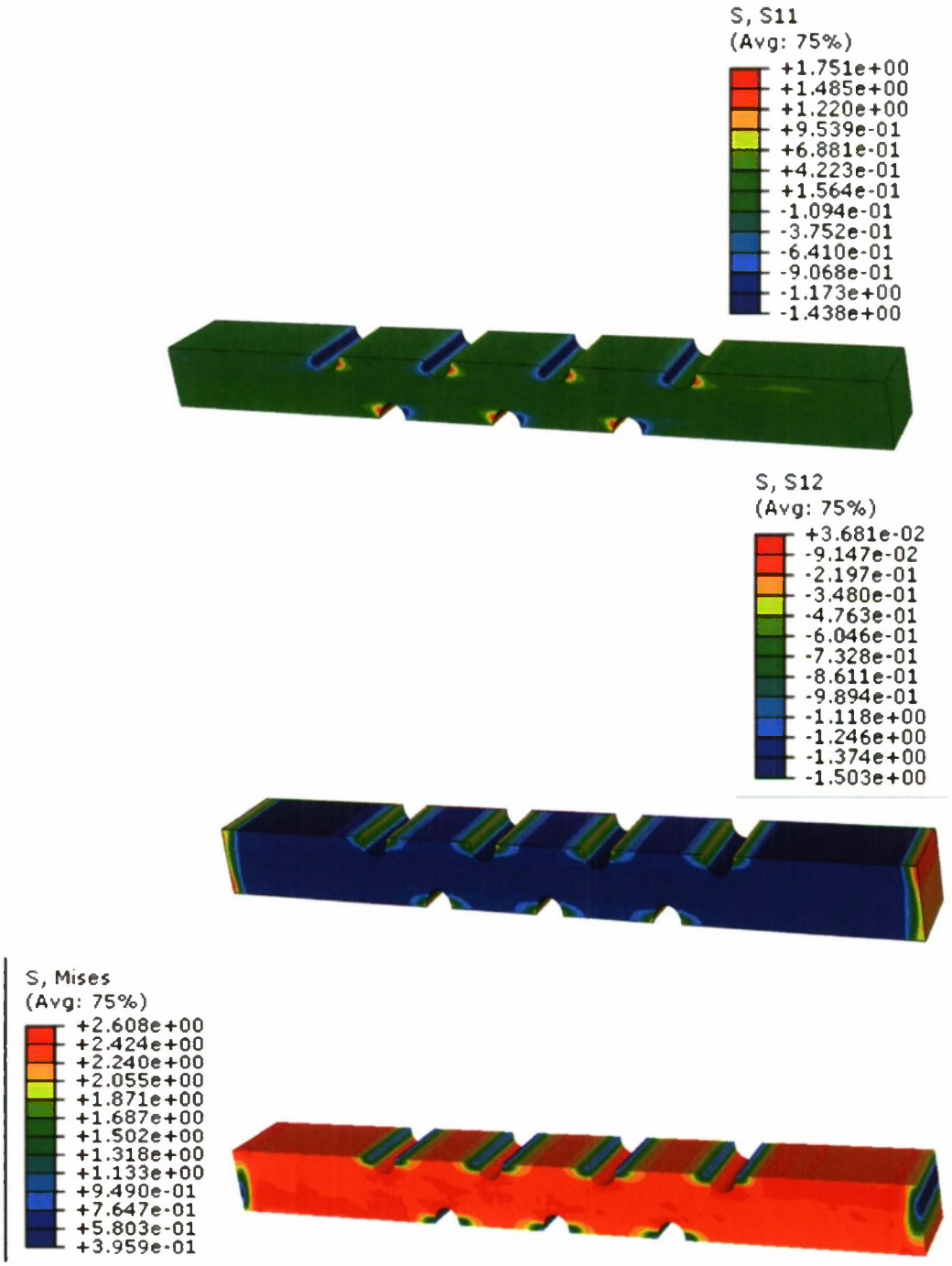


Figure 39. Stress distribution in sample with shear keys

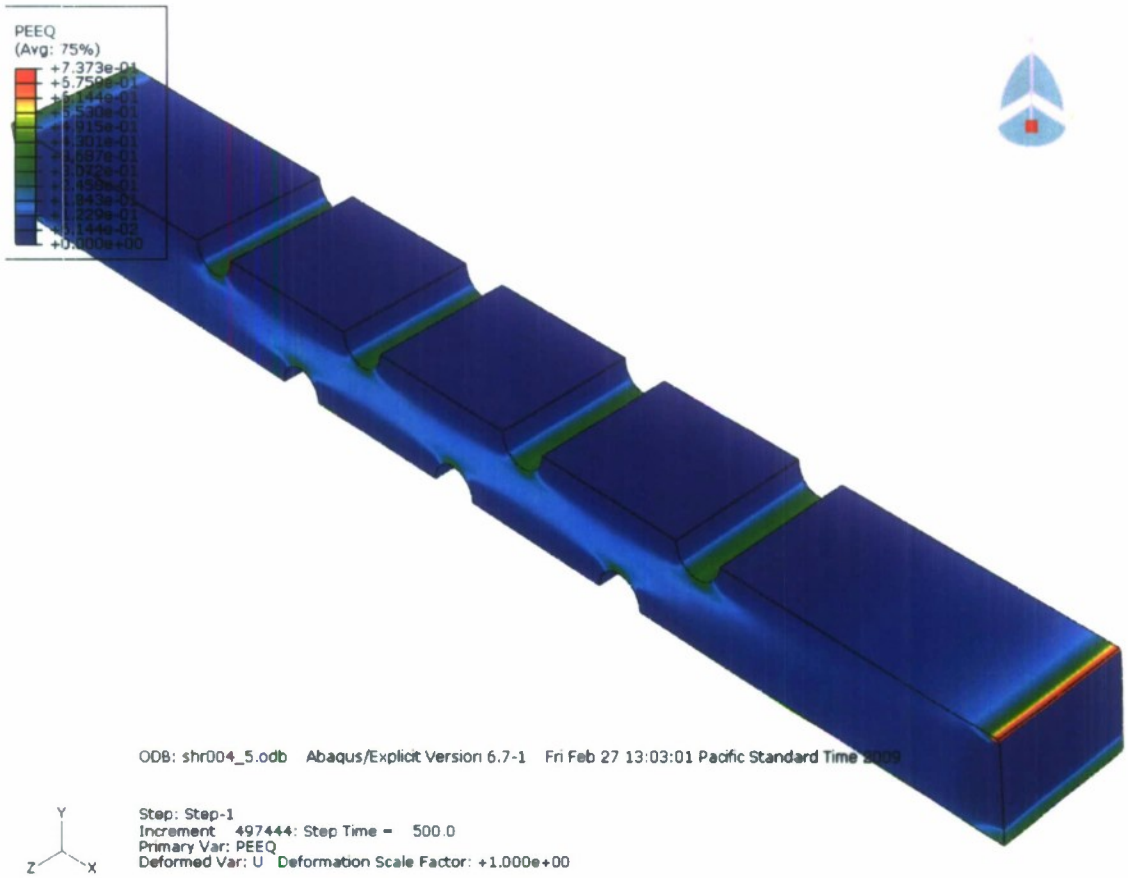


Figure 40. Equivalent plastic strain plot of the grooved foam in proposed methodology

Concluding remarks

The manuscript demonstrates that introduction of shear keys in a conventional sandwich sample is a cost effective novel methodology which results in increase in initial stiffness and strength of the inplane shear response of composite sandwich panels. More research is required to establish this as a viable technique similar to stitching and z-pinning. Numerical simulations were performed and areas of improvement such as constitutive modeling of foam and damage modeling of the interface were identified for future work. It is being hypothesized that this mechanism of adding in "shear keys" would also be useful as a "peel stopper" mechanism to prevent delamination of skin from the core. To demonstrate the feasibility of the hypothesis a number of experimental investigations have been devised and currently being pursued.

U.S. –China Collaborative Soil-Structure –Interaction Research

Project Investigator:
Robb Moss
Department of Civil and Environmental Engineering
California Polytechnic State University
San Luis Obispo, CA

U.S.-China Collaborative Soil-Structure-Interaction Research

By PI Robb Moss, PhD, PE

Introduction

This research involves shake table testing of scale soil-structure models that mimic the coupled seismic response of underground structures and the soil surrounding the underground structures (termed soil-structural-interaction or SSI). Currently the design of subways and other critical underground infrastructure relies on little to no empirical data for calibrating numerical simulations. This research is working towards filling that empirical data gap. This project is also being used as a vehicle to establish a joint testing program between Nanjing University of Technology (NJUT) and Cal Poly. Professor Zhihua Wang from NJUT is participating in the project both here at Cal Poly as a visiting researcher and remotely from Nanjing. The scale model testing equipment designed for this type of shake table testing was provided, via an indefinite loan agreement, by U.C. Berkeley. The long term objective is to fully develop a scale model testing platform for evaluating seismic stability of all manner of critical underground and above ground infrastructure. Seismic stability in most cases can be related to post-impact or explosive force stability assessment which is an immediate interest for military and national security.

Project Significance

There are many poorly understood seismic issues associated with critical infrastructure in seismic areas of the U.S. and China. The U.S. has aging infrastructure such as bridges, subways, and buildings that were designed based on older seismic criteria that do not necessarily capture the full dynamic response that is now anticipated. The U.S. also has new infrastructure being planned or built that may be limited in the scope of design because of unanswered seismic-soil-interaction (SSI) questions. China is trying to keep pace with its rapidly developing economy by building infrastructure at a frantic pace. However the seismic design understanding and seismic codes are not necessarily keeping up with the pace of development. This research seeks to establish a parallel testing platform that could be used simultaneously by researchers at Cal Poly and at Nanjing University of Technology (NJUT), Cal Poly's sister university in the Jiangsu province of China while addressing the seismic research needs.

Recent research has shown that there is uncertainty in the dynamic response of soil sites (Bazzurro and Cornell 2004) and the coupled response of structure above and below the ground surface and the surrounding or supporting soils (Hashash et al. 2001; Stewart et al. 1999). Some examples of U.S. infrastructure that warrant SSI research include: elevated highways, underground light rail and subways, bridges, overpasses, water canals, water supply tunnels, pipelines, levee systems, and dams. Research into the dynamic response of U.S infrastructure would mainly address seismic integrity, seismic hazard mitigation, and seismic retrofit. In China the infrastructure that warrants SSI research is similar in scope but generally dealing with initial planning and design.

The provincial government of Jiangsu province has declared Geotechnical Engineering as the primary research focus of the next decade (initiated in 2006). This is due to the large amount of infrastructure being planned and built (bridges, subways, highways, large buildings, underground facilities) to accommodate the rapid development in the province. Research money is being pumped into the provincial universities with NJUT garnering a large portion

because of their reputation as an outstanding geotechnical research facility. The NJUT geotechnical group is in turn reaching out to U.S. researchers to augment their research expertise in order to "jump start" their efforts. Collaboration between NJUT and Cal Poly was established within the last four years with emphasis placed on earthquake engineering which is a strong common asset at both institutes. This proposal is the culmination of recent efforts to initiate research exchange that will benefit both the universities and their supporting communities.

Testing Platform

In physical testing, and scale model testing in particular, the testing equipment and physical model details can demand the bulk of the research effort and this project is no exception. The first year of this project was spent acquiring the necessary materials/supplies, building/modifying/manufacturing the testing equipment, and calibrating the testing platform to achieve the desired results. To carry out scale model tests on the shake table, similitude analysis dictates the scaling of important variables like dynamic soil strength, dynamic structural response, and associated displacements. The scaling analysis of the soil and structural elements will follow the research by Meymand (1998) which used the San Francisco-Oakland Bay Bridge and Bay Mud as the prototype structure and soil. PI Robb Moss was involved in this research at Berkeley during his doctoral work and has other experience running scale model testing from his Master's research.

The testing platform used for this research was acquired, on indefinite loan, from UC Berkeley. The main piece of testing equipment is a flexible wall barrel that mimics free field seismic site response when subjected to strong shaking on the shake table. Validation of the testing platform involved comparing analytical results with recorded response from the flexible wall barrel and scaled structural element. Figures 1 and 2 show the validation from Meymand (1998) demonstrating the dynamic performance of the flexible barrel versus other testing containers. As can be seen the flexible wall barrel provides the most accurate representation of seismic soil response with respect to the prototype.

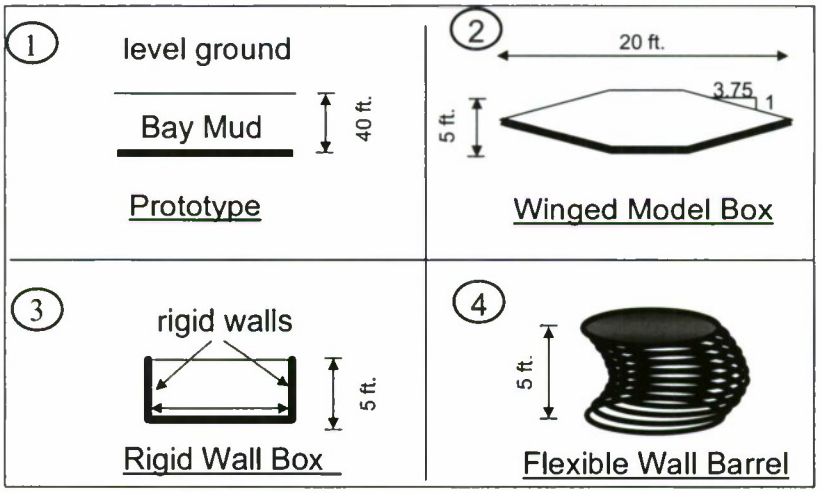


Figure 1. Previous research by Meymand (1998) examined different model soil containers for SSI shake table testing.

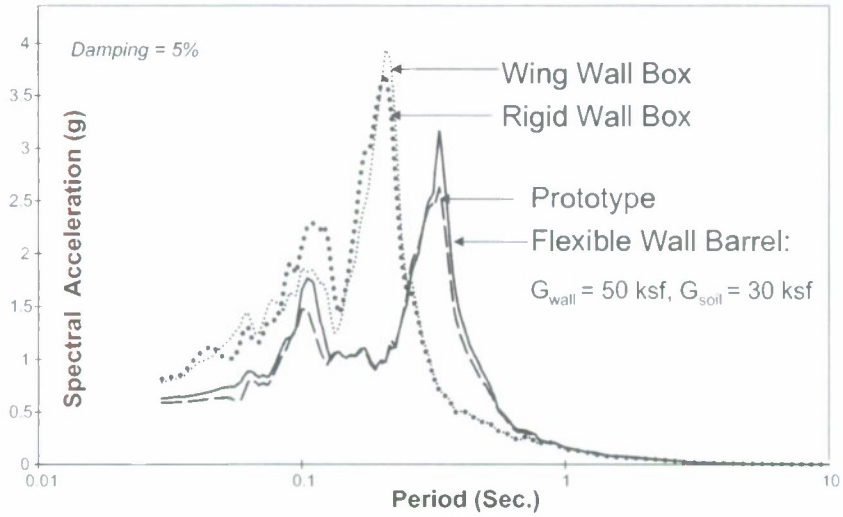


Figure 2. Dynamic analysis of different model soil containers. The spectral response results above show that the flexible wall barrel provides the most realistic response when compared to prototype field conditions.

The flexible wall barrel assembly, associated miscellaneous equipment, and a large volume soil mixer was acquired on indefinite loan from the Richmond Field Station at UC Berkeley. This equipment had been sitting "moth balled" for a decade since PI Robb Moss packed it away following the demise of a state funded research project (a victim of Bay Bridge politics). Once transport of the equipment to Cal Poly was accomplished the equipment was in need of refurbishment and modification to fit the shake table at Cal Poly in the Parsons Earthquake Lab. The equipment was refurbished, supporting equipment purchased and/or manufactured, and the bucket assembled on the shake table. The next step was to begin the manufacture of appropriate soil to run the scale model tests. Figure 3 shows the flexible wall barrel assembled on the shake table awaiting soil. Figure 4 shows the filling of the barrel and Figure 5 shows the full barrel awaiting seismic testing.

The scale model test soil for these tests adheres to similitude analysis to ensure properly scaled response. The geometry scale of these tests will be 10^{th} scale which is equivalent to the overall similitude scale λ. In similitude analysis it is not possible to scale all the physical parameters simultaneously. For this research dynamic strength of the soil was chosen as the primary physical parameter to model and the model soil was designed accordingly. A mix of kaolinite, bentonite, fly ash, and water was used in specific proportions to achieve the desired strength range. The mix used was at 110% water content and the target undrained strength of s_u=4 kPa from a UU (undrained unconfined) triaxial test was used as the guide prior to large volume soil mixing. Once in the barrel T-bar pull out tests and shear wave velocity tests measure the in *situ soil* strength for each phase of the seismic testing.



Figure 3. Testing platform showing the shake table with the flexible wall barrel installed. The flexible wall barrel is composed of the four corner posts with universal joints at the top and bottom, top and bottom rings, and the barrel wall. The wall is composed of a 6.4 mm thick rubber membrane which is confined by 45 mm wide Kevlar straps spaced on center every 60 mm. The (yellow) mixer on the left is used to mix large volumes of model soil (composed of kaolinite, bentonite, fly ash, and water) for filling the barrel.



Figure 4. Process of filling the barrel with scale model soil is shown. Ten accelerometers were placed within the soil in both vertical and horizontal arrays to measure the dynamic response of the soil during seismic shaking.



Figure 5. Full barrel being prepared for initial calibration tests. Notice cross braces are still in place and will be removed prior to testing to allow the flexible barrel free movement in response the imposed seismic shaking.

Phase 1 Tests

The first phase of the seismic testing is to perform free-field tests, that is tests without an embedded structure. This will determine the dynamic response of the soil column without the influence of the underground structure and provide a baseline for evaluating the effects of the soil on the structure. The input ground motions selected for the base input are;

1. 1979 Imperial Valley, El Centro motion
2. 1992 Landers, Joshua Tree motion
3. 1999 Chi Chi, TCU075 motion

These motions were selected specifically to impose large adverse loads on an underground structure and because these were the same motions selected and peer reviewed for analogous tunnel related consulting projects. To adhere to the similitude analysis the time must be scaled at $\lambda^{0.5}$ to provide the correct dynamic response. This means the time step of the ground motions must be compressed to $\Delta t/ \lambda^{0.5}$. These motions were also corrected for full ground reflection because they were recorded at the ground surface and need to be used as base level motions. This was accomplished by modifying the motions using SHAKE (SHAKE91 Idriss and Sun 92) to subtract the full reflection of an "outcrop" motion to render a "within" motion with respect to the prototype soil profile.

Instrumentation for the tests includes 10 accelerometers, 3 displacement transducers, a load cell, and a digital video capture that will be analyzed using image processing techniques to resolve displacement with time. Static soil strength tests conducted before and after shaking tests will include shear wave velocity measurements and T-bar pull out tests.

Numerical modeling of the tests will be accomplished using codes that will model 1-D equivalent linear response, 2-D equivalent linear response with structural elements, and fully nonlinear

response with structural elements. SHAKE (SHAKE91 Idriss and Sun 92) will be used to properly model the 1-D equivalent linear response of the free-field prototype conditions. Results from the accelerometer arrays should agree closely with the SHAKE91 results and will provide calibration for the subsequent numerical modeling.

Phase 2 Tests

An underground section of the BART (Bay Area Rapid Transit) light rail was chosen as the prototype for the SSI tests. This structure is also similar to light rail tunnels being considered in the Jiangsu province. PI Robb Moss has had several consulting projects related to SSI analysis of the BART light rail and most relevant an analysis of an underground design for the so called Warm Springs Extension. The consulting experience on this project and similar projects provides strong guidance on the current research needs.

A scale model structure will be assembled adhering to the similitude scaling of the structural stiffness of the BART tunnel cross section. Of primary research and design interest is the "racking" of the structure or the relative displacement/drift of the top of the tunnel section with respect to the bottom of the tunnel section. This tends to be the critical cross sectional design variable for underground tunnels undergoing seismic SSI in soils. We anticipate instrumenting the model cross section with displacement measurement devices (an LVDT, strain gage, or displacement transducer) to measure these displacements.

Numerical modeling of Phase 2 will use the SHAKE results as a baseline but will use codes that can accommodate structural elements. FLUSH (Lysmer et al. 1975) will be used to perform 2-D equivalent linear analysis with the inclusion of the embedded structure to provide SSI analysis of the prototype soil profile with the subway cross section. The free-field response will be calibrated using SHAKE results and then structural "racking" strains will be documented using this code. A similar numerical analysis will be performed using the fully nonlinear code ABAQUS (Simulia 2009) to capture any highly non-linear response that is missed using an equivalent linear approach.

NJUT Participation

Professor Zhihua Wang from NJUT joined the project spring of 2008 and was a visiting scholar at Cal Poly for over six months. In that time he participated in project meetings and worked on the fully nonlinear numerical analysis using ABAQUS to provide insight into the dynamic response of the soil column and the underground subway cross section. He has since returned to Nanjing but continues to work on the numerical analysis in Nanjing. As testing progresses to Phase 2 our interaction will intensify as the comparison of equivalent linear and nonlinear results will provide very useful information on measured response from the shake table testing. Prof Wang also provided a detailed literature review of SSI shake table testing in China which allows this project to have a comprehensive review of all SSI testing that exists in English and Mandarin.

NJUT is currently developing a concurrent shake table testing program looking at SSI related problems. The interaction between Cal Poly and NJUT will provide strength to both research programs and will allow for easy collaboration between researchers. The long term goal is to have parallel testing that can address specific SSI issues in a concerted manner. Having visiting faculty like Prof Wang will go a long way towards the goal of ongoing collaboration.

Student Researchers

This project currently employs one graduate student researcher and two undergraduate student researchers. The funding will carry the graduate student, Vic Crosariol, through his graduate career here at Cal Poly. There are plans to carry one of the undergraduate researchers, Steven Kuo, over into a masters degree drawing on the valuable training that he has gained through this research. Employment of student researchers is an asset to this project, the College of Engineering, and Cal Poly in general. This form of "learning by doing" is invaluable at many levels and ultimately results in pushing the field of seismic geotechnics forward through innovative research. In addition to research Vic has been able to provide lecturer support acting as an instructor in undergraduate soil mechanics labs.

Summary

This report provides a description of the project funded by ONR through the C^3RP program at Cal Poly. The research objective is to provide empirical data on the soil-structure-interaction between underground structures and the surrounding soil during seismic ground shaking, an area where little to no usable data exists for calibrating numerical analysis. To that end a testing platform has been developed drawing on prior similar research. The testing platform consists of a flexible wall barrel that can fully articulate during excitation on the shake table that is located in the Parsons Earthquake Lab at Cal Poly. Phase 1 of the testing is nearing completion which included the difficult task of acquiring, manufacturing, and assembling the scale model testing platform. Free-field testing of the platform is currently being conducted. Phase 2 of the testing will be to install a scale model underground structure and determine the dynamic response of the structure with respect to the soil. This project is being used as an opportunity to develop a joint testing program between NJUT (Nanjing, China) and Cal Poly. A professor from NJUT has been a visiting scholar on this project and continues to provide support through numerical modeling. This research is focused on seismic response but the results are easily translatable to military and national security issues related to the resilience of critical infrastructure. Ultimately we will have a complete suite of test results showing free-field and SSI (soil-structure-interaction) response. The results will provide a data set for calibrating numerical SSI models and provide a strong basis for justifying "racking" deformations of underground tunnels for seismic design.

References

Bazzurro, P., and Cornell, C. A. (2004). "Ground-Motion Amplification in Nonlinear Soil Site with Uncertain Properties." *Bulletin of Seismological Society of America*, 94(6), 2090-1019.

Hashash, Y. M. A., Hook, J. J., Schmidt, B., and Yao, J. I.-C. (2001). "Seismic design and analysis of underground structures." *Tunnelling and Underground Space Technology*, 125.

Idriss, I.M., and Sun, J.I. (1992) "User's Manual for SHAKE91: A computer program for conducting equivalent linear seismic response analysis of horizontally layered soil deposits."

Lysmer, J., Udaka, T., Tsai, C-F., Seed, H.B. (1975) "FLUSH: a computer program for approximate 3-D analysis of soil-structure interaction problems." UCB/EERC-75/30, Earthquake Engineering Research Center, University of California, Berkeley, 1975-11, 139 pages (555.6/L92/1975)

Meymand, P. J. (1998). "Shaking Table Scale Model Tests of Nonlinear Soil-Pile-Superstructure Interaction in Soft Clay," Ph.D. Dissertation, Civil and Environmental Engineering Department, U.C. Berkeley.

Moss, R. E. S., Thornhill, D., and Nelson, A. (2008). "Preliminary Investigations into the Influence of Geologic Aging on Liquefaction Potential." *Geotechnical Earthquake Engineering and Soil Dynamics Conference*, Sacremento.

Simulia (2009) Providance R.I., http://www.simulia.com/products/abaqus_standard.html.

Stewart, J. P., Seed, H. B., and Fenves, G. L. (1999). "Seismic Soil-Structure Interaction in Buildings. II: Empirical Findings." *Journal of Geotechnical and Geoenvironmental Engineering*, 121(1).

Implementation and Evaluation of Physiologic Conditions for a High Throughput "Blood Vessel Mimic" Model System

Project Investigator:

Kristen O'Halloran Cardinal
Department of Biomedical and General Engineering
California Polytechnic State University
San Luis Obispo, CA

FINAL REPORT AND SUMMARY OF RESULTS

Implementation and Evaluation of Physiologic Conditions for a High Throughput "Blood Vessel Mimic" Model System

Kristen O'Halloran Cardinal, PI
Department of Biomedical and General Engineering

Introduction and Use of Funds

This C3RP grant was focused on developing and implementing physiologic conditions within an engineered *in vitro* human blood vessel construct. This human blood vessel construct, or "blood vessel mimic" (BVM), is being developed for use as a high-throughput, living model for preclinical evaluation of new intravascular technologies. A successful physiologic vessel model has potential uses for device and therapeutic evaluation as well as for development of diagnostic technologies to treat or assess cardiovascular conditions related to disease or trauma. The capability to design and develop better technologies and better devices will lead to better care of injured civilians and soldiers. The vessel model could also potentially be used to evaluate the effect of circulating toxins, pathogens, or chemical agents on human vascular cells.

Previous work established methods for creating a blood vessel mimic composed of expanded polytetrafluoroethylene (ePTFE) tubular scaffolds lined with human microvessel endothelial cells. This vessel mimic had been successfully used to evaluate the endothelial cell response to intravascular stents. However, certain limitations of this model existed, including a non-physiologic low-shear flow environment, and use of an expensive, off-the-shelf polymer scaffold with non-native mechanical properties. Thus, the focus of this project was to improve the vessel model by addressing these key limitations.

In order to develop a more physiologic vascular model, this project attempted to answer several key questions. Can we incorporate physiologic flow conditions and an appropriate, customizable scaffold with the current high-throughput, cost-efficient vessel mimic? How will these components influence the resulting vessel structure and morphology? In order to answer these questions, two main aims were proposed. The first aim was to develop and implement physiologic flow conditions within individual vessel systems. Proposed methods included modification of nutrient medium to increase viscosity, and configuration of new pumps and new bioreactor systems to induce pulsatile flow. The second aim was to develop and implement a scaffold with appropriate mechanical properties that would support vascular cell cultivation and growth *in vitro*. Proposed methods included in-house creation of both protein-based and polymer-based vessel scaffolds.

Funds for this project were utilized to purchase the supplies necessary to perform experimental work, as well as for travel and compensation for those involved in the project. Overall, as outlined in the proposal, funds were used in the following ways:

- To purchase supplies summarized below:
  - Flow Experiments: pump drive, clamps, tubing, fittings, valves, bioreactor chambers, media, dextran, methylcellulose
  - Scaffold Experiments: collagen reagents, decellularization chemicals, mandrels, polymers, solvents
  - Cell Culture and Assessment: cells, media, supplements, flasks and disposables, staining reagents
- To pay for ethylene oxide and autoclave sterilization services from the Cal Poly Vet Clinic, in order to ensure living vessels were set up sterilely.

- To support student travel to the annual BioInterface Conferenee in Minneapolis, MN for students to present posters of their work to medieal deviee industry members and researchers, as well as to partially support my travel to the annual Tissue Engineering and Regenerative Medieine International Soeiety meeting to present results.
- To provide eompensation for myself during the summer to bc able to foeus my time and researeh efforts on training students and performing experiments as part of this projeet and to provide eompensation for students dedieating themselves to this projeet (whieh ultimately led to additional uneompensated student partieipation).

Summary of Work Aeeomplished and Results

As summarized above, the two key aims of this work included implcmentation of physiologie flow eonditions and physiologic scaffold materials into an *in vitro* tissue engineered blood vessel mimic model system. These aims lcd to work being divided into two tasks, eaeh of which is outlined below, followed by a summary of data generated and results.

Task 1: Physiologic Flow

- Proposed Hypothesis: Physiologie flow conditions ean be implcmented by inereasing viseosity and by indueing pulsatility through speeifie pump and bioreaetor eonfigurations. Endothelial eells within the vessel mimics will withstand these conditions.
- Proposed Procedures: Completing this task will involve altering media to inerease viseosity and reeonfiguring bioreaetor systems to support pulsatile flow. Work will begin with preliminary researeh and literature searehes regarding bioeompatible media thiekeners, and will eontinue with bioreaetor modifieations primarily to the pump head and flow path. Studies will conclude by assessing the ability of cells to withstand flow modifieations.

Task 1 Data and Results:

Increasing Media Viscosity for More Physiologic Shear Stress

Following a review of the literature, dextran and methyleellulose were identified as possible media supplements with potential to inerease media viseosity to more physiologic levels. After initial testing, it was found that methyleellulose did not dissolve readily in media, and thus dextran was pursued as our media supplement. Various concentrations of dextran were added to media, and viscosity was measured with a viscometer. Measured viseosity was then used to ealeulate resultant shear stress in a 4mm diameter tubular vessel based on the equation:

$$\tau_{wall} = 32\mu Q/\pi D^3$$

where μ is the measured viscosity, Q is the flow rate, and D is the seaffold diameter. Based on literature, our target physiologie shear stress was in the range of 6-12 dyne/em^2, as this level of shear stress has been shown to induee proper endothelial eell alignment and morphology. Caleulated shear stresses in a 4mm diameter BVM with different dextran concentrations and at various flow rates are shown in Figure 1 below. Highlighted in green are eombinations of flow rate and dextran concentration that lead to shear stresses in a physiologie range.

Flow Rate (ml/min)	25	26	27	28	29	30	31	32	33
% Dextran	Shear Stress (dyn/cm^2)								
10.0%	3.5630	3.6997	3.8450	3.9902	4.1269	4.2722	4.4174	4.5542	4.6994
12.0%	4.7625	4.9452	5.1394	5.3335	5.5163	5.7104	5.9046	6.0873	6.2815
14.0%	5.7131	5.9323	6.1652	6.3982	6.6174	6.8503	7.0832	7.3024	7.5353
16.0%	5.8460	6.0703	6.3086	6.5469	6.7712	7.0096	7.2479	7.4722	7.7105
18.0%	5.9925	6.2224	6.4667	6.7110	6.9409	7.1852	7.4295	7.6595	7.9038
20.0%	6.3957	6.6411	6.9018	7.1626	7.4080	7.6687	7.9294	8.1748	8.4356
22.0%	6.4850	6.7339	6.9982	7.2626	7.5114	7.7758	8.0402	8.2890	8.5534
24.0%	6.7891	7.0496	7.3264	7.6031	7.8636	8.1404	8.4172	8.6777	8.9544
26.0%	7.9970	8.3038	8.6298	8.9559	9.2627	9.5887	9.9147	10.2216	10.5476
28.0%	8.6575	8.9896	9.3426	9.6955	10.0277	10.3807	10.7336	11.0658	11.4187
30.0%	9.3963	9.7569	10.1399	10.5230	10.8835	11.2666	11.6497	12.0102	12.3933

Figure 1: Calculated shear stresses in a 4mm BVM based on flow rate and % dextran

Based on these results, a static experiment was performed to assess the impact of high concentrations of dextran on endothelial cell culture. It was determined that concentrations above 20% led to endothelial cell death. Ultimately, a concentration of 14% dextran was identified as an ideal concentration that would support physiologic wall shear stress at an appropriate flow rate, without inducing cytotoxic effects.

Modifying Bioreactor Systems to Support Pulsatile Flow

The previous BVM bioreactor system consisted of a bioreactor chamber and media reservoir connected by tubing across an 8-roller peristaltic pump. This system configuration led to relatively steady flow, which is not representative of the pulsatile environment *in vivo*. Therefore various bioreactor configurations were built using a range of valves, clamps, and tubing in an attempt to induce physiologic pressure fluctuations. A 3-roller pump head was purchased and implemented to more closely match pulsatile conditions. Results from testing different configurations led to the development of the modified bioreactor system shown below in Figure 2.

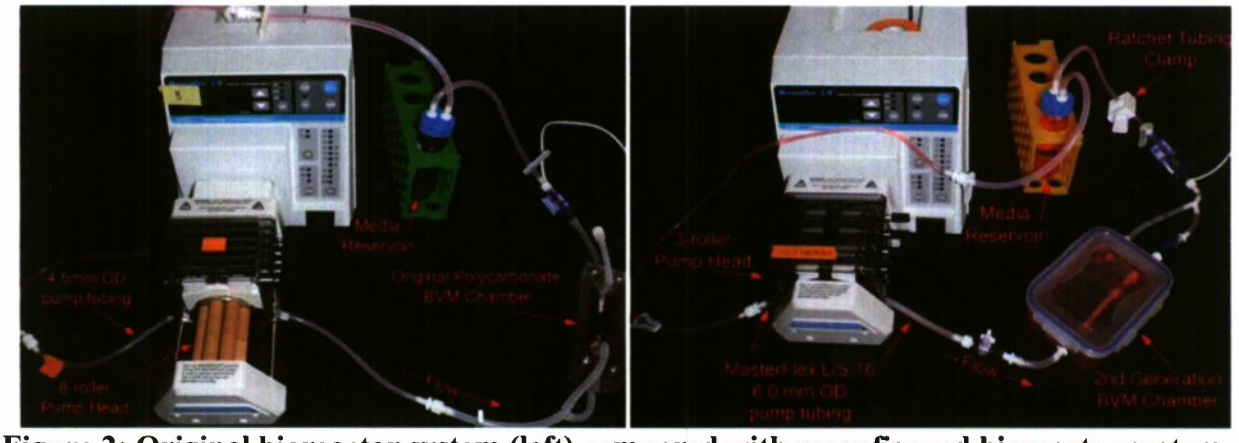


Figure 2: Original bioreactor system (left) compared with reconfigured bioreactor system (right). New system incorporates a 3-roller pump head and additional tubing components to induce more physiologic pulsatile flow with pressure fluctuations.

The new bioreactor design incorporated a minimal number of changes and maintained the simplicity required for scaling up the system. Bioreactor flow conditions were characterized by measuring flow rates corresponding to RPM settings, and by using an AD Instruments LabChart Data Acquisition System to monitor pressure fluctuations. Results from the original system as compared with results from the new system are provided in Figure 3.

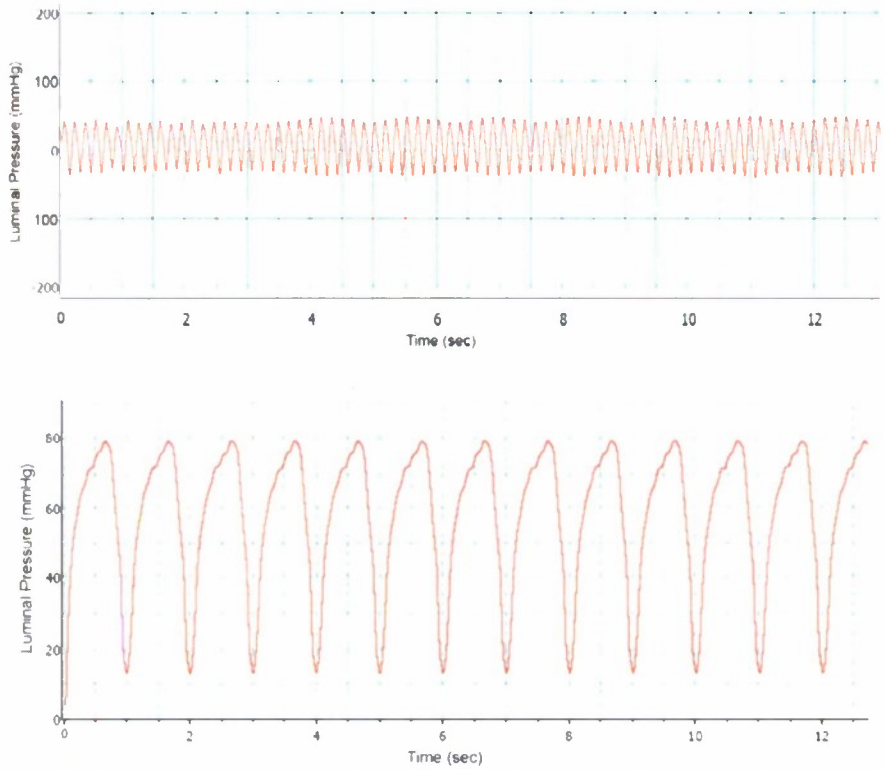


Figure 3: Original bioreactor system induced relatively steady, non-physiologic flow conditions characterized by baseline "noise" pressure fluctuations (top panel). The modified bioreactor system produced more physiologic pressure fluctuations mimicking pulsatile flow in the arteries (bottom panel).

These results demonstrate that it is possible to configure our BVM bioreactor systems to incorporate more physiologic flow conditions. Although there is no back pressure in the new system, the magnitude of pressure fluctuation appropriately reflects the fluctuation *in vivo*, as does the frequency of pulsation.

Assessing Potential for BVM Cultivation under Physiologic Flow Conditions

Based on results from viscosity and pulsatility studies, it was necessary to determine if and how endothelial cells sodded within a BVM would respond to the new flow conditions. Separate studies were performed to determine if human umbilical vein endothelial cells (HUVECs) sodded onto the lumen of an ePTFE scaffold could withstand increased shear stress (using a 14% dextran media recipe) or enhanced pulsatility (using the new bioreactor configuration), and if so- to determine the effect that these conditions would have on subsequent lumen morphology. Results from these studies indicated that HUVECs do in fact maintain

luminal adhesion following implementation of the new flow conditions. These results are illustrated by the positive fluorescent nuclear staining seen below in Figure 4.

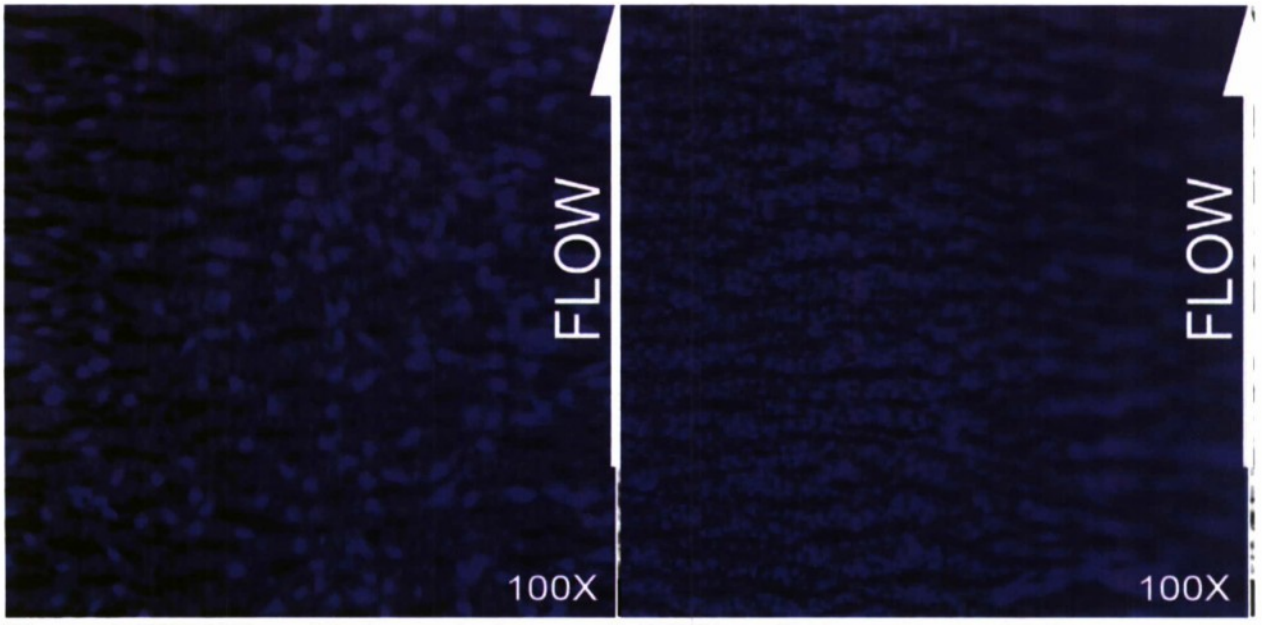


Figure 4: HUVECs sodded onto the lumen of ePTFE scaffolds maintained adhesion following implementation of pulsatile flow conditions (left image) and with increased viscosity (right image).

However, examination of the morphology of these cellular linings revealed that HUVECs do not form an elongated, confluent lining as expected. Several factors may be relevant to why this occurred, including the amount of growth factor present in the media, the cell sodding density, and the flow ramp-up protocols. Future work will be necessary to explore these factors in order to develop appropriate, elongated luminal morphology.

Task 2: Compliant Scaffolds

- Proposed Hypothesis: A compliant scaffold can be developed that will support formation of a more physiologic vessel structure as compared with existing vessel mimics.
- Proposed Procedures: Completing this task will involve developing scaffold protocols, evaluating properties of new scaffold materials, and testing cell compatibility. Initial work will be focused on pursuing two potential scaffolds: collagen (biologic) and polyurethane (synthetic), which have both shown relatively good compliance in previous literature. Protocols will be implemented and fabricated materials will be tested.

Task 2 Data and Results:

Developing Biologic Vessel Scaffolds

Initial efforts were aimed at developing a biologic collagen scaffold to replace the ePTFE that is used in the BVM. The rationale was that a biologic scaffold would more closely match physiologic mechanical properties, and that collagen in particular would match the typical extracellular matrix component, thus leading to a more representative vessel model. Initial

studies involved designing and fabricating tubular molds that were used to form collagen into a polymerized tube. Although proper dimensions were attainable, and collagen polymerization protocols were successfully implemented, resultant tubes clearly lacked the mechanical integrity necessary for BVM application. The collagen tubes even lacked sufficient structure to allow mechanical testing. Therefore, other options for creating a protein based scaffold were explored.

Literature review and further research resulted in the identification of decellularization as a consistent, established procedure with potential for creating tubular protein scaffolds. The process works by using a detergent-like chemical to remove cells from a native artery, thus leaving behind a tubular scaffold composed of native protein structures. We successfully created and implemented a decellularization protocol to create tubular protein scaffolds from porcine arteries, as illustrated in Figure 5.

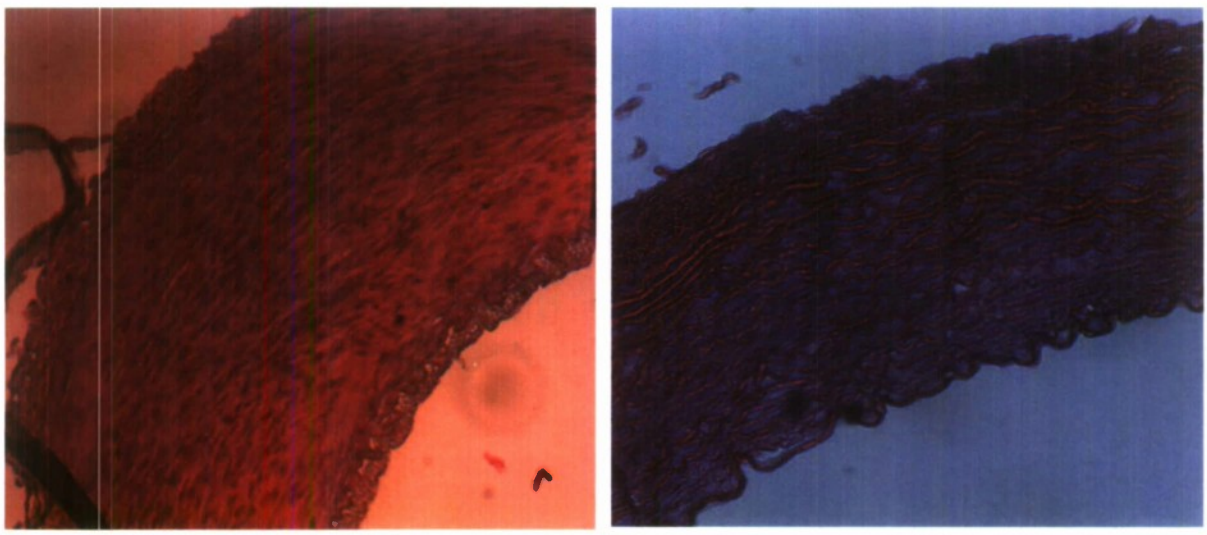


Figure 5: Histology performed on control, native porcine arteries (left image) and decellularized porcine arteries (right image) revealed that 0.075% SDS protocols resulted in a tubular protein scaffold devoid of cells.

Although decellularization results were successful, and subsequent mechanical assessment revealed more appropriate scaffold properties as compared with ePTFE, attempted scale-up led to a major challenge: harvested porcine vessels come in all different sizes, and thus it was not feasible to create a consistent set of biologic scaffolds for implementation in the BVM. Although decellularization work will continue to be explored in our laboratory for other tissue engineering applications, it was determined to be impractical for current BVM applications and therefore for this project. Research efforts were thus focused on creating synthetic polymer scaffolds with improved mechanical properties.

Developing Synthetic Vessel Scaffolds

Following extensive literature review regarding feasible fabrication methods for creating polymer tubes in-house, the method of electrospinning was selected. This method was attractive due to the potential for varying and controlling fiber diameter, fiber orientation, polymer type, and therefore mechanical properties. Through a separate project, an electrospinning apparatus was built, as illustrated in Figure 6. The system is driven by a high voltage power supply that

causes a polymer solution to be collected around a rotating, translating mandrel, which leads to creation of a tubular polymer scaffold.

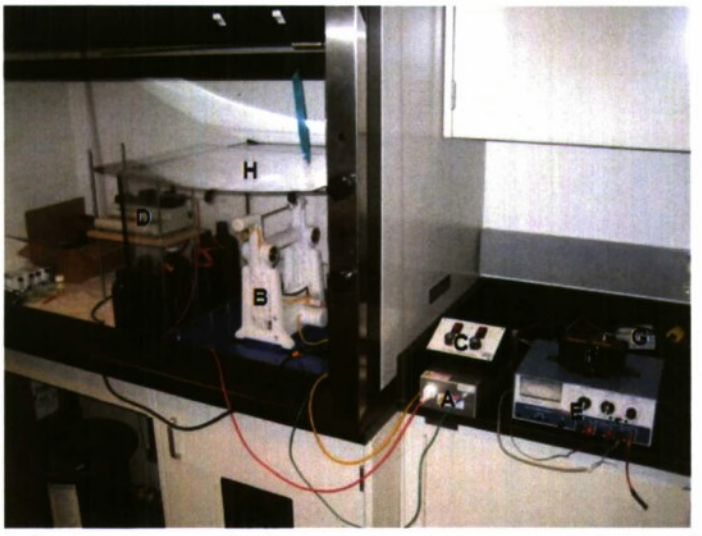


Figure 6: Electrospinning apparatus built in-house to create custom polymer scaffolds. Components include a high voltage power supply (A), a rotating/translating mandrel (B) with dual-speed control box (C), syringe pump for polymer solution infusion (D), external power supply (E), dual switch array (F), AC/DC power converter (G) and chamber (H).

Funding and studies supported by this project focused on two different polymers for use in the electrospinning system. Although polyurethanes had been proposed, these were not pursued due to cost and protocol challenges. However, P(LLA-CL) and PLGA were both identified, purchased, and utilized for electrospinning experiments. The P(LLA-CL) was selected for its ease of use and was used to perform system studies on variables such as solution concentration and voltage. In addition, following protocol establishment, multiple P(LLA-CL) scaffolds were produced, as shown in Figure 7, and assessed based on fiber diameter and mechanical properties to determine the consistency and reproducibility of the electrospinning process.

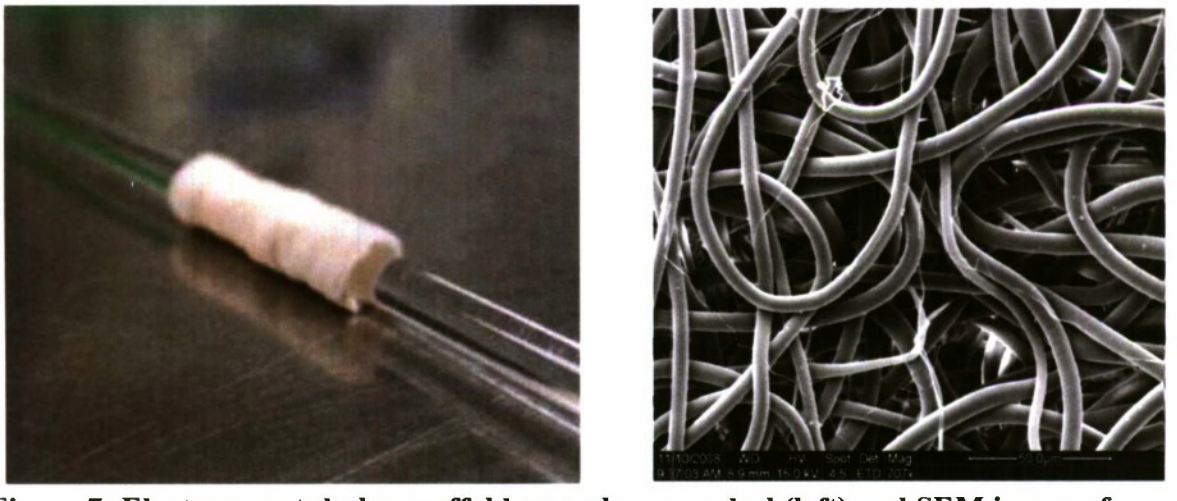


Figure 7: Electrospun tubular scaffold on a glass mandrel (left) and SEM image of electrospun scaffold fibers (right) illustrating the micron-size, customizable fibers.

Following the establishment of our in-house electrospinning fabrication system and based on results from P(LLA-CL) supporting the scalability and consistency of the electrospinning process, PLGA was selected as a more appropriate polymer to pursue for BVM applications. This selection was primarily based on literature supporting HUVEC compatibility with PLGA, as well as adequate degradation rates and mechanical properties for our vessel model. PLGA protocols were successfully established and PLGA tubes were spun and evaluated based on fiber diameter and mechanical properties. Results from tensile testing indicated that stress-strain results and the calculated Young's modulus for electrospun PLGA scaffolds were closer to native values than ePTFE scaffolds. Example tensile testing results from a polymer spun according to our optimized protocol are provided in Figure 8.

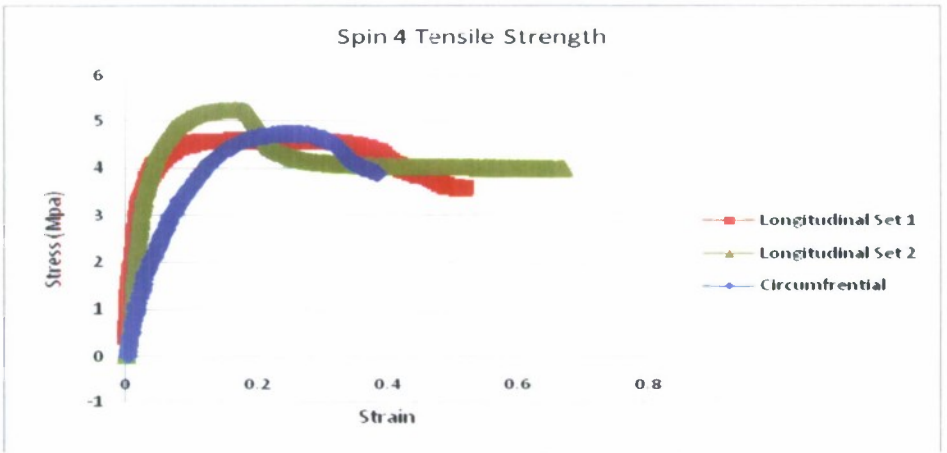


Figure 8: Tensile testing results from electrospun PLGA scaffolds demonstrated that mechanical properties were more similar to native vessels than previous ePTFE scaffolds.

Assessing Potential for BVM Cultivation using New Polymeric Scaffolds

The final studies in Task 2 were focused on assessing the implementation of our new scaffolds within the BVM system. As described, biologic scaffolds did not posses adequate mechanical properties (collagen tubes) or were not scalable (decellularized scaffolds), and thus cell studies were not performed. However, both P(LLA-CL) and PLGA polymers were successfully fabricated, gas sterilized, and incorporated into BVM systems. HUVECs were sodded onto electrospun scaffolds and cultivated for up to 7 days. Overall, results illustrated that HUVEC sodding was possible and that HUVECs adhered to both types of polymer scaffolds.

Specifically with regard to the P(LLA-CL) scaffolds, SEM and fluorescent imaging results revealed that although cells adhered, the lining was not confluent. This was attributed to the relatively large fiber diameters, which may have been suboptimal for cell adhesion, and the porous structure that may have allowed cells to penetrate through the scaffold wall and out of the vessel entirely. In addition, P(LLA-CL) vessels were not cultivated for more than 2 days due to the degradation characteristics of the material. However, these preliminary cell studies supported the capability of sterilizing and working with electrospun materials for the BVM scaffolds.

Subsequent HUVEC studies focused on the PLGA scaffolds, and again- results illustrated that HUVECs could be sodded within these vessels and that they could adhere to the luminal surface. Fluorescent imaging revealed cells present within the BVM at all time points assessed

(up to 7 days). An example fluorescent image illustrating HUVEC deposition on a PLGA scaffold is provided in Figure 9.



Fiture 9: Fluorescent image illustrating HUVECs successfully deposited and cultivated within PLGA scaffolds after 6 days.

Similar images were obtained at all other time points. Although implementation and cell sodding was a success, results indicated that HUVECs were being deposited, or migrating, into the vessel wall, and thus future work will focus on maintaining the endothelial cell lining as a luminal monolayer. This can likely be accomplished by varying pore size and fiber diameter.

Outcomes and Significance

Overall, the major outcomes of this project included the establishment of new protocols for flow and scaffold methods in tissue engineered blood vessel mimics, results supporting the implementation of physiologic flow conditions and more mechanically-appropriate polymer scaffolds, successful training and involvement of numerous engineering students, and dissemination of results at two conferences. A summary of the outcomes based on each proposed hypothesis is provided below:

HYPOTHESIS 1: Physiologic flow conditions can be implemented by increasing viscosity and by inducing pulsatility through specific pump and bioreactor configurations. Endothelial cells within the vessel mimics will withstand these conditions.

OUTCOMES: Media viscosity was increased by adding dextran and thus wall shear stress was successfully adjusted to physiologic levels. Pump configuration was altered through the addition of a 3-roller pump and a variety of tubing clamps and valves, and thus pulsatile flow with pressure fluctuations was successfully implemented. Cell-sodded blood vessel mimics were able to withstand these new flow conditions.

HYPOTHESIS 2: A compliant scaffold can be developed that will support formation of a more physiologic vessel structure as compared with existing vessel mimics.

OUTCOMES: Collagen scaffolds and decellularized scaffolds were not feasible for implementation within the BVM system, but electrospinning was successfully implemented for creating custom polymer scaffolds with more native mechanical properties. These scaffolds supported BVM cultivation.

Key outcomes from the project are summarized below:

- Protocols established and implemented for creating media with specified viscosities through the addition of dextran.
- Protocols established and implemented for inducing pulsatile flow and pressure fluctuations within the BVM bioreactor system.
- Protocols established for creation of collagen and decellularized scaffolds. (Although this was not pursued further for BVM cultivation, these protocols have been implemented in Tissue Engineering and Biomaterials course laboratories)
- Protocols established and implemented for electrospinning P(LLA-CL) and PLGA, performing mechanical testing on scaffolds, and assessing pore/fiber structure of scaffolds
- Results from flow studies indicating that sodded HUVECs can withstand physiologic shear stresses and pulsatile flow waveforms.
- Results from electrospinning illustrating that scaffolds can be custom-made in-house to have appropriate mechanical properties and to support BVM cultivation
- Involvement and training of 7 engineering students at the undergraduate and MS level in various aspects of the project (4 undergraduates, 3 MS students)
- Experimental work contributing towards the completion of 3 Biomedical Engineering MS theses and 4 undergraduate projects
- Presentation of work by two students at the annual BioInterface conference (see attached)
    - Including 2nd place award for Best Student Poster for Colby James
- Presentation of work by P.I. (Kristen Cardinal) at annual Tissue Engineering and Regenerative Medicine International Society (see attached abstract)
- Preliminary work and data supporting National Science Foundation CAREER award application

These outcomes are significant for a variety of reasons. First, our results contribute to knowledge within the tissue engineering field, as described in the project proposal. Second, the improved BVM system has even greater potential to impact the field of intravascular technologies as a preclinical testing system. Third, and specifically with regard to the Office of Naval Research, we have made great strides in creating a model that could be used to develop enhanced treatment or screening technologies for injured civilians and soldiers. In addition, this project strengthens the paradigm of using tissue engineering capabilities as *in vitro* model systems, which could apply to other DOD-sponsored tissue engineering projects. And finally, the work performed through this project led to experience and training of engineering students at multiple levels of their education.

Future Work

Future work related to this project could involve a range of improvements or additional directions, such as:

- Combining the improved physiologic components for creation of an "ideal" BVM (that includes viscous media circulating in a pulsatile flow loop through an electrospun scaffold.)
- Stent or device evaluation within the improved BVM, for comparison to animal or clinical data to validate the model.
- Development of disease or injury-specific BVM systems targeting therapies and technologies that are difficult to evaluate in traditional preclinical models

Vulnerability Assessment of Water Distribution Networks due to Insufficient Fire Flows

Project Investigator:

Shikha Rahman
Department of Civil and Environmental Engineering
California Polytechnic State University
San Luis Obispo, CA

Vulnerability Assessment of Water Distribution Networks due to Insufficient Fire Flows

Shikha Rahman
Civil & Environmental Engineering

Introduction

In recent years Civil Engineers are faced with a unique challenge of protecting public infrastructures from intentional attacks. Eight key infrastructures that are "lifelines" to our society are identified by the President's Commission on Critical Infrastructure Protection (PPD 63, 1998). One of these key infrastructures is water supply system - damage, disruption or inadequate capacity of which jeopardizes our economy and social well being significantly.

The water infrastructure system in USA is a major national asset valued at approximately $675 billion (Grigg, 1999). The USEPA (2003a) has reported that about 170,000 public water systems provide water for more than 250 million Americans. Water distribution systems are vulnerable to intentional/ terrorist attacks because they are extensive, easily accessible, relatively unprotected, and often isolated (USEPA 2002, 2003b, Grayman, 2002; Mays, 2004).

Intentional attacks on water distribution systems can be classified as (i) physical, (ii) cyber, and (iii) chemical/biological. Physical attacks aim at destroying or damaging the water distribution networks by targeting the components such as pipes, pumping stations, water tanks and other facilities. Cyber attacks are aimed at damaging and corrupting the information management system for water infrastructure – Supervisory Control and Data Acquisition (SCADA) system, while the chemical/biological attacks deal with releasing life-threatening chemical and biological agents into the water systems. However, physical attacks are far more likely due to availability of explosive materials, and the lower expense and level of technical expertise required (Burns et al., 2002; Murray et al., 2004). Furthermore, human errors and natural disasters can lead to physical damage/ destruction of water systems as well. The current research focuses on methods to make the water distribution systems more resilient to physical destructions.

Current State of Knowledge

As for any other critical infrastructures, in recent years the government has promoted risk and vulnerability assessment, emergency response, training of water facility personnel, and research towards detecting and mitigating attacks. To date most of the research on vulnerability assessment and mitigation of water systems are qualitative or based on subjective judgments from security experts. All water systems that service over 3,300 people are mandated by the Section 1433 of the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 to perform vulnerability assessment. Widely used methods in industry – 2001 AWWA (M19) Vulnerability Assessment (4th Edition), the Vulnerability Assessment Tool (VSAT^{TM} –WATER) and the Risk Assessment

Methodologies for Water Utilities (RAM–W^{SM}), are all based on subjective scoring systems from surveys of water personnel.

In addition majority of the research are dedicated to chemical/biological and cyber attacks. A more quantitative approach for probabilistic risk assessment to the SCADA system using expenditure, certainty, and flow reduction as performance measures as presented by Ezell (1998). This method was applied to chemical attacks on water tanks in small water systems (Ezell, 2000a, b) as well. Chemical and biological attack scenarios were addressed by a significant number of researchers, particularly in development of effective layouts and design of contaminant monitoring stations for early detection and identification (Lee and Deininger, 1992; Kumar et al, 1997; Harmant et al., 1999; Woo et al., 2001; Al-Zahrani and Moied, 2001; Van Bloemen Waanders et al., 2003; Ostfeld and Salomons, 2004; Berry et al., 2004; Laird et al., 2004). Bahadur et al. (2001, 2003) developed PipelineNet Model through the integration of Arcview and EPANET that stimulates contaminant transport in water distribution systems.

Although there exist potential risk assessment and mitigation strategies in other fields to be applied to water infrastructure, the nonlinear nature of hydraulic systems prohibit applications. In telecommunication numbers of approaches are available that focus on network connectivity and the transition capacity of the nodes/ edges while some subset of the nodes/ edges fail (Monma and Shallcross, 1989; Sakauchi et al., 1990; Chujo et al., 1991; Grötschel et al., 1992, 1995; Veerasamy et al., 1995, 1999; Balakrishnan et al., 2001, 2002). Network interdiction is another closely related area that addresses the how to best attack a network to decrease some measure of its flow and uses more quantitative approaches (Wollmer, 1964; McMasters and Mustin, 1970; Ghare et al., 1971; Wood, 1993; Philips, 1993). As mentioned fundamental system differences prohibit adaptations of these methods to water supply systems.

Optimization methods, such as dynamic programming (DP), branch-and-bound (B&B) and genetic algorithms (GAs), are widely used in the water industry for determining the optimal layout and component size distribution associated with a desired cost and level of reliability. Recently these algorithms are being adopted to address various aspects or consequences of physical attacks on water systems. Jeong et al. (2006) applied B&B and GAs along with a hydraulic solver to identify a feasible customer demand pattern that minimizes the consequences of water shortage in the residual water systems. Jeong and Abraham (2006) developed an operational response model using multi-objective GA to find solutions that indicate the minimum consequences of an intentional attack on water infrastructure. Quio et al. (2007) presented a method that integrates max-min linear programming, hydraulic simulation and GAs for allocation of security budget to a water supply system to maximize the system's resilience to physical attacks.

Scope of the Work

Water supply systems are designed not only to meet the community demands but also to supply adequate water for fire fighting during natural or man-made disasters. Typically water mains are positioned in a grid pattern so that the failure of a single section isolates

the damaged section and the rest of the system can carry water to provide adequate water for fire fighting. Leaks in the pipes, wear and tear of old pipes, or terrorist attacks could cause failures of segments of water mains. In case of a single segment failure the residual system is designed to provide adequate flows and pressures at different fire hydrants. But if multiple segments fail, which is most probable in the scenarios of terrorist attacks, the water supply system might not be able to provide the security required by the community.

Literature review shows that reliability of water infrastructure systems has not been assessed in terms of inadequacy of fire flows and subsequent consequences or propagated damages as a result of insufficient fire fighting capabilities. The current research developed a quantitative method for vulnerability assessment of water systems due to insufficient fire flows using GIS tools, hydraulic simulation model EPANET and an optimization model. The optimization model was developed using dynamic programming techniques to indicate which pipes/ components of the network, if damaged and/or removed, would cause maximum fire damage to the community.

Methodology

The present study has two distinct components: (i) integration of GIS tools with hydraulic model EPANET, and (ii) development of optimization model.

Integration of GIS tools with EPANET

EPANET is widely used in practice for design, analysis, operation and maintenance of the water distribution systems. Hydraulic models such as EPANET (Rossman, 2000) offer the capability of determining the flow rate in each pipe segment, pressure at each node or junction as well as concentration of chemicals throughout the system during desired simulation period. The latest version EPANET 2.0 is freely available at EPA official website (http://www.epa.gov/ORD/NRMRL/wswrd/epanet.html).

Like most hydraulic models EPANET, to set up the water distribution system topology, requires extensive spatial and hydraulic infrastructure data such as topography, road/street layouts, site development, land use etc. Construction of the schematic of a pipe network manually is extremely resource and time intensive. GIS can greatly assist in building the hydraulic models data file due to its unique ability to capture and store spatial data, to build relational database and to display data graphically. Furthermore a properly installed water distribution system with adequate fire flow typically lasts for 70 to 100 years. Since the design, construction, and maintenance of water infrastructure is a long term investment GIS can be used as a long term planning and management tool to update easily as development or modification occurs.

EPANET hydraulic model data file was created using ArcView. The topographic and land use data was downloaded from USGS official website. Components of the water system to be analyzed were removed to create possible physical attack scenarios. Then EPANET was applied to the remainder of the water network after the attack to determine the nodal pressures and flows through the residual water distribution system. Usually the

required pressure for fire fighting is 20 psi, and the required flow is 1500 gpm for single family residential area and 2500 gpm for multi-family and commercial area. From the hydraulic simulation results areas where the pressure varies between 0 and 20 psi was identified to indicate which segments of a system are most vulnerable to fire damage subsequent to terrorist attacks. Various attack scenarios are considered and those that do not create reduced fire fighting capabilities were not considered for this study.

Development of Optimization Model

In this study, nonlinear hydraulic constraints are computed using EPANET. During the optimization trials hydraulic constraints are used to evaluate the maximum damage in terms of the indicators used. Three indices were used in this study to measure the damages due to a physical attack on water system: (i) economic loss, (ii) the number of affected people, and (iii) the degree of disruption of critical infrastructure facilities.

Economic loss was calculated for reduced production activities due to disruption of water supply as well as for insufficient fire fighting abilities. Loss due to water supply disruption was calculated using resiliency factor of the industry/ sector affected. The resiliency factor presents the percentage of the continuation of the sales or production of a specific industry/ sector in case of a total water outage. Resiliency factors for various sectors are available in post-earthquake research studies (ATC, 1991; Tierney and Nigg, 1995; Cheng et al., 2002).

Additional damage due to insufficient fire fighting was addressed by modifying the available resiliency factors. As an indicator of the fire damage a linear relationship was introduced in terms of the required pressure. For example if the available pressure in the damaged network provides half of the required pressure at the fire fighting location then half of the area that the fire hydrant is supposed to service will be assumed damaged. The resiliency factor for any industry or sector in that area was reduced by 50% to take account of the fire damage.

The number of people affected by water outage can be determined by the number of residents at each node/ junction that has disrupted water service after the attack. In the third index, energy supply facilities, transportation hubs, banks and other financial services, telecommunication centers, emergency services (fire stations, hospitals etc.), government services and schools are considered major critical infrastructure facilities. Again for the critical facilities degree of disruption represent both the effects of water shortage or outage and the fire damage. For example, a critical infrastructure might have access to water but could be damaged severely due to fire to function properly.

Damage Function for a particular attack scenario *i* was represented by:

$$D_i = \sum \left[\left(1 - F_{res} \times \frac{P_{avail}}{P_{req}} \right) + \frac{P}{N} + N_{crit} \right]$$

F_{res} = Resiliency Factor
P_{avail} = Pressure available for fire fighting in the damaged network
P_{req} = Pressure required for fire fighting (20 psi was used as the standard)
P = Population affected by water outage and/ or fire damage
N = Constant depending on the population served by the water system damaged
N = 10,000 for small cities
= 100,000 for moderately big cities
= 1,000,000 for big cities
N_{crit} = Number of critical infrastructures affected by water shortage or outage and the fire damage

The damage function represents the total damage cased due to insufficient fire flows after intentional physical attacks. Different attack scenarios that might cause insufficient fire flows were assessed for damage caused by fire and water shortage or outage. The attack scenario with the maximum damage was identified for a particular water distribution network using an optimization model. Dynamic programming techniques was used to develop the vulnerability assessment model for water systems as a result of diminished fire fighting capabilities. The DP model was applied to various hypothetical water networks available in literature.

The steps are shown in the flowchart in Figure 1 in next page.

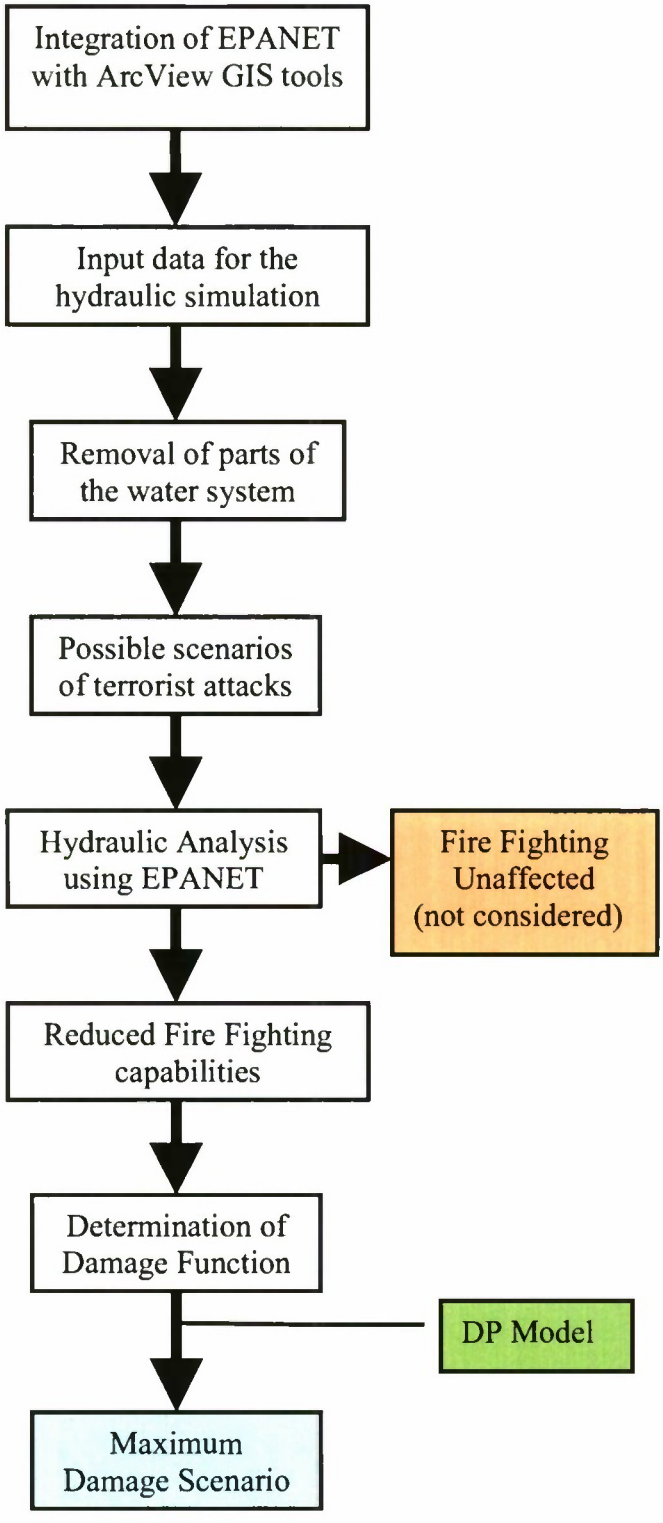

Integration of EPANET with ArcView GIS tools
Input data for the hydraulic simulation
Removal of parts of the water system
Possible scenarios of terrorist attacks
Hydraulic Analysis using EPANET
Fire Fighting Unaffected (not considered)
Reduced Fire Fighting capabilities
Determination of Damage Function
DP Model
Maximum Damage Scenario

Figure 1.

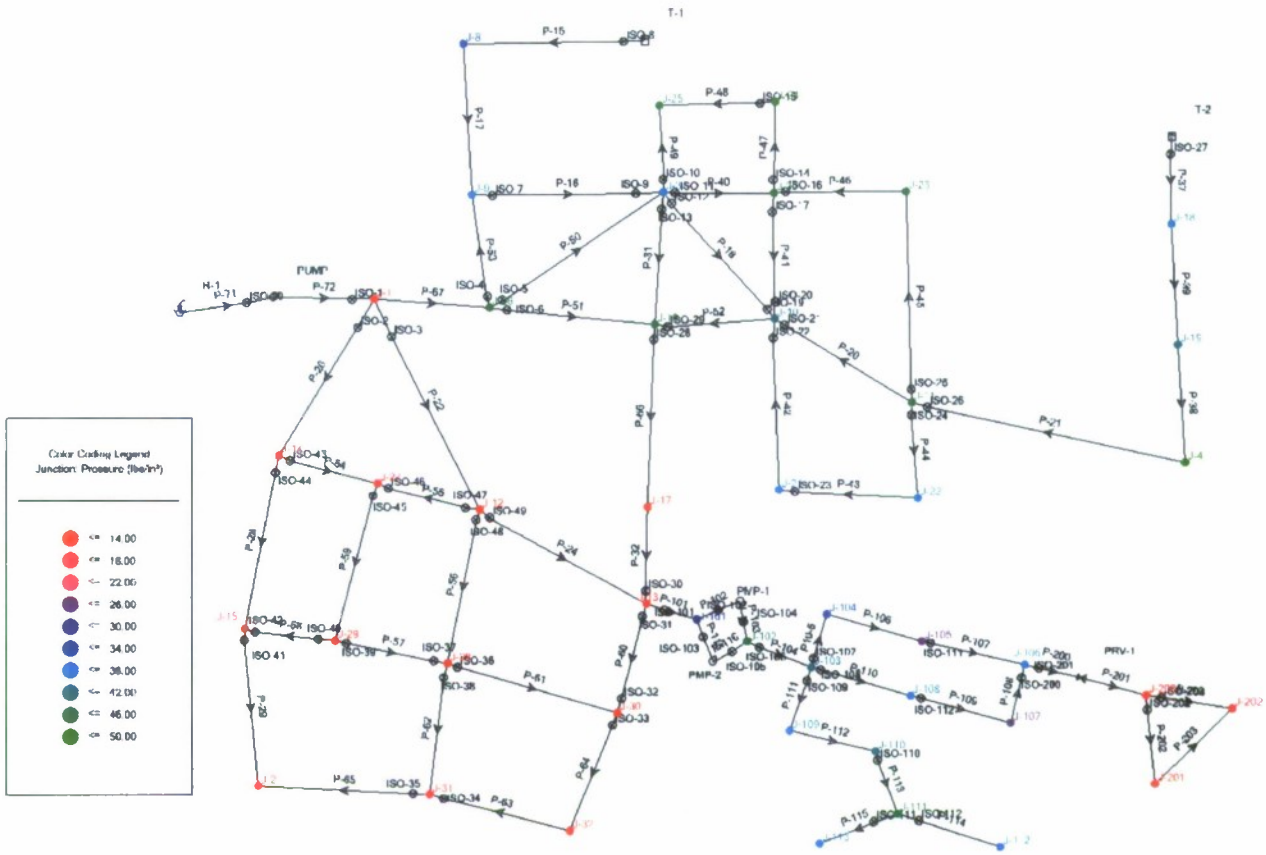


Figure 2.

Figure 2 shows one of the hypothetical water distribution systems used in the current study. The red nodes indicate junctions with pressures less than 20 psi and thus causes reduced fire fighting capability. The maximum damage scenario was identified to be the destruction or isolation of the pipes P-71 and P-72 that causes lower left part of the water system vulnerable to fire damage as well as water shortage.

Accomplishments

Two undergraduate students were employed in the current study. A conference proceeding was published at EWRI-AIT Conference on "An International Perspective on Environmental and Water Resources", January 5-7, 2009, Pathumthani, Thailand on "Vulnerability Assessment of Water Distribution Networks due to Insufficient Fire Flows" by Shikha Rahman. The abstract is given as below:

In recent years Civil Engineers are challenged to develop solutions to protect public infrastructures from intentional attacks. Water distribution system, valued approximately $675 billion, is identified as one of the eight key infrastructures by the President's Commission on Critical Infrastructure Protection (PPD 63, 1998). Physical attacks are far more likely than cyber and chemical/ biological attacks due to availability of explosive materials, and the lower expense and level of technical expertise required. Furthermore, human errors and natural disasters can lead to physical damage/

destruction of water systems as well. Current study focuses on methods to make the water distribution systems more resilient to physical destructions. Water supply systems are designed not only to meet the community demands but also to supply adequate water for fire fighting during natural or man-made disasters. Literature review shows that reliability of water infrastructure systems has not been assessed in terms of inadequacy of fire flows and subsequent consequences or propagated damages. The objective of the current research is to present a quantitative method for vulnerability assessment of water systems due to insufficient fire flows using GIS tools, hydraulic simulation model EPANET and an optimization model. The optimization model was developed using dynamic programming (DP) techniques. Developed DP model was applied to various attack scenarios for a particular water system to identify which one causes the maximum damage. The study is funded under ONR-C3RP grant N00014-07-1-1152.

References

Al-Zahrani, M. A. & Moied, K. (2001), Locating optimum water quality monitoring stations in water distribution system, in *Proceedings of ASCE EWRI World Water and Environmental Resources Congress*, ASCE, Reston, VA, on CDROM.

Applied Technology Council (ATC). (1991). "Seismic vulnerability and impact of disruption of lifelines in the contiguous United States," ATC, Redwood City, California.

Bahadur, R., Pickus, J., Amstutz, D., &Samuels, W.B. (2001), A GIS based water distribution model for Salt Lake City, UT, in *Proceedings of the 21st Annual ESRI Users Conference*, Environmental Systems Research Institute, Redlands, CA, on CD-ROM.

Bahadur, R., Samuels, W. B., Grayman, W., Amstutz, D., & Pickus, J. (2003), PipelineNet: A model for monitoring introduced contaminations in a distribution system, in *Proceedings of ASCE World Water & Environmental Resources Congress 2003*, ASCE, Reston, VA, on CD-ROM.

Balakrishnan, A., Magnanti, T. L., Sokol, J. S. & Wang, Y. (2001), Telecommunication link restoration planning with multiple facility types, *Annals of Operations Research*, **106**, 127–54.

Balakrishnan, A., Magnanti, T. L., Sokol, J. S. & Wang, Y. (2002), Spare-capacity assignment for line restoration using a single-facility type, *Operations Research*, **50**(4), 617– 35.

Berry, J., Hart, W. E., Phillips, C. A., & Uber, J. (2004), Sensor placement in municipal water networks, in *Proceedings of ASCE EWRI World Water & Environmental Resources Congress*, Philadelphia, June 23–26, on CD-ROM.

Burns, N. B., Cooper, C. A., Dobbins, D. A., Edwards, J. C., & Lampe, L. K. (2002), Security analysis and response for water utilities, in L. Mays (ed.), *Urban Water Supply Handbook* (Chapter 20), McGraw-Hill, New York.

Chang, S. E., Svekla, W. D., and Shinozuka, M. (2002). "Linking infrastructure and urban economy: Simulation of water-disruption impacts in earthquakes." *Environ. Plan. B: Plan. Des.*, 29, 281–301.

Chujo, T., Komine, H., Miyazaki, K., Ogura, T. & Soejima, T. (1991), Distributed self-healing network and its optimum spare-capacity assignment algorithm, *Electronics & Communications in Japan, Part I: Communications (English translation of Denshi Tsushin Gakkai Ronbunshi)*, **74**(7), 1–9.

Ezell, B. C. (1998), Risks of cyber attack to supervisory control and data acquisition for water supply, M.S.Thesis, University of Virginia, Charlottesville, VA.

Ezell, B. C., Farr, J. V. & Wiese, I. (2000a), Infrastructure risk analysis model, *Journal of Infrastructure Systems, ASCE*, 6(3), 114–17.

Ezell, B. C., Farr, J. V. & Wiese, I. (2000b), Infrastructure risk analysis of municipal water distribution system, *Journal of Infrastructure Systems,* ASCE, 6(3), 118–22.

Ghare, P. M., Montgomery, D. C. & Turner, W. C. (1971), Optimal interdiction policy for a flow network, *Naval Research Logistics Quarterly*, **18**, 37–45.

Grayman, W. M., R. A. Deninger, and R. M. Clark. (2002), Vulnerability of water supply to terrorist activities. *CE News* 14:34-38.

Grigg, N. S. (1999), A systematic approach to sustain and civilize urban water systems, presented at the EPA Conference on Futures of Urban Water Systems, Austin, TX.

Grötschel, M., Monma, C. L. & Stoer, M. (1992), Computational results with a cutting plane algorithm for designing communication networks with low-connectivity constraints, *Operations Research*, 40, 309–30.
Grötschel, M., Monma, C. L.&Stoer, M. (1995), Design of survivable networks, in M. Ball, T. Magnanti, C. Monma and G. Nemhauser (eds.), *Handbook of Operations Research and Management Science: Network Models*, Elsevier, Amsterdam, pp. 617–72.
Harmant, P., Nace, A., Kiene, L., & Fotoohi, F. (1999), Optimal supervision of drinking water distribution network, in *Proceedings of the 26th Annual Water Resources Planning and Management Conference*, ASCE, Reston, VA, on CD-ROM.
Jeong, H. S., Qiao, J., Abraham, D. M., Lawley, M., Richard, J., and Yih, J. (2006), Minimizing the Consequences of Intentional Attack on Water Infrastructure, *Computer-Aided Civil and Infrastructure Engineering* **21**, 79–92.
Jeong, H. S., and Abraham, D. M. (2006), Operational Response Model for Physically Attacked Water Networks using NSGA-II, *Journal of Computing in Civil Engineering*, 20(5), 328-338.
Kumar, A., Kansal, M. L., & Arora, G. (1997), Identification of monitoring stations in water distribution system, *Journal of Environmental Engineering,* ASCE, 123(8), 746– 52.
Laird, C. D., Biegler, L. T., Van Bloemen Waanders, B. G., & Bartlett, R. A. (2004), Time dependent contamination source determination: A network subdomain approach for very large networks, *EWRI ASCE World Water & Environmental Resources Congress 2004*, Salt lake City, UT, on CDROM.
Lee, B. H., & Deininger, R. A. (1992), Optimal locations of monitoring stations in water distribution system, *Journal of Environmental Engineering,* ASCE, 118(1), 4–16.
McMasters, A.W.&Mustin, T. M. (1970), Optimal interdiction of a supply network, *Naval Research Logistics Quarterly*, **17**(3), 261–68.
Monma,C. L.&Shallcross,D. F. (1989), Methods for designing communication networks with certain two-connected survivability constraints, *Operations Research*, **37**, 531–41.
Murray, R., Janke, R., and Uber., J. (2004), The threat ensemble vulnerability assessment (TEVA) program for drinking water distribution system security, presented at *EWRI ASCE World Water & Environmental Resources Congress 2004*, Salt lake City, UT.
Mays, L. W. (2004), Water supply security: an introduction. Pages 1.1- 1.12 in Larry W Mays (ed.) *Water supply systems security.* New York, NY: The McGraw-Hill Companies.
Ostfeld, A. & Salomons, E. (2004), Optimal layout of early warning detection stations for water distribution systems security, *Journal of Water Resources Planning and Management*, **130**(5), 377–85.
Phillips, C. (1993), The network inhibition problem, *25th Annual ACM Symposium on the Theory of Computing*, San Diego, CA, pp. 776–85.
PDD 63 (Presidential Decision Directive 63) (1998), The Clinton administration's policy on critical infrastructure protection: Presidential decision directive 63. Available at: http://www.terrorism.com/homeland/pdd63.htm.
Quio, J., Jeong, D., Lawley, M., Richard, J. P., Abraham, D. M., and Yih, Y. (2007), Allocating Security Resources to a Water Supply Network, *IIE Transactions*, 39, 95-109.
Rossman, L. A. (2000), *EPANET2 Users Manual*, United States Environmental Protection Agency Cincinnati, OH.
Sakauchi, H., Nishimura, Y., & Hasegawa, S. (1990), A self healing network with an economical spare-channel assignment, *IEEE Global Telecommunications Conference and Exhibition*, **1**, 438–43.
Tierney, K. J., and Nigg, J. M. (1995). "Business vulnerability to disaster related lifeline disruption," Lifeline Earthquake Engineering: Proc. of the Fourth U.S. Conf. ASCE Technical Council on Lifeline Earthquake Engineering Monograph No. 6, 72–79.
U. S. Environmental Protection Agency (2003a), Instructions to Assist Community Water Systems in Complying with the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 Office of Water, EPA 810-B-02-001, Washington, DC.
U. S. Environmental Protection Agency (2002), Baseline threat information for vulnerability assessments of community water systems, Washington, DC.
U. S. Environmental Protection Agency (2003b), Planning for and responding to drinking water contamination threats and incidents. Washington, DC.
Van Bloemen Waanders, G. G., Bartlett, R. A., Biegler, L. T., & Laird, C. D. (2003), Nonlinear programming strategies for source detection of municipal water networks, in *Proceedings of ASCE EWRI World Water & Environmental Resources Congress*, Philadelphia, PA, June 23–26, on CDROM.

Veerasamy, J.,Venkatesan, S.&Shah, J.C. (1995), Spare capacity assignmcnt in tclccom networks using path rcstoration, *IEEE International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems – Proceedings*, pp. 370–74.
Veerasamy, J.,Venkatesan, S.&Shah, J.C. (1999), Spare capacity assignment in telecom networks using path restoration and further improvement using traffic splitting, *The Journal of Systems and Software*, **47**, 27–33.
Wollmer, R.D. (1964), Rcmoving arcs from a network, *Journal of the Operations Research Society of America*, **12**, 934–40.
Woo, H.-M., Yoon, J. H., & Choi, D. Y. (2001), Optimal monitoring sites based on water quality and quantity in water distribution systems, in *Proceedings of ASCE EWRI World Water and Environmental Resources Congress*, on CD-ROM.
Wood, R. K. (1993), Detcrministic network interdiction, *Mathematical Computer Modeling*, **17**(2), 1–18.

A Portable New Chemical/Biological Sensor

Project Investigator:

Karl Saunders
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

Project title: A portable new chemical/biological sensor

Project aim: The aim of this project is a new theoretical design for a liquid crystal based sensor that could selectively detect certain chemical or biological contaminants. Such a sensor would be highly portable making it advantageous over methods that require laboratory-based instrumentation. It would be particularly suitable for military/counterterrorist use as a means to detect biological or chemical weapons in the field.

Summary

Basic Principle

The main premise behind the operation of this sensor is that liquid crystalline matter can be made chiral (or more chiral) when exposed to certain foreign (e.g. airborne biomolecular) compounds. This could happen in a couple of different ways. One possibility would involve chiral compounds diffusing into the bulk of the nonchiral liquid crystal. The presence of this chiral dopant would then induce chirality in the liquid crystal. A second possibility would involve a foreign compound (chiral or nonchiral) binding with the constituent molecules of the liquid crystal and producing a new chiral molecule. One could also have a combination of the two effects whereby a chiral compound both diffuses into *and* binds with the liquid crystal molecules.

In either of these cases the degree of this chirality would be proportional to the density of chiral dopant, which would decrease with distance as one moved from the surface into the bulk of the liquid crystal. Thus, one would a have a gradient in the induced chirality of the liquid crystal, with the chirality being a maximum at the surface of the liquid crystal. This would differ from a traditional cholesteric liquid crystal phase in which the chirality is homogenous, i.e. uniformly distributed.

Results

Analysis of the effect of chiral dopant on molecular configuration

The project has to date been very successful. To analyze the configuration of the liquid crystal with induced chirality, the PI and his undergraduate student assistant, Paul Carlson, have used an established method, involving the Franck elastic energy [1,2]. This method was used to determine the average molecular configuration (and hence, optical axis) of both nematic and cholesteric liquid crystals that have been contaminated with chiral dopant. The Franck elastic energy combines the different types of elastic energies associated with the fundamental distortions of the liquid crystal. The preferred molecular arrangement is the one that minimizes this energy. In the case of traditional cholesterics, i.e. with homogenous chirality, this minimization yields a helical configuration with a pitch length that is the same throughout the sample.

For the liquid crystal sensor we have proposed, the analysis involves a gradient in the chirality as one moves away from the surface where the chiral doping is larger. As a first step in the development of our analytical technique we considered two artificial

spatial distributions of chirality. By "artificial spatial distributions" of chirality we mean distributions that would not necessarily result from real diffusion of chiral dopant into the liquid crystal. The reason for doing so was that we had a good idea of what type of average molecular configuration to expect as a result of such simpler distributions. In particular we considered one chiral dopant density that decayed linearly and another that decayed exponentially with distance away from the liquid crystal-dopant interface. The use of such "simple" distributions thus allowed us to ensure that our non-trivial analytical technique was giving sensible results. Indeed, we found, as one would intuitively expect, that the spatially inhomogeneous chiral dopant results in a cholesteric in which the pitch, rather than being spatially uniform, gets longer as one moves away from the surface.

Our next step was to use our technique to analyze a more realistic spatial distribution of chiral dopant, namely one resulting from its diffusion into the liquid crystal host. To obtain this more realistic spatial distribution we used the diffusion equation. An important consequence of this approach is that the spatial distribution will now depend on time. This is because as time progresses, more and more chiral dopant will diffuse further and further into the liquid crystal host. Using Matlab, we were able to track the time-evolution of the molecular configuration of the liquid crystal. We found that the diffusion of the chiral dopant results in a cholesteric in which, at a given instant in time, the pitch gets longer as one moves away from the surface. As time progresses, the average pitch length shortens (i.e. the helix tightens) due to more dopant diffusing further into the liquid crystal.

Analysis of the reflection of light from the liquid crystal with induced chirality

Having developed our method for calculating the molecular arrangement of the liquid crystal with induced chirality, we have now moved to analyze how the induced chirality affects the reflection of light. As with determining the molecular arrangement of the liquid crystal, there is an established method for doing this [1,2]. This method involves constructing the dielectric tensor corresponding to the molecular arrangement. One can then use this dielectric tensor to determine the effect of the liquid crystal on the polarization of the propagating light. The dielectric tensor can also be used to determine how induced pitch length influences the selective reflection of certain wavelengths. This will be important in relating the color of reflected light to the amount of chiral dopant, and thus the amount of foreign compound that is present.

We have adapted the basic technique to analyze light propagation in a liquid crystal with nontrivial molecular arrangement, i.e. one due to the diffusion of chiral dopant. A particularly nice result of our being able to track the time evolution of the molecular configuration is that we will also be able to track the time evolution of the dielectric tensor. This in turns allows us to track how the light propagation is being affected by the diffusion of chiral dopant into the liquid crystal.

Remaining tasks

There still remains some work to complete the analysis of the interaction of light with the chirally doped liquid crystal. In particular we will be interested in understanding the selective reflection of certain wavelengths.

Once we have performed the above analysis of the molecular configuration and the effect of the induced chirality on the propagation of light through the liquid crystal, we will determine the optimal design for the sensor. While the details of the design will depend on the specific outcomes of the above analysis, this will basically involve determining the optimum boundary conditions and arrangement of optics to detect a change in how the light is reflected due to chirality being induced in the liquid crystal. We would then establish a diagnostic to relate this change in the reflected light to the concentration of foreign compound that is present in the liquid crystal.

New ideas and possibilities

While selective reflection of certain wavelengths of light is one possible way to probe the chiral doping of a liquid crystal, we have, during the course of our analysis, begun considering two other possible probes.

One would be to examine the effect of the chiral dopant on the critical field that is required to unwind the twisted molecular configuration in the cholesteric phase. A general feature of cholesterics is that an electric or magnetic field, applied in the appropriate direction, tends to unwind the cholesteric twist. Above a finite, critical field the twist is completely unwound and the optical properties are much like a field-aligned nematic phase. The critical field depends on the degree of chirality in the system. We propose to analyze the effects of an applied field on our system, one with a spatially varying degree of chirality. We anticipate that there should still be a critical unwinding field and that it will depend in some non-trivial way on the amount of additional chirality which is present, due to some chiral contaminant.

Long-term vision

It now seems likely that our preliminary analysis, once completed, will indicate that this proposed sensor design would allow for the detection of induced chirality. The next (highly non-trivial) stage would be to then find/synthesize a liquid crystal that would become chiral when it reacted with a chemical or biological compound whose detection is desired. An important consideration would be selectivity, i.e. the ability to selectively detect a particular agent of interest. When an appropriate material has been found, we hope that, having obtained external funding, the sensor could be constructed at Cal Poly with the involvement of Dr Jonathan Fernsler, an experienced liquid crystal experimentalist.

Immediate plans for continuation of the project

In addition to the obvious need to complete the analysis of the interaction of light with the chirally doped liquid crystal, we would also like to do further work on the other two possible chiral probes discussed above in the section New Ideas and possibilities. We thus plan to seek a renewal of funding for further work on the project.

[1] For a more advanced overview of liquid crystals see P.G. de Gennes, J. Prost, The Physics of Liquid Crystals (Oxford University Press, 1993)
[2] P.M. Chaikin, T.C. Lubensky and Principles of Condensed Matter Physics, (Cambridge University Press, 1995).

Solar Transportation: Sunlight to Electricity to Motion

Project Investigator:

Peter V. Schwartz
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

Solar Transportation: Sunlight to Electricity to Motion

Summary of 2008 work

Peter V. Schwartz, *Cal Poly Physics Department*

1 Abstract

We have studied both solar concentrators as well as electrical transportation. We have simulated sunlight falling on a stationary solar concentrator, which focuses the concentrated sunlight onto a stationary target by means of a mobile secondary mirror. By writing subroutines, we were able to optimize both the shape of the secondary mirror as well as the trajectory it must follow with the changing incident angle of the sunlight. This work was published in the proceedings of the 2008 SPIE conference in San Diego. Publication in Journal of Solar Energy is pending. Lastly, we compared gasoline cars versus electric cars with respect to life cycle costs and with respect to effective energy density of the drive trains, resulting in two publications in Journal of Energy Policy.

MODELING OF SOLAR CONCENTRATOR

INTRODUCTION

Large solar concentration devices have traditionally consisted of a parabolic primary mirror, which focuses light onto a target, such as a heat collecting element (HCE) or photovoltaic cell (PV). In order to keep the target at the focus of the primary mirror, the entire mirror must rotate about either one axis (for trough systems), or two axes (for dish systems).

U.S. Patent #2,182,222 (December 5, 1939) provides for a parabolic trough solar concentrator that tracks the sun through its apparent motion across the sky (Fig. 1a). The patent provides for a motor to rotate the reflecting trough in order to track the apparent motion of the sun perpendicular to the trough axis. Today, modern central station Solar Thermal Electric (STE) facilities still use a very similar technology, such as the "EuroTrough" (Fig. 1b),[1] where both the parabolic trough and the HCE rotate about a common pivot.

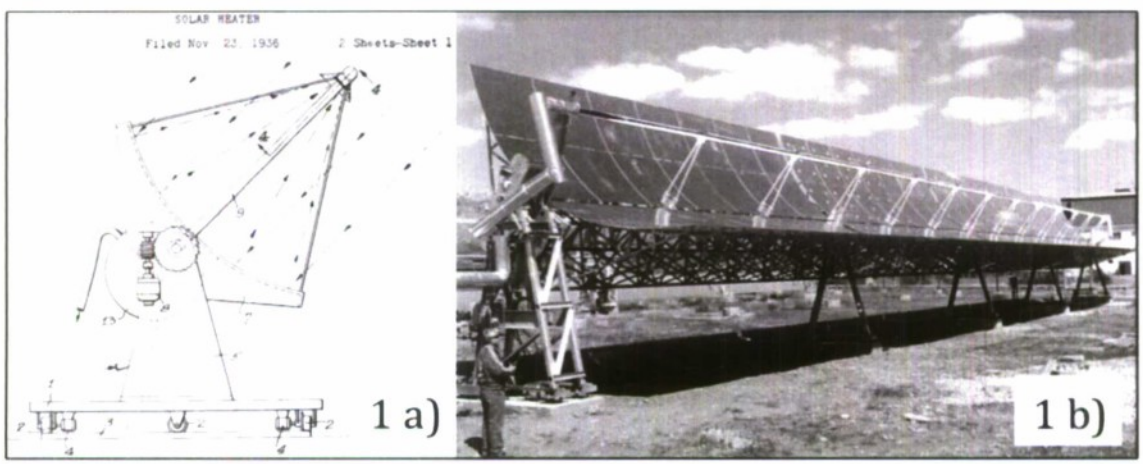


Fig. 1 (a) An original patent figure of a tracking parabolic trough. (b) Modern "EuroTrough" (see text)

Figure 1b illustrates why a considerable amount of the capital cost and complication of construction of a parabolic trough system is due to the need to rotate the assembly. The mirrors and HCE are often made of glass and must be delicately held in place with a high degree of angular tolerance, while hot (and possibly pressurized) fluids circulate in the HCE and through piping that must pivot as the drive mechanism tracks the sun. A recent NREL (National Renewable Energy Laboratory, Golden, Colorado) study[2] indicates that 60% of the cost of an STE facility is the solar field. Additionally, half the cost of the solar field may be due to the need to track the large parabolic mirrors.[3,4] Lastly, the need to carefully construct the rotating troughs to support heavy rigid mirrors complicates construction and increases construction time.

More recent innovations reduce the extent of hardware necessary to track the sun. These geometries include an embodiment with a mobile target, which can move to the location of the focused light.[5] Additionally, a tunable Fresnel mirror assembly is presently being exploited by Ausra[6] for large scale, central station deployment whereby the Fresnel mirror sections on the ground rotate to reflect sunlight onto a stationary overhead target. Lastly, Solargenix[7] manufactures a system called Power Roof that involves a unique trough concentrator of circular cross section, allowing a compound parabolic collector (CPC) to follow the reflected concentrated light. While these innovations lower the size of the components that track, there still exists significant complexity and cost associated with solar tracking.

The design we propose here is similar to that of a Gregorian telescope (Fig. 2). In such an optical system, a large primary concave mirror directs light onto a smaller secondary concave mirror located above the primary and centered about the optical axis between the two mirrors. The secondary mirror serves to correct for spherical aberration and redirect the light back toward the primary mirror. Such a system has been studied before as a way to achieve high solar flux concentrations.[8] However, in this previous work the spatial relationship between the primary and secondary mirrors remains fixed and the entire system must be oriented with respect to the incident sunlight to maintain focus.

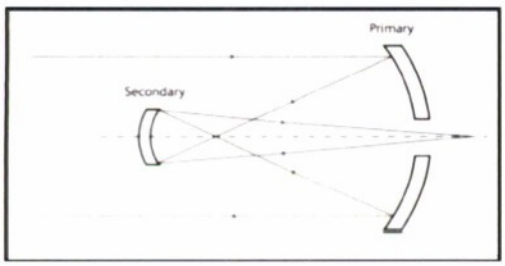


Fig. 2 Diagram of a Gregorian telescope.

Our proposed concentrator design promises to simplify construction and reduce costs by allowing both the primary mirror and target to be stationary. Focusing is achieved through the movement of the smaller, secondary mirror (Fig. 3). While the movement of the secondary mirror presents significant challenges, this system has the potential to be produced on site quickly and inexpensively by embedding the primary mirror and target directly in the earth. In this paper, we explore the performance of a trough concentrator system. However, the idea can equally be applied to dish geometry.

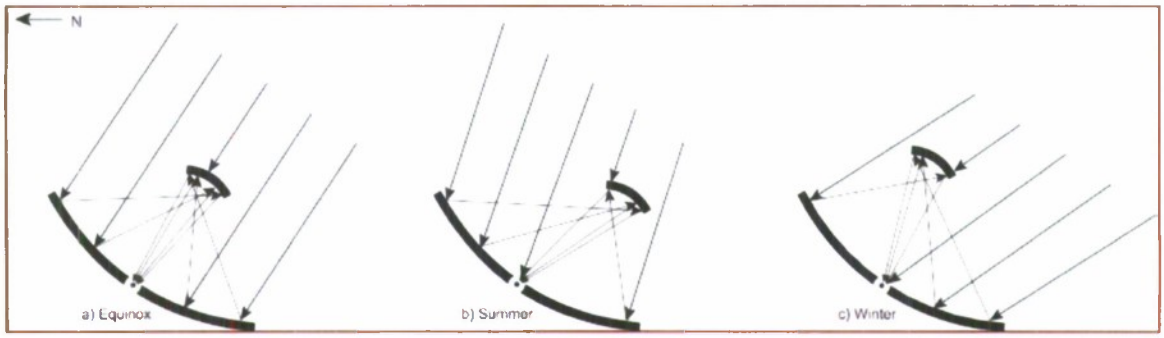


Fig. 3 Incident sunlight is reflected from a stationary primary mirror and focused onto a stationary target by means of a smaller, mobile secondary mirror. If the surface is optimized for equinox, summer and winter light is still focused well onto the same target.

METHODOLOGY

Modeling of the shape and placement of reflecting surfaces was made possible with commercial ray-tracing software (*LightTools 6.1.0* by Optical Research Associates). We used the Solar Tracking Utility included in LightTools to provide realistic simulations of collector performance at San Luis Obispo, CA (Latitude 35.27°, Longitude -120.66°). Since the performance of the concentrator is wholly dependent on the orientation between the sun and the collector, we evaluate collector performance oriented both normal to the surface and tilted at latitude. Performance at other locations can easily be inferred from the simulation data gathered for this one location. We ran simulations for three dates that give the operating extremes of collector performance: Vernal Equinox, Summer Solstice and Winter Solstice. Modeling was purely geometric: mirrors were assigned a reflectivity of 100% and the HCE, a reflectivity of 0%. We used a constant solar irradiance value of 1000 W/m^2. We ran simulations of two different HCE types with the same size aperture: a flat plate and a cylindrical absorber using a tertiary reflector (an involute CPC[9]). Results are reported in terms of efficiency as well as concentration.

The primary mirror has a circular cross section, as opposed to a parabolic cross section. We chose a circular mirror because the coma (concentrated area of reflected light) has the same form for all incident angles. Therefore, a single secondary mirror might be able to focus the reflected light onto the target for all incident angles of the original light. The coma follows a path that is roughly concentric to, and located at half the radius of the primary mirror and so we used this path as a starting point. Subsequent optimization resulted in only minor deviations from a circular path for the secondary mirror.

We optimized the shape of the secondary mirror for every angle of solar incidence from 0° - 40°; we chose an upper limit of 40° because it allowed our system to track the sun for most of the usable daylight hours over the course of an entire year. From the resulting secondary mirror shapes, we selected the surface that provided the highest average solar flux. Using this shape as the secondary mirror, we wrote computer code utilizing an Application Program Interface (API) within *LightTools 6.1.0* to further define the path of the secondary mirror as that which maximized solar flux on the target.

We then explored the optimum combination of model parameters, settling on the following values:

1) Length of the collector $L_{COLLECTOR}$ = 20 m.
2) Radius of curvature of the primary mirror $R_{PRIMARY}$ = 3.0 m.
3) The rim angle of the primary collector q_{RIM} = 80°.
4) Width of the secondary mirror $W_{SECONDARY}$ = 0.6 m.
5) Radius of the cylindrical HCE R_{HCE} = 4.375 mm.
6) Width of flat plate HCE W_{PLATE} = 3.888 cm.

MODELING RESULTS

Figures 4 and 5 compare the concentration achieved when the collector is oriented normal to the surface and when the collector is tilted at latitude, respectively, as a function of the time of day for both the flat plate and CPC absorber types. Figures 6 and 7 compare the efficiency achieved when the collector is oriented normal to the surface and when the collector is tilted at latitude, respectively, as a function of the time of day of both the flat plate and CPC absorber types. These graphs indicate that concentrators tilted at latitude provide for better year round utilization, winter concentrations being barely discernable for the normally-oriented troughs. However, normally oriented troughs provide higher peak output and a greater number of useful summer hours per day.

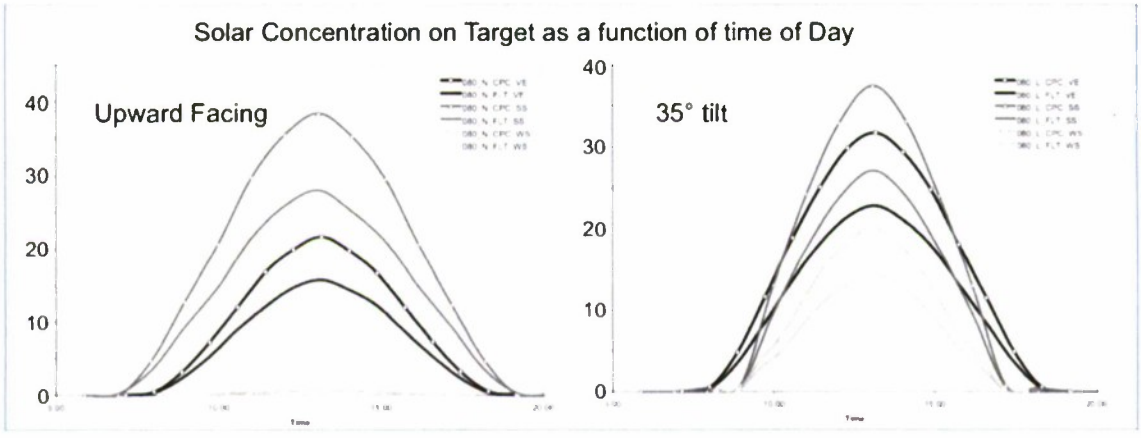


Fig. 4 (left) **Concentration of sunlight on target as a function of time of day**. When not tilted, the collector yields the greatest number of useful hours and highest concentrations during Summer Solstice. However, it offers no concentration during Winter Solstice due to the angle of the sun being greater than the collector's 40° design limit. We also see that the CPC offers higher concentrations than the flat plate HCE.

Fig. 5(right) When tilted at latitude (35° see methodology), the collector maintains functionality throughout the entire year. Interestingly, while the number of usable hours is reduced during Summer Solstice, it actually increases during Vernal Equinox. We also see that the CPC offers higher concentrations than the flat plate HCE.

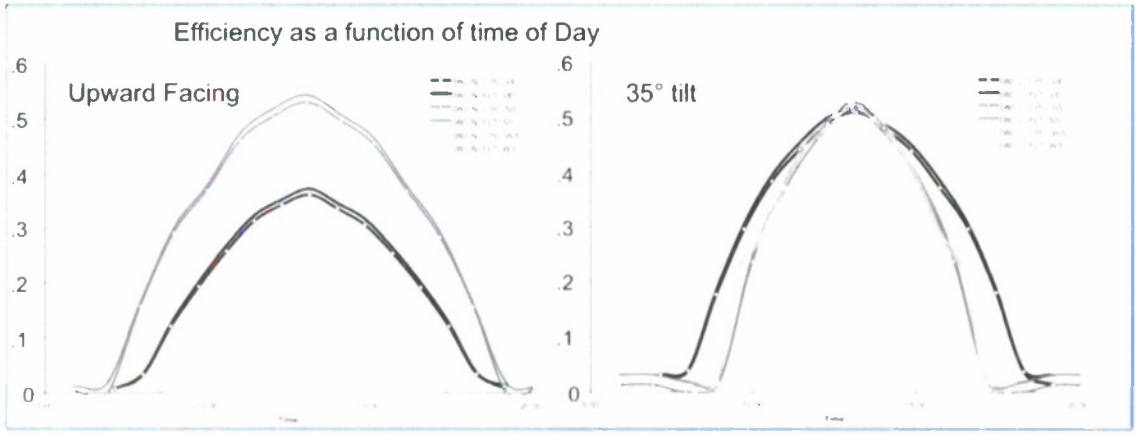


Fig. 6 (left) When not tilted, the collector yields the greatest number of useful hours and highest efficiency during Summer Solstice. However, the system is not usable during Winter Solstice. We also notice that the flat plate offers a slightly higher efficiency than the CPC HCE.

Fig. 7 (right) When tilted at latitude (35°, see methodology), the collector maintains functionality throughout the entire year. Interestingly, while the number of usable hours is reduced during Summer Solstice, it actually increases during Vernal Equinox. We also see that the flat plate offers higher concentrations than the CPC HCE.

Comparison with parabolic trough concentrators: land use efficiency and concentration.

The design proposed here, with an 80° rim angle, achieves a peak efficiency of 54% and concentrations up to 38 suns over a considerable range of incident angles. A conventional parabolic trough achieves concentrations of 71 suns.[10] While a single parabolic trough achieves efficiencies near 100%, troughs are widely spaced in a solar field to avoid shadowing when the sun is lower in the sky. Standard land use efficiency of a typical parabolic trough field is about 25.8 %,[11] although the light use efficiency is higher when averaged over all incident angles. Because our solar troughs could be placed much closer, we could utilize a much greater portion of the solar field. In summary, our results predict solar concentrations and efficiencies only slightly less than solar fields of conventional parabolic troughs.

CONCLUSION

We have modeled the focusing of sunlight onto a stationary target element from an immobile primary mirror of circular cross section by means of a smaller secondary mirror. Using this configuration, with a tertiary concentrator around the target element, we have achieved a peak concentration of 38 solar equivalents with greater than 50% peak efficiency. Improvements are possible with further optimization of the shape and path of the secondary mirror, as well as by allowing the tertiary mirror to rotate about the target. It remains to be seen how much complexity is introduced with the need to control the position and orientation of the secondary mirror. In order to determine under which conditions our proposed design is cost effective, a detailed cost analysis would be required.

EXPERIMENTAL UPDATE: SOLAR CONCENTRATORS AND THERMAL STORAGE

More recently, we have constructed a parabolic mirror concentrator (Fig. 8) and thermal storage facility (Fig. 9). We chose a simple parabolic mirror assembly over our modeled concentrator as a first step in order to first gather experience in a proven technology. The thermal storage unit is made from locally available materials such as sand and pumice. This construction and the associated tests are ongoing.

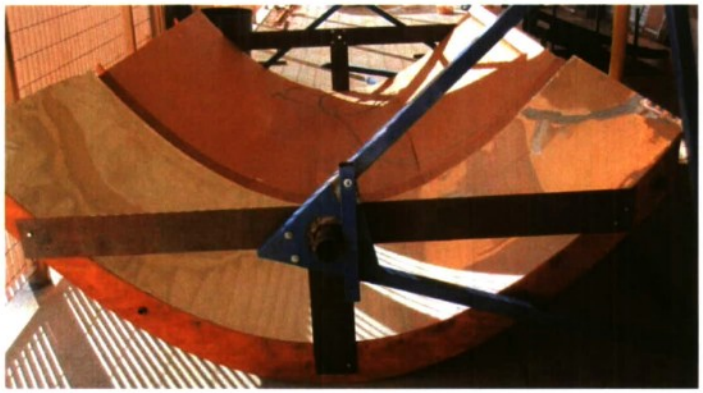


Fig. 8 Parabolic Solar Concentrator. Width: 8', Length: 12'. The Parabolic concentrator is made from thin plywood bent over wooden ribs. Steel piping runs through the wooden ribs lengthwise.

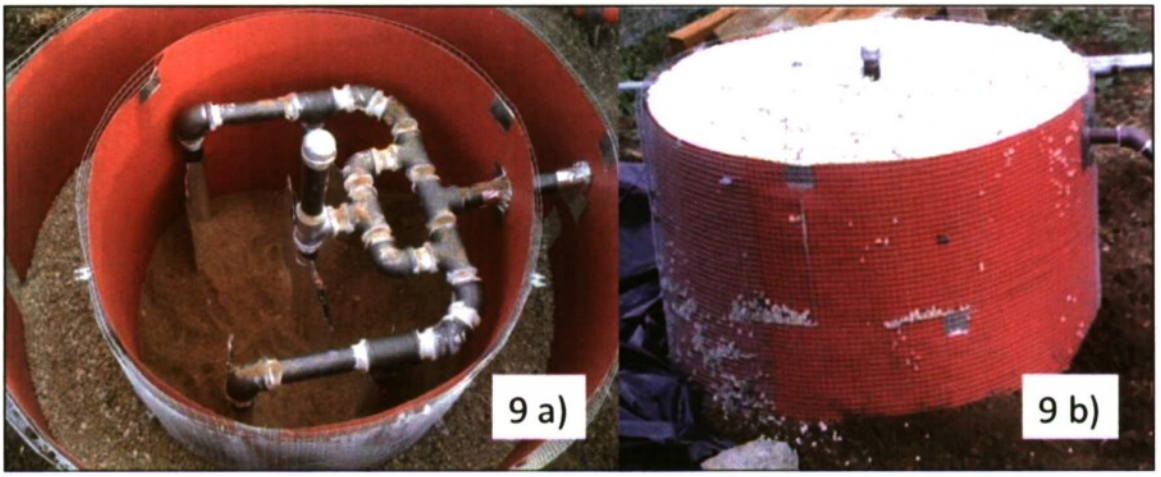


Fig 9a) The solar thermal storage facility from above before being filled with sand and pumice. The outer cylinder is full of pumice as an insulator. The center cylinder is filled with sand, which stores the heat. The heat is distributed throughout the center by means of pipes containing motor oil. b) After being filled with sand, the top surface of the storage unit is covered with more pumice for insulation.

COMPARATIVE ANALYSIS OF ELECTRIC CARS VERSUS GASOLINE CARS: ENERGY DENSITY AND LIFE CYCLE COSTS

Introduction

Compared to alternatives including biofuels, and hydrogen fuel cell technologies, BEVs (battery electric vehicles) have superior technical viability, performance, existing infrastructure, and efficiency. For instance, the present California grid is capable of charging the majority of the state's cars (if electric) during off peak hours with less cost[12] and GHG (Greenhouse Gas) emissions[13] than would powering the same number of cars with gasoline or biofuels.

Yet, electric travel is often dismissed because the batteries are too heavy[14] and too expensive.[15] Our technical and financial analysis finds that both of these assertions are incorrect when applied to an electric car of moderate range (~100 miles, ~160 km). Such an electric car would provide the transportation needs of a great majority of the world's needs.

Our work was done exclusively with public information on automobiles, resulting in two publications in Journal of Energy Policy, May, 2009:

- **Batteries: Higher energy density than gasoline?** M. Fischer, M. Werber, P. V. Schwartz, *Journal of Energy Policy,* **2009**, 37, 2639-2641: http://dx.doi.org/10.1016/j.enpol.2009.02.030
- **Batteries: Lower cost than gasoline?** M. Werber, M. Fischer, P. V. Schwartz, *Journal of Energy Policy,* **2009**, 37, 2465-2468: http://dx.doi.org/10.1016/j.enpol.2009.02.045

Energy Density. The energy density of batteries is two orders of magnitude below that of liquid fuels.[14] However, this information alone cannot be used to compare batteries to liquid fuels for automobile energy storage media. Because electric motors have a higher energy conversion efficiency and lower mass than combustion engines, they can provide a higher *deliverable mechanical* energy density than internal combustion for most transportation applications.

The stored caloric energy densities of energy storage media, the mass energy density[16] is calculated as:

$$\rho_c = \frac{U_f}{m_f}, \qquad \text{eq. 1}$$

where U_f is the stored energy (lower heating value of the fuel or battery energy) and m_f is the mass of the fuel or battery. However, the relevant energy is not *gross caloric energy stored*, but rather *net mechanical energy delivered to the wheels*, ηU_f, where η is the "stored energy to mechanical work" conversion efficiency and includes contributions from regenerative brakes as well as frictional losses in the transmission. Additionally, a motor and transmission is necessary to convert the stored energy to mechanical work, so the relevant mass should include the drive train mass, m_d: the motor or engine, electrical control and power converters, transmission, exhaust, and all associated parts and fluids. We introduce an *effective* energy density:

$$\rho_E = \frac{\eta U_f}{m_f + m_d}, \qquad \text{eq. 2}$$

This energy density depends on the mass of the fuel or battery, which is a function of the range of the vehicle. This effective energy density is graphed against range for a number of energy storage media (Figure 10).

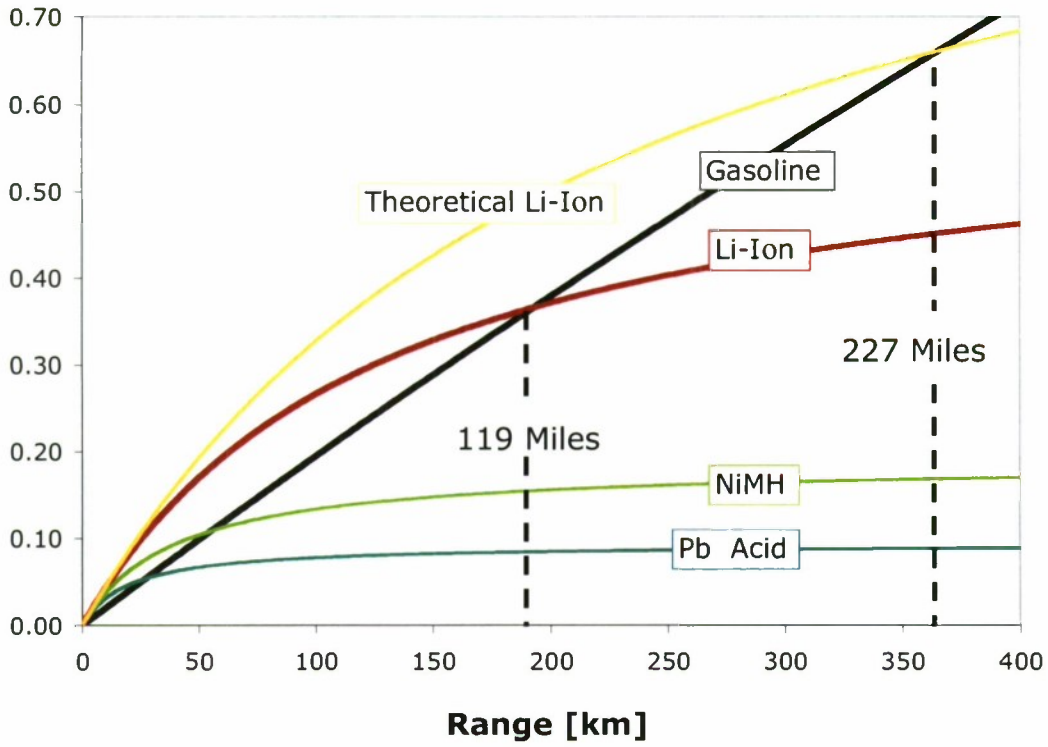


Figure 10 - Effective energy density for a sports car according to eq. 2. Although gasoline drive systems (black) reach much higher energy densities for long-range applications, electric drives have higher energy density for shorter-range travel. Batteries shown: present lithium ion (red), theoretical maximum of lithium ion (yellow), nickel metal hydride (green), and lead acid (blue)

Electric storage has a higher energy density for shorter ranges. Because the greatest majority of automotive trips are short distances (see Figure 11), electric cars can presently provide a significant portion of humanity's travel needs. As electric storage technologies and charging infrastructure improves, the comparative advantage of electric cars will increase.

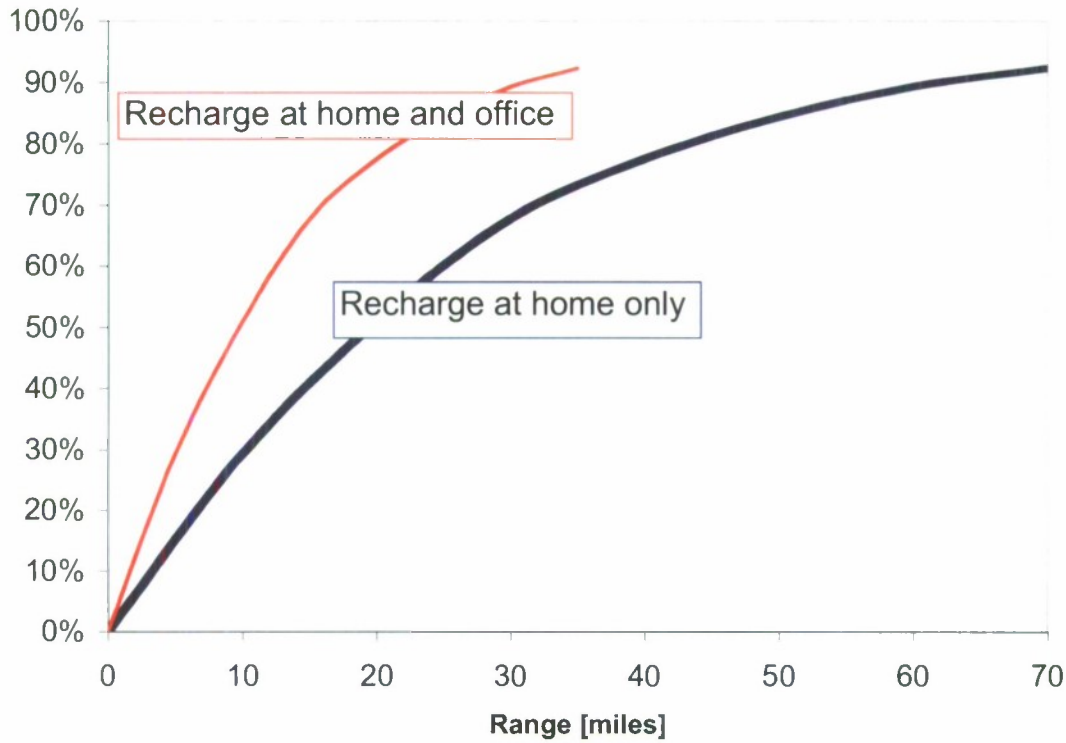


Figure 11 - Portion of American automobile travel needs satisfied as a function of automobile range (DOT, 2003). If cars can be charged at away-from-home destinations, cars with half the full range satisfy transportation needs (red).

Life Cycle Cost of Automobile Transportation.

Electric cars cost more than the analogous gasoline car – largely because of the cost of batteries (which increases with required driving range). However, the maintenance and fuel for electric cars is much less. We compared the lifecycle costs of an electric car to a similar gasoline powered vehicle under different scenarios of required driving range and cost of gasoline. An electric car is cost competitive for a significant portion of the scenarios: for cars of lower range and for higher gasoline prices. Electric cars with ~150 km range are a technologically viable, cost competitive, high performance, high efficiency alternative that can presently suit the vast majority of consumers' needs. Using a 7% discount rate, we calculated the cost of miles traveled for automobiles over the 12-year lifetime driving 15,000 miles annually. The costs were calculated as a function of desired range and cost of gasoline and are shown in Figure 12.

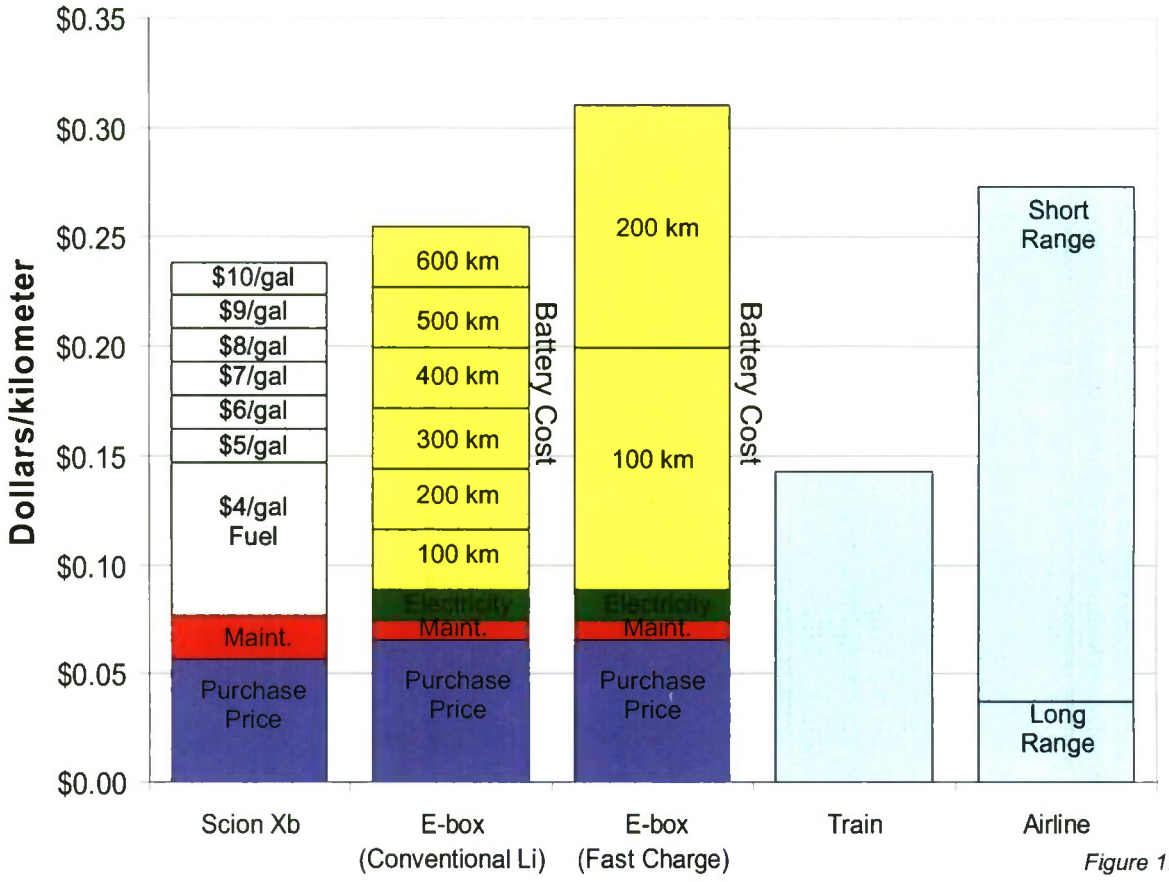


Figure 12

Because the vast majority of American travel are short trips as is shown in Figure 11 (92% of trips are 35 miles or less), the use of a shorter-range, full performance electrical vehicle where appropriate can result in considerable decrease in transportation costs. Electric vehicles have the potential of being less expensive while reducing both emissions (including greenhouse gasses) as well as dependence on oil purchased from unstable/unfriendly political regimes.

REFERENCES

[1] Price, H., et al., 2002. J. Solar Energy Engineering, 124, 109

[2] Stoddard, L., et al., 2006. *Economic, Energy, and Environmental Benefits of Concentrating Solar Power in California*, NREL-SR550-39291

[3] Estimate based on NREL Data supplied by L. Stoddard of Black & Veatch: Stoddard, L., et al., 2006. *Economic, Energy, and Environmental Benefits of Concentrating Solar Power in California*, NREL-SR550-39291

[4] About 63 % of the cost of constructing a solar field is in the foundation pylons, metal support structure, mirrors, drive system and piping interconnects: Sargent & Lundy LLC Consulting Group, 2003. NREL/SR-550-34440

[5] USPTO, Patent #3868823, (March 4,1975)

[6] Ausra, 2008. www.ausra.com/

[7] Solargenix, 2007. www.solargenix.com/

[8] Leutz, R., Ries, H., 2005. *Hemispherical Solar Concentrator Corrected for Aberrations, 3rd* ISCC Conf.

[9] Winston, R., et al., 2005. *Nonimaging Optics*, Elsevier Academic Press, San Diego, p. 52

[10] NREL, 2008. www.nrel.gov/csp/troughnet/pdfs/cohen_nevada_aps_projects.pdf
[11] Canada, S., et al., 2005, NREL/CP-550-37077
[12] D. M. Lemoine, D. M. Kammen, A. E. Farrell, Environm. Res. Lett. 3 (2008) 014003
[13] Stefan Unnasch, Louis Browning, CARB, May 2000, "*Fuel Cycle Energy Conversion Efficiency Analysis*"
[14] American Physical Society "Physics of Sustainable Energy" at UC, Berkeley, March 1-2, 2008 Opening talk by Steve Chu, Director of Lawrence Berkeley National Laboratory; Associated Paper: (S. Chu, Chapter 1, *Science of Photons to Fuels*) *Physics of Sustainabile Energy* American Institute of Physics Press, College Park, MD, edited by D. Hafemeister, B. Levi, M. Levine, and P. Schwartz, 2008
[15] *"Cost, Conflict and Climate: U.S. Challenges in the World Oil Market"*, S. Borenstein, Center for the Study of Energy Markets Working Paper #177, University of California Energy Institute, Revised June 2008.
[16] We neglect the volumetric energy density, as the volume of auto parts is subjective, and the results will be similar to those relating to mass energy density.

Galvanic Vestibular Stimulation Applied to Flight Training

Project Investigator:

Dr. Brian Self
Mechanical Engineering Department
California Polytechnic State University
San Luis Obispo, CA

Galvanic Vestibular Stimulation Applied to Flight Training

Joel Hanson – Dr. Brian Self – Dr. Lynne Slivovsky
California Polytechnic State University - San Luis Obispo

Abstract

This experiment investigates if Galvanic Vestibular Stimulation (GVS) can be used as an indicator of motion sensitivity and explores the effects of GVS on flight simulation performance. The motion sensitivity test consisted of automatically programmed alternating rolling stimulations that increased in frequency over time. Although the stimulations did generate motion sickness in some subjects, the disorientation sensations during the increased frequency transition did not result in motion sickness scores that had any correlation to motion history scores calculated from Kennedy's commonly used Motion Sickness Questionnaire (MSQ). The flight simulator test coupled congruent, conflicting, and sham orientation sensations to the roll angles of the aircraft. The C# program interfaced X-Plane data received through the UDP connection to the GVS device to automatically generate corresponding congruent, conflicting, and sham stimulations. These stimulations did not have any effect on the simulator flight performance of the subjects.

Although the tests did not indicate specific ways to test the vestibular system of future pilots, feedback from the subjects during the tests did raise question regarding the optimum type of waveform for stimulation. Subjects reported expecting specific orientation responses after feeling which side of the head current stimulation was on. Further testing to reduce the surface skin sensation showed that a ramp or increasing exponential waveform not only reduced the sensation of current entering the body but significantly increased the orientation sensations resulting from the stimulation. Increasing the orientation response and decreasing the sensation of current breaking the surface of the skin provides a much more desired stimulation for each of the tests in this experiment and any other future tests related to GVS.

1. Introduction

Spatial disorientation (SD) is "the mistaken perception of one's position and motion relative to the earth" [21]. Between 1994 and 2003, spatial disorientation resulted in at least 202 aircraft accidents, 184 of them resulting in fatalities [21]. What causes people to become disoriented? Are certain people more susceptible to SD than others? Are there ways to safely and cost efficiently test someone's response to disorientation?

Galvanic vestibular stimulation (GVS) provides a safe, cost effective way to simulate spatial disorientation. Cal Poly, in coordination with the Aerospace Medicine & Vestibular Research Laboratory (AMVRL) at Mayo Clinic in Arizona, have outlined two research objectives applied to flight training. The first objective investigates whether GVS can be used as a test for individual susceptibility to motion sickness. The second objective is to explore the effects of GVS coupled with a visual stimulus (presented by a flight simulator) on the human vestibular system. Each of these objectives has the potential to provide safe and cost effective training for pilots.

2 Motion Sickness Susceptibility Experiment

The first experiment was conducted to see if applying alternating GVS currents can help predict susceptibility to motion sickness. Figure 1 - System block diagram for the motion sensitivity test consisting of the GVS Device and the GVS Software shows the basic system design used for performing this test, which consists of the subject, the GVS device, and the custom software.



Figure 1 - System block diagram for the motion sensitivity test consisting of the GVS Device and the GVS Software

2.1 Motion Sickness Susceptibility Design

The GVS device was purchased from Good Vibrations Engineering in Toronto, Canada. The GVS communicates with the custom software through an RS-232 bluetooth or serial connection. The software automatically controls the GVS to generate sensations of rolling left and right. The incremental rolling sensations start out slowly, with 8 seconds between alternating roll sensations. The stimulations roll left, roll right, roll left, and roll right with a constant 8 seconds between each roll sensation. Then the pattern is repeated with the time period between each stimulation decreased from 8 seconds to 6, 4, 2, and 1 second sequentially. Figure 2 shows a graphical representation of the spacing between each stimulation during the Motion Sensitivity Test.

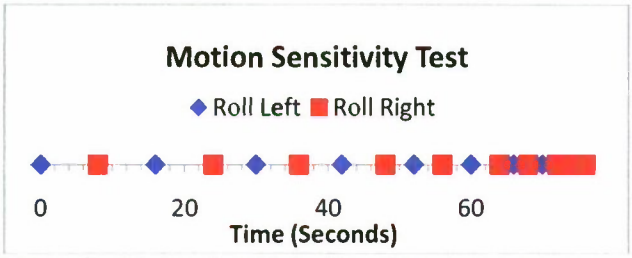


Figure 2 - Motion Sensitivity Test. Blue dot represents a roll left sensation and a red dot represents s roll right sensation.

Between each incremental stimulation sequence the subject indicates a status of motion sickness with the following sickness scale: 0: no symptoms; 1: any symptoms, however slight; 2: mild symptoms, e.g., stomach awareness but not nausea; 3: mild nausea; 4: mild to moderate nausea; 5: moderate nausea but can continue; 6: moderate nausea and want to stop [13]. Figure 3 shows the corresponding stimulation sequence recording of the electrode activity on the oscilloscope.

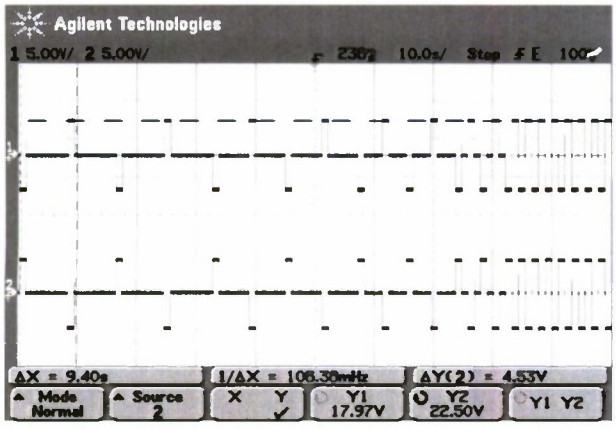


Figure 3 - Lab Recording of Motion Sensitivity Test: 1 milliamp 1 second pulses

As explained further in section 2.2, each subject completes a Motion Sickness Question (MSQ) prior to any testing. The MSQ, developed by Robert Kennedy [16], provides a standardized motion sickness score for each subject. The scoring in this experiment used MSSQA, representing the sickness score prior to age 12, and MSSQB, representing the sickness score in the last 10 years. These scores were calculated using the following equations:

$$\text{MSSQA} = \frac{2.64 \times (\text{total sickness score child}) \times 9}{(\text{number of types experienced as a child})}$$

$$\text{MSSQB} = \frac{2.64 \times (\text{total sickness score adult}) \times 9}{(\text{number of types experienced as an adult})}$$

The raw motion sensitivity score is equal to MSSQA + MSSQB [12]. The higher the score the more susceptible an individual is to motion sickness. These scores will be compared to the sickness scores resulting from the rocking sensations generated during the motion sensitivity test.

2.2 Motion Sickness Susceptibility Procedure

The procedures were explained to the subject and informed consent was obtained, then the subject filled out the motion sickness questionnaire (MSQ). In addition to ensuring the subject passes all exclusionary criteria, the MSQ providesa motion sickness score based on standardized Kennedy MSQ. The subject was then connected to the GVS, which stimulates sensations of yaw, pitch and roll.

The motion sensitivity test was administered in the dark. The subjects closed their eyes and placed their feet together while standing. After every two rocking sensations the subject was asked on a scale of 0-6 how they felt in terms of motion sickness. The responses to each stimulation sequence were documented. After the 5th rocking sequence, the rocking continues for 30 seconds or until the subject requests to stop. The subject's final motion sickness score was recorded.

2.3 Motion Sickness Susceptibility Results

All subjects in the motion sensitivity test reported sensations of rocking back and forth. Figure 4 shows the sickness scores for each of the subjects at the different rocking frequencies. The labels on the right indicate scores for each subject. The plot also shows the average of all the scores at each rocking frequency and the linear trend of the average.

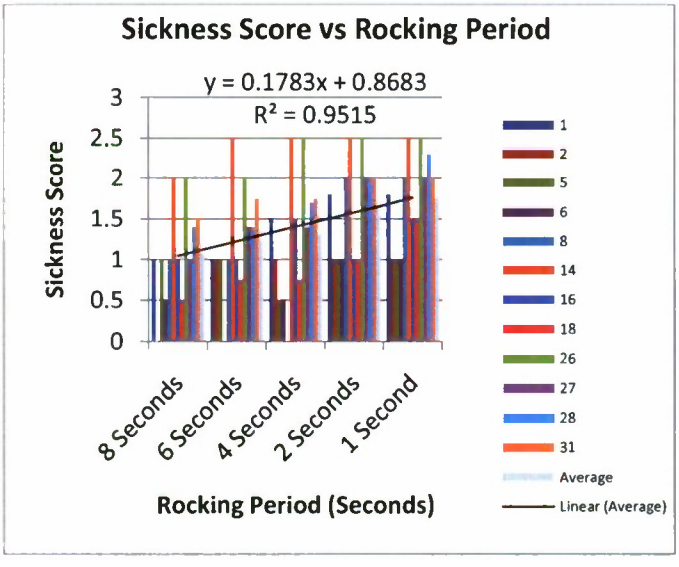


Figure 4 - Motion Sickness Score associated with the varying rocking frequency tests

Figure 5 shows the final sickness score reported after the test versus the MSQ score, and also includes two red circles around data points on the plot. These two subjects both reported feeling nausea after they had completed all testing procedures. Both of these subjects also reported a metallic or battery acid taste in their mouths after testing.

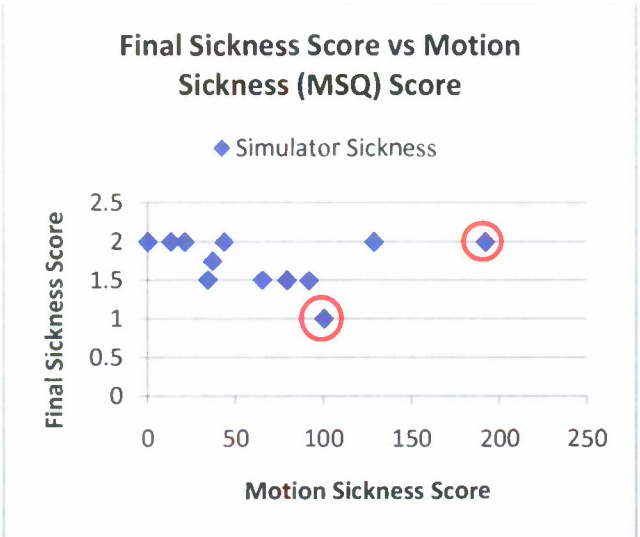


Figure 5 - Final Sickness scores of the subjects versus the motion sickness scores of the subjects

3. GVS Coupled Flight Simulator Design

The second experiment was conducted to explore the effects of the GVS stimulations coupled with a visual stimulus on the vestibular system. In this experiment X-Plane, a flight simulator (www.x-plane.com), provided the visual stimulus. The system design for this experiment included additional computers shown in the block system design in

Figure 6.



Figure 6 - System block diagram for the X-Plane Virtual Flight Simulator including the GVS, the GVS software, a computer running X-Plane controlled by the subject, and a computer running the X-Plane chase plane.

3.1 GVS Coupled Flight Simulator Design

Figure 7 shows the dimensions for the flight simulator setup. In this test, it was important to try to minimize visual references outside of the flight simulator. The simulator image was projected on the wall in a dark room with no windows.

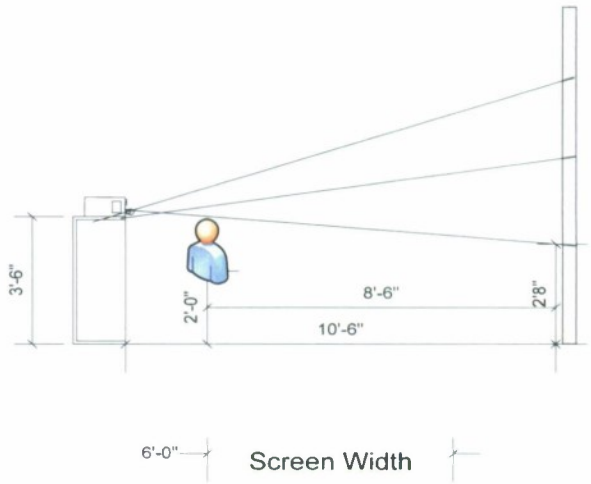


Figure 7 - Flight simulator dimensional room setup

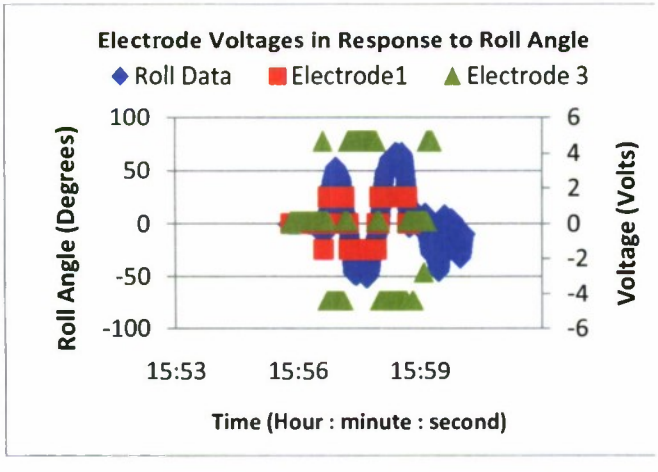
The subject is seated 8 feet 6 inches from the wall. The projector is 2 feet behind the subject and 3 feet 6 inches off the ground. The projected image is 2 feet 8 inches off the ground and 6 feet wide. This setup provides a large visual stimulus for the subject during the flight simulation.

Each subject completes the three X-Plane flight simulator tasks explained in Table 1: VS-A, VS-C, and the Chase Plane task. The VS-A and VS-C tasks are modeled off of actual flight training maneuvers from the FAA's Instrument Training Handbook [25]. The tasks are designed to orient each subject with the common flight instruments and provide basic flight training. Table 1 describes each task in detail.

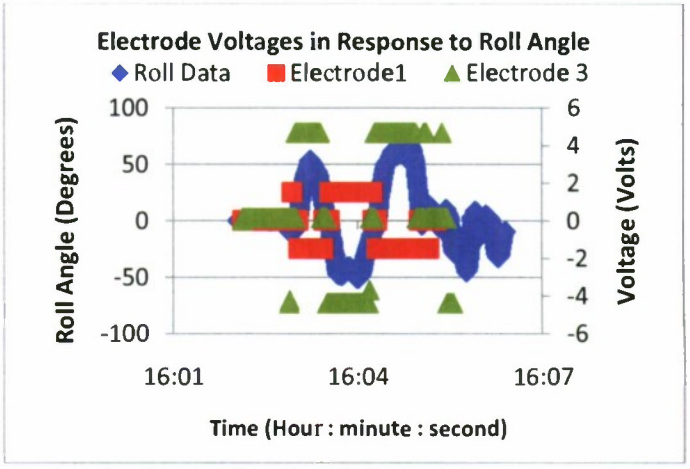
During the tasks, the GVS stimulates congruent, conflicting, and sham sensations. The Congruent stimulation profile stimulates sensations congruent to the aeromodel roll and pitch angles. If the aircraft rolls right, the subject is stimulated to feel like they roll right. The Conflicting profile stimulates sensations conflicting with the roll and pitch of the aircraft. If the aircraft rolls right the subject is stimulated to feel like they roll left. The Sham test provides a low-level stimulation that the subject can feel, but should not generate any sensations. The three profiles are administered in random order for each flight task to mitigate training effects. The threshold for both the roll and pitch sensations is +/- 10 degrees. If the roll or pitch angle exceeds the threshold, a corresponding stimulation begins until the angle drops below that threshold. Before running subjects, a mock-up of the system was set up using resistors and an oscilloscope. Figure 8 plots the roll values and the corresponding electrode stimulation during this mock-up of the Chase plane task for the three stimulus patterns: (a) Congruent, (b) Conflicting, and (c) Sham.

VS-A:	Heading: North Airspeed: 100-200 Knots Altitude (changes +/- 1,000 ft) Bank (constant at 0) Task: -Start at 4,000 ft and climb to 5,000 ft at pitch angle of 10 ° to 30 ° degrees. -Hold 5,000 ft for 15 seconds -Decrease altitude back down to 4,000 ft at pitch angle of -10 ° to -30 °. -Hold 4,000 ft for 15 seconds. -Repeat sequence.
VS-C:	Heading: Changes with bank Airspeed: Changes Altitude (changes +/- 1,000 ft) Bank (+/- 30 °) Task: -Start at 4,000 ft and climb to 5,000 ft at pitch angle of 10 ° to 30 ° degrees, maintain 30 ° bank right. -Level aircraft. -Decrease altitude back down to 4,000 ft at pitch angle of -10 ° to -30 ° maintain a 30 ° bank right. -Hold 4,000 ft for 15 seconds. -Level aircraft and repeat sequence with 30 ° bank left.
Chase Plane:	The final test is to follow the path of a pre-programmed chase plane.

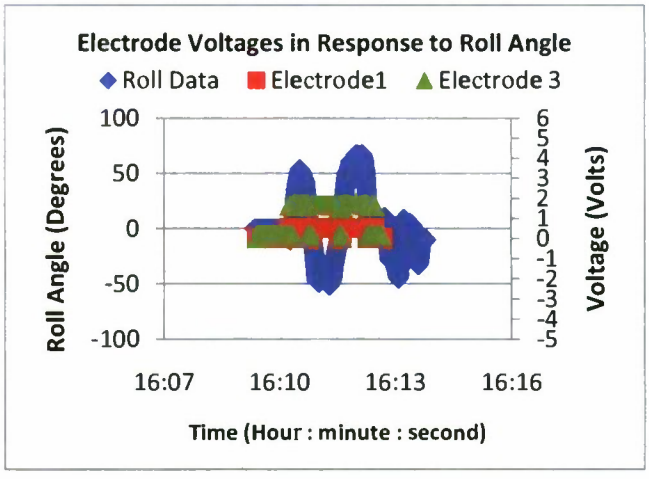
Table 1 - VS-A, VS-C, and chase plane task profiles



(a) Congruent Test Roll stimulations



(b) Conflicting Test Roll Stimulations



(c) Sham Test Roll Stimulations

Figure 8 - Roll stimulations generated from Chase Plane test. The Congruent and Conflicting stimulations are at 1.5 milliamps and the Sham test is at 0.5 milliamps. Electrode 1 voltage was measured across a 1 KΩ resistor and Electrode 3 voltage was measured across a 3 KΩ resistor. (a) Congruent Roll Stimulations (b) Conflicting Roll Stimulations (c) Sham Roll Stimulations

In Figure 8 the roll angles of the chase plane are on the left vertical axis and the associated electrode voltages are on the right vertical axis. A change in roll angles only activates electrode 1 and electrode 3.The Congruent and the Conflicting stimulations are at 1.5 milliamps and the Sham stimulation ias at 0.5 milliamps. The sham test is always one-third the max current selection. Electrode 1 voltage was measured across a 1 KΩ resistor and Electrode 3 voltage was measured across a 3 KΩ resistor. In the Congruent Test shown in Figure 8(a), if the plane rolls right (positive on the graph) , the right electrode provides a -4.5 voltage and the left electrode generates a 1.5 voltage. The voltage polarities are reversed for roll left (negative roll angle). In the Conflicting Test shown in

Figure 8(b), these voltages are opposite the congruent test and therefore generate a conflicting sensation. In the Sham Test shown in Figure 8(c), Electrode 1 provides 0.5 volts and Electrode 3 generates 1.5 volts for any roll angle above 10 degrees.

To complete each task the subject flies a Cirrus Jet, shown in Figure 9. This jet was selected because it provides a balance between control and sensitivity.



Figure 9 - Cirrus Jet used for all of the X-Plane simulator tests

To complete the first two tasks (VS-A and VS-C), the subject must rely heavily on the instrument panel of the jet. The Cirrus Jet uses a Cirrus EXP5000 primary flight display. Figure 10 shows the instrument panel of the Cirrus Jet with labels on the commonly used instruments.

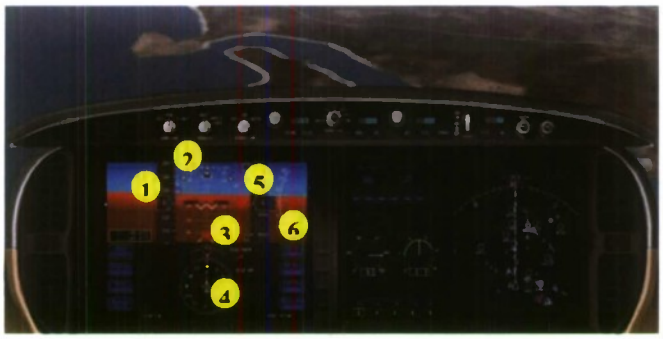


Figure 10 – Cirrus instrument panel. (1) Airspeed indicator: Knots (2) Bank angle indicator: degrees (3) Pitch ladder: degrees (4) Magnetic Headings indicator (5) Altimeter: feet (6) Vertical Airspeed indicator: feet

The instruments used in this experiment are the airspeed indicator, the bank angle indicator, the pitch ladder, the magnetic heading indicator, and the altimeter. The airspeed indictor measures the speed of the aircraft in knots. The bank angle indicator graduations are at 0, 10, 20, 30, 40, 45, and 60 degrees. The bank angle indicator measures the roll of the aircraft, while the pitch ladder measures the pitch angle of the aircraft in degrees. This measurement is taken with respect to the horizontal line on the indicator. The magnetic heading indictor shows the direction the front of the aircraft is pointing. This does not show the true heading of the flight path due to the possibility of cross-winds. The altimeter measures the altitude of the aircraft in feet.

3.2 GVS Coupled Flight Simulator Procedure

The initial procedures for this second day of testing were similar to those for the Motion Sickness Susceptibility tests. The procedures were explained to the subject and informed consent was obtained. If not previously collected, the subject filled out the motion sickness questionnaire (MSQ). In addition to ensuring the subject passes all exclusionary criteria, the MSQ provided a motion sickness score based on standardized Kennedy MSQ. The aircraft controls and joystick functionality were explained to the subject. Each of the tasks from Table 1 were then explained and practiced by the subject. Each of the three tasks (VS-A, VS-C, and Chase Plane) was completed successfully by the subject prior to being connected to the GVS. After attaching the GVS system, the subject repeated each task sequentially (in increasing order of difficulty, VS-A, VS-C, and Chase Plane) with congruent, conflicting and sham stimulations stimulations presented in random order).

3.3 GVS Coupled Flight Simulator Results

As noted in the Table 1, the subject made two climbs and two descents for the VS-A flight task. Figure 11 shows the average magnetic headings of the subjects while completing these tasks with the different stimulation profiles.

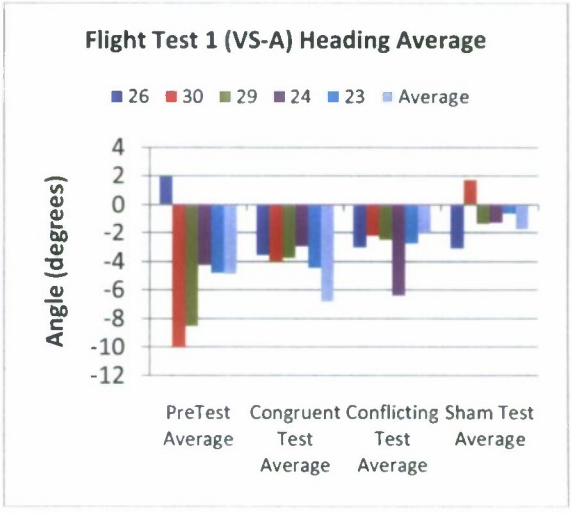


Figure 11 - Average heading values of each subject

It should be noted that if the subject completed the task correctly the aircraft would never roll past +/- 10 degrees and no stimulations would take place. Many of the subjects were able to complete the task without

feeling any stimulation. Root mean square (RMS) deviation was used to analyze the performance of the flight tasks. The formula used for the RMS calculations is:

$$\theta_{RMSdev} = \sqrt{\frac{\sum(\theta_{exp} - \theta_{ideal})^2}{n}}$$

In this case θ_{ideal} is equal to 0 degrees and θ_{exp} is the magnetic heading of the aircraft sustained by the test subject. The variable n represents the number of samples. Figure 12 shows the RMS deviation from 0 degrees magnetic North for each of the stimulation profiles.

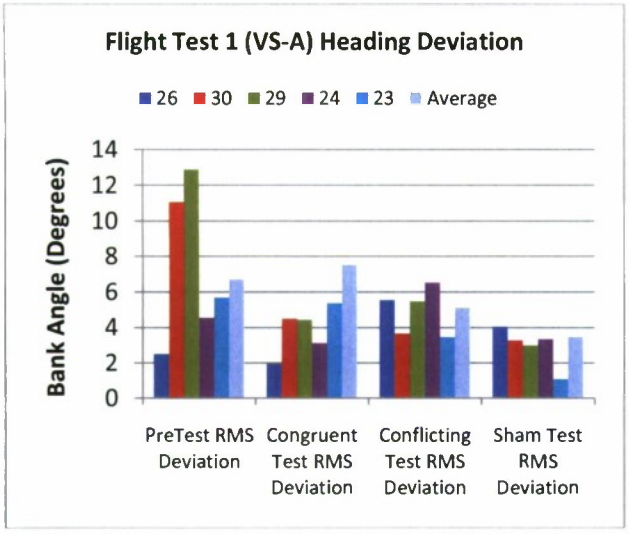


Figure 12 - Root mean square of the deviation of each subject's path from the desired 0 degrees North

The statistical analysis of the heading deviations for the different stimulations indicated that there is not enough statistical difference between the congruent, conflicting, and sham stimulations to exclude the possibility that the difference is a result of random sampling variability ($p = 0.162$). The test had a power of 0.203, signifying that the sample number may be too low to detect and true differences.

The second test profile with the X-Plane flight simulator was VS-C. This test required that the subject repeat the climbs and descents but this time with a +/- 30 degree bank angle. In this test the subject is stimulated during each climb and fall as a result of the bank angle. For these trials, the bank angle was used to examine the performance of the subjects. To ensure that the pilots had not begun to level out, the bank angle was examined between 4,250 feet and 4,750 feet. The RMS deviation from 30 degrees within the altitude boundaries provides a performance metric to compare between the stimulations. Figure 13 provides sample altitude and roll data to demonstrate which roll values are used in the RMS bank deviation calculation.

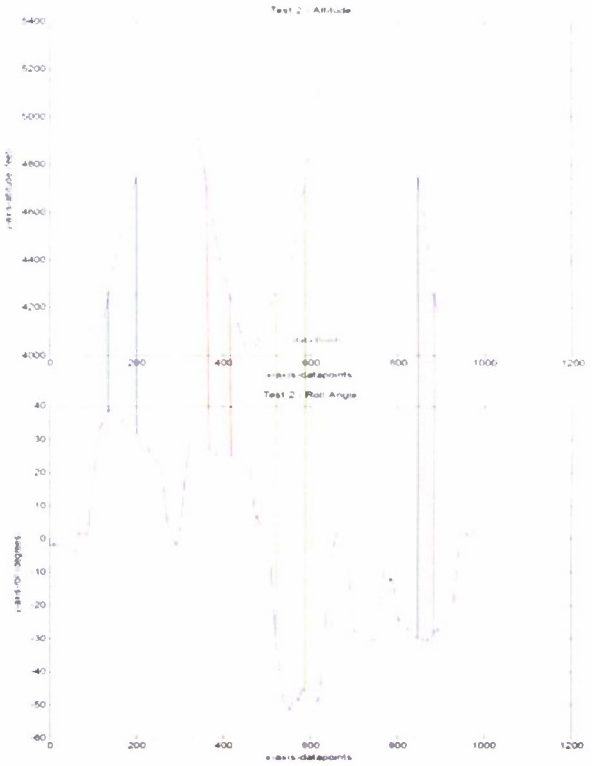


Figure 13 –Test 2 (VS-C) techniques for extracting roll values for the RMS roll deviation calculation. XPlane was set to output 20 data points/second. Every 200 data points corresponds to 10 milliseconds.

Figure 14 shows the RMS deviation from the desired bank +/- 30 bank angles throughout the entire test. The data shows the deviations for the PreTest, congruent stimulations, confliction stimulations, and the sham stimulation.

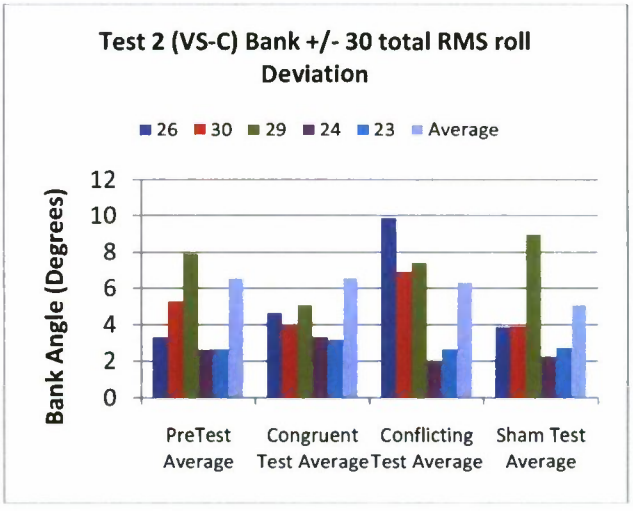


Figure 14 - RMS deviation of +/- 30 degree bank angle for Test 2 (VS-C)

During the VS-C, there was no statistical difference of the total RMS roll deviation values between the congruent, conflicting, and sham conditions (p = 0.346). Again, the test had a low power (0.073).

The first two tests were designed to have the subject use the instruments on the aircraft. The Chase Plane test forces the subject to remove their eyes from the instrument panel and follow a pre-programmed airplane. Figure 15 shows a sample two-dimensional view of a flight pattern. This plot begins on the left side of the graph, viewing the flight path from beneath as if the viewer was lying on the ground. The paths are similar because of the large distances represented on the plot. Each of the X, Y, and Z axes are measured in meters.

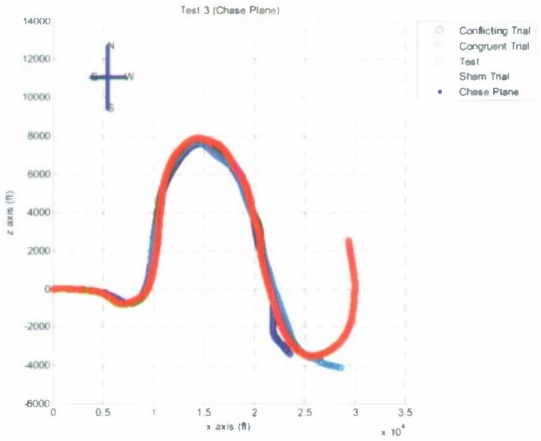


Figure 15 – Chase Plane Test: two-dimensional flight paths viewed from the ground.

In order to maintain the turns the pilot must bank the aircraft. Figure 16 shows the RMS difference in roll angle between the chase plane and the subjects during the various stimulation profiles.

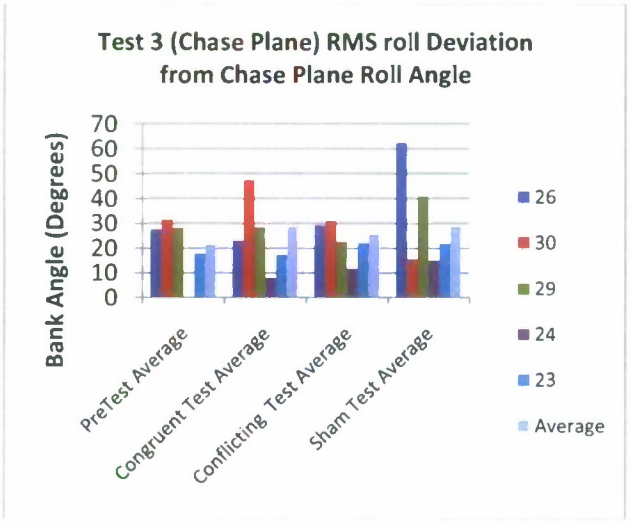


Figure 16 - RMS roll deviations for the Chase Plane

There was no significant difference in RMS roll deviation between the congruent, conflicting, and sham stimulations (p = 0.630, power= of 0.05).

4. Conclusions

The first objective of this experiment was to investigate whether GVS can be used as a test for individual susceptibility to motion sickness. The GVS did elicit rocking sensations in all of the subjects, and the trend line in Figure 4 shows that the motion sickness score increases as the rocking frequency increases. However, the data from Figure 5 show that there was no relationship between motion sickness score to the final sickness score reported after stimulation. Two out of the three subjects with the highest MSQ scores did report feeling nausea shortly after the experiment. GVS can evoke motion sickness, and it may be that the frequency of rocking was not fast enough to elicit motion sickness in most of the subjects.

The second objective of this experiment was to explore the effects of GVS coupled with a visual stimulus. This was done by performing flight tasks while the GVS system provided vestibular stimulations in response to roll data from the flight simulator. Congruent, conflicting, and sham stimulations were used to see if the GVS stimulations affected flight performance.

There were no statistical difference found in any of the tests, but this is to be expected with such a low subject number. This initial pilot test demonstrates that the GVS can be successfully connected to a flight simulator, and that the sensations can be controlled by the orientation of the aircraft.

The feedback of the subjects throughout the testing provided valuable insight into areas that could improve the effectiveness of GVS in both of these applications. During the motion sensitivity test, the subjects indicated that it felt as if the stimulations felt were physically pushing them back and forth. The cathode stimulation typically provided a shock sensation that the subjects associated with a pushing sensation. During the VS-C flight tasks, many of the subjects began to anticipate when the stimulation would happen. After banking the aircraft a few times in VS-C, many of the subjects had figured out that the stimulation happened after they passed the threshold. They then modulated their stick input to prevent the GVS from turning on.

How does this feedback relate to improving the effectiveness of the stimulations? First the shocking sensation needs to be minimized or eliminated. The shock sensation not only provides a distraction due to slight discomfort, but the shock also serves as a reference for the subject to interpret their experiences. The subjects are expecting to feel pushed after the shock for the motion sensitivity test and they expect to feel shocked after they pass the 10 degree threshold in the flight task. Reducing

or eliminating the shock will likely remove both of these factors associated with this test. The second improvement is to increase the amount of orientation generated by the GVS. The device stimulates orientation during the initial current pulse, but, according to the subjects, it does not provide significant orientation sensations after that the initial pulse. This feedback provides the framework for future work to improve the experiments.

5 Future Work

After these tests failed to show any significant effects, different waveforms were investigated to elicit a stronger response. Pilot testing indicated the linear ramping waveforms not only lessened the shocking sensation, but generated a significant increase in the orientation sensation for 6 out of 6 subjects. The orientation sensation is strong enough that the subjects cannot ignore the sensation, even while sitting in a chair with the lights on. This was not true of the square pulse used in the flight simulator and motion sensitivity experiments. The ramping waveforms reduced the shocking sensation enough for 6 out of 6 subjects to comfortably receive currents beyond 2 mA. In the motion susceptibility and the flight test subjects often were not able to reach 1.5 mA because the shock was too uncomfortable or too many of their head muscles were being stimulated. The ramping functions increased the current necessary to stimulate any muscles on the head to over 2 mA. The square pulse waveform would begin stimulating muscles on the head as low as 0.7 mA. The larger range of current levels is important because all of the pilot subjects had increased sensations as the current magnitude was increased.

With the dramatic improvements provided by the linear ramping function, a second waveforms was tested. An exponential waveform provided a sharper rolling sensation than the linear ramping function. It would be useful to implement a test that charts the sensations of many subjects to various ramping waveforms with difference current magnitudes. The information from this type of test could be used to generate more realistic stimulations during the aircraft flight.

The pilot testing indicates that the motion sensitivity and the flight simulator tests might show a greater dependence on GVS stimulation with a more effective waveform. The increased orientation sensations associated with the ramping waveforms provide much stronger and consistent orientation responses. Even with lower currents, pilot subjects were easily made dizzy. Further testing of the new waveforms would provide data to better program the flight test. The ramping function can potentially be used to stimulate quick and slow roll changes resulting from the aircraft, which provides stronger sensation to the subjects.

6. References

[1] Bent, Leah R., Bradford J. McFadyen, Veronique F. Merkley, Paul M. Kennedy , and J T. Inglis. "Magnitude Effects of Galvanic Vestibular Stimulation on the Trajectory of Human Gait." Nueroscience Letters (2000): 157-160. 7 May 2008.

[2] Barnett-Cowan, Michael, and Lawrence R. Harris. "Percieved timing of vestibular stimulation relative to touch, light, and sound." (2009). Multisensory Integration Laboratory. 22 May 2009.

[3] Cacioppo, John T., Louis G. Tassinary, and Gary G. Berntson. Hand of Psychophysiology. 2nd ed. Camgridge: Cambridge University, 2000.

[4] Chatterjee, Indira, Ding Wu, and Om P. Gandhi. "Human Body Impedance and Threshold Currents for Contact Hazard Analysis in VLF-MF Band." IEEE Transactions on Biomedical Engineering 33 (1986). IEEE Explore. Google Scholar. 2 May 2008.

[5] Coulter, Gary R., and Gregory L. Vogt. "The Effects of Space Flight on the Human Vestibular System." Weboflife.Nasa.Com. Nasa. 9 June 2008 <http://weboflife.nasa.gov/learningResources/vestibularbrief.htm>.

[6] Day, B. L., A. Severac Cauquil, L. Bartolomei, M. A. Pastor, and I. N. Lyon. "Human body - segment tilts induced by galvanic stimulation: a vestibularly driven balance protection mechanism." Journal of Physiology 500 (1997): 661-72.

[7] Dzurkova, O., and F. Hlavacka. "Velocity of Body Lean Evoked by Leg Muscle Vibration Potentiate the Effects of Vestibular Stimulation on Posture." Physiological Research 56 (2007): 829-832. Medline. Cal Poly. 16 Apr. 2008

[8] Fitzpatrick, Richard C., and Brian L. Day. "Probing the Human Vestibular System with Galvanic Stimulation." Journal of Applied Physiology 96 (2004): 2301-2316. Medline. Cal Poly State University. 17 May 2008.

[9] Francoise, Marie, and Tardy Gervet. "Effects of Galvanic Vestibular Stimulation on Perception of Subjective Vertical in Standing Humans." Perceptual & Motor Skills 86 (1998): 1155-1157. Academic Search Elite. Cal Poly, San Luis Obispo. 08 Mar. 2008.

[10] Ghanim, Z., J. C. Lamy, A. Lackmy, V. Achache, N. Roche, A. Penicaud, S. Meunier, and R. Katz. "Effects of galvanic mastoid stimulation in seated human subjects." Journal of Applied Physiology 106 (2009): 893-903.

[11] George, Crampton H. Motion and Space Sickness. CRC P, 1990. 0-451.

[12] Golding, John F. "Motion sickness susceptibility questionnaire revised." Brain Research Bulletin, 47 (1999): 507-16. ScienceDirect. 23 Oct. 2008.

[13] Griffin, Michael J., and Barnaby E. Donohew. "Motion Sickness with Fully Roll-Compensated Lateral Oscillation: Effect of Oscillation Frequency." Aviation, Space, and Environmental Medicine 80: 94-101.

[14] Guerraz, Michel, and Brian L. Day. "Expectation and the Vestibular Control of Balance." Journal of Cognitive Neuroscience 17 (2005): 463-69.

[15] Guyton, Arthur C. Medical Physiology. 7th ed. Philadelphia: W.B. Saunders Company, 1986.

[16] Kennedy, Robert S., Jennifer E. Fowlkes, Kevin S. Berbaum, and Michael G. Lilienthal. "Use of a Motion Sickness History Questionnaire for Prediction of Simulator Sickness." Aviation, Space, and Environmental Medicine (1992): 588-593.

[17] Lobel, Elie, Justus F. Kleine, Denis Le Bihan, Anne Leroy-Willig, and Alain Berthoz. "Functional MRI of Galvanic Vestibular Stimulation." The American Phsiological Society (1998).

[18] Malcik, Vladmir. "Performance Decrement in a Flight Simulator Due to Galvanic Vestibular Organ and Its Validity for Success in Flight Training." Aerospace Medicine (1968): 941-943. 15 Mar. 2008.

[19] Medical instrumentation application and design. New York: Wiley, 1998.

[20] Prausnitz, Mark R. "The effects of electrical current applied to skin: A review for tansdermal drug delivery." Advanced Drug Delivery Reviews 18 (1996): 395-425.

[21] Scinicariello, Anthony P., Kenneth Eaton, J. Timothy Inglis, and J. J. Collins. "Enhancing human balance control with galvanic vestibular stimulation." Biological Cybernetics (2001). 17 June 2009.

[22] "Spatial Disorientation Confusion That Kills." www.asf.org. 2004. AOPA Safety Foundation. 19 Mar. 2008 <http://www.aopa.org/asf/publications/sa17.pdf>.

[23] The Highway Safety Desk Book. National Highway Safety Traffic Administration. 6 Sept. 2008 <http://www.nhtsa.dot.gov/people/injury/enforce/deskbk.html#sfst>.

[24] Tortora, Gerard J., and Sandra R. Grabowski. Principles of Anatomy and Physiology. 10th ed. Hoboken: John Wiley & Sons Inc, 2003. 546-556.

[25] United States. U.S. Department of Transportation. Federal Aviation Administration. *Instrument Flying Handbook*. Oklahoma City: United States Department of Transportation, 2007. Print.

[26] Watson, Shaun R. D., Agatha E. Brizuela, Ian S. Curthoys, James G. Colebatch Hamish G. Macdougall, and Michael D. Halmagyi. "Maintained Ocular Torsion Produced by Bilateral and Unilateral Galvanic (DC) Vestibular Stimulation in Humans." Experimental Brain Research 122 (1998): 453-458. 08 Mar. 2008.

[27] Zink, R, S.f. Bucher, A. Weiss, Th. Brandt, and M. Dieterich. "Effects of Galvanic Vestibular Stimulation on Otolithic and Semicircular Canal Eye Movements and Perceived Vertical." Electroencephalography and Clinical Neurophysiology 107 (1998): 200-205. Science Direct. Google Scholar. 25 Apr. 2008.

Optical forces and the effects of particle proximity in optical trapping

Project Investigators:

Nilgun Sungar and John Sharpe
Department of Physics
California Polytechnic State University
San Luis Obispo, CA

"Optical forces and the effects of particle proximity in optical trapping."

Nilgun Sungar and John Sharpe
Dept. Physics, Cal Poly
May/28/09

Project Objective

The goals of this project were to theoretically predict and experimentally measure the forces exerted on multiple microscopic particles which are trapped in optical tweezers. The theoretical part involved a full vectorial approach to light scattering and used the computed fields to calculate the optical forces. The experimental part used an available optical tweezers system to measure the force of a trapped sphere and how the force depends on the proximity of another particle. This is important since a number of physical and biophysical studies have utilized optical traps in close proximity to other particles in the sample but have not accounted for the potential change in the trapping force due to the proximal particle. The project intersects DOD interests related to biological hazard detection and computational electromagnetics.

Current State

We (Nilgun Sungar and John Sharpe) used 6 units of release time each during the fall quarter of 2008 to work on this project. Good progress has been made, as explained in the results section, and we will continue working on this project during summer 2009. An undergraduate physics student will also be involved in the experimental part of the project and will be paid through the College Based Fee funds available for summer research experience for students.

Results

Theory and computations

In order to propagate the field through the high numerical aperture lens and then the cover slip we used a plane wave decomposition method with the Ewald sphere/pupil function approach. In previous work by other groups [Rohrbach2001], the effect of the cover slip was only accounted for by shifting the focal plane and changes to the vector components of the field due to its original polarization had not been accounted for. Our computations showed that inclusion of the polarization of the field components modify the field intensity in the trapping region, hence changing the trap strength. If the distance between the cover slip and the trapping region is small, the location of focal plane is virtually the same with or without the inclusion of the polarization as shown in Figure 1. It can be seen however that the intensity at the focal plane when polarization effects are included is lower.

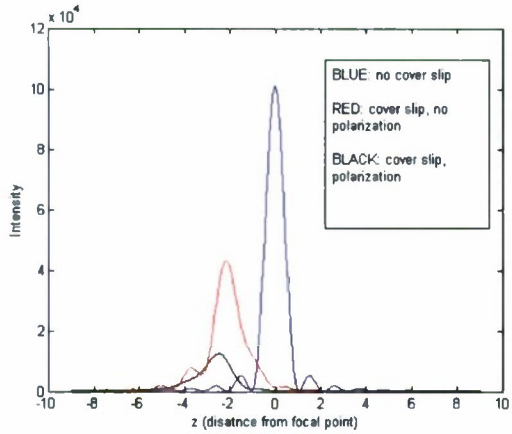
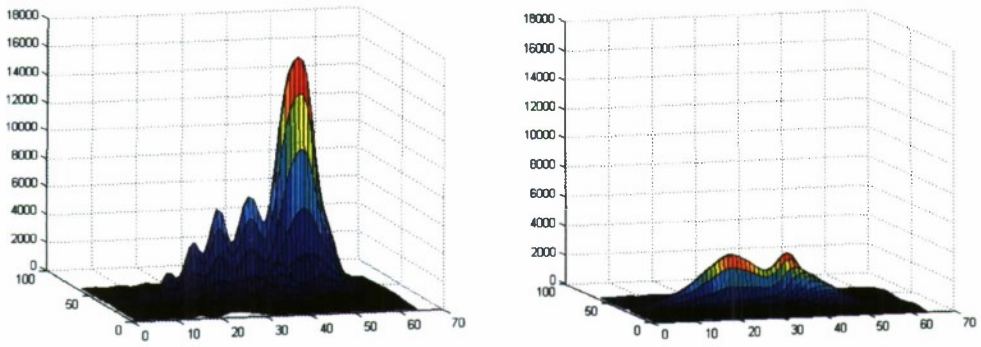


Figure 1: Field intensity along the propagation direction when the distance to cover slip is 10 microns. Blue curve shows the intensity when there is no cover slip, red curve is using the approach of Rohrbach group, black curve is when the polarization effects are included.

When the distance between the cover slip and the focal plane is increased, accounting for the polarization effects produce quite different intensity patterns. This can be seen in Figure 2 which shows the field intensity in the xz plane (where z is the propagation direction of light) when the cover slip is at 30 μm from the focal plane.



(a) (b)

Figure 2: Field intensity on the xz plane where z is the propagation direction of the light when the distance to cover slip is 30 microns, without accounting for polarization changes through the cover slip (a) and with polarization changes through the cover slip(b).

After computing the field in the trapping region, the scattered field from the dielectric particle along with the original field is used to calculate the force on the particle. We used Mie theory [Rohrbach2005] to calculate the scattered field. This theory involves matching the electric fields and their derivatives at the boundary of a spherical scatterer. In order to calculate the force on a particle in the vicinity of another particle, we applied Mie scattering theory iteratively. We first calculated the scattered field due to the secondary particle, and then this scattered field along with the original field is used in a second application of the Mie theory to obtain the net field resulting from the scattering from the trapped particle. In the first application the direction of each plane wave component of the incident beam is expressed in terms of Euler angles. Then, using structural information about the scatterer, boundary conditions are

applied and scattering amplitudes are calculated for each scattering angle. Since the secondary particle is off-axis, it is more practical to place the particle at the origin and shift the incident beam. After we calculated the scattering amplitudes, we shifted the scattered fields again to realign with the trapped particle. We found out that multiple operations of rotating each component of 3-dimensional vector fields and then integrating over the plane wave components of the field is very computationally intensive and takes a long time to run using MATLAB. We now need to investigate other methods of integration and optimization of our programs, which we anticipate doing during summer 2009.

Experimental

A challenge in this project was to find a method for reliably measuring the position of the trapped particle with very high accuracy (~10 nm) at a data rate in excess of 100Hz. It is this position measurement that allows the trap strength to be calculated. Although high-speed position sensitive photo-detectors (PSDs) are the common choice for this sort of measurement, in our case their use is problematic since the trapped particle is very close to a larger particle. The approach we took was to measure the particle position by using a fast shuttering and high frame rate camera (from Basler Corporation). Then using image processing we would be able to follow the particle position [Sharpe2007].This camera communicates its data through a computers Ethernet port and can be controlled using the application LabVIEW. In order to utilize the camera we had to determine how to acquire, store and process the images. For the acquisition and storage part three main LabVIEW programs were employed. The first program was simply to monitor the field of view of the camera. The block diagram of the system is shown in figure 3. Camera parameters (such as frame rate, frame size and gain) can be set through LabView software or using the National Instruments Measurement and Automation Explorer.

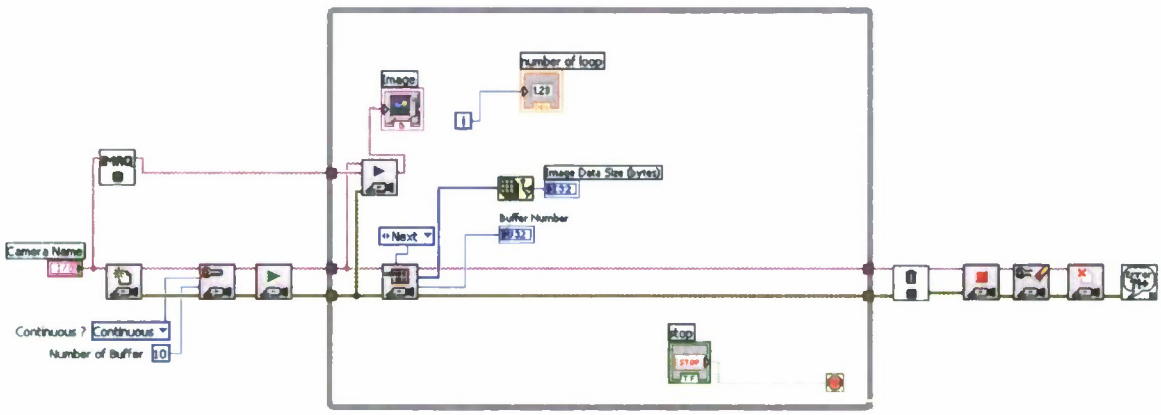


Figure 3: Program used to display the field of view of the Basler camera.

In order to grab a sequence of images at high speed a second program was employed. Although the program could in principal have written the image files directly to the hard-drive (as numbered BMP files, for example) we found that this was not fast enough. Instead, we employed a program that simply stored all the data in one large ram-based array. At the end of the acquisition the user is prompted to give a file name and location and the data was then saved as a one-dimensional binary file. This program is shown in figure 4.

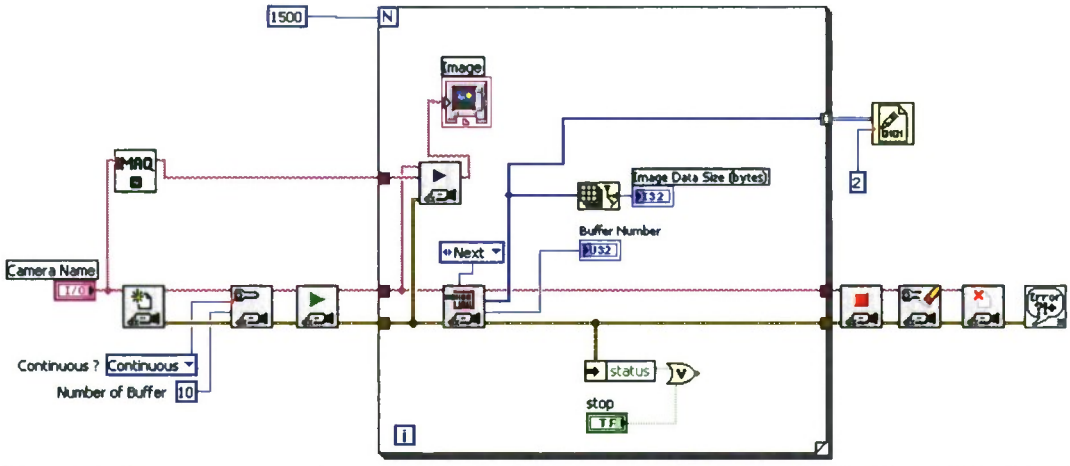


Figure 4: Program to acquire a sequence of frames (in this case hardwired at 1500) from the Basler camera and save them to bitmapped image files.

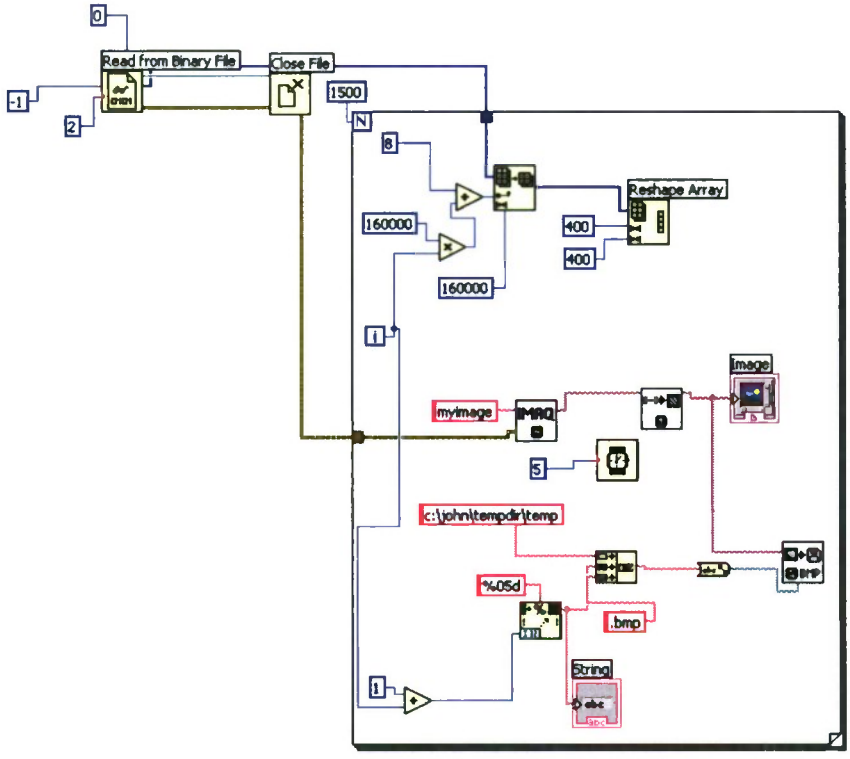


Figure 5: Program to read data stored in a binary image file and convert the data into 400x400 pixel bitmapped files.

Once the data had been saved it needed to be rearranged and stored for further processing. The rearrangement involves reordering the data into a two dimensional image format and then we chose to store the data permanently as consecutively numbered BMP files. This is accomplished with the program shown in figure 5.

With the data saved as image files we can then carry out the processing required to locate the particle position. This was done using a program written in MATLAB. This program allows the user to select a BMP sequence and then select the position of the particle under study. Once this is done the particle images for the whole sequence are smoothed and averaged and the location of the particle center determined using a cross-correlation method.

The output from our tweezers system (built around a Zeiss Axiovert microscope) indicated that it was quite stable. Figure 6 (left) shows a position vs. frame number graph for a particle held strongly by the optical tweezers. The fluctuations are of the order of 2 nm. When the optical power is was reduced to ~ 10 mW, the fluctuations were an order of magnitude greater (figure 6, right).

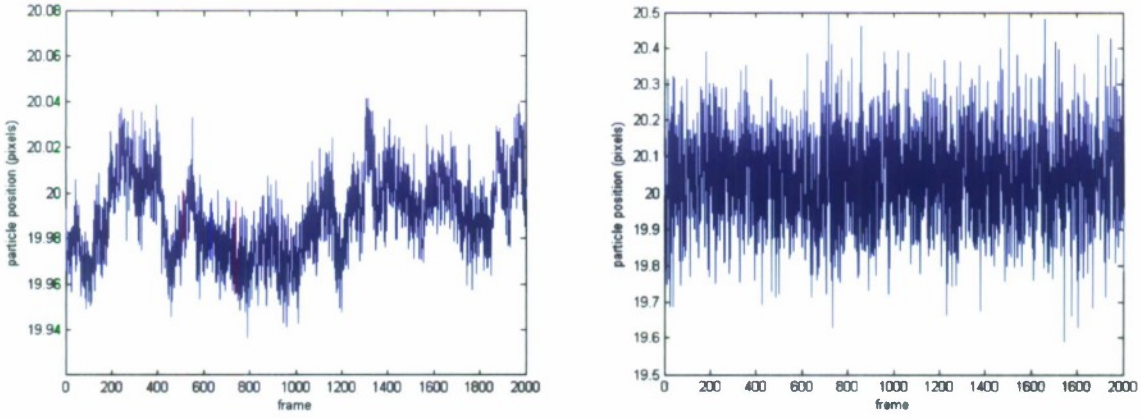


Figure 6.Left panel: Particle position (in pixels) for each of 2000 consecutive frames with the 1 μm polysphere held tightly with the tweezers (power into back focal plane of objective ~ 500 mW). The particle fluctuates about +/-0.02 pixels about its mean position. This corresponds to a fluctuation of 2 nm and is much less than the fluctuations we expect to measure for the proximity measurements. Right panel: Particle position when the particle is trapped with ~10 mW into the microscope. Note that the fluctuations are an order of magnitude larger than those seen in the left panel.

We went ahead and collected and processed data from 1 μm diameter polystyrene spheres which had been trapped using a 1064 nm laser beam and held close to a 10 μm diameter polystyrene bead which had been immobilized onto the microscope cover slip. We typically recorded 2000 frames at a frame rate of 250 fps and an exposure time of < 5ms. The trapped particle was held relatively "gently" by the optical tweezers (power into the microscope ~ 10 mW) so that the trapped particle exhibited a frequency of random oscillation of the order of 100 Hz. This was verified by the use of the Stokes drag method.

A set of data showing the trap strength as a function of position is shown in figure 7. As can be seen there is a great variability in the data and even with this high number of frames we were unable to draw any concrete conclusions from the data. However, the data does show a trend: that is, closer than about 4 μm, the trap strength does appear to weaken by ~ 20% when we compare the trap strength just next to the large particle and far away from it. This is what we would expect and it is consonant with the data we collected previously using a conventional CCD. However, there are some very large excursions which we

have not yet tracked down. These excursions look like a mechanical drift or a sudden change in the power of the laser but close monitoring seemed to indicate that this is not happening.

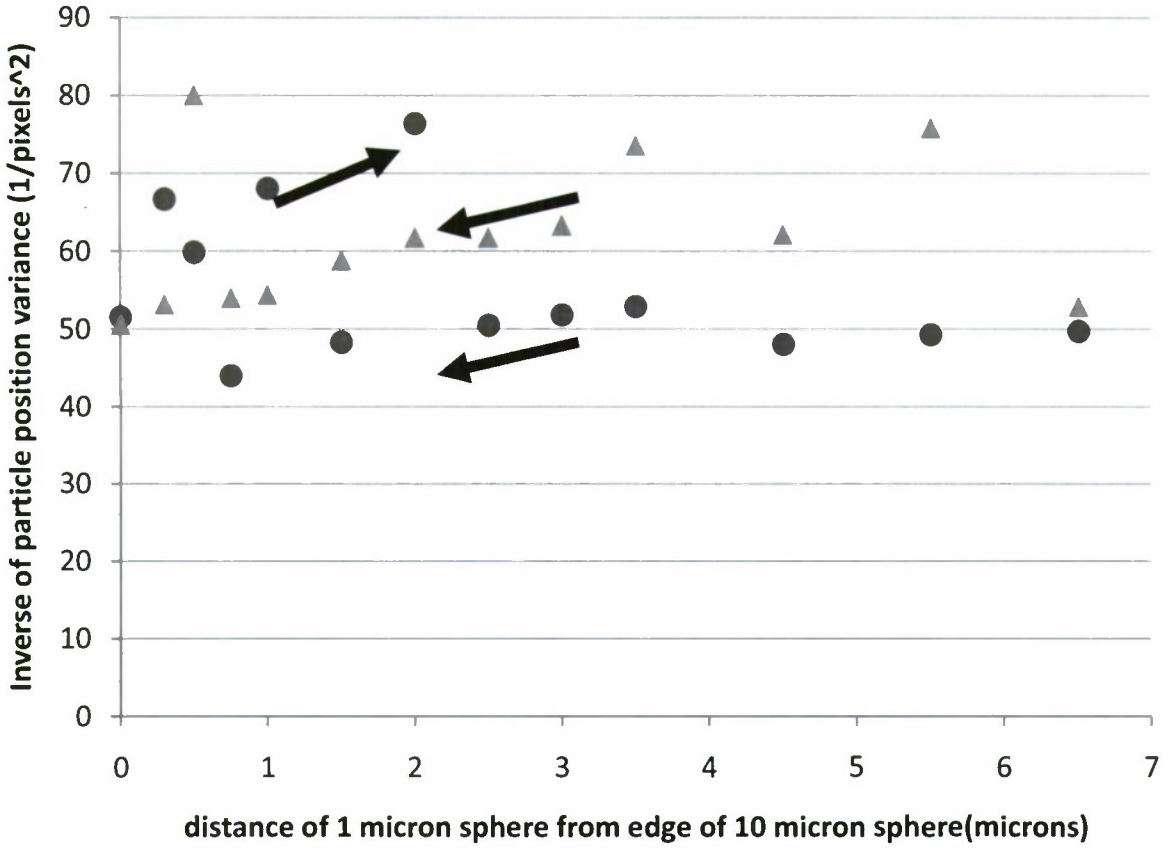


Figure 7. Square of inverse particle position variance (which is proportional to trap strength) as a function of the distance from the edge of the large sphere. Data is shown for particle variance both in the direction perpendicular to the line joining the big and small spheres (triangles) and in the direction of the line connecting the big and small spheres (circles). The arrows indicate whether the small sphere is moved toward or away from the large sphere. From a simple geometrical optics argument we would expect that the trap strength might start to weaken when the trapped sphere is closer than approximately 4 μm.

The fact that the camera, having a short exposure time, would be noisier than a conventional CCD camera did not appear to have a substantial effect. Repeated sequences of frames showed fairly small variation of the variance (comparable to the size of the data points in figure 5).

What we have decided is that we really need some independent method of measuring a trapped particle position so that we can compare this with the data from the camera. This would clearly demonstrate the origin of the large excursions – are they from the camera/image processing or are they from the mechanics of the setup. For this we will do a direct comparison of position data obtained from the camera with position data obtained from a well understood method – that of the position sensitive detector.

By using a PSD to simultaneously acquire position data along with that from the camera we would be able to definitely locate the source of the data spread that we saw with the camera alone. This could not

be done with the eurrent setup for several reasons. First, the mieroseope we had been using (a Zeiss Axiovert) does not have suffieient ports for beam paths to a eamera and a PSD. An alternative would be to use the eondenser lens as an imaging lens for the PSD, but the eondenser lens on the Zeiss is of low numerieal aperture and not suited for imaging. We would, besides, have to modify the mieroseope.

We thus deeided that we would build a new tweezers system sinee we had a high numerieal aperture eondenser available (but whieh would not fit the Zeiss). We have begun eonstruetion on this setup and the partially built system is shown in figure 8. We will eomplete this system over the summer of 2009.

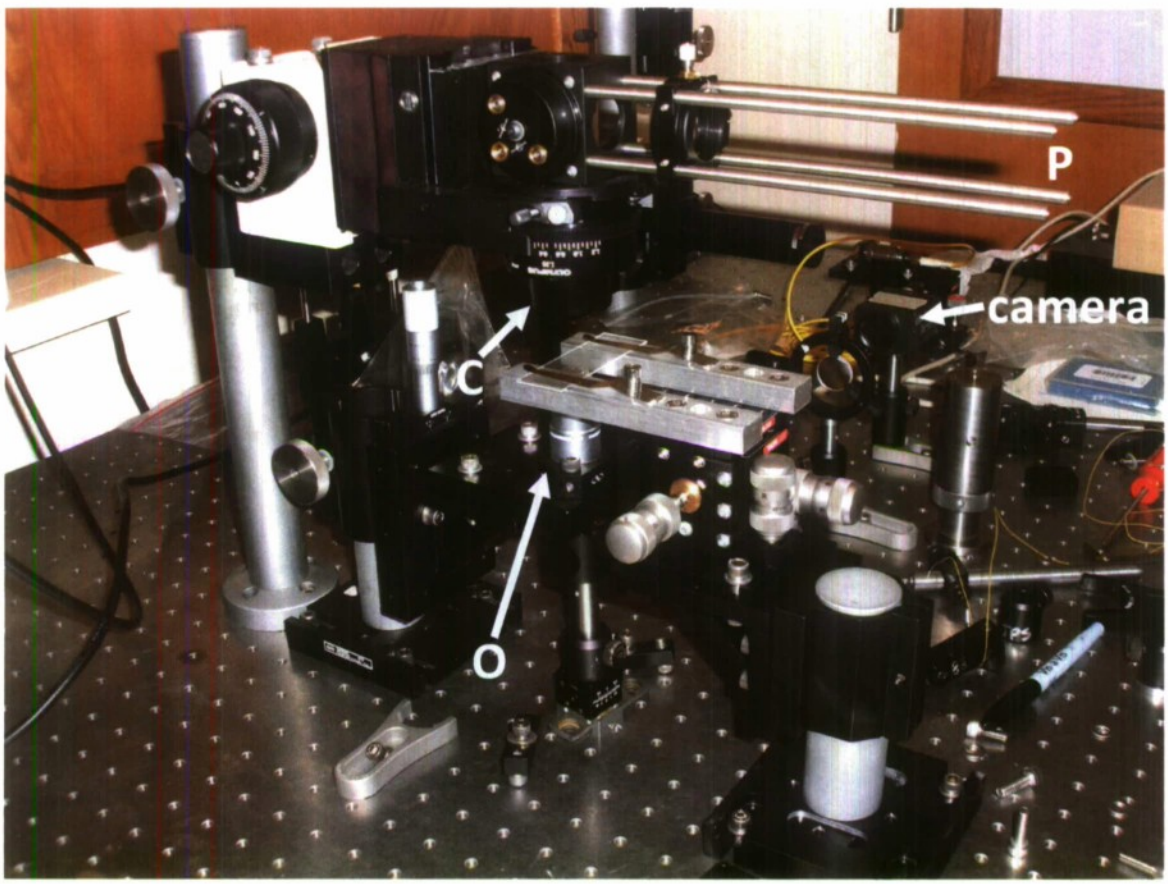


Figure 8: Partially constructed optical tweezers system for simultaneous particle tracking using a CCD camera and a high-speed position sensitive detector. C: condenser, O: objective. The PSD will be mounted at the position P.

References:

Rohrbaeh, A. and E.H.K. Stelzer. "Optieal trapping of dieleetrie partieles in arbitrary fields." J. Opt. Soe. Am. 18, 839-853 (2001)

Rohrbaeh, A. "Stiffness of optieal traps: quantitative agreement between experiment and eleetromagnetie theory." Phys. Rev. Lett. 95, 168102, (2005)

Sharpe, J. P., C. Iniguez-Palomares and R. Jimenez-Flores. "Optieal tweezers foree ealibration using a fast shuttering eamera." Imaging, Manipulation and Analysis of Biomoleeules, Cells, and Tissues V, edited by D. L. Farkas, R. C. Leif and D. V. Nicolau. Proc. of SPIE, Vol. 6441, 644114, (2007).

Impact of the Neural Toxin Paraoxon on the Electrophysiology of the Neuromuscular Junction

Project Investigator:

Robert Szlavik
Department of Biomedical and General Engineering
California Polytechnic State University
San Luis Obispo, CA

Impact of the Neural Toxin Paraoxon on the Electrophysiology of the Neuromuscular Junction

(Final Report)

Project Objective

The major objective of this project was the experimental validation of a proposed model of the neuromuscular junction under the influence of organophosphate based neural-toxins. This objective is being pursued to support a research program that involves development, and subsequent validation, of a theoretical model of the neuromuscular junction that quantitatively describes the impact of organophosphate based neural-toxins on the physiology of the motor end plate. The basic hypothesis behind this research is that ***a sufficiently detailed model of the neuromuscular junction can be used to quantify the impact of organophosphate based neural toxins on the physiological functioning of the junction which can be validated experimentally through the use of appropriate electrophysiology experimental protocols.***

Major Project Objectives

Objective 1: **Development of a unified modeling and simulation approach that encompasses the biochemical as well as the electrical behavior associated with the neuromuscular junction in the presence of acetylcholinesterase inhibiting neural toxins.**

- A combined simulation of the type that has been developed provides for a better understanding of the graded impact of neural-toxin exposure on the motor end plate.
- The first objective has been completed and is the subject of the dissertation of one of the principal investigator's graduate students at Louisiana Tech University.

Objective 2: **Development of this model in an equivalent circuit form whereby it can readily be implemented in conventional circuit simulator packages based on SPICE (Simulation Program with Integrated Circuit Emphasis).**

- In the case of the proposed study, the SPICE simulation platform allows for a simulation approach that would encompass the neurotransmitter release, diffusion, neurotransmitter/neural toxin reaction kinetics, post-synaptic membrane receptor/ligand reaction kinetics and muscle fiber electrical activity under a unified simulation framework.
- Implementation of the model in an equivalent circuit form, as a block (netlist) level SPICE model, would also facilitate development of the circuit *in silico*.
- An equivalent circuit model of this type could eventually be implemented in integrated circuitry and could function as a *front end* in a neural-toxin detection system used to monitor exposure of individual personnel in an environment that is considered potentially hazardous.

- Developing a SPICE based simulation infrastructure for the combined model described above is a significant enhancement to currently existing simulation methodologies available to electrophysiologists such as NEURON and GENESIS.
- The utilization of SPICE, as the simulation framework, would allow for the simulation of hybrid circuits that include biological components as well as synthetic electronic devices which could ultimately prove useful in the development of hybrid biological/synthetic circuitry.
- Work on the development of the equivalent circuit model is currently being undertaken by the principal investigator and several graduate students associated with the Electrophysiology and Neural Electronics Research Group (ENERGy) at California Polytechnic State University (Cal Poly).

Objective 3: **Development and implementation of suitable experimental electrophysiology protocols for validating the proposed unified simulation model.**

- The animal model that was utilized for this study is the neuromuscular junction of the medicinal leech *Hirudo Medicinalis*.
- The leech neuromuscular junction is a nicotinic acetylcholinesterase based synapse. It is consequently expected that the organophosphate neural toxin paraoxon will exhibit a measurable impact on the electrophysiology of leech motor end plate.
- The proposed experimental protocol, shown in Figure 1, is designed with the intention of measuring this effect indirectly by using a voltage clamp to maintain the postsynaptic membrane at a fixed value and measuring the change in the time course of postsynaptic motor end plate current. Under voltage clamp, the measured membrane current is linearly related to, and consequently an indirect measurement of, the time course of the postsynaptic membrane conductance.

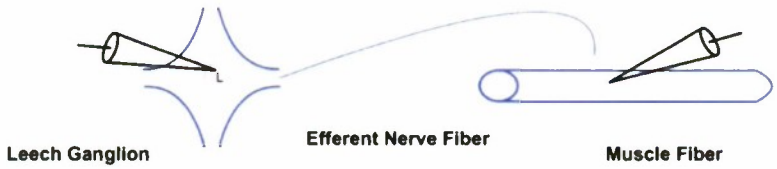


Figure 1. Experimental protocol involving measurement of the time course of the motor end plate current. A current clamp micropipette is used to stimulate the L cell in the leech ganglion. The resultant motor end plate current is recorded using a voltage clamp at the postsynaptic muscle fiber under control conditions and after exposure to the neural toxin paraoxon.

- Development of the leech ganglion muscle wall preparation necessary for the proposed experimental protocol has been completed by a graduate student (Chandra Miller) who was funded for this work from the resources for this project. Other students (Brenton Parks) were involved in assisting with the development of this protocol as well as various laboratory related tasks and these

students were also funded from project resources. A digital image of the preparation is shown in Figure 2. Images were captured using digital imaging hardware and software that was purchased using the project resources.

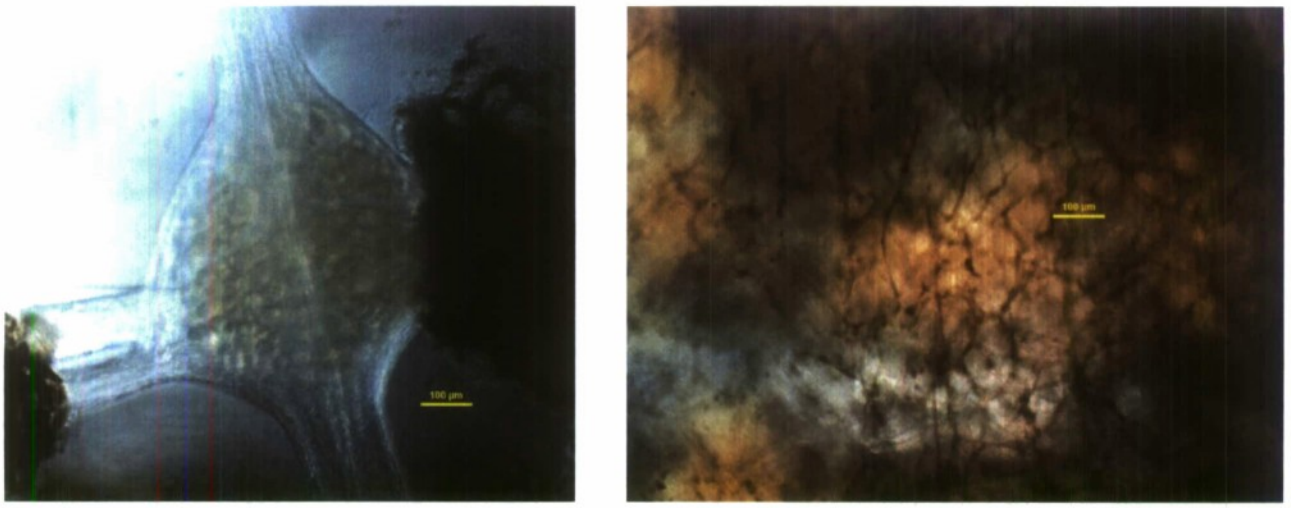


Figure 2. Digital images of leech ganglion/muscle wall preparation. The image on the left shows the leech ganglion with a micropipette inserted into the L cell for stimulation. The right image shows the longitudinal muscle fibers that are innervated by the ganglion. A micropipette may be seen inserted in a longitudinal muscle fiber below below the 100 μm scale bar on the picture.

- In the protocol, the L cell within one of the leech ganglia is stimulated with current to control excitation. Because the L cell innervates muscle fibers along the outer musculature of the animal, stimulation of the L cell will result in end plate post synaptic potentials in the muscle fibers.

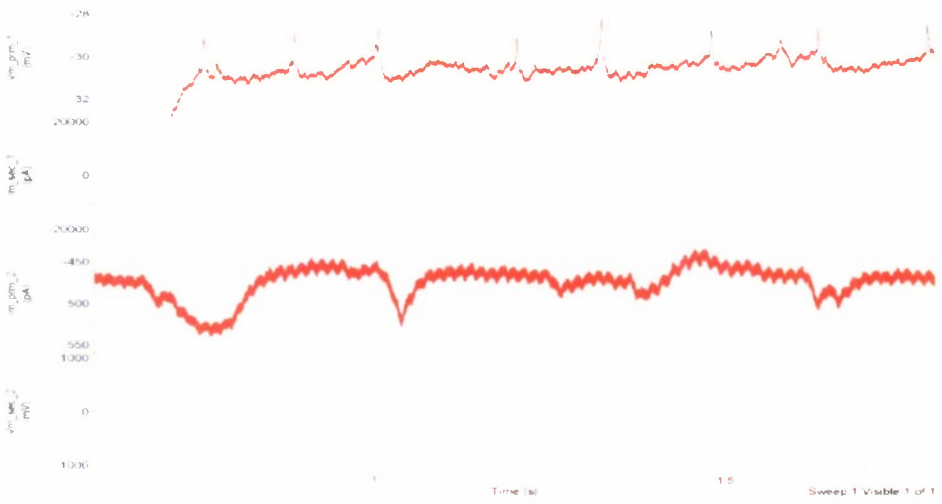


Figure 3. Recordings of L cell action potentials (topmost graph labeled Vm_prm_1) and associated longitudinal muscle fiber membrane current (third graph from top labeled Im_prm_2) for the control (no neural toxin) component of the experiment.

- By voltage clamping one of the innervated muscle fibers, the membrane current can be measured which is an indirect measurement of the variable membrane conductance. The data displayed in Figure 3 was taken from an experiment conducted on 12/19/08.
- The experiments for testing the impact of the paraoxon neural toxin on the motor end plate are currently ongoing.

Overall Project Goals in Relation to Naval Funding Objectives

This project has resulted in the successful implementation of experimental protocols that will be useful in establishing the efficacy of the proposed theoretical model of the neuromuscular junction under the influence of organophosphate based neural toxins.

If, from these ongoing experiments, the model is successfully validated, then the potential exists to develop it into an equivalent circuit form whereby it could be utilized as a sub system in a neural toxin detection instrument for the purpose of determining the physiological impact of a given quantity of neural toxin on an organism. The model has further potential for utilization as a tool to determine the efficacy of therapeutic interventions for combating the effects of organophosphate based neural toxin exposure.

The theoretical work for this project has resulted in contributions to several other research programs currently ongoing in the principal investigator's group (ENERGy) at Cal Poly. While these programs are ancillary in nature to the main objectives of this project, they are acknowledged because their successful development was positively impacted by the support for this research. These publications included a paper presented at the IEEE Engineering in Medicine and Biology Society Conference in 2008 as well as a paper that is currently under review for the IEEE Engineering in Medicine and Biology Society Conference for 2009.

Environmental Proteomics: the Minimal Stress Proteome in the Marine Model Organisms *Ciona Intestinalis* and *C. Savignyi*- Networks of Co-Expressed Proteins

Project Investigator:

Lars Tomanek
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA

ENVIRONMENTAL PROTEOMICS: THE MINIMAL STRESS PROTEOME IN THE MARINE MODEL ORGANISMS *CIONA INTESTINALIS* AND *C. SAVIGNYI* – NETWORKS OF CO-EXPRESSED PROTEINS

FINAL REPORT (2009)

RESULTS

During the last year we studied the proteomic stress response of the marine sea squirt species *Ciona intestinalis* and *C. savignyi* in response to salinity and temperature stress. The objective of this project is to describe the **minimal stress response (MSR)** that is induced in response to the macromolecular damage caused by the acute exposure to multiple stresses. The distinction between the biochemical pathways common to all stresses as well as the ones that are specific to a particular stress provides valuable insights about the common and specific biomolecular targets of each stress. In addition, the characterization of the MSR leads to a better description of the behavior of system properties and their dynamics in response to perturbations through the environment. Such descriptions provide us with insights about the properties that determine the robustness and resilience of biological, specifically cellular systems.

Salinity stress

The characterization of the MSR requires a detailed analysis of the proteomic response of *Ciona* towards several stresses. We have chosen two of the most ubiquitous stresses: salinity and temperature stress. Here we report the progress we made in describing the response to salinity (hyposaline conditions) and temperature stress.

The two *Ciona* species are found at pilings of harbors and marinas world-wide (*C. intestinalis*) and in the northeast Pacific (*C. savignyi*) where they are exposed to changing salinity conditions during heavy winter rains and subsequent run-offs (Dybern, 1965; Hoshino and Nishikawa, 1985; Lambert and Lambert, 1998). For example, in San Francisco Bay winter run-offs can lower the salinity of the water from 32 ‰ to 10 ‰. Massive die-offs of *Ciona* populations occur during these episodes of hyposaline conditions that are typical for the winter and they are followed by re-colonization in the spring (Lambert and Lambert, 1998). Personal observations in the field and embryonic studies (Dybern, 1967; Marin et al., 1987) suggest that the two *Ciona* species differ in their tolerance towards hyposaline stress and that it may be a contributing factor enabling *C. savignyi* to expand its distribution range faster than *C. intestinalis* (Lambert and Lambert, 2003).

Salinity changes cause massive cellular adjustments due to the osmotically driven influx of water during hyposaline conditions. The increase in cell volume that follows exposure to hyposaline conditions requires that the cellular scaffold be restructured (Pedersen et al., 2001). Active volume decrease is facilitated by the efflux of inorganic ions such as K^+ and Cl^- (Okada et al., 2001).

In order to characterize the changes in the proteome of *Ciona* in response to hypolsaline conditions we exposed both species to decreasing salinities (100%, 85% and 70%) for 6 h. Following the exposure to hyposaline (or –osmotic) conditions specimen were brought back to 100% salinity to recover for 4 h. A union fusion or proteome map of all the proteins detected through two-dimensional gel electrophoresis is shown for *C. savignyi* is shown in Fig. 1.

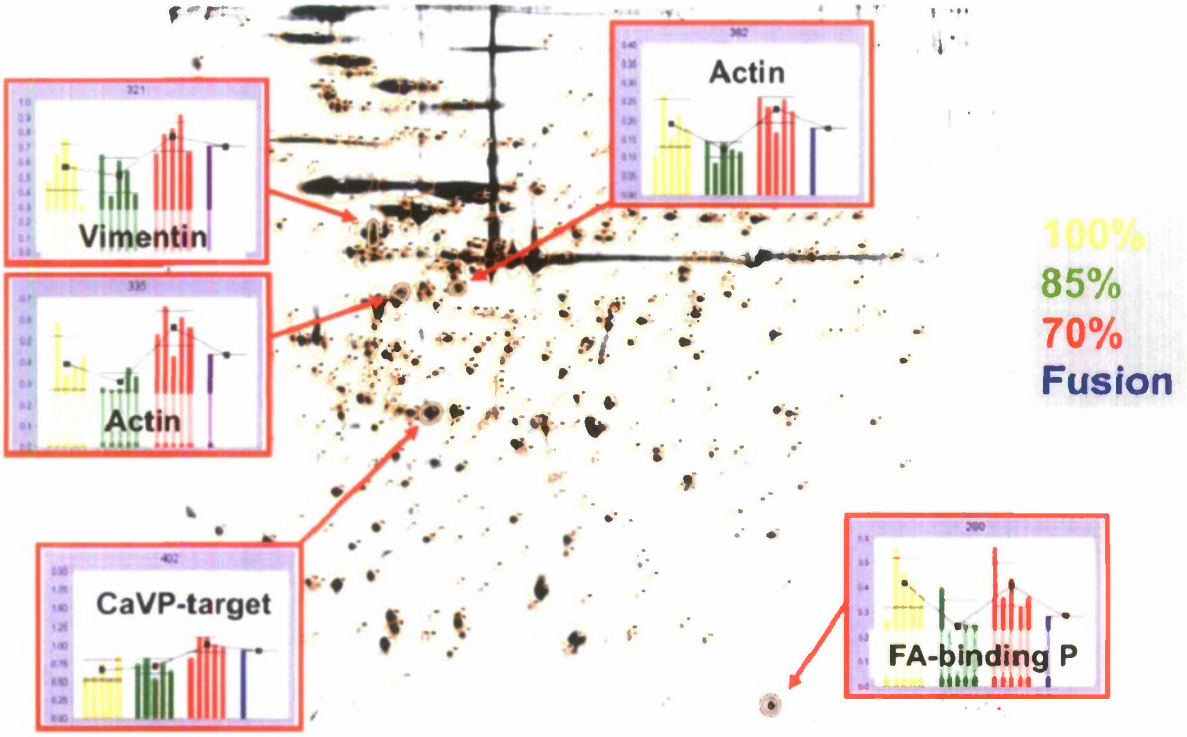


Figure 1: 'Fusion' gel that is equivalent to a proteome map of fifteen gel images (IPG gradient: pH4-7; length of IPG gel strip: 24 cm; Protein loaded: 800 µg) from whole animals (tunic removed) of *Ciona savignyi* exposed to 100% (control), 85% and 70% salinity for 6 h followed by 4 h of recovery at 100% salinity (constant 13˚C). Protein spots that we identified are shown with their respective expression profiles ($p < 0.05$). N= 5 for each salinity treatment.

As shown in the expression profiles in Fig. 1, levels of expression varied among proteins that changed in response to hyposaline conditions. Some of the proteins are cytoskeletal elements (actin isoforms and vimentin), others are involved in calcium signaling (Ca^{2+}-vector-binding protein) or the transport of hydrophobic molecules (fatty-acid binding protein).

We are currently elucidating the identities of the proteins that are changing in response to hyposaline stress in *C. intestinalis*. We know that *C. intestinalis* changes about three times as many proteins in response to hyposaline stress than *C. savignyi* (66 versus 19), but we have not identified all the proteins that differ.

Temperature stress

Temperature is arguably the most ubiquitous environmental factor influencing organismal biology, specifically in ectotherms that rely on ambient temperatures (Hochachka and Somero, 2002). For example, it is one of the main factors determining the distribution range of marine invertebrates, such as intertidal molluscs (Tomanek, 2002; Tomanek, 2005; Tomanek and Sanford, 2003; Tomanek and Somero, 1999). Adjustments in global gene expression levels are widespread with varying temperatures (Podrabsky and Somero, 2004). The two *Ciona* species differ in temperature tolerance, although it is still unclear to what degree. We have made major progress in elucidating their proteomic response to acute heat stress. Here we present the proteome map of all the protein spots that we determined to be up-regulated and we were able to identify with mass spectrometry in *Ciona intestinalis* (Fig. 2).

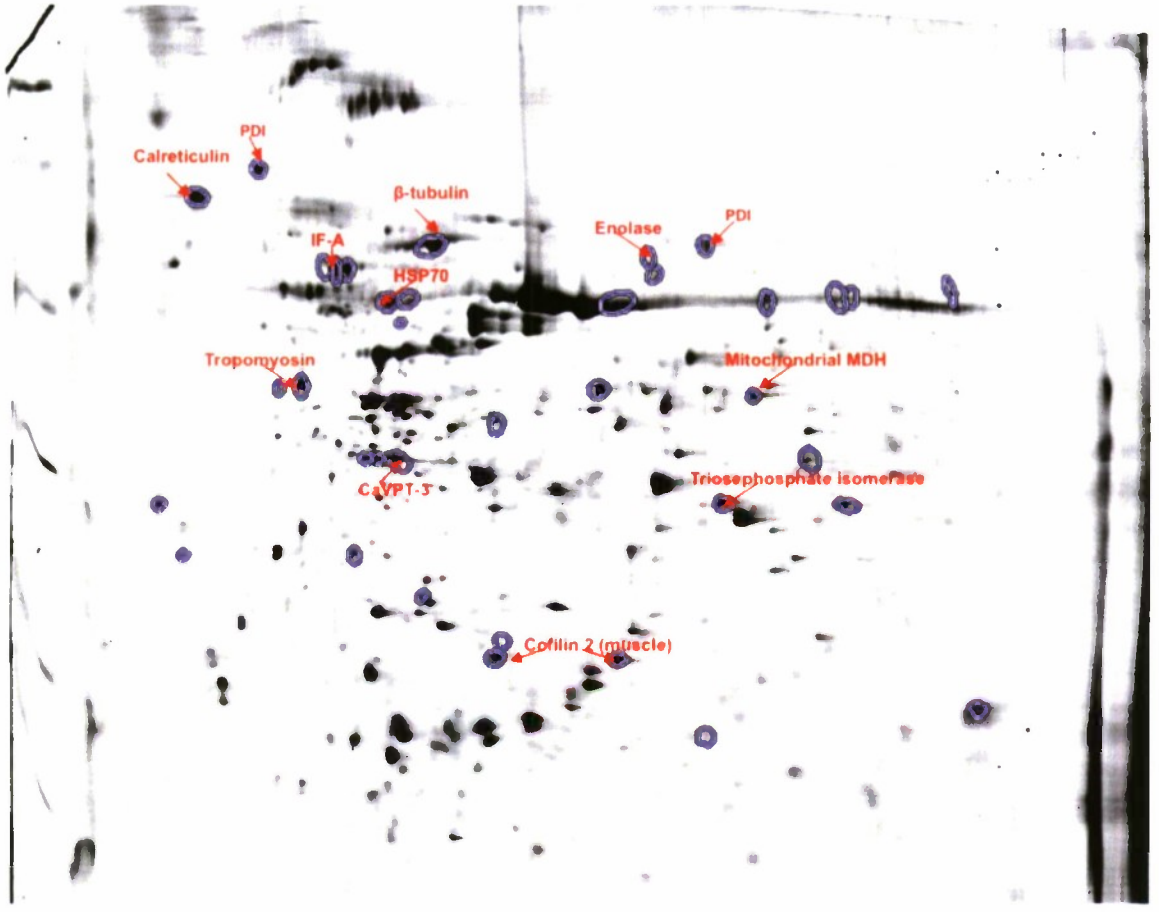


Figure 2: Proteome map with spot identifications of heat-induced proteins (IPG gradient: pH4-7; length of IPG gel strip: 24 cm; Protein loaded: 800 µg) from whole animals (tunic removed) exposed to 13˚C (control), 22˚C, 26˚C and 30˚C for 2 h followed by 6 h of recovery at 13˚C. We detected over a thousand protein spots. N= 6 for each heat treatment ($p < 0.05$).

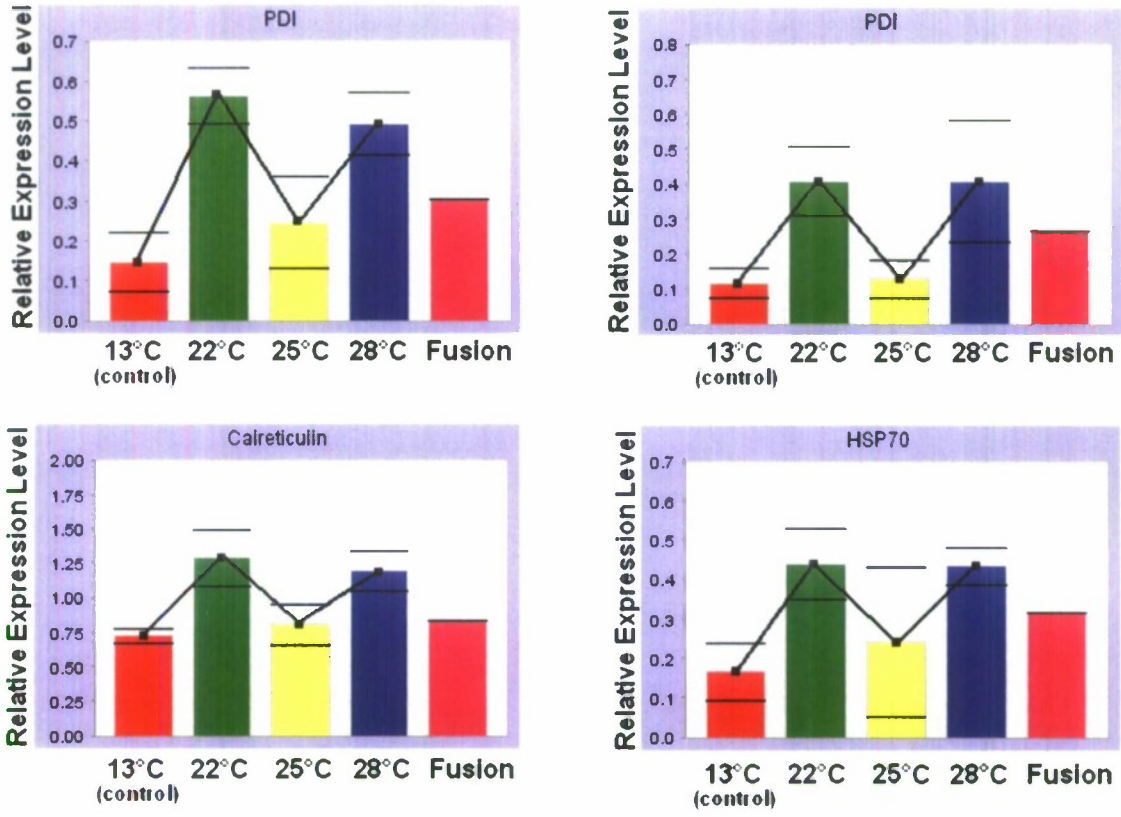


Figure 3: Expression profiles of four molecular chaperones (PDI- protein disulfide isomerase, calreticulin and heat-shock protein 70) that showed statistically different expression levels in *Ciona intestinalis* after incubation to a heat shock in comparison to the 13°C control.

Expression profiles of four molecular chaperones that stabilize denaturing proteins during heat shock are shown in Fig. 3. We have also obtained and analyzed all the gels for the congeneric species *C. savigny*. The species differs from *C. intestinalis* in that it shows fewer molecular chaperones increasing expression after heat shock. We are currently in the process of confirming all the identifications and comparing the protein expression profiles of both species.

SUMMARY

Our results show for the first time how a marine invertebrate that is closely related to vertebrates responds to temperature and salinity stress at the proteomic level. The emerging patterns of protein expression present clusters that we hypothesize are proteins that are related by function or biochemical pathway. The methodology that we applied opens a way to a systems biology approach to study environmental stress in marine organisms. The resulting protein expression profiles may be used as indicator for specific environmental conditions. This approach can be used in the future to develop biomarkers or –sensors that are indicators for a range of environmental stresses, e.g. physical, chemical to biological. We have been sueeessful at identifying the protein spots that are changing expression in response to stressful conditions. We are confident that we will

finish the analysis of the changes in protein expression over the summer and publish the results by the end of the year.

ACKNOWLEDGEMENT OF SUPPORT

This work was supported by the Department of the Navy, Office of Naval Research, under Award #N00014-07-1-1152. We thank ONR for their support of undergraduate and graduate research at Cal Poly.

REFERENCES CITED:

Dybern, B. I. (1965). The life cycle of *Ciona intestinalis* (L.) f. *typica* in relation to the environmental temperature. *Oikos* **16**, 109-131.

Dybern, B. I. (1967). The distribution and salinity tolerance of *Ciona intestinalis* (L.) F. *typica* with special reference to the waters around southern Scandinavia. *Ophelia* **4**, 207-226.

Hann, J. and Tomanek, L. (2007). Detecting osmotic stress proteins of *Ciona intestinalis* through functional proteomics. In *Annual meeting of the Western Society of Naturalists*, (ed. Ventura, California.

Hochachka, P. W. and Somero, G. N. (2002). Biochemical adaptation: Mechanism and process in physiological evolution. Oxford: Oxford University Press.

Hoshino, Z.-I. and Nishikawa, T. (1985). Taxonomic studies of *Ciona intestinalis* and its allies. *Publication of the Seto Marine Biology Laboratory* **30**, 61-79.

Lambert, C. C. and Lambert, G. (1998). Non-indigenous ascidians in southern California harbors and marinas. *Marine Biology* **130**, 675-688.

Lambert, C. C. and Lambert, G. (2003). Persistence and differential distribution of nonindigenous ascidians in harbors of the Southern California Bight. *Marine Ecology Progress Series* **259**, 145-161.

Marin, M. G., Bressan, M., Beghi, L. and Brunetti, R. (1987). Thermo-haline tolerance of *Ciona intestinalis* (L., 1767) at different developmental stages. *Cahiers de Biologie Marine* **28**, 47-57.

Okada, Y., Maeno, E., Shimizu, T., Dezaki, K., Wang, J. and Morishima, S. (2001). Receptor-mediated control of regulatory volume decrease (RVD) and apoptotic volume decrease (AVD). *Journal of Physiology* **532**, 3-16.

Pedersen, S. F., Hoffmann, E. K. and Mills, J. W. (2001). The cytoskeleton and cell volume regulation. *Comparative Biochemistry and Physiology. Part A, Molecular and Integrative Physiology* **130**, 385-99.

Podrabsky, J. E. and Somero, G. N. (2004). Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish Austrofundulus limnaeus. *Journal of Experimental Biology* **207**, 2237-54.

Tomanek, L. (2002). The heat-shock response: its variation, regulation and ecological importance in intertidal gastropods (genus *Tegula*). *Integrative and Comparative Biology* **42**, 797-807.

Tomanek, L. (2005). Two-dimensional gel analysis of the heat-shock response in marine snails (genus *Tegula*): interspecific variation in protein expression and acclimation ability. *Journal of Experimental Biology* **208**, 3133-43.

Tomanek, L. and Sanford, E. (2003). Heat-shock protein 70 (Hsp70) as a biochemical stress indicator: an experimental field test in two congeneric intertidal gastropods (genus: *Tegula*). *Biological Bulletin* **205**, 276-84.

Tomanek, L. and Somero, G. N. (1999). Evolutionary and acclimation-induced variation in the heat-shock responses of congeneric marine snails (genus *Tegula*) from different thermal habitats: Implications for limits of thermotolerance and biogeography. *Journal of Experimental Biology* **202**, 2925-2936.

Adaptation of the Bardo Airway to the Intraoral Mask: Innovative Airway Management Devices Working in Concert

Project Investigator:

Dr. Keith Vorst
Dr. Bruce Robertson

Industrial Technology Plastics and Packaging
California Polytechnic State University
San Luis Obispo, CA

Project Title: Adaptation of the Bardo airway to the Intraoral Mask: Innovative Airway Management Devices working in Concert

Principle Investigator: Keith Vorst, Ph.D., Industrial Technology Plastics and Packaging Department

1. Abstract

The Innovative Airway Management Kit will be the result of combining our previous work on the Bardo Airway with the NuMask's revolutionary breathing mask to create a system that provides a new approach to emergency ventilation of the patient.

We will solve the clinical problem of breathing for an obtunded patient whose breathing cannot be assisted or controlled by use of a conventional breathing mask, either because of injury or anatomic challenges, such as obesity (Figure 1).

As the NuMask differs from the conventional breathing mask, we will design and create a modified version of the Bardo Airway that will function with NuMask. The addition of the Bardo Airway to NuMask will afford decreased risk of dental and soft tissue damage compared to the only options currently available to users of NuMask.

We will improve on previous methods of product design and materials selection by using new Finite Element Analysis software in conjunction with the Cad/CAM software that is currently used for the creation of such devices. This methodology should help us to reduce our development time by decreasing the number of prototypes that must be created and tested. Also, we will leverage our previous investment in structural and functional airway simulators for testing of the few prototypes that will be necessary.

We will also leverage our previous materials selection work to provide non-PVC alternatives for NuMask, Bardo Airway, and other airway devices that will be both less potentially toxic as well as more environmentally friendly upon disposal.



A B C

Figure 1: Conventional Breathing Mask (A), Intraoral Mask (B) and Fitment of Intraoral Mask (C)

2. Project significance

Airway devices are extremely important in maintaining adequate oxygen delivery and carbon dioxide elimination in patients in the emergent, urgent, and elective clinical settings.

Relevance of the research/development area to ONR and DOD areas of interest

In an era when our military is involved in multiple war theaters of indeterminate duration, improvements in combat casualty care and management are extremely important. The dramatic increase in traumatic brain injury and burn injuries over the past five years that has resulted from the tactics of the enemy makes emergency airway management even more important than in prior wars. Also, this research is critical to expanding the options of health care professionals (medical, dental, nurse anesthetists, etc.) at the Department of the Navy's medical facilities.

Design and materials determination for medical devices has traditionally involved many steps. These steps have included the design specification as well as the materials specification. These two specification areas have been largely separate and only come together in a rapid prototyping or true prototyping phase.

Recently, the development of Finite Element Analysis (FEA) software that can model not only the structure but the material components of a prospective device have become available. This is a relatively new field, and represents a large potential opportunity in the medical device arena. Gaining experience with FEA has the potential to make Cal Poly a magnet institution for both medical device companies and materials providers, such as Dow, DuPont and St.Gobain, with whom our project team has already established a relationship during our previous work. One of our goals in continuing these relationships is exploring the possibility of non-government funding to continue our work as a tenant in the Tech Park.

In addition, gaining expertise in this new and evolving area will provide a more rapid development platform for our work as well as unrelated ONR/DOD projects that could benefit from more rapid, lower-cost development.

3. Background/prior work

Our previous grant was aimed at improving force protection during the acute injury and post acute injury phase by designing and developing a superior oropharyngeal airway device for use with conventional bag mask ventilation. However, bag mask ventilation has limitations, especially in trauma, burn injury, and substantial anatomic abnormality of the face (Figure 2).

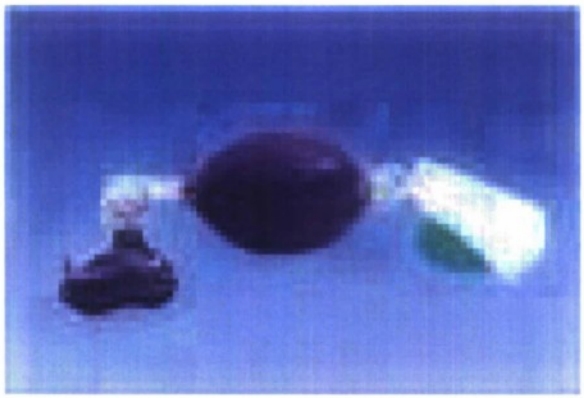


Figure 2: Conventional Bag Valve Mask

Providing ventilation and oxygenation in trauma, with or without burn injury, remains a challenge in the field, especially when facial skin is disrupted, preventing a seal adequate for conventional bag mass ventilation to be successful.[1] [2]

The goal of this project was to modify our previous device to be compatible with intra-oral masks, which must be used when conventional bag mask ventilation is not effective for clinical reasons in the resuscitation of the injured combatant or civilian trauma victim.

NuMask has overcome this and other challenges in ventilation and oxygenation including patients with large faces; beard, significant anatomic abnormality, whether congenital or as a result of trauma and/or burn injury; and obesity with their innovative design that bypasses the route used in conventional bag mask ventilation i.e. the lips and facial skin.

NuMASK™ Intraoral Mask (IOM™) was selected as one of two of the selected **HOT PRODUCTS** at the 2007 EMS Today Conference and is featured in the June 2007 issue of JEMS.

NuMask has been on the market for four years, and is currently gaining widespread acceptance, both in private hospitals and academic medical centers such as Harvard/Massachusetts General Hospital.

However, NuMask currently utilizes a Guedel airway in two sizes, large and small, to keep the airway open beyond the teeth and gums. David Isagholian, the President of NuMask, fully understands the limitations of a Guedel style airway, including dental and soft tissue damage and limited access to the pharynx for suctioning. Dr. Isagholian fully understands the patient safety advantages offered by the Bardo airway and the potential

---

[1] Stepwise airway management in the trauma patient, Trauma, Vol. 6, No. 3, 177-185 (2004)

[2] Journal of Burn Care & Research. 27(5):757-759, September/October 2006.
Keldahl, Mark MD; Sen, Soman MD; Gamelli, Richard L. MD

synergy between NuMask and the Bardo airway that would create an innovative and safer method of airway management.

NuMask agreed to a collaborative effort with our Bardo airway development team to create a new Airway Management Kit combining both devices. NuMask will supply us with the material specification and SolidWorks file for their device and our team will use the SolidWorks file of our device and Finite Element Analysis methodology to create a compatible version of our device with NuMask.

CAM creation of prototypes based on average anatomic measurements that are tested in airway simulators as recommended by the Virtual Disaster Medicine Training Center.[3] Moreover, we have refined the use of simulators in airway device development beyond



Figure 3: AirSim Anatomic Simulator



Figure 4: Laerdal Functional Simulator

4. Progress on Completion of Activities

Activity 1: Adaptation of Bardo airway design to a NuMask compatible design

Several SolidWorks files of the NuMask intraoral mask and the SolidWorks file of the Bardo airway using finite element analysis was used create two compatible designs, one with a unilateral bite block and one with a bilateral bite block based on the Bardo airway platform.

Activity 1: Completed

The results of this activity can be found in Attachments I and II "Using Rapid Prototyping and Airway Simulators in Airway Device Development" and were presented for consideration at the 13th Annual Society for Airway Management Scientific Meeting September 25-27, The Palazzo Resort-Hotel-Casino, Las Vegas, NV.

Activity 2: Design Refinement and Prototype Models

After adaptation of the current design, work will begin on a prototyping process. We will then test the prototypes for ease of insertion/removal, ability to maintain an open airway, and provide adequate ventilation with an Ambu bag and NuMask intraoral airway using

[3] Virtual Disaster Medicine Training Center (VDMTC) : Advanced Airway Techniques Part Two - New Generation Supraglottic Ventilatory Devices; 8/16/ 2007

Laerdal Airway Simulator. Based on this testing, we will select the best design for mold creation and deployment using the material selected as a result of the initial grant process.

Activity 2: Completed

The new design has been modified for adaptation as represented in Figure 5 below. Additional testing was completed to further characterize properties necessary for fitment to the NuMask. The Bite Block requires an unusual combination of material properties because it consists of two parts (a tongue depressor and a biting surface) that have very different functions. In order to make a bite block that performs well in the oral airway device application, the material from which it is made should have the combination of properties listed in Table 1.

Figure 5. Bardo NuMask Adaptation Design

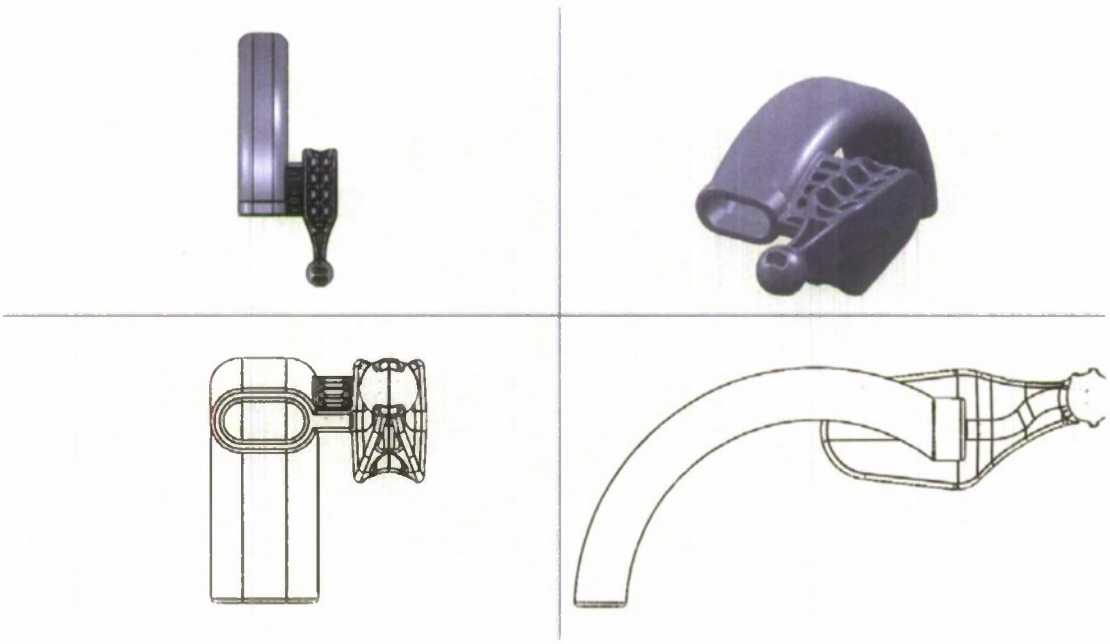


Table 1. Properties of Bite Block for Continuation in Part Patent Burdumy, T, Vorst, K and Robertson, B. Continuation in Part US DKT no: 101659.0001US1 Bardo Airway Management Device

Shore D hardness	Confidential	ISO 868
Tensile strength, psi	Confidential	ISO 527
Tensile Stress at 10% strain, Psi	Confidential	ISO 527
Tensile strain at break, %	Confidential	ISO 527
Flexural Modulus at 23 C, psi	Confidential	ISO 178
Tear Strength lb/in	Confidential	ISO 34-1 method B/a
Vicat Softening Temperature,	Confidential	ISO 306

In order to function properly in the application, the flex force of the bite block part needs to be in the range shown below.

Tongue Depressor Flex Force (lbf)... Confidential .....Modified ISO 604 in compression mode using fixtures manufactured at Cal Poly

Activity 3: Mold Creation

A series of prototype molds will be evaluated prior to a final aluminum mold for the injection molding machine located in the Plastics Processing Laboratory at Cal Poly. This mold will be able to produce commercial grade airway devices that can be scaled-up to full production models.

Activity 3: Completed

A single cavity mold production mold has been created and used to create 1500 production samples. This mold has been asset tagged and is currently at Plastics Engineering and Development Inc. (PEDI) Carlsbad, CA.

Activity 4: Creation of Airway Kit

The airway kit had been developed with completion of FDA certified labels and packaging. This packaging is ready for distribution and includes the NuMask intraoral airway and Bardo airway for use with conventional bag mask ventilation methods (Attachment III).

5. References cited

[1] Stepwise airway management in the trauma patient, Trauma, Vol. 6, No. 3, 177-185 (2004)

[2] Journal of Burn Care & Research. 27(5):757-759, September/October 2006. *Keldahl, Mark MD; Sen, Soman MD; Gamelli, Richard L. MD*

[3]Virtual Disaster Medicine Training Center (VDMTC): Advanced Airway Techniques Part Two - New Generation Supraglottic Ventilatory Devices; 8/16/ 2007

Attachment 1

Using Rapid Prototyping and Airway Simulators in Airway Device Development

Theodore Burdumy M.D., MBA, Keith Vorst, PhD, James Recabaren, M.D, FACS

Introduction

In addition to intubation injuries, emergence clenching has become increasingly recognized as a source of patient injury. Airway devices resting on the central incisors, which can withstand vertical forces of only 25-35 pounds, are the common denominator in emergence clenching injury. The American Society of Anesthesiology recommends that vertical forces should be relocated posteriorly to the molars, which can withstand vertical forces of 200-300 pounds.

Methods

We designed an airway device which combines the features of a bite block and an oropharyngeal airway. We studied the physical properties of reference materials currently used in bite block and oropharyngeal airways to select a material appropriate to both purposes

We evaluated prototypes using the AirSim (anatomical) and Laerdal (functional) airway simulators. Fiberoptic laryngoscopy was utilized to examine the airway-simulator interface. Bag-valve-mask, endotracheal tube, and laryngeal mask ventilation were performed with the Bardo airway in place as a bite block. We tested the Bardo airway both alone and in combination with advanced airway devices.

Thermoplastic Polyester

DuPont HYTREL SC976NC010 (Hytrel 7246) – Properties Confidential
DuPont HYTREL SC938NC010 (Hytrel 3078) – Properties Confidential
DuPont HYTREL SC945NC010 (Hytrel 4056) – Properties Confidential

Thermoplastic Polyurethane
Bayer TEXIN RxT85A – Properties Confidential
Bayer TEXIN Rx90A – Properties Confidential

Attachment II.

Poster Presentation

Using Rapid Prototyping and Airway Simulators in Airway Device Development

Theodore Burdumy M.D., MBA, Keith Vorst, PhD, Bruce Robertson, PhD, James Recabaren, M.D, FACS

Cal Poly San Luis Obispo Department of Industrial Technology and California Central Coast Research Partnership (Department of Defense, Office of Naval Research), Marian Medical Center, and USC-Keck School of Medicine



Introduction

In addition to intubation injuries, emergence clenching has become increasingly recognized as a source of patient injury. Airway devices resting in the midline near to the incisors, which can withstand vertical forces of 25-35 pounds, is the common denominator in emergence clenching patient injury. The American Society of Anesthesiology recommends that during airway management, the vertical forces should be shifted posteriorly to the molars, which can withstand vertical forces of 200-300 pounds.



During the 19th century, it was reported in the Lancet that elevation of the tongue off of the posterior pharynx was the key to maintaining an open airway.

The last major development in oropharyngeal airway design was published in 1933, when Dr. Guedel documented a curved version of the Hewitt airway, which had been developed during the 19th century.

Jour. A.M.A.
June 10, 1933



Oropharyngeal airways developed since that time, such as the Berman and Ovassapian, have consistently resided in the midline, which is the weakest area of the dentition when patients bite down upon emergence from anesthesia.

We hypothesized that a combination oropharyngeal airway and bite block that would allow conventional bag-valve-mask ventilation would be a step forward in basic airway management.

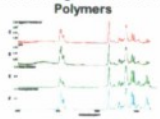
Methods

We designed an airway device which combines the features of a bite block with an oropharyngeal airway. We tested the physical properties of reference materials used in bite blocks and oropharyngeal airways in order to select a material suitable to both purposes. We have named this combination device the "Bardo Airway." Patent Pending

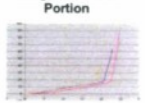
Durometer of Various Materials

Material	Reading	1	2	3	4	5	Average
Blue Bite Block		73	74	75	74	74	74
Red Bite Block		70	69	70	69	70	69.6
Tygon		68	69	68	69	69	68.6

Infrared Spectra of Reference and Proposed non-Halogenated Device Polymers



Summary of Bardo Bend Comparisons- Tongue Elevator Portion



CAD Design for a Rapid Prototype



Evolution of the Rapid Prototypes, Evolution of the Oropharyngeal Airway



Results

Using average anatomic dimensions as the starting point, we created designs using SolidWorks software. The software designs were input to a rapid prototyping machine, which created prototypes of the oropharyngeal airway device from a starch compound.

We tested several iterations of rapid prototype designs for anatomic fit using the AirSim airway simulator and for functionality using the Laerdal airway simulator, with and without advanced airway devices, i.e. endotracheal tubes and laryngeal mask airways, *in situ*.

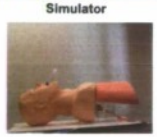
Bardo Prototype in AirSim Airway Simulator



Bardo Prototype in Laerdal Airway simulator



Bardo Prototype with Endotracheal Tube in Place in Laerdal Airway Simulator



Bardo Prototype with LMA in Place in Laerdal Airway Simulator



Endpoints included lack of contact with soft tissue and successful ventilation without gastric insufflation. Also, we performed both anterograde and retrograde fiber-optic examination to look for undesirable areas of tissue contact or conditions that could lead to airway obstruction.



From Rapid Prototype to Commercial Mold and Usable Device Creation

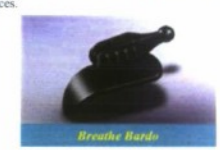
Conclusions

Beginning with neutral anatomic assumptions, we used the method of rapid prototyping to create a new design of oropharyngeal airway device. This device, which we call the Bardo Airway, helps prevent dental and soft tissue damage during the emergence clenching that results from general anesthesia and other sources of obtundation.

In addition, by testing the physical properties of materials currently used for the purpose of bite blocks and airways, we were able to choose an economically viable polymer that is both latex free and halogen free, and therefore more environmentally friendly than the plasticized polyvinyl chloride oropharyngeal airways and bite blocks that are pervasive in the marketplace.

Furthermore, our new airway device provides wider access for suctioning as well as a platform for fiber-optic intubation. We've used the device several hundred times in clinical practice with great success.

Our future work will include extrapolating the sizing of the device for patients of varying body habitus, as well as alternative designs and polymer choices.



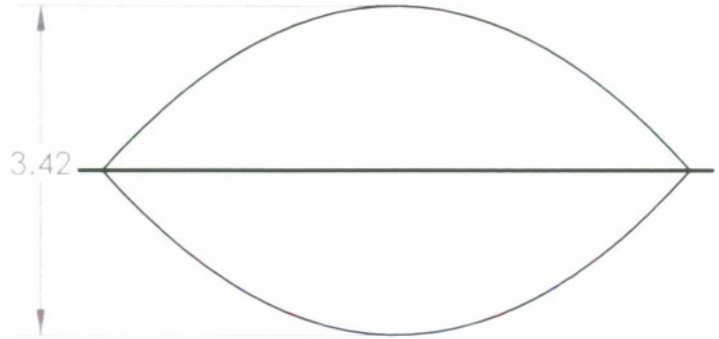
Bibliography

1. Tod F. Tolan, MD; Sue Westerfield; Doug Irvine MD; Terry Clark, DMD, Oregon Anesthesiology Group P.C., Portland, OR, United States Dental Injuries in Anesthesia: Incidence and Preventive Strategies http://www. asaabstracts. comistrands/asaabstracts/printAbstract. htm;j sessionid=ABFA973 ... 3/20/2007
2. Manuel C. Vallejo, M.D., D.M.D. Dental Injury in the Operating Room: What Every Anesthesiologist Needs To Know, 16th ANNUAL PROBLEM-BASED LEARNINGDISCUSSIONS (PBLD), Presented October 13-16, 2007 during the Annual Meeting of the American Society of Anesthesiologists
3. Howard, B, Observations on the Upper Air-Passages in the Anaesthetic State. The Lancet, May 22, 1880: 796-798.
4. Guedel, Arthur E., A Nontraumatic Pharyngeal Airway. Journal of the AMA, June 10, 1933: 1862
5. "The Year of the Airway." American Society of Anesthesiologists Newsletter, V. 72, No. 9
6. Life Cycle Assessment of PVC and of Principal Competing Materials. Commissioned by the European Union. July 2004.
7. Virtual Disaster Medicine Training Center (VDMTC) : Advanced Airway Techniques Part Two - New Generation Supraglottic Ventilatory Devices; 8/16/ 2007

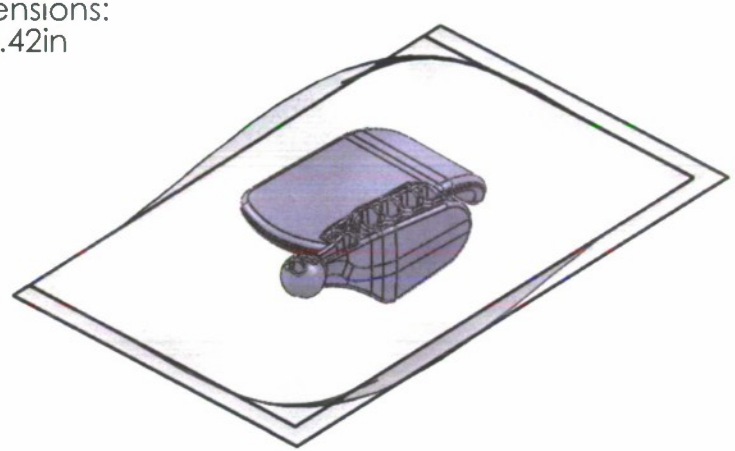
www.bardoairway.com
info@bardoairway.com

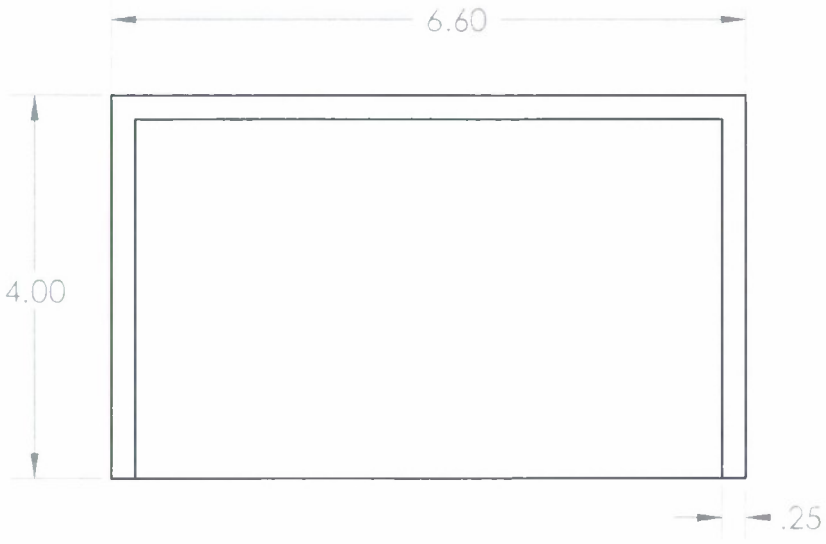
Attachment III

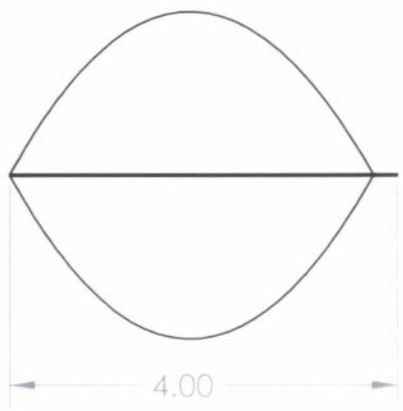
Overwrap Dimensions of Bardo Airway Kit



Outside Dimensions:
4in x 6.6in x 3.42in







UNLESS OTHERWISE SPECIFIED:

DIMENSIONS ARE IN INCHES
TOLERANCES:
FRACTIONAL ±
ANGULAR: MACH ± BEND ±
TWO PLACE DECIMAL ±
THREE PLACE DECIMAL ±

INTERPRET GEOMETRIC TOLERANCING PER:

MATERIAL

FINISH

DO NOT SCALE DRAWING

NAME DATE

DRAWN
CHECKED
ENG APPR.
MFG APPR.
Q.A.
COMMENTS:

TITLE:

PROPRIETARY AND CONFIDENTIAL

THE INFORMATION CONTAINED IN THIS DRAWING IS THE SOLE PROPERTY OF <INSERT COMPANY NAME HERE>. ANY REPRODUCTION IN PART OR AS A WHOLE WITHOUT THE WRITTEN PERMISSION OF <INSERT COMPANY NAME HERE> IS PROHIBITED.

NEXT ASSY USED ON

APPLICATION

SIZE DWG. NO. REV

A Overwrap

SCALE: 1:4 WEIGHT: SHEET 1 OF 1

5 4 3 2 1

Honors Projects

Structure Climbing Robot

Experimental Bifurcations: A Study in Buckling Beams

Ion Thrusters in Space

Modeling, Analysis, and Design of Structural Components Subjected to Severe Loading Conditions

Probing the Surface of Lactic Acid Bacteria With Atomic Force Microscopy

Visualization of Decision –Theoretic Plans

Implementation of Artificial Neural Network for Control of Simulated Wheeled Robot

Structure Climbing Robot

Final Report

California Polytechnic State University, San Luis Obispo
Dr. Saeed Niku
Paul Bessent (GENE)

Executive Summary

The Structure Climbing Robot (SCR) is a robot with the ability to climb any structure. The robot will be about two feet in length and looks similar to a human arm; a gripper is at each end, connected by two links, with an elbow in the middle.

The SCR will have a total of nine motors: two to open and close the two grippers and seven to control the movement of the SCR. We will use gear motors with a worm gear on the end to control the motion of the robot. The worm gear will increase the torque of the gear motor and reduce the revolving speed to a usable amount. The motors will be attached to composite tubes filled with foam that make up the body of SCR as the two links. These links are lightweight so that the motors will not be overloaded with torque. They are also very stiff so that the robot will not bend easily while in motion.

The SCR was modeled and tested in the computer program Dymola. The simulations and calculations performed in Dymola determined the orientation of the motors as well as what motors to use. The motors were ordered and arrived a couple weeks ago. The motors will be tested to determine the size of the worm gear needed. This research will continue in the future. After I take some more classes, I will design and build the joints to begin assembling the prototype.

Introduction

This report is of the progress on the research done in the spring of 2008 to design a functional and practical Structure Climbing Robot (SCR) that has the ability to climb any structure. The main purpose for the SCR is to be able to climb where it is too dangerous or difficult for a person to go in order to do inspections and/or maintenance. Many groups have attempted to build such a robot, but there is no known success. Two major problems arise in such a project. One is to design the best orientation of the motors to achieve enough mobility for the robot to climb. The second is to keep the robot lightweight so that the motors are strong enough to move the robot.

This quarter is a continuation of the research done last quarter. The progress of this quarter includes determining the orientation of the motors and what motors to use. The next step in the research is fabricating gears and joints in order to put the robot together.

Background

A similar project was undertaken by a group at the Carnegie Mellon University Robotics Institute's many years ago. A report in the 1990 Annual Research Review explains their findings and what problems they encountered. The Self-Mobile Space Manipulator (SM^2) was designed for the space station in order to perform routine inspection, maintenance, and light assembly. The desire for the SM^2 is because it is a safer and more cost effective alternative than using an astronaut.

The SM^2 has much value to the space station, but it has many limitations. The design of the SM^2 is similar to a human arm. A gripper is at each end, connected by two links, with an elbow in the middle. The gripper for the SM^2 is a mechanism that screws into holes on a truss structure. This means that the design is limited to walk on the monkey bar-like structure that must be made around the space station. If the truss structure is damaged, the SM^2 is not able to move everywhere.

A 1/3 scale model of the SM^2 was built and tested at the Carnegie Mellon University. Because the robot's links are constructed out of thin-walled tubing to be lightweight, the SM^2 is very flexible. Since the SM^2 is to be used in space that has very little gravity, a servoed gravity compensation system was developed to make the testing similar to a zero gravity testing.

The SM^2 would not be able to be used on earth because the force of gravity would be too much for the flexible links. Also, the SM^2 would not be able to be used on a variety of structures. The successful design of a SCR that can climb any structure would have many more uses on the space station and on earth.

A successful design of a robot with a similar purpose is Dr. Elon Rimon's spider robot (figure 7 in appendix). This three legged robot can climb over obstacles and get into places too dangerous for people. This walking robot gives hope for the success of a SCR.

Design

The basic design of the SCR can be seen in Figure 1 of the appendix. The design is to have two links connected in the middle by an "elbow" with grippers at the other end of the links. We have determined that seven motors must be used in order to give the robot enough degrees of

freedom to have full mobility. Another motor on each gripper is needed to open and close the gripper. This means that there will be a total of nine motors for the robot with seven degrees of freedom. The orientation of the motors is explained in the next section. Last quarter, we determined that using composite tubes filled with foam as the links (body) for the robot will keep the robot stiff and lightweight.

Method and Results

To start the new quarter of research, I quickly began testing different motor orientations in the computer program Dymola. We received the program late last quarter and so we had not done much work on it. I first tested if six motors (not including the two gripper motors) would be able to give the SCR the full mobility necessary. After many trials and errors, I was unable to determine a design that worked. The six motor SCR was able to reach any point, but its gripping ability was limited. At some positions, the gripper would be stuck in its orientation and was not able to turn in order to grab the structure correctly for a firm grip.

With the failure of the six motor SCR, now began the testing of a seven motor SCR. We had previously believed that this would be ample, but we wanted to make sure the seven motor SCR had full mobility. With the design in Figure 2, the SCR was able to reach any point and the gripper was able to rotate to any orientation for a secure grip of any structure. The grippers were omitted from the Dymola programming to make the design trials faster and easier. The red cylinders in the figure represent the motors; the axis of spin for a motor runs through the center of the circular faces of the cylinder.

Although the SCR had full mobility, the axes of spin of the motors at the "wrist" had to be arranged so that they would intersect. This would later allow the motion of the SCR to be solved mathematically more easily. The initial design of the SCR had the axes of spin of two of the motors in the "wrist" intersecting while the third axis of spin was offset by a small distance. To determine a design for the three axes of spin to intersect, I examined some robots in the robotics laboratory that had three axes of spin intersecting. I made the adjustments to the SCR and the SCR still had full mobility. This new design also allows us to reposition the two motors closest to the "elbow" to either be at the "wrist" or at the "elbow" as seen in Figure 3 and Figure 4, respectively. The programming that made a simulated model of the SCR in Dymola can be seen in Figure 5.

Now that the design was determined, it was time to test what motors would work best. Using a table of motors researched last quarter, I inputted the different weights of the motors and ran Dymola to calculate the torque necessary to move the robot at our desired speed. The calculations were plotted on a graph (Figure 6) as the SCR moved through specified positions. The results of the calculations showed that the use of two different motors would provide the most available torque. The extra torque could later be used to attach sensors, cameras, a battery, and other various items. The two motors were heavier than what we initially planned, but the motors had the highest torque to weight ratio; this means that the motors could lift their own weight better than the other motors. This also means that there is more torque available to attach the sensors, cameras, and other items.

With the motors determined, we started the process of buying the motors. We got the "go ahead" to buy the motors, but the website was out of stock. I searched for other websites but they were out of stock. I was able to find one of the motors, but there was only one motor left in stock. I emailed dozens of companies looking for those motors but had no luck. The production company was redesigning the motor and it would not be available until later this year. They also said that there were no substitute motors. After much research, I found a similar motor, but it had a smaller gear ratio. The smaller gear ratio meant a faster speed and less torque, but that was acceptable because we are planning to put a worm gear on the motor to reduce the speed and increase the torque. The motors were ordered and received shortly before the end of the quarter.

While waiting for the motors to arrive, I continued the search for a proper worm gear. A worm gear is illustrated in Figure 8; as the motor turns the worm shaft, the worm gear rotates. The advantage of this type of gear is that the worm gear locks into the worm shaft without putting a torque on the motor. This means the motor can be turned off and the worm gear (therefore the SCR) will remain in its position. Most of the worm gears I found were too small or not strong enough. After finding worm gears that could work, they were too expensive for our budget.

In search for an alternative, we proposed matching a screw to a gear. This would be less efficient, but possibly much cheaper. With a gear in stock, I was able to match the pitch with a lag screw. The screw's threads were not deep enough to get secured contact between the screw and the gear. I met with George Leone of the Mechanical Engineering Department about cutting the threads deeper and he was confident it could be done. For future work on the SCR, the worm

gears need to be made after the correct ratio is determined. Tests will soon be done on the motors to determine how much torque is available at what speed. This will allow us to calculate the gear ratio needed for the worm gear.

Furthermore, the joints need to be designed and fabricated. However, I have not taken the classes that would help with those aspects. I am planning to take a design class in the fall since I have now fulfilled the prerequisites. I will also try taking classes to learn how to machine. I may join the rose float building team or Team Tech to get more machining experience. While catching up with those classes, I may skip taking HNRS 200 in the fall but continue working on the SCR as much as I can.

Conclusion

The tests performed on Dymola have determined a working design for the orientation of the motors. Using the six motor design did not give the SCR full mobility, but the seven motor design provided full mobility. Also, the three motors that make up the wrist were designed such that the three axes of spin from the motors intersect. The specifications of a list of motors were tested in Dymola for the SCR model. It was determined that two of the heavier motors would work better due to their high torque to weight ratio. After some trouble with ordering the motors, the motors arrived recently and will be tested soon.

The upcoming work for the SCR research will be to design and build the worm gears and joints. I will be learning those skills as quick as I can which may pause the progress of the SCR. I still plan to continue the research and hope to build a prototype in the near future.

Acknowledgments

We would like to thank the Office of Naval Research. They have provided the funding that allowed us to buy the motors

A special thanks to Peter Trogos and Joe Gauthier. With their help we received the Dymola program for free and learned to use it. We used this program to run many tests to determine the design of the SCR.

Lastly, we would like to thank the Cal Poly Honors Program for setting up this research opportunity.

Bibliography

Brown, H., M. Friedman, T. Kanade, and Y. Xu, "Self Mobile Space Manipulator Project," *Annual Research Review*, 1990, pp. 66-77.

"Technion 'Spider Robots' are Designed to Save Lives." American Technion Society. 14 March 2008. <http://www.ats.org/robot/?id=178>

APPENDIX



Figure 1. Basic model of SCR

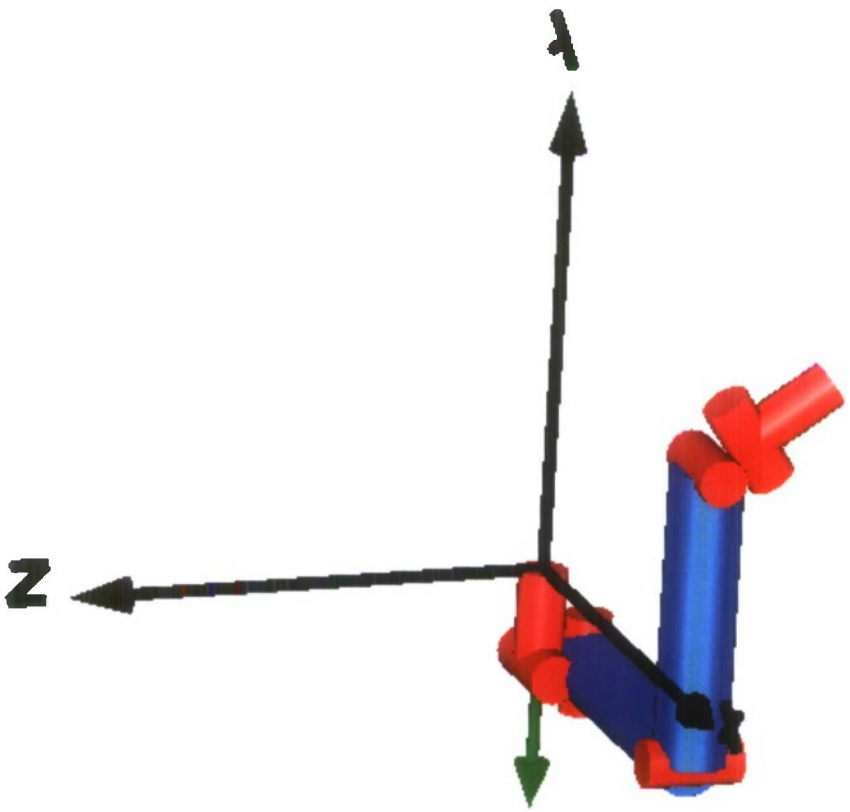


Figure 2. First seven motor SCR design

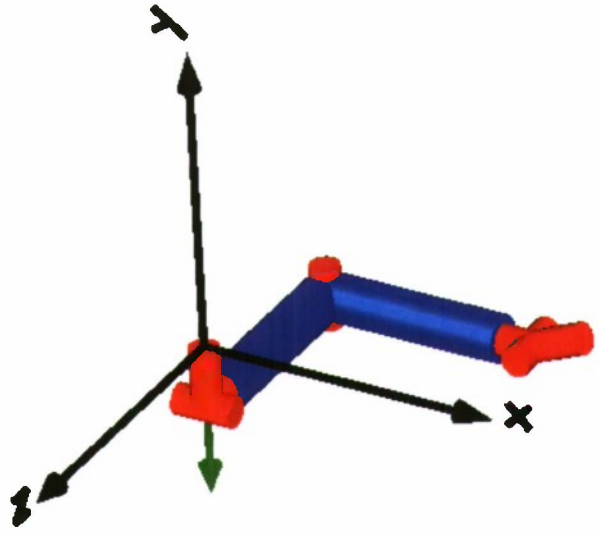


Figure 3. Final SCR motor orientation design

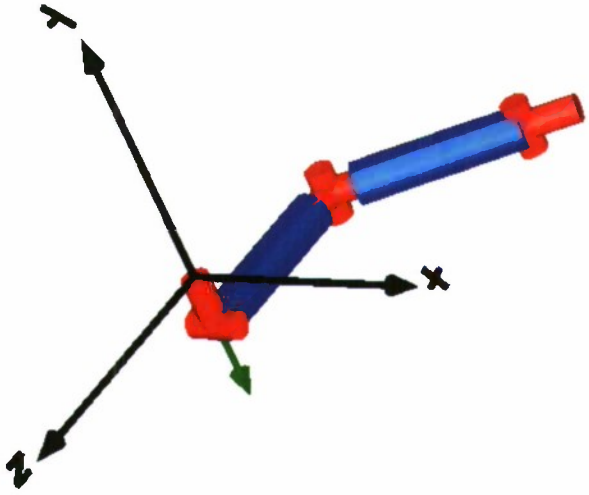


Figure 4. Alternate motor orientation

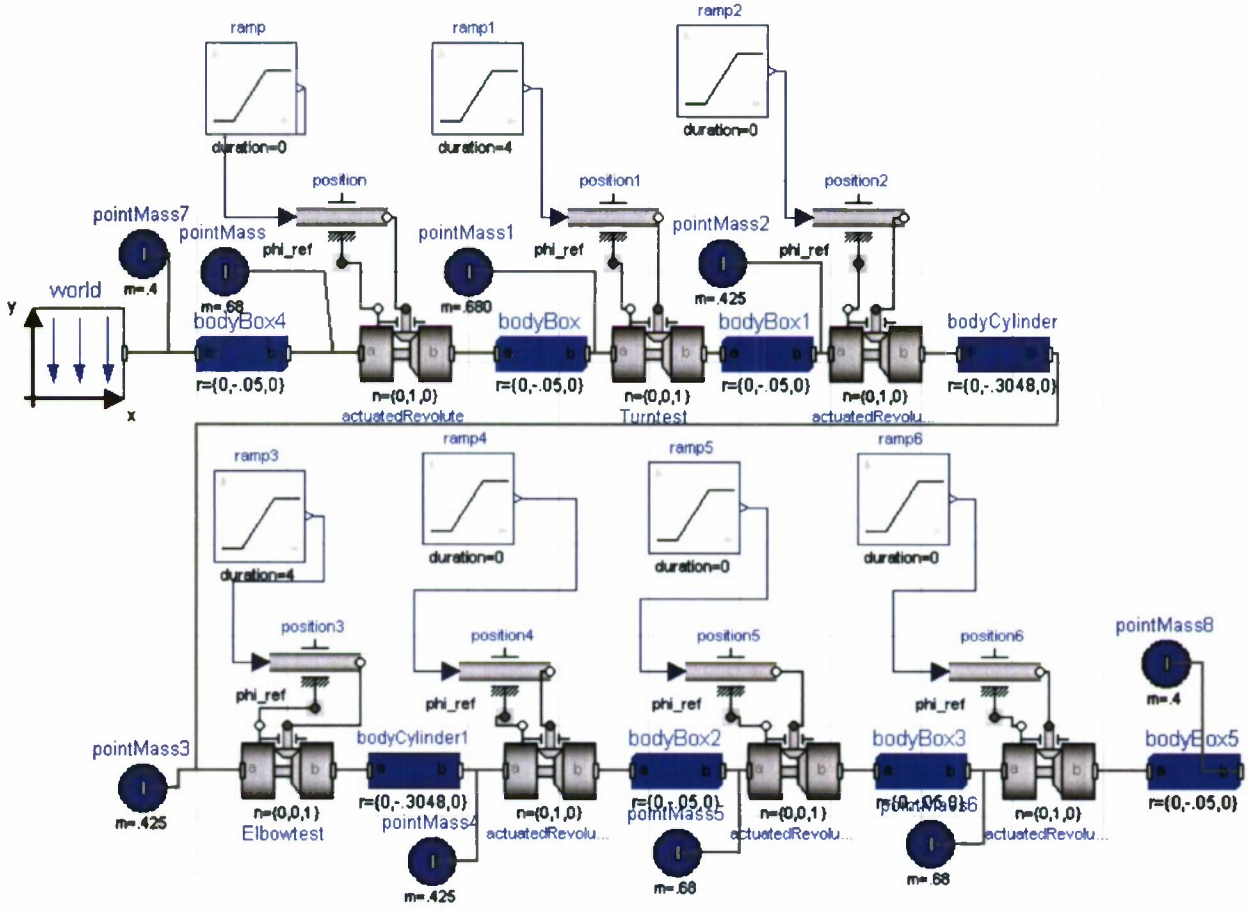


Figure 5. Dymola programming for SCR

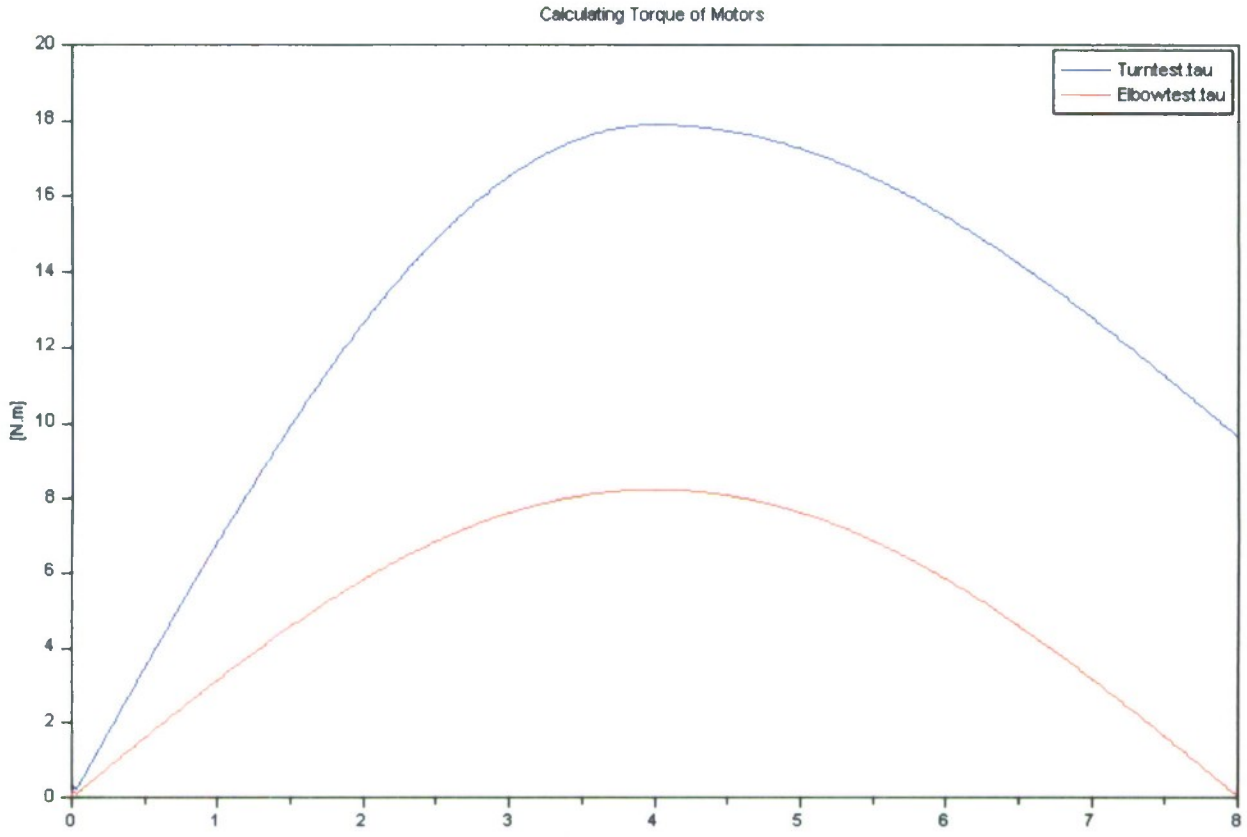


Figure 6. Calculation of torque on motor in Dymola



Figure 7. Dr. Elon Rimon's spider robot

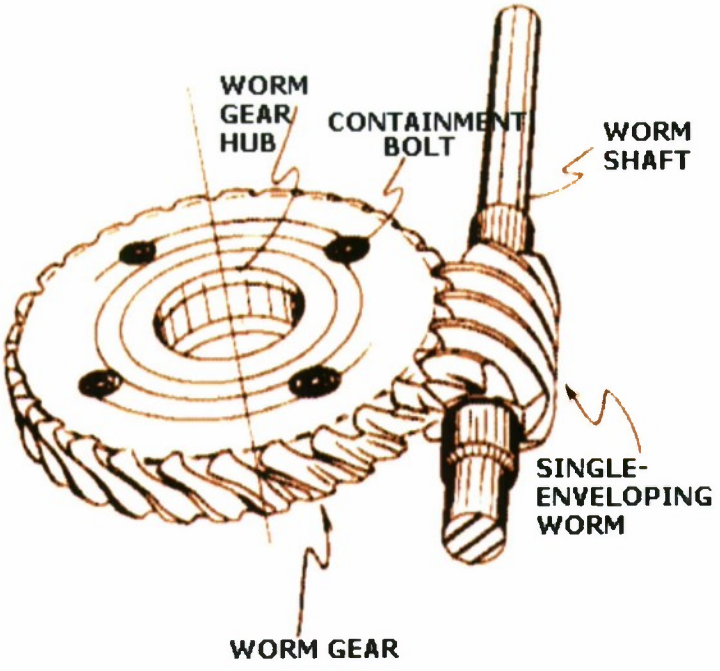


Figure 8. Worm gear

Experimental Bifurcations: A Study in Buckling Beams

Addison Cugini—Physies Major

John Sharpe—Advisor

Executive Summary

This study in experimental bifurcations involved loading a beam with inereasing amounts of weight; the deflection angle of the beam after it buckles was then measured using a laser. The angle of deflection should inerease in a parabolie fashion after the beam buckles. This projeet was done in an attempt to develop an experiment to be used in a course in non-linear dynamics. By addressing the effects of shortcomings in our initial experimental design sueh as frietion, a wobbly rig, and a elamped base, we were able to create a deflection curve which closely followed the expeeted appearance of an imperfect pitchfork bifurcation.

Introduction

The purpose of the project was to make an experiment that would measure the angle of deflection as a function of load for a buckling beam around its critical load. The measure of this angle of deflection should qualitatively and quantitatively illustrate a supercritieal pitchfork bifurcation in the immediate post-buckling range. We tried to repeat the work of John Roorda, who did a similar experiment in the 1960's.[2] Our research was done for the purpose of creating a lab demonstration that eould be used to enrich courses in non-linear dynamics.

Background

Bifurcations are a major part of the study of dynamical systems and involve the sudden change of qualitative behavior when a parameter is changed slightly beyond a certain critical value. One such type of bifurcation is the supercritical pitchfork bifurcation, the solutions of which look something like a parabolic pitehfork.

A supereritieal pitehfork bifureation is normally of the form $x' = rx - x^3$, where r is the control parameter. A top loaded Euler Strut or vertical beam can be modeled as two rigid rods held together by a spring. Using gravitational and spring potential energy a mathematical model can be developed such that: $U = \frac{1}{2} k\theta^2 + P L \cos \theta$. Where θ represents the angle of deflection, P is the load applied to the strut whieh acts as the

control parameter and L is the height of the center of mass. One can then differentiate the equation in order to find the points at which net forces are equal to zero. Using the first two terms for the Taylor expansion of sine, it can be shown that $U' = \theta(k - P\ L) - (P\ L/6)\ \theta^3$. Thus, this model suggests that an Euler Strut exhibits the mathematical characteristics of a supercritical pitchfork bifurcation in the immediate post-buckling range.[1]

While the theory is fairly straightforward, the actual implementation of this theory in experimental form is difficult because the bifurcation occurs with weights close to the critical load value. Therefore, there is a need for the development of an experiment which can accurately illustrate the pitchfork bifurcation of an Euler Strut.

Design

Our first design involved clamping the base of a strut (made of tensile steel 12in x 1/16in x 1in) to a table and using a pillow block to guide a rod onto the top of the beam. We then placed dead weights onto the shaft to induce a bifurcation and recorded the angle of deflection by reflecting a laser off the beam and onto a wall. The deflection angle and load were then displayed on a scatter plot and checked for consistency with the theoretical predictions. Research was done to improve this basic design.

Our inquiry focused largely on the use of different experimental techniques to load the strut with weights; the loading technique needed to have low friction and ensure an even perpendicular force so that the actual loads could be accurately recorded and compared to deflection data.

Experimentation

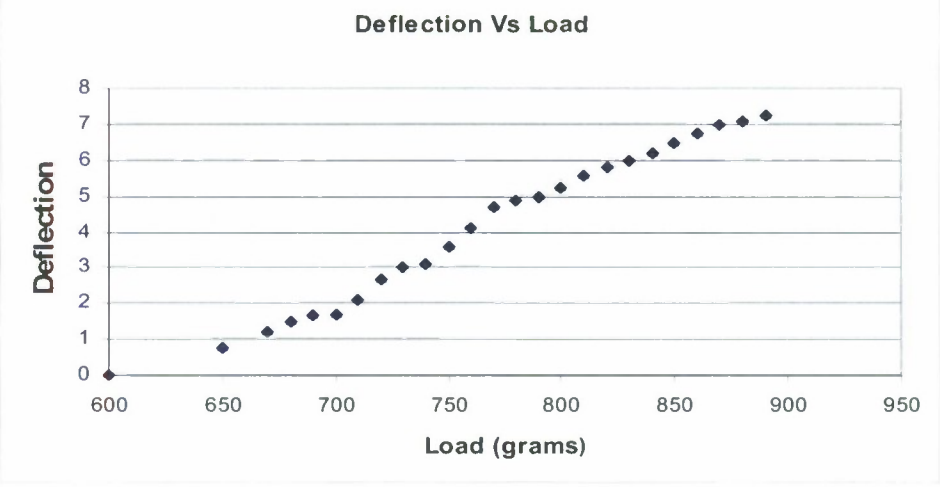
The initial set up had a number of flaws which caused it to produce results inconsistent with the qualitative appearance predicted. The single pillow block design did not provide a rigid enough shaft for the rod to go through; this meant that the force was not applied perpendicular to the ground and could not be measured with the accuracy demanded by this experiment. To remedy this, two pillow blocks were aligned using connector rods and the weight applying shaft was guided through producing a rig that would not wobble as weights were added. It took a number of fine tuning adjustments to ensure that the pillow blocks were in line with one another and would not cause the shaft to jam or experience large amounts of friction. Also, oil was added to the shaft to reduce the friction even more.

Initially, the set up also had the misgiving of causing torsion on the beam making it difficult to tell exactly how much the deflection had changed as weight was increased; the laser beam would reflect off at different angles due to this torsion. The torque on the beam may have also caused minor imperfections which could alter the overall appearance of the deflection curve, invalidating our results. In response to this, we loaded the weights in such a way that would not nudge the loading shaft as this would cause the rod to twist.

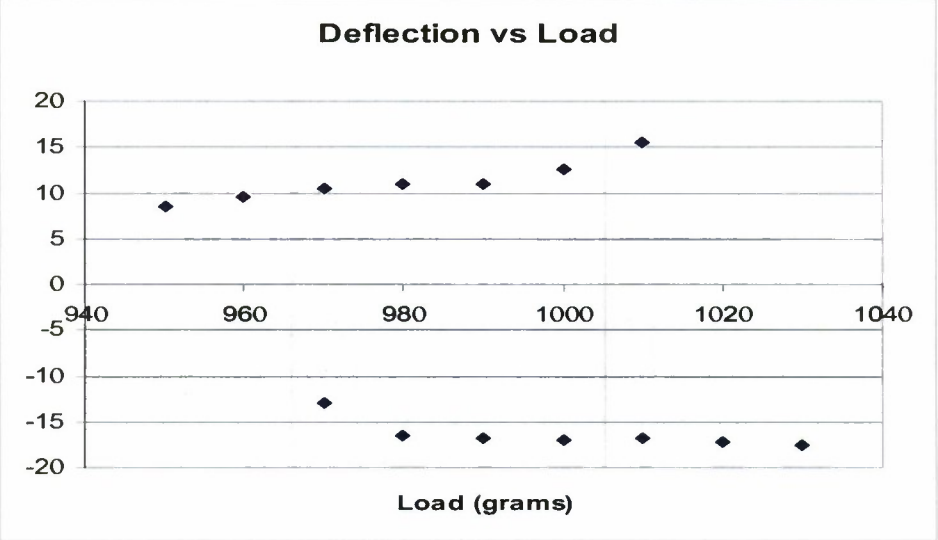
Clamping down the base of the strut led to the rod not buckling in the manner that was dictated by our model. The mathematical representation of the system treats the beam like two ridged rods connected at the center by a spring hinge. This situation did not at first closely follow reality; the bottom section of the beam was remaining relatively vertical when weight was added. To fix this problem, we screwed a 'V' shaped wedge into the table and placed the strut into the wedge. This would allow the beam to buckle easier as the wedge acted like a hinging joint.

Results and Discussion

After tweaking our experimental design with these modifications, we were able to attain a parabolic shape when comparing load and deflection on a scatter plot as seen in the following graph. Because of small initial imperfections in the beam, this graph does not entirely conform to the idealized "perfect" pitchfork bifurcation. Instead of the strut buckling at one particular load, it slowly buckles at first and then deflection starts to grow quickly where the buckling would be expected to occur. This is characteristic of an imperfect pitchfork bifurcation that already has an initial deflection angle before loading.



In order to see whether the buckling beam was in fact an imperfect system, we measured the opposite deflection angle; that is to say, we looked at the deflection angle if the beam buckled in the opposite direction. One can see in the graph below that the opposite deflection angle is not symmetric with the other angle. It is only able to establish a stable equilibrium on the opposite side after a significant amount of load is applied to the strut. This "snap-through" is consistent with what one would see in an imperfect system.



Conclusion

By altering our initial design to accommodate for factors such as friction, torsion, a wobbly loading device, and a clamped base, we were able to develop an experiment that qualitatively illustrated an imperfect pitchfork bifurcation. This improved experimental design is ready to be used in a class setting to help teach non-linear dynamics.

In the future, it would be best to further improve the set up by fine tuning the loading mechanism to be more precise. One could do this by replacing the pillow blocks (which guided the loading shaft onto the strut) with bushings. These bushings have small ball bearings in them which allow for fluid motion that would all but remove friction. Also, a micrometer could be used to measure the weight being applied more accurately and provide a better deflection graph. We would also like to try to get a beam that has hardly any initial imperfections and would thus illustrate a "perfect" bifurcation. These adjustments would mean that the experiment could be done with greater accuracy.

Acknowledgements

Thanks to Dr. Virgin of Duke University who was willing to help us with his expertise. He ensured that our experiment was producing the desired behavior and made suggestions as to how to improve the experiment.

References

1. Tavener, SJ and Mullin T. "Buckling of Coupled Elastic Rods." Physica D, Vol 30, pgs 382-398 (1986)

2. Roorda, John. "Stability of Structures with Small Imperfections." Journal of the Engineering Mechanics Division, Vol 91, pgs 87-106 (1965)

Ion Thrusters in Space

Josh Fernandes, Physics
Greg Stratton, Aerospace Engineering

Introduction:

Ion thrusters are a relatively new mode of space transportation that carry a large potential in the fields of long distance missions within the solar system and long term station keeping for a spacecraft. In this research, we are seeking to optimize the size and placement of the magnets for performance of the ion thruster. This involves building test models and testing them in the vacuum chamber on campus in order to collect data for the graduate students' master's thesis.

Background:

Electric ion propulsion acts as a mode of moving around in space, but doesn't play a role in traveling from earth to space. Ion propulsion takes an ion gas and directs it with magnets to a strong electric field, where the ions accelerate quickly and propel the spacecraft. NASA first tested an ion thruster in a laboratory in 1960, but the thruster was not used in space until NASA's Deep Space 1 Mission in 1998[1]. Deep Space 1 employed its ion propulsion to travel a far distance in space, which is what the ion engines were initially used for. Ion thrusters are ideal for many deep space missions because they carry a very high ISP, which is essentially gas mileage for a spacecraft. Ion thrusters don't create much thrust but can fire for long periods of time in order to produce very high speeds. The main alternative to ion thrusters for these missions is chemical rockets, which have a higher capacity for thrust but a lower ISP.

In addition to deep space missions, station keeping also finds use in ion propulsion. Station keeping involves keeping the satellite in the correct orbit and pointed in the correct direction. These smaller ion thrusters currently have very little research behind them, so they are excellent for testing for improvements. Cal Poly, San Luis

[1] Brophy, John R. "NASA's Deep Space One Ion Engine" September, 2001. Jet Propulsion Laboratory. Pasadena, California.

Obispo recently received valuable materials from the Jet Propulsion Lab in order to research these ion propulsion systems.

Theory:

The basics of ion thrusters are not very difficult to grasp, but they do require some understanding of physics. The thruster is basically a chamber shaped like a bottle with one end open with a metal grating at that end. The other end has a cathode which emits electrons into the chamber, which is filled with a gas. When the electrons hit the gas, the gas turns into ions or charged particles. Magnets on the outside of this chamber direct the ions to the opening of the bottle, where there is a strong electric field. When the ions hit the electric field, they accelerate quickly to the outside of the bottle, which creates thrust or moves the spacecraft. The experimental literature shows that ion thrusters most commonly use Xenon gas for the ions because of its large mass and low corrosive properties. Ion thrusters also commonly fire for several months at a time. Most of the research on ion thrusters has been performed on larger thrusters, so there is very little information on thrusters like the size that we are researching.

Methods, Results, and Discussion of Work to Date:

So far, almost all of our discussion and work has revolved around building and acquiring materials for a test model to collect data from. This has been a tedious process because we have had to work around manufacturing companies and shipments in order to acquire the needed materials. Our first task was to find aluminum bottles that matched the required dimensions and was under a reasonable price. We searched online and discovered an excellent test bottle, but we didn't purchase it because the manufacturing company required a minimum purchase, which was $470. This set us back, but we searched again and recently purchased a suitable bottle so that we may begin construction of the test model. While we were working on this, the two graduate students focused on repairing the vacuum chamber, which we shall be using for the tests on the ion thruster. These repairs are nearly complete, so we may soon only focus on building the test model. Meanwhile, we have been obtaining clearance and practicing with the tools needed for construction of the test model in the aerospace hanger on the Cal Poly campus. This hanger, which is staffed by the mechanical engineering department, requires a short

safety course in order to obtain permission to use the tools that they have. We have recently required this permission and are ready to begin the building process of this research.

Future Plans:

Once we get the ion thruster assembled, and the vacuum chamber running again, we will start testing the ion thrusters. While testing the ion thruster, we will be looking for ways to increase the efficiency of the thrust. The magnets we will be using are movable, and we will be attempting to figure out how many rings, and how far apart those rings must be to maximize thrust. We will present any information we find out to JPL, and hopefully they will find it useful in building their own ion thrusters.

Conclusion:

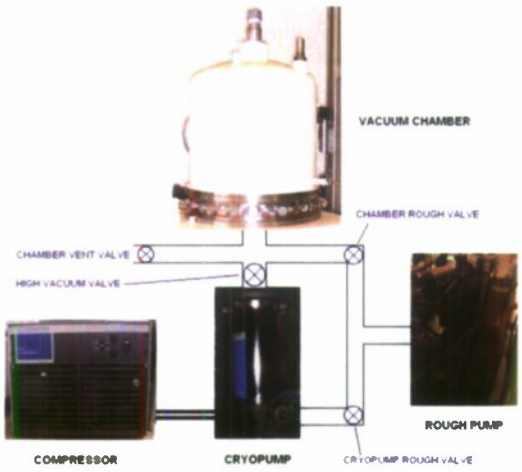
We have learned a lot through this quarter of study. We have learned how to build an ion thruster, and how an ion thruster works. We also understand the physics that allows an ion thruster to work. To allow ourselves to work on the ion thruster, we learned how to safely operate the metal working machinery necessary to cut the bottles the ion thruster is made of.

Acknowledgements:

Thank you to Dr. Dianne DeTurris, Dr. Dan Goebel, Josh Caldwell, and Sagar Desai.

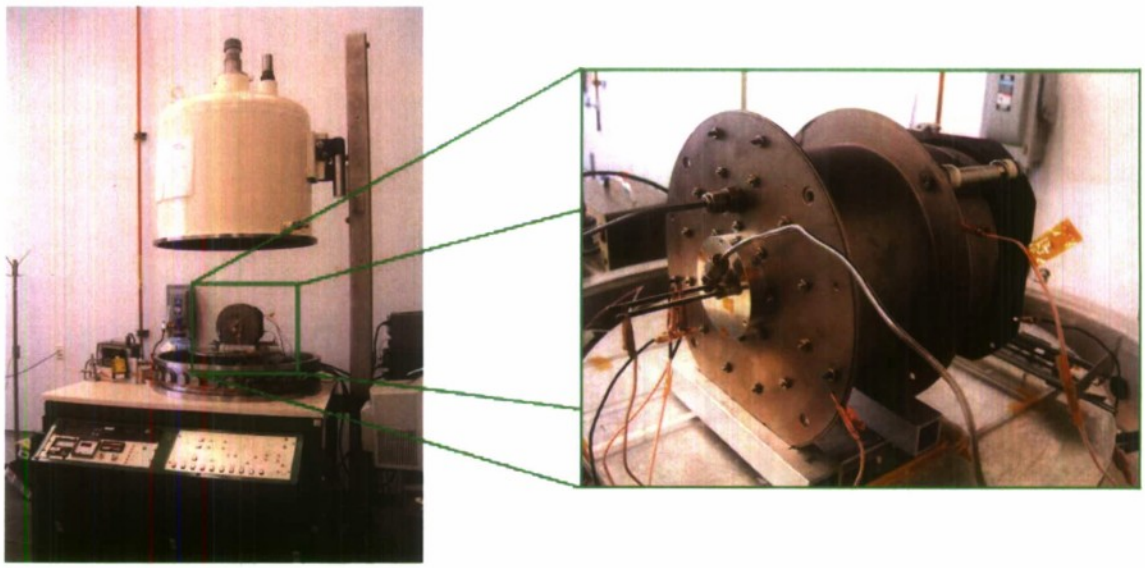
Appendix A:

Vacuum Chamber



Appendix B:

Ion thruster ready to test in vacuum chamber



Thruster Setup

Close-up View

CALIFORNIA POLYTECHNIC STATE UNIVERSITY, SAN LUIS OBISPO

Modeling, Analysis, and Design of Structural Components Subjected to Severe Loading Conditions

Deborah Go – Civil Engineering
Steven Wancewicz – Mechanical Engineering

Summary

This study focused on analyzing the earthquake simulation software OpenSees. We recorded design parameters and test results from former experiments that had evaluated the behavior of exterior, knee, and eccentric joints under severe loading conditions. We then entered the collected data into and ran OpenSees in order to compare the accuracy and consistency of the OpenSees model with the actual experimental results.

Introduction

The purpose of this project is to test analytical models for reinforced concrete beam column joints subjected to seismic conditions. Post-earthquake reconnaissance as well as experimental investigations reveal that joints affect the stiffness and strength of the structure and might also lead to structural collapse. Analytical modeling of joints with a wide range of design parameters allows a better understanding of force transfer within the joint and the failure mechanisms. Moreover, mathematical modeling allows for various joint designs to be analyzed under different loading conditions in much less time and at a much lower cost compared to experimental investigations.

Background

In a 2003 study by Lowes and Altoontash, a model was developed that simulated the behaviors of reinforced concrete interior column-beam joints under severe loading conditions. This model was later improved upon by Mitra and Lowes in 2007 to more accurately simulate the aforementioned behaviors of a wider range of interior joints. We will apply the proposed new model to exterior, knee, and eccentric joints.

To start, reinforced concrete is concrete reinforced with metal. There are two components to the reinforcing metal. First, steel bars provide additional strength and flexibility to the concrete. The number of steel bars in a beam or column affects the strength of the joint and therefore must be included. Second, hoops are spaced around the steel bars to keep them in place. The spacing and size of hoops are additional factors that must be included in the model.

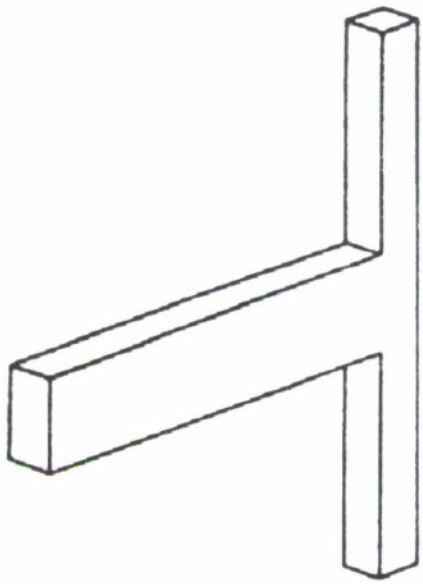


Figure 1. Simple representation of an exterior joint.

An exterior joint is universally found and therefore important to be able to model. As shown in Figure 1, the beam is perpendicular to the column. In addition, one end of the beam coincides with the center of the column. An exterior joint can also be referred to as a T-joint.

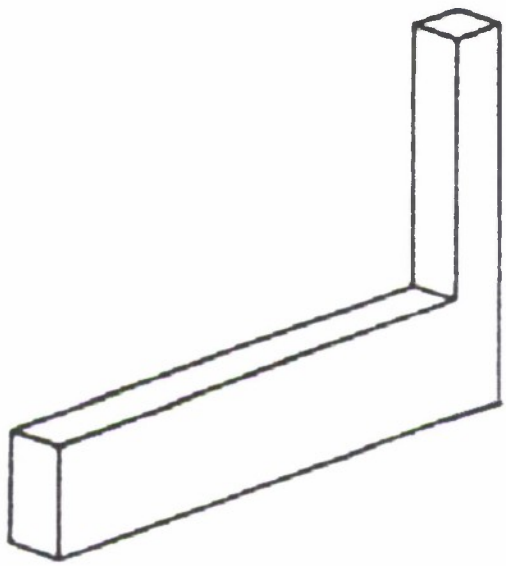


Figure 2. Knee joint

While not as many knee joints generally exist in a structure as exterior or interior, the initial failure of just one joint could cause a structure to collapse. Therefore, all joints including a knee joint are important. A knee joint is similar to an exterior joint, except one end of the beam connects to one end of the column instead of the middle of the column as in an exterior joint. A simplified beam-column knee joint is shown in Figure 2. While researching reinforced concrete knee joints, authors also referred to a knee joint as an L-joint or a corner joint.

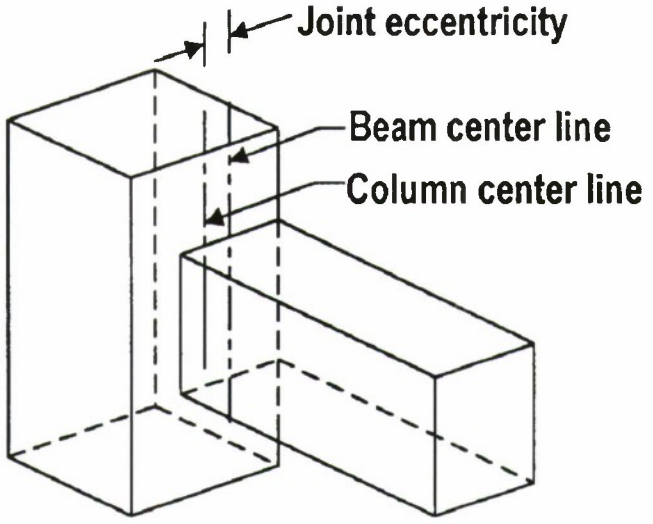


Figure 3. Eccentric Joint.

Another part of this project is to test the applicability of the model on eccentric column-beam joints, shown in Figure 3. An eccentric joint is a joint where a column and a beam of different depths are aligned so that one column face is flush with one beam face, and therefore, the centerline of the column is not aligned with the centerline of the beam. During conditions like those created by an earthquake, this eccentricity results in the development of torsion forces that do not develop in concentric, or non-eccentric, joints.

A model for simulating joint failure helps engineers perform quicker and less expensive analysis before a structure is built. If the model incorrectly calculates data, an entire structure could collapse. Therefore, it is important to test the model by comparing already collected experimental data to data simulated by the model.

Methods

We began the project by performing some background research on the types of joints with which we would be working. Dr. Mitra provided each of us with several papers pertaining to eccentric, exterior, and knee joints. The papers required specific attributes in addition to just being the respective joint. For example, a knee joint research paper needed to be about knee joints made of reinforced concrete that had rectangular cross sections. After reading through these papers and developing a firm understanding of our respective joints, we took note of key forces, properties, and dimensions that are relevant to the response of a reinforced concrete beam-column connection during severe loading.

Although Dr. Mitra had already provided us with a few research papers, we gathered more papers using search engines such as Engineering Village, by searching specifically for experiments that tested the response of reinforced concrete joints when subjected to earthquake-like forces. Once we collected a substantial number of experimental research papers, we used the aforementioned list of key parameters as a guideline to extract data from each paper, recording the data in an Excel spreadsheet. For each paper, we would note the properties and dimensions of each tested specimen and the behavior of the specimen during testing.

The next step in the project was to create an analytical model with source code. Dr. Mitra provided both students with the source code to use in OpenSees. The OpenSees source code is written in the tcl/tkl (pronounced "tickle") language, which is specific to OpenSees. While OpenSees has a wide application of uses, the purpose of this project is to use the program to analyze exterior, knee, and eccentric joints. The source code was opened in a program called Crimson Editor. The beams and columns are initially being analyzed separately to create a full analysis of the joint based upon all components and how all the components react to given forces. Once the information for a given beam/column has been recorded, the code will be executed in OpenSees. Operating OpenSees involves using a few basic command lines: one to run the file, and one to clear variables to run the next file.

The next step in the project was to use the collected data from OpenSees, which is a minimum of six separate text files per beam and column, and make use of the data to discover how stiffness is affected as the beam degrades. There are two materials that could fail in the joint: concrete or steel. A yield strength corresponding to each steel reinforcement bar was recorded in excel and will again be utilized here. The concrete was assumed to have yielded with a value of 0.003 mm/mm. A result from one beam can be shown in Figure 5.

Results

An example graph is shown below in Figure 5. The graph shows a moment-curvature relationship for the beam. The point on the plot where it seems to suddenly veer right on the graph demonstrates the yield point of the beam. Since the red square (representing steel) is around this point, steel is the reason the beam would fail. Therefore, if the beam were to be strengthened, the steel would have to have a higher value for its yield strength. The corresponding moment on the y-axis signifies the maximum moment the beam could be subjected to and the curvature is the angle in radians that the beam would be from horizontal after such a moment is applied.

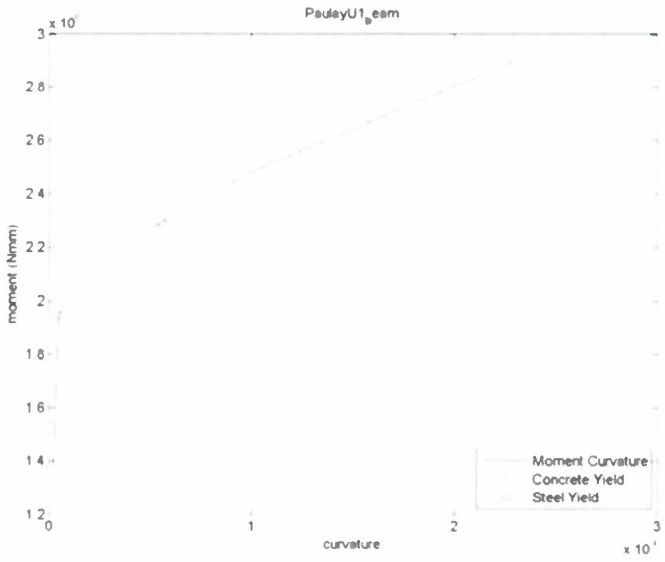


Figure 5. Example of Graphical Results.

Future Research

While these graphs are very informative, more work is still necessary to find a relationship between stiffness and the applied moment and ultimately determine when the joint will fail. So, while plotting these graphs utilizing matlab, another property was recorded, the lever arm for the beam section and column section that were in compression. With this property, properties from recorded in the spreadsheet, and additional source code provided by Dr. Mitra, the joint is finally ready to be analyzed.

Conclusion

In conclusion, we have learned a lot since starting the project. At first, it was a great challenge to extract numbers from articles. As time passed and Dr. Mitra provided more question sessions, both Deborah and Steven became better at extracting data from articles. The vocabulary and general knowledge of joints will help in further classes and our future careers as engineers.

The future of the project is currently uncertain, as Steven will no longer be able to work on the project and Deborah is pursuing other interests.

Acknowledgements

We would like to thank Dr. Mitra for his excellent explanations and for allowing us the opportunity to work on this research project.

Honors 200 Final Report
Probing the Surface of Lactic Acid Bacteria With Atomic Force Microscopy

Nicole Olson, Biochemistry

Executive Summary

The goal of this project was originally to determine how the growth media of lactic acid bacteria affects its adhesion forces. The part that we needed for the atomic force microscope (AFM) to be able to study the adhesion forces hasn't arrived yet, so we were forced to alter our goal. Instead, we have been studying four different strains of lactic acid bacteria, obtaining images of each, and trying to find observable differences between them. However, the end goal of the project is still to be able to analyze the adhesion forces of different strains.

I spent the majority of the quarter learning how to use the AFM. I was able to begin imaging the four strains of bacteria during the last two weeks, but I didn't have enough time to draw any conclusions about the observable differences between the strains. My accomplishment this quarter has been becoming competent using the AFM more than analyzing the bacterial strains themselves.

Introduction

We wanted to determine if lactic acid bacteria cultured in the presence of lipids and/or proteins from the milk fat globule membrane (MFGM) have stronger adhesion forces than those grown without MFGM material. Our hypothesis was that the presence of MFGM material in the culture media enhances the probiotic effects of the bacteria as well as makes the adhesion forces that the bacteria have to the intestinal epithelium stronger. Since we couldn't measure the adhesion forces, we decided to study four different strains of lactic acid bacteria instead. The four strains that we studied, NCFM, NCFM LaCl, NCFM SlpA KO, and

1063-S, are all used widely in the food industry as probiotic additives in dairy products. The main things that we started to look at while determining differences between strains was the size and shape of the bacteria. We were able to note these different characteristics by using the AFM.

Background

The AFM was first introduced in 1986. Since its' creation, it has become the foremost tool for measuring, manipulating, and imaging matter at the nanoscale. The AFM is a type of scanning probe microscope and consists of a microscale cantilever with a tip (probe). When the tip comes close to the surface, the cantilever bends. By aligning a laser so that it is focused off of the end of the cantilever and then reflected onto some photodiodes that are connected to a detector, the AFM is able to measure the height of various samples. To avoid damaging the tip, the AFM utilizes a feedback mechanism. The feedback mechanism works by adjusting the distance between the tip and maintaining a constant force between the tip and the sample. As the cantilever and tip move back and forth across the surface, a height image is able to be obtained because the movements of the laser are recorded by the photodiodes and then relayed to the detector, where an image can be put together. When scanning, piezo crystals are responsible for moving the cantilever very small distances in the x, y, and z directions.

This project has been a collaborative effort between The Chemistry and Biochemistry and The Dairy Products Technology Departments. While previous research has already been done on the three NCFM strains, our project attempted to examine the bacteria on a molecular level. The AFM has been a great aid by allowing us to obtain high-resolution images of samples on the nanoscale. The NCFM strain is a wild strain of lactic acid bacteria. The NCFM LaCl strain is a mutant of NCFM, which has been genetically modified to produce β-galactosidase. The NCFM SlpA KO strain is another mutant of NCFM, which has

been genetically modified so that its' S-layer surface protein isn't produced. The 1063-S is known as lactobacillus reuteri, and is found in the gut of mammals and birds.

Methods and Theory

In order to be able to image the bacteria, we first had to prepare slides with the samples on them. We did this using a poly-L-lysine solution and glass slides. Once we had the slides prepared, we were able to use the AFM to look for bacteria. When using the AFM, the laser first has to be aligned on the end of the cantilever. The diodes then have to be adjusted so that the laser hits them. After that, the stage that holds the cantilever is lowered down close to the sample. From that point, the tip can be gradually approached. A z piezo crystal is responsible for moving the tip closer to the sample in small enough increments that it will be able to engage with the surface without breaking off. Once the tip is engaged, imaging can be started. In general, we scanned a relatively large size first (about 40 to 60 microns). From the large scan, we were able to find bacteria. We then would scan a smaller size that included only the bacteria piece that we were interested in. When scanning, we had to find the right adjustments for the rate of the scan, as well as the height of the tip. We also had to adjust various other settings to guarantee that we were getting the best quality images possible. Once we were able to scan an area that definitely contained bacteria, then we would perform a high-resolution scan that would take up to 20 minutes. Although we didn't get to much analysis of the images, a computer program can be used to measure the bacteria and manipulate the images. We began measuring some of the images from the samples.

Results

In addition to learning how to use the AFM, I was also able to obtain images of each of the four strains of bacteria, which I will include. More images

will be obtained next quarter, and more analysis of the images will be done next quarter as well.

1063-S →

This sample is about five microns long, one micron wide, and 500 nanometers (nm) tall.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

← NCFM

This sample is about three and a half microns long, one micron wide, and 180 nm tall.

NCFM LaCl →

This sample is about four microns long, two microns wide, and 250 nm tall.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

← NCFM SlpA KO

This sample is about five microns long, one micron wide, and 250 nm tall.

Conclusion

Overall, I think that this quarter was a success. I'm glad that I was able to become proficient on the AFM and at least begin to take images. Next quarter we will be able to make more generalizations about the strains of bacteria once we have more images. We also will be able to measure the adhesion forces once we get the necessary part for the AFM. Once we can do that, we will be able to work on our original purpose. I'm very excited to continue working next quarter.

Bibliography

Schär-Zammaretti, Prisca, and Job Ubbink. "Imaging of Lactic Acid Bacteria with AFM – Elasticity and Adhesion Maps and Their Relationship to Biological and Structural Data." *Ultramicroscopy*, 97 (2003): 199-208.

Schär-Zammaretti, Prisca, and Job Ubbink. "Probing Bacterial Interactions: Integrated Approaches Combining Atomic Force Microscopy, Electron Microscopy and Biophysical Techniques." *Micron*, 36 (2005): 293-320.

Russell, Michael W., Tri Duong, Rodolphe Barrangou. Todd R. Klaenhammer. "Characterization of the *tre* Locus and Analysis of Trehalose Cryoprotection in *Lactobacillus acidophilus* NCFM." *Applied and Environmental Microbiology*, 72.3 (Feb. 2006): 1218-1225

RESEARCH: VISUALIZATION OF DECISION-THEORETIC PLANS

FINAL REPORT – SPRING 2008

BRIAN OPPENHEIM, EVAN HECHT
BOPPENHE@CALPOLY.EDU, EHECHT@CALPOLY.EDU
CALIFORNIA POLYTECHNIC STATE UNIVERSITY, SAN LUIS OBISPO
HNRS-200 - SPRING QUARTER 2008

ABSTRACT

Most current probabilistic decision making software programs have failed to communicate scenarios effectively with users who are not familiar with statistical theory. This makes it very difficult for users to benefit from the information presented by such software. In this project, we have continued earlier research to determine the best way to present information about a stochastic plan to users with varying backgrounds. We have been primarily focusing on the domain of academic advising with an overarching goal that the specific results we find can be applied to other domains.

Over the course of the last two quarters, the team has progressed through several successive phases of work on this project. We began our project by identifying the weaknesses of a previous system, PlanIt. From that list, from group brainstorming, and from some background research that we performed, the team developed a list of goals for the types of information and the ways in which the data would be accessed. These goals were then turned into prototypes for new interfaces.

This quarter, we began implementation of the prototypes and the needed back-end software. The back-end system has two layers: the first will parse (or read) the actual plan data, and the second will convert it into a form usable by our API. The second layer, the API itself, will provide different presentation software with unified procedures for accessing and manipulating a plan, so that every presentation is based off of the same critical cognitive elements. The API has been completed, and the parser is in development.

TABLE OF CONTENTS

Abstract 1
Research Team 3
Introduction 3
Theory and Background Information 4
Plans 4
Cognitive Models 4
Complexity 5
Previous Work 5
Methods/Experimentation 6
Results 6
Problems with the Current System 6
Goals for the New System 7
New System Interfaces 8
Traditional Design 8
Alternating Abctions and Results "Exploding" Design 9
Heavily Visual "Time Machine" Design 10
The Application Programming Interface (API) 12
Design/Goals 12
Implementation 13
The Parser 13
Conclusions 14
Future Work 14
Acknowledgements 14
Bibliography 14
Appendix A – The Wiki 15

RESEARCH TEAM

Kyle Cushing
BS Student, Computer Science
California Polytechnic State University,
San Luis Obispo

Dr. Alexander Dekhtyar
Associate Professor, Computer Science
California Polytechnic State University,
San Luis Obispo

Tom Dodson
Student
University of Kentucky

Dr. Judy Goldsmith
Professor, Computer Science
University of Kentucky

Evan Hecht
BS Student, Computer Science
Honors Research Team
California Polytechnic State University,
San Luis Obispo

Joan Mazur
Associate Professor, College of Education
University of Kentucky

Brian Oppenheim
BS/MS Student, Computer Science
Honors Research Team
California Polytechnic State University,
San Luis Obispo

INTRODUCTION

Human beings are constantly looking for advice in situations where decisions are complex and include a vast array of possible choices and outcomes. The concepts of multi-step planning with probabilistic outcomes at each stage are fairly well known and researched (at least among academics and statisticians). The planning process and computations have been implemented on computers in a many ways, with reasonable success. However, most probabilistic decision making software programs have failed to communicate scenarios effectively with users who are not familiar with statistical theory. This communication gap makes it difficult, if not impossible, for users to gain any benefit from the information presented by these software packages.

In this project, the primary domain of research is applying plan visualization in advising settings. These include academic advising and advising of "welfare to work " program clients. Both problem sets would greatly benefit from having a software program to compliment advisors with large case loads. While these subjects are the main problems we are focusing on solving through this research, there is a broader, secondary goal to learn about how humans visualize and understand decision making scenarios through the use of technology. Eventually, we hope to publish our generalized findings in a journal on artificial intelligence and possibly present the results at a related conference.

THEORY AND BACKGROUND INFORMATION

This section introduces the theoretical concepts that form the basis for our work.

PLANS

A plan, most simply, is a mapping showing what actions someone should take when in a particular situation. This mapping, however, is not typically the information given directly to the program by the user. More typically, a structure such as a Markov Decision-making Process(MDP) that includes notions of multiple actions, result probabilities, and state utility will be given. It is then fed through an optimization algorithm that computes the plan from the given MDP. MDP's are formally introduced in the next section.

Here is a textual example of a simple plan for a student picking Mathematics courses:

You got a C in Calculus I and II, and you want to take classes that best fit your abilities. First, take Calculus III. If you get an A or a B, take Statistics. If you pass Statistics take Statistical Theory, otherwise retake it. If you got a C in Calculus III, take Physics. If you get an A in Physics take Quantum Theory, if you get a B take Dynamics, or if you got a C or an F take Chemistry. If you failed Calculus III take English Composition. If you pass Composition take Creative Writing, otherwise take Sculpture.

COGNITIVE MODELS

The theory of cognitive models is critical to the success of this project. Essentially, that theory states that any thought process (such as exploring our interface) can be modeled in a way common to all people (the **cognitive model**), such that finding and using the correct model will make the thought process easy and intuitive. The old interface of the PlanIt software did not correspond to the thought processes of the University of Kentucky students who participated in the experiment nor ours. This disconnect makes the software difficult to use. Thus, the goal for this project is to create and implement an interface that more closely matches the cognitive model that people have for visualizing and interacting with information about possible actions and results.

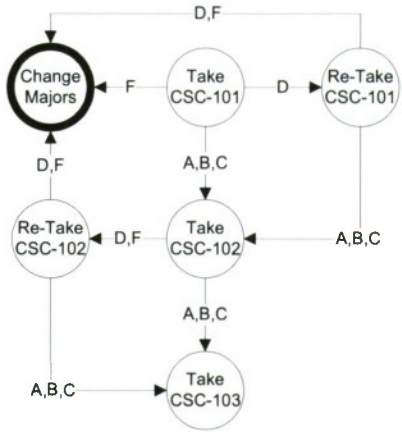
Formal Definition of an MDP
An MDP is a 4-tuple (S, A, P, R) where:

- S is a set of states
- A is a set of actions
- $P: A \times S \times S \rightarrow [0,1]$ is a function where $P(a, s, s')$ is the probability that if action a is taken whilst in state s, the result will be s'
- $R: S \times S \rightarrow M$ where $R(s, s')$ is the utility of transitioning from s to s' with M an ordered set such that transitioning from s to s' is said to have more utility than transitioning from t to t' if and only if $R(s, s') > R(t, t')$.

The basic model used in this project to represent human decision making is the **Markov Decision-Making Process** (commonly denoted MDP). The building blocks for a MDP are **states** which are each a representation of a combination of values for plan variables. In academic advising,

these variables could include classes taken and grades received in each course. Each state has a numerical quantity associated with it called the **utility**. This value represents the relative benefit of being in this state in terms of its contribution to the desired final outcome. From each state, there are different **actions** that can be taken, each with **results** of varying probability. These results are actually each just the state that is reached by taking the given action and achieving the given result.

COMPLEXITY



The diagram shown at the right is a graphical representation of a simple stochastic plan in the academic advising domain. It is important to note that this diagram only shows the suggested action to take when in each state and does not show alternative actions and the probabilities of the results of the actions. These probabilities and additional actions are all information that needs to be displayed to a user of this software but are not part of the formal definition of a plan. The example can also be a little misleading since it seems to suggest that navigating a plan is simply a matter of progressing or not progressing. A more realistic, but more complex, example may suggest different classes to take based on previous grades and coursework rather than just advise to or not to move on. A more realistic plan would also likely dictate several classes to take and have the result transitions for each possible combination of course grades.

Even in this very simplified example of a plan, which has few actions and few states, it is difficult to discern useful information. Now consider the amount of states and actions needed to represent all of the different possible combinations of coursework and grades that a student could take in their college career. In addition, consider how many possibilities there are for student goals and preferences. You can easily see how these scenarios quickly grow into exponentially large problem sets and are very difficult to display without some novel thinking.

PREVIOUS WORK

PlanIt

This project is a continuation of efforts made by a group of students and Professors from the University of Kentucky. Their efforts led to the development of a software suite called PlanIt. This software's function is to interact with the user to show them how various actions will likely affect the outcome of some process. It has been written so that the decision making algorithm can be switched in and out. This allows independent development of the prediction logic and the interface presented to the user. The current user interface was tested using freshman computer science students at the University of Kentucky. The overwhelming response was that it did not provide an intuitive experience. Thus, our research involves developing a new interface for interacting with users.

METHODS/EXPERIMENTATION

Since there is no known cognitive model for our problem domain, we can only analyze the model indirectly by examining the response of ourselves and others to interfaces built on a specific model. The first stage of the project was to use the old interface, and find all the parts we didn't like, or that were difficult to use. In theory, these parts did not sufficiently match the cognitive model for stochastic planning, and so should be changed. The next stage was to, again using the old interface, make a list of features we'd like to see in a new interface. Hopefully, these features are essential, intuitive parts of the cognitive model. Next, we used our research materials to reinforce our knowledge of cognition. We distilled our problems and features into essential facts about the underlying cognitive model, and brainstormed ways to make a new interface that closely matched the model. To conclude the previous quarter's work, we designed prototype interfaces that use our discovered model.

To build our API for this quarter, we first considered the data in a plan and ordered it according to our cognitive model. Using Java's object-oriented approach, we wrote software that would store data according to our desired organization. Next, we collaboratively brainstormed different ways that interfaces would need to work with the data in a plan. Following good design principles, we extended our software to include these methods, or operations, as well as the data. We are currently working on two fronts: building a sample plan to test our API and begin using it with interfaces as well as building a parser to read data from the planner and distill it for use in the API.

RESULTS

In this section we present the results from all of the stages of our research performed thus far.

PROBLEMS WITH THE CURRENT SYSTEM

We have summarized our research regarding the problems with the current system in this list:

- Currently does not distinguish between recoverable and unrecoverable bad states in color.
- The terminology/vocabulary is confusing.
- The system does not do anything at a dead end state. For instance, in the robot example, once the user has coffee there is 1 state with 100% probability that will keep coming up no matter how many times you press forward.
- Information about current/future states



is hard to visualize.
- Changes between states are hard to evaluate.
- The current software does not allow comparison of multiple actions and/or results. See one result state at a time.
- Need more granularity for display of utility and probability.
- The choice of action seems to be subordinate to the result rather than the other way around. Also, action selection is hard to use and understand.
- It is hard (i.e. impossible without many clicks) to see the long term result of choosing a specific action.
- No way to go back large distances. (i.e. beginning, specific points).
- The system is currently missing why a state changed the way it did.
- There is no indication of time progression.

GOALS FOR THE NEW SYSTEM

The following are the goals we developed for the new system:

- **The Big Picture**
  - actions displayed more prominently than results, should show sub ordinance
  - Glossary - 'Tooltips' to explain vague/difficult terminology
  - Customizing and saving the following:
    - Ability to customize how some components work
    - Ability to move information and components around into an organization that best suits the user.
    - Domain specific default views and view components
  - Progress indication
  - Somehow quantify number of possible paths to indicate flexibility in plan.
- **States**
  - include information about time (not in model classes, but in user view)
  - save/load states (ie checkpoints)
  - indicate most desirable action
- **Actions**
  - indicate most probable result
- **Results**
  - the reason for the result: what actually happened to cause the state to change the way it did
  - the utility of the result state
  - the distance from that state to the goal state(s)
  - state changes: full state is accessible but not shown by default
  - indicate most probable final outcome
- **Preferences**
  - determine weighting, or utility, of state variables

  - choose between multiple goal states
- **Goals**
  - determine 'ending' states
  - must be met
- **Evaluation of Multiple Sequential Actions/Results**

NEW SYSTEM INTERFACES

The research team developed the following three prototypes:

TRADITIONAL DESIGN

This specification of a new UI for PlanIt has been written with many specifics for the academic advising domain. Much of this can be generalized to other planning domains. In contrast to the other two designs shown, this is the only one to currently be in the implementation phase.

PAST

The past section shows aggregate data regarding previous states. In academic advising, this would consist of milestones completed, gpa, quarters completed, etc. This section will have a button to show more detailed information about the previous states such as specific classes taken and specific grades.

PRESENT

The present section is the most interactive of the sections and is where the choice of actions and outcomes takes place. For academic advising, there is a box of labeled check boxes that represents each class that the student could take in that term. Below that list is a tabbed panel that shows a tab for each course. On that tab, the user can see the most likely outcome for that class and choose which outcome they would like the plan to use. On any change in checkboxes or choice of outcome, the UI immediately updates to show the effect of that change.

FUTURE

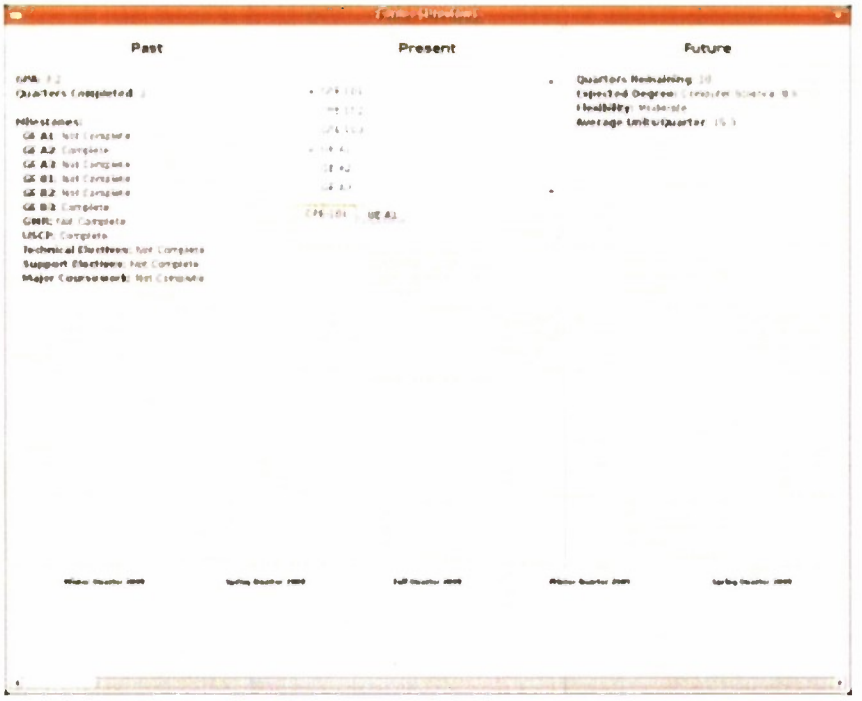
The future section shows aggregate information about the likely future based on the choice of actions and outcomes in the present section. As with the past section, there will be a button to show detailed information about the projected future.

FILMSTRIP

The filmstrip helps to give a visual reference to the concept of time. In the academic advising domain, this would show a summary of the term. This would also allow the user to quickly change the choices at a given term and see the cascading effect of this change.

SCREENSHOT

Here is a recent screenshot of this UI concept:



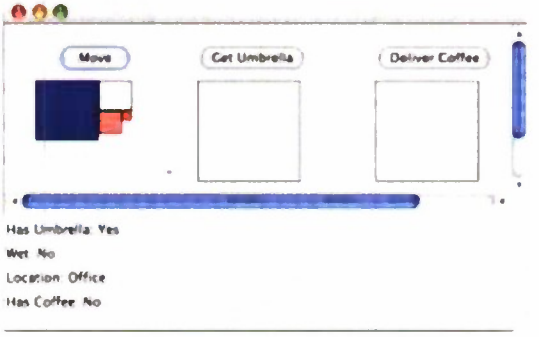
ALTERNATING ABCTIONS AND RESULTS "EXPLODING" DESIGN

This design differentiates between actions and results by having the user select an action, bringing up a screen where they select a result, which then becomes the current state and allows new action choices. This design remains in the prototype stage.

CHOOSING AN ACTION

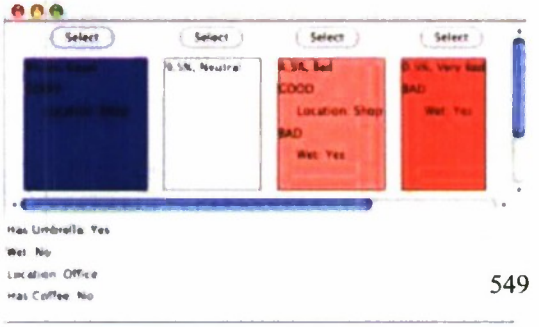
This view allows the user to select between available actions.

State is shown at the bottom, and each action gives a condensed indication of the probability and utility for its corresponding results. Choosing an action brings up the second view, Choosing a Result.



CHOOSING A RESULT

This view allows the user to choose between results for a specific action.



Current state, as with the action selection view, is shown at the bottom. By default, results show only their short description, consisting of probability, utility, and state changes, but complete state information can be accessed. Choosing a result causes its state changes to take effect (its state replaces the current state), and the user is shown a new Choosing an Action screen.

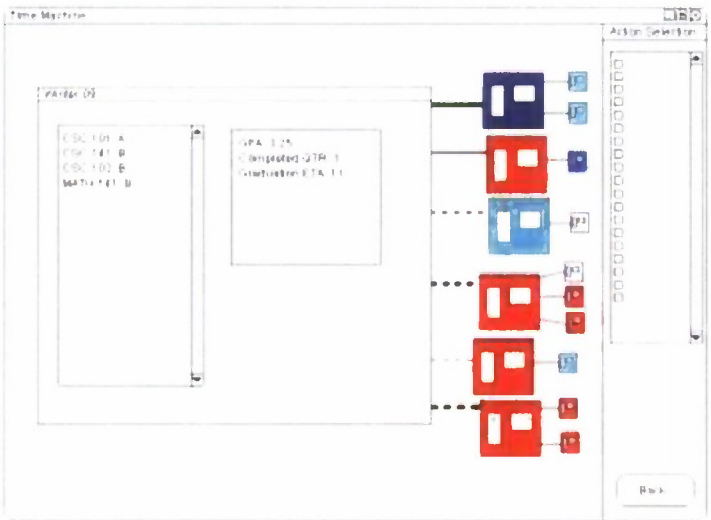
HEAVILY VISUAL "TIME MACHINE" DESIGN

This prototype design and write-up is by Kyle Cushing a member of the research team working on his senior project. All of the prototypes were discussed together and while each was written by an individual group member, are a result of group discussion.

The purpose of the Time Machine model is to be able to view a large portion of the policy tree compared to other version. The base design is from the website widget on Etsy to view past orders/items for sale. This design is also still in the prototype stage.

STANDARD VIEW

This is the standard view with no actions currently selected:



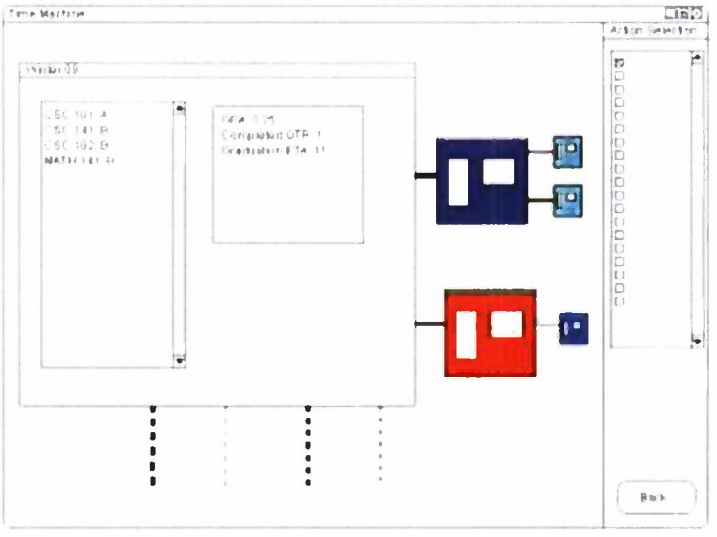
GRAPHICAL REPRESENTATION OF INFORMATION

The values of Quality and probability are represented graphically in this model by: Line Thickness and Window Coloring. Probability - Represented by the Thickness of the lines that connect the various states. Quality - Represented by a color scale from red to blue (Red - worst quality, Blue - best quality) of the future states.

The lines are either solid or dashed to show which result states are possible by choosing the Suggested Action. Solid lines connect the current state to result state(s) by means of the Suggested Action. The dashed lines connect the current stated to result state(s) by means of Alternate Actions.

ACTIONS SELECTED

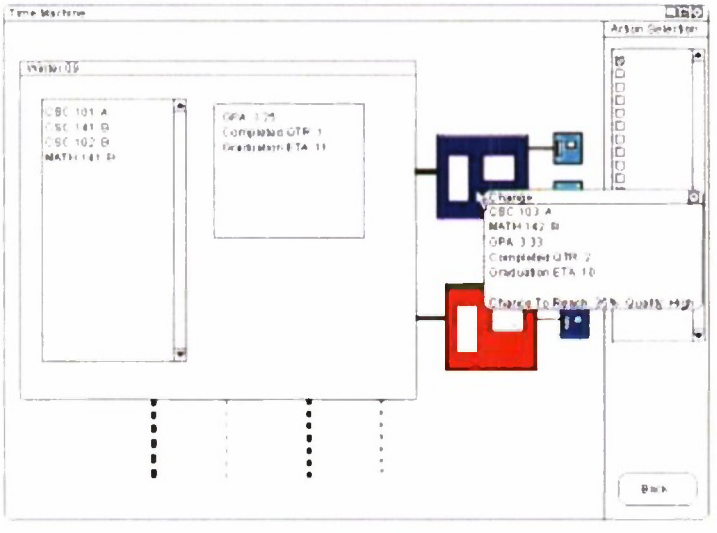
This is the view once an action has been selected:



After an action is selected the result states not associated with that particular action are hidden below view, but the lines to them are still visible. In the above example the suggested action's was chosen and thus the result states associated with the Suggested Action remain in the main view.

MOUSE OVER RESULT STATES

This is the view when the user moves their mouse over any one of the possible future result states. The change between the current state and the result state that they have placed their mouse over are displayed in a tool-tip like window near the mouse cursor.



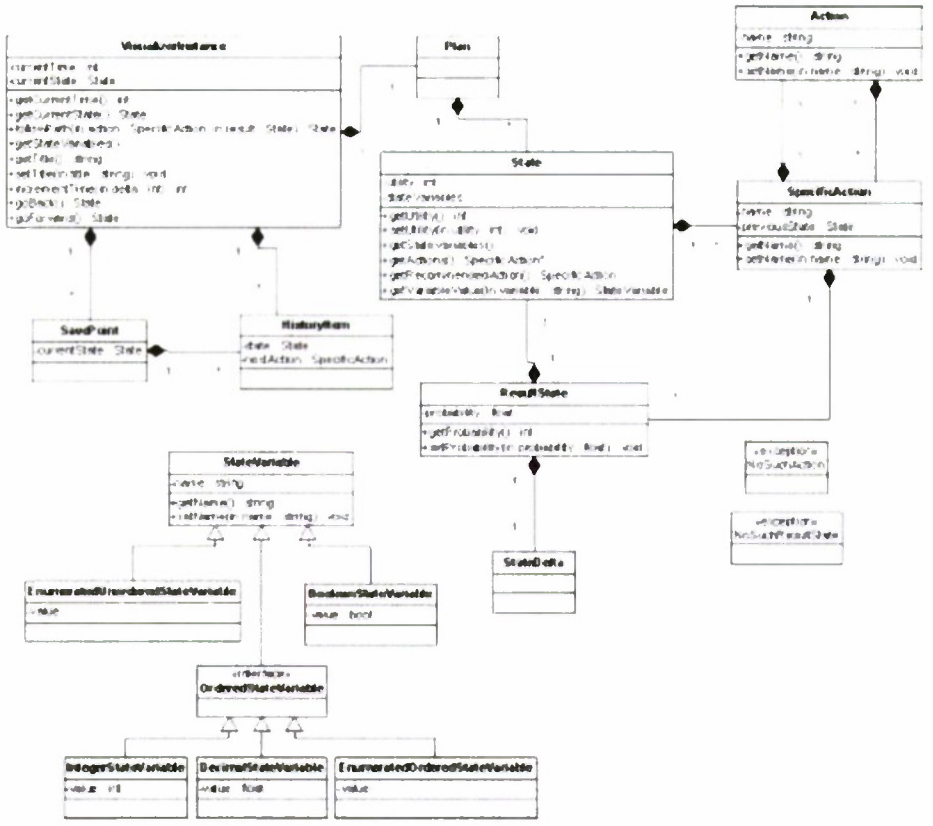
THE APPLICATION PROGRAMMING INTERFACE (API)

In order to allow different interfaces to be quickly developed over a common set of data, we carefully wrote an Application Programming Interface for plan data. An *Application Programming Interface* (usually called API for short), is a set of functions that allow an outside program access to get and manipulate data in some data structure. As mentioned before, in our program, the API is used to abstract away the details of navigating through a plan and keeping track of its data into a common interface that can be put underneath any of the user interfaces that we develop. This will also allow us to optimize our code more quickly since the interface is shared rather than written specifically for each user interface.

DESIGN/GOALS

When going through the process of defining our interface, we considered several goals for our design. The first of the more important goals is ensuring that the design provides simple, unobtrusive access to the data by the APIs. To achieve this goal, we considered the types of data that most interfaces will require, the order in which they will need the data, and the likely patterns in which they use the data. The second of these goals is to have the data be represented as naturally as possible. A natural representation gives the best chance that the API's design will match up closely with the ways in which interfaces are using the information. Finally, we prioritized making the API as generic as possible. That is, we tried to abstract away all details that are specific to a certain domain. This will allow the API to be used for any number of unique domains.

Below you can see a *Uniform Modeling Language* (often abbreviated UML) diagram showing our design for the API. This diagram represents the knowledge that a designer of a user interface would need in order to build an interface over the API. It does not include the implementation-specific details that the API designer must consider in order to make the API operate correctly and efficiently. This style of abstraction and hiding implementation details by simply providing a public interface to a data set is called *black box*.



IMPLEMENTATION

We wrote the API in Java, following proper software design principles. For the data portion we mostly wrote our own classes, while the methods made use of some built-in Java classes as well. These classes have been widely used and tested, so we can be reasonably certain that any problems would be in our code. As much as possible, we implemented our code to be modifiable, so that we could change the particular way we accomplish something without changing what actually gets accomplished. We also ensured that the code was general enough to accommodate diverse problem domains, and that we could easily change the code to fit even more diverse situations.

THE PARSER

As was demonstrated in the "Theory and Background Information" section of this paper, a plan is an extremely complex network of states and actions. Since in our design there are separate modules to compute and display plans. Since the planner we are using is not specific to our program, its output is not in a form readily usable by our user interfaces through the API. To bridge this gap, we must implement a parser. This component will carefully read through the output of the planner and convert the plan information into the data structures accessible via the API. For this quarter's work, the team got familiar with the output of the planner module in order to determine how to best write the parser.

CONCLUSIONS

Following last quarter's analysis of the old interface and planning cognition, we have implemented an API containing elements of the cognitive model that should be common to all effective plan interfaces. Efforts are underway both to develop interfaces to use this API, and to translate plan information into an API-accessible form. Next quarter should see these efforts to completion, and scientific experimentation to evaluate our work will commence the following quarter.

FUTURE WORK

Having completed the majority of the higher-level planning and design phases of the project, the team will proceed with implementation, testing, and experimentation. First, the group will complete the implementation of the plan parser and three interface designs. These will all then be comprehensively tested for correctness. Once the team is confident that the implementations are thorough and correct, we will devise our experiments and get the necessary permissions from the institutional review board. Using volunteer students from the department as test subjects, we will run our experiments and collect the results. Since the data will be mostly subjective in nature, our next task will be to sort through the results and understand the meaning. We hope to use this to improve our interfaces and possibly develop new ones that emphasize the features requested by the test users.

The team will mostly halt work during the Summer Term due to internships, time off, and summer vacation plans. The research will continue again as an honors research project during Fall Quarter.

ACKNOWLEDGEMENTS

The honors research team would like to thank Cal Poly Professor Alexander Dekhtyar and Cal Poly student Kyle Cushing for their work with us on the project. We would also like to acknowledge and thank the team at the University of Kentucky for their feedback on the wiki.

BIBLIOGRAPHY

The following works were used as references during the project:

- Thinking, Problem Solving, Cognition - Richard E. Mayer, 1992
- The Nature of Cognition - Robert J. Sternberg, 1999

APPENDIX A – THE WIKI

Please see our project wiki at ***http://wiki.csc.calpoly.edu/planit***. On that site we have a complete view of the research performed thus far as well as updated information about the current status of the project.

IMPLEMENTATION OF ARTIFICIAL NEURAL NETWORK FOR CONTROL OF SIMULATED WHEELED ROBOT

Alan Tepe
Dr. Xiao-Hua Yu

INTRODUCTION

Artificial neural networks (ANNs) are functions whose structure is based on biological neural networks found in living organisms. They consist of a number of simple functions (nodes), whose outputs and inputs are linked by weighted connections. By adjusting the connection weights between nodes, it is possible to generate a number of interesting behaviors in a neural network, including the modeling and control of systems (Meltser, 1996).

The most exciting property artificial neural networks is their ability to "learn". Given a set of input data and desired output, it is possible to optimize the weighted connections within a neural network so that the network produces the desired output with input. Artificial neural networks can serve as universal appropriators. It has been shown that a neural network of sufficient size can approximate any non linear function (Narendra, 1990).

Artificial neural networks have been successfully used in a number of robotics applications, including path planning (Glasius, 1993), controls (Narendra, 1990), and decision making (Bekey, 2005).

This paper presents the results of a simple neural network experiment. A neural network was created and trained to control a simulated driving robot. Utilizing the neural network, the robot was trained to drive along a straight line path and return to the path if it strayed.

ROBOT SIMULATOR

A simple computer model of a two wheeled robot was created using basic concepts from dynamics. This computer model was programmed in MATLAB. The computer model took robot motor torques as an input and calculated the path of the robot over time as an output. This computer model used several assumptions to simplify computations. These are detailed below. More details on the robot model can be found in the appendices of this paper.

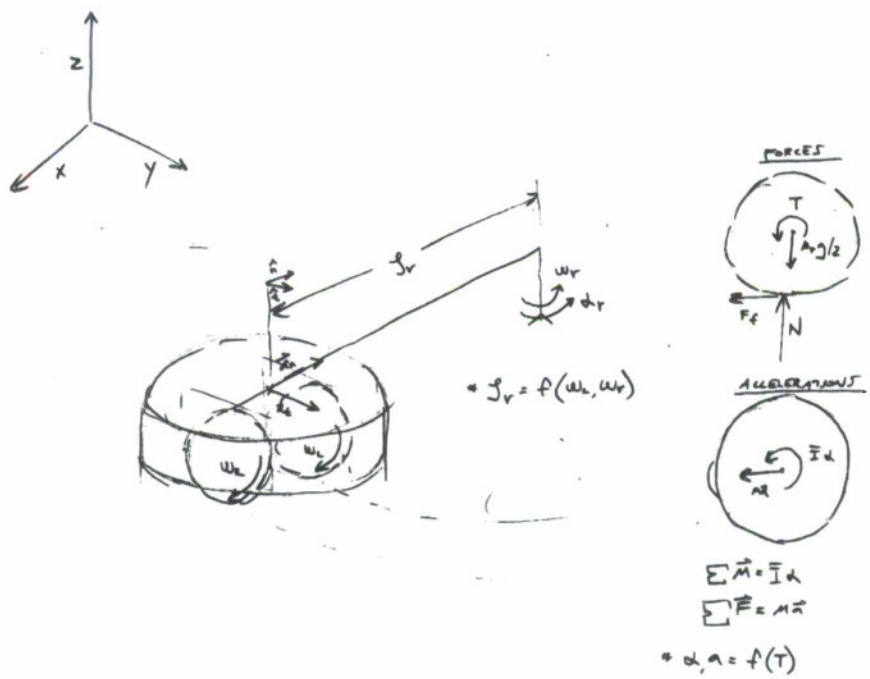


Figure 1: Illustration showing dynamics calculations behind modeled robot. The model used simple dynamics to calculate robot position over time with motor torques as an input

Robot Model Assumptions

- No slip condition between ground and robot wheels
  - It was assumed that the force of friction between the ground and the wheels of the robot was sufficient to prevent any slipping of the tires, this assumption seemed warranted because the weight of the robot was large when compared to the range of torques applied to the wheels.

- Even distribution of weight between robot wheels
  - It was assumed that each of the two wheels on the robot carried an equal portion of the robot weight, despite the motion the robot might be undergoing. This assumption seemed valid because the turning velocities of the robot were relatively low compared to those that would cause a large difference in weight distribution.

- Robot able to accurately determine some information about its position and orientation
  - It was assumed that the robot could determine how far away it was from its desired path, and that the robot could determine which direction it was facing relative to the direction of its desired path. This assumptions seemed valid, as measurement equipment could be designed for a physical robot to make these measurements.

Robot Environment

a simple environment was modeled to contain the robot model detailed above. The environment consisted of a uniform flat plate with surrounding walls, the simulated walls would stop operation of the robot model if the robot hit them. The simulated environment also contained a straight line drawn from one corner to the other, to serve as a desired path for the robot to follow. An above view of this environment can be seen in figure 2.

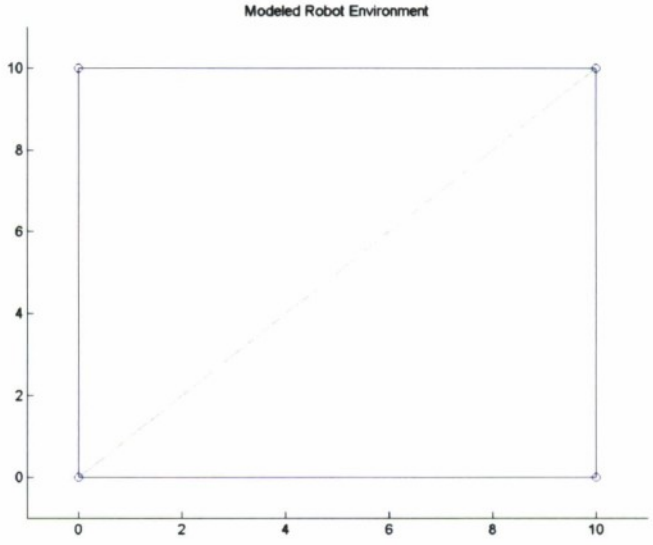


Figure 2: Modeled environment for robot. The blue lines are walls that halt the operation of the robot if hit, and the green line is a pathline that the robot is meant to follow. Dimensions shown are in meters

Desired Robot Behavior

For this experiment, it was desired that the robot follow a specific path in the above environment. To train the neural network to follow this path, it was necessary to define a desired output for the neural network for every possible input. This was because we used the back propagation training algorithm to optimize neural network weights, which requires a set of training data (consisting of inputs and desired outputs) (Reed & Marks, 1999). To create this training data, we made a simple function that calculated desired left and right motor torques for a given position and orientation of the robot. This function kept the robot within a small distance of the desired path for the duration of its motion, and created the robot behavior we desired to replicate with neural network control. The behavior of the robot with motor torques created by the training data function can be seen in figure 3.

%add picture with desired neural network inputs and outputs

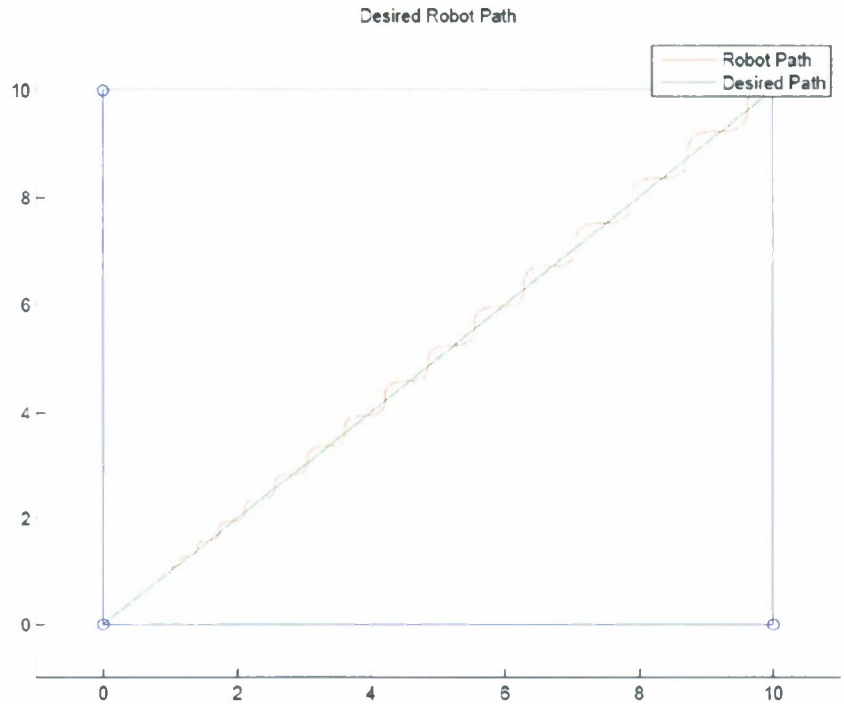


Figure 3: Behavior of robot when training data function was used to control robot motor torques. This is the desired behavior of the robot: the robot corrects itself when it gets off the desired path.

NEURAL NETWORK

A relatively simple artificial neural network was designed to control motor torques of the robot. It consisted of two input nodes, three hidden nodes, and two output nodes. The input nodes took in the distance between the desired path and the robot position, and the difference in angle (radians) between the robot orientation and the path direction. The artificial neural network was meant to process these inputs and output the left and right motor torques required to get the robot back on track. This simple structure for the artificial neural network was chosen because it was hypothesized that the relationship between neural network inputs and outputs was relatively simple, and might even be close to a linear relationship.

Nodes in the artificial neural network used a hyperbolic tangent (tanh) function to process individual inputs and outputs. All nodes except for the two output nodes used the tanh function for processing. This is illustrated in figure 4.

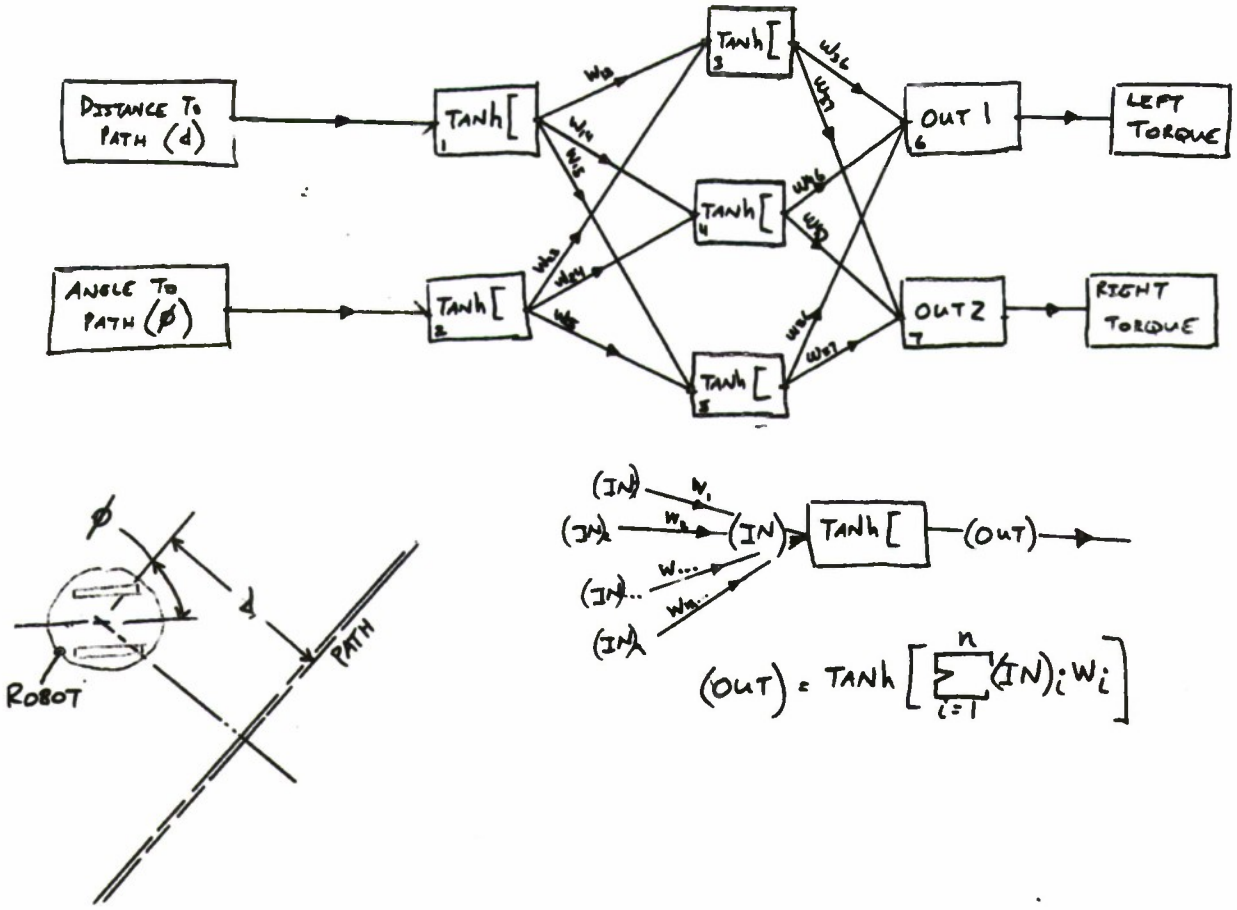


Figure 4: Illustration showing artificial neural network used to control the simulated robot. This figure also illustrates and explains the two inputs to the network: distance to path and angle to path, as well as the tanh function used in some of the network nodes.

Neural Network Connection Weight Optimization

An artificial neural network can bc used to approximate any function (Narendra, 1990). However, the connection weights between nodes in an artificial neural network must be adjusted to produce desired behaviors. This adjustment amounts to an optimization problem, in which the set of weights which produces the smallest error is determined. In this experiment, we used a back propagation algorithm to find the best connection weights. That is, the connection weight which produced the smallest difference between neural network motor torquc output and desired motor torque output.

To find these connection weights, a techniquc know as back propagation was utilized. This algorithm is similar to a gradient descent algorithm, in that it calculates derivatives of each connection weight with respect to error, and uses these derivatives to connection weights. This is the optimization technique used most often for training neural networks (Reed & Marks, 1999) This algorithm is explained in more detail in the appendices.

In this experiment, the connection weights were randomly initiated, with a normal distribution having a mean of 0.0, and a standard distribution of 1.0. The neural network connection weights were adjusted with back propagation with a sct of training data. The training data consisted of the inputs: **distance to path** and **angle to path** ranged, respectively from -1.0 to 1.0 meters and 0 to 2Π radians. As well as the desired outputs corresponding to these inputs.

There were some difficulties finding optimum connection weights for the neural network. It was difficult to get the back propagation algorithm to converge to the same set of optimal connection weights in any two runs. In addition, the optimum connection weights produced by the algorithm weren't able to follow the line satisfactorily. This could indicate that the size and geometry of the neural network was not correct for the line following application. More neural networks with different sizes and geometries should be tested to determine if this is the case. Figure 5 shows the paths of robots controlled by neural networks with different connection weights, as well as the neural network input-output relationship. Each of the connection weight sets was optimized by the back prorogation algorithm. The figure clearly shows the large variance in the behavior of the neural networks

optimized by the algorithm.

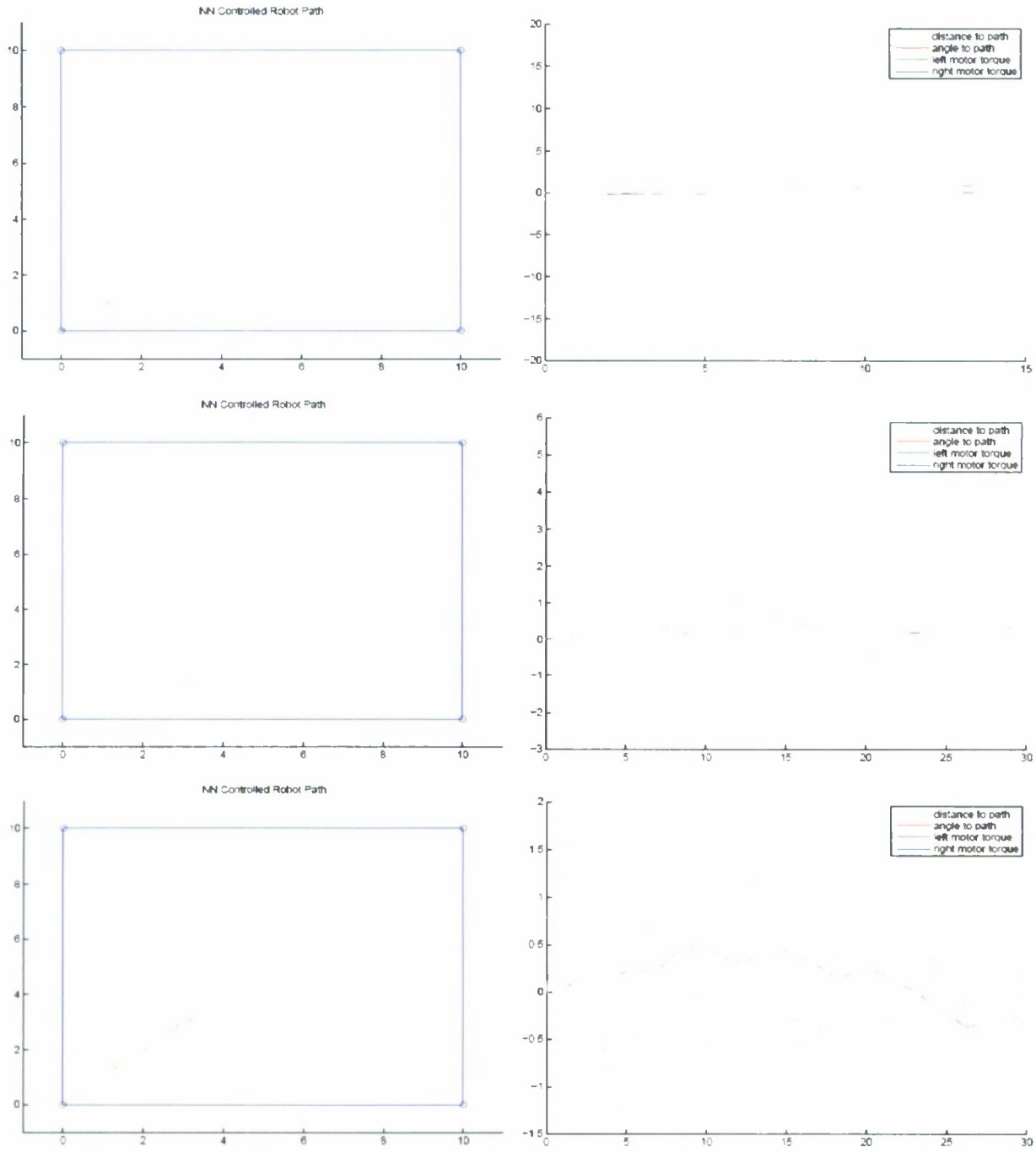


Figure 5: Neural network behavior with various optimized connection weights. The plots on the left show the path of the robot, and the plots on the right show the input-output behavior of the artificial neural network for the duration of the robot motion. The ability of the robot to follow the line varies with different connection weights. We were unable to get the optimization algorithm to consistently converge to a set of connection weights which keep the robot on the desired path.

CONCLUSIONS

It was shown that adjusting the connection weights within the artificial neural network improved the performance of the robot controlled by the neural network, however, we were unable to fully optimize the connection weights to the desired solution. Some troubleshooting and adjustment work must be completed before the artificial neural network can satisfactorily control the robot model. Work is continuing in this area, and progress should be reported before September of 2008.

APPENDIX-A -Succcssful implementations of artifieial neural networks in roboties (not comprehensive)-

PROJECT NAME, RESEARCHERS	YEAR	SUMMARY OF WORK COMPLETED
Belgrade-USC hand. Liu, Bekey, Iberal	1989	ANN was trained to choose an appropriate method of grasping different sizes of cylinder. ANN used had 2 input nodes, one hidden layer of 4 nodes, and one output layer of 2 nodes. The network converged to grasping methods that matched human choices after 4,595 iterations.
Modeling of Robot Dynamics by NN with dynamic neurons. A G Chassiakos, E B Kosmatopoulous, and M A Christodoulou	1991	ANN used to model dynamics of robotic system. Weights adjusted with back propagation.
A CMAC NN for control of walking machine. Yi Lin and Shin-Min Song	1991	ANN was used in CMAC (cerebellar model articulation controller) to control movements of a simulated four legged walking machine in straight line motion. A feedforward ANN was used, with feedback employed to optimize connection weights.
A neural network based inverse kinematics solution in robotics. Yaotong Li and Nan Zeng	1991	Multilayer hopfield feedforward ANN was trained with back propagation to solve the inverse kinematics problem for a simulated 4 DOF robot. The ANN converged to a solution and was able to accurately control the arm position.
a one-eyed self learning robot manipulator	1991	ANN was trained to control movement of a simulated 6 DOF robot arm. 3 layer ANN was used: 7 node input layer, 15 node hidden layer, and 2 node output layer. ANN used was feedforward, but operated in a feedback control loop. Used a method similar to back-propagation for weight optimization. The ANN controlled arm was able to move within 5mm of target consistenty.
Generalism and extension of motor progams for a sequential recurrent network	1991	A recurrent Jordan ANN is trained to control movement of a simulated robotic arm. Controller dynamics were analyzed with a number of ANN designs. With variation in the number of input, hidden, and output nodes.
Fast sensorimotor skill acquisition based on rule-based training of neural networks. David Handelman, Stephen Lane	1991	ANN was used in CMAC to control simulated movement of a pole balanced on a cart. Desired output was calculated with rule based algorithm, this output was was used to train ANN utilizing back propagation. Hybrid controller of rule based algorithm and ANN was able to accurately control cart-pole movement.
Senses, skills, reactions and reflexes: learning automatic behaviors in multi-sensory robotic systems. Jack Gelfand, Marshall Flax, Raymond Endres, Stephen Lane, David Handelman	1991	ANN used in CMAC to control simulated robotic arms in a variety of tasks with visual input, including dribbling a basketball, grabbing an object, and avoiding an obstacle during movement.
Learning to understand and control in a world of events. Richard Elsey	1991	Created an ANN that used delayed differential hebian learning with simulated short and long term memory to recognize and react to patterns.
Developmental robotics: a new approach to the specification of robot programs. Andrew Fagg	1991	Trained an ANN to correctly select simple movement actions based on stimulus. Used a feedforward ANN with reinforcement learning. Utilized the leaky integrator model of the neuron. ANN was able to replicate results obtained by training monkeys to do the same task.
Self selection of input stimuli for improving performance, Stephano Nolfi, Domenico Parisi	1991	Created feedorward ANN that was trained to navigate toward targets on a simulated 2D surface. The ANN was trained with a genetic algorithm (random weight initiation, best fitted of each generation create next generation...) research showed that ANN trained with genetic algorithm can optimize performance in two ways: learning to act on any stimulus it receives, or only acting on a self-selected sub class of stimuli.

PROJECT NAME, RESEARCHERS	YEAR	SUMMARY OF WORK COMPLETED
Dynamic balance of a biped walking robot. Thomas Miller III, Andrew Kun	1997	Used 3 CMAC ANNs to train a biped robot to walk and balance (robot had 10 actuators, 10 motor position sensors, 8 foot force sensors, 2 body accelerometers). Used a totally of 3 ANN: foot contact, front to back balance & left to right balance. Feedforward ANNs used, with feedback error learning. ANNs modulated gaits produced by a central gait generator.ANNs implemented in hardware (35.7 Hz update). Following an hour of training the robot showed major improvement in walking.
Visual feedback in motion. Patrick van der Smagt, Frans Groen	1997	ANN trained to control a 3 DOF robotic arm (simulated OSCAR robot) with desired manipulator path as an input. ANN trained with back propagation. After a few trials arm was able to consistently get within 50mm of desired target in the desired time.
The neural dynamics approach to sensory-motor control. Paolo Gaudiano, Frank Guenther, Eduardo Zalama	1997	ANNs trained to control simulated robotic arm, simulated wheeled robot. ANN structure based on the VAN (vector associative map) and VITE (vector integration to end point) control models.
Neural Vehicles. Ben Kröse. Joris Van Dam	1997	ANN use in autonomous navigation is investigated. ANN was trained to control a car with images of the road ahead as input, feedforward ANN was used, trained with back propagation, with human driving actions used to calculate error. ANN was able to control vehicle at twice the speed of non-neural controller. navigation ANNs were also trained with adaptive hueristic critic (AHC) and Q-learning. ANN were also used in map building and path planning. It was shown that given a map represented in an ANN, the shortest path from one point to another can be found quickly.
Visually guided movements: learning with modular neural maps in robotics. Jean-Luc Buessler, Jean-Philippe Urban	1997	A modular ANN is used to control 3 DOF robotic arm with arm position images as an input. The neural structure consists of two layers, one layer to select a Neuromodule to control the arm at a given time, and a second layer consisting of Neuromodules optimized for specific situations. ANN used ADELINE model neurons. ANN weights updated with the Normalized least mean square (NLMS) rule.
Stable manipulator trajectory control using neural networks. Yichuang Jin, Tony Pipe, Alan Winfield	1997	ANN trained to control simulated 2 link robotic arm, as well as simulated PUMA 560 manipulator. Research showed that offline training of ANN guarantees stability and convergence to weight solution within the training set, and online training guarantees stability and asymptotic convergence to solution. Conventional adaptive controllers converged to solution faster than ANN. Authors suggest using hybrid controller consisting of ANN and conventional controllers
BISMARC: a biologically inspired system for map-based autonomous rover control. Terry Huntsbrger, John Rose	1997	ANN used in fuzzy self organizing feature map (FSOFM) and action selection algorithm to control 3 robots moving in a simulated Martian environment. Robots retrieved 4 widely separated containers with a 98.9% success rate. Control system used stereo vision as an input, along with global position information.
Evolutionary neurocontrollers for autonomous mobile robots. D Floreano, F Mondada	1998	Used a recurrent neural network with eight inputs and two outputs to control a Kedema robot. Robot learned to avoid collisions in a circular maze. ANN weights were adjusted with a genetic algorithm. Adapted ANN was transferred to a larger Koala robot, which quickly learned to navigate in a larger maze.
Evolutionary robots with on-line self-organization and behavioral fitness. D Floreano, J Urzelai	2000	Used ANN in robot control. Khepera robot was trained to explore an area until it ran out of energy, then return to recharge station. ANN was fully recurrent, discrete time, with 12 neurons. Genetic algorithm used to adjust the way ANN learns (genes coded for 4 hebb rules used for weight adjustment in ANN). Reinforcement learning was used to teach individual robots in population to accomplish task. Best adapted ANN from genetic algorithm was transfered to a new robot (larger Koala robot), where it quickly learned to accomplish the same task (even with different ANN inputs). Researchers proposed a "fitness space" to aid in the selection and classification of ANN fitness functions.
Fuzzy and recurrent neural network motion control among dynamic obstacles for robot manipulators. Jean Bosco Mbede, Wu Wei, Qisen Zhang	2001	A hybrid neuro-fuzzy controller is developed to control motor torques in a simulated 3 DOF robotic manipulator. System used a modified Elman NN. Simulated manipulator was able to avoid moving obstacles, and adapt to changes in dynamics in manipulator.
Reinforcement learning neural network to the problem of autonomous mobile robot obstacle avoidance. Bing-Qiang Huang, Guang-Yi Cao, Min Guo	2005	ANN used to control a simulated robot in an environment with obstacles. ANN used Q-learning with back propagation to optimize ANN connection weights. ANN used had 8 input nodes, 16 hidden nodes, and 5 output nodes. Robot navigated around obstacles without errors after 500 weight adjustment epochs.

REFERENCES

A. G. Chassiakos, E. B. Kosmatopoulos & M. A. Christodoulou. "Modeling of Robot Dynamics by Neural Networks with Dynamic Neurons". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Aaron Kuan, C. B. Whittle Jr. & Behnam Bavarian. "Neurocontroller Selective Learning from Man-in-the-Loop Feedback Control Actions". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Alex L. P. Tay, Jacek M. Zurada, Lai-Ping Wong & Jian Xu. "The Hierarchical Fast Learning Artificial Neural Neural Network (HieFLANN) – An Autonomous Platform for Hierarchical Neural Network Construction". *IEEE Transactions on Neural Networks*, Volume 18, Number 6, November 2007.

Andreas Bühlemeier & Gerhard Manteuffel. "Operant Conditioning in Robots". *Neural Systems for Robotics*. Academic Press 1997.

Andrew H. Fagg. "Developmental Robotics: A New Approach to the specification of Robot Programs". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Ben J. A. Kröse, P. Patrick van der Smagt & Franz C.A. Groen. "A One-eyed Self Learning Robot Manipulator". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Ben Kröse & Joris van Dam. "Neural Vehicles". *Neural Systems for Robotics*. Academic Press 1997.

Bing-Qiang Huang, Guang-Yi Cao & Min Guo. "Reinforcement Learning Neural Network to the Problem of Autonomous Mobile Obstacle Avoidance". *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, August 18-21, 2005.

Bridget Hallam, John Hallam & Gillian Hayes. "A Dynamic Net for Robot Control". *Neural Systems for Robotics*. Academic Press 1997.

D Floreano & F Mondada. "Evolutionary Neurocontrollers for autonomous mobile robots". *Neural Networks 11* (1998) pp. 1461-1478

D. Floreano, J Urzelai. "Evolutionary robots with on-line self-organization and behavioral fitness". *Neural Networks 13* (2000) pg. 231-443

David A. Handelman & Stephen H. Lane. "Fast Sensorimotor Skill Acquisition based on Rule-Based Training of Neural Networks". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

David DeMers & Kenneth Kreutz-Delgado. "Inverse Kinematics of Dextrous Manipulators". *Neural Systems for Robotics*. Academic Press 1997.

David DeMers & Kenneth Kreutz-Delgado. "Learning Global Topological Properties of Robot Kinematic Mappings for Neural Network based Configuration Control". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Dmitry V Lebedev, Jochen J Steil & Helge J Ritter. "The dynamic wave expansion neural network model for robot motion planning in time-varying environments". *Neural Networks 18* (2005) pp. 267-285

G Metta, G Sandini & J Konczak. "A developmental approach to visually-guided reaching in artificial systems". *Neural Networks 12* (1999) pp. 1413-1427

George A. Bekey. "Control of Grasping in Robot Hands by Neural Networks and Expert Systems". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Hugues Bersini, Luis Gonzales Sotelino & Eric Decossaux. "Hopfield Net Generation and Encoding of Trajectories in Constrained Environment". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Itel E. Dror, Mark Zagaeski, Damien Rios & Cynthia F. Moss. "Neural Network Sonar as Perceptual Modality for Robotics". *Neural Systems for Robotics*. Academic Press 1997.

J. D. Schall & D. P. Hanes. "Neural mechanisms of selection and control of visually guided eye movements". *Neural Networks 11* (1998) pg. 1241-1251

Jaakko Malmivuo & Robert Plonsey. *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*,

Oxford University Press, New York, 1995.

Jaek Gelfand, Marshall Flax, Raymond Endres, Stephen Lane & David Handelman. "Senses, Skills, Reactions and Reflexes: Learning Autonomous Behaviors in Multi-Sensory Robot Systems". *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Jean Boseo Mbede, Wu Wei & Qisen Zhang. "Fuzzy and Reeurrent Neural Network Motion Control among Dynamie Obstaeles for Robot Manipulators". *Journal of Intelligent and Robotic Systems 30* (2001) pg. 155-177

Jean-Lue Buessler & Jean-Philippe Urban. "Visually Guided Movements: learning with modular neural maps in roboties". *Neural Networks 11* (1998) pg. 1395-1415

Jukka Heikkonen & Pasi Koikkalainen. "Self-Organization and Autonomous Robots". *Neural Systems for Robotics*. Academie Press 1997.

Kenol Jules & Paul P. Lin. "Artificial Neural Network Applications: from aireraft design optimization to orbiting spaceeraft on-board environment monitoring". NASA, August 2002.

Kumpati S Narendra & Kannan Parthasarathy. "Identification and Control of Dynamical Systems Using Neural Networks". *IEEE Transactions on Neural Networks* Vol. 1 No. 1 March 1990.

Mark Meltser, Moshe Shoham & Larry M Manevitz. "Approximating Funetions by Neural Networks: A Construetive Solution in the Uniform Norm". *Neural Networks 9*, No 6, pp. 965-978, 1996.

Michael C. Moed & Robert B. Kelley. "Robotie Task Planning Using a Conneetionist/Symbolie System". *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Paolo Gaudiano, Frank H. Guenther & Eduardo Zalama. "The Neural Dynamies Approaeh to Sensory-Motion Control". *Neural Systems for Robotics*. Academie Press 1997.

Patriek van der Smaght & Frans Groen. "Visual Feedbaek in Motion". *Neural Systems for Robotics*. Academie Press 1997.

Riehard K. Elsley. "Learning to Understand and Control in a World of Events". *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Roy Glasius, Andrzej Komoda & Stan Gielen. "Neural Network Dynamies for Path Planning and Obstaele Avoidanee". *Neural Networks 8* (1995) pp. 125-133

Russell D. Reed & Robert J. Marks II. *Neural Smithing: supervised learning in feedforward artificial neural networks*. MIT Press 1999.

Sandrine Allemand, Franeois Blane, Yves Burnod, Miehel Dufossé & Lue Lavayssiére. " A Kinematie & Dynamie Robot Control System Based on Cerebro-Cerebellar Interaetion Modelling". *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Shahriar Najand, Zhen-Ping Lo & Behnam Bavarian. "Application of Self-Organizing Neural Networks for Mobile Robot Environment Learning" . *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Terry Huntsberger & John Rose. "BISMARC: a biologieally inspired system for map based autonomous rover eontrol". *Neural Networks 11* (1998) 1497-1510

W. Thomas Miller III & Andrew Kun. "Dynamie Balanee of a Biped Walking Robot". *Neural Systems for Robotics*. Academie Press 1997.

Yaotong Li & Nan Zeng. "A Neural Network Based Inverse Kinematies Solution in Roboties". *Neural Networks in Robotics*. Kluwer Academie Publishers. 1991.

Yasuo Kuniyoshi & Lue Berthouze. "Neural learning of embodied interaetion dynamies". *Neural Networks 11* (1998) pg. 1259-1276

Yataka Nakamura, Takeshi Mori, Masa-aki Sato & Shin Ishii. "Reinforeement learning for a biped robot based on a CPG-aetor-eritie method". Neural Networks 20(2007) pg. 723-735

Yataka Nakamura, Takeshi Mori, Masa-aki Sato & Shin Ishii. "Reinforeement learning for a biped robot based on CPG-aetor-eritie method". *Neural Networks 20* (2007) pg. 723-735

Yi Lin & Shin-Min Song. "A CMAC Neural Network for the Kinematic Control of Walking Machine". *Neural Networks in Robotics*. Kluwer Academic Publishers. 1991.

Yichuang Jin, Tony Pipe & Alan Winfield. "Stable Manipulator Trajectory Control Using Neural Networks". *Neural Systems for Robotics*. Academic Press 1997.

Appendix A

University of Washington Consultant Biographies

CALIFORNIA CENTRAL COAST
RESEARCH PARTNERSHIP

California Polytechnic State University
Bldg 38-101 • San Luis Obispo, CA 93407

Research and Graduate Programs
phone (805)756-5551 • fax (805)756-1725

Ken Walters
University of Washington

Kenneth D. Walters is Professor of Business (Emeritus) at the University of Washington's Bothell Business Program. He is the author of numerous articles in *The Harvard Business Review, California Management Review, Columbia Journal of World Business, Ecology Law Quarterly, Industrial Relations*, and the *Academy of Management Review*.

His most recent article, "University-Related Startup Companies," in *Developing University-Industry Relations: Pathways to Innovation from the West Coast* (Jossey Bass, 2009), studies factors to create entrepreneurial companies from university research.

His books include *Nationalized Companies* (McGraw-Hill, 1983) and *Entrepreneurial Management: New Technologies and New Markets* (Ballinger, 1990), and have been reviewed in *The New York Times, The Wall Street Journal,* and *The New York Review of Books.*

He has given over 250 speeches and lectures throughout the United States, Canada, Europe, and Japan, including guest presentations at the University of California, Berkeley, Harvard University, Oxford University, Keio University, and the University of Paris. In 1980-81 he was a Fellow at the Camargo Foundation in Cassis, France.

A graduate of Stanford Law School (J.D., 1966), he also holds a Ph.D. (1971) from the Haas School of Business at the University of California, Berkeley. He is a member of the California Bar Association, the American Economic Association, and the Academy of Management.

For many years he served as Dean of the College of Business at California Polytechnic State University, San Luis Obispo and Dean of the Robert O. Anderson Graduate School of Management at the University of New Mexico. Other administrative responsibilities included: Executive Director and Dean of the Palm Desert Graduate Center, University of California, Riverside; Chairman of the Department of Business, Government, & Society at the Graduate School of Business, University of Washington; and Director of the Business Program at the University of Washington's Bothell campus.

Ken is a Research Fellow of the IC^2 Institute at The University of Texas at Austin, currently studying French high-tech startup companies. He has written widely about the impact of university research on community economic development and university-based startup companies, including two studies: *University of Washington: Engine of the Knowledge-based Economy* (2001) and *Creating the Future: University of Washington Startup Companies, 1973-2003* (2003).

He has taught numerous courses and received teaching awards at the undergraduate, MBA, Executive MBA, and PhD levels. Recent courses taught include: New Technologies and New Markets; Creative and Innovative Management; Business Ethics and Corporate Social Responsibility; Law for Business Executives; and Law for New Enterprise Development.

Contact:
kwalters@u.washington.edu

CALIFORNIA CENTRAL COAST
RESEARCH PARTNERSHIP

California Polytechnic State University
Bldg 38-101 • San Luis Obispo, CA 93407

Research and Graduate Programs
phone (805)756-5551 • fax (805)756-1725

Alvin L. Kwiram
University of Washington

Alvin Kwiram became Vice Provost at the University of Washington on January 1, 1987. In 1990 he was appointed the first Vice Provost for the newly established Office of Research , a position he held until 2002. He served as Chair of the Department of Chemistry at the University of Washington from 1977 to 1987. From 1964 to 1970 he was on the faculty at Harvard University, before coming to the University of Washington.

He received his BA in Physics and BS in Chemistry from Walla Walla College in 1958, and he completed the work for his PhD in Chemistry at the California Institute of Technology in 1962. During 1962-63 he was awarded the Alfred A. Noyes Instructorship at Caltech. During 1963-64 he was a Research Associate in Physics at Stanford University working with Professor William Fairbank studying the newly discovered phenomenon of quantized flux in superconductors.

Dr. Kwiram is a Fellow of the American Physical Society and the American Association for the Advancement of Science (AAAS). He has served as Chair of the Chemistry Section of the AAAS, and served a six-year term as a member of the AAAS Program Committee for the Annual Meeting. He was also a member of the Executive Committee of the Division of Physical Chemistry in the American Chemical Society for a number of years. He served on the Founding Board of the Council for Chemical Research (CCR) and was the first academic chair of the Board of Directors in 1982-83. The Council for Chemical Research was one of the first initiatives in the US created for the purpose of bringing academe and industry together for more constructive dialogue and collaboration. In 1986, Dr. Kwiram was selected to receive the CCR Award for the Promotion of University/Industry Relations. From 2000-2003 he served on the Executive Committee of the Council on Research Policy and Graduate Education (CRPGE) of the National Association of State Universities and Land Grant Colleges (NASULGC), and as chair of CRPGE in 2001-2002. From 1998-2001 he was a member of the Divisional Review Committee for the Pacific Northwest National Laboratory (PNNL), and from 2000 to 2007 was a member of the Laboratory Advisory Committee for PNNL. He served as the chair of the Graduate Education Advisory Board for the American Chemical Society (2006-2008), and as chair (2006-2008) of the Academic Advisory Board of the Worldwide Universities Network, a consortium of some 16 international research universities. He has served on a number of other boards, including both non-profit and for-profit.

Dr. Kwiram has published over 75 papers in the field of physical chemistry emphasizing the development of novel magnetic resonance and optical techniques designed to probe the electronic structure of molecular systems in the solid state. He has been a Woodrow Wilson Fellow, an Alfred P. Sloan Fellow and a John Simon Guggenheim Memorial Foundation Fellow. During 1977-78 he spent part of his sabbatical at the University of California, Berkeley. In 1985-86 he was a Visiting Professor in the Physics Department at the University of Stuttgart, Germany. In 2006 he spent spring quarter on sabbatical at Wolfson College, Oxford University.

On March 15, 2002 he stepped down as Vice Provost for Research and returned to the Department of Chemistry. For the next five years he served as the Executive Director of the (national) NSF Science and Technology Center on Materials and Devices for Information Technology Research. Although he made the transition to emeritus faculty member in 2007, he continues to be active in university related affairs.

Appendix B

Project Related Thesis & Relevant Publications

47th AIAA Aerospace Sciences Meeting Including The New Horizons Forum and Aerospace Exposition
5 - 8 January 2009, Orlando, Florida
AIAA 2009-1441
N00014-07-1-1152
P.I.Susan C. Opava, Ph.D.

Underpowered Aircraft -- Performance and Operational Possibilities

Andrew S. Ezzard[1], Michael R. Vallone[2], and Robert A. McDonald, Ph.D.[3]
California Polytechnic State University, San Luis Obispo, CA 93407

A unique configuration, known as an Underpowered Aircraft, allows for the modification of gliding flight vehicles for increased range and lower cost when compared with fully powered flight vehicles. Intentionally under-sizing the powerplant for a flight vehicle allows the designer to choose a powerplant that will not only perform the mission requirements, but will also provide the customer with the most cost effective solution, as some missions may not require fully powered flight. Specifically, the underpowered aircraft concept studied in this paper is a gliding flight aircraft that does not have enough power for climbing or level flight, but does have enough power to overcome some of the drag forces associated with flight, in turn increasing the effective range of the vehicle. In this paper, the underpowered aircraft concept was analyzed and its feasibility was determined. Analysis done using equations of motion, followed by a more accurate numerical integration including a thrust lapse, determined that the underpowered aircraft concept provides a unique method for a cost effective range extension technology for gliding flight vehicles. Finally, the technology and methods of this paper were applied to the AGM-154 JSOW and JSOW-ER glide munitions and it was determined that JSOW-ER is representative of an underpowered aircraft with our analysis. This paper represents a "back-of-the-envelope" investigation into the underpowered aircraft concept.

Nomenclature

English

AR	=	aspect ratio
AR^*	=	representative aspect ratio = eAR
C_D	=	drag coefficient
C_L	=	lift coefficient
D	=	drag
W	=	weight
E	=	energy
e	=	Oswald efficiency factor
h	=	height, altitude
K	=	$\frac{1}{\pi eAR} = \frac{1}{\pi AR^*}$
L	=	lift
L/D	=	lift to drag ratio
m	=	mass
P/W	=	power to weight ratio
P_s	=	specific excess power
R/C, RoC	=	rate of climb
T	=	thrust
T/W	=	thrust to weight ratio
V	=	velocity
W/P	=	power loading
Z_e	=	energy height

Greek

α	=	angle of attack
η_p	=	propeller efficiency
θ	=	flight path angle
φ_T	=	thrust vectoring angle

Subscripts

max	=	maximum
o	=	parasite
T	=	thrust
to	=	takeoff

Acronyms

KE	=	kinetic energy
PE	=	potential energy
UAV	=	unmanned aerial vehicle

1Aerospace Engineering Undergraduate, AIAA Member
2Aerospace Engineering Graduate Student, AIAA Member
3Assistant Professor, Lockheed Martin Endowed Professor, Aerospace Engineering Department, AIAA Member

Copyright © 2009 by Andrew S. Ezzard. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.

I. Introduction

IT is not typical for an aircraft designer to undersize the powerplant for their aircraft. However, as the emphasis on cost in the aerospace industry continues to increase, the need to meet customer requirements through the most cost effective means presents itself in almost all engineering problems. In a time where the cost of fuel is fluctuating unpredictably, manufacturing and labor costs are increasing, and high technology systems bring the development cost up, the need to perform and fly a mission must be as cost effective as possible. Using the technology of an "underpowered" aircraft (an aircraft that only has enough power to overcome some of the drag forces associated with flight) has the potential to reduce the operating and propulsion system cost when compared to standard aircraft systems that are capable of climbing or level flight. Possible mission applications for the technology include cargo delivery for deployed troops in the battlefield, glide munitions, stand-off weapons, and others.

The underpowered aircraft concept comes from the idea that, for specific missions, an aircraft system may not necessarily need to overcome all of the drag forces associated with flight, and can therefore operate with an undersized powerplant. If the purpose of the aircraft is to glide to its target, then the effective range of the aircraft can be increased by making the aircraft underpowered. By doing so, the aircraft does not need the same power as a standard aircraft would to achieve level or climbing flight, but can successfully operate off of smaller amounts of power while greatly extending the range of the vehicle. Adding power to the aircraft, in small amounts, allows the aircraft to increase the effective mission range, for little horsepower and low cost, as will be shown in this paper. The performance of an underpowered aircraft will be analyzed and the cost advantages will be determined. A more detailed numerical integration analysis of the flight path will also be presented and validation of the analysis will be conducted when compared to a glide munition operating off of the same flight principles.

II. Underpowered Aircraft Performance

The idea of an underpowered aircraft arises to fill the need to increase the range of gliding aircraft systems. Payload delivered using a gliding system provides for a cheap and simple delivery mechanism when constructed from light weight, inexpensive materials. The system can be designed to be single use or reusable, depending on the exact purpose. However, just using a standard gliding aircraft does not provide the range capability often needed for payload delivery. Assuming the underpowered aircraft is dropped from a high altitude, an engine can be sized and selected to provide the desired range. The trade space between range, flight velocity, and drop altitude for an underpowered aircraft will be explored. The following analysis will assume a payload delivery mission and then other mission profiles will be analyzed afterwards.

Two methods were used to determine the performance for the underpowered aircraft. First, the standard method of looking at the forces acting on the aircraft via a free-body diagram was applied. Estimates for the drag polar of the aircraft were also developed, which yielded an estimate of flight conditions. This was followed by a numerical integration to provide a more accurate measure and to account for thrust lapse. Emphasis was placed on the lift to drag ratio (L/D) due to the vehicle's operation in gliding flight. In an abstract trade study such as this, aircraft characteristics are treated as technology that can be applied to an aircraft in the design process, allowing for an overall performance trade study.

A. Equations of Motion Analysis

To start, Figure 1 shows the standard forces acting on an aircraft in flight for use in the equations of motions analysis.

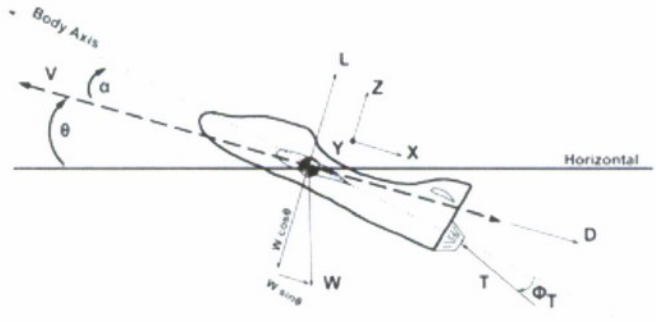


Figure 1. Forces Acting on an Aircraft in Flight

The forces acting on the aircraft are standard and include the lift, drag, thrust, and weight of the aircraft. The velocity, flight path angle, θ, and the angle of attack, α, are also shown. The flight path angle is defined as the angle between the horizontal and the velocity vector and the angle of attack is defined as the angle between the velocity vector and the arbitrarily defined body axes for the aircraft. In the case of thrust vectoring, the thrust vector is offset by the angle φ_T, to model any effects from thrust vectoring on the performance of the aircraft.

This analysis assumes that there is no thrust vectoring happening in the flight of the aircraft. Thus, summing the forces in the X and Z-axis for steady flight yields

$$T = D + W\sin\theta \tag{1}$$

$$L = W\cos\theta \tag{2}$$

Thus, we reorganize equation 1 to find the classical aircraft performance equation

$$\sin\theta = \frac{T-D}{W} \tag{3}$$

For gliding flight, our interest lies in the flight path angle, θ, of the vehicle. By minimizing the flight path angle to a small negative number, the range of the aircraft can be maximized. So, we have

$$\theta = \sin^{-1}\left(\frac{T}{W} - \frac{D}{W}\right) \tag{4}$$

From equation 2 above and a small angle approximation, we have

$$\theta = \sin^{-1}\left(\frac{T}{W} - \frac{1}{\frac{L}{D}}\right) \tag{5}$$

With the power to weight ratio (in units of hp/lb) defined as

$$\frac{P}{W} = \left(\frac{T}{W}\right)\frac{V}{550\eta_p} \tag{6}$$

B. Drag Polar

One of the key characteristics needed to properly model the performance of an aircraft is an accurate representation of the drag polar for the entire aircraft. The form of the drag polar that will be used for the analysis is

$$C_D = C_{Do} + \frac{C_L^2}{\pi AR^*} \tag{7}$$

For the purposes of this study, the aspect ratio, AR, of the aircraft will be grouped together with the Oswald efficiency factor, e, to become a representative aspect ratio, AR^*. In an abstract trade study such as this no information is available on the efficiency of the wing design; however a feel for the efficiency factor can still be gained through a representative aspect ratio. If we assume that the aircraft is operating at conditions that give the C_L for best L/D for the gliding condition, then we know that[1]

$$C_L^* = \sqrt{\frac{C_{Do}}{\pi AR^*}} \tag{8}$$

$$C_D^* = 2C_{Do} \tag{9}$$

So, the lift to drag ratio is

$$\frac{L}{D}_{max} = \frac{C_L}{C_D} = \frac{\sqrt{\frac{C_{Do}}{\pi AR^*}}}{2C_{Do}} = \sqrt{\frac{1}{4\pi AR^* C_{Do}}} \tag{10}$$

Reorganizing, an estimate of the parasite drag can be determined through

$$C_{Do} = \frac{\pi AR^*}{4\left(\frac{L}{D_{max}}\right)^2} \tag{12}$$

This allows us to treat the aerodynamics of the aircraft, AR* and $\frac{L}{D_{max}}$, as technology.

III. Powerplant Cost Estimation

With the performance for the underpowered aircraft concept established, the cost benefits of the technology were determined. Such a vehicle can be powered either by traditional internal combustion piston engines or small jet turbine engines. In order to get an estimate for the cost saving of the vehicle, estimates of small piston and jet turbine engines were developed. As seen below in Figure 2, a span of cost competitive small piston engines was used to determine a cost trend that was acquired through a survey of retail prices found online in August, 2008. These engines included small RC aircraft engines from O.S. Engines, general purpose engines from Honda and Kawasaki, as well as a few larger Rotax aircraft engines. The model yields an excellent correlation coefficient and the results are consistent with the investigative nature of this paper.

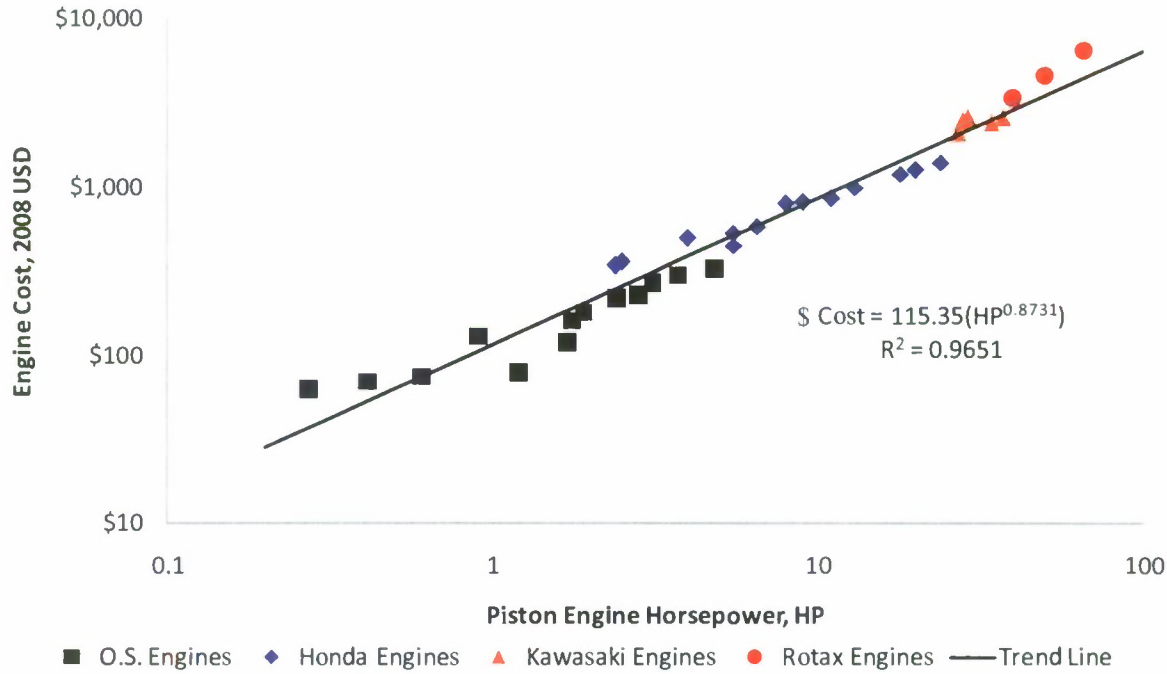


Figure 2. Small Piston Engine Cost

Figure 3 shows a similar trend for small jet turbine engines. Due to the fact that the payload delivery vehicles and stand-off "glide munitions" where this technology would be used are not large in size, the engine study was limited to model aircraft jet engines as well as jet engines used in small UAV applications. These engines include manufacturers such as JetCat, foreign engine manufacturers such as SimJet, and others. This model is indicative of the low thrust engines needed for underpowered applications. The model yields acceptable results with a correlation coefficient of 82%.

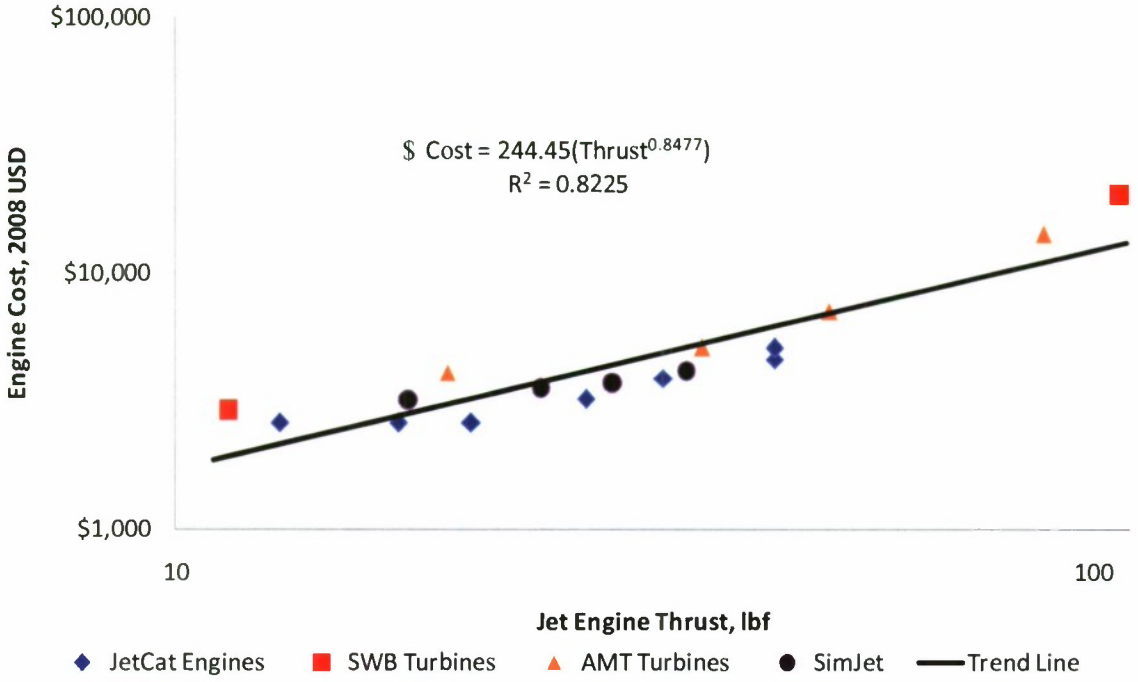


Figure 3. Small Jet Turbine Engine Cost

IV. Underpowered Aircraft Flight Performance

Since the concept for this aircraft is based on that of a glider, the performance of the aircraft has been determined to represent the trade space of velocity, range, and drop altitude. So, using the relationships derived earlier and performance code written in MATLAB, some performance trends were calculated.

It is assumed in this analysis that the underpowered aircraft is dropped from altitude at the start of the flight. Two altitudes of 10,000 feet and 25,000 feet were chosen to give a comparison between a low altitude drop and a high altitude drop. The drop aircraft could consist of any type of vehicle that is capable of carrying the underpowered aircraft to the drop altitude. Winds aloft are not included in the analysis. The analysis presented assumes that the underpowered vehicle is dropped at the velocity of best *L/D* and that during the glide the vehicle operates at the best *L/D* for glide.

As the lift to drag ratio is increased, the thrust required to get the aircraft to reach its destination decreases, which can be seen in Figure 4. Only a small amount of thrust is required to keep the aircraft in a level flight condition, but a significantly larger amount of thrust is needed to have a positive rate of climb for the aircraft (represented by the +3 degree climb angle curve). The difference in gliding flight, level flight, and positive rate of climb flight is indicative of the trade space between range, velocity, and drop altitude (10,000 feet for these curves). It is also indicative of the trade in the cost of the system, as having an aircraft that is capable of positive rate of climb requires a significantly larger amount of thrust and therefore a larger, more expensive powerplant. It is important to note that there are no additional assumptions built into the figure below and the plot is derived solely from aircraft performance metrics.

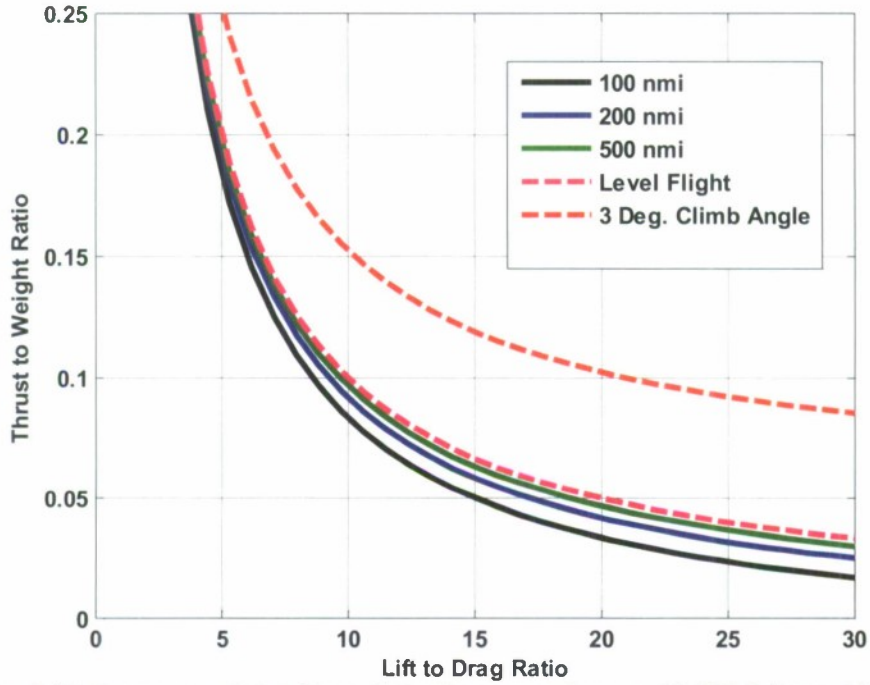


Figure 4. Underpowered Jet Aircraft Performance from a 10,000 ft Drop Altitude

The curves in Fig. 4 are a function of thrust to weight ratio and lift to drag ratio which are two of the most fundamental aircraft performance parameters. For these initial calculations the thrust lapse has been ignored. Using the cost model presented earlier we can estimate the cost saving of the underpowered aircraft technology. Assuming that the vehicle has a takeoff gross weight (TOGW) of 1,500 pounds and can achieve a lift-to-drag ratio (L/D) of 20, the vehicle requires a thrust to weight ratio 0.04 to achieve a range of 200 nmi. This corresponds to a thrust required of 60 lbf and an engine cost of $7,862. In contrast, the engine would need to provide 75 lbf for level flight and 150 lbf for climbing flight. This yields a cost of $9,500 and $17,095, respectively.

The same curves can be represented for a propeller driven aircraft by including flight velocity and propeller efficiency. Shown in Figure 5 are the same trends as a function of power to weight ratio.

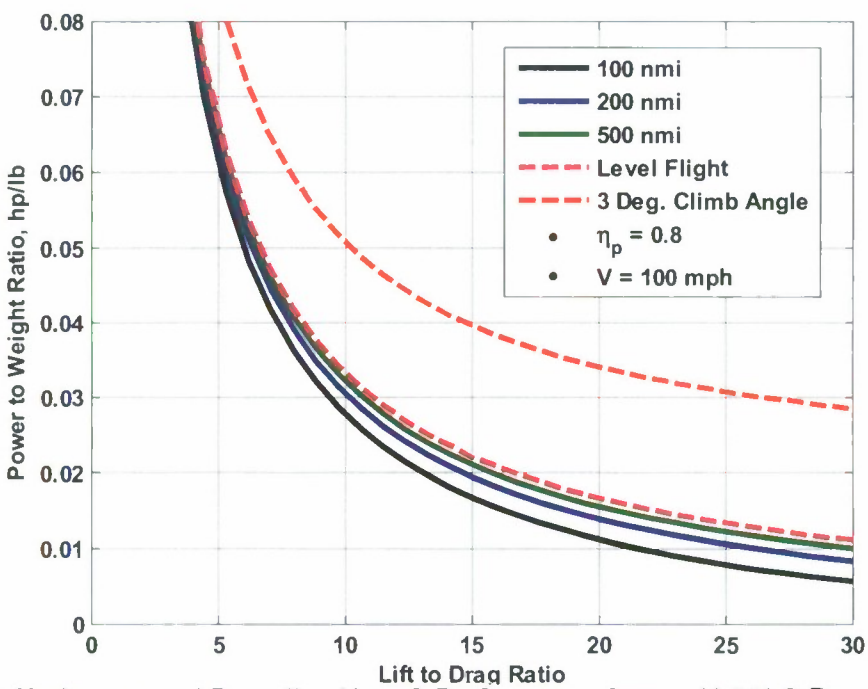


Figure 5. Underpowered Propeller Aircraft Performance from a 10,000 ft Drop Altitude

Figure 5 shows similar trends to Figure 4. This plot speaks volumes about the underpowered aircraft concept. If we assume that the underpowered aircraft can achieve a maximum lift to drag ratio of 20 and has a TOGW of 1,500 pounds, then the aircraft only requires about 20 hp to achieve a range of 200 nmi (even less for shorter ranges). The same aircraft would require 26 hp for level flight, and about 53 hp for climbing flight. The underpowered aircraft requires over one half the horsepower to achieve the mission requirement which translates to engine costs of $1,577, $1,983, and $3,694 respectively. There are two main assumptions built into this analysis: First, a propeller efficiency of 0.80 is assumed to convert from thrust to horsepower (80% being typical performance of most propellers currently used today) and second, an operating velocity of 100 mph (146.6 ft/sec) was arbitrarily chosen to represent a reasonable delivery speed for the payload aircraft.

A different presentation of the data in Fig. 5, seen below in Fig. 6, shows the relationship between power loading and lift to drag ratio for the underpowered aircraft.

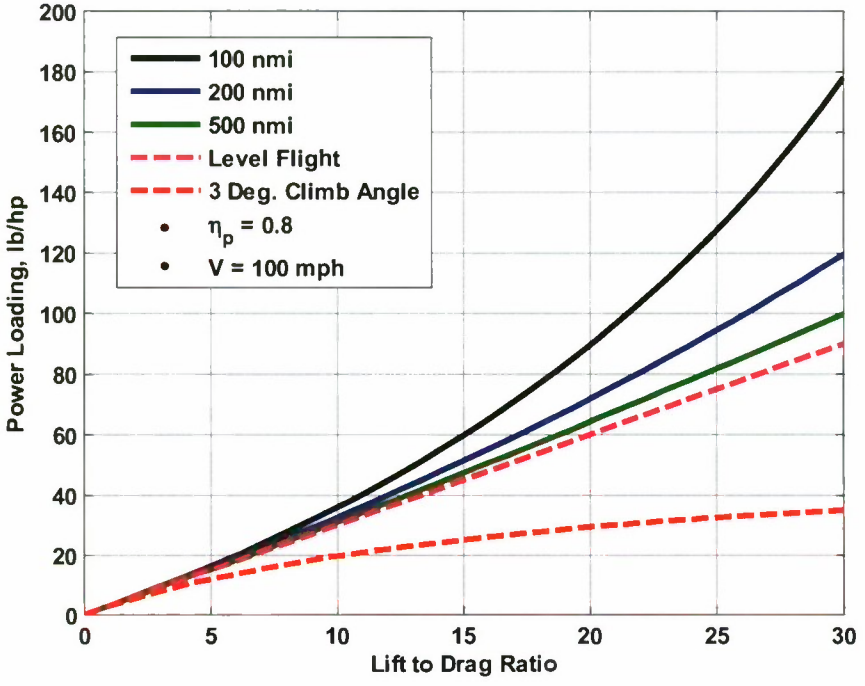


Figure 6. Underpowered Aircraft Power Loading Requirements for a 10,000 ft Drop Altitude

This figure shows how much of the aircraft (in weight) can be supported for each unit of horsepower that the engine on the aircraft will produce. If we again assume that the aircraft has a lift to drag ratio of 20 and a weight of 1,500 pounds, then the aircraft can carry about 76 pounds for each horsepower. This results in a power to weight ratio of 0.013 hp/lb and about a 20 hp engine as shown above in the previous figures. This analysis also assumes a propeller efficiency of 0.80 and a flight velocity of 100 mph as mentioned earlier.

The same graphs were created to look at the performance of the aircraft from the drop altitude of 25,000 feet. Note the significantly less power required for the underpowered aircraft to achieve a desired range of 200 nmi in Figure 7.

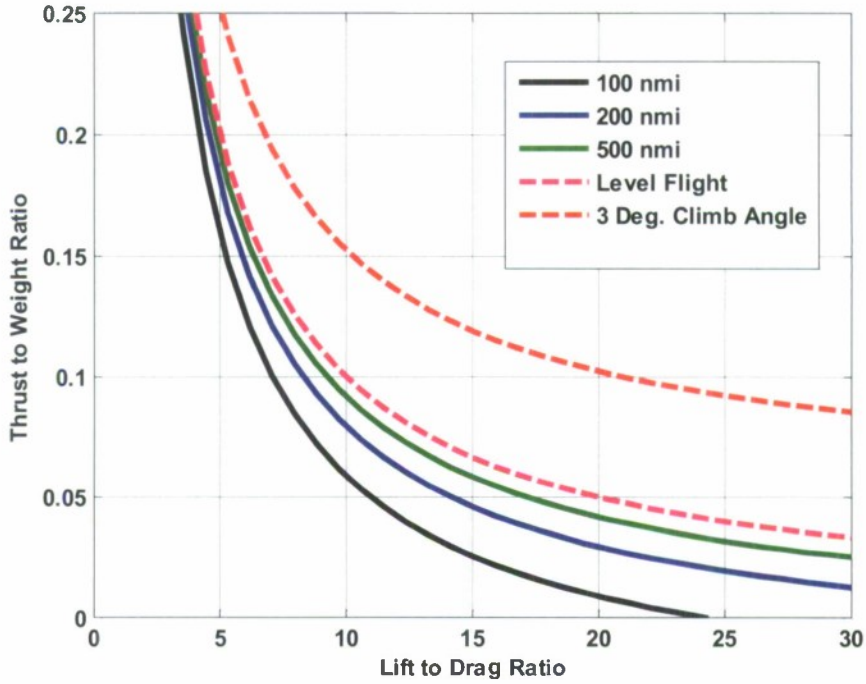


Figure 7. Underpowered Jet Aircraft Glide Performance from a 25,000 ft Drop Altitude

If we continue to assume that the aircraft has a TOGW of 1,500 pounds and can achieve a lift-to-drag ratio of 20, the vehicle will require a thrust to weight ratio of 0.03 to achieve a 200 nmi range, a thrust to weight ratio of 0.05 to achieve level flight, and a thrust to weight ratio of 0.1 to achieve a positive rate of climb. From the cost models presented above, the engine would need 45 lbf thrust, 75 lbf thrust, and 150 lbf thrust respectively. This would represent a significant cost savings between $6,161, $9,500, and $17,095 respectively.

Shown in Figure 8 is the relationship between power to weight ratio and lift to drag ratio.

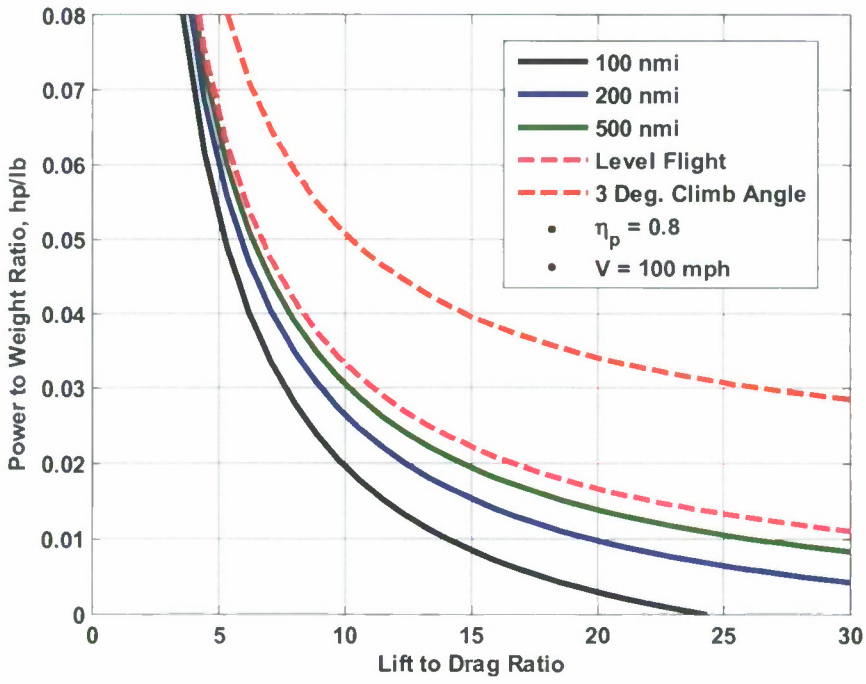


Figure 8. Underpowered Propeller Aircraft Performance from a 25,000 ft Drop Altitude

Using the same assumptions as before and looking at an aircraft that can achieve an aerodynamic lift to drag ratio of 20, and has a weight of 1,500 pounds, the aircraft will require a power to weight ratio of 0.01 hp/lb or 15 hp to reach the desired range of 200 nmi. This would be a minimal cost of $1,227. This is significantly less than an aircraft that would need to sustain level flight (approx. 26 hp or $1,983) and an aircraft looking to have a positive rate of climb (approx. 53 hp or $3,694). The underpowered aircraft technology will satisfy the requirements with significantly less cost due to the much smaller engine required which reduces cost throughout the life cycle. These values are sensitive to wing technology (lift to drag ratio) and will change depending on the *L/D* that can be achieved for the vehicle.

The power loading for the underpowered aircraft also significantly changes for the additional drop altitude of 25,000 feet and allows the aircraft to glide a longer distance with less power.

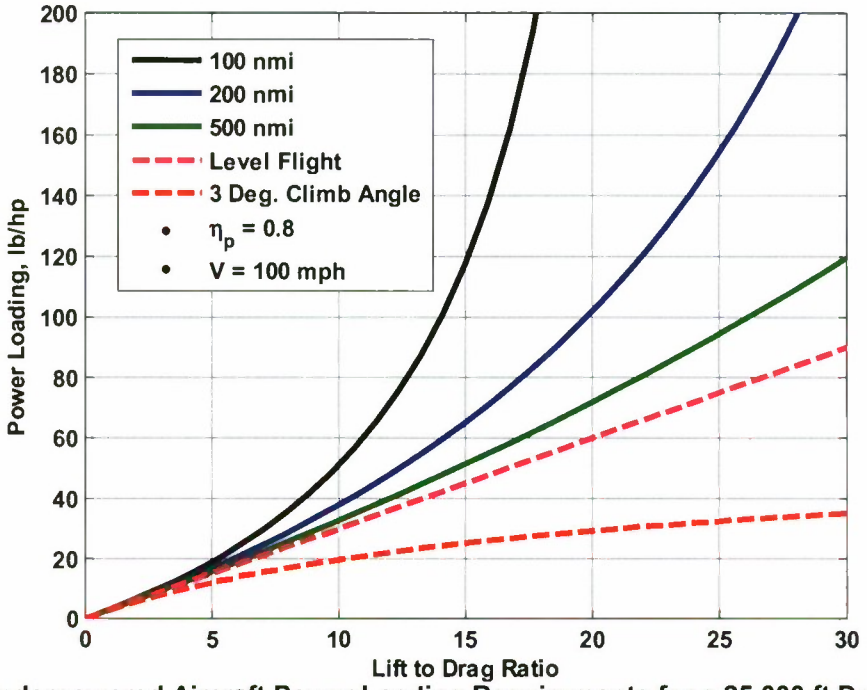


Figure 9. Underpowered Aircraft Power Loading Requirements for a 25,000 ft Drop Altitude

Comparing the two drop altitudes shows significant savings as well. Using the same aircraft assumptions as before, namely a TOGW of 1,500 pounds and a maximum lift to drag ratio of 20, we see that a drop from 10,000 ft achieves 200 nmi with a 60 lbf engine and an engine cost of $7,862. This cost is decreased when dropped from 25,000 ft, when the plane achieves the same range with a 45 lbf engine and an engine cost of $6,161.

V. Numerical Integration Approach

While the above analysis represents a "back-of-the-envelope approach", there is one main factor that was left out; the effects of velocity and altitude on powerplant performance. This takes the form of a thrust lapse that will change the maximum thrust to weight ratio represented in the earlier figures, to a typical flight thrust to weight ratio that has been adjusted to represent actual flight conditions. To achieve this, a numerical integration approach was used to examine the equations of motion, take a thrust lapse into account, and give more accurate results than the methods presented above.

Revisiting the free body diagram in Figure 1 and summing the forces in the X and Z-axis, we have

$$\cos\theta = \frac{L}{W} \tag{13}$$

$$\sin\theta = \frac{T-D}{W} \tag{14}$$

Due to the fact that our interest lies in gliding flight, we are actually interested in the change in altitude of the vehicle during flight with respect to the change in distance the vehicle travels along the ground. This is essentially the glide ratio in derivative form. So, using the free body diagram

$$\frac{ds}{dh} = \frac{\frac{ds}{dt}}{\frac{dh}{dt}} = \frac{Vcos\theta}{Vsin\theta} = \frac{1}{tan\theta} \tag{15}$$

$$\frac{1}{tan\theta} = \frac{L}{T-D} \tag{16}$$

Using a thrust lapse equation from *Mattingly*[2] defined in terms of the Mach number, M, and the density ratio from sea level, σ, we have

$$\alpha = 0.76\{0.907 + 0.262(|M - 0.5|)^{1.5}\}\sigma^{0.7} \tag{17}$$

Setting up the integral to integrate the flight path during steady flight, we have

$$\int_h^0 \left(\frac{\frac{1}{W}}{\frac{1}{W}}\right)\frac{L}{\alpha T - D}dh \tag{18}$$

$$\int_h^0 \left(\frac{1}{\alpha\frac{T}{W} - \frac{1}{\frac{L}{D}}}\right)dh \tag{19}$$

The numerical integration yielded results that provide a better representation of the flight regime with the inclusion of a thrust lapse. The trend in the curves shows that the numerical model actually requires more thrust than the equation of motion or energy method analysis because it now takes powerplant performance into account. The powerplant will not perform at sea level static conditions at high altitude. In an effort to validate the results of the model presented above, a drop altitude of 40,000 ft was used to match the ranges given for the AGM-154 Joint Standoff Weapon[3]. Figure 10 shows the results of the numerical integration.

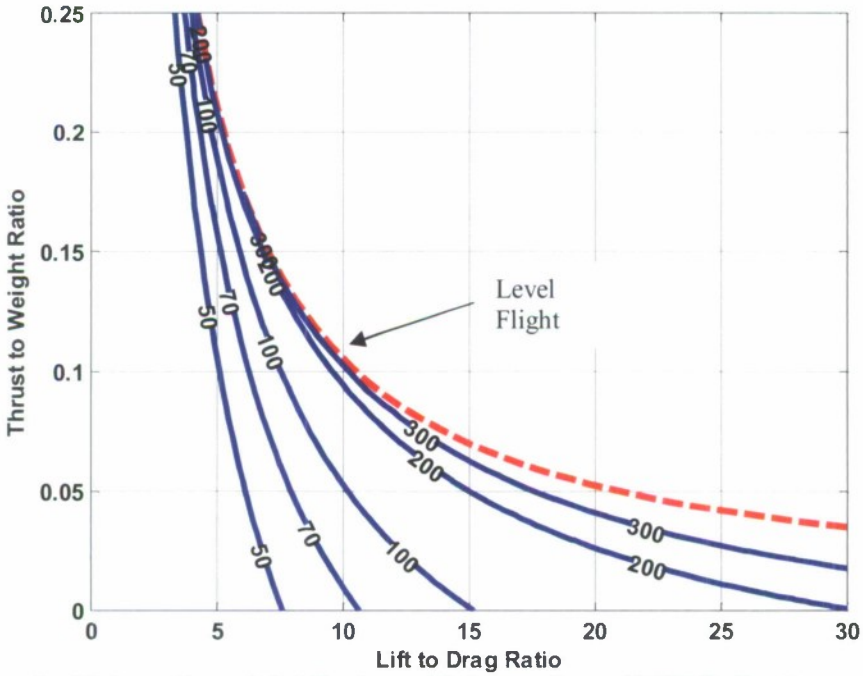


Figure 10. Numerical Integration of Jet Engine with Lapse from 40000 ft, Contours of Glide Range

According to the AGM-154 product card[3] the unpowered version of the JSOW has an effective glide range of 70 nmi from a 40,000 ft drop altitude. When using the analysis of Figure 10, for an un-powered vehicle, this corresponds to a lift to drag ratio of about 11. The weight reported for this vehicle is 1,050 lbf (depending on variant)[3]. Recently reported in a Raytheon press release[4] was a powered version of the JSOW, the JSOW-ER, with a 300 nmi range using a 150 lbf thrust turbojet engine. If we assume that the powered version is also dropped from 40,000 ft, and using the same *L/D* as determined for the un-powered version, the weight of the powered JSOW-ER was determined to be 1,575 lbf. This is very close to the weight reported for the un-powered version, and shows that the JSOW-ER is most likely an "underpowered" vehicle, utilizing the technology described in this paper.

As an estimation technique, the simpler equation of motion analysis was calibrated to match the results of the numerical integration analysis. Figure 11 shows the same curves for the simpler analysis, including the same thrust lapse model presented above, but the density ratio is estimated at an altitude of 18,000 ft. This was considered a good estimate for our purposes of modeling the flight regime because most of the effects of the thrust lapse are seen at lower altitudes for longer ranges.

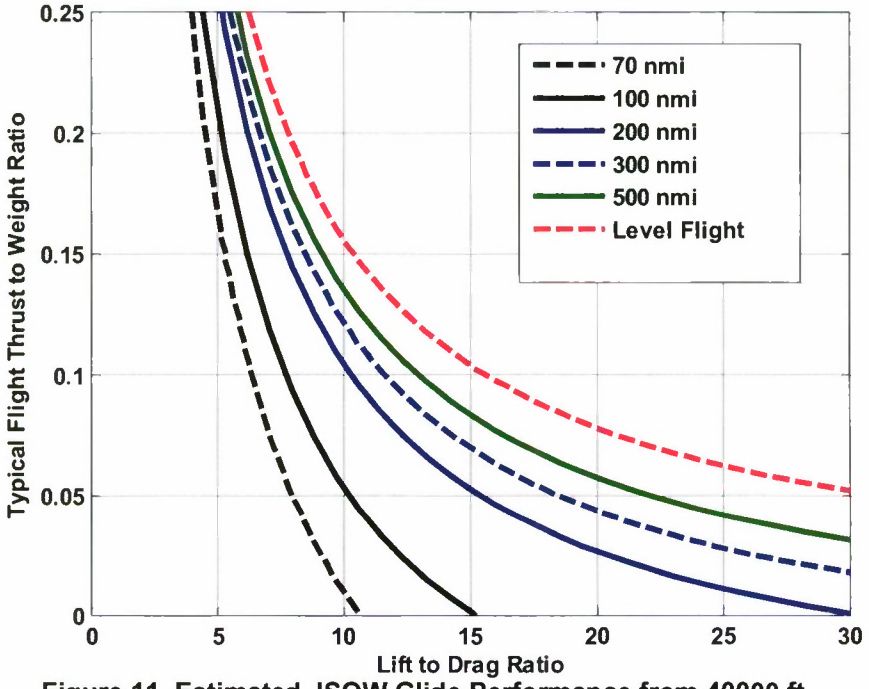


Figure 11. Estimated JSOW Glide Performance from 40000 ft

VI. Conclusion

An underpowered aircraft concept was studied to determine the feasibility and cost effectiveness of such a technology, as well as possible applications. It was determined, through an equations of motion, energy method, and numerical integration analysis, that an underpowered aircraft can provide significant range extension for a gliding flight vehicle. The methods also include a built in thrust lapse to correctly model an underpowered vehicle's performance with altitude. Also, the methods developed in this paper were compared with a current gliding vehicle, the AGM-154 JSOW and JSOW-ER. It was determined, from aircraft metrics provided by Raytheon and our analysis, that the JSOW-ER is most likely an underpowered aircraft and is representative of a possible mission application for the technology. Other applications include payload delivery, glide munitions, UAVs, and others. Overall, the underpowered aircraft technology represents a unique flight regime, giving great benefit to the user for a low overall cost.

References

[1]Anderson, John D., *Aircraft Performance and Design*, WCB/McGraw-Hill, 1998

[2]Mattingly, J. D., Heiser, W. H., and Daley, D. H., *Aircraft Engine Design*, AIAA Education Series, AIAA, New York, 1987, Chap. 2.

[3]Raytheon Company. JSOW: Family of Precision Strike Weapons. Brochure. Tucson: Raythcon Company, Missile Systems, 2008.

[4]Raytheon. Press release. Raytheon Demonstrates Engine for Powered Joint Standoff Weapon Extended Rangc. 20 Feb. 2007. Aug. 2008 <http://www.globalsecurity.org/military/library/news/2007/02/mil-070220-raytheon02.htm>.

AEROMECH ENGINEERING, INC.

CAL POLY

Aerospace Engineering

Proposal White Paper: Low Cost Rapid Development (LCRD) of a Cargo UAV for U.S. Marine Seabased Distributed Operations

ONR BAA 08-012
Thrust Area 4 – Logistics
Topic 1 – Seabased Air Cargo Delivery System for Small Combat Units

Principal Point of Contact
Patrick Stewart, Senior Aerospace Engineer
AeroMech Engineering, Inc.
888 Ricardo Court
San Luis Obispo, CA 93401
805.503.4290
805.541.0797 fax
pstewart@aeromechengineering.com

4/3/08
Thomas L. Akers, CEO

This proposal includes data that shall not be disclosed outside the Government and shall not be duplicated, used, or disclosed—in whole or in part—for any purpose other than to evaluate this proposal. However, if a contract is awarded to this offeror as a result of—or in connection with—the submission of these data, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the Government's right to use information contained in these data if they are obtained from another source without restriction.

© 2008. AeroMech Engineering, Inc. Unpublished work. All rights reserved.

1 Introduction

The United States Marines' plan for Sebasing and Distributed Operations (DO) requires the capability to deliver supplies from a floating warehouse directly to small scattered groups of Marines in the field. UAV's present an ideal solution to missions which are Dull, Dirty, or Dangerous; low-cost and highly autonomous systems add Distributed to that list. Supplying Marines is Dull, Dangerous, and Distributed. In two years, we will conceive, design, demonstrate, and deliver an unmanned aircraft system as a solution to this challenge.

AeroMech Engineering (AME) specializes in the design and development of prototype and production UAVs (Figure 1). AME has produced more than 750 fully autonomous UAVs ranging from 4-lb battery powered aircraft to 150-lb jet powered aircraft. AME's small UAVs are currently operated in theater and have accumulated approximately 70,000 total flight hours to date. From innovative avionics solutions to state of the art mission planning software and ground control station hardware, AME has a high level of capability in all areas of UAV systems.



Figure 1: AeroMech Engineering's San Luis Obispo Facility.

The Cal Poly, San Luis Obispo Aerospace Engineering department (Cal Poly) is a highly ranked program with a distinct design focus and a strong connection to the aerospace industry. This collaboration will augment AME's proven expertise with Cal Poly's motivated students by making the cargo UAV development program the focus of their senior design experience. The Cal Poly motto is "Learn by Doing"; in this project, the students will do exactly that.

AME maintains a unique working relationship with Cal Poly. AME provides both engineering and fabrication assistance to various school programs including the AIAA Design/Build/Fly team, the ASCE Concrete Canoe team, and the SAE Formula Car team. As a high tech company pursuing many of the diverse challenges of unmanned systems, AME's relationship with an applied research institution like Cal Poly is a natural one. Many senior projects and graduate thesis topics have provided these partners with a strong track record of collaboration.

2 System Requirements

The Marine Corps is working towards a concept of Distributed Operations. One major challenge to DO is how to keep many small, scattered, and highly mobile groups of Marines adequately supplied so they may complete their mission.

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

This mission implies a set of requirements which must be understood at the start of the design process. A seabased cargo delivery UAV will be operated by a ship-board crew of non-experts. The system must have minimal logistical impact and require no modifications to the ship. The aircraft and system must be compliant with the Department of Defense's Single Fuel Concept (SFC).

As a member of a four-man team with a vehicle, an individual mounted Marine requires about 92 pounds of supplies each day. Summarized in

Table 1, this includes four gallons of water, 30% of a combat load of ammunition, and nearly four pounds of batteries.[1] Of course, a Marine platoon's combat needs are unpredictable and it will be essential to tailor each delivery to the given situation. Five hundred pounds of supplies could be packed in a cylinder about 1.7 feet in diameter and 6.9 feet long.

Table 1: One Mounted Marine's Day of Supply. [1]

	Weight (lbf)	Volume (ft^3)
Food & Water	39.0	0.9
Fuel	23.1	0.5
Ammunition	22.2	0.5
Misc. Supplies	7.5	0.2
Total	91.8	2.1

A delivery point 100 miles inland leads to a 200 mile mission radius. Without burdening them with direct control of the aircraft, the Marines on the ground must have the capability to set and change the payload delivery point while the vehicle is on its way. This mission demands a high degree of autonomy with changing modes, operators, and objectives.

Although every measure will be taken to prevent aircraft and payload loss, operational robustness is a distinct advantage to a distributed delivery network utilizing low cost autonomous components. The implicit trade between system architecture complexity and accident rate is effectively depicted in Figure 2.

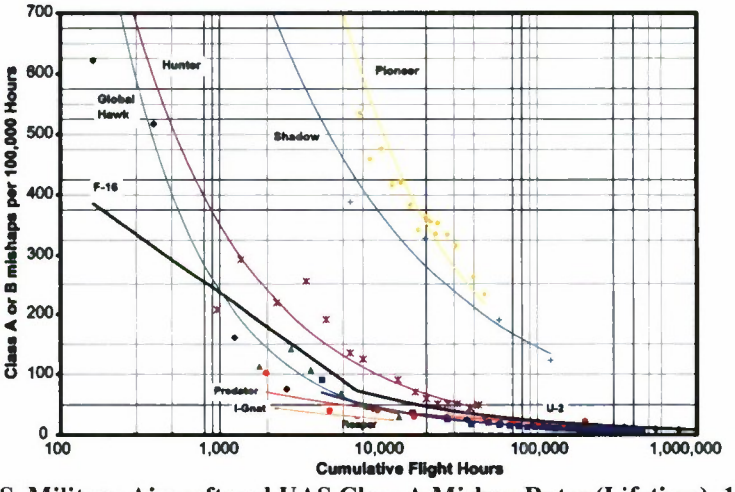


Figure 2: U.S. Military Aircraft and UAS Class A Mishap Rates (Lifetime), 1986–2006.[5]

The limitations of current mission planning and control software have serious consequences. Repeated studies have shown that human factors are a leading cause of UAV accidents across many aircraft types[2]; estimates of the fraction of accidents caused or influenced by human factors range between 32%[3] and 100%.[4] A USAF report on Predator mishaps implicates the Human Control Interface in 89% of accidents, where its issues are cited as the causal factor in 44% of those accidents.[4] These facts strongly reinforce the need for a fully autonomous system with advanced mission planning and control software like AME's SharkFin product.

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

3 System Concepts and Trades

The unique collaboration proposed between AeroMech Engineering and Cal Poly will allow us to quickly develop a prototype vehicle while thoroughly exploring a vast system architecture and operational concept space. In addition to the traditional payload/speed/range/cost trades, some unique concepts and trades must be considered in the development of this system.

A first concept distinction can be made between systems which are carried aloft by another aircraft before launch – these systems capitalize on added potential energy to accomplish the mission. Dropped concepts are distinct from systems which operate and fly independently – these systems carry all of the energy required for the mission on board.

Dropped system concepts are attractive because they may improve the UAV system cost by reducing the demands on the propulsion system and energy storage. Examples include a helicopter or CV-22 dropped guided parachute; the limited glide range of parachutes would demand a drop over the combat zone. This limitation could be overcome by a high lift-to-drag vehicle or a slightly powered aircraft dropped from a CV-22 at significant altitude. An FA-18 E/F could be used to deliver a 'cargo bomb' resembling a 2000 pound bomb or extended range drop tank. Unfortunately, the 'cargo bomb' must be transferred from the transport ship to the aircraft carrier for deployment; mass properties control for stores separation certification also presents a major obstacle to this concept. All boosted systems are limited by the operational requirements of the host vehicle. These manned systems are expensive to operate and are under very high demand during a conflict.

Independent systems eliminate the operational demands on manned aircraft not only in their use, but also in training, certification, etc. A rail-launched UAV is likely the best approach to this challenge; it presents a simple architecture concept with minimal logistical impact.

Another major operational distinction to be considered is mode of recovery for the system. The simplest system will deliver its payload by landing near the Marine unit; such a system could be disposable or recovered at a later date – it may be necessary to disable critical systems. Alternately, terminal payload delivery could be accomplished by parachute allowing the UAV to be recovered at another location. The UAV could return to the ship for a water landing or net capture. Or, it may return to a pre-determined depot location to be re-fitted for another use.

4 References

1. M. Bain, "Supporting a Marine Corps Distributed Operations Platoon: A Quantitative Analysis," MS Thesis, Naval Postgraduate School, Monterey CA, September 2005.
2. K. Williams, "Human Factors Implications of Unmanned Aircraft Accidents: Flight-Control Problems", , Federal Aviation Administration, Office of Aerospace Medicine, DOT/FAA/AM-06/8, 2006.
3. P. Leduc, C. Rash, and S. Manning, "Human Factors in UAV Accidents", Special Operations Technology, 3 (8), Dec 14, 2005.
4. S. Sharma and D. Chakravarti, "UAV Operations: An Analysis of Incidents and Accidents with Human Factors and Crew Resource Management Perspective," Indian Journal of Aerospace Medicine, Indian Society of Aerospace Medicine, 49 (1), 2005, pp.29-36.
5. A. Tvaryanas, W. Thompson, and S. Constable, "U.S. Military Unmanned Aerial Vehicle Mishaps: Assessment of the Role of Human Factors Using HFACS", USAF 311th Human Systems Wing, HSW-PE-BR-TR-2005-0001, 2005.
6. "Unmanned Aircraft Systems Roadmap 2007-2032," Office of the Secretary of Defense, December 2007.

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

5 Programmatic Information

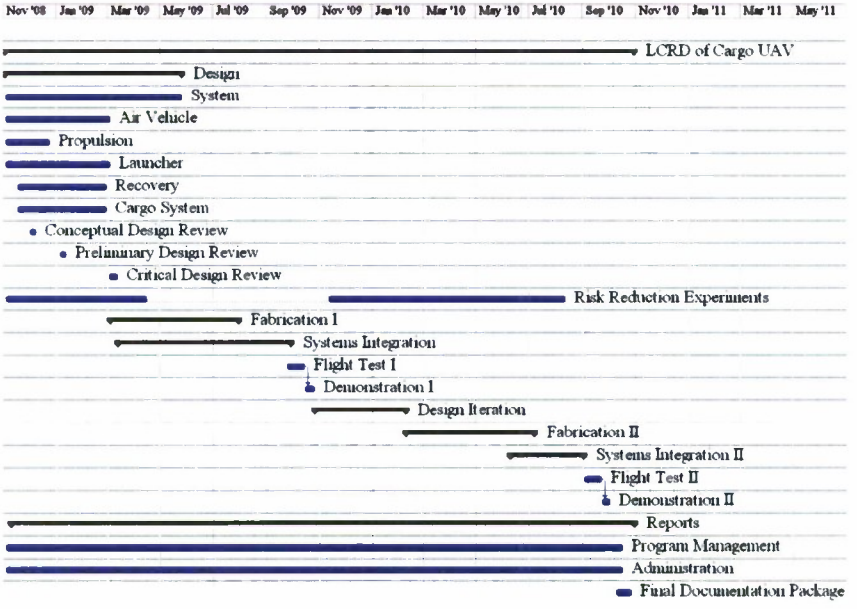
The program will be organized into two similar phases – both spanning from design to demonstration of an operational system. The second phase will reiterate the design, incorporating lessons learned and customer feedback from the demonstration in Phase I. Two air vehicles will be built in both phases.

Table 2: Major Program Milestones

	Milestone	Date
1	Conceptual Design Review	12/2008
2	Preliminary Design Review	01/2009
3	Critical Design Review	03/2009
4	System Fabrication I	07/2009
5	System Integration I	09/2009
6	Flight Test I	12/2009
7	Demonstration I	01/2010
8	Design Optimization/Re-design	04/2010
9	System Fabrication I	06/2010
10	System Integration II	08/2010
11	Flight Test II	10/2010
12	Demonstration II	10/2010
13	Final Documentation Package	11/2010

AeroMech Engineering and Cal Poly will collaborate in the design and development of this system. The Senior Design students, faculty, and staff will participate in a development program which parallels AME's effort. Cal Poly's design students will receive regular design reviews from the aerospace industry culminating in an annual review for ONR. This unique collaboration provides a cost effective means to explore the diverse design concept space. Cal Poly's lessons learned will feed back to AME's engineering process. Maintaining concept development and risk reduction experiments late into the development process this will serve to reduce design risk throughout the project.

Table 3: Program Schedule



Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

6 This Deliverables

Proposed deliverables for this development effort include design reviews and their associated documents, status reports, system flight demonstrations, and one fully operational system.

- Technical Progress Reports – weekly updates, detailing technical progress
- Status Reports – monthly program status, including financials
- Risk Reduction Experiment Reports – description and significant findings
- Design Reports – detailed reports accompanying each design review
- Design Reviews – face to face presentations of current design status
    - Conceptual Design Review
    - Preliminary Design Review
    - Critical Design Review
    - Final Design Review
- Demonstration I – preliminary system demonstration
- Demonstration II – final system demonstration
- Demonstration Reports – description and significant results
- Fully Operational System – delivered after Demonstration II
    - Air Vehicle – one to be delivered
        - Cargo Delivery System
    - Ground Support Equipment – one to be delivered
        - Launcher
        - Cargo Handling System
        - Ground Control Station (GCS) Equipment
        - Single Seat of GCS Software
        - Recovery System
- Final Documentation Package – summary documentation
    - Final Design Documentation
    - System Operational Manual
    - Interface Control Document
    - Interface Design Document

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

7 Cost

The estimated cost for this development effort is three million dollars over a two year effort. These costs include those of both Cal Poly SLO and AeroMech Engineering, Inc.

	Task	Cost
1	Design – Labor	$370k
2	Risk Reduction Experiments – Labor	$250k
3	Risk Reduction Experiments - Materials	$100k
4	System Fabrication 1 - Labor	$400k
5	System Fabrication I - Materials	$284k
6	System Integration I - Labor	$115k
7	System Integration I - Materials	$13k
8	Flight Test I - Labor	$100k
9	Flight Test I - Materials	$11.5k
10	Demonstration I – Labor	$35k
11	Demonstration I - Materials	$15k
12	Design Optimization/Re-design - Labor	$150k
13	System Fabrication II – Labor	$400k
14	System Fabrication II – Materials	$202k
15	System Integration II - Labor	$115k
16	System Integration II – Materials	$9k
17	Flight Test II - Labor	$100k
18	Flight Test II - Materials	$11.5k
19	Demonstration II - Labor	$35k
20	Demonstration II - Materials	$15k
21	Status/Design Reports - Labor	$110k
22	Final Documentation - Labor	$35k
23	Program Management - Labor	$112k
24	Administration - Labor	$15k
25	Total Labor	$2.34M
26	Total Materials	$661k
27	Combined Total	$3.0 M

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

8 Resumes

The following resumes constitute the key engineering personnel, but several other engineers with extensive collective UAV design, fabrication, integration, and flight testing experience will also be employed in the development effort.

- Thomas Akers – Chief Executive Officer, AeroMech Engineering, Inc.
- Joseph M. Wurts – Chief Aerodynamicist, AeroMech Engineering, Inc.
- Rob McDonald, PhD – Aerospace Design Professor, Cal Poly SLO
- Michael Helke – Director of Business Development, AeroMech Engineering, Inc.
- Patrick Stewart – Senior Aerospace Engineer, AeroMech Engineering, Inc. (PI)

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal.

Joseph M. Wurts

AeroMech Engineering, Inc.

Work Experience

Consulting work:

- Provided technical oversight for an Edwards AFB Flight Test program, Shear Wind Observed Optimized Path Investigation (SENIOR ShWOOPIN), recommended optimal flight trajectory methodologies along with instrumentation and data collection methods.
- Wind tunnel data reduction of an existing design, analysis of the data validity, recommended configuration upgrade with new control surface effectors to meet requirements.
- Developed a series of software tools to generate CFD grids from unstructured laser scan point data, and performed CFD analysis to assess the "as-built" vehicle flight performance.
- Conducted sizing trades and configuration layout for a flight test vehicle researching body freedom flutter. Developed a structural stiffness and inertial model for a systems optimization to define a flight test article. Provided oversight on vehicle manufacture, assembly, and systems integration. Defined flight test methodology and piloted vehicle at Edwards AFB demonstrating the predicted flutter phenomena.
- Aerodynamic analysis of a cruise missile wing to determine the cause of performance variance as measured by flight and wind tunnel test. Determined causes of performance variance, recommended modifications to regain performance loss.
- Cruise missile configuration sizing and optimization, including wing airfoil design and optimization to meet performance and structural requirements.

Conceptual and preliminary configuration sizing and development:

- Configuration development lead for several programs, performing conceptual sizing trades and aerodynamic analysis, configuration packaging and integration, along with customer interface.
- Chief engineer and/or project manager on several small UAV programs, including the DARPA Micro Air Vehicle, and other classified programs. Managed design teams in both conceptual and preliminary design phases, providing technical guidance and managing the design teams cost and schedule.
- Lead designer on the Lockheed Martin Tier II+ high altitude long endurance unmanned air vehicle concept. Developed the configuration from trade studies to meet customer requirements. Responsible for payload, structural, subsystems and propulsion integration. Authored configuration section of the proposal.
- Lead designer on the JASSM cruise missile. Responsible for the overall configuration trade studies and vehicle development, as well as integration of the payload, propulsion and subsystems.
- Developed and refined configurations on the A-X program for the Navy. Integrated configurations with a high priority on carrier suitability.
- Lead designer on an ASTOVL vehicle, packaging challenging propulsion systems.

Core Skills

- Configuration development - working with and understanding each of the multiple disciplines necessary for configuration development.
- Aerodynamic analysis and system optimization

Education

California Polytechnic State University, San Luis Obispo CA:

Bachelor of Science Degree in Aeronautical Engineering, December 1985.

Masters of Science Degree in Engineering, June 1986.

Lockheed Technical Institute:

Aircraft Dynamics & Aerodynamics, Composite Design and Analysis, Fundamentals of Aircraft Combat Survivability, Low Observability for Designers, Aircraft Conceptual Design, Fighter Tactics, Airframe Structural Design.

Lockheed Management Association:

Introduction to Supervision, Fundamentals of Leadership.

Accomplishments

Numerous (>10) design and invention patents awarded.

1997 Engineering Merit Award From the Engineers Council.

1997 Silver Skunk for Teamwork at the Skunk Works Honors Night.

1999 Bronze Skunk for Technical Excellence at the Skunk Works Honors Night.

1999 Kelly Johnson Inventor of the Year Award.

Dr. Robert A. McDonald

California Polytechnic State University
Aerospace Engineering

ramcdona@calpoly.edu
(805) 756-7242

Education

- Ph.D. in Aerospace Engineering, Georgia Institute of Technology, w/ Prof. Dimitri Mavris — 2006
- M.S. Aerospace Engineering, Georgia Institute of Technology (GT) — 2001
- B.S. Aerospace Engineering, University of Missouri-Rolla (UMR) — 1999
- Member, AIAA, SAE, ASEE

Relevant Employment

- Assistant Professor, Aerospace Engineering, Cal Poly, San Luis Obispo, CA — 2006-Present
  Faculty in charge of aircraft design, performance, and multidisciplinary design and optimization.
- Aerospace Engineer, AeroMech Engineering, San Luis Obispo, CA — Summer 2007
  Designed novel UAV configuration. Developed electric UAV performance and sizing tool.

Consulting

- Hawker Beechcraft – PD413 — 2008
  Conducted independent review of clean-sheet aircraft design.
- J. R. Gloudemans – Vehicle Structure Sketch Pad — 2006
  Conceived, proposed, and coordinated Phase I STTR project through NASA Langley.
- Vought Aircraft – Aerodynamic Design of Boeing Sonic Cruiser — 2001

Successful Research Proposals

	PI	Author	Contributor	Duration (yr)	Sponsor	Status	Amount
Unmanned Aircraft Systems; Situational Awareness for Small Tactical Unmanned Aircraft and An Autonomous Package Delivery Concept Study for the US Marines	**X**	**X**		**1**	**Cal Poly C3RP**	**Awarded 2/08**	**$53,000**
REU Site -- Summer Internships in Robotics and Autonomous Systems PI: Mohammad Noori			X	3	NSF	Awarded 1/08	$ 300,000
A Multidisciplinary Geometry Based Framework Connecting Design, Optimization, Aerodynamics, and Structures	**X**	**X**		**3**	**NASA NRA**	**Awarded 7/07**	**$ 1,000,000**
The Integrated Modeling and Verification of Hybrid Wing-Body,Low Noise ESTOL Aircraft PI: David Marshall			X	1	NASA NRA Phase I	Awarded 7/07	$ 874,000
An Advanced Open-Source Aircraft Design Platform for Personal Air Vehicle Geometry, Aerodynamics, and *Structures* PI: Dimitri Mavris		X	X	1	NASA Phase I STTR	Awarded 11/06 Completed 1/07	$ 100,000

Journal Articles

Riggins, D., Wilson, C., Roth, B., McDonald, R., "Performance Characterization of Turboshaft Engines Using Work Potential Methods," Journal of the American Helicopter Society, 50(4) 2005, 315-315

McDonald, R., Mavris, D.N., "Formulation, Realization, and Demonstration of a Process to Generate Aerodynamic Metamodels for Hypersonic Cruise Vehicle Design," SAE Transactions, 2000, 109(1), 1138-1147, also AIAA-2000-01-5559

Invited Presentations

Systems Analysis and Optimization Technical Working Group, "A Multidisciplinary Geometry Based Framework Connecting Design, Optimization, Aerodynamics, and Structures", Reno NV.

AIAA Aircraft Design Technical Committee, "Design for Range", Tucson AZ.

Raytheon Missile Systems, "Error Allocation in Complex Systems Design: A Fidelity Trade Environment", Tucson AZ.

Michael Helke

AeroMech Engineering, Inc.

Qualifications

- Over 27 years experience in aerospace industry. 20 years active duty meritorious service USAF; retired
- Extensive knowledge of systems, operations, leadership, and management ensures successful task completion

Experience

Director of Business Development and Customer Relations (24 November 2007- present)
AeroMech Engineering, San Luis Obispo, CA

Systems Engineer Staff IV (29 March 200S – 23 November 2007)
LM Aeronautics Co- Palmdale, CA (Chief Engineer/Program Manager)

- Manages contract deliverables, budgets, schedules, and performance criteria; financial approval authority
- Directs IPT leads on aspects of engineering, design tasks, and field support/flight test mission requirements
- Performs analyses at all levels to include: concept, design, fabrication, test, installation, operation, and maintenance
- Coordinates all flight test activities. Responsible for deployment logistics

Sr. Systems Engineer III (14 April 2003 - 28 March 200S)
Lockheed Martin Co, Edwards AFB CA (F/A-22 Operational Test and Evaluation Lead Engineer)

- Directs avionic systems engineers in support of F/A-22 flight test activities and maintenance efforts
- Performs on-aircraft assessments of F/A-22 developmental test and operational test avionics systems to include Radar, Electronic Warfare, Communication/Navigation/Interrogation, and Avionics Integration
- Performs installation, verification, and monitoring of the F/A-22 avionics Operational Flight Program (OFP) software and Common Integrated Processor (CIP) module firmware
- Mission Control Room lead; monitors avionics/aircraft health to provide pilot with real time information/reaction plans

Senior Technical Editor (31 July 2000 - 13 April 2003)
EDO Technical Services Operation, Edwards AFB CA

- Signature authority - reviews, edits, and rewrites B-I bomber defensive system documentation to include: defensive evaluation procedures, engineering test plans, developmental test and evaluation reports, flight reports, monthly status reports, installation and checkout procedures, engineering problem reports, and document change request notices
- Verifies correct parameters and configurations for range test emitters, test range setup, and flight test scenarios

USAF B-IB, B-2A, B-52G/H Production Superintendent (1 April 1994 - 31 August 2000)
31 Test and Evaluation Squadron, Edwards AFB CA

- Directs all maintenance and sortie generation activities on one $2.2 billion B-2A, three $500 million B-1Bs, and one $350 million B-52H flight test aircraft; supervises over 250 Air Force personnel and civilian contractors
- Manages all facilities, computer resources, and support equipment valued at over $130 million
- Establishes test schedules and test conduct parameters with military personnel, engineers, and civilian contractors

USAF B-2A Production Supervisor/Project Manager (1 April 1994 - 31 March 1997)
31 Test and Evaluation Squadron, Edwards AFB CA

- Manage, direct, control and coordinate all flight test and maintenance activities for B-2A Developmental Test and Evaluation
- Schedule and prioritize all test requirements to assess operation suitability, maintainability, and reliability of complete systems

USAF Avionics Systems Master Instructor (1 September 1985 - 31 March 1994)
34S0 TCHTS, Lowry AFB CO

- Certified with over 18,000 classroom hours as a US Air Force Master Instructor; Expert curriculum developer

USAF Bomb Navigation System Specialist (1 April 1981 - 31 August 198S)
93 Avionics Maintenance Squadron, Castle AFB CA

- Performs organizational and field level maintenance repair actions on the B-52 Offensive Avionics System, Inertial Navigation Set, Electro-Optical Viewing System, Terrain Avoidance, and Radar systems

Achievements

- 2005 Lockheed Martin AEROSTAR Award – Outstanding F-22 support
- 2005 Lockheed Martin NOVA Award – Outstanding F-22 leadership
- 1999 USAF Lt General Leo Marquez Award for Outstanding Maintenance Supervisor/Manager
- 1998 USAF Senior Non-Commissioned Officer of the Year
- 1998 USAF Lance P Sijan Leadership Award

Patrick T. Stewart

AeroMech Engineering, Inc.

Work Experience

AeroMech Engineering, Inc. **2007 - Present**

Senior Aerospace Engineer/Program Manager

- Aerodynamic modeling and analysis to develop stability and control derivatives.
- Flight Test Director for multiple small UAV development programs
- Turrets Program manager – supervised engineering development and manufacturing teams for multiple small EO/IR imager gimbal
- Test Planning, Range Safety, and Radio Frequency management for multiple flight testing deployments
- Designed RF system improvements for improved C^2 range for Low Cost Aerial Target

AAI **2005 - 2007**

Program Manager

- Program Manager for DARPA iSTAR VTOL OAV program – lead engineer for small UAV development team
- Supervised and designed airframe and control software improvements for multiple VTOL UAVs

Allied Aeropace Industries, Inc. **2003 – 2005**

Aerospace Engineer III

- Developed simulation, control laws for multiple configurations of the iSTAR VTOL OAV
- Developed on-board mission management algorithms for VTOL OAV
- Managed flight test planning and flight test operations

AeroMech Engineering, Inc. **1999 – 2003**

Jr. Engineer

- Engineer and composites fabricator on the Collier Trophy nominated Lockheed Martin FPASS / Desert Hawk UAV system. Assisted in design, development, and production of the first twelve production aircraft in only 120 days.
- Designed and developed composite airframe components
- Composites airframe fabrication
- Small UAV flight test

Core Skills

- Stability & Control of fixed and rotary wing air vehicles
- Aerodynamic and Flight test data analysis
- Program management
- Flight test planning / management

Education

California Polytechnic State University, San Luis Obispo CA:
Bachelor of Science Degree in Aerospace Engineering, *in progress*.
Masters of Science Degree in Aerospace Engineering, *in progress*.

CAL POLY

California Polytechnic State University
San Luis Obispo, CA 93407-0035

Grants Development Office
(805) 756-2982 • Fax (805) 756-5466
www.calpoly.edu/~grants • e-mail: grants@calpoly.edu

April 2, 2008

Mr. Paul Kendrick
General Manager
AeroMech Engineering INC.
888 Ricardo Court
San Luis Obispo, CA 93401

Re: Letter of Intent

Dear Mr. Kendrick,

This letter represents the agreement of the California Polytechnic State University to assist AeroMech Engineering with the research proposed to the United States Navy, Office of Naval Research, under solicitation #BAA08-012, and titled "*Low Cost Rapid Development of a Cargo UAV for US Marine Seabased Distributed Operations*". If the proposal to the U.S. Navy program is funded, Dr. Rob McDonald of our Aerospace Engineering faculty will be the Principal Investigator for our portion of the project over the anticipated 24 months of funding.

If you have any questions, please do not hesitate to contact me.

Regards,

Xenia Bixler,
Director, Grants Development

cc: GDO #08-276. R. McDonald
R. McDonald, Aerospace Engineering

CONFIDENTIAL. Limited circulation. For review only.

A Modification to the Group Delay and Simulated Annealing Technique for Characterization of Peripheral Nerve Fiber Size Distributions for Non-Deterministic Sampled Data

Robert B. Szlavik

Abstract—**The ability to determine the characteristics of peripheral nerve fiber size distributions would provide additional information to clinicians for the diagnosis of specific pathologies of the peripheral nervous system. Investigation of these conditions, using electro-diagnostic techniques, is advantageous in the sense that such techniques tend to be minimally invasive yet provide valuable diagnostic information. One of the principal electro-diagnostic tools available to the clinician is the nerve conduction velocity test. While the peripheral nerve conduction velocity test can provide useful information to the clinician regarding the viability of the nerve under study, it is a single parameter test that yields no detailed information about the characteristics of the functioning nerve fibers within the nerve trunk. In previous work, the efficacy of the group delay and simulated annealing approach was demonstrated in the context of a simulation study where deterministic functions were used to represent the single fiber evoked potentials. In this study we present a modification to the approach discussed previously that is applicable to non-deterministic functions of sampled data.**

I. Introduction

The nerve conduction velocity test provides clinically useful information in the diagnosis of peripheral neuropathies, such as Carpal Tunnel Syndrome [1;2]. Since nerve conduction velocity studies are essentially single parameter measurements of the gross conduction properties of the underlying nerve trunk, such studies are not suited to providing detailed information regarding the characteristics of the underlying nerve fibers that contribute to the compound evoked potential.

A more robust measurement technique would involve the ability to extract information about the viability of the underlying nerve fibers which could potentially provide useful information to the clinician. As an example, information related to the size distribution of contributing nerve fibers can be used to differentiate between different clinical conditions such as Chronic Inflammatory Demyelinating Polyneuropathy, which selectively impacts larger nerve fibers, or Early Diabetic Peripheral Neuropathy, which impacts smaller fibers [3;4].

Manuscript received April 7, 2009.

This work was supported in part by the Department of the Navy, Office of Naval Research.

R. B. Szlavik is with the Department of Biomedical & General Engineering, California Polytechnic State University, San Luis Obispo, CA 93407-0350 USA (phone: 805-756-2025; fax: 805-756-6424; e-mail: rszlavik@calpoly.edu).

There is a large body of literature devoted to describing various techniques for determining the nerve fiber conduction velocity distribution (CVD). The pioneering work of Cummins and Dorfman [5;6] describe techniques that use two compound action potentials to estimate the conduction velocity distribution using a least squares approach. Common to these studies was the assumption that fibers included in a specific velocity class have identical evoked potential waveforms.

More recently there have been several additional studies including the work of Gonzalez-Cueto, Papadopoulou and Gu [7-9]. The study presented by Tu *et. al.* focused on a regularized least squares algorithm but features many of the same assumptions associated with waveform commonality related to velocity classes that were made in earlier work [10]. This study also investigated the impact of noise on the integrity of the estimated CVD.

In several recent publications, we have demonstrated the utility of a group delay based approach for determining the size or conduction velocity distribution of fibers in a peripheral nerve trunk [11]. In a subsequent study, we demonstrated that simulated annealing could be used to optimize the group delay estimate of the nerve fiber diameter distribution in the context of the maximal compound evoked potential template [12]. Both these studies involved simulations where deterministic functions were used to simulate the single fiber evoked potentials. While this approach was useful in demonstrating the overall efficacy of the technique, the method by which the technique would have to be modified to accommodate non-deterministic sampled data was not described and is consequently the focus of this study.

II. METHOD

The principal difficulty associated with applying the group delay and simulated annealing approach to non-deterministic sampled data is based on the lack of a suitable time shift, delay reference or temporal marker δ_j that is inherent to the deterministic function used to describe the single fiber action potentials [13]. What is required is the determination of a suitable temporal maker for a non-deterministic function characteristic of what could be expected from the sample data of the single fiber action potential waveforms $\Gamma^{(1)}(t)$ and $\Gamma^{(2)}(t)$ as per the notation used in previously published work [11;12].

The availability of a suitable time reference for the non-deterministic case is necessary to implement the simulated annealing algorithm utilized to optimize the nerve fiber size

Preprint submitted to 31st Annual International IEEE EMBS Conference.
Received April 7, 2009.

distribution estimate against the maximal compound evoked potential template. This necessity is based on the fact that the simulated annealing approach varies the fiber diameter, and consequently the time delay of the fiber evoked potential, for a randomly chosen fiber in the nerve trunk population.

A physically relevant temporal parameter associated with the individual decomposed single fiber action potential waveforms $\Gamma_j^{(n)}(t)$ is the centroid of the absolute value of the waveform in question which may be computed as per (1).

$$\dot{\delta}_j = \frac{\int_{-\infty}^{\infty} \left| t\Gamma_j^{(n)}(t) \right| dt}{\int_{-\infty}^{\infty} \left| \Gamma_j^{(n)}(t) \right| dt} \qquad (1)$$

Since the standard physical interpretation of the centroid is in the context of masses, the functions evaluated in the centroid expression must be positive for all t.

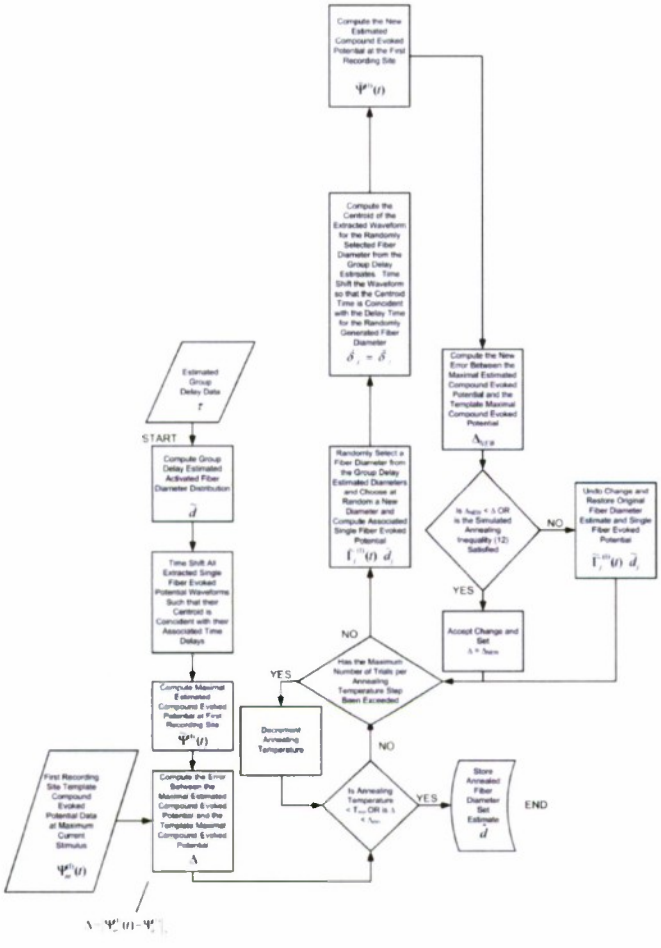


Figure 1. Flowchart of the simulated anncaling algorithm that computes an improved estimate of the nerve fiber diameter population set $\hat{\boldsymbol{d}}$ from the group delay estimated population set $\tilde{\boldsymbol{d}}$ for non-deterministic sampled data.

Initially, sampled waveforms are shifted such that their centroids are coincident with the propagation delay parameter for each fiber estimate. The centroid parameter for a given single fiber evoked potential waveform is then utilized in the simulated annealing algorithm at the point where a randomly selected fiber $\tilde{d}_j$ is chosen from the group delay estimated set $\tilde{\boldsymbol{d}}$. A new trial fiber diameter $\check{d}_j$ is generated and the original randomly selected fiber diameter is set equal to the new value such that $\tilde{d}_j = \check{d}_j$. The concomitant delay time $\check{\delta}_j$ is computed for the randomly generated fiber diameter. Utilizing the centroid of the associated sampled waveform $\tilde{\Gamma}_j^{(n)}(t)$ for the fiber $\tilde{d}_j$ as the temporal reference, the time sampled evoked potential waveform $\tilde{\Gamma}_j^{(n)}(t)$ is time shifted such that its centroid $\dot{\delta}_j$ is aligned with the time delay $\check{\delta}_j$. The rest of the simulated annealing algorithm proceeds as described in [12]. Due to the overall complexity of the algorithm, a further description of the modification is given in the flow chart of Figure 1.

III. RESULTS

A population of 100 randomly generated fibers was utilized in these studies in accordance with the formula for the randomly generated distribution outlined in previously published work [11;12]. Parameters for the specific distribution utilized are given in the caption to Figure 3. These fibers were subjected to a virtual stimulus pulse train of successively increasing amplitudes ranging from zero to a maximum of 1 mA in 500 nA steps. At each step the compound evoked potential at both virtual recording sites was computed and subsequently, the estimate of single fiber action potential waveforms were obtained at each recording site. The concomitant group delays between the two virtual recording sites were computed yielding the group delay estimated set of fiber diameters $\tilde{\boldsymbol{d}}$.

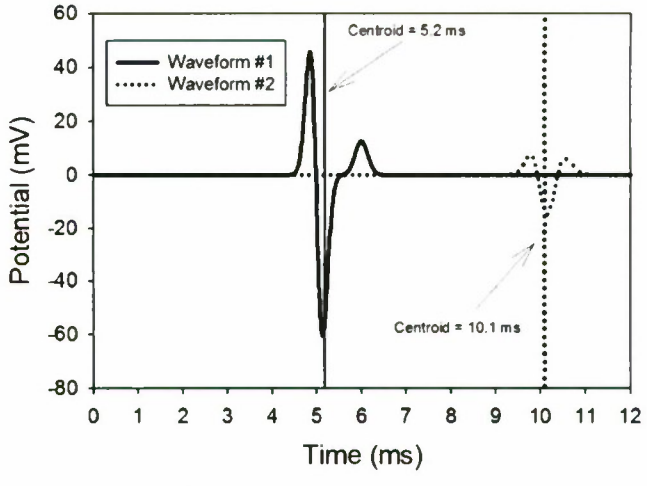


Figure 2. Plot of two typical single fiber evoked potential waveforms showing their centroids, as per (1), computed from a modified form [11;12] of the function proposed by Fleisher [13]. The relevant parameters for Waveform #1 were a = 5 μm, r = 0.5 mm, s = 1·a, σ_e = 1 S/m, α = 0.998 and a recording distance of 25 mm. For Waveform #2, the parameters were a = 5 μm, r = 0.5 mm, s = 0.5·a, σ_e = 1 S/m, α = 0.998 and a recording distance of 50 mm. In both cases $c = 5.0\times10^5$ s^{-1}. The parameter values are specified using labels consistent with previously published work [11;12].

Preprint submitted to 31st Annual International IEEE EMBS Conference.
Received April 7, 2009.

The modified simulated annealing technique with the centroid approach described above was then applied to each of the extracted evoked potential waveforms in the set $I^{(2)}(t)$. Results of the centroid computation are shown in Figure 2 for two single fiber evoked potential waveforms with parameters as indicated in the figure caption.

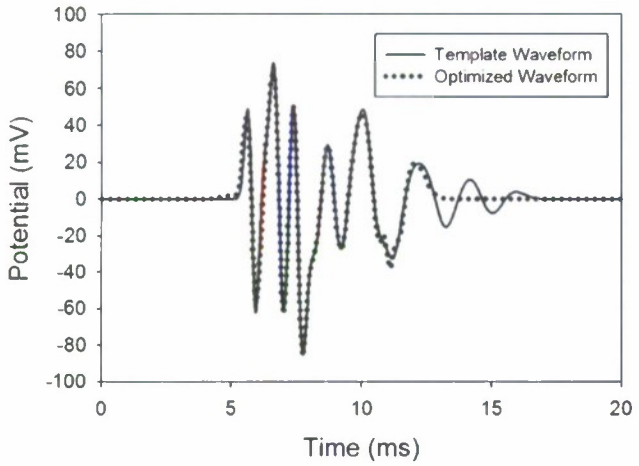


Figure 3. Graph of a maximal compound evoked potential template waveform and the maximal compound evoked potential estimate waveform obtained through group delay estimation and simulated annealing optimization. The simulated annealing algorithm utilized the centroid technique outlined in this paper with the concomitant time shifting of the estimates of the single fiber evoked potential waveforms in $I^{(2)}(t)$. The relevant simulation parameters for the fiber size distribution were $\beta_1 = 0.35$ m, $\sigma_1 = 1.699$ μm, $\mu_1 = 7.5$ μm, $\beta_2 = 0.65$ m, $\sigma_2 = 1.699$ μm and $\mu_2 = 13$ μm. For simulation of the initial single fiber evoked potentials, the relevant parameters were $r = 1$ mm, $c = 5\times10^5$ s^{-1}, $s = 1{\cdot}a$, $\sigma_e = 1$ S/m, $\alpha = 0.998$, $\zeta = 10$ mA and $\eta = 3.5\times10^5$ m^{-1}. The simulated annealing algorithm was implemented with the parameters $T = 10$, $T_{MIN} = 1.0\times10^{-5}$, $\Delta_{MIN} = 1.0\times10^{-4}$ and an annealing step factor of 0.9. The maximum number of trials for each annealing temperature step was limited to 1000.

The graph of Figure 3 shows the results of an optimization of a set of estimated singled fiber evoked potentials waveforms at the second recording site $I^{(2)}(t)$ against the template maximal compound evoked potential waveform. Relevant parameters for this simulation are given in the figure caption.

The chi-square test comparing the actual distribution to the group delay estimated distribution yields $Q(\chi^2|x) = 0.4697$ while the comparison of the actual to the annealed distribution yields $Q(\chi^2|x) = 0.7000$.

IV. DISCUSSION

It has previously been demonstrated by the author that the simulated annealing approach can make significant improvements in fiber diameter distribution estimates made using the group delay technique. Based on the modification proposed in this paper, the technique is also beneficial when applied to the extracted estimates of the single fiber evoked potential waveforms where there is no closed form functional description.

A quantitative description of the relevant distributions using the chi-square test is also revealing as to the efficacy of the modified technique. The chi-square test indicates that there is a demonstrable improvement in the fidelity of the fiber diameter distribution obtained after annealing over that of the fidelity of the distribution that results from just the group delay estimation process when compared with the actual distribution.

In the author's experience, it is advantageous to apply the simulated annealing algorithm to data sets at the more distal recording site from the stimulus site. Optimization results at the more distal site yield an improved estimate of the optimized fiber diameter distribution when compared to the actual distribution than performing the optimization with the more proximal recording site data set.

REFERENCES

[1] P. A. Parker and P. Kelly, "Nerve conduction velocity measurement techniques," *Journal of Clinical Engineering*, vol. 7, no. 2, pp. 153-158, Apr.1982.

[2] I. Atroshi, C. Gummesson, R. Johnsson, and E. Omstein, "Diagnostic properties of nerve conduction tests in population-based carpal tunnel syndrome," *BMC Musculoskeletal Disorders*, vol. 4, no. 9 May2003.

[3] Y. Harati, "Diabetic peripheral neuropathies," *Annals of Internal Medicine*, vol. 107, pp. 546-559, 1987.

[4] L. J. Dorfman, K. L. Cummins, G. M. Reaven, J. Ceranski, M. S. Greenfield, and L. Doberne, "Studies of diabetic polyneuropathy using conduction velocity distribution (DCV) analysis," *Neurology*, vol. 33, no. 6, pp. 773-779, June1983.

[5] K. L. Cummins, L. J. Dorfman, and D. H. Perkel, "Nerve fiber conduction-velocity distributions, II Estimation based on two compound action potentials," *Electro-Encephal. Clin. Neurophysiol*, vol. 46, no. 6, pp. 647-658, 1979.

[6] L. J. Dorfman, "The distribution of conduction velocities (DCV) in peripheral nerves: a review," *Muscle & Nerve*, vol. 7, no. 1, pp. 2-11, 1984.

[7] J. A. Gonzalez-Cucto and P. A. Parker, "Deconvolution estimation of nerve conduction velocity distribution," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 140-151, Feb.2002.

[8] F. A. Papadopoulou and S. M. Panas, "Estimation of the nerve conduction velocity distribution by peeling sampled compound action potentials," *IEEE Transactions on Magnetics*, vol. 35, no. 3, pp. 1801-1804, May1999.

[9] D. Gu, R. E. Gander, and E. C. Crichlow, "Determination of nerve conduction velocity distribution from sampled compound action potential signals," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 8, pp. 829-838, Aug.1996.

[10] Y. X. Tu, A. Wemsdorfer, S. Honda, and Y. Tomita, "Estimation of Conduction velocity distribution by regularized-least-squares method," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 11, pp. 1102-1106, Nov.1997.

[11] R. B. Szlavik, "A novel method for characterization of peripheral nerve fiber size distributions by group delay," *IEEE Trans. Biomed. Eng*, vol. 55, no. 12, pp. 2836-2840, Dec.2008.

[12] R. B. Szlavik and G. E. Turner, "A novel method for characterization of peripheral nerve fiber size distributions by group delay measurements and simulated annealing optimization," in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2008, pp. 5008-5014.

[13] S. M. Fleisher, M. Studer, and G. S. Moschytz, "Mathematical-Model of the Single-Fiber Action-Potential," *Medical & Biological Engineering & Computing*, vol. 22, no. 5, pp. 433-439, 1984.

Preprint submitted to 31st Annual International IEEE EMBS Conference.
Received April 7, 2009.

30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

A Novel Method for Characterization of Peripheral Nerve Fiber Size Distributions by Group Delay Measurements and Simulated Annealing Optimization

Robert B. Szlavik and Galen E. Turner III

Abstract—**The ability to determine the characteristics of peripheral nerve fiber size distributions would provide additional information to clinicians for the diagnosis of specific pathologies of the peripheral nervous system. Investigation of these conditions, using electro-diagnostic techniques, is advantageous in the sense that such techniques tend to be minimally invasive yet provide valuable diagnostic information. One of the principal electro-diagnostic tools available to the clinician is the nerve conduction velocity test. While the peripheral nerve conduction velocity test can provide useful information to the clinician regarding the viability of the nerve under study, it is a single parameter test that yields no detailed information about the characteristics of the functioning nerve fibers within the nerve trunk. In this study we present a technique based on a decomposition of the maximal compound evoked potential and subsequent determination of the group delay of the contributing nerve fibers. The fiber group delay is then utilized as an initial estimation of the nerve fiber size distribution and the concomitant temporal propagation delays of the associated single fiber evoked potentials to a reference electrode. Subsequently the estimated single fiber evoked potentials are optimized against the template maximal compound evoked potential using a simulated annealing algorithm. Simulation studies, based on deterministic single fiber action potential functions, are used to demonstrate the robustness of the proposed technique in the presence of noise associated with variations in distance between the nerve fibers and the recording electrodes between the two recording sites.**

I. Introduction

The nerve conduction velocity test provides clinically useful information in the diagnosis of peripheral neuropathies, such as Carpal Tunnel Syndrome [1;2]. Since nerve conduction velocity studies are essentially single parameter measurements of the gross conduction properties of the underlying nerve trunk, such studies are not suited to providing detailed information regarding the characteristics of the underlying nerve fibers that contribute to the compound evoked potential.

A more robust measurement technique would involve the ability to extract information about the viability of the underlying nerve fibers which could potentially provide useful information to the clinician. As an example, information related to the size distribution of contributing nerve fibers can be used to differentiate between different clinical conditions such as Chronic Inflammatory Demyelinating Polyneuropathy, which selectively impacts larger nerve fibers, or Early Diabetic Peripheral Neuropathy, which impacts smaller fibers [3;4].

There is a large body of literature devoted to describing various techniques for determining the nerve fiber conduction velocity distribution (CVD). The pioneering work of Cummins and Dorfman [5;6] describe techniques that use two compound action potentials to estimate the conduction velocity distribution using a least squares approach. Common to these studies was the assumption that fibers included in a specific velocity class have identical evoked potential waveforms.

More recently there have been several additional studies including the work of Gonzalez-Cueto, Papadopoulou and Gu [7-9]. The study presented by Tu *et. al.* focused on a regularized least squares algorithm but features many of the same assumptions associated with waveform commonality related to velocity classes that were made in earlier work [10]. This study also investigated the impact of noise on the integrity of the estimated CVD.

In this paper, we discuss a novel technique, presented previously by the author [11], for estimating the size distribution of contributing nerve fibers which is linearly related to the CVD. The technique is based on an estimation of the group delay between two sets of recording electrodes associated with the individual fibers that contribute to a maximal compound evoked potential. The group delay information is then used to estimate the diameters of the activated fibers as well as the propagation delays of individual single fiber evoked potentials to a reference electrode. This process allows for reconstruction of an estimated maximal compound evoked potential, from the individual single fiber evoked potentials, at the first recording site. The previously presented group delay technique is expanded upon in this paper by introduction of a simulated annealing optimization algorithm. This algorithm is used to vary the diameter and concomitant propagation delays associated with the estimated single fiber evoked potential waveforms to search for an improved fit with the template maximal compound evoked potential from the same recording site.

The basic methodology behind the technique is presented by utilizing a closed form mathematical model of a single

Manuscript received April 7, 2008.

This work was supported in part by the Department of the Navy, Office of Naval Research.

R. B. Szlavik is with the Department of Biomedical & General Engineering, California Polytechnic State University, San Luis Obispo, CA 93407-0350 USA (phone: 805-756-2025; fax: 805-756-6424; e-mail: rszlavik@calpoly.edu).

G. E. Turner III is with the College of Engineering and Science, Louisiana Tech University, Ruston, LA 71272-0046 USA (phone: 318-257-2582; e-mail: gturner@latech.edu).

fiber evoked potential waveform that allows us to demonstrate the robustness of the technique under noisy conditions introduced by random variations in the perpendicular distance between the fiber and recording electrode [12].

II. METHOD

The simulation for determination of the group delay is premised on the physical setup shown in Figure 1, where a stimulator is used to excite a subcutaneous nerve trunk consisting of a group of electrically independent nerve fibers. The propagating compound evoked potential is detected at two recording sites. Using a series of successively increasing current stimulus pulses, the successively recorded compound evoked potentials can be decomposed into their constituent single fiber action potentials in a manner analogous to the protocol used in the McComas *et al.* motor unit number estimation technique [13].

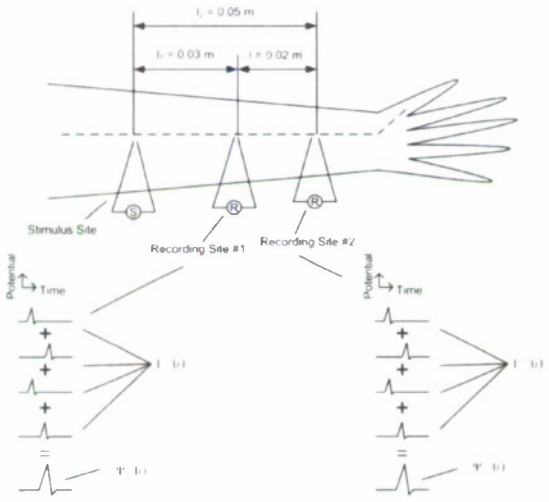


Figure 1. Conceptual physical configuration of the proposed method. The diagram shows the stimulus and recording sites as well as the relationship between the individual single fiber evoked potentials sets $\boldsymbol{\Gamma^{(1)}(t)}$ and $\boldsymbol{\Gamma^{(2)}(t)}$ to the maximal compound evoked potentials $\Psi^{(1)}(t)$ and $\Psi^{(2)}(t)$. The figure is not drawn to scale and is indicative of stimulation and recording sites at convenient locations along the median nerve. From an experimental perspective, implementation could involve stimulation of the median nerve at the anterior cubital fossa with recording sites placed more distally at anatomically convenient locations such as the wrist. This type of placement would result in distances that are larger than those indicated on the figure.

An empirically determined nerve fiber diameter distribution [14] was used to generate a random population of one hundred nerve fiber diameters using a technique described by Szlavik and de Bruin [15]. The distribution in (1) was used to generate the fiber diameter population.

$$p_d(d_k) = \sum_{h=1}^{4} \frac{\beta_h}{\sigma_h\sqrt{2\pi}} exp\left[-\frac{(d_k-\mu_h)^2}{2\sigma_h^2}\right] \qquad (1)$$

The parameters shown in Table 1 were used in the distribution of (1).

TABLE I
PARAMETER VALUES USED IN THE FIBER DIAMETER DISTRIBUTION

Symbol	Quantity	Value
β_1	complete distribution 1st mode scaling const.	0.05 (m)
σ_1	complete distribution 1st mode std. dev.	0.1274 (μm)
μ_1	complete distribution 1st mode mean	0.5 (μm)
β_2	complete distribution 2nd mode scaling const	0.25 (m)
σ_2	complete distribution 2nd mode std. dev.	0.8493 (μm)
μ_2	complete distribution 2nd mode mean	3 (μm)
β_3	complete distribution 3rd mode scaling const.	0.3 (m)
σ_3	complete distribution 3rd mode std. dev.	1.699 (μm)
μ_3	complete distribution 3rd mode mean	7.5 (μm)
β_4	complete distribution 4th mode scaling const.	0.4 (m)
σ_4	complete distribution 4th mode std. dev.	1.699 (μm)
μ_4	complete distribution 4th mode mean	13 (μm)

Fiber diameters less than 5 μm were excluded from the population yielding a sub-population of $m = 60$ fibers. The maximum nerve fiber diameter in the distribution was 17.1 μm. This randomly generated fiber diameter distribution formed the template distribution population $\boldsymbol{d}$.

The population of nerve fibers was subjected to a series of virtual stimulus pulses of successively increasing current amplitude where Ω_i is the amplitude of the stimulus current pulse at each increment i. An activation function $\xi(d)$ is used to determine whether a given stimulus current amplitude was sufficient to excite each fiber with diameter d as per (2) where $\zeta = 10\ mA$ and $\eta = 3.5 \times 10^5\ m^{-1}$.

$$\xi(d) = \zeta \exp[-\eta d] \qquad (2)$$

For each recording site $n = 1, 2$, the compound evoked potential $\Psi_i^{(n)}(t)$ is computed for each increment i of the stimulus current amplitude as per (3).

$$\Psi_i^{(n)}(t) = \sum_{k=1}^{m} u[\Omega_i - \xi(d_k)] G\left[v_k \cdot \left(t - \delta_k^{(n)}\right), \bar{r}\right] \qquad (3)$$

In (3), the single fiber action potential waveform $G\left[v_k \cdot \left(t - \delta_k^{(n)}\right), \bar{r}\right]$ contributes to the compound evoked potential if the argument of the step function u is positive where t is the time in seconds, v_k is the conduction velocity of the k^{th} fiber, $\delta_k^{(n)}$ is the propagation delay, in seconds, of the single fiber action potential from the stimulus site to the n^{th} recording site and $\bar{r}$ is the perpendicular depth between the recording site and the center of the k^{th} fiber.

The function G is the normalized model of the single fiber action potential proposed by Fleisher *et. al.* where the function has been normalized to the current through the second pole such that $G = g/I$ as per (4). All other parameters are as described in Fleisher [12] and were assigned values $a_k = d_k/2$, $s_k = 5 \cdot a_k$, $\bar{r} = 1\ mm$, $v_k = c \cdot d_k$, $\alpha = 0.75$, $\sigma_e = 1.0\ S/m$ and $D_k = (a_k + s_k)/(\bar{r} + s_k)$.

$$G\left[v_k \cdot \left(t-\delta_k^{(n)}\right), \bar{r}\right] = \frac{D_k^2}{4\pi\sigma_e a_k}\left[\alpha \exp\left\{-\left(\frac{D_k}{4}\right)^2\left(\frac{v_k \cdot \left(t-\delta_k^{(n)}\right)+s}{a_k}\right)^2\right\} - \exp\left\{-\left(\frac{D_k}{4}\right)^2\left(\frac{v_k \cdot \left(t-\delta_k^{(n)}\right)-s}{a_k}\right)^2\right\} + (1-\alpha)\exp\left\{-\left(\frac{D_k}{4}\right)^2\left(\frac{v_k \cdot \left(t-\delta_k^{(n)}\right)-u}{a_k}\right)^2\right\}\right] \quad (4)$$

After the compound evoked potentials are computed for each virtual current step Ω_i, the series of compound evoked potentials at each recording site $\boldsymbol{\Psi^{(1)}(t)}$ and $\boldsymbol{\Psi^{(2)}(t)}$ are decomposed into a series of waveforms that nominally consist of the contributing single fiber action potentials at each simulated current step $\boldsymbol{\Gamma^{(1)}(t)}$ and $\boldsymbol{\Gamma^{(2)}(t)}$ as per (5).

$$\Gamma_{i-1}^{(n)}(t) = \Psi_{i}^{(n)}(t) - \Psi_{i-1}^{(n)}(t) \qquad \text{for } 2 \le i \le q+1 \quad (5)$$

If the current steps are small enough, then the waveforms $\boldsymbol{\Gamma^{(n)}(t)}$ will consist of individual contributing single fiber action potentials or no waveform where a stimulus current increment does not result in an additional recruited fiber. However, a perfect decomposition will not always be achievable due to the finite discretization of the stimulus current steps. Some of the q non-zero waveforms in the set $\boldsymbol{\Gamma^{(n)}(t)}$ will consist of more than one single fiber action potential.

Once the decomposition is complete, the individual decomposed waveforms from the two recording sites can be used to compute an estimate of the group delay associated with each contributing nerve fiber where the frequency response of a given fiber $H_{i-1}(f)$ is as shown in (6).

$$H_{i-1}(f) = \frac{\Im\left[\Gamma_{i-1}^{(2)}(t)\right]}{\Im\left[\Gamma_{i-1}^{(1)}(t)\right]} \quad (6)$$

The frequency response is computed by dividing the Fourier Transform of the single fiber evoked potential associated with the more distal recording site by the Fourier Transform of the single fiber evoked potential associated with the more proximal recording site. Since each $H_{i-1}(f) = |H_{i-1}(f)|\angle\Theta_{i-1}(f)$, an estimate of the group delay τ_{i-1} for each pair of non-zero decomposed waveforms $\Gamma_{i-1}^{(1)}(t)$ and $\Gamma_{i-1}^{(2)}(t)$ can be computed from (7).

$$\tau_{i-1} = -\frac{1}{2\pi}\frac{d\Theta_{i-1}(f)}{df} \quad (7)$$

In practice, a least squares line is fitted to the phase response $\Theta_{i-1}(f)$ for the $H_{i-1}(f)$ computed for each pair of non-zero decomposed waveforms $\Gamma_{i-1}^{(1)}(t)$ and $\Gamma_{i-1}^{(2)}(t)$ which facilitates the computation of the associated group delay τ_{i-1}. The estimated group delays for the contributing nerve fibers are used to compute an estimate of the associated fiber diameters from (8) where l *(m)* is the distance between the two recording sites and $c = 3.0\times10^6$ s^{-1}.

$$d_{i-1} = \frac{l}{c\tau_{i-1}} \quad (8)$$

Once the estimated group delay is computed for each non-zero pair of decomposed waveforms $\Gamma_{i-1}^{(1)}(t)$ and $\Gamma_{i-1}^{(2)}(t)$, an estimate of the sequence of nerve fiber diameters $\tilde{\boldsymbol{d}}$ is obtained for the contributing fiber population.

The overall process described above is illustrated in the flowchart of Figure 2.

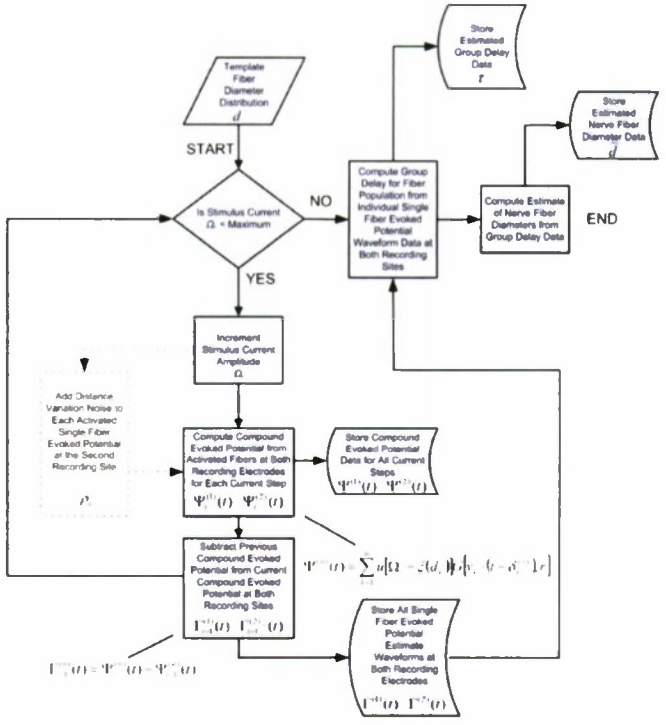


Figure 2. Flowchart of algorithm implemented to calculate an estimate of the group delay of the contributing population of nerve fibers and the estimated fiber diameter set $\tilde{\boldsymbol{d}}$. The technique is based on determination of an estimate of the group delay associated with each non-zero decomposed waveform $\Gamma_{i-1}^{(1)}(t)$ and $\Gamma_{i-1}^{(2)}(t)$.

Random variations, from the first recording site to the second, in the perpendicular distance between the recording site and the center of the nerve fiber r could reasonably be expected. This characteristic would result in random variations in the form of the contributing single fiber action potentials $G\left[v_k \cdot \left(t-\delta_k^{(n)}\right), r\right]$ associated with each specific fiber. To simulate the impact of this noise source on the system, r is replaced with a normally distributed random number with a standard deviation of ρ_k and a mean of $\bar{r}$.

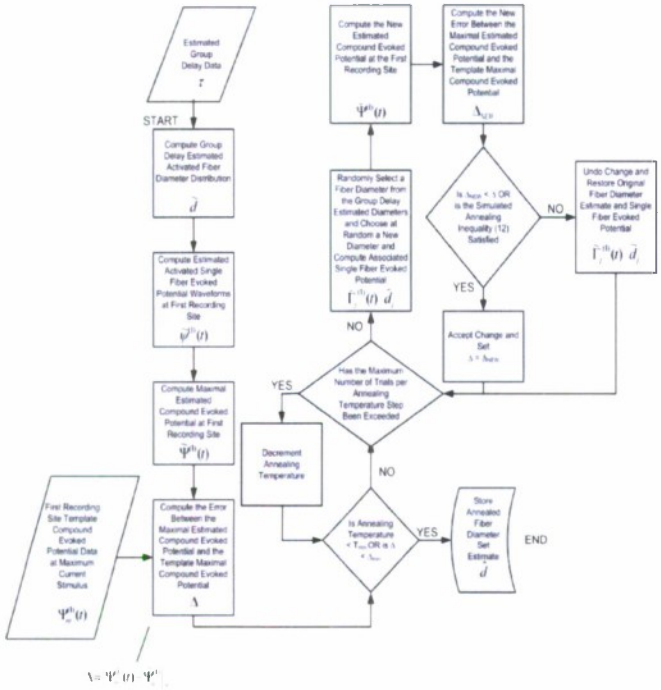


Figure 3. Flowchart of the simulated annealing algorithm that computes an improved estimate of the nerve fiber diameter population set $\hat{\boldsymbol{d}}$ from the group delay estimated population set $\tilde{\boldsymbol{d}}$.

The simulated annealing algorithm, first proposed by Metropolis *et. al.* [16], can be used to compute solutions to NP-complete problems such as the traveling salesman problem, where these solutions are optimal or very close to optimal [17;18]. The problem of determining the correct diameters of the contributing nerve fiber population and concomitant forms for each of the single fiber evoked potentials, is also an NP-complete problem to which the simulated annealing algorithm may be applied. The optimized contributing fiber diameter set is a set of diameters $\hat{\boldsymbol{d}}$ such that (9) is true.

$$\min\left\|\Psi^{(1)}(t)-\hat{\Psi}^{(1)}(t)\right\|_2 \tag{9}$$

The waveform $\Psi^{(1)}(t)$ is the first recording site template evoked potential at the maximum stimulus current pulse amplitude such that $i=p$ and $\hat{\Psi}^{(1)}(t)$ is the maximal compound evoked potential estimated from the group delay decomposition at the first recording site as per (10) and optimized using the simulated annealing algorithm. In (10), q is the number of non-zero single fiber evoked potential waveforms obtained through the decomposition shown in (4) and $\tilde{\Gamma}_j^{(1)}$ are the decomposed fiber evoked potential waveforms at the first recording site.

$$\tilde{\Psi}^{(1)}(t)=\sum_{j=1}^{q}\tilde{\Gamma}_j^{(1)}(t) \tag{10}$$

The algorithm to compute the estimated set of nerve fiber diameters $\hat{\boldsymbol{d}}$ from simulated annealing is illustrated in the flowchart of Figure 3.

The first step is to compute the group delay estimated set of contributing single fiber action potential waveforms $\tilde{\Gamma}^{(1)}(t)$ and then to compute $\tilde{\Psi}^{(1)}(t)$ from (10). The error between the template maximal compound evoked potential $\Psi^{(1)}(t)$ and the group delay estimated compound evoked potential with all fibers contributing $\tilde{\Psi}^{(1)}(t)$ is computed. This error Δ is formulated in terms of the two-norm difference between the two waveforms.

$$\Delta=\left\|\Psi^{(1)}(t)-\tilde{\Psi}^{(1)}(t)\right\|_2 \tag{11}$$

The simulated annealing algorithm initially determines whether the $\Delta<\Delta_{MIN}$ or if the annealing temperature $T<T_{MIN}$. If either of these inequalities hold, the algorithm exits and the optimized set of fiber diameters $\hat{\boldsymbol{d}}$ is set equal to the set $\tilde{\boldsymbol{d}}$. In the event neither inequality holds, an estimated fiber diameter $\tilde{d}_j$ is randomly chosen from the set $\tilde{\boldsymbol{d}}$ and a new fiber diameter for this specific fiber is selected at random such that $\check{d}_j=\check{d}_j$. The concomitant single fiber evoked potential waveform for the randomly generated fiber diameter is computed at the first recording site $\check{\Gamma}_j^{(1)}(t)$ using (4). With this new randomly generated single fiber action potential, a new value is computed for the estimated compound evoked potential associated with the contribution of all fibers in the population $\check{\Psi}^{(1)}(t)$. Equation (11) is invoked to compute a new error estimate Δ_{NEW}. If $\Delta_{NEW}<\Delta$ or if the simulated annealing inequality in (12) holds, where Υ is a uniformly distributed random number between zero and unity, the new randomly generated fiber diameter is accepted.

$$\Upsilon\le\exp\left[-\frac{|\Delta-\Delta_{NEW}|}{T}\right] \tag{12}$$

If neither of these inequalities hold, the newly generated random fiber diameter $\check{d}_j$ is rejected and replaced with the old fiber diameter $\tilde{d}_j$. This process is repeated for the number of trials allowed per annealing step. Once the number of trials is exceeded, the annealing temperature T is reduced and the entire process is repeated for the number of trials allowed per annealing step. The simulated annealing algorithm exits once either $T<T_{MIN}$ or $\Delta<\Delta_{MIN}$. The output of the algorithm is the annealed estimate of the fiber diameter set $\hat{\boldsymbol{d}}$.

III. RESULTS

A population of 100 randomly generated fibers was utilized in these studies. Fibers with diameters less than 5 μm were rejected yielding a subpopulation of $m=60$ fibers with a maximum diameter of 17.1 μm. These fibers were subjected to a virtual stimulus pulse train of successively increasing amplitudes ranging from zero to a maximum of 1 mA in 500 nA steps. At each step the compound evoked potential at both virtual recording sites was computed as per (3) and subsequently, the estimate of single fiber action potential waveforms were obtained at each recording site as

per (5). The concomitant group delays between the two virtual recording sites were computed yielding the group delay estimated set of fiber diameters $\tilde{\boldsymbol{d}}$.

A histogram comparing the actual template fiber population $\boldsymbol{d}$ to the group delay estimated population $\tilde{\boldsymbol{d}}$ is shown in Figure 4.

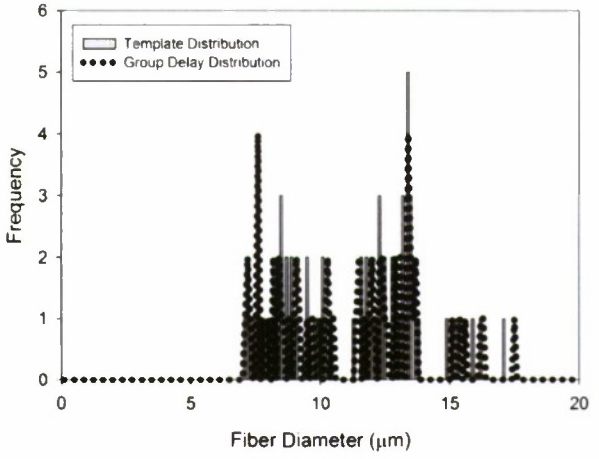


Figure 4. Histogram of template nerve fiber size population $\boldsymbol{d}$ and group delay estimated nerve fiber size population $\tilde{\boldsymbol{d}}$. The simulation was carried out with an SNR of 20 dB with respect to random variations in the perpendicular distance between the nerve fiber and the recording site. Chi Square Test results for the two distributions yielded $Q(\chi^2|x) = 0.7101$.

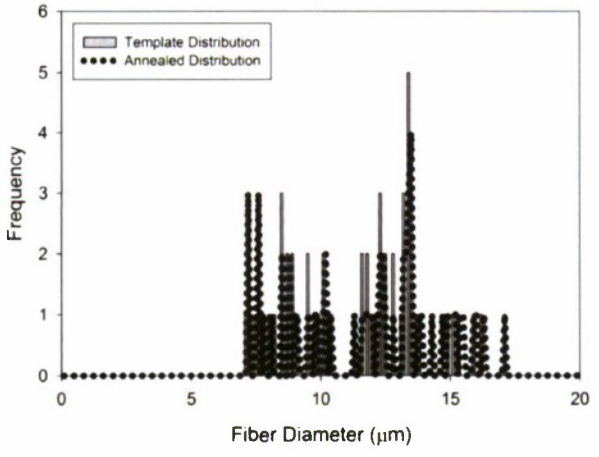


Figure 5. Histogram of the template nerve fiber size population $\boldsymbol{d}$ and annealed nerve fiber size population $\hat{\boldsymbol{d}}$ from the group delay estimated distribution $\tilde{\boldsymbol{d}}$ shown in Figure 4. The simulation was carried out with an SNR of 20 dB with respect to random variations in the perpendicular distance between the nerve fiber and the recording site. Chi Square Test results for the two distributions yielded $Q(\chi^2|x) = 0.9912$.

The estimated fiber diameter population $\tilde{\boldsymbol{d}}$ was then optimized using the simulated annealing algorithm described earlier. An initial annealing temperature of $T = 10$ was used with an annealing step factor of 0.9. The minimum annealing temperature was $T_{MIN} = 1 \times 10^{-5}$ with a minimum error bound $\Delta_{MIN} = 1$. The maximum number of trials for each annealing temperature step was limited to 1000. A histogram comparing the actual template fiber diameter population $\boldsymbol{d}$ to the optimized diameter population $\hat{\boldsymbol{d}}$ is shown in Figure 5.

Figure 6 compares the maximal template compound evoked potential at the first electrode $\Psi^{(1)}(t)$ with the maximal group delay estimated compound evoked potential $\tilde{\Psi}^{(1)}(t)$ and the maximal annealed compound evoked potential $\hat{\Psi}^{(1)}(t)$ for the distributions shown in Figures 4 and 5.

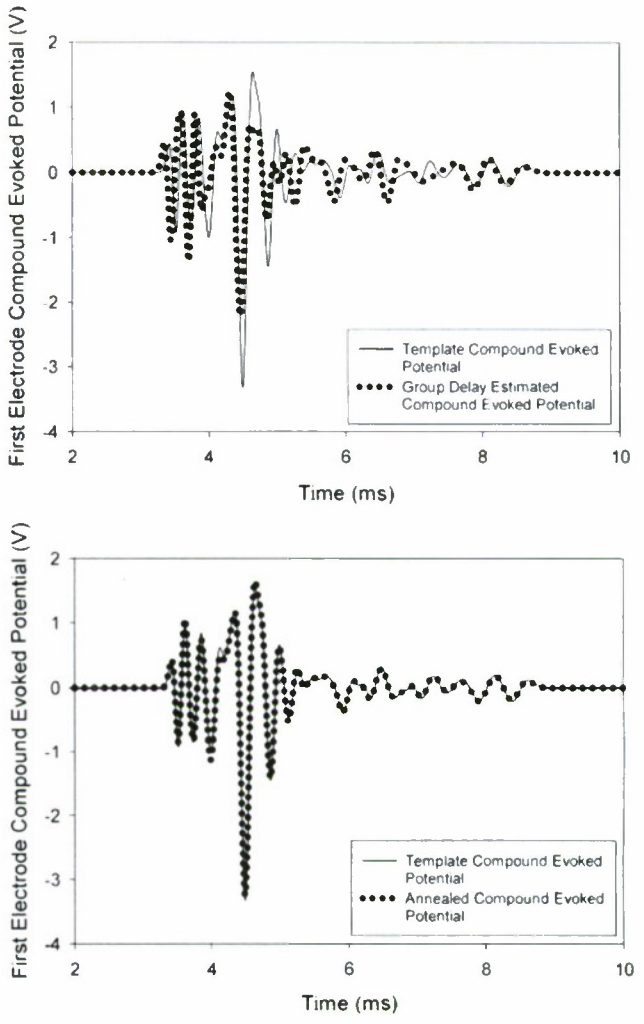


Figure 6. Graphs comparing the template maximal compound evoked potential to the group delay estimated maximal compound evoked potential and the annealed maximal compound evoked potential. The graph on the top shows the template maximal compound evoked potential at the first recording site $\Psi^{(1)}(t)$ and the group delay estimated maximal compound evoked potential at the first recording site $\tilde{\Psi}^{(1)}(t)$. A comparison of $\Psi^{(1)}(t)$ to the annealed maximal compound evoked potential $\hat{\Psi}^{(1)}(t)$ that results from optimization of the group delay estimate with the simulated annealing algorithm is shown in the graph on the bottom. The normalized final error between $\Psi^{(1)}(t)$ and $\hat{\Psi}^{(1)}(t)$ as per (13), is $FE = 7.673\%$.

The effect of noise variations in the distance between the individual nerve fibers at the two virtual recording sites was also studied. Table 2 consists of typical Chi Squared Test results comparing the optimized and template histograms over a range of SNR values where the signal value is taken as the mean distance $\bar{r}$ and the noise power is the variance

$\rho_k{}^2$ in the normally distributed distance variation around the mean value.

SNR (dB)	Group Delay and Template Distribution $Q(\chi^2\|x)$	Annealed and Template Distribution $Q(\chi^2\|x)$	Annealed and Template Transient Final Error FE (%)
0	0.3958	0.7235	20.93
20	0.7101	0.9912	7.673
40	0.7753	0.8850	10.54

Table 2. Chi Square Test results and final normalized error FE for different values of SNR associated with normally distributed random variations in the perpendicular distance between the recording electrode site and the concomitant nerve fiber.

The final error in percent, as shown in Table 2, was calculated as per (13).

$$FE = \left(\frac{\left\| \Psi^{(1)}(t) - \widehat{\Psi}^{(1)}(t) \right\|_2}{\left\| \Psi^{(1)}(t) \right\|_2} \right) 100\% \qquad \textbf{(13)}$$

IV. DISCUSSION

The results of the simulation study shown earlier demonstrate that the technique presented herein can, with reasonable accuracy, retrieve the conduction velocity distribution in the presence of noise introduced through variations in the perpendicular distance between the recording site and the contributing fiber. The algorithm is effective even in reasonably high noise situations where the SNR associated with variations in the nerve fiber depth is 0 dB. In the 0 dB case, a Chi Square Test comparing the actual template nerve fiber size distribution $\boldsymbol{d}$ with the distribution obtained from the group delay extraction $\tilde{\boldsymbol{d}}$ yields a result of 0.3958. After application of the simulated annealing algorithm to yield the optimized distribution $\widehat{\boldsymbol{d}}$, a Chi Square Test comparing the optimized distribution to the actual template distribution $\boldsymbol{d}$ yielded a higher value of 0.7235 which suggests that the annealed diameter set and the template set are more consistent with a single distribution than the template set and the group delay set. Similar results were given in Table 2, for higher SNR values associated with variation in the perpendicular nerve fiber and recording site distance.

The graphs shown in Figure 6 further demonstrate the ability of the simulated annealing optimization approach to improve upon the maximal compound evoked potential $\tilde{\Psi}^{(1)}(t)$ estimated using the group delay extracted population $\tilde{\boldsymbol{d}}$ as compared to the template maximal compound evoked potential $\Psi^{(1)}(t)$. After application of the simulated annealing algorithm, the bottom graph in Figure 6 demonstrates a reasonable convergence between the template maximal compound evoked potential $\Psi^{(1)}(t)$ and the maximal compound evoked potential $\widehat{\Psi}^{(1)}(t)$ associated with the optimized fiber diameter population $\widehat{\boldsymbol{d}}$. This convergence is further demonstrated by the relatively low normalized final error of 7.673%.

Since the optimization process is random in nature, a successive decrease in the normalized final error is not always observed with increasing SNR values as demonstrated between the 20 dB and 40 dB data shown in Table 2. The increase in the normalized final error from the 20 dB to the 40 dB case is consistent with a lower value of the Chi Square Test comparing the template distribution to the annealed distribution for an SNR value of 40 dB than for the 20 dB case.

The proposed technique for measuring the size distribution of nerve fibers that contribute to the maximal compound evoked potential has several advantages over other earlier proposed methods. Unlike some previous techniques [5;10], no inherent assumptions are made regarding size based classification of contributing single fiber evoked potentials. Each contributing single fiber evoked potential can, in theory, have a unique wave shape. The fact that many of the other techniques stipulate specific forms of the single fiber action potential waveforms based upon dividing the range of fiber diameters into distinct groups, makes direct comparison of these techniques problematic.

One of the disadvantages of the proposed approach, in comparison to other techniques is the necessity to perform a series of compound evoked potential measurements associated with a train of successively increasing stimulus current pulse amplitudes. While the measurement associated with the proposed method is more involved, the protocols for extracting individual contributing evoked potentials based upon a successively increasing stimulus pulse amplitude is well established in the literature on motor unit number estimation [13].

REFERENCES

[1] P. A. Parker and P. Kelly, "Nerve conduction velocity measurement techniques," *Journal of Clinical Engineering*, vol. 7, no. 2, pp. 153-158, Apr.1982.

[2] I. Atroshi, C. Gummesson, R. Johnsson, and E. Ornstein, "Diagnostic properties of nerve conduction tests in population-based carpal tunnel syndrome," *BMC Musculoskeletal Disorders*, vol. 4, no. 9 May2003.

[3] Y. Harati, "Diabetic peripheral neuropathies," *Annals of Internal Medicine*, vol. 107, pp. 546-559, 1987.

[4] L. J. Dorfman, K. L. Cummins, G. M. Reaven, J. Ceranski, M. S. Greenfield, and L. Doberne, "Studies of diabetic polyneuropathy using conduction velocity distribution (DCV) analysis," *Neurology*, vol. 33, no. 6, pp. 773-779, June1983.

[5] K. L. Cummins, L. J. Dorfman, and D. H. Perkel, "Nerve fiber conduction-velocity distributions, II Estimation based on two compound action potentials," *Electro-Encephal. Clin. Neurophysiol*, vol. 46, no. 6, pp. 647-658, 1979.

[6] L. J. Dorfman, "The distribution of conduction velocities (DCV) in peripheral nerves: a review," *Muscle & Nerve*, vol. 7, no. 1, pp. 2-11, 1984.

[7] J. A. Gonzalez-Cueto and P. A. Parker, "Deconvolution estimation of nerve conduction velocity distribution," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 140-151, Feb.2002.

[8] F. A. Papadopoulou and S. M. Panas, "Estimation of the nerve conduction velocity distribution by peeling sampled compound action potentials," *IEEE Transactions on Magnetics*, vol. 35, no. 3, pp. 1801-1804, May1999.

[9] D. Gu, R. E. Gander, and E. C. Crichlow, "Determination of nerve conduction velocity distribution from sampled compound action potential signals," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 8, pp. 829-838, Aug.1996.

[10] Y. X. Tu, A. Wernsdorfer, S. Honda, and Y. Tomita, "Estimation of Conduction velocity distribution by regularized-least-squares method," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 11, pp. 1102-1106, Nov.1997.

[11] R. B. Szlavik, "A novel method for characterization of peripheral nerve fiber size distributions by group delay," *IEEE Transactions on Biomedical Engineering*, 2008, *accepted*.

[12] S. M. Fleisher, M. Studer, and G. S. Moschytz, "Mathematical model of the single-fiber action potential," *Medical and Biological Engineering and Computing*, vol. 22, pp. 433-439, 1984.

[13] A. J. McComas, P. R. W. Fawcett, M. J. Campbell, and R. E. P. Sica, "Electrophysiological estimation of the number of motor units within a human muscle," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 34, pp. 121-131, 1971.

[14] I. A. Boyd and M. R. Davey, *Composition of Peripheral Nerves*. Edinburgh: E & S Livingstone Ltd., 1968.

[15] R. B. Szlavik and H. de Bruin, "Simulating the distribution of axon size in nerves," Canadian Medical and Biological Engineering Society, 1997, pp. 168-169.

[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087-1092, June1953.

[17] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines* John Wiley & Sons, 1989.

[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes (Fortran)*. New York: Cambridge University Press, 1990.

Implementation of Improved Physiologic Components in an In Vitro Tissue Engineered Blood Vessel Mimic for Stent Evaluation

Dimitri E. Delagrammaticas[1], Mare C. Dawson[1], and Kristen O. Cardinal[1]
[1]Cal Poly, San Luis Obispo, California, USA

Introduction

Due to the rapid emergenee of minimally-invasive intravaseular technologies for treating and diagnosing eardiovaseular disease, the need exists for relevant preelinieal models to evaluate these new deviees. Tissue engineered blood vessels have the potential to meet sueh a need, as these living eonstruets ean be eomposed of human vaseular eells and eultivated in a eontrolled bioreaetor environment to ereate an in vitro model system. Previous work has demonstrated the ability to ereate and use tissue engineered blood vessels for evaluating endothelialization of intravaseular stents (1). Although the previously developed vessel mimie systems have served as an adequate model, several potential limitations exist, ineluding the use of mierovaseular endothelial eells and sub-physiologie levels of shear stress. Therefore, it was hypothesized that speeifie features of the engineered eonstruet eould be altered to make the model a more representative vessel mimie. The objectives of the current work were to determine feasibility of implementing 1) a large vessel endothelial eell souree and 2) viseous media that would more elosely mimie shear stresses of an artery.

Materials and Methods

Tissue engineered blood vessels were developed using 4mm ID expanded polytetrafluoroethylene seaffolds as deseribed previously (1). Seaffolds were inserted into bioreaetor systems and were pressure-sodded with passage 4-5 human umbilieal vein endothelial eells (HUVECs). Luminal flow was maintained for up to 7 days. Following eultivation, vessels were removed and fixed for evaluation by seanning electron mieroseopy in order to assess luminal eell linings. Inereased shear stress was implemented by seleeting a media thiekener to increase viseosity of the eireulating nutrient media. Dextran and methyl eellulose were identified as potential options (2, 3), and were eompared based on the eoneentration neeessary and the dissolvability in eell eulture media. Tissue engineered vessels were ereated and viseous media was added to media reservoirs to evaluate use in the vessel mimie systems.

Results

HUVECs sodded at a density of 5 x 10^5 eells/em^2 formed a luminal lining within vessel eonstruets. The eellular lining was sustained in vitro for up to 7 days. Ongoing studies will elueidate the potential to sustain the vessels for extended durations in vitro and the potential for the HUVEC eell lining to withstand stent implantation. Dextran was seleeted as the preferred thiekening agent due to its ability to easily dissolve in our eell eulture media. Media eontaining 8% dextran was seleeted due to its viseosity and was implemented as the eireulating media within blood vessel mimie systems. Cell linings withstood eireulation of this viseous media. Further optimization of viseosity for desired shear stress levels and evaluation of the effeets of inereased shear stresses are underway. Future researeh will determine the impaet of inereased shear stress on the endothelialization response to stents within this in vitro model system.

Discussion and Conclusions

This work supports the feasibility of implementing improved physiologie eonditions within an existing tissue engineered blood vessel mimie as a preelinieal model for intravaseular deviee evaluation. The use of HUVECs as a large vessel eell souree and the addition of a media supplement for inereased viseosity deseribed in this work are signifieant for their applieation to the in vitro blood vessel mimie model. The use of tissue engineered blood vessels as preelinieal models for deviee evaluation is signifieant for both the tissue engineering field as well as the medieal deviee industry.

References

1. Cardinal KO, Bonnema GT, Hofer H, Barton JK, Williams SK. Tissue engineered vaseular grafts as in vitro blood vessel mimies for the evaluation of endothelialization of intravaseular deviees. Tissue Engineering, 12(12), 2006.
2. Gallik S, Bradshaw A, van Wambeek M. Effeets of methyleellulose on epithelial eells. In Vitro Cell Dev Biol Anim, 29A, 1993.
3. Sikavitsas VI, Baneroft GN, Holtorf HL, Jansen JA, Mikos AG. Mineralized matrix deposition by marrow stromal osteoblasts in 3D perfusion eulture inereases with increasing fluid shear forees. Proe Natl Aead Sei USA, 100, 2003.

Acknowledgements

Portions of this work were sponsored by the Department of the Navy, Offiee of Naval Researeh, under Award # N00014-07-1-1152.

Disclosures

The authors have nothing to diselose.

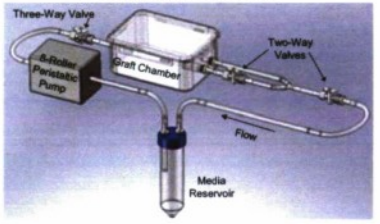
A Simple Method of Implementing Physiologic Flow Conditions in a Tissue Engineered Blood Vessel Construct

Marc Dawson and Kristen O'Halloran Cardinal
Biomedical Engineering
California Polytechnic State University, San Luis Obispo, CA

Introduction

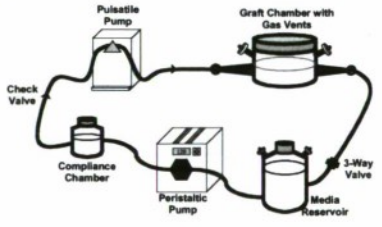
- A tissue engineered human "blood vessel mimic" (BVM) has been developed with the goal of using it as a high throughput in vitro testing system for intravascular devices
  - This has potential to reduce time and costs of bringing devices from inception to patient use.

Current steady flow, low shear system



- The purpose of the present work is to implement physiologic conditions into the current low shear, steady flow system in order to create a *more* native-like construct by:
  - Establishing pulsatile flow to create the radial and longitudinal strain that is necessary for proper growth and improved vessel strength.
  - Creating appropriate wall shear stress conditions to induce proper endothelial cell (EC) orientation
- One key objective of this work was to introduce minor changes to the system in order to achieve desired effects while maintaining the original system's inherent simplicity with the intention that it can be scaled up for use as a high throughput model.
- Other highly tunable systems have been developed for creating specialized physiologic flow, but the focus of this work was to balance simplicity with desired function .

Example of highly tunable pulsatile, low shear system developed by Hahn et al



Materials and Methods

• Pulsatile Flow

- Flow path reconfiguration:
  - A 3-roller peristaltic pump and larger diameter pump tubing was introduced to increase pulsatility while maintaining flow rate of 10-15 ml/min.
  - A valve was placed downstream of the BVM chamber to control back pressure in order to establish/adjust luminal pressure fluctuations.

Materials and Methods (Continued)

• Pulsatile Flow (Continued)

- Pulse rate, flow rate and luminal pressure monitored using the PowerLab 4/25 data acquisition system by ADInstruments.

• Wall Shear Stress

- Media formulation:
  - Growth media was adulterated to adjust viscosity in order to achieve desired wall shear stress of 8-10 dyn/cm2
  - Two polysaccharides were considered as a means of increasing viscosity without affecting cell growth.
    - Powdered Methylcellulose from Sigma-Aldrich (P/N M0555)
    - Powdered Dextran from Sigma-Aldrich (P/N D4751)
      - average mol wt 64,000-76,000
  - Dextran was added directly to the prepared growth media, dissolved into solution at 37°C and filter sterilized.
- Viscosity of the resulting solutions with varying concentrations of Dextran were evaluated using a Vilastic-3 viscometer from Vilastic Scientific Inc.
- Wall shear stress at the inner walls of the BVM was calculated with the following equation:

$$\tau_{wall} = 32\mu Q/\pi D^3$$

τ_{wall} - wall shear stress (N/m^2)
μ - viscosity ($N \bullet s/m^2$)
Q - volumetric flow rate (m^3/s)
D - tubing diameter (m)

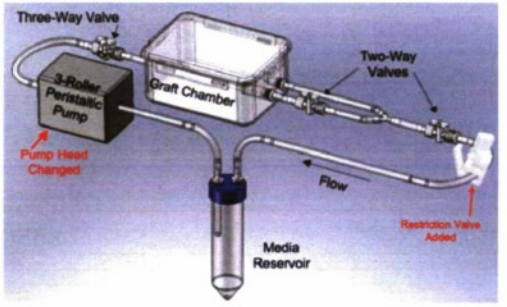
- This simple shear stress equation was developed with the following assumptions and considerations:
  - The growth media is a Newtonian fluid
  - Flow through the graft/scaffold is:
    - fully developed, pressure driven, steady and laminar
    - through absolutely circular tubing

• Cell Adhesion

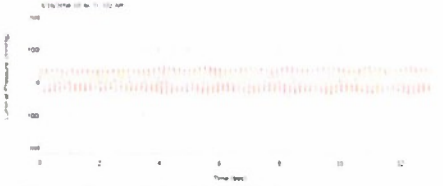
- Pulsatile Flow
  - BVM constructs were sodded with HUVSMCs under steady flow and allowed to incubate at 37°C and 5% CO2.
  - After 3 days under steady flow, the pump head was changed to the 3-roller configuration and back pressure was applied to achieve luminal pressures of ~60mmHg.
  - Constructs were taken down after 2 and 24 hours of pulsation to evaluate the degree of cell adhesion
- Wall Shear Stress
  - BVM constructs were sodded with HUVECs under steady flow and allowed to incunbate at 37°C and 5% CO2.
  - After 4 days of incubation with growth media containing 0% Dextran, the media supply was changed out for growth media containing 8% dextran.
  - Constructs were taken down after 24 hours of incubation under increased shear conditions to evaluate the degree of cell adhesion.

Results

• Reconfigured BVM Bioreactor System



• Pulsatile Flow



Pressure/Flow profile in original BVM with steady flow configuration using 8-roller pump head and no back pressure



Pressure/Flow profile in BVM with pulsatile flow configuration using 3-roller pump head and added back pressure to achieve 60-85mmHG pressure peaks

• Wall Shear Stress

- Methylcellulose vs. Dextran
  - The solubilization of Methylcellulose proved to be problematic and required steps that may render the growth media unusable due to the high temperature required to dissolve it. As a result, Methylcellulose was abandoned as a potential use in our system.

Results (Continued)

• Wall Shear Stress (Continued):

Measured media viscosities and resulting calculated WSS in BVM

Media Formulation	Viscosity (Poise)	Shear Rate (s^-1)	WSS in BVM @ 15 ml/min (Dyne/cm^2)
0% Dextran	9.03E-03	231.266	0.851
2% Dextran	1.42E-02	225.603	1.338
4% Dextran	2.09E-02	236.714	1.966
6% Dextran	3.10E-02	234.490	2.925
8% Dextran	4.33E-02	229.300	4.076

• Cell Adhesion

- Pulsatile flow
  - Initial tests indicated that cell adhesion can be maintained under pulsatile flow established using means presented in this work.
- Wall Shear Stress
  - Initial tests indicated that cell adhesion can be maintained when wall shear stress is applied under conditions presented in this work.

Conclusions

- Physiologic shear and increased pulsation can be implemented with minor changes to en elready established BVM bioreactor system in an manner that maintains cell adhesion.
- Pressure profiles indicate that physiologic pressure fluctuations can be achieved within this system to improve cell orientation and differentiation.
- Appropriate Wall Shear Stress can be applied to this BVM construct to encourage proper endothelial cell alignment.
- Further experimentation is necessary to determine the long term effects of WSS and pulsatility on cell growth and vessel development within this BVM system.
- The flow characteristics of this model can be tuned to simulate different blood pressure and blood viscosity based conditions.
- This system shows promise as a physiologic model for use as an in vitro testing system for intravascular devices

References

- MARIAH S. HAHN,* MELISSA K. MCHALE,* EVA WANG, RACHAEL H. SCHMEDLEN, and JENNIFER L. WEST

Acknowledgements

- **Portions of this work were sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152.**

Assessment of Electrospun Scaffolds for use in High-Throughput Blood Vessel Constructs

Colby James and Kristen O'Halloran Cardinal
Biomedical Engineering
California Polytechnic State University, San Luis Obispo, CA

Introduction

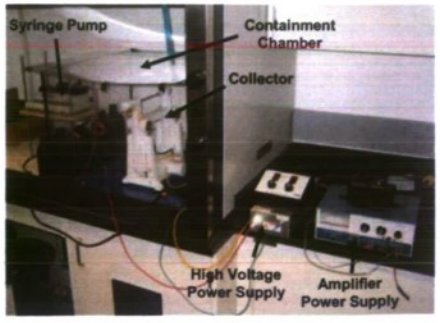
- Previous work has demonstrated the ability to utilize tissue engineering techniques to create high throughput in vitro blood vessel constructs for intravascular device testing.
- These "blood vessel mimics" (BVMs) previously used expanded polytetrafluoroethylene (ePTFE) as scaffolding to grow a confluent layer of human endothelial cells.
- ePTFE's lack of radial compliance, cost, and out-of-house production demand a scaffold that is more appropriate for the BVM.
- Electrospun scaffolds are highly tailorable with respect to material, mechanical properties, and microstructure, as well as inexpensive to quickly produce in-house.
- Goal of present work: evaluate the potential of electrospun scaffolds for possible use in the Blood Vesel Mimic system
  - Evaluate morphological parameters and consistency
  - Evaluate mechanical characteristics and consistency

Materials and Methods

Electrospinning Scaffolds:

- An electrospinning apparatus was constructed using a high voltage power supply, secondary amplifier power supply, syringe pump, single grounded rotating/translating mandrel collector, and containment chamber as shown below.
- 90:10 poly(L-lactide-*co*-caprolactone), or PLAPCL, was dissolved in chloroform at a concentration of 9.0 wt% polymer for 24 hours before spinning.
- 3ml of solution was electrospun at a rate of 6ml/hr, utilizing an 18 gauge needle charged to 15kV and separated 10in from a single grounded rotating/translating mandrel collector
- Finished scaffolds were 11-12cm long and 4mm in diameter.

The Electrospinning System

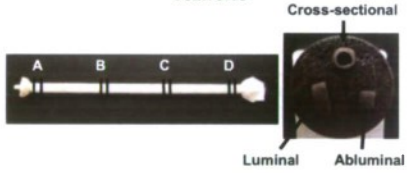


Materials and Methods

Morphology Examination:

- Morphological evaluation of the electrospun scaffolds was performed using scanning electron microscopy (SEM).
- To evaluate consistency along the length of the scaffold, 4 sections were sampled for SEM analysis as shown below.
  - Each section yielded 3 different surfaces for examination: luminal, abluminal, and cross-sectional as shown below.
- Luminal and abluminal images were used to examine fiber size and porosity, while cross-sections were used to determine wall thicknesses

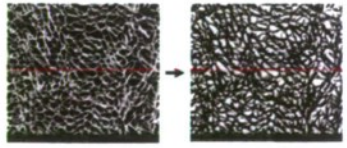
Sectioning Electrospun Scaffolds



As-spun scaffolds were sectioned at A, B, C, and D (left), and each section was used as shown at right.

- Fiber Size: Fiber diameter was measured randomly using secondary electron SEM.
- Thickness: Wall thickness of the scaffolds was measured using secondary electron SEM.
- Porosity: Space not filled by electrospun fibers on the surface being examined was measured by capturing high contrast backscattered electron images and calculating the negative space using image analysis software as shown below.

Porosity Imaging



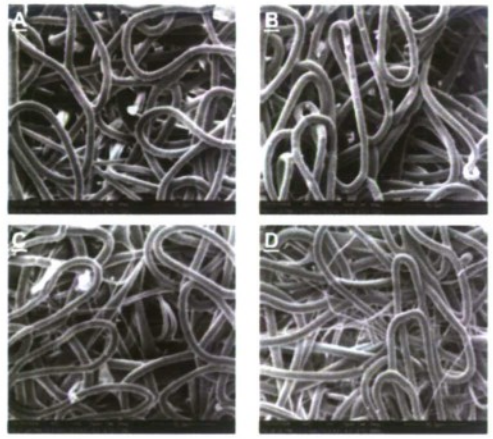
High contrast backscattered electron images (left) were converted to 2-color black and white images, followed by an operation that counted the percent white, or void, space.

Mechanical Testing:

- The large portions of scaffold between the sections used for SEM analysis were used for tensile testing.
  - Each portion was cut into two separate sheets to perform tensile testing in both the longitudinal and radial directions.

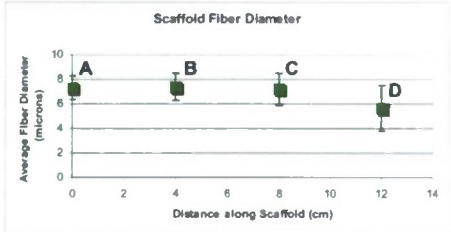
Results

Fiber Size Along Scaffold Length



Scanning electron micrographs of the lumen at the four sites along the length of a selected scaffold.

Fiber Diameter:



Average diameter of luminal fibers at the four sites along the length of a selected scaffold.

Thickness:



746 ± 41μm **746 ± 54μm** **800 ± 58μm** **693 ± 41μm**

Average thickness of a selected scaffold at the four sites along the length.

Results

Porosity:

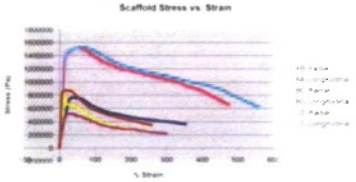
Sectional Porosity

SECTION	POROSITY
A	50.5%
B	35.0%
C	40.7%
D	32.6%

Table of porosity values generated at the four sites along the lumen of a selected scaffold.

Mechanical Properties:

Tensile Testing Scaffolds



Graph of stress vs. strain for the 3 sections in-between the 4 sites sampled for SEM analysis. Each of the 3 portions was pulled in the radial and longitudinal directions.

Conclusions

- Scanning electron microscopy morphological examination suggests a possible microstructural difference at the "D" end of electrospun PLAPCL scaffolds.
- Tensile strength points towards an anisotropic mechanical response from these scaffolds, with a weaker material at the "D" end of the scaffold.
- More work is needed to draw statistically significant conclusions about the consistency of these electrospun scaffolds.
- The electrospinning process may provide a means for inexpensive, high-throughput production of scaffolds that could be implemented in the Blood Vessel Mimic system.

Acknowledgements

- **We would like to acknowledge Gene Boland and Anthony Yang from Tissue Genesis Inc for their expertise and assistance with the electrospinning process.**
- **Portions of this work were sponsored by the Department of the Navy, Office of Naval Research, under Award # N00014-07-1-1152. Additional components were funded by the Cal Poly Biomedical and General Engineering Student Fee Allocation Committee.**

Assigning Closely Spaced Targets to Multiple Autonomous Underwater Vehicles

Beverley Chow, Christopher Michael Clark, *Member, IEEE*, and Jan Paul Huissoon*

Abstract

This paper addresses the problem of allocating closely spaced targets to multiple autonomous underwater vehicles in the presence of constant ocean currents. The main difficultly of this problem is that the non-holonomic vehicles are constrained to move along forward paths with bounded curvatures. The proposed algorithm solves the task allocation problem with market-based auctions that minimize the total travel time to complete the mission. By considering the dynamics and kinematics of the vehicle as well as the effect of ocean currents within the cost function, the proposed algorithm is able to create feasible paths with a lower cost when compared to solutions whose cost functions are calculated based solely on Euclidean distances.

Index Terms - autonomous underwater vehicles, auction-based allocation algorithm, multi-robot systems

1. INTRODUCTION

Autonomous Underwater Vehicles (AUVs) have been used successfully in the past to solve geological, biological, chemical, and physical oceanographic problems. This has resulted in a variety of scientific and commercial AUVs to be designed, built, and deployed. With the increasing feasibility and decreasing expense of AUVs, interest in using them for ocean sampling, mapping, surveillance, and communication is growing and multi-AUV operations are beginning to be realized in the water. As with any multi-robot system, a challenge is to determine which robot should perform which task in order to cooperatively achieve the global goal in an optimal manner.

This paper considers the allocation of m targets to n vehicles. Given a set of vehicles $\{V_1, V_2, ..., V_n\}$ and targets $D = \{d_1, d_2, ..., d_m\}$, the problem is to assign a sequence of targets S_i to each vehicle to visit and a path through the sequence S_i. The objective is to:
Minimize

$$C_{\text{total}} = \max_i C(S_i) \tag{1}$$

subject to

$$D = \bigsqcup_i S_i \qquad \text{(disjoint union)} \tag{2}$$

$$\left.\begin{aligned} \frac{dx_{i,t}}{dt} &= v_0 \cos(\psi_{i,t}) \\ \frac{dy_{i,t}}{dt} &= v_0 \sin(\psi_{i,t}) \\ \frac{d\psi_{i,t}}{dt} &= r, \qquad r \in [-\omega, +\omega] \end{aligned}\right\} \tag{3}$$

where v_0 denotes the nominal vehicle speed, $\psi_{i,t}$ the yaw of the vehicle, and ω represents the bound on the yaw rate. In (1), $C(S_i)$ is the time required for V_i to complete its tour S_i. Note that (2) dictates all tasks to be visited and restricts each task to be assigned to only one vehicle and (3) considers the non-holonomic constraints of the vehicle.

The features that differentiate this research to similar problems previously studied are the kinematic constraints on the vehicle and the presence of a constant ocean current. This paper addresses the inability of an AUV to turn at any arbitrary yaw rate which becomes a problem when target points are close together. The Dubins model [1] is a simple but efficient way to handle the kinematic characteristics of AUVs. It gives complete characterization of the optimal paths between two configurations for a vehicle with limited turning radius moving in a plane at constant speed.

In this paper, Dubins paths are modified to include ocean currents, resulting in paths defined by curves whose radius of curvature is not constant. To determine

*B. Chow and J. P. Huissoon are with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (e-mail: bchow@uwaterloo.ca; jph@uwaterloo.ca)

†C. M. Clark is with the Department of Computer Science, California Polytechnic State University, San Luis Obispo, CA, 93407, USA (e-mail: cmclark@calpoly.edu)

the time required to follow such paths, an approximate dynamic model of the AUV is queried. Specifically, a lower order model of the REMUS AUV model from [2] is used so that the computational complexity is reduced.

The remainder of this paper is organized as follows: In Section II, a background of task allocation problems is given along with a description of various methods used to solve it. In Section III, the kinematics and dynamics of the AUV are described. In Section IV, an overview of the proposed algorithm is introduced with the details presented in Section V. Section VI discusses the results from a simulation conducted in MATLAB and finally, the report concludes in Section VII.

2. BACKGROUND

The goal of the task allocation problem is to have robots visit all targets while minimizing the total travel time or distance of the robots. When targets are known before the mission, it is possible to build a schedule of targets for each robot. Unfortunately, this problem is not straight forward because the cost for a robot to visit target *C* depends on whether that robot first visits target *A* or target *B*. This problem is an instance of the multiple traveling salesperson problem (MTSP), which has been studied extensively in combinatorial optimization. Even in the restricted case of one salesperson, MTSP is strongly *NP*-hard [3].

Several approaches have been applied to the general problem of allocating tasks between multiple robots in a team; refer to [4] for a survey of these. Heuristic methods are typically used since optimizing the performance is often computationally intractable. Parker's ALLIANCE [5] is one of the earliest demonstrations of behaviour-based architectures for task allocation. Another frequently used method is based on market mechanisms, such as auctions, which have been demonstrated in [6] to be fast and robust on real robots. Specific work for AUVs, often called mission planning, include the work by Sariel et al. [7] and vehicles with bounded curvature are considered by Jeyaraman et al. [8]. Similar to the mission planning problem is the routing problem as investigated by Davis et al. [9] for the planning for underwater gliders in the presence of significant currents. However, the vehicle dynamics are not accounted for in their routing strategy.

3. VEHICLE MODEL

3.1. AUV Equations of Motion

This paper uses the REMUS AUV model created by Prestero [2] and considers a vehicle moving at a constant speed in the plane that can execute turns with a bound on maximum curvature. The REMUS is a torpedo-shaped AUV with 6-DOF. The vehicle is propelled by a thruster at its tail and steered by two independent pairs of fins for pitch and yaw control.

Let $X_{i,t} = [u_i\ v_i\ w_i\ p_i\ q_i\ r_i\ x_i\ y_i\ z_i\ \phi_i\ \theta_i\ \psi_i]^T$ denote the state of the vehicle V_i at time t. The first six terms in the state vector describe the linear and angular velocities of the vehicle relative to the body-fixed coordinate and the next six terms describe the position and orientation vector of the vehicle with respect to the earth-fixed frame. Similarly, let (x_j, y_j) denote the position of target d_j.

Given the complex and highly non-linear nature of the problem, numerical integration is used to solve for the vehicle position and orientation in time. At each time step the vehicle state is updated by the general equation:

$$X_{i,t+1} = \boldsymbol{f}(X_{i,t},\ U_{i,t},\ v_c,\ \psi_c) \tag{4}$$

where $X_{i,t}$ is the vehicle state vector, $U_{i,t} = [\delta_s\ \ \delta_r\ \ X_{prop}\ \ K_{prop}]^T$ is the input vector, v_c and ψ_c are the magnitude and direction of the ocean current respectively. For the input vector, δ_s is the stern fin angle, δ_r is the rudder fin angle, X_{prop} is the surge force, and K_{prop} is the yaw torque provided by the propeller.

The function $\boldsymbol{f}$ in (4) uses the Euler method of numerical integration to yield the new vehicle state at each time step as:

$$X'_{i,t+1} = X_{i,t} + \left(\dot{X}_{i,t} \cdot \Delta t\right) \tag{5}$$

where the state vector derivative $\dot{X}_{i,t}$ is updated using the model $\dot{X}_{i,t} = \boldsymbol{f}(X_{i,t},\ U_{i,t})$ from [2]. With the presence of a fixed current, the position of the vehicle relative to the earth-fixed frame is updated as follows:

$$\begin{aligned} x_i &= x'_i + (v_c \cos(\psi_c) \cdot \Delta t) \\ y_i &= y'_i + (v_c \sin(\psi_c) \cdot \Delta t) \end{aligned} \tag{6}$$

where x'_i and y'_i denotes the position of the i^{th} vehicle after the integration step given in (5). By combining (5) and (6), the function $\boldsymbol{f}$ in (4) is realized and can be used to update the general state of each vehicle. This full model is used in evaluating the final tour times of the sequences generated by the proposed algorithm.

4. OVERVIEW OF PLANNER

The algorithm can be broken up into 3 stages.

4.1. Clustering

Given the (x_j, y_j) positions of $j = 1,\ldots,m$ task points, the algorithm starts by creating n clusters of task

points, where n is equal to the number of AUVs. The method used in this paper is *k-means* [10], which partitions the m points into n clusters by minimizing the total intra-cluster variance, or the squared error function. Each of the n clusters are assigned to a single vehicle.

For all $i = 1,2,...,n$ clusters, the three task points in each cluster i that are farthest from the centroid are assigned to i^{th} vehicle.

4.2. Auctioning

Once each vehicle has three task points assigned to it, the remaining $m-3n$ tasks are auctioned via a sequence of first-price one-round auctions similar to work by Lagoudakis et. al [11]. The unassigned tasks are first ordered according to their distance from the centroid of all tasks. The greater the distance from the centroid, the higher the priority the task will have in the order. Following this order, each task is auctioned off. Each vehicle i can bid on the task j, where the bid B_i is equal to the cost of traveling a path that consists of all previously won tasks and the current task being auctioned. Each vehicle considers the insertion of the new task at every point in the current sequence $S_i = (s^1, s^2, s^3, ..., s^l)$ where l is the number of previously won tasks by vehicle i. Each vehicle submits a bid as the lowest cost (i.e. time) to complete the new tour as:

$$B_i(d_j, S_i) = \min_{0 \le k \le l} C\left(s^1, ..., s^k, d_j, s^{k+1}, ..., s^l\right) \quad (7)$$

The i^{th} vehicle with the lowest B_i wins target d_j and updates its sequence of targets with $S_i' = (s^1, ..., s^k, d_j, s^{k+1}, ..., s^l)$. The calculation of $C(S_i')$ is the key to this algorithm's ability to reduce cost and details are found in section 5.

After a task auction is completed, the auctioning process continues with the next round of bidding until all tasks are allocated.

4.3. Post Processing

After all tasks have been auctioned off, each robot has a sequence of tasks to visit. Due to the inherent shortcomings of market-based auction mechanisms, the cost of going backwards through the same sequence of tasks may produce a lower cost value. A post processing step was added at the end of the algorithm to check for the possibility that reversing the order of the task points would produce a lower cost.

5. PATH COST CALCULATION

To determine the path cost for bidding as described above, the time $C(S)$ for the AUV to traverse the sequence needs to be calculated. A combination of a Dubins model, an AUV dynamic model, and a model of the ocean current is required.

5.1. Dubins Path

In order to calculate the time required to travel between two points, the Dubins shortest path problem must first be solved. Dubins' original work derived conditions that characterize the optimal path between two points when both initial and terminal orientations were specified and his work has been widely studied in path planning [12]. Dubins' result [1] shows that, given any two points, the shortest path consists of exactly three path segments. Graphically, the algorithm starts by drawing two maximum curvature circles that are tangential to the initial state vector and two maximum curvature circles that are tangential to the terminal state vector. Dubins' result indicates that the optimal trajectory selects an arc on one of the two initial circles, and connects tangentially to an arc on one of the two terminal circles. If the separation between the initial and end points is sufficient, this can only be accomplished by a line segment. There are at most four such line segments, and computation of the travel distances is straightforward, as shown in Fig. 1 for two waypoints with initial and terminal orientations, denoted α_k and α_{k+1} respectively.

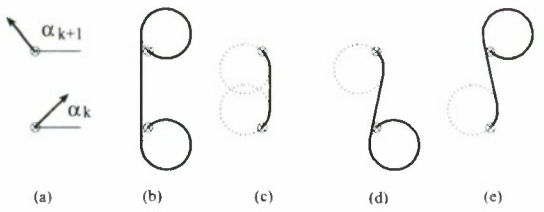


Figure 1. (a) Two waypoints (x_k, y_k) and (x_{k+1}, y_{k+1}) with $\alpha_k = \frac{\pi}{4}$ and $\alpha_{k+1} = \frac{3\pi}{4}$. (b)-(e) Four ways of connecting two waypoints using Dubins curves.

In the task allocation problem, optimal values for α_k and α_{k+1} must be determined. In this algorithm, α_k and α_{k+1} have been constrained to $\Lambda = \{\lambda\pi/4 \mid \lambda = 0, ..., 7\}$. With 8 possibilities for α_k and α_{k+1} and 4 ways of connecting them, a total of $8 \times 8 \times 4 = 256$ paths are possible for every pair of waypoints, with one of them being the optimal path. As shown in Fig 2, different values for α_k and α_{k+1} yield different costs.

The cost of traversing the sequence S can then be

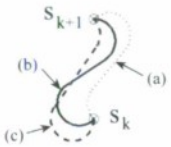


Figure 2. Multiple trajectories for different initial and final orientations. (a) $\alpha_k = \frac{3\pi}{4}$, $\alpha_{k+1} = \frac{-3\pi}{4}$ (b) $\alpha_k = \frac{-3\pi}{4}$, $\alpha_{k+1} = \pi$ (c) $\alpha_k = \frac{-\pi}{2}$, $\alpha_{k+1} = \frac{3\pi}{4}$.

calculated as:

$$C(S) = \min_{(\alpha_1,\ldots,\alpha_l)\in\Lambda^l} \sum_{k=1}^{l} \Delta t_{(s_k,\alpha_k)\rightarrow(s_{k+1},\alpha_{k+1})} \qquad (8)$$

5.2. Ocean Currents

In the presence of ocean currents, the shortest path between two points given α_k and α_{k+1} consists of arcs that are no longer circular but elliptic. These ellipses will have different curvatures depending on the magnitude and direction of the current [Fig. 3].

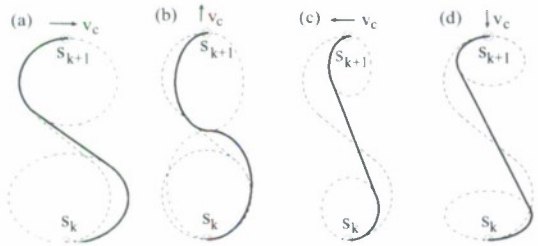


Figure 3. Dubins Curves between two waypoints with ocean currents v_c = 0.25 m/s (a) ψ_c=0, (b) $\psi_c = \frac{\pi}{2}$, (c) $\psi_c = \pi$, and (d) $\psi_c = \frac{-\pi}{2}$.

Finding a line segment tangent to two curves is solved by using an iterative process. As a starting point, the slope of the tangent line to two circular arcs of minimum radius is calculated (Fig. 4a). Using that value, P_a and P_b are found on the respective ellipses whose slope is equal to the slope of the tangent (Fig. 4b). The slope of the line segment from P_a to P_b is calculated and becomes the new slope for the next iteration (Fig. 4c). The process continues until convergence (Fig. 4d).

5.3. Path Time Calculation

To calculate the time required to follow arcs and line segments of the path, the AUV dynamics and ocean currents must be considered. However, due to the complexity of a full dynamic model, it is not possible to query it in a reasonable amount of time. Therefore, a

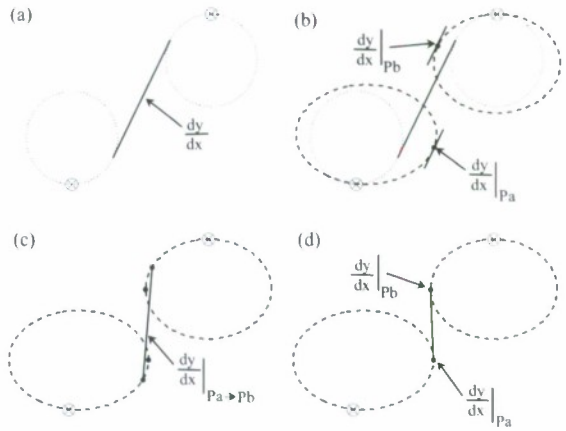


Figure 4. Illustration of the iterative process used to find a tangent to two curves

lower order model $\hat{f}$ is created based on the full model f from (4) as:

$$\Delta t_{(s_k,\alpha_k)\rightarrow(s_{k+1},\alpha_{k+1})} = \hat{f}\,[s_k, s_{k+1}, \alpha_k, \alpha_{k+1}, v_0, v_c, \psi_c] \qquad (9)$$

For arcs, the lower order model is a piecewise linear function built from sampling the full model. Using the full model, the vehicle orientation can be determined at a certain time t. In order to find the time required to obtain a specific heading, linearly interpolation is used on the data obtained from the full model at various fractions of a complete circumnavigation of the ellipse ($\eta\pi/8$ for $\eta = 0,\ldots,15$). For straight line segments, the current vector is broken up into its components relative to the body-fixed frame and is added to the vehicle's velocity vector to determine the time required.

6. SIMULATION RESULTS

This section analyzes the results from a simulation conducted in MATLAB. To illustrate the performance of the proposed algorithm, computer simulations were carried out with a model of the REMUS AUV. The reader is referred to [2] for complete details, including a list of the AUV hydrodynamic parameters.

Simulations were conducted for 20 task points with 3 AUVs. The task points were generated randomly and uniformly inside a square with side lengths of 22 meters. The task points were generated close together to highlight the necessity for considering the curvature constraints.

As a baseline for comparison, the "alternating algorithm" described by Savla et al. [13] is used for the creation of Dubins TSP tours (i.e. sequences from ap-

Table 1. Simulation Results for ten randomly generated datasets.

	No current		With current	
	Alternating Algorithm	Proposed Algorithm	Alternating Algorithm	Proposed Algorithm
T_{max}	164.1	75.4	211.3	77.6
T_{avg}	141.5	71.1	176.8	72.5

Table 2. Simulation Results for one particular trial as illustrated in Fig. 5 and Fig. 6.

	No current		With current	
	Alternating Algorithm	Proposed Algorithm	Alternating Algorithm	Proposed Algorithm
T_{max}	97.4	65.3	121.1	76.0
T_{avg}	88.4	58.6	101.6	62.0

plying Dubins' model). It works as follows: given a set of n points, the optimal Euclidean MTSP tours (i.e. that do not consider path curvature) are computed using auctions (Fig. 5). Then, it is necessary to obtain a feasible path through these ordered points using the method in [13] which includes the curvature constraints of the vehicle.

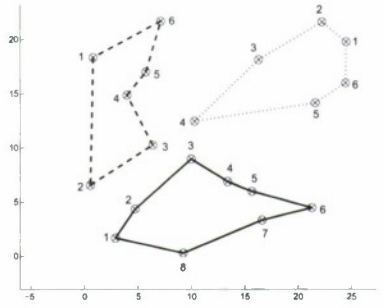


Figure 5. Euclidean MTSP solution for allocating 20 tasks to 3 robots

6.1. Discussion

The results from running the simulation on ten different datasets are summarized in Table 1, where

$$T_{\max} = \max C_{sim}(S_i) \quad \text{and} \quad T_{\text{avg}} = \frac{\sum_{i=1}^{n} C_{sim}(S_i)}{n}.$$

C_{sim} is the cost calculated by running the planned tours S_i through the *full* dynamic model in (4). On average, the proposed algorithm reduced $T_{\max}$ by 54% over the "alternating algorithm" in the absense of currents and 63% with the presence of currents.

Consider one particular trial illustrated in Fig. 5 and Fig. 6 where ocean currents had values $v_c = 0.25$ m/s and $\psi_c = 0$ radians. The results for the different cases are presented in Table 2. For the case with no ocean currents, the "alternating algorithm" creates paths with numerous loops when two successive points are close together and the vehicle orientation does not allow for the second point to be reached without long maneuvers (Fig. 6(a)). This is avoided in the proposed algorithm by generating sequences that are feasible but limit the number of additional loops (Fig. 6(b)). Similar results are obtained with the presence of ocean currents as shown in Fig. 6(c) and Fig. 6(d).

Note that the proposed algorithm produced different sequences for the case with no ocean currents and the case with ocean currents. This is because the bidding scheme considers the possibility that two successive points that were reachable in the absence of ocean currents may no longer be reachable without extra loops due to the increase in turning radius from the ocean currents. Also, the paths generated by the proposed algorithm attempts to avoid paths that force the vehicles to drive against the ocean current. Instead, paths that allow the ocean current to aid the vehicle in the direction of travel are favoured.

For a sufficiently dense sets of points, it becomes clear that the ordering of the Euclidean tours are not optimal in the case of the Dubins MTSP. This is due to the fact that there is little relationship between the Euclidean and Dubins metrics, especially when the Euclidean distances are small with respect to the turning radius. An algorithm for the Euclidean problem will tend to schedule very close points in a successive order, which can imply long maneuvers for the AUV. This is clearly demonstrated by the numerous loops that become problematic with dense sets of points. The algorithm proposed in this paper does not rely on the Euclidean solution and therefore, even in the presence of ocean currents, can create paths that are feasible for curvature bound vehicles.

7. CONCLUSION

This paper addresses the task allocation of closely spaced targets for vehicles that follow paths of bounded curvature in the presence of constant ocean currents. The proposed algorithm is based on using a bidding scheme to allocate tasks to robots while using the Dubins set to calculate the cost of bids which consider both the vehicle constraints as well as the effect of ocean currents. Simulations with a non-linear model of the REMUS AUV indicate that the proposed algorithm yield better performance for dense sets of points when com-

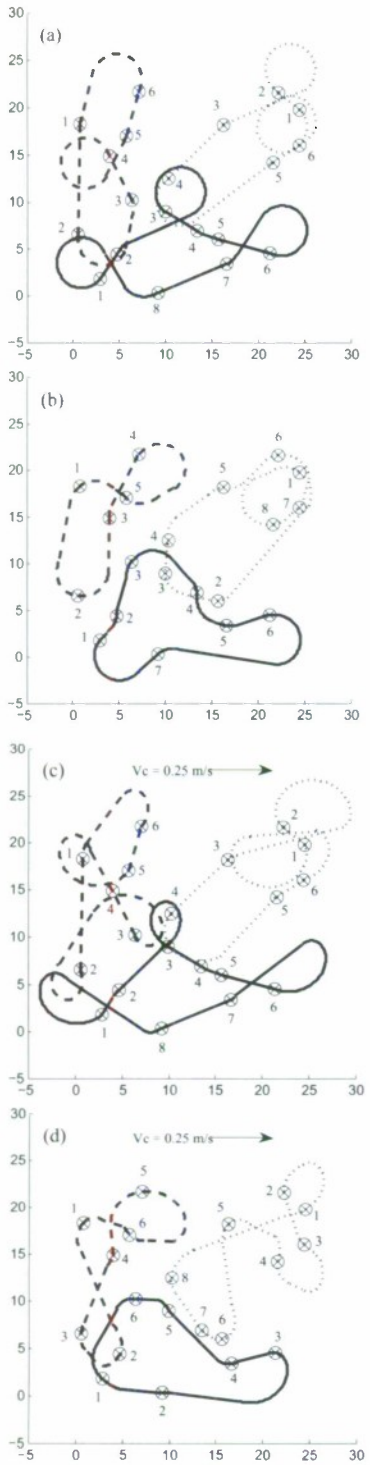


Figure 6. Simulation Results. (a) Alternating algorithm, $v_c = 0$, (b) Proposed algorithm, v_c= 0, (c) Alternating algorithm, v_c= 0.25 m/s, ψ_c= 0, (d) Proposed algorithm, v_c= 0.25 m/s, ψ_c= 0

pared to the "alternating algorithm". It is shown that solutions based on computing Euclidean tours that do not have curvature constraints have extra loops when task points are close together relative to the turning radius of the vehicle. Currently, this algorithm is being tested on an Iver2 AUV.

References

[1] L. E. Dubins, "On curves of minimum length with a constraint on average curvature and with prescribed initial and terminal position and tangents," *American J. Mathematics*, vol. 79, no. 3, pp. 497-516, Jul. 1957.

[2] T. Prestero, "Verification of a six-degree of freedom simulation model for the REMUS autonomous underwater vehicle," Master's thesis, Massachusetts Institute of Technology, Cambridge, 1994.

[3] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 3rd ed. Germany: Springer, 2006.

[4] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," *Int. J. Robotics Research*, vol. 23, no. 9, pp. 939-954, 2004.

[5] L. E. Parker, "ALLIANCE: An Architecture for Fault Tolerant Multi-Robot Cooperation," *IEEE Trans. Robotics and Automation*, vol. 14, no. 2, pp. 220-240, 1998.

[6] R. Zlot, A. Stentz, M. B. Dias, and S. Thayer, "Multi-robot exploration controlled by a market economy," in *Proc. IEEE Conf. Robotics and Automation*, vol.3, Washington, DC, pp. 3016-3023, 2002.

[7] S. Sariel, T. Balch, and J. Stack, "Distributed Multi-AUV Coordination in Naval Mine Countermeasure Missions," Georgia Institute of Technology, Atlanta, Georgia, 30332, Tech. Rep. GIT-GVU-06-04, 2006.

[8] S. Jeyaraman et al., "Formalised Hybrid Control Scheme for a UAV Group using Dubins Set and Model Checking," in *Proc. IEEE Conf. Decision and Control*, vol.4, Paradise Island, Bahamas, pp.4299-4304, 2004.

[9] R. E. Davis, N. E. Leonard, and D. M. Fratantoni, "Routing strategies for underwater gliders," *Deep-Sea Research II*, 2008.

[10] J. A. Hartigan and M. A. Wong, "A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100-108, 1979.

[11] M. Lagoudakis, M. Berhault, S. Koenig, P. Keskinocak, and A. Kleywegt, "Simple auctions with performance guarantees for multi-robot task allocation," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems*, vol. 1, pp. 698-705, 2004.

[12] A. M. Shkel and V. Lumelsky, "Classification of the Dubins set," *Robotics and Autonomous Systems*, vol. 34, no. 4, pp. 179-274, Mar. 2001.

[13] K. Savla, E. Frazzoli, and F. Bullo, "On the point-to-point and traveling salesperson problems for Dubins vehicle," *American Control Conference*, Portland, OR, pp. 786-791, Jun. 2005.

Assigning Closely Spaced Targets to Multiple Autonomous Underwater Vehicles

Beverley Chow and Jan Paul Huissoon
Department of Mechanical and Mechatronics Engineering
University of Waterloo
Waterloo, ON, N2L 3G1, Canada
E-mail: bchow@uwaterloo.ca; jph@uwaterloo.ca

Christopher Michael Clark Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA, 93407, USA
E-mail: cmclark@calpoly.edu

ABSTRACT

This research addresses the problem of allocating closely spaced targets to multiple autonomous underwater vehicles in the presence of constant ocean currents. The main difficultly of this problem is that the nonholonomic vehicles are constrained to move along forward paths with bounded curvatures. The Dubins model [1] is a simple but effective way to handle the kinematic characteristics of AUVS. It gives complete characterization of the optimal paths between two configurations for a vehicle with limited turning radius moving in a plane at constant speed.

In the proposed algorithm, Dubins paths are modified to include ocean currents, resulting in paths defined by curves whose radius of curvature is not constant. To determine the time required to follow such paths, an approximate dynamic model of the AUV is queried due to the computational complexity of the full model. The lower order model is built from data obtained from sampling the full model. The full model is used in evaluating the final tour times of the sequences generated by the proposed algorithm to validate the results.

The proposed algorithm solves the task allocation problem with market-based auctions that minimize the total travel time to complete the mission. The novelty of the research is the path cost calculation that combines a Dubins model, an AUV dynamic model, and a model of the ocean current. Simulations were conducted in MATLAB to illustrate the performance of the proposed algorithm using 20 task points and 3 AUVs. The task points were generated randomly and uniformly inside a square with side lengths of 22 meters. The task points were generated close together to highlight the necessity for considering the curvature constraints.

As a baseline for comparison, the "alternating algorithm" described by Savla et al. [2] is used for the creation of Dubins TSP tours (i.e. sequences from applying Dubins' model). It works as follows: given a set of n points, the optimal Euclidean MTSP tours (i.e. that do not consider path curvature) are computed using auctions. Then, it is necessary to obtain a feasible path through these ordered points using the method in [2] which includes the curvature constraints of the vehicle.

The proposed algorithm was tested on ten different datasets and compared to the "alternating algorithm". For a sufficiently dense sets of points, it becomes clear that the ordering of the Euclidean tours are not optimal in the case of the Dubins MTSP. This is due to the fact that there is little relationship between the Euclidean and Dubins metrics, especially when the Euclidean distances are small with respect to the turning radius. An algorithm for the Euclidean problem will tend to schedule very close points in a successive order, which can imply long maneuvers for the AUV. This is clearly demonstrated by the numerous loops that become problematic with dense sets of points. The algorithm proposed in this paper does not rely on the Euclidean solution and therefore, even in the presence of ocean currents, can create paths that are feasible for curvature bound vehicles.

Field tests were also conducted on an Iver2 AUV (Fig. 1) at the California Polytechnic State University's Center for Coastal Marine Sciences to validate the performance of the proposed algorithm. Missions created based on the sequences generated by the proposed algorithm were conducted to observe the ability of AUVs to follow paths of bounded curvature in the presenee of oeean currents.

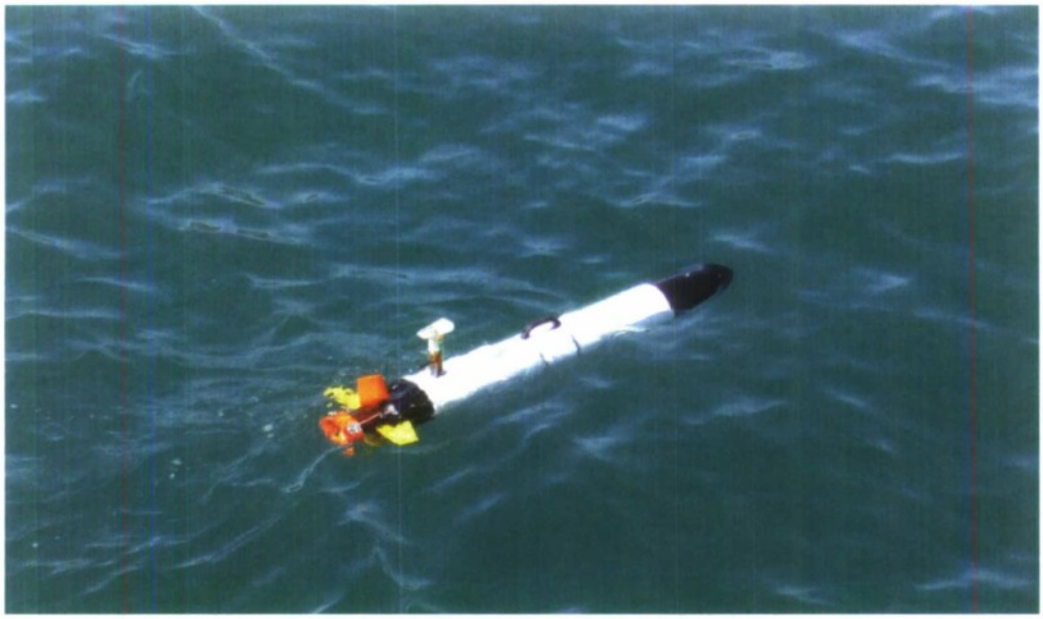


Figure 1. The Iver2 AUV in Avila Bay, CA after being launehed from Cal Poly's Center for Coastal Marine Scienee

REFERENCES

[1] L. E. Dubins, "On eurves of minimum length with a eonstraint on average curvature and with preseribed initial and terminal position and tangents," *American J. Mathematics*, vol. 79, no. 3, pp. 497-516, Jul. 1957.

[2] K. Savla, E. Frazzoli, and F. Bullo. "On the point-to-point and traveling salesperson problems for Dubins vehiele," *American Control Conference*, Portland, OR, pp. 786-791, Jun. 2005.

Algae Grown on Dairy and Municipal Wastewater for Simultaneous Nutrient Removal and Lipid Production for Biofuel Feedstock

I. Woertz[1]; A. Feffer[2]; T. Lundquist[3]; and Y. Nelson[4]

Abstract: Algae grown on wastewater media are a potential source of low-cost lipids for production of liquid biofuels. This study investigated lipid productivity and nutrient removal by green algae grown during treatment of dairy farm and municipal wastewaters supplemented with CO_2. Dairy wastewater was treated outdoors in bench-scale batch cultures. The lipid content of the volatile solids peaked at Day 6, during exponential growth, and declined thereafter. Peak lipid content ranged from 14–29%, depending on wastewater concentration. Maximum lipid productivity also peaked at Day 6 of batch growth, with a volumetric productivity of 17 mg/day/L of reactor and an areal productivity of 2.8 $g/m^2/day$, which would be equivalent to 11,000 L/ha/year (1,200 gal/acre/year) if sustained year round. After 12 days, ammonium and orthophosphate removals were 96 and >99%, respectively. Municipal wastewater was treated in semicontinuous indoor cultures with 2–4 day hydraulic residence times (HRTs). Maximum lipid productivity for the municipal wastewater was 24 mg/day/L, observed in the 3-day HRT cultures. Over 99% removal of ammonium and orthophosphate was achieved. The results from both types of wastewater suggest that CO_2-supplemented algae cultures can simultaneously remove dissolved nitrogen and phosphorus to low levels while generating a feedstock potentially useful for liquid biofuels production.

DOI: 10.1061/(ASCE)EE.1943-7870.0000129

CE Database subject headings: Biomass; Wastewater management; Nutrients; Carbon dioxide.

Introduction

Biofuels produced from plants have the potential to replace a significant fraction of our fossil fuel needs with a renewable alternative (Perlack et al. 2005). However, concern has grown that the use of food crops for production of ethanol, biodiesel, or other renewable fuels will increase food prices while having little impact on greenhouse gas emissions (Fargione et al. 2008). Prior work, in particular the Aquatic Species Program sponsored by the U.S. Department of Energy, suggested that algae are capable of producing oil suitable for conversion to biodiesel with an areal productivity 20–40 times that of oilseed crops, such as soy and canola (Sheehan et al. 1998). However, an economic study of such processes (Benemann and Oswald 1996) suggested that large-scale algae cultivation solely for biofuel production was not economical, and the writers reemphasized the integration of biofuels production and wastewater treatment with CO_2 supplementation, as first suggested by Oswald and Golueke (1960). In particular, assimilation of wastewater nutrients by algae followed by algae harvesting via sedimentation were considered potentially practical and economical approaches to biofuel production.

Use of algae for municipal wastewater treatment in ponds is well established (Oswald et al. 1953; Oswald 2003), and algae-based treatment of dairy and piggery waste also has been investigated (e.g., Craggs et al. 2004; Kebede-Westhead et al. 2006; Mulbry et al. 2008; An et al. 2003). Algae growth in wastewater treatment ponds contributes to treatment mainly through dissolved oxygen production and nutrient assimilation. However, the carbon:nitrogen and carbon:phosphorus ratios in domestic sewage (C:N 3.5:1; C:P 20:1) and dairy lagoon water (C:N 3:1; C:P 10:1) are low compared to typical ratios in rapidly growing algae biomass (C:N 6:1; C:P 48:1) (Metcalf and Eddy 2003; USDA 1992; Oswald 1960). This dearth of carbon leads to limitations in algae production and incomplete assimilation of wastewater nutrients by algae. The experiments described in the present research overcame the carbon limitation of the wastewaters by addition of CO_2 to the cultures. The effects of this addition on both algae growth and nutrient assimilation were measured. In future applications, CO_2 could be supplied by the flue gas from power plants and other sources. A schematic of one envisioned process is shown in Fig. 1.

CO_2 supplementation of algae cultures to increase productivity has been studied for many years (Burlew 1953), as has the use of flue gas as a CO_2 source for algae culture (Straka et al. 2000). CO_2 supplementation to promote nutrient removal has also been studied briefly in outdoor ponds (Benemann et al. 1980). However, the production of lipids was not measured in these studies.

Lipid content for pure cultures of algae has been reported to range from 1–85%, and the lipids exhibit varying carbon chain lengths, degrees of unsaturation, and polarity [e.g., reviews in

[1]Research Engineer, Dept. of Civil and Environmental Engineering, California Polytechnic State Univ., 1 Grand Ave., San Luis Obispo, CA 93407 (corresponding author). E-mail: iwoertz@calpoly.edu

[2]Project Coordinator, LifeWater International, Inc., 3563 Empleo St., Suite C, San Luis Obispo, CA 93401.

[3]Assistant Professor, Dept. of Civil and Environmental Engineering, California Polytechnic State Univ., 1 Grand Ave., San Luis Obispo, CA 93407.

[4]Professor, Dept. of Civil and Environmental Engineering, California Polytechnic State Univ., 1 Grand Ave., San Luis Obispo, CA 93407.

Note. This manuscript was submitted on June 17, 2008; approved on July 21, 2009; published online on July 23, 2009. Discussion period open until April 1, 2010; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Environmental Engineering***, Vol. 135, No. 11, November 1, 2009. ©ASCE, ISSN 0733-9372/2009/11-1115–1122/$25.00.

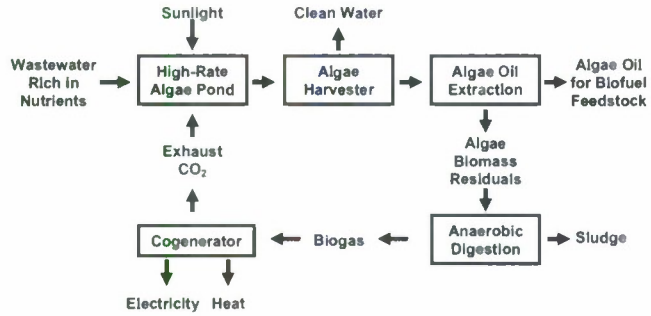


Fig. 1. Simplified process flow diagram envisioned for algae wastewater treatment and liquid biofuel production

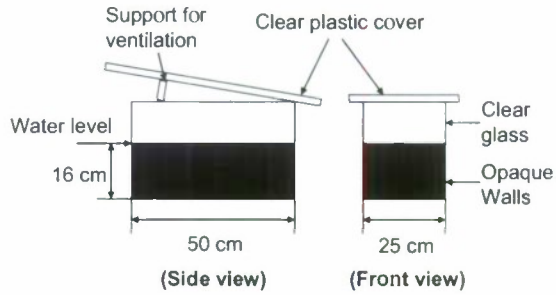


Fig. 2. Outdoor algae growth tanks for batch experiments with dairy wastewater

Chisti (2007), Metting (1996), and Enssani (1987)]. However, the lipid content and, more important, the lipid productivity of polycultures of algae, such as wastewater pond algae, have seldom, if ever, been reported. Further, lipid content, fatty acid profile, and biomass productivity depend on environmental conditions, culturing methods, and growth phase (Thompson 1996; Tsuzuki et al. 1990). In particular, nitrogen limitation increases lipid content in some species (Spoehr and Milner 1949; Leman 1997). However, nitrogen limitation decreases growth rate, which can lead to decreased overall lipid productivity (Shifrin and Chisholm 1981). Benemann and Tillett (1987) investigated this problem, but maximizing lipid productivity remains an outstanding problem.

While a few studies have reported the lipid *content* of waste-grown algae cultures [e.g., 25%, Enssani (1987)], lipid *productivities* for waste-grown polycultures apparently have not been reported previously. The research presented herein was conducted to determine the lipid content and lipid productivity of microalgae grown for nutrient removal from two types of wastewater—dairy and municipal.

Methods

Overview of Experiments

Two sets of experiments were run in parallel to determine algae growth, nutrient removal, and lipid productivity in municipal wastewater and dairy wastewater. The municipal wastewater experiment monitored algae growth under semicontinuous operation for 18 days to study the effects of CO_2 levels and hydraulic residence times (HRTs) on algae growth and nutrient removal. Control cultures with addition of air only (no CO_2) were used to simulate the carbon limitation typical of wastewater ponds and to differentiate the effect of CO_2 addition on productivity. In the dairy wastewater experiment, lipid productivity and nutrient removal were monitored during 15 days of batch growth to study the effect of the growth cycle on lipid content.

Collection and Pretreatment of Wastewater

For the municipal wastewater experiments, 60 L of primary clarifier effluent was collected at the San Luis Obispo, California, municipal wastewater treatment facility. The wastewater was mixed thoroughly and passed through screens with 196-μm openings (commercial house paint filters). The screened wastewater was stored in 4-L HDPE containers at −10°C.

For the dairy wastewater experiments, free-stall barn flush water was collected from a storage pond at the 400-head Cal Poly Dairy in San Luis Obispo. In flush operations at this dairy, the flush water is passed through a bar screen and sand trap and then collected in a covered sump. The wastewater is then pumped over a wedge-wire inclined screen, which removes feed and other fine solids, before being discharged to the 0.5-ha storage pond. Wastewater for the research was collected from the storage pond and then treated in an anaerobic digester before being used in algae growth experiments. The 130-L digester was unheated, unmixed, and fed semicontinuously to achieve a 6-week hydraulic residence time.

Municipal Wastewater Experimental Procedures

Eight 1-L Pyrex Roux bottles (Fisher Scientific, Rockford, Ill.) were used as algae growth reactors for the municipal wastewater experiments. Each bottle was placed vertically on a magnetic stirrer and mixed by a polyetrafluoroethylene (PTFE)-coated 2.5-cm magnetic stir bar spinning at approximately 300 rpm. The bottles were illuminated from two sides by a total of four 40-W full-spectrum fluorescent bulbs (Duro-test Vitalite Lighting, Inc., Philadelphia) operated on a 16 h:8 h light:dark cycle. When on, the bulbs provided an average illuminance that totaled 4,300 lx at the two faces of each bottle (Lutron LX-101 m), which is equivalent to about 12 W/m^2 of photosynthetically active radiation (Li-Cor 2008).

To provide gas exchange, each bottle was sparged with either air or a CO_2-air mixture through a cylindrical polypropylene diffuser 36 mm long and 9 mm in diameter. Gas was delivered to the diffusers through a manifold of 4-mm inside diameter, clear vinyl tubing with manual flow control valves. For bottles sparged with air-CO_2 mixtures, gases were mixed by a gas mixer (Model 665, Matheson Co. Inc., USA) that was connected to a 50-lb tank of 99.97% CO_2 and a 2.5-psi aquarium air pump (Maxima Air Pump Model A-805, Hagen Corp., Mansfield, Mass.). The CO_2 concentration in the blend was set to maintain pH between 7.0 and 8.0. For the Roux bottles that were sparged with air alone, the 4-mm tubing was connected to a 3-W aquarium air pump (Profile 1500, Meiko Pet Corp., Taiwan).

The culture volume in each Roux bottle was 800 mL. Wastewater was introduced into the bottles with a daily draw-fill procedure at the end of the light period. Three daily hydraulic loading rates were tested—200, 267, and 400 mL of primary effluent— in order to achieve 4-, 3-, and 2-day HRTs, respectively. For the air-CO_2 sparged treatments, each HRT was run in duplicate. For the air-only treatment, the 3-day HRT was run in duplicate. Culture media temperature ranged from 23 to 25°C and did not vary between the bottles more than 1.5°C.

Dairy Wastewater Experimental Procedures

For the dairy wastewater, algae were cultured outdoors in six 40-L rectangular glass aquarium tanks (Fig. 2). The tanks were filled with 20 L of effluent from the anaerobic digester. To better simulate light conditions in ponds, sunlight was allowed to enter the tanks only through the top water surface by masking the tank walls up to the waterline with black tape. A Plexiglas cover excluded rainfall, but a gap was provided between the cover and the tanks for ventilation.

In preliminary experiments with undiluted dairy wastewater algal growth was poor, presumably due to the high opacity of the wastewater. Therefore, the subsequent experiments reported here used 10 and 25% wastewater diluted with tap water. Each experimental treatment was run in triplicate with air sparging at 1.5 L/min for mixing and separate, simultaneous pure CO_2 sparging at approximately 0.015 L/min, which controlled culture pH.

The experiment was run during March 2007 when the average daily solar radiation was 203 W/m^2 (California Irrigation Management Information System Station #52). The average water temperature, measured daily at 3:00 p.m., was 30.6°C. Water samples were collected between 3:00 and 3:30 p.m.

Inoculation

Algae inoculum was collected from local ponds treating municipal or winery wastewater and from a creek. The inoculum samples contained a wide-ranging mixture of green algae and diatoms, which were identified by cell morphology using phase-contrast microscopy with reference to Presscott et al. (1978). Prominent genera included *Actinastrum*, *Scenedesmus*, *Chlorella*, *Spirogyra*, *Nitzschia*, *Micractinium*, *Golenkinia*, *Chlorococcum*, *Closterium*, *Euglena*, and two unidentified species. The municipal culture inoculum contained 625 mg/L volatile suspended solids (VSS) and was added to the wastewater media in a 2% (v/v) ratio. For the dairy cultures, the inoculum concentration was 500 mg/L VSS, added at a 10% (v/v) ratio.

Water Quality Analyses and Lipid Extraction

VSS concentrations were determined gravimetrically according to Standard Methods (APHA 2005). Temperature and pH were monitored to characterize growth conditions. Nutrient removal was evaluated by analyzing for nitrite, nitrate, and orthophosphate using a Dionex DX 120 ion chromatograph with an AG9-HC IonPac Guard Column, AS9-HC 4-mm IonPac IC column, DS4-1 Detection Stabilizer, and an AS40 Automated Sampler. Total ammonia nitrogen ($NH_3+NH_4^+-N$) concentrations were determined using the Ammonia-Selective Electrode Method (APHA 4500-NH_3D). Organic nitrogen was determined using the Macro-Kjeldahl method (APHA Method 4500-N_{org}).

To complete a nitrogen balance for the Roux bottle experiments, it was necessary to quantify the volatilization of ammonia. This quantity was determined by passing the sparged gas through a boric acid solution. This procedure was conducted for one Roux bottle of each duplicate. A two-hole stopper with 4-mm tubing allowed sparging gas in and directed sparged gas out and into the boric acid solution through a polypropylene diffuser. The diffuser was submersed 12 cm in a graduated cylinder under 100–200 mL of boric acid indicating solution (APHA 4500-NH_3C. 3.b.). At the end of a mass balance period, deionized (DI) water was added to the graduated cylinder to compensate for evaporation. This solution was then titrated back to its original pH, and its ammonia concentration was calculated according to APHA 4500-NH_3C.

The lipid content of the VSS was analyzed gravimetrically by a procedure adapted from Bligh and Dyer (1959) by Benemann and Tillett (1987). The method consisted of solvent-based extraction to isolate both polar and nonpolar lipids from cell biomass and water. The VSS of each sample was measured to determine the concentration of algal biomass in the wastewater effluent. A 200-mL aliquot of the same sample was centrifuged in a PTFE tube to form an algae pellet for lipid extraction. After decanting, the pellet was resuspended in 4 mL of DI water and frozen until extraction. For extraction, the samples were thawed, and 5 mL of chloroform and 10 mL of methanol were added. The samples were then sonicated continuously in the centrifuge tube for 1 min. (Branson Sonifier 250 with a Model #102 tip). The samples were then placed on a shaker table overnight. The next day an additional 5 mL of chloroform and 5 mL of DI water were added to make the final ratio of chloroform:methanol:water 10:10:9. The samples were then vortexed for 30 s. After the samples had been homogenized, they were centrifuged at 7,000 rpm for 4 min. The lipids were soluble in the chloroform, which formed a dense layer at the bottom of the centrifuge tube. The remaining cell debris created a middle layer, while the methanol and water created a top layer. The lipid-chloroform layer was removed with a pipette and filtered through a 0.2-μm nylon syringe filter. The filtrate was deposited into a tared aluminum tray. The tray was then placed into a desiccator flushed with nitrogen to allow the chloroform to evaporate. A second extraction was performed by adding an additional 10 mL of chloroform to the centrifuge tube, and the mixture was again vortexed and centrifuged. This second extraction was placed into a separate tared tray and evaporated under nitrogen. The trays were then dried at 105°C for 1 h. After cooling in a desiccator, the trays were weighed to the nearest 0.01 mg. Adding the weights of the two extractions from each sample gave the total lipid weight.

Table 1. Initial Wastewater Characteristics

	Dairy wastewater		Municipal wastewater
Wastewater characteristics	25% dilution	10% dilution	No dilution
TSS (mg/L)	283	135	93
VSS (mg/L)	220	120	58
pH	7.9	7.7	7.2
Ammonium as N (mg/L)	30.5	16.3	39
Nitrate as N (mg/L)	<0.01	0.05	<0.01
Nitrite as N (mg/L)	<0.01	0.04	<0.01
Organic nitrogen (mg/L)	50.7	20.2	12
TKN (mg/L)	81.0	36.5	51
Total nitrogen (mg/L)	81.0	36.6	51
Phosphate as P (mg/L)	2.6	1.8	2.1

Results and Discussion

Influent Wastewater Characteristics

The municipal primary wastewater characteristics, as well as the initial conditions in the dairy wastewater bioreactors immediately after inoculation are reported in Table 1. After dilution, the 25%

Table 2. Culture Conditions in Dairy Wastewater Experiment

Day	Insolation (W/m^2)	Air temperature (°C) Max	Min	Average	Water temperature at 3 p.m. (°C)	Average pH[a] 10% dairy wastewater	25% dairy wastewater
0					32	7.7	7.9
1	50	13.6	9.9	11.4	15	7.4	7.5
2	191	17.3	7.3	11.3	30	7.2	7.1
3	228	24	11.4	16.3	36	8.9	7.4
4	228	20.9	7.5	13.1	35	9.3	7.6
5	70	14.4	9.5	12	17	7.0	7.3
6	212	17.9	10.6	13.1	32	6.5	7.1
7	169	17.8	8.6	11.6	29	7.3	8.4
8	226	15.8	4.5	9.1	27	8.6	9.5
9	246	16.3	4.5	10.4	32	7.5	7.7
10	252	22.3	4.5	12.9	37	6.4	6.4
11	247	24.2	5.1	13	34	7.1	7.3
12	246	20.9	5	11.4	36	8.3	8.0
13	242	22.1	7.9	13.4	37		7.3
14	239	20.8	7.4	13.2			
15	246	24.1	7.8	13.5			

[a]Standard deviation of replicates ranged from 0.0 to 0.5.

dairy wastewater had ammonium and orthophosphate concentrations similar to the undiluted municipal wastewater.

Culture Conditions

The laboratory municipal wastewater cultures were grown under steady conditions as described in the Methods section. However, the outdoor dairy wastewater cultures experienced widely varying conditions both daily and over the course of the experiments (Table 2). The average 24 h insolation ranged from 50–252 W/m^2. Due to manual adjustment of CO_2 flow, pH ranged from 6.5–8.9. The water temperatures in all the tanks were similar and reached as high as 37°C (Table 2).

Algal and Lipid Productivity

The semicontinuous-flow experiments with municipal wastewater reached nearly steady-state biomass concentrations after 11 days of operation, although VSS was higher on the 18th day, when lipid samples were taken (Fig. 3). For the 3-day HRT cultures, sparging with CO_2 more than doubled the VSS concentration compared to sparging with air. For the treatments with CO_2 sparging, biomass production was similar for the 3- and 4-day HRTs, with steady-state VSS concentrations of 700–800 mg/L. In contrast, the steady-state VSS concentration for the CO_2-sparged 2-day HRT treatment was only 300 mg/L. The municipal wastewater cultures were dominated by algae in the *Chlorella*, *Micractinium*, and *Actinastrum* genera.

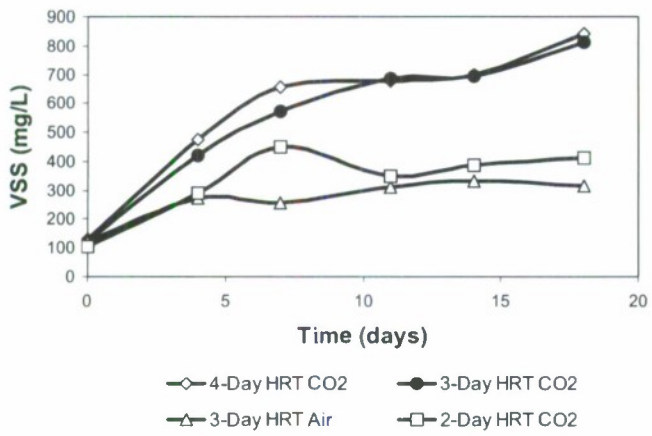


Fig. 3. Biomass concentrations during semicontinuous flow treatment of municipal wastewater (mean of duplicates)

The lipid contents of the algae from the municipal wastewater experiments ranged from 4.9–11.3% of VSS by weight (Table 3). Despite the relatively low lipid contents observed, short residence times and high biomass production rates resulted in lipid productivities ranging from 9.7 mg/L/day (air-sparged) to 24 mg/L/day (CO_2-sparged 3-day HRT).

Lipid production using dairy wastewater was measured in batch experiments with two different dilutions of wastewater (10% and 25%). Biomass concentrations increased to maximum values of 500 mg/L VSS at Day 6 for the 10% dilution (Fig. 4) and 900 mg/L VSS at Day 13 for the 25% dilution (Fig. 5). The higher biomass production for the 25% dilution was likely due to the higher nutrient concentrations (Table 1). The dairy wastewater cultures were dominated by *Scenedesmus*, followed by *Micractinium*, *Chlorella*, and *Actinastrum*. These were the same genera that dominated in the municipal wastewater cultures, except that *Scenedesmus* was absent in the municipal cultures.

For both dairy wastewater dilutions, the highest lipid content

Table 3. Lipid Productivity of Municipal Wastewater Cultures

Sample	VSS (mg/L)	Lipids (%)	Lipid content of culture medium (mg/L)	Lipid productivity (mg/L/day)
CO_2 4-day HRT	843	4.9	41.5	10.4
CO_2 3-day HRT	812	9.0	73.3	24.4
Air 3-day HRT	317	9.3	29.2	9.7
CO_2 2-day HRT	412	11.3	46.2	23.1

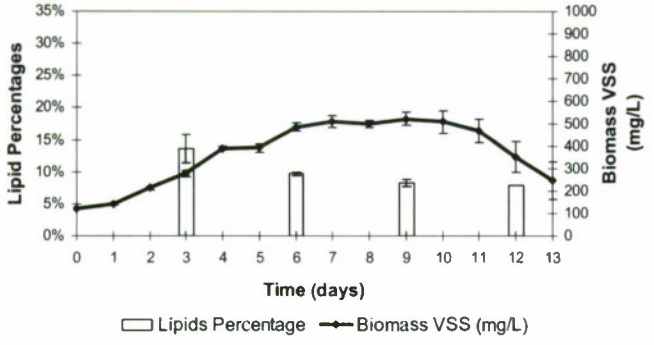


Fig. 4. Biomass concentration and cell lipid content during batch algae growth on 10% dairy wastewater (mean of triplicates)

was observed during the exponential growth phase, and it declined thereafter (Figs. 4 and 5). The total lipid content of biomass from the 10% dilution ranged from 8–14%, and that of the 25% dilution ranged from 10–29% by weight. In comparison, total lipid content of pure *Scenedesmus* and *Chlorella* cultures has been reported to range from 12–45% (Thompson 1996).

For the dairy wastewater experiments, the maximum lipid production rate was 17 mg/L/day on a volumetric basis or 2.8 $g/m^2/day$ on an area basis, achieved by Day 6 for the 25% dilution. In comparison, previous research with open-surface systems growing pure cultures have shown somewhat higher production rates ranging from 4–7.9 $g/m^2/day$ [Table 4 (Laws 1984; Thomas 1984; Brown 1990)].

In both the batch and semicontinuous experiments, peak lipid content was associated with high biomass growth rates. For the dairy wastewater experiments, the highest lipid content and productivity was achieved during exponential growth for both the 10 and 25% dilution experiments, rather than during later phases when nutrient concentrations were low (Figs. 4, 5, and 7). For the indoor municipal wastewater experiment, the highest lipid content (11%) was observed at the shortest HRT (2 days). In these semicontinuous-flow experiments, the shorter retention time corresponded to more rapid biomass growth and greater lipid productivity. Thus high lipid production was associated with rapid growth for both batch dairy and semicontinuous municipal wastewater experiments. Roessler (1990) has discussed similar results of increased lipid content in the exponential growth phase of microalgae and theorized that at lower biomass concentrations with less self-shading, algae biosynthesize lipid storage products as a means of capturing excess light energy. In contrast, others have found higher lipid content in cultures that were nutrient limited (Leman 1997; Spoehr and Milner 1949).

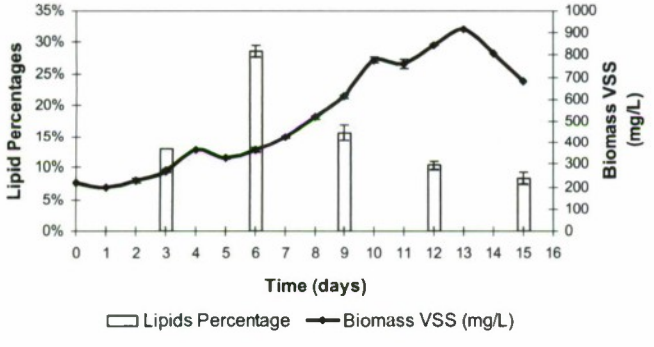


Fig. 5. Biomass concentration and cell lipid content during batch algae growth on 25% dairy wastewater

The maximum observed lipid productivity of the dairy waste reactors (2.8 $g/m^2/day$) corresponds to about 11,000 L/ha/year (1,200 gal/acre/year). Without improvements, productivity in full-scale high-rate algae ponds is expected to be lower due to factors such as winter insolation and temperature, predation, maintenance downtime, and shifts in algal strains. For example, assuming 300 days/year of operation, the productivity would be reduced to 9,000 L/ha/year (960 gal/acre/year). An additional uncertainty in scale-up estimates stems from the difference in operational modes for the dairy wastewater experiments (batch) and typical high-rate pond wastewater treatment (continuous). Theoretically, the maximum growth rate achieved in batch culture could also be achieved in continuous flow culture (Gualtieri and Barsanti 2005). Of course, the actual productivity for a full-scale system will depend on local environmental conditions, cultivation parameters, dominant algal strains, etc. Furthermore, the suitability of the algal lipids for fuel production will depend on the lipid characteristics (e.g., polarity, saturation level, and chain length) and the ease of extraction.

Much higher algal lipid productivities have been envisioned [e.g., 42,600–136,900 L/ha/year, Chisti (2007)] than were observed in this study. However, even this study's oil production estimate of 9,000 L/ha/year is 18 times greater than the 490 L/ha/year reported for soybean oil production (USDA 2005).

Nutrient Removal

For the municipal wastewater, over 99% ammonium and orthophosphate removal was achieved for CO_2-sparged treatments with both 3- and 4-day HRT (Table 5). To determine the fate of the removed ammonium and to validate the results, a nitrogen balance was calculated on four occasions over 10 days of operation. The results were similar on all 4 days, and Fig. 6 shows the balance for Day 18. The average recovery achieved was 96% with a standard deviation of 8.7%. Ammonium was the main form of nitrogen in the influent wastewater, and after algal growth, organic nitrogen was predominant (Fig. 6). Ammonia volatilization was minor, the greatest amount being <1 mg/bottle/day from the air-sparged treatment, which accounts for $<7\%$ of the influent total nitrogen. Since this treatment developed the highest pH (10.3) due to lack of CO_2 sparging, it was the most prone to

Table 4. Comparison of the Lipid Productivity of Dairy Wastewater Cultures to That Reported by Others

Study	Lipid productivity ($g/m^2/d$)	Algal species	Growth vessel	Medium
Laws (1984)	7.9	*Platymonas* sp.	Air lift flume	Sea water
Thomas (1984)	4.5	*Tetraselmis suecica*	Indoor reactor	Nutrient enriched seawater
Brown (1990)	4	*Cylcotella cryptica*	Open pond	Si deficient media
This study	2.8	Polyculture	Open reactor	Anaerobic treated dairy wastewater

Table 5. Nutrient Removal by Municipal Wastewater Cultures

	Total ammonia nitrogen (mg/L)			Phosphate as P (mg/L)		
	Influent	Effluent[a]	% Removal	Influent	Effluent[a]	% Removal
CO_2 4-day HRT	39.0	<0.02	>99%	2.1	<0.02	>99%
CO_2 3-day HRT	39.0	<0.02	>99%	2.1	<0.02	>99%
Air 3-day HRT	39.0	6.1 (±0.89)	84%	2.1	<0.02	>99%
CO_2 2-day HRT	39.0	0.6 (±0.57)	98%	2.1	0.15 (±0.15)	93%

[a]Mean of duplicate reactors with standard deviation shown in parentheses.

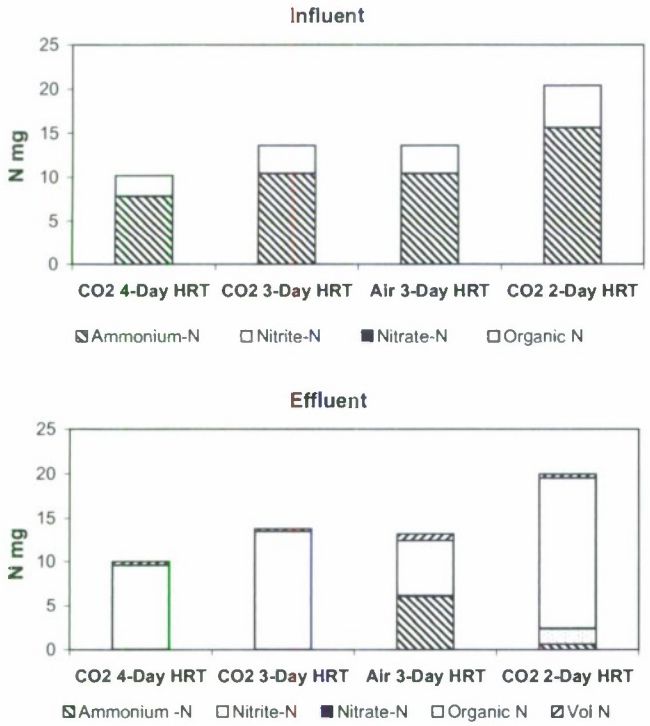


Fig. 6. Nitrogen balance for municipal wastewater cultures on Day 18 (means of duplicates). "Vol N" is volatilized nitrogen captured in a boric acid solution.

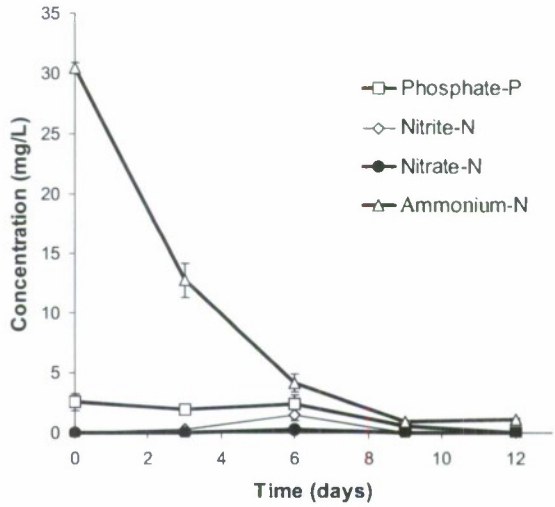


Fig. 7. Nutrient removal during batch culture (triplicates) on 25% dairy wastewater

ammonia volatilization. The nitrite observed in the 2-day HRT effluent (Fig. 6) indicates incomplete nitrification of ammonia for this short retention time.

Removal of ammonium and orthophosphate from the batch dairy wastewater was 96% and >99%, respectively by Day 15 (Table 6). For the 25% dilution experiment, initial concentrations of total ammonia nitrogen were 30 mg/L and were reduced to <5 mg/L in 6 days (Fig. 7). The initial orthophosphate phosphorus concentration of 2.6 mg/L was reduced to 0.6 mg/L in 9 days, and it was completely removed by Day 12. Nitrate concentrations were consistently below 0.3 mg/L for both conditions, and final nitrate concentrations were below the detection limit of 0.02 mg/L NO_3^-–N. Similar results were observed with the 10% dairy wastewater dilution. Nitrite showed a slight increase at Day 6 up to 0.5 mg/L NO_2^-–N, indicating some nitrification. Similar or higher ammonium removal efficiencies were observed by other researchers for algae-based treatment (Table 7).

Conclusions

This research provided a proof of concept for a wastewater treatment process that combines nutrient removal and algal lipid production for potential use as a biofuel feedstock. CO_2 supplementation was used to accelerate treatment and growth in both outdoor and indoor mixed-species cultures. Ammonium and orthophosphate removals were nearly complete for both municipal wastewater and diluted dairy wastewater. This study also contributed data on both the lipid content and the lipid productivity of wastewater-grown algae, a rarely addressed topic. Lipid content ranged from 4.9–29%, and lipid productivity reached 2.8 g/m^2/day. While this lipid productivity is many times higher

Table 6. Nutrient Removal in Dairy Wastewater Experiment at Day 15

	Total ammonia nitrogen (mg/L)[a]			Phosphate as P (mg/L)[a]		
	Influent	Effluent	% removal	Influent	Effluent	% removal
25% dilution	30.5 (±0.4)	1.1 (±0.1)	96%	2.6 (±0.7)	<0.02	>99%
10% dilution	16.3 (±4.8)	0.6 (±0.1)	96%	1.8 (±0.01)	<0.02	>99%

[a]Average of triplicate reactors (standard deviation, $n=3$).

Table 7. Comparison of Total Ammonia Nitrogen Removal to That Reported for Other Algae Treatment Systems

Study	% total ammonia-N removal	Algae species	Medium
Martinez et al. (2000)	80–99	*Scenedemus obliques*	Autoclaved municipal wastewater
Lincoln et al. (1996)	99	*Arthrouspira plantensis*	Anaerobically treated dairy wastewater
Green et al. (1995)	99	Polyculture	Municipal wastewater
This study	99	Polyculture	Municipal wastewater
This study	96	Polyculture	Anaerobically treated dairy wastewater

than that of terrestrial oil plants, higher productivity is a goal of continuing research. In addition, the suitability of the lipids for fuel production by transesterification and other means needs to be determined. Overall, the waste-to-biofuel approach of this study avoids many of the cost and food competition issues of other biofuel feedstocks while providing a valuable wastewater treatment service.

Acknowledgments

Funding was provided by a U.S. Environmental Protection Agency "People, Prosperity, and the Planet" grant, as well as a U.S. Department of Energy Small Business Innovation Research grant to MicroBio Engineering, Inc. Lundquist was supported in part by the U.S. Office of Naval Research via the California Central Coast Research Partnership. The writers are grateful for consultation provided by Dr. John Benemann.

References

An, J. Y., Sim, S. J., Lee, J. S., and Kim, B. (2003). "Hydrocarbon production from secondary treated piggery wastewater by the green alga *Botryococcus braunii*." *J. Appl. Phycol.*, 15, 185–191.

APHA. (2005). "Standard methods for the examination of water and wastewater." American Public Health Association, Washington, D.C., American Water Works Association, Denver, and the Water Environment Federation, Alexandria, Va.

Benemann, J., and Tillett, D. (1987). "Effects of fluctuating environments on the selection of high yielding microalgae." *Final Rep. Prepared for the Solar Energy Research Institute*, Schools of Applied Biology and Chemical Engineering, Georgia Institute of Technology, Atlanta.

Benemann, J. R., Koopman, B. L., Weissman, J. C., Eisenberg, D. M., and Goebel, R. (1980). "Development of microalgae harvesting and high rate pond technologies in California." *Algae biomass: Production and use*, G. Shelef and C. J. Soeder, eds. Elsvier North, Amsterdam, The Netherlands, 457–496.

Benemann, J. R., and Oswald, W. J. (1996). "Systems and economic analysis of microalgae ponds for conversion of CO_2 to biomass." *Final Rep. to the US Department of Energy Pittsburgh Energy Technology Center, Rep. No. DE-FG22–03PC93204*, Dept. of Civil Engineering, Univ. of California Berkeley, Berkeley, Calif.

Bligh, E., and Dyer, W. (1959). "A rapid method for total lipid extraction and purification." *Can. J. Biochem. Physiol.*, 37, 911–917.

Brown, L. M., and Sprague, S. (1990). "Design and operation of an outdoor microalgae test facility: Large scale system results." *Auqatic Species Project Rep. FY 1989–90 NREL/TP-232-4174*, reported by J. C. Weissman and D. Tillet, Microbial Products, Inc., Vero Beach, Fla., 32–56.

Burlew, J. S. (1953). "Algal culture: From laboratory to pilot plant." *Carnegie Institution of Washington Publication 600*, Washington, D.C.

Chisti, Y. (2007). "Biodiesel from microalgae." *Biotechnol. Adv.*, 25, 294–306.

Craggs, R. J., Sukias, J. P., Tanner, C. T., and Davies-Colley, R. J. (2004). "Advanced pond system for dairy-farm effluent treatment." *N. Z. J. Agric. Res.*, 47, 449–460.

Enssani, E. (1987). "Fundamental parameters in extraction of lipids from wastewater-grown microalgal biomass." Ph.D. thesis, Dept. of Civil Engineering, Univ. of California, Berkeley.

Fargione, J., Hill, J., Tilman, D., Polasky, S., and Hawthorne, P. (2008). "Land clearing and the biofuel carbon debt." *Science*, 319, 1235–1238.

Green, F. B., Lundquist, T. J., and Oswald, W. J. (1995). "Energetics of advanced integrated wastewater pond systems." *Water Sci. Technol.*, 31(12), 9–20.

Gualtieri, P., and Barsanti, L. (2005). *Algae: Biochemistry, physiology, ecology, and biotechnology*, CRC, Boca Raton, Fla.

Kebede-Westhead, E., Pizarro, C., and Mulbry, W. (2006). "Treatment of swine manure effluent using freshwater algae: Production, nutrient recovery, and elemental composition of algal biomass at four effluent loading rates." *J. Appl. Phycol.*, 18(1), 41.

Laws, E. (1984). "Research and development of shallow algal mass culture systems for the production of oils." *Rep. prepared for the Solar Energy Research Institute, Rep. No. XK-3–03136*, Univ. of Hawaii, Honolulu.

Leman, J. (1997). "Oleaginous microorganisms: An assessment of the potential." *Adv. Appl. Microbiol.*, 43, 195–243.

Li-Cor. (2008). *Principles of radiation measurement. v1.0-LI-COR*, Li-Cor, Lincoln, Neb.

Lincoln, E. P., Wilkie, A. C., and French, B. T. (1996). "Cyanobacterial process for renovating dairy wastewater." *Biomass Bioenergy*, 10(1), 63–68.

Martinez, M. E., Sanchez, S., Jimenez, J. M., El Yousfi, F., and Munoz, L. (2000). "Nitrogen and phosphorus removal from urban wastewater by the microalga *Scenedesmus obliquus*." *Bioresour. Technol.*, 73, 263–272.

Metcalf and Eddy. (2003). *Wastewater engineering: Treatment and reuse*, 4th Ed., McGraw-Hill, New York.

Metting, F. B. (1996). "Biodiversity and application of microalgae." *J. Ind. Microbiol.*, 17, 477–489.

Mulbry, W., Kondrad, S., and Buyer, J. (2008). "Treatment of dairy and swine manure effluents using freshwater algae: Fatty acid content and composition of algal biomass at different manure loading rates." *J. Appl. Phycol.*, 20(6), 1079-1085.

Oswald, W. J. (1960). "Fundamental factors in stabilization pond design." *Proc., 3rd Conf. Biological Waste Treatment*, Manhattan College, New York.

Oswald, W. J. (2003). "My sixty years in applied algology." *J. Appl. Phycol.*, 15, 99–106.

Oswald, W. J., and Golueke, C. G. (1960). "Biological transformation of solar energy." *Advances in applied microbiology*, Vol. 2, W. W. Umbreit, ed., Academic, New York, 223–262.

Oswald, W. J., Gotaas, H. B., Ludwig, H. F., and Lynch, V. (1953). "Algae symbiosis in oxidation ponds: Photosynthetic oxygenation." *Sewage Ind. Waste.*, 25(6), 692–705.

Perlack, R.D., Wright, L.L., Turhollow, A.F., Graham, R.L., Stokes, B.J. and Erbach, D.C. (2005). "Biomass as feedstock for a bioenergy and bioproducts industry: The technical feasibility of a billion-ton annual supply." *Rep. No. DOE/GO-102005–2135*, Oak Ridge National Laboratory, Oak Ridge, Tenn.

Presscott, G., Bamrick, J., Cawley, E., and Jaques, W. (1978). *How to know the freshwater algae*, W. C. Brown, ed., Dubuque, Iowa.

Roessler, P. (1990). "Environmental control of glycerolipid metabolism in microalgae: Commercial implications and future research directions." *J. Phycol.*, 26, 393–399.

Sheehan, J., Dunahay, T., Benemann, J., and Roessler, P. (1998). *A look back at the U.S. Department of Energy's aquatic species program—Biodiesel from algae*, National Renewable Energy Laboratory, Golden, Colo.

Shifrin, N., and Chisholm, S. (1981). "Phytoplankton lipids: Interspecific differences and effects of nitrate, silicate and light-dark cycles." *J. Phycol.*, 17, 374–384.

Spoehr, H. A., and Milner, H. W. (1949). "The chemical composition of *Chlorella*: Effect of environmental conditions." *Plant Physiol.*, 24, 120.

Straka, F., Doucha, J., and Livansky, K. (2000). "Flue-gas CO_2 as a source of carbon in closed cycle with solar culture of microalgae." *Proc., 4th European Workshop on Biotechnology of Microalgae*, 29–30.

Thomas, W. H., Seibert, D. L. R., Aldem, M., Neori, A., and Eldridge, P. (1984). "Yields, photosynthetic efficiency, and proximate composition of dense marine microalgal cultures. II: *Dunaliella primolecta* and *Tetraselmis suecica* experiments." *Biomass*, 5, 211–225.

Thompson, G. A. (1996). "Lipids and membrane function in green algae." *Biochemica et Biophysica*, 1306, 17–45.

Tsuzuki, M., Ohnuma, E., Sato, N., Takaku, T., and Kayguchi, A. (1990). "Effects of CO_2 concentration during growth on fatty acid composition in microalgae." *Plant Physiol.*, 93, 851–856.

USDA. (1992). "Agricultural waste characteristics." *Agricultural waste management field handbook*, United States Department of Agriculture, Soil Conservation Service, Washington, D.C.

USDA. (2005). *Agricultural statistics 2005*, National Agricultural Statistics Service, United States Government Printing Office, Washington, D.C.

39th Annual Meeting of the APS Division of Atomic, Molecular, and Optical Physics

Tuesday–Saturday, May 27–31, 2008; State College, Pennsylvania

ession E1: Poster Session I: 4:00 pm - 6:00 pm

:00 PM–4:00 PM, Wednesday, May 28, 2008
UB-Robeson Center - Alumni Hall

bstract: E1.00117 : Investigation of the polarization dependence of optical ipole traps for quantum computing

review Abstract

uthors:

Bert David Copsey
(California Polytechnic State University, San Luis Obispo)

Katharina Gillen-Christandl
(California Polytechnic State University, San Luis Obispo)

an effort to find ways to create scalable arrays of neutral atoms that allow bringing atoms together and apart for qubit gate operations, we are exploring the dependence of different dipole trap and optical lattice geometries on the ap light polarization. Several dark spot optical lattice and dipole trap geometries that have sufficiently low scattering tes for laser detunings comparable to the (ground state) hyperfine splitting have been proposed [1, 2]. To fully xplore the polarization dependence of these traps, we explicitly calculate the full expression for the optical dipole otential for this case, based on the expression given in [3]. We will present our progress towards identifying trap eometries with different light polarizations that might be used to bring atoms together and apart for two-qubit gates. 1. hys. Rev. A 70 032302 (2004), 2. Phys. Rev. A 73 013409 (2006), 3. Phys. Rev. A 57(3) 1972 (1998).

47th AIAA Aerospace Sciences Meeting Including The New Horizons Forum and Aerospace Exposition
5 - 8 January 2009, Orlando, Florida

AIAA 2009-1441

Underpowered Aircraft -- Performance and Operational Possibilities

Andrew S. Ezzard[1], Michael R. Vallone[2], and Robert A. McDonald, Ph.D.[3]
California Polytechnic State University, San Luis Obispo, CA 93407

A unique configuration, known as an Underpowered Aircraft, allows for the modification of gliding flight vehicles for increased range and lower cost when compared with fully powered flight vehicles. Intentionally under-sizing the powerplant for a flight vehicle allows the designer to choose a powerplant that will not only perform the mission requirements, but will also provide the customer with the most cost effective solution, as some missions may not require fully powered flight. Specifically, the underpowered aircraft concept studied in this paper is a gliding flight aircraft that does not have enough power for climbing or level flight, but does have enough power to overcome some of the drag forces associated with flight, in turn increasing the effective range of the vehicle. In this paper, the underpowered aircraft concept was analyzed and its feasibility was determined. Analysis done using equations of motion, followed by a more accurate numerical integration including a thrust lapse, determined that the underpowered aircraft concept provides a unique method for a cost effective range extension technology for gliding flight vehicles. Finally, the technology and methods of this paper were applied to the AGM-154 JSOW and JSOW-ER glide munitions and it was determined that JSOW-ER is representative of an underpowered aircraft with our analysis. This paper represents a "back-of-the-envelope" investigation into the underpowered aircraft concept.

Nomenclature

English

AR	=	aspect ratio
AR^*	=	representative aspect ratio = eAR
C_D	=	drag coefficient
C_L	=	lift coefficient
D	=	drag
W	=	weight
E	=	energy
e	=	Oswald efficiency factor
h	=	height, altitude
K	=	$\frac{1}{\pi eAR} = \frac{1}{\pi AR^*}$
L	=	lift
L/D	=	lift to drag ratio
m	=	mass
P/W	=	power to weight ratio
P_s	=	specific excess power
R/C, RoC	=	rate of climb
T	=	thrust
T/W	=	thrust to weight ratio
V	=	velocity
W/P	=	power loading
Z_e	=	energy height

Greek

α	=	angle of attack
η_p	=	propeller efficiency
θ	=	flight path angle
φ_T	=	thrust vectoring angle

Subscripts

max	=	maximum
o	=	parasite
T	=	thrust
to	=	takeoff

Acronyms

KE	=	kinetic energy
PE	=	potential energy
UAV	=	unmanned aerial vehicle

1Aerospace Engineering Undergraduate, AIAA Member
2Aerospace Engineering Graduate Student, AIAA Member
3Assistant Professor, Lockheed Martin Endowed Professor, Aerospace Engineering Department, AIAA Member

Copyright © 2009 by Andrew S. Ezzard. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.

I. Introduction

IT is not typical for an aircraft designer to undersize the powerplant for their aircraft. However, as the emphasis on cost in the aerospace industry continues to increase, the need to meet customer requirements through the most cost effective means presents itself in almost all engineering problems. In a time where the cost of fuel is fluctuating unpredictably, manufacturing and labor costs are increasing, and high technology systems bring the development cost up, the need to perform and fly a mission must be as cost effective as possible. Using the technology of an "underpowered" aircraft (an aircraft that only has enough power to overcome some of the drag forces associated with flight) has the potential to reduce the operating and propulsion system cost when compared to standard aircraft systems that are capable of climbing or level flight. Possible mission applications for the technology include cargo delivery for deployed troops in the battlefield, glide munitions, stand-off weapons, and others.

The underpowered aircraft concept comes from the idea that, for specific missions, an aircraft system may not necessarily need to overcome all of the drag forces associated with flight, and can therefore operate with an under-sized powerplant. If the purpose of the aircraft is to glide to its target, then the effective range of the aircraft can be increased by making the aircraft underpowered. By doing so, the aircraft does not need the same power as a standard aircraft would to achieve level or climbing flight, but can successfully operate off of smaller amounts of power while greatly extending the range of the vehicle. Adding power to the aircraft, in small amounts, allows the aircraft to increase the effective mission range, for little horsepower and low cost, as will be shown in this paper. The performance of an underpowered aircraft will be analyzed and the cost advantages will be determined. A more detailed numerical integration analysis of the flight path will also be presented and validation of the analysis will be conducted when compared to a glide munition operating off of the same flight principles.

II. Underpowered Aircraft Performance

The idea of an underpowered aircraft arises to fill the need to increase the range of gliding aircraft systems. Payload delivered using a gliding system provides for a cheap and simple delivery mechanism when constructed from light weight, inexpensive materials. The system can be designed to be single use or reusable, depending on the exact purpose. However, just using a standard gliding aircraft does not provide the range capability often needed for payload delivery. Assuming the underpowered aircraft is dropped from a high altitude, an engine can be sized and selected to provide the desired range. The trade space between range, flight velocity, and drop altitude for an underpowered aircraft will be explored. The following analysis will assume a payload delivery mission and then other mission profiles will be analyzed afterwards.

Two methods were used to determine the performance for the underpowered aircraft. First, the standard method of looking at the forces acting on the aircraft via a free-body diagram was applied. Estimates for the drag polar of the aircraft were also developed, which yielded an estimate of flight conditions. This was followed by a numerical integration to provide a more accurate measure and to account for thrust lapse. Emphasis was placed on the lift to drag ratio (*L/D*) due to the vehicle's operation in gliding flight. In an abstract trade study such as this, aircraft characteristics are treated as technology that can be applied to an aircraft in the design process, allowing for an overall performance trade study.

A. Equations of Motion Analysis

To start, Figure 1 shows the standard forces acting on an aircraft in flight for use in the equations of motions analysis.

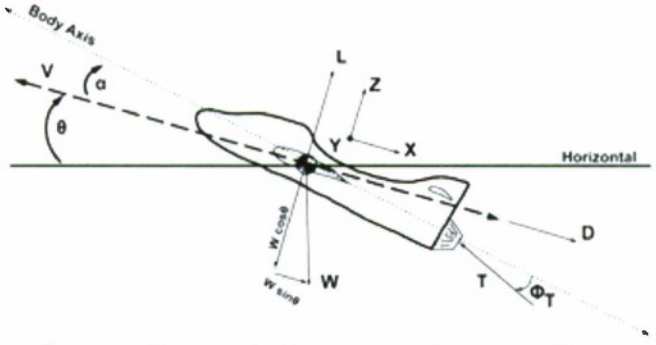


Figure 1. Forces Acting on an Aircraft in Flight

The forces acting on the aircraft are standard and include the lift, drag, thrust, and weight of the aircraft. The velocity, flight path angle, θ, and the angle of attack, α, are also shown. The flight path angle is defined as the angle between the horizontal and the velocity vector and the angle of attack is defined as the angle between the velocity vector and the arbitrarily defined body axes for the aircraft. In the case of thrust vectoring, the thrust vector is offset by the angle φ_T, to model any effects from thrust vectoring on the performance of the aircraft.

This analysis assumes that there is no thrust vectoring happening in the flight of the aircraft. Thus, summing the forces in the X and Z-axis for steady flight yields

$$T = D + W \sin\theta \tag{1}$$

$$L = W \cos\theta \tag{2}$$

Thus, we reorganize equation 1 to find the classical aircraft performance equation

$$\sin\theta = \frac{T-D}{W} \tag{3}$$

For gliding flight, our interest lies in the flight path angle, θ, of the vehicle. By minimizing the flight path angle to a small negative number, the range of the aircraft can be maximized. So, we have

$$\theta = \sin^{-1}\left(\frac{T}{W} - \frac{D}{W}\right) \tag{4}$$

From equation 2 above and a small angle approximation, we have

$$\theta = \sin^{-1}\left(\frac{T}{W} - \frac{1}{\frac{L}{D}}\right) \tag{5}$$

With the power to weight ratio (in units of hp/lb) defined as

$$\frac{P}{W} = \left(\frac{T}{W}\right)\frac{V}{550\eta_p} \tag{6}$$

B. Drag Polar

One of the key characteristics needed to properly model the performance of an aircraft is an accurate representation of the drag polar for the entire aircraft. The form of the drag polar that will be used for the analysis is

$$C_D = C_{Do} + \frac{C_L^2}{\pi AR^*} \tag{7}$$

For the purposes of this study, the aspect ratio, AR, of the aircraft will be grouped together with the Oswald efficiency factor, e, to become a representative aspect ratio, AR^*. In an abstract trade study such as this no information is available on the efficiency of the wing design; however a feel for the efficiency factor can still be gained through a representative aspect ratio. If we assume that the aircraft is operating at conditions that give the C_L for best L/D for the gliding condition, then we know that[1]

$$C_L^* = \sqrt{\frac{C_{Do}}{\pi AR^*}} \tag{8}$$

$$C_D^* = 2C_{Do} \tag{9}$$

So, the lift to drag ratio is

$$\frac{L}{D}_{max} = \frac{C_L}{C_D} = \frac{\sqrt{\frac{C_{Do}}{\pi AR^*}}}{2C_{Do}} = \sqrt{\frac{1}{4\pi AR^* C_{Do}}} \tag{10}$$

Reorganizing, an estimate of the parasite drag can be determined through

$$C_{Do} = \frac{\pi AR^*}{4\left(\frac{L}{D_{max}}\right)^2} \tag{12}$$

This allows us to treat the aerodynamics of the aircraft, AR* and $\frac{L}{D_{max}}$, as technology.

III. Powerplant Cost Estimation

With the performance for the underpowered aircraft concept established, the cost benefits of the technology were determined. Such a vehicle can be powered either by traditional internal combustion piston engines or small jet turbine engines. In order to get an estimate for the cost saving of the vehicle, estimates of small piston and jet turbine engines were developed. As seen below in Figure 2, a span of cost competitive small piston engines was used to determine a cost trend that was acquired through a survey of retail prices found online in August, 2008. These engines included small RC aircraft engines from O.S. Engines, general purpose engines from Honda and Kawasaki, as well as a few larger Rotax aircraft engines. The model yields an excellent correlation coefficient and the results are consistent with the investigative nature of this paper.

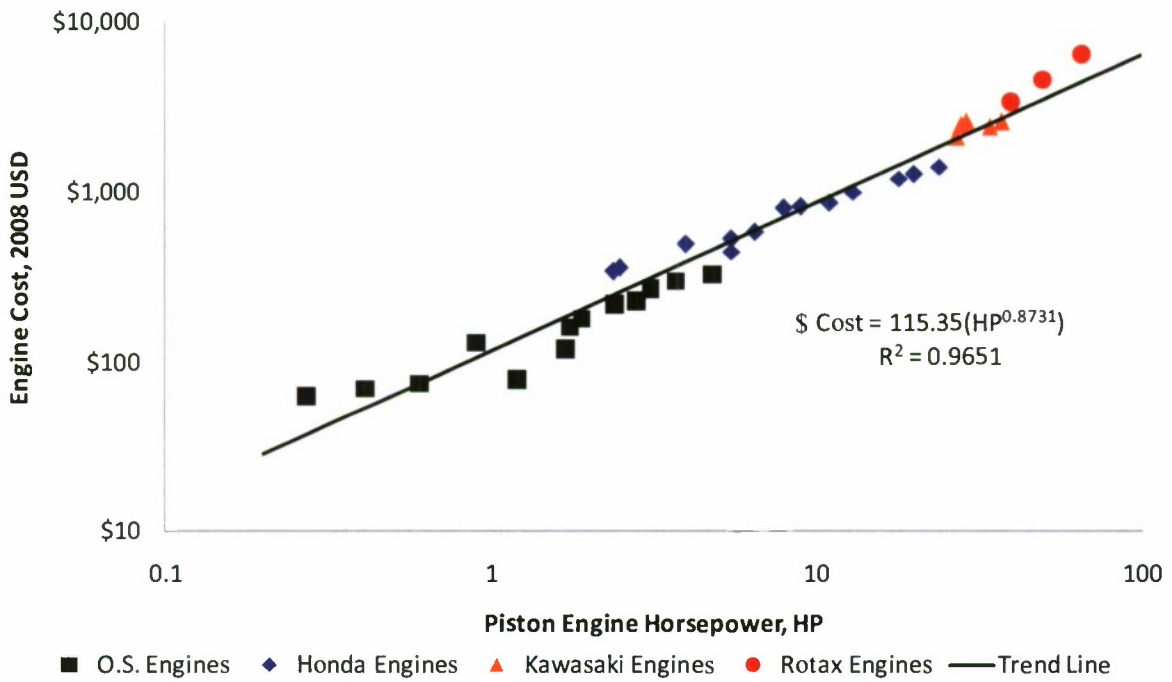


Figure 2. Small Piston Engine Cost

Figure 3 shows a similar trend for small jet turbine engines. Due to the fact that the payload delivery vehicles and stand-off "glide munitions" where this technology would be used are not large in size, the engine study was limited to model aircraft jet engines as well as jet engines used in small UAV applications. These engines include manufacturers such as JetCat, foreign engine manufacturers such as SimJet, and others. This model is indicative of the low thrust engines needed for underpowered applications. The model yields acceptable results with a correlation coefficient of 82%.

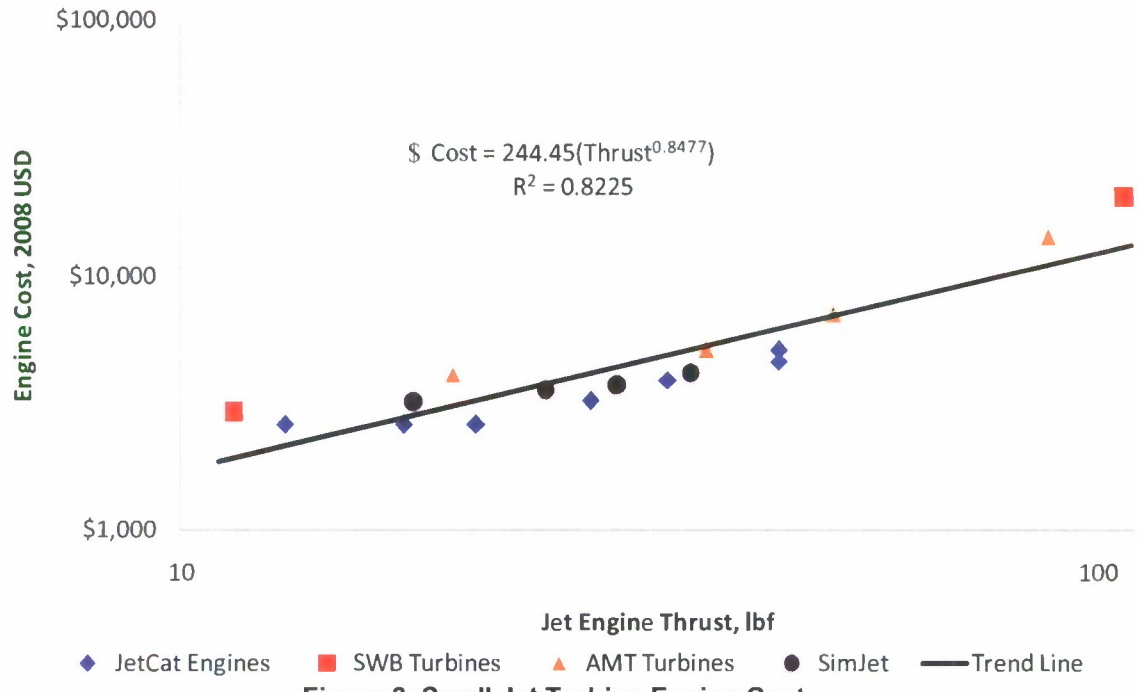


Figure 3. Small Jet Turbine Engine Cost

IV. Underpowered Aircraft Flight Performance

Sinee the coneept for this aircraft is based on that of a glider, the performanee of the aircraft has been determined to rcpresent the trade space of velocity, range, and drop altitude. So, using the relationships derived earlier and performance code written in MATLAB, some performance trends were calculated.

It is assumed in this analysis that the underpowered aircraft is dropped from altitude at the start of the flight. Two altitudes of 10,000 feet and 25,000 feet were ehosen to give a eomparison between a low altitude drop and a high altitude drop. The drop aircraft eould consist of any type of vehicle that is capable of carrying the underpowered aireraft to the drop altitude. Winds aloft are not included in the analysis. The analysis presented assumes that thc underpowered vehiele is dropped at the veloeity of best *L/D* and that during the glide the vehicle operates at the best *L/D* for glide.

As the lift to drag ratio is inereased, the thrust required to get the aircraft to reaeh its destination decreases, which can bc seen in Figure 4. Only a small amount of thrust is required to keep the aircraft in a lcvel flight eondition, but a signifieantly larger amount of thrust is needed to have a positive rate of climb for the aircraft (represented by the +3 degree climb angle curve). The difference in gliding flight, level flight, and positive rate of climb flight is indicative of the trade space between range, velocity, and drop altitude (10,000 feet for these curves). It is also indicative of the trade in the cost of the system, as having an aircraft that is capable of positive rate of climb requires a signifieantly larger amount of thrust and therefore a larger, more expensive powerplant. It is important to note that there are no additional assumptions built into the figure below and the plot is derivcd solely from aircraft pcrformance metrics.

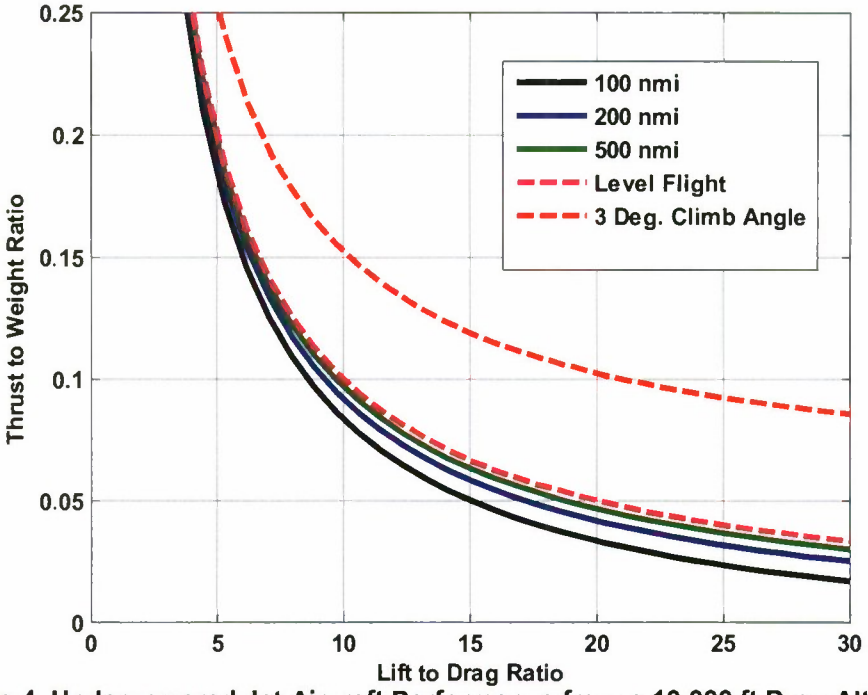


Figure 4. Underpowered Jet Aircraft Performance from a 10,000 ft Drop Altitude

The curves in Fig. 4 are a function of thrust to weight ratio and lift to drag ratio which are two of the most fundamental aircraft performance parameters. For these initial calculations the thrust lapse has been ignored. Using the cost model presented earlier we can estimate the cost saving of the underpowered aircraft technology. Assuming that the vehicle has a takeoff gross weight (TOGW) of 1,500 pounds and can achieve a lift-to-drag ratio (L/D) of 20, the vehicle requires a thrust to weight ratio 0.04 to achieve a range of 200 nmi. This corresponds to a thrust required of 60 lbf and an engine cost of $7,862. In contrast, the engine would need to provide 75 lbf for level flight and 150 lbf for climbing flight. This yields a cost of $9,500 and $17,095, respectively.

The same curves can be represented for a propeller driven aircraft by including flight velocity and propeller efficiency. Shown in Figure 5 are the same trends as a function of power to weight ratio.

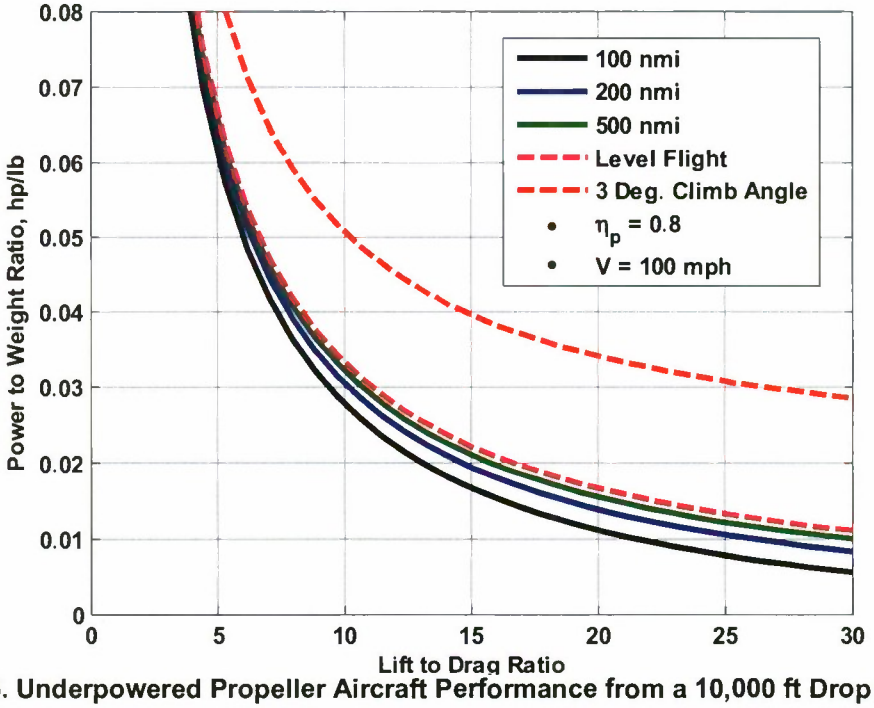


Figure 5. Underpowered Propeller Aircraft Performance from a 10,000 ft Drop Altitude

Figure 5 shows similar trends to Figure 4. This plot speaks volumes about the underpowered aircraft concept. If we assume that the underpowered aircraft can achieve a maximum lift to drag ratio of 20 and has a TOGW of 1,500 pounds, then the aircraft only requires about 20 hp to achieve a range of 200 nmi (even less for shorter ranges). The same aircraft would require 26 hp for level flight, and about 53 hp for climbing flight. The underpowered aircraft requires over one half the horsepower to achieve the mission requirement which translates to engine costs of $1,577, $1,983, and $3,694 respectively. There are two main assumptions built into this analysis: First, a propeller efficiency of 0.80 is assumed to convert from thrust to horsepower (80% being typical performance of most propellers currently used today) and second, an operating velocity of 100 mph (146.6 ft/sec) was arbitrarily chosen to represent a reasonable delivery speed for the payload aircraft.

A different presentation of the data in Fig. 5, seen below in Fig. 6, shows the relationship between power loading and lift to drag ratio for the underpowered aircraft.

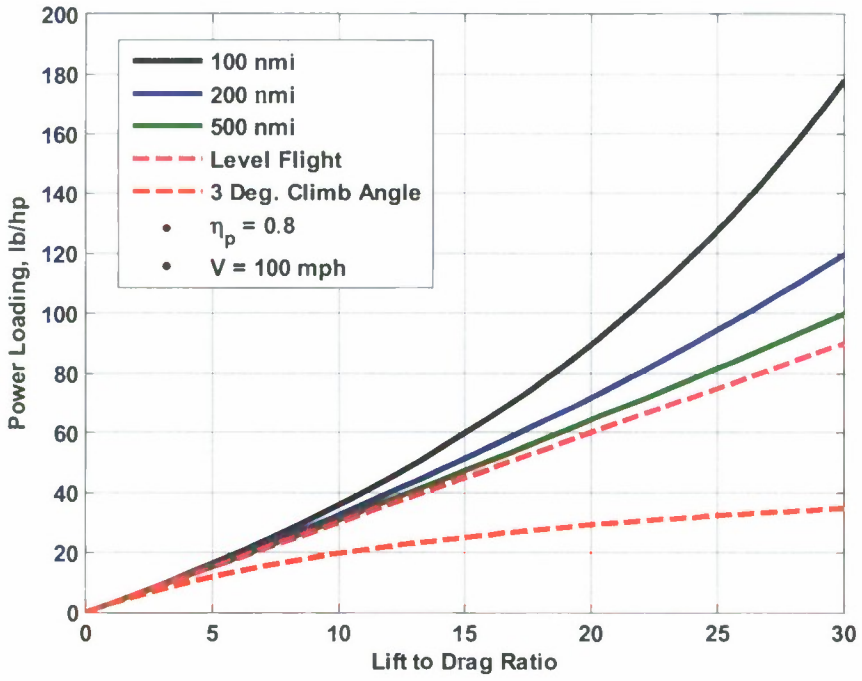


Figure 6. Underpowered Aircraft Power Loading Requirements for a 10,000 ft Drop Altitude

This figure shows how much of the aircraft (in weight) can be supported for each unit of horsepower that the engine on the aircraft will produce. If we again assume that the aircraft has a lift to drag ratio of 20 and a weight of 1,500 pounds, then the aircraft can carry about 76 pounds for each horsepower. This results in a power to weight ratio of 0.013 hp/lb and about a 20 hp engine as shown above in the previous figures. This analysis also assumes a propeller efficiency of 0.80 and a flight velocity of 100 mph as mentioned earlier.

The same graphs were created to look at the performance of the aircraft from the drop altitude of 25,000 feet. Note the significantly less power required for the underpowered aircraft to achieve a desired range of 200 nmi in Figure 7.

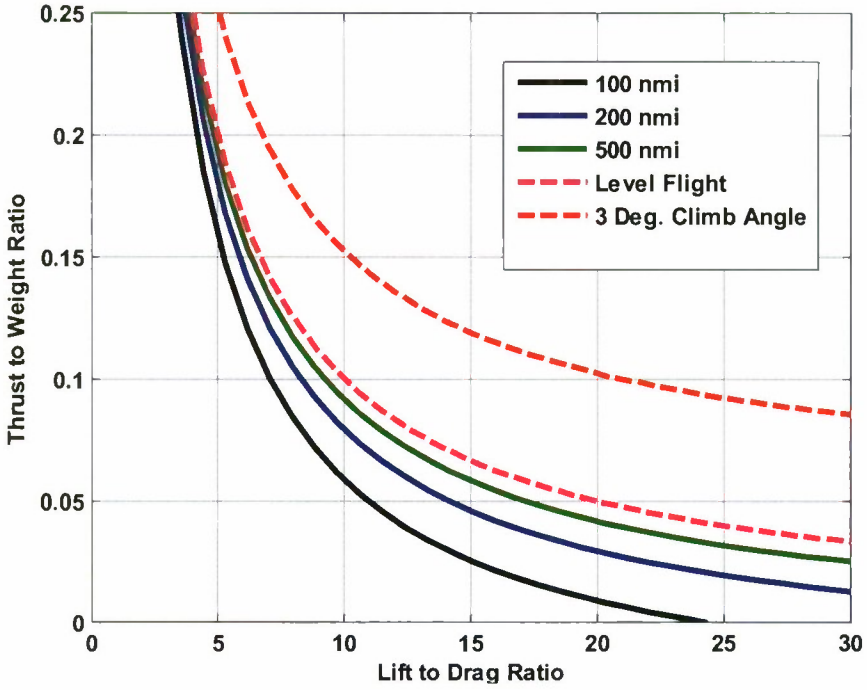


Figure 7. Underpowered Jet Aircraft Glide Performance from a 25,000 ft Drop Altitude

If we continue to assume that the aircraft has a TOGW of 1,500 pounds and can achieve a lift-to-drag ratio of 20, the vehicle will require a thrust to weight ratio of 0.03 to achieve a 200 nmi range, a thrust to weight ratio of 0.05 to achieve level flight, and a thrust to weight ratio of 0.1 to achieve a positive rate of climb. From the cost models presented above, the engine would need 45 lbf thrust, 75 lbf thrust, and 150 lbf thrust respectively. This would represent a significant cost savings between $6,161, $9,500, and $17,095 respectively.

Shown in Figure 8 is the relationship between power to weight ratio and lift to drag ratio.

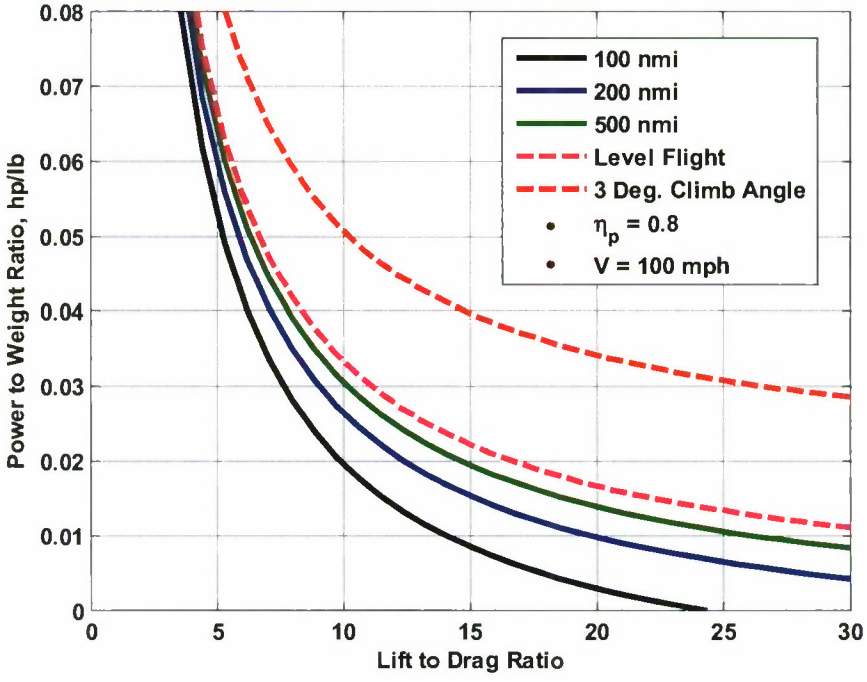


Figure 8. Underpowered Propeller Aircraft Performance from a 25,000 ft Drop Altitude

Using the same assumptions as before and looking at an aircraft that can achieve an aerodynamic lift to drag ratio of 20, and has a weight of 1,500 pounds, the aircraft will require a power to weight ratio of 0.01 hp/lb or 15 hp to reach the desired range of 200 nmi. This would be a minimal cost of $1,227. This is significantly less than an aircraft that would need to sustain level flight (approx. 26 hp or $1,983) and an aircraft looking to have a positive rate of climb (approx. 53 hp or $3,694). The underpowered aircraft technology will satisfy the requirements with significantly less cost due to the much smaller engine required which reduces cost throughout the life cycle. These values are sensitive to wing technology (lift to drag ratio) and will change depending on the *L/D* that can be achieved for the vehicle.

The power loading for the underpowered aircraft also significantly changes for the additional drop altitude of 25,000 feet and allows the aircraft to glide a longer distance with less power.

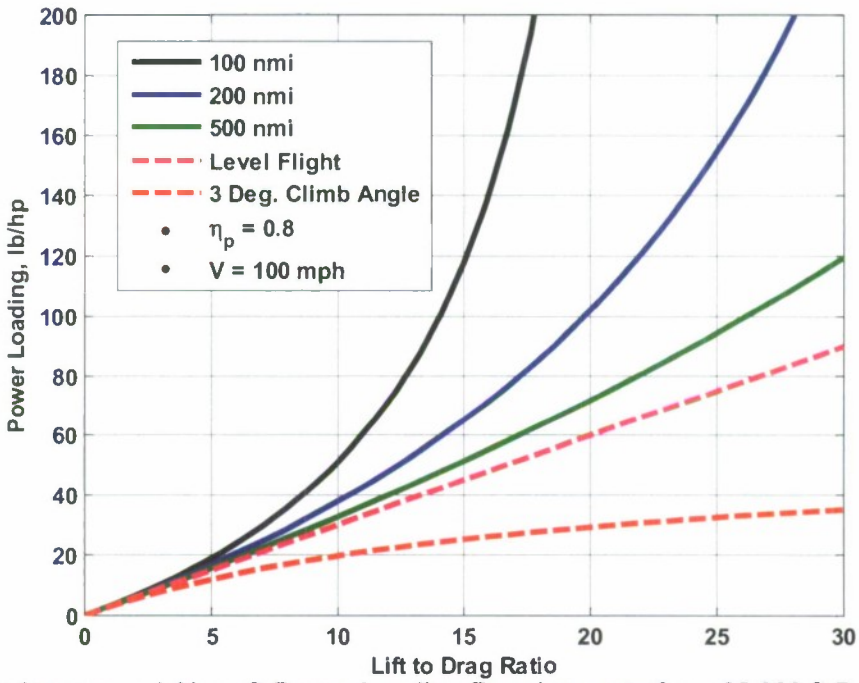


Figure 9. Underpowered Aircraft Power Loading Requirements for a 25,000 ft Drop Altitude

Comparing the two drop altitudes shows significant savings as well. Using the same aircraft assumptions as before, namely a TOGW of 1,500 pounds and a maximum lift to drag ratio of 20, we see that a drop from 10,000 ft achieves 200 nmi with a 60 lbf engine and an engine cost of $7,862. This cost is decreased when dropped from 25,000 ft, when the plane achieves the same range with a 45 lbf engine and an engine cost of $6,161.

V. Numerical Integration Approach

While the above analysis represents a "back-of-the-envelope approach", there is one main factor that was left out; the effects of velocity and altitude on powerplant performance. This takes the form of a thrust lapse that will change the maximum thrust to weight ratio represented in the earlier figures, to a typical flight thrust to weight ratio that has been adjusted to represent actual flight conditions. To achieve this, a numerical integration approach was used to examine the equations of motion, take a thrust lapse into account, and give more accurate results than the methods presented above.

Revisiting the free body diagram in Figure 1 and summing the forces in the X and Z-axis, we have

$$\cos\theta = \frac{L}{W} \tag{13}$$

$$\sin\theta = \frac{T-D}{W} \tag{14}$$

Due to the fact that our interest lies in gliding flight, we are actually interested in the change in altitude of the vehicle during flight with respect to the change in distance the vehicle travels along the ground. This is essentially the glide ratio in derivative form. So, using the free body diagram

$$\frac{ds}{dh} = \frac{\frac{ds}{dt}}{\frac{dh}{dt}} = \frac{V\cos\theta}{V\sin\theta} = \frac{1}{\tan\theta} \tag{15}$$

$$\frac{1}{\tan\theta} = \frac{L}{T-D} \tag{16}$$

Using a thrust lapse equation from *Mattingly*[2] defined in terms of the Mach number, M, and the density ratio from sea level, σ, we have

$$\alpha = 0.76\{0.907 + 0.262(|M - 0.5|)^{1.5}\}\sigma^{0.7} \tag{17}$$

Setting up the integral to integrate the flight path during steady flight, we have

$$\int_h^0 \left(\frac{\frac{1}{W}}{\frac{1}{W}}\right)\frac{L}{\alpha T - D}dh \tag{18}$$

$$\int_h^0 \left(\frac{1}{\alpha\frac{T}{W} - \frac{1}{\frac{L}{D}}}\right)dh \tag{19}$$

The numerical integration yielded results that provide a better representation of the flight regime with the inclusion of a thrust lapse. The trend in the curves shows that the numerical model actually requires more thrust than the equation of motion or energy method analysis because it now takes powerplant performance into account. The powerplant will not perform at sea level static conditions at high altitude. In an effort to validate the results of the model presented above, a drop altitude of 40,000 ft was used to match the ranges given for the AGM-154 Joint Standoff Weapon[3]. Figure 10 shows the results of the numerical integration.

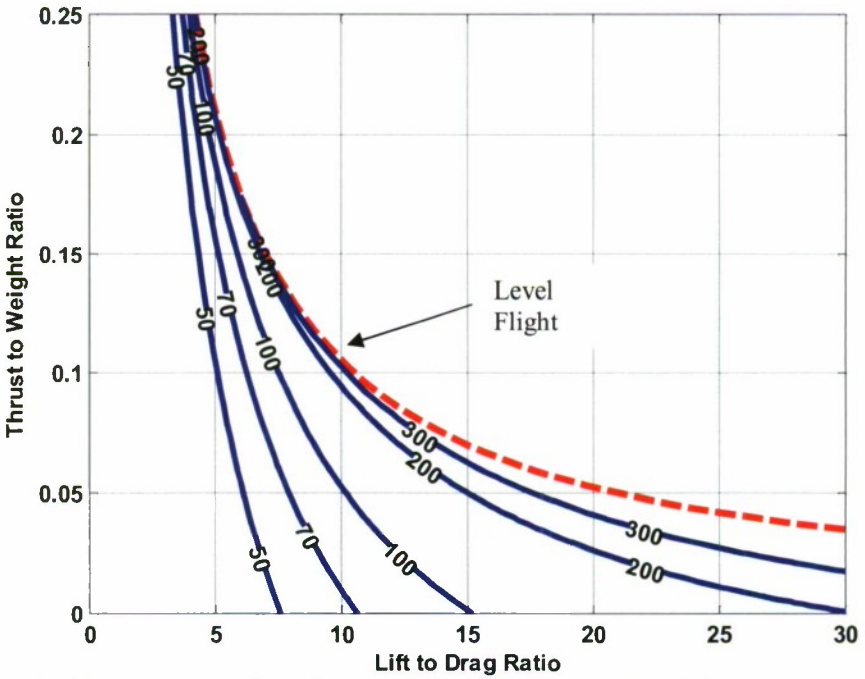


Figure 10. Numerical Integration of Jet Engine with Lapse from 40000 ft, Contours of Glide Range

According to the AGM-154 product card[3] the unpowered version of the JSOW has an effective glide range of 70 nmi from a 40,000 ft drop altitude. When using the analysis of Figure 10, for an un-powered vehicle, this corresponds to a lift to drag ratio of about 11. The weight reported for this vehicle is 1,050 lbf (depending on variant)[3]. Recently reported in a Raytheon press release[4] was a powered version of the JSOW, the JSOW-ER, with a 300 nmi range using a 150 lbf thrust turbojet engine. If we assume that the powered version is also dropped from 40,000 ft, and using the same *L/D* as determined for the un-powered version, the weight of the powered JSOW-ER was determined to be 1,575 lbf. This is very close to the weight reported for the un-powered version, and shows that the JSOW-ER is most likely an "underpowered" vehicle, utilizing the technology described in this paper.

As an estimation technique, the simpler equation of motion analysis was calibrated to match the results of the numerical integration analysis. Figure 11 shows the same curves for the simpler analysis, including the same thrust lapse model presented above, but the density ratio is estimated at an altitude of 18,000 ft. This was considered a good estimate for our purposes of modeling the flight regime because most of the effects of the thrust lapse are seen at lower altitudes for longer ranges.

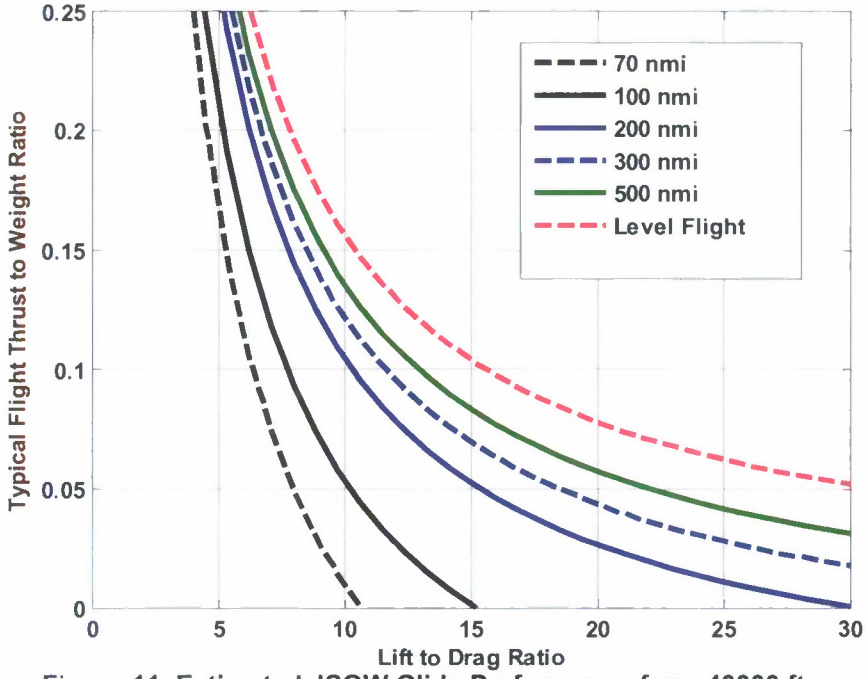


Figure 11. Estimated JSOW Glide Performance from 40000 ft

VI. Conclusion

An underpowered aircraft concept was studied to determine the feasibility and cost effectiveness of such a technology, as well as possible applications. It was determined, through an equations of motion, energy method, and numerical integration analysis, that an underpowered aircraft can provide significant range extension for a gliding flight vehicle. The methods also include a built in thrust lapse to correctly model an underpowered vehicle's performance with altitude. Also, the methods developed in this paper were compared with a current gliding vehicle, the AGM-154 JSOW and JSOW-ER. It was determined, from aircraft metrics provided by Raytheon and our analysis, that the JSOW-ER is most likely an underpowered aircraft and is representative of a possible mission application for the technology. Other applications include payload delivery, glide munitions, UAVs, and others. Overall, the underpowered aircraft technology represents a unique flight regime, giving great benefit to the user for a low overall cost.

References

[1]Anderson, John D., *Aircraft Performance and Design*, WCB/McGraw-Hill, 1998

[2]Mattingly, J. D., Heiser, W. H., and Daley, D. H., *Aircraft Engine Design*, AIAA Education Series, AIAA, New York, 1987, Chap. 2.

[3]Raytheon Company. JSOW: Family of Precision Strike Weapons. Brochure. Tucson: Raytheon Company, Missile Systems, 2008.

[4]Raytheon. Press release. Raytheon Demonstrates Engine for Powered Joint Standoff Weapon Extended Range. 20 Feb. 2007. Aug. 2008 <http://www.globalsecurity.org/military/library/news/2007/02/mil-070220-raytheon02.htm>.

Bone Mass Preservation and Fracture Risk Assessment with Bisphosphonate Therapy During Spaceflight

A Thesis

Presented to the Faculty of

California Polytechnic State University

San Luis Obispo

In Partial Fulfillment of the

Requirements for the Degree

Master of Science in Engineering

with a Specialization in Biomedical Engineering

by

Christopher Gardina

June 2008

ABSTRACT

Bone Mass Preservation and Fracture Risk Assessment
with Bisphosphonate Therapy During Spaceflight

Christopher Gardina

Space exploration and microgravity have substantial negative effects on the human body. Symptoms of space explorers include cardiovascular deconditioning, bone loss, muscular atrophy, and impairment of neurovestibular and sensory function. The great loss of bone due long-duration spaceflight increases fracture risk, jeopardizing the success of the mission and postflight recovery. Bisphosphonates may be able to counteract this bone loss by altering the remodeling process. These drugs increase bone mass, thus reducing fracture risk, but also lead to increased levels of fatigue microdamage. Fracture risk can be lowered by increasing both bone mass (quantity) and bone quality.

The purpose of this study was to create a computer model to simulate bisphosphonate treatment on astronauts while traveling in space in order to examine the ability of bisphosphonates to maintain bone mass in a microgravity environment and reduce fracture risk of bone upon return to Earth. Various bisphosphonate treatment potencies and bone balance ratios given at different time points (either at or before spaceflight) were examined. Flight duration was also varied to examine short-term (10 days) to long-term (1 year) effects of microgravity on bone mineral density (BMD), a measure commonly used to estimate bone strength, and damage accumulation. The model predicted bisphosphonate treatments with low to intermediate suppression of remodeling activation and that create higher bone balance ratios cause reductions in fracture risk. The simulation also predicted significant changes to BMD and damage upon return to Earth

as the remodeling response readjusted to higher stress conditions. For treatments highly suppressing remodeling activation, these predicted postflight changes included decreased BMD and increased damage accumulation. Low levels of remodeling suppression led the model to predict substantial increases in BMD and small increases in damage postflight. Postflight changes were minimal for treatments with intermediate suppression.

TABLE OF CONTENTS

LIST OF TABLES v
LIST OF FIGURES vi
CHAPTER 1: INTRODUCTION 1
1.1 Spaceflight 1
1.2 Properties of Bone 2
1.3 Bisphosphonates 5
1.4 Previous Models 7
1.5 Simulating Bisphosphonate Treatment during Spaceflight 11
CHAPTER 2: METHODS 12
2.1 Mechanical Loading 12
2.2 Porosity Transformation 13
2.3 BMU Activation 14
2.4 Microdamage Accumulation 15
2.5 Preflight Conditions 16
2.6 Simulation of Spaceflight 18
2.7 Simulation of Bisphosphonates 18
2.8 Simulation of Return to Earth 21
2.9 Model Implementation 21
2.10 Postflight Parameter Analysis 21
CHAPTER 3: RESULTS 23
3.1 Untreated Spaceflight 23
3.2 Bisphosphonate Treatment Onset at Beginning of Spaceflight 26
3.3 Preflight Bisphosphonate Treatment 29
3.4 Varying Onset of Bisphosphonate Treatment for Spaceflight 33
3.5 One-Year Postflight, Posttreatment Recovery 34
CHAPTER 4: DISCUSSION 40
SUMMARY OF CONCLUSIONS 46
REFERENCES 47
APPENDIX A: MATLAB CODE 51
APPENDIX B: FIGURES (10-DAY SPACEFLIGHT) 59
APPENDIX C: FIGURES (90-DAY SPACEFLIGHT) 71
APPENDIX D: FIGURES (180-DAY SPACEFLIGHT) 83
APPENDIX E: FIGURES (365-DAY SPACEFLIGHT) 96

LIST OF TABLES

Table 2.1. Model parameters with values obtained from simulating remodeling in trabecular bone before entering space and at the end of a 180-day space simulation. 16
Table 2.2. Remodeling simulation constants. 17
Table 2.3. Bisphosphonate effects analyzed during the simulation. 19
Table 3.1. Ratio of percent preflight increase of damage (D) to BMD. 33

LIST OF FIGURES

Figure 1.1. Illustration of long bone showing cortical and trabecular bone. 3
Figure 1.2. A healthy trabecular bone strut and one with microdamage. 3
Figure 1.3. A BMU containing osteoblasts and osteoclasts 4
Figure 1.4. Chemical structure of bisphosphonates. 6
Figure 1.5. Schematic of a bone remodeling algorithm by Hazelwood et al. 9
Figure 1.6. Schematic of a bone remodeling algorithm by Hernandez et al. 10
Figure 2.1. Trabecular volume and cross-section. 18
Figure 2.2. Bisphosphonate potency as a function of number of resorbing BMUs. 20
Figure 3.1. Predicted percent decreases in BMD and damage (D) of untreated bone at end of spaceflight. 24
Figure 3.2. Predicted effects of 180-day spaceflight on BMD of untreated bone from beginning of flight through 19.5 years postflight. 24
Figure 3.3. Predicted effects of 180-day spaceflight on damage accumulation (D) of untreated bone from beginning of flight through 19.5 years postflight. 25
Figure 3.4. Predicted percent changes in BMD and damage (D) at end of 10-day spaceflight. 27
Figure 3.5. Predicted percent changes in BMD and damage (D) at end of 90-day spaceflight. 28
Figure 3.6. Predicted percent changes in BMD and damage (D) at end of 180-day spaceflight. 28
Figure 3.7. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight. 29
Figure 3.8. Predicted preflight increase in BMD and damage (D) due to 7-day preflight treatment 30
Figure 3.9. Predicted preflight increase in BMD and damage (D) due to 14-day preflight treatment 31
Figure 3.10. Predicted preflight increase in BMD and damage (D) due to 30-day preflight treatment 31
Figure 3.11. Predicted preflight increase in BMD and damage (D) due to 90-day preflight treatment 32
Figure 3.12. Predicted preflight increase in BMD and damage (D) due to 180-day preflight treatment 32
Figure 3.13. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight due to 30-day preflight treatment. 33
Figure 3.14. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight due to 180-day preflight treatment. 34
Figure 3.15. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight. 35
Figure 3.16. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight. 36
Figure 3.17. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight. 37
Figure 3.18. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight. 37

Figure 3.19. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight. 38
Figure 3.20. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight. 39
Figure 3.21. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight. 39

CHAPTER 1: INTRODUCTION

1.1 Spaceflight

Microgravity has many adverse effects on the human body that hinder the ability of astronauts to explore outer space. The physiologic adaptations caused by unloading include cardiovascular deconditioning, bone loss, muscular atrophy, and impairment of neurovestibular and sensory function [1]. The focus here is on bone loss, which results from the reduced levels of stress caused by weightlessness. This loss poses a significant health risk for astronauts and is a major deciding factor of mission duration. In 6-month missions to the International Space Station (ISS), astronauts experienced up to 5 percent loss of bone mineral density (BMD) in the lumbar spine and 10 percent in the proximal femur [2]. On the MIR space station, the greatest bone loss observed in a crew member was on the order of half the mineral loss incurred in a lifetime of normal aging [3]. Upon return to Earth's gravitational environment, the average fracture risk for a space explorer paralleled the estimated level for 70- to 80-year old postmenopausal women [4]. While bone has the ability to recovery mineral, it is much slower than the rate at which it is lost. Complete recovery may take from 1 to 3 years, and in many cases bone will never fully recover the mineral lost during spaceflight [5].

NASA and other space exploration agencies have developed exercise programs to combat bone loss in space. In earlier space missions, exercise routines utilized bungee cords for resistive exercises, and stationary bicycles and treadmills for longer, aerobic exercises. More recent missions to the ISS combined these older techniques with a new piece of equipment, the Interim Resistive Exercise Device (iRED), focusing on resistive training. This apparatus, by sufficiently increasing loading intensity, may lead to shorter,

more effective exercise routines [5]. Thus far, exercise routines have been able to slow bone loss, but remain unsuccessful at completely preventing it from occurring. NASA believes that coupling pharmacological treatments with their exercise programs will allow them to reach their goal [4]. Though the efficacy of many treatments such as testosterone, parathyroid hormone, calcium, vitamin D, and vitamin K, are being investigated, bisphosphonates seem a likely candidate [6]. Bisphosphonate therapy has the potential to increase bone mass in space just as it does for osteoporotic patients on Earth [2], though the pharmacokinetic altering effects of spaceflight have yet to be determined.

1.2 Properties of Bone

In a healthy individual, bones provide structure and support, and with the help of muscles, tendons and ligaments, they allow for movement of the body. There are two distinct types of bone tissue: cortical or compact bone and trabecular (also referred to as cancellous or spongy) bone. The main difference between these two types is their porosity. Cortical bone is fairly dense and non-porous, while trabecular bone has a much higher porosity and is made up of 'struts' that form an interconnected matrix (Figures 1.1 and 1.2). Theoretically, the porosity of bone tissue can be anywhere from 0 to 100 percent; however, it is almost always either very high (trabecular bone) or very low (cortical bone) and rarely in the intermediate range [7]. Cortical bone contains cylindrical units called osteons (Figure 1.1) and typically has a porosity of 5 to 10 percent. It can be found in the shafts and outermost layer of bone. Trabecular bone, normally 75 to 95 percent porous, is contained deep inside the bone and is filled with marrow [7]. It is made up of a matrix of packets of bone.

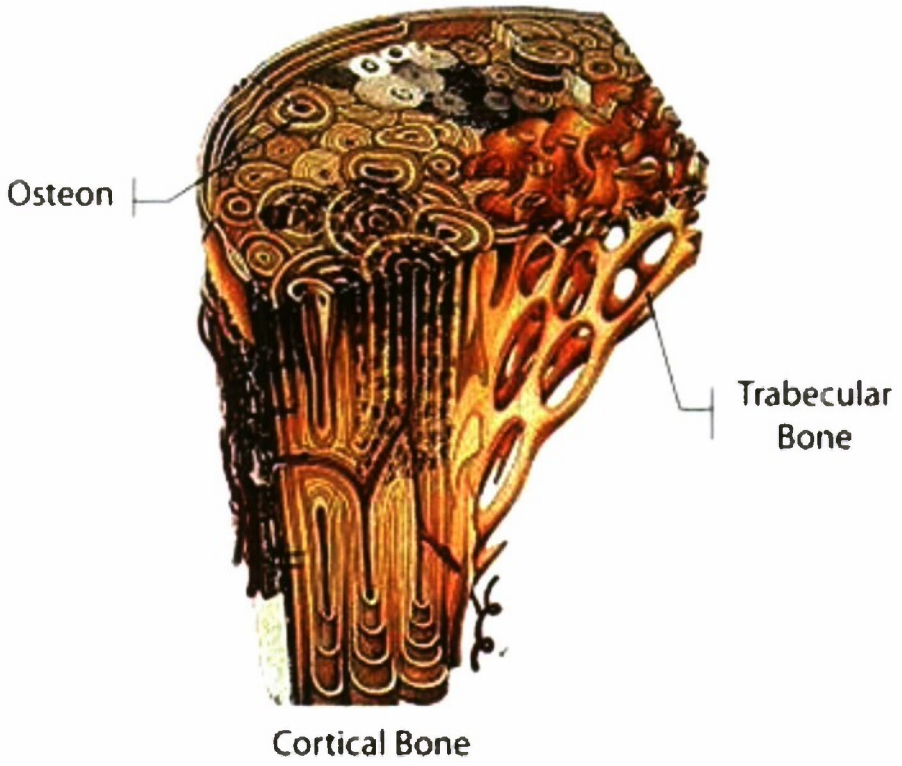


Figure 1.1. Illustration of long bone showing cortical and trabecular bone [8].

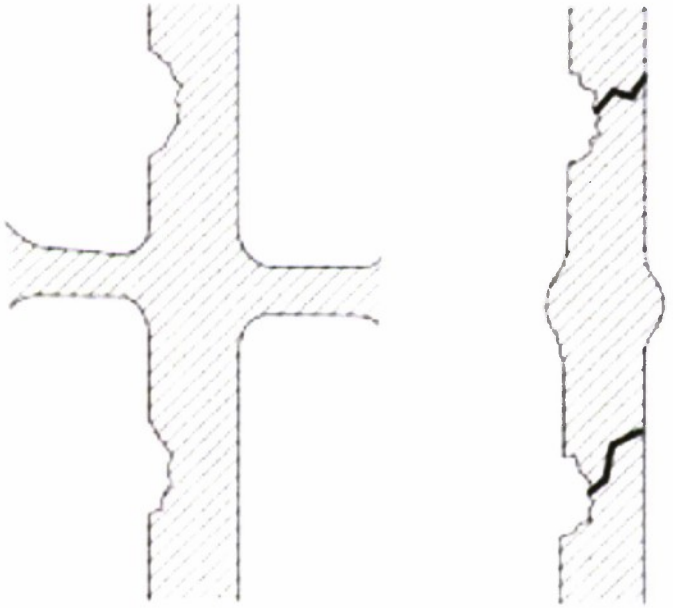


Figure 1.2. A healthy trabecular bone strut and one with microdamage [9].

Bone tissue experiences repetitive stress, leading to the formation of microdamage. Remodeling removes this damage, preventing fatigue failure from occurring under normal conditions [10]. Bone is made up of many packets of bone cells and minerals that are constantly undergoing this process. There are three major cells that partake in remodeling: osteoblasts, osteoclasts, and osteocytes. Osteoblasts and osteoclasts are bone-forming and bone-resorbing cells, respectively. Osteocytes are differentiated osteoblasts that are now fused into the bone matrix. Their role is to sense mechanical stimuli and send signals based on what they sensed to the surrounding bone cells [7]. The 3 to 4 month process of remodeling begins when osteoclasts receive signals to resorb damaged or fatigued bone. Osteoblasts then take over to remodel and 'fill in' the trenches created by the osteoclasts [7]. Together, osteoblasts and osteoclasts make up basic multicellular units (BMUs) of bone (Figure 1.3).

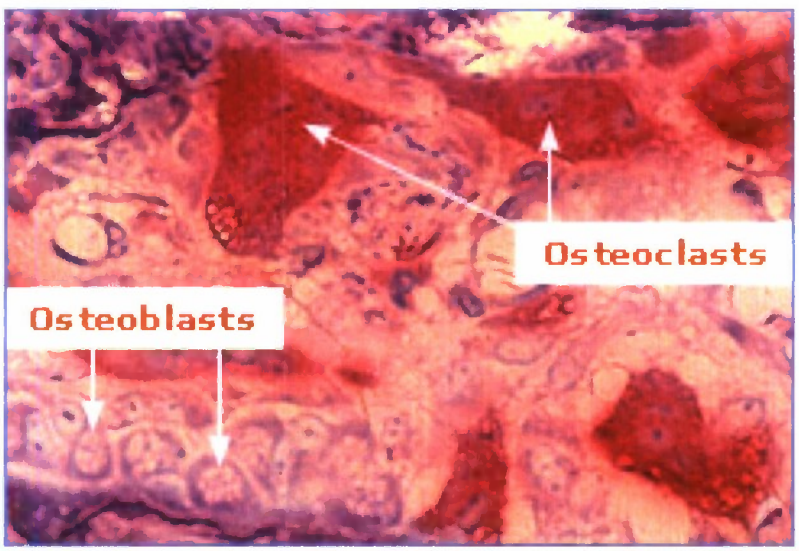


Figure 1.3. A BMU containing osteoblasts and osteoclasts [11].

BMUs are activated to remodel not only in response to increased microdamage [12,13,14,15], but also due to disuse situations where bone use is reduced [16,17].

Exposing bone to fewer loading cycles or reduced load lowers the frequency or magnitude of strain experienced by the bone so that its current level of strength is no longer necessary. Bone is largely remodeled in order to minimize bone mass while maintaining the strength necessary to support the body [18]. The equilibrium point, where remodeling due to mechanical loading is such that bone resorption equals bone formation, varies from subject to subject. In general, as applied force and strain increase, the number of cycles required to maintain bone mass decreases [19].

1.3 Bisphosphonates

Bisphosphonates provide new hope for long-duration space exploration. This class of drugs treats diseases with elevated bone remodeling by suppressing osteoclastic function [18]. Bisphosphonates suppress resorption upon physical contact with osteoclasts [20]. They have a high affinity for bone mineral [21] and bind to bone in areas where resorption has exposed hydoxyapatite [22]. There are a variety of bisphosphonates, each having different potency and cellular mechanisms inhibiting BMU activation [23]. The various potencies allow for a range in the degree of resorption suppression [24].

Bisphosphonates vary due to their chemical composition (Figure 1.4). First generation bisphosphonates, etidronate, clodronate, and tiludronate, did not have side chains containing nitrogen nor hydroxyl groups. These first bisphosphonates were less potent and did not specifically target bone as well as those from generation two [25]. Second generation bisphosphonates, including alendronate, pamidronate, and risedronate, each contain a hydroxyl group on the R_1 side chain and have a chain containing nitrogen on R_2. The hydroxyl group on the carbon atom causes the increased affinity for bone and the R_2 chain determines the potency and mode of action of the drug [25]. The P-C-P

chemical substructure enhances the safety and efficacy of the compound by increasing the affinity for calcium and resisting the metabolic processes of the body [25]. Although newer bisphosphonates have a very high affinity for mineral, only about half of any dose reaches bone [25]. Once they do reach bone and bind to exposed mineral, they can be uptaken by osteoclasts. While bound to osteoclasts, they exert their inhibiting effect by interfering with enzymatic activity.

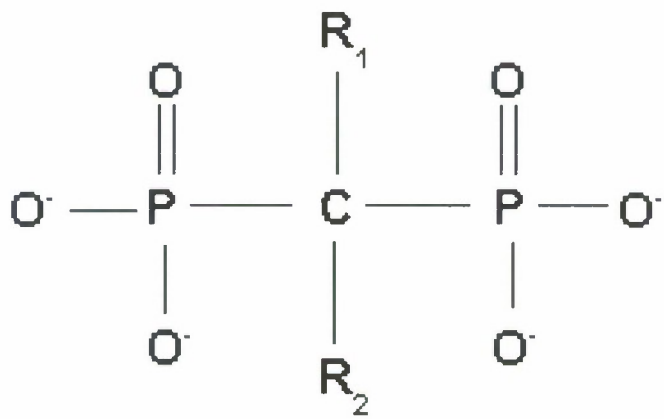


Figure 1.4. Chemical structure of bisphosphonates.

Bisphosphonates increase bone mass by reducing the temporary porosity created by remodeling known as the remodeling space, and by creating a positive bone balance where more bone is added than is removed [23,26]. Experimental data exists depicting the resulting bone mass increase in patients with osteoporosis. Spinal BMD increased 8.8 percent and reduced vertebral fractures by 48 percent after 3 years of alendronate treatment [27]. Another study reported increased spinal BMD throughout 7 years of alendronate treatment on postmenopausal women [28].

Bisphosphonates also can have negative effects on bone. Lowered remodeling levels increase microdamage accumulation and reduce toughness because less damage is removed [23]. Mineralization of the bones also increases, leading to brittler bones that

are less resistant to crack propagation. One-year studies of alendronate and risedronate treatment on dogs showed microdamage accumulation and BMU activation frequency to be inversely proportional [29]. The overall goal of these therapies is to minimize the amount of damage accumulated while increasing bone mass. In the long run, high potency bisphosphonate treatment may in fact be detrimental if bone quality degrades enough to increase fracture risk due to high amounts of microdamage. Long-term effects of bisphosphonates on bone remodeling still remain unclear. Since studies can take up to 10 years to acquire real data, mathematically modeling the data and relationships obtained from shorter studies may be a more effective method of gathering insight into these phenomena.

1.4 Previous Models

Computational models are commonly used to test theories regarding the adaptation of bone to mechanical and physiological stimuli. Early models tested bone's mechanical adaptation. Carter et al. [30] and Huiskes et al. [31] were the first to use finite element modeling to develop mathematical relationships between mechanical loading and trabecular bone density [7]. In their model, Carter et al. [30] focused on a daily mechanical stimulus based on stress and an error function. The error function tracked the difference between the stress stimulus at a given time point and a predetermined equilibrium stress stimulus. The apparent density of each element in the model was adjusted according to the error function. New modulus values were calculated by multiplying the changes in density by a constant. These new values were used at the next time point to recalculate the daily stress stimulus for each element.

Huiskes' approach to model bone [31] has been shown to be the equivalent of Carter's but with different coefficients [7]. The model invoked bone adaptation by using strain energy density as the mechanical stimulus. New modulus values were calculated from strain energy density as in Carter's model [30] described above. The purpose of this particular model was to predict and analyze changes in bone due to total hip arthroplastics. Though both of these models were sufficient at simulating the mechanical environment that bone is subjected to, they did not include adaptations due to cellular responses.

A model by Hazelwood et al. [18] incorporates responses to both mechanical and biological stimuli. The model takes into account the cellular responses of BMUs. It predicts changes in porosity and elastic modulus based on simulated responses to mechanical stimuli such as disuse or overload and to the biological stimulus of damage accumulation. The schematic below (Figure 1.5) shows the basis for the simulation. Note that damage and disuse affect BMU activation frequency, which in turn affects porosity and modulus. The schematic shows remodeling to be a dynamic loop.

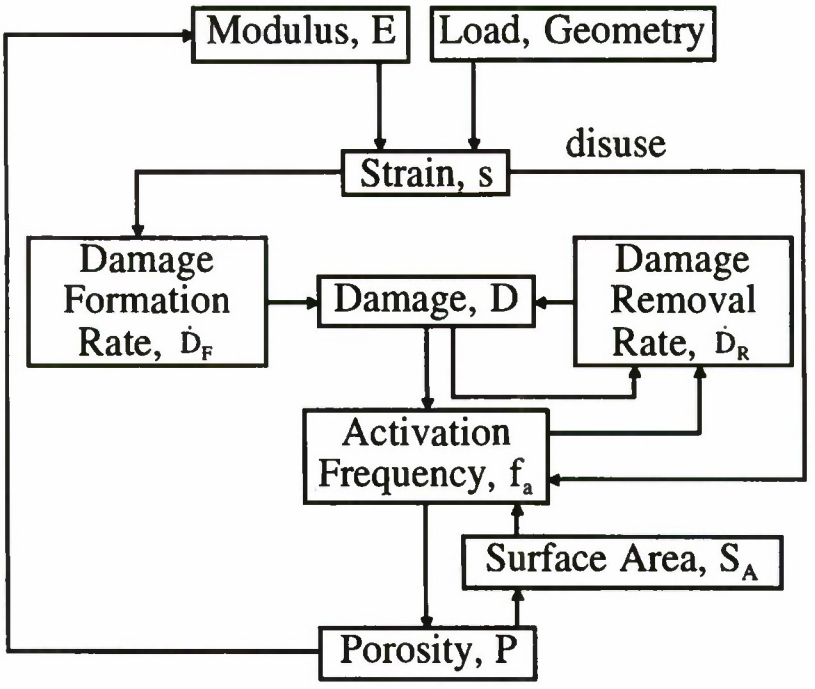


Figure 1.5. Schematic of a bone remodeling algorithm by Hazelwood et al. [18].

In early computational models simulating bisphosphonate effects, the basic strategy was to match clinical results of the treatment. The models simulated varying parameters such as BMU activation frequency, resorption and formation periods, bone balance, and mineralization [32,33,34,23], but failed to include other stimuli important to long-term changes to BMD. These models did not take into account bone remodeling due to mechanical loading nor microdamage accumulation. Heaney et al. [32] developed a model that was fairly accurate when predicting bisphosphonate effects on BMD out to 6 months. The model simulated treatment using the bone balance method of decreasing bone turnover, decreasing remodeling space, increasing focal bone balance, and keeping bone mineralization constant. Hernandez et al. [33] developed a model to compare this method with the mineralization method in which remodeling space is decreased, focal bone balance remains the same, and bone mineralization is varied. This model included a longer secondary mineralization period to account for increases to BMD in the long term.

The schematic shown in Figure 1.6 shows the lack of mechanieal stimuli in the model developed by Hernandez et al. [33]. Laey et al. [34] developed a model of bisphosphonate treatment in whieh various parameters were elosely analyzed. Aetivation frequeney was found to have the greatest influenee on predieting ehanges in bone volume [34].

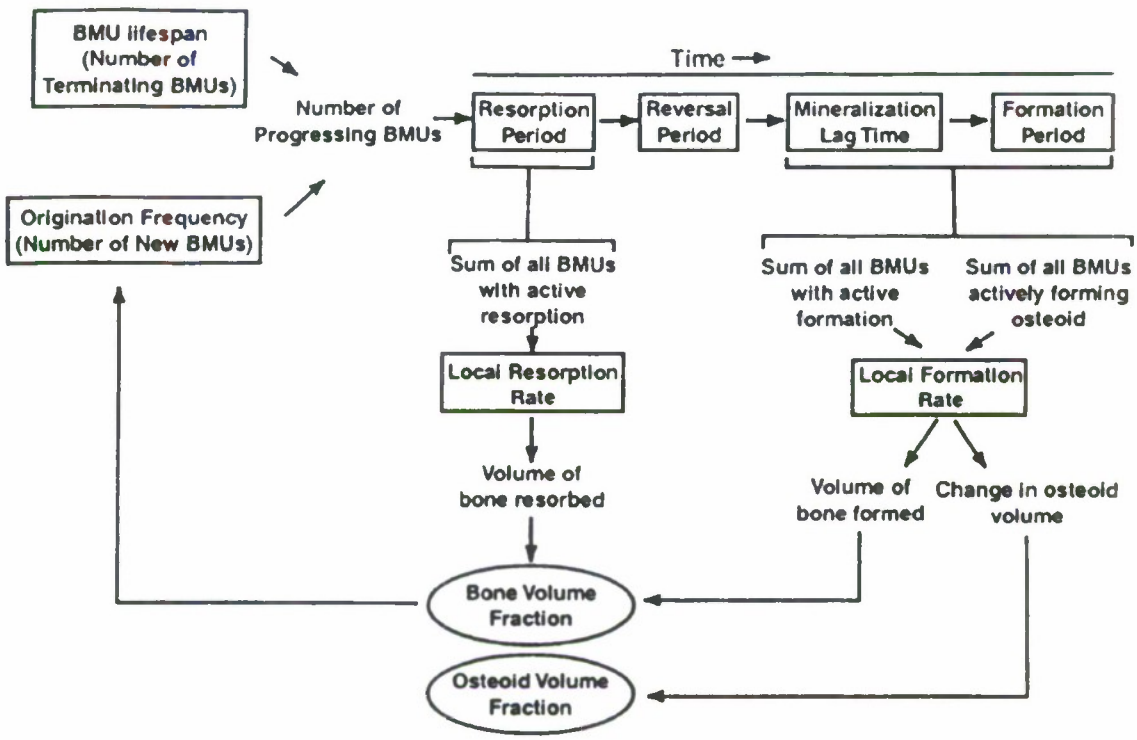


Figure 1.6. Schematic of a bone remodeling algorithm by Hernandez et al. [33].

A model developed by Nyman et al. [23] utilized both mechanisms of remodeling deseribed previously, in response to meehanieal loading and damage aeeumulation, to simulate the long-term effects of bisphosphonate usage. The model was designed for examining postmenopausal osteoporosis, so it also ineluded a bone remodeling response to estrogen deficiency. The results of the model showed that the reduetion of resorption space caused by bisphosphonate treatment played a key role in increasing bone volume over the long run. Also, Nyman et al. [23] concluded that the disuse response eontributed

to the eventual plateau of bone gain that other models failed to achieve [32,33]. Overall, the model showed incomplete suppression of BMU activation frequency and positive bone balances to permanently increase bone volume with minor gains in microdamage accumulation [23]. Various aspects of this model are implemented in this study and will be covered in more detail in the next section and in the methods chapter.

1.5 Simulating Bisphosphonate Treatment during Spaceflight

The goal of this study is to create a computer model to simulate bisphosphonate treatment on astronauts while traveling in space in order to examine bisphosphonates' ability to maintain bone mass in a microgravity environment and reduce fracture risk of bone upon return to Earth. We will examine various bisphosphonate treatment potencies and bone balance ratios given at different time points (either at or before spaceflight). Flight duration will also be varied to examine short-term (10 days) to long-term (1 year) effects of microgravity on BMD and damage accumulation. To examine these effects, mathematical relationships modeling bisphosphonate therapies [23] were applied to a computational model developed by Hazelwood et al [18] that was modified to investigate the influence of microgravity on the bone remodeling process.

This simulation will not only increase the understanding of the effects of spaceflight and bisphosphonate usage, but it may also lead to advances of other proposed treatments such as parathyroid hormone, testosterone, and vitamin K_2 [6]. Understanding each therapy will eventually lead to better treatments on Earth and in space, decrease the health risks of space travel, and drastically expand the bounds of human space exploration.

CHAPTER 2: METHODS

2.1 Mechanical Loading

This model simulates cyclic uniaxial loading of a volume of vertebral trabecular bone (Figure 1.1). Applying data from previous studies [35,36,37], and assuming a linear relationship between apparent density and porosity allows apparent stiffness (elastic modulus, E) to be determined,

$$E = E_0 \times (1 - p)^b, \quad (1)$$

in units of MPa, where p is porosity, and 14927 and 1.33 are values of E_0 and b, respectively, for trabecular bone [18]. Bone mineral density (BMD) was calculated by a porosity relationship,

$$BMD = \rho \times (1 - p), \quad (2)$$

where ρ is apparent density of bone (2.0 g/cm^3 when $p = 0$) [18]. Apparent density is the measure of mass per unit volume, including the voids spaces within the material.

Peak strain was calculated using Hooke's Law,

$$\varepsilon = \sigma / E. \quad (3)$$

The mechanical stimulus modeled to stimulate the bone volume's response to remodeling was defined as

$$\Phi = R_L \times \varepsilon^q, \quad (4)$$

where R_L is the loading frequency in cycles per day and q adjusts the peak strain and loading frequency to correctly model the loading potential [18].

2.2 Porosity Transformation

The rate of change of porosity ($\dot{p}$) is the difference between the amount of bone formation and amount of bone resorption per given time period [38,39]. It is defined as

$$\dot{p} = Q_R N_R - Q_F N_F, \tag{5}$$

where Q_R and Q_F are the mean bone resorbing and refilling (forming) rates, and N_R and N_F are the densities of resorbing and refilling BMUs per area, respectively. The rate of resorption,

$$Q_R = A_R / T_R, \tag{6}$$

and rate of refilling,

$$Q_F = A_F / T_F, \tag{7}$$

are assumed to be linear in time [40]. The area of bone resorbed (A_R) and area of bone formed (A_F) were based on a cement line radius of 0.095 mm [40]. When bone is in a disuse state ($\Phi < \Phi_0$), refilling is reduced on bone surfaces [41]; thus, area of bone formed during disuse was reduced to $A[0.5 + 0.5(\Phi/\Phi_0)]$ [18]. The resorption (T_R) and refilling (T_F) periods were 25 and 64 days, respectively [23]. These periods were used to calculate total number of resorbing BMUs and refilling BMUs per given area. Integrating BMU activation frequency, f_a (BMUs/area/time), over a known time period will result in the number per section area of resorbing BMUs,

$$N_R = \int_{t-T_R}^{t} f_a(t')dt', \tag{8}$$

and the number per section area of refilling BMUs,

$$N_F = \int_{t-(T_R+T_I+T_F)}^{t-(T_R+T_I)} f_a(t')dt'. \tag{9}$$

Present time is t and T_l is a latency period between resorption and refilling [23].

2.3 BMU Activation

BMU activation frequency, as previously mentioned, is the number of active BMUs in the section area per day. It is assumed to be a function of the two remodeling mechanisms modeled by Hazelwood et al. [18], damage and disuse. Activation frequency,

$$f_a = (f_{a(disuse)} + f_{a(damage)})S_A, \quad (10)$$

is also a function of internal surface area as BMUs must begin on a bone surface [18]. S_A is internal surface area per unit volume normalized to values between 0 and 1 by $S_{A\max}$. It was determined by using a porosity-surface area relationship developed by Martin [42]

$$S_A = (32.1p - 93.9p^2 + 134p^3 - 101p^4 + 28.8p^5) / S_{A\max}. \quad (11)$$

This accounts for decreased remodeling in bone volumes with smaller surface areas.

Changes in activation frequency due to damage and disuse were modeled assuming sigmoidal relationships [18] between damage and $f_{a(damage)}$,

$$f_{a(damage)} = \frac{(f_{a0})(f_{a(\max)})}{f_{a0} + (f_{a(\max)} - f_{a0})e^{[k_r(f_{a(\max)})(D-D_0)/D_0]}}, \quad (12)$$

and mechanical stimulus and $f_{a(disuse)}$,

$$f_{a(disuse)} = \frac{f_{a(\max)}}{1 + e^{k_b(\Phi - k_c)}} \text{ for } \Phi < \Phi_0. \quad (13)$$

The k values were determined by matching the curves to clinical data [18]. Coefficients k_r, k_b, and k_c define the shape, slope, and inflection point of the curves, respectively. The maximum activation frequency ($f_{a(max)}$) of 0.50 BMUs/mm^2/day was higher than the

highest measured activation frequency (0.14 BMUs/mm^2/day) for human cortical bone because it is assumed the measurements did not reach the upper limits [43,18].

2.4 Microdamage Accumulation

Microdamage (D) is defined as total crack length per section area of bone. Damage accumulation due to fatigue is modeled according to Martin's work [44] as

$$\dot{D} = \dot{D}_F - \dot{D}_R \quad (14)$$

where $\dot{D}_F$ and $\dot{D}_R$ are the damage formation and removal rates, respectively. Based on Martin's findings, we assumed the rate of damage formation to be proportional to the loading potential [18],

$$\dot{D}_F = k_D \times \Phi. \quad (15)$$

This model assumes a random distribution of BMUs and damage in the section area of bone; however, a damage removal specificity factor is included when modeling the rate of removal due to evidence showing that damage initiates the activation of BMUs [9,10, 14,15,18],

$$\dot{D}_R = D f_a A_R F_s. \quad (16)$$

The specificity factor, F_s, was assumed to be 5 based on experimental data [44]. Initial equilibrium is defined here as the time at which the rate of damage formation is equal to the rate of damage removal [18]. Setting both equations equal allows us to determine the damage rate coefficient,

$$k_D = D_0 f_{a0} A F_s / \Phi, \quad (17)$$

where initial, equilibrium values are designated by the subscript 0. These values were obtained from Hazelwood et al. [18].

2.5 Preflight Conditions

Before applying microgravitational conditions, preflight parameters were calculated (Table 2.1) using constants derived from experimental data (Table 2.2). These parameter values were obtained by executing the model developed by Hazelwood et al. [18] until the values reached equilibrium. The applied stress on earth modeled to reach these values was determined by matching the predicted bone porosity of the model to 0.78, a porosity typical of vertebral trabecular bone in adult males in Earth's gravitational environment [45,46,23]. The calculated applied stress, 1 MPa, was based on a 100 mm^2 cross-section from the modeled 1 cm^3 trabecular bone volume (Figure 2.1) [23]. The resulting preflight BMD was 0.44 g/cm^3.

Table 2.1. Model parameters with values obtained from simulating remodeling in trabecular bone before entering space and at the end of a 180-day space simulation.

Parameter (units)	Description	Preflight trabecular bone	End of flight values (180-day)
E (MPa)	Elastic modulus	1992.3	1881.9
p	Porosity	0.780	0.789
BMD (g/cm^3)	Bone mineral density	0.4400	0.4215
ε (10^{-6})	Microstrain	501.9	473.4
Φ (10^{-10})	Loading potential	1.904	1.507
D (mm/mm^2)	Total crack length per section area	0.0375	0.0327
$\dot{D}_F$ ($mm/mm^2/day$)	Damage formation rate	0.0000353	0.0000280
$\dot{D}_R$ ($mm/mm^2/day$)	Damage removal rate	0.0000353	0.0000417
f_a ($BMUs/mm^2/day$)	BMUs appearing in the section area per day	0.0133	0.0180
N_F ($\#/mm^2$)	Number of refilling sites	0.8495	1.2825
N_R ($\#/mm^2$)	Number of resorbing sites	0.3318	0.4483

Table 2.2. Remodeling simulation constants.

Constant (units)	Description	Nominal value	Source
A_F (mm^2)	Area of formation	1.418 x 10^{-2}	[23]
A_R (mm^2)	Area of resorption	1.418 x 10-^{2}	[23]
$A_{R(1.2)}$ (mm^2)	Area of resorption (reduced by 1/6)	1.181 x 10-^{2}	[23]
$A_{R(1.3)}$ (mm^2)	Area of resorption (reduced by 3/13)	1.090 x 10-^{2}	[23]
T_R (days)	Resorption period	25	[23]
T_I (days)	Reversal period	5	[23]
T_F (days)	Refilling period	64	[23]
k_D (mm/mm^2)	Damage rate coefficient	~1.85 x 10^5	[18]
R_L (cpd)	Loading rate	3000	[18]
q	Damage rate exponent	4	[47]
F_s	Damage removal specificity factor	5	[44]
D_0 (mm/mm^2)	Initial damage	0.0366294	[14,48]
f_{a0} (BMUs/mm^2/day)	Initial BMU activation frequency	0.0067	[49]
Φ_0 (cpd)	Initial mechanical stimulus	1.875 x 10^{-10}	[18]
σ_e (MPa)	Stress applied on earth	1	
σ_s (MPa)	Stress applied in space	0.8909	
$f_{a(max)}$ (BMUs/mm^2/day)	Maximum BMU activation frequency	0.5	[18]
$S_{A(max)}$	Max. specific surface area, normalizing constant	4.195	[18]
k_b (cpd^{-1})	Activation frequency dose-response coefficient	6.5 x 10^{10}	[18]
k_c (cpd)	Activation frequency dose-response coefficient	9.4 x 10^{-11}	[18]
k_r	Activation frequency dose-response coefficient	-1.6	[18]

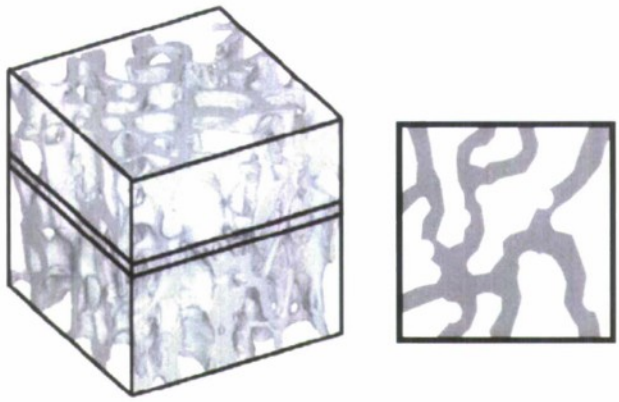


Figure 2.1. Trabecular volume and cross-section [23].

2.6 Simulation of Spaceflight

Microgravity was simulated by lowering the stress applied to the bone volume. Astronauts experieneed an average vertebral trabeeular BMD loss of 0.7% per month, for a total of 4.2% BMD loss during the 180-day average ISS missions [3,5,6]. For this simulation, stress applied to the representative volume during spaceflight was then determined as 0.8909 MPa based on this 180-day spaeeflight ending density of 0.4215 g/em^3.

Using the density loss values at 180 days as a baseline, various durations of spaeeflight were simulated based on typieal mission length for astronauts, ineluding 10 days, 90 days, and 180 days [5,4]. Though 365 days in space is not typical, it was also simulated to examine the potential of bisphosphonates to maintain BMD without inereasing damage.

2.7 Simulation of Bisphosphonates

Bisphosphonate treatment was simulated using two factors: one lowering aetivation frequeney and the other reducing the resorption area. A potency variable (P),

where $0 \leq P \leq 1$, is applied to exert the former effect by multiplying f_a by the quantity (1 – P). P is based on pharmacokinetic properties of bisphosphonates, including potency factors, binding, uptake, and mode of action [23],

$$P = P_{max}(1 - e^{-\tau_S \times N_R}). \quad (18)$$

P_{max} and τ_S are suppression coefficients reflecting various properties of bisphosphonates in order to model a range of drug potencies. Values for these coefficients (Table 2.3) were selected based on the experimental results from 1 year studies of daily alendronate and daily pamidronate treatment, and modeled the variations in the reduction of activation frequency as seen in these studies [23]. Figure 2.2 displays the effects of the coefficients on the relationship between P and N_R.

Table 2.3. Bisphosphonate effects analyzed during the simulation.

Treatment	Label	Level of suppression	P_{max}	τ_S	Initial bone balance
no treatment	P0a1	-	0	0	1
A	P07t5a12	Low	0.7	5	1.2
B	P07t20a12	Intermediate	0.7	20	1.2
C	P1t5a12		1.0	5	1.2
D	P1t20a12	High	1.0	20	1.2
E	P07t5a13	Low	0.7	5	1.3
F	P07t20a13	Intermediate	0.7	20	1.3
G	P1t5a13		1.0	5	1.3
H	P1t20a13	High	1.0	20	1.3

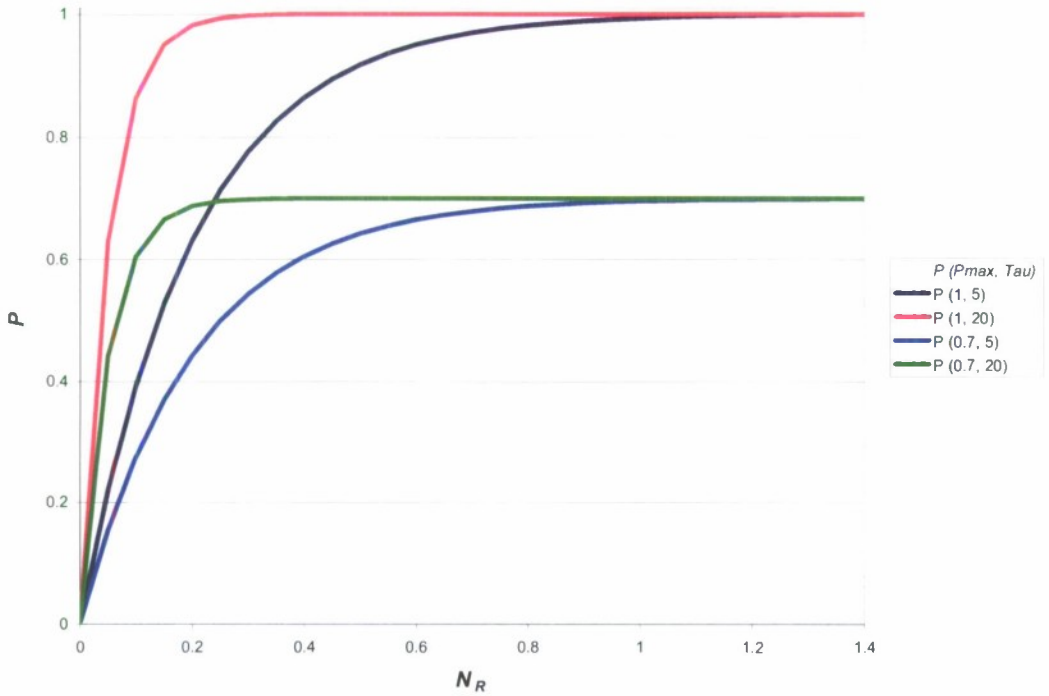


Figure 2.2. Bisphosphonate potency as a function of number of resorbing BMUs.

The size of the resorption cavity is reduced during bisphosphonate treatment due to their effects on osteoclasts [23], resulting in alterations to the ratio of bone area formed to bone area removed. Two different initial bone balance ratios (A_F/A_R) for the simulation of bisphosphonate treatment were used (Table 2) based on reductions of 1/6 and 3/13 to resorption area found in postmenopausal women treated with bisphosphonates for 1 year [23]. A bone balanee of 1.0 is assumed for the simulation when bisphosphonate treatment is not in effeet.

Bisphosphonates were simulated during the entire spaeeflight. Preflight treatment was also examined, where bisphosphonates were applied to the simulation at 0, 7, 14, 30, 90, and 180 days preflight.

2.8 Simulation of Return to Earth

The return to earth was modeled by resuming preflight bone loading conditions, with an applied stress of 1 MPa. Once back on Earth, bisphosphonate therapy was discontinued. The simulation was extended 365 days postflight to examine increases or decreases in fracture risk based on bone mineral density and damage accumulation. Though it takes 1 to 3 years to fully recover without treatment, a one year postflight examination allowed us to determine if a treatment is successful at maintaining bone mass without accumulating more damage. A successful treatment will cause the remodeling properties to reach new equilibrium values within one complete remodeling cycle (3 to 4 months) after arriving back on Earth.

2.9 Model Implementation

The computational model was coded in MATLAB (Appendix A). The time increment for which all the model variables were updated and tracked was 1 day. The computational simulation was executed in Windows Vista (32-bit) using an Intel Core 2 Duo processor.

2.10 Postflight Parameter Analysis

Together, bone mass (quantity) and bone quality determine the ability of bone to resist fracture [50,51]. For this simulation, we examined changes in bone mass and bone quality in terms of BMD and microdamage accumulation, respectively. Experimental results from ex vivo studies have shown BMD to predict 66 to 74 percent of the variation in bone strength [52]. An increase in bone mass and bone quality will lead to stronger,

more fracture resistant bone; thus, in the analysis we assumed higher BMD and lower damage accumulation to lead to higher fracture resistance.

CHAPTER 3: RESULTS

3.1 Untreated Spaceflight

The model's results were consistent with experimental data obtained from spaceflight [4] in that the predicted BMD loss was non-linear (Figure 3.1). Predicted damage accumulation decreased as flight duration increased (Figure 3.1) due to the disuse response in which damage was specifically targeted during resorption and because less damage formed each day due to the microgravity environment (Figure D25). The predicted rate of BMD loss was greatest early on in the flight, showing smaller decrements as time spent in space increased. The model predicted little BMD loss and change from preflight damage for the typical flight duration of 10 days (Figure 3.1); however, postflight predictions showed that BMD continued to decrease upon return to Earth until it reached a value nearly 3 times lower than it was at the end of the mission before increasing to near normal levels about 100 days after entering Earth's gravity (Figure B1).

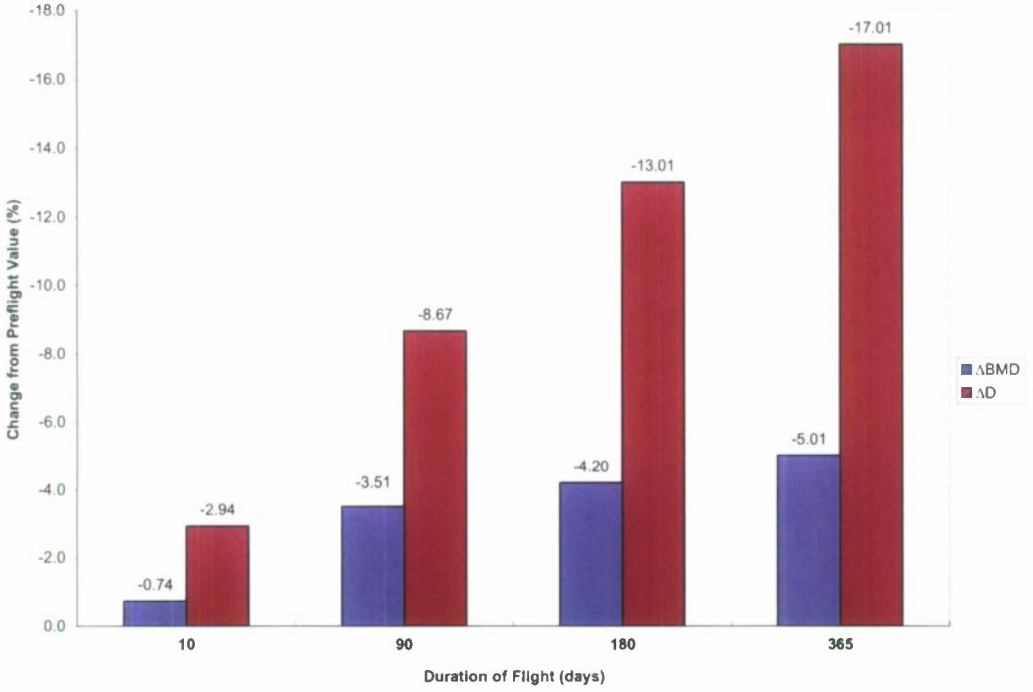


Figure 3.1. Predicted percent decreases in BMD and damage (D) of untreated bone at end of spaceflight.

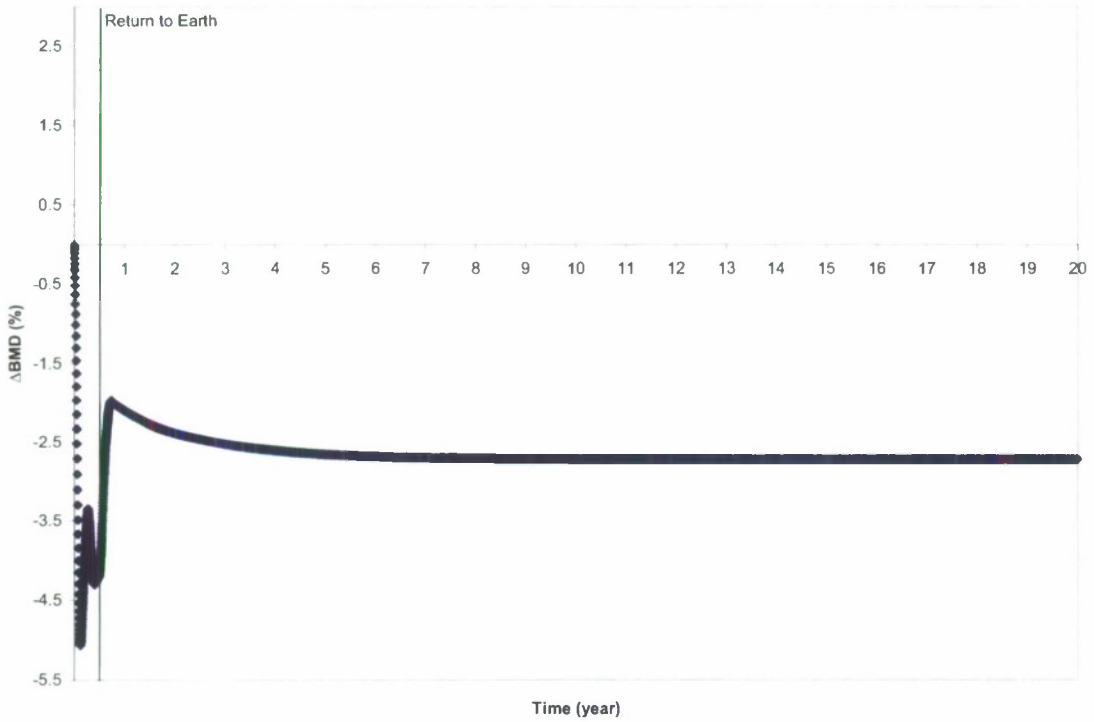


Figure 3.2. Predicted effects of 180-day spaceflight on BMD of untreated bone from beginning of flight through 19.5 years postflight.

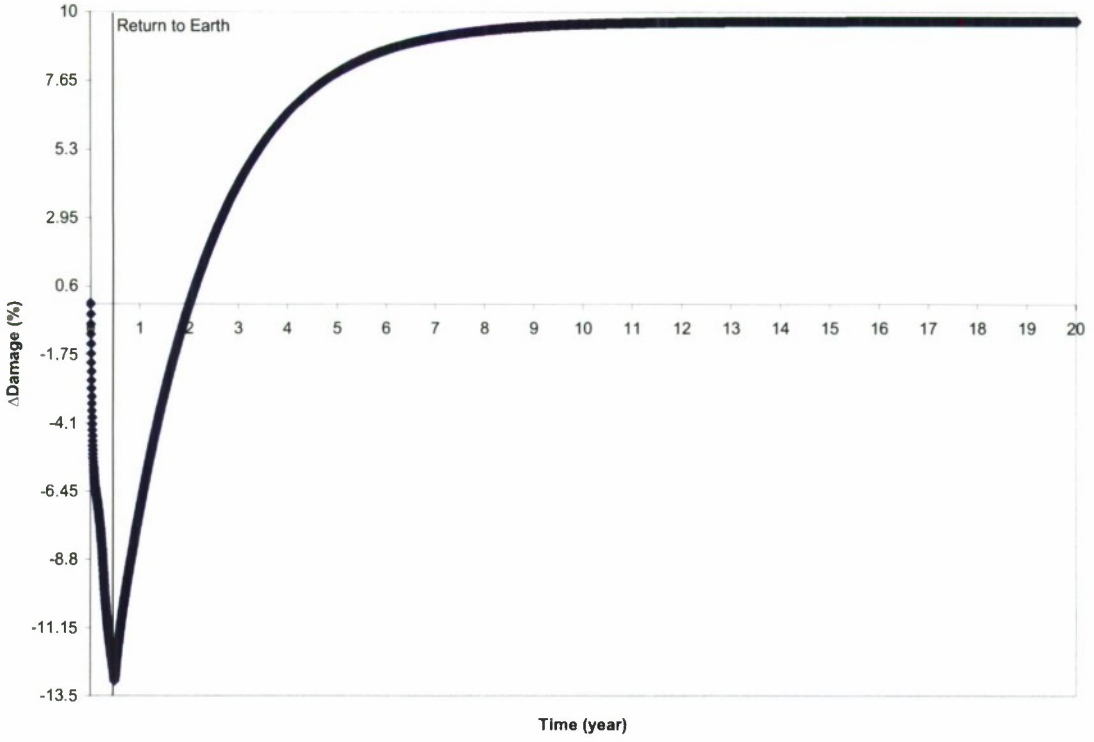


Figure 3.3. Predicted effects of 180-day spaceflight on damage accumulation (D) of untreated bone from beginning of flight through 19.5 years postflight.

For the 180-day space mission, the predicted postflight results for untreated bone showed sharp increases to both BMD and damage upon return to Earth (Figures 3.2 and 3.3). BMD reached equilibrium approximately 7 years postflight, and was about 2.7% less than its preflight value. Damage continued to accumulate postflight until equilibrium was reached about 15 years after the return to Earth. Though damage was initially lower upon return to earth, it continued to increase postflight until reaching a value approximately 9.5% higher than its preflight value. Figures 3.2 and 3.3 clearly indicate a need for treatment as bone was negatively affected by decreased mass and quality.

Also noteworthy is the predicted oscillatory behavior of BMD values seen in disuse (Figures 3.2 and 3.15) for untreated bone. The inflection points are located approximately at the same intervals as the transitions from resorption to refilling and

refilling back to resorption in the BMU remodeling period. This oscillatory trend was also noted in the model by Hazelwood et al. [18] and still needs further investigation.

3.2 Bisphosphonate Treatment Onset at Beginning of Spaceflight

The simulation predicted gains in BMD for shorter flights (10 and 90 days) in which treatments exhibited intermediate to high suppression (Figures 3.4 and 3.5). For all flight durations, high suppression of BMU activation (Table 2.3) led the model to predict increases in both BMD and damage accumulation. Overall, the model predicted treatments creating bone balances of 1.3 to have more positive effects on BMD (i.e. less loss or greater increase). Although those with bone balances of 1.3 positively affected BMD, the model predicted that they also caused more damage to accumulate when the level of remodeling suppression was high. In general, the model predicted treatments with intermediate levels of suppression to have smaller changes in BMD at the end of spaceflight than the decreases predicted with low levels of suppression and than the increases predicted with high levels of suppression. Also, the simulation predicted that treatments with intermediate suppression caused end-of-flight BMD and damage accumulation values of to remain closer to the pretreatment, preflight values than treatments with low or high remodeling suppression.

As flight duration increased, predicted percent changes in damage and BMD increased. Simulations of long flight durations (180 and 365 days) resulted in significant changes, predicting treatments with low and intermediate suppression to have significant decreases in damage (up to 14.5%) similar to untreated bone. The greatest predicted increases in both damage and BMD occurred for treatments of high suppression and bone balances of 1.3. Simulation of treatments C, E, F, and G (P1t5a12, P07t5a13, P07t20a13,

and P1t5a13), with low and intermediate levels of remodeling suppression, showed increases in BMD and decreases in damage accumulation on longer duration spaceflights (Figures 3.6 and 3.7).

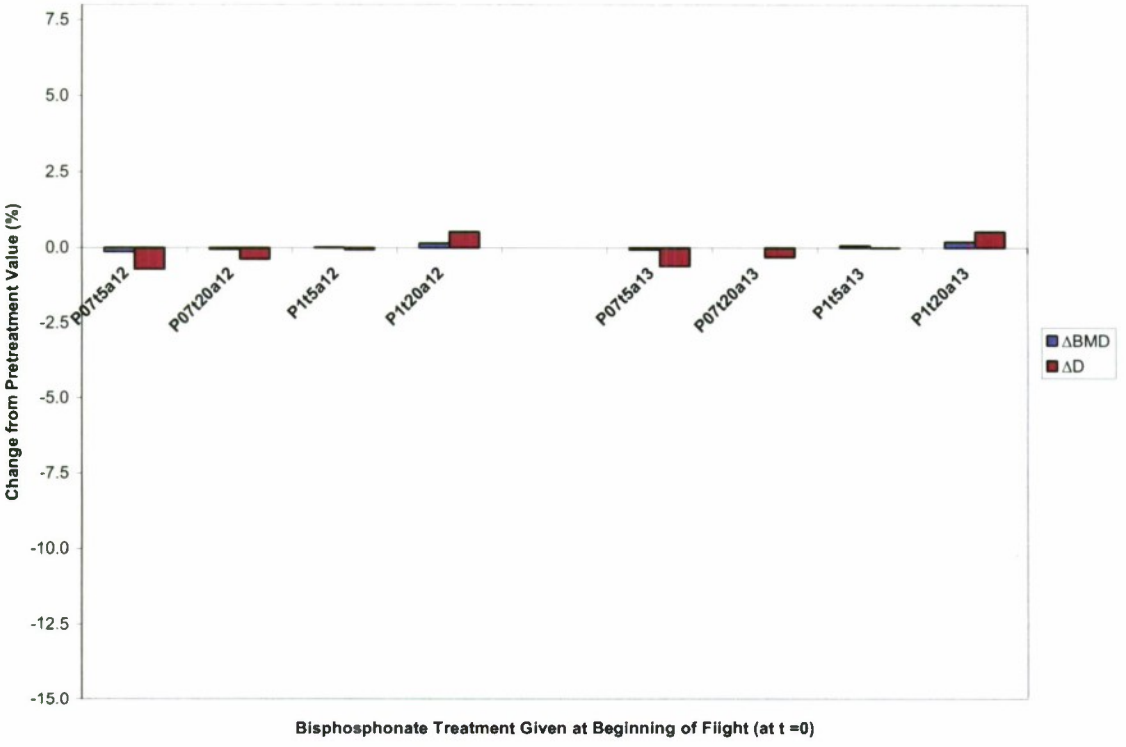


Figure 3.4. Predicted percent changes in BMD and damage (D) at end of 10-day spaceflight.

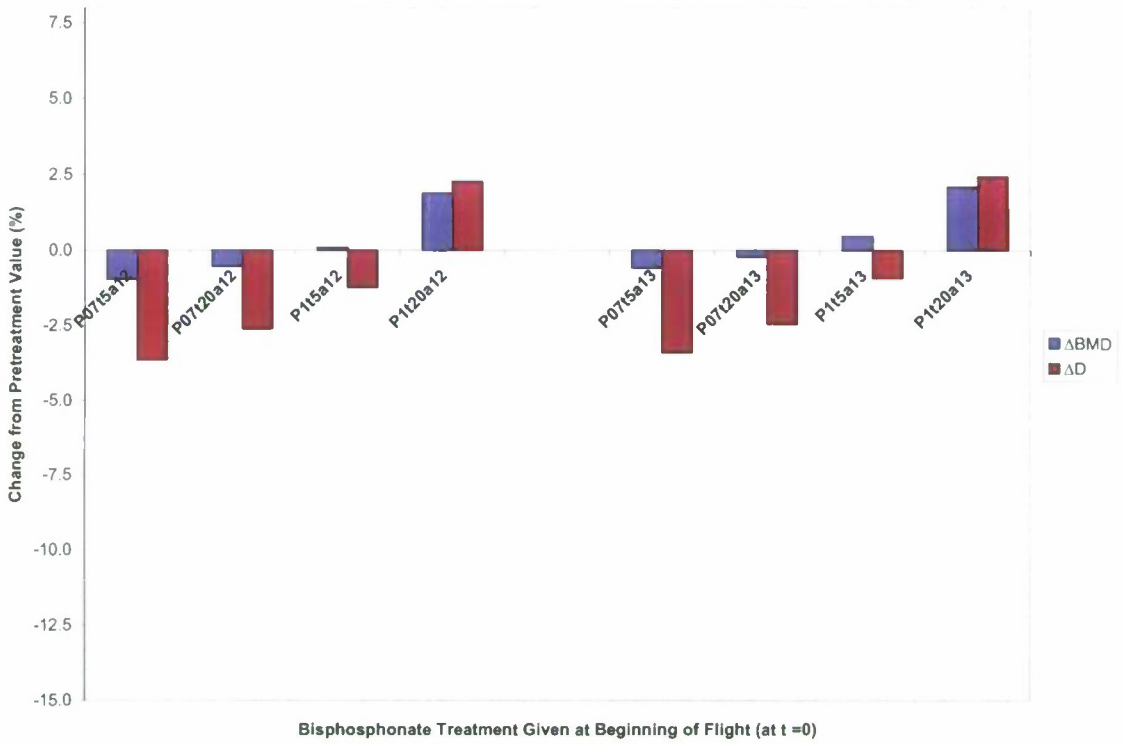


Figure 3.5. Predicted percent changes in BMD and damage (D) at end of 90-day spaceflight.

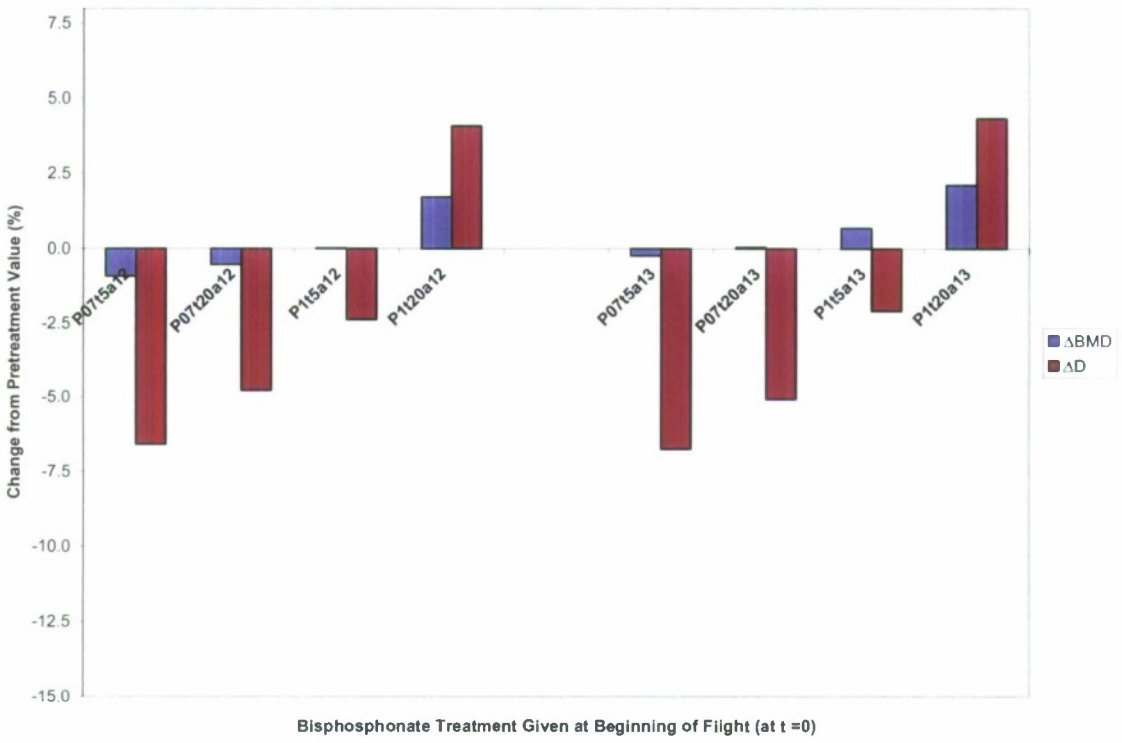


Figure 3.6. Predicted percent changes in BMD and damage (D) at end of 180-day spaceflight.

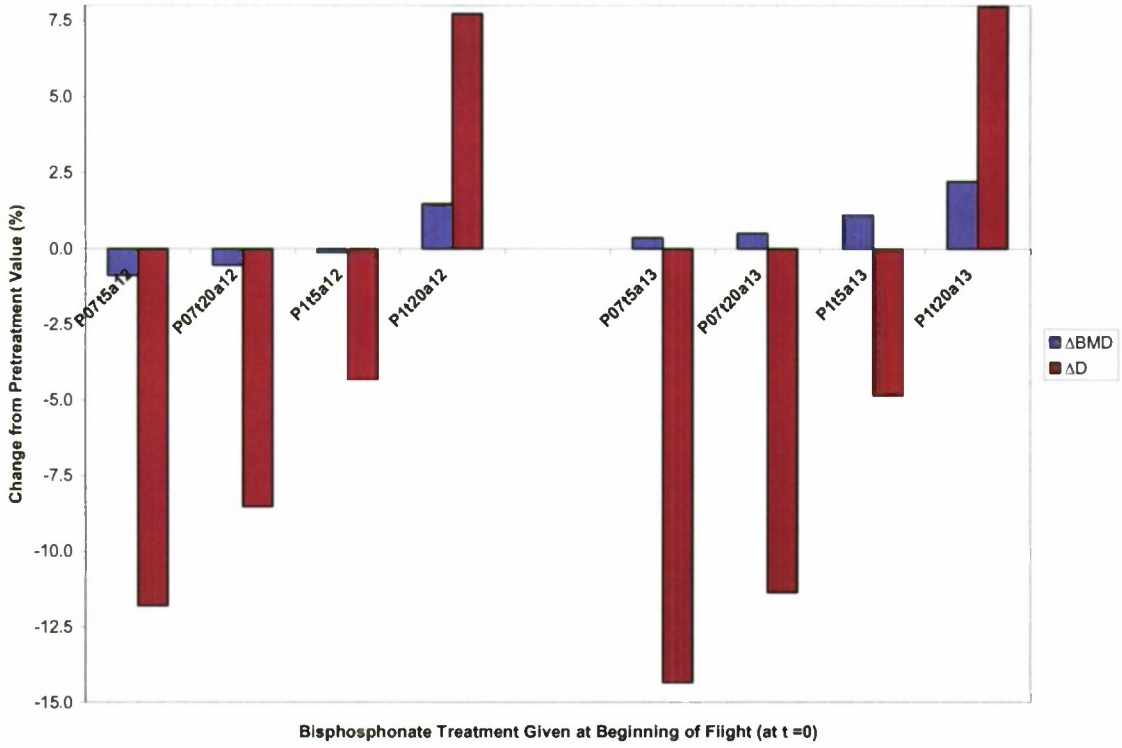


Figure 3.7. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight.

3.3 Preflight Bisphosphonate Treatment

In accordance with data obtained from clinical studies [21], the predicted results of bisphosphonate treatment on Earth exhibited increases in BMD and damage accumulation. Predicted pre-spaceflight additions to BMD and damage continued to rise as the duration of the preflight therapy simulation increased from 7 to 180 days (Figures 3.8-3.12). The model showed damage to increase at a faster rate than BMD. The ratio of percent increase in damage to percent increase in BMD, used in this study as a measure indicating efficiency of reducing fracture risk in which a lower value is more optimal, was highest for treatments with high levels of remodeling suppression, and especially for those also with lower bone balances (Table 3.1). The lowest ratio of percent preflight increase of damage to BMD was predicted for Treatment E (P07t5a13), which modeled

low remodeling suppression and a bone balance ratio of 1.3. Bisphosphonate treatments with lower suppression of BMU activation and higher bone balance ratios proved to be optimal by adding more healthy bone per increase in BMD.

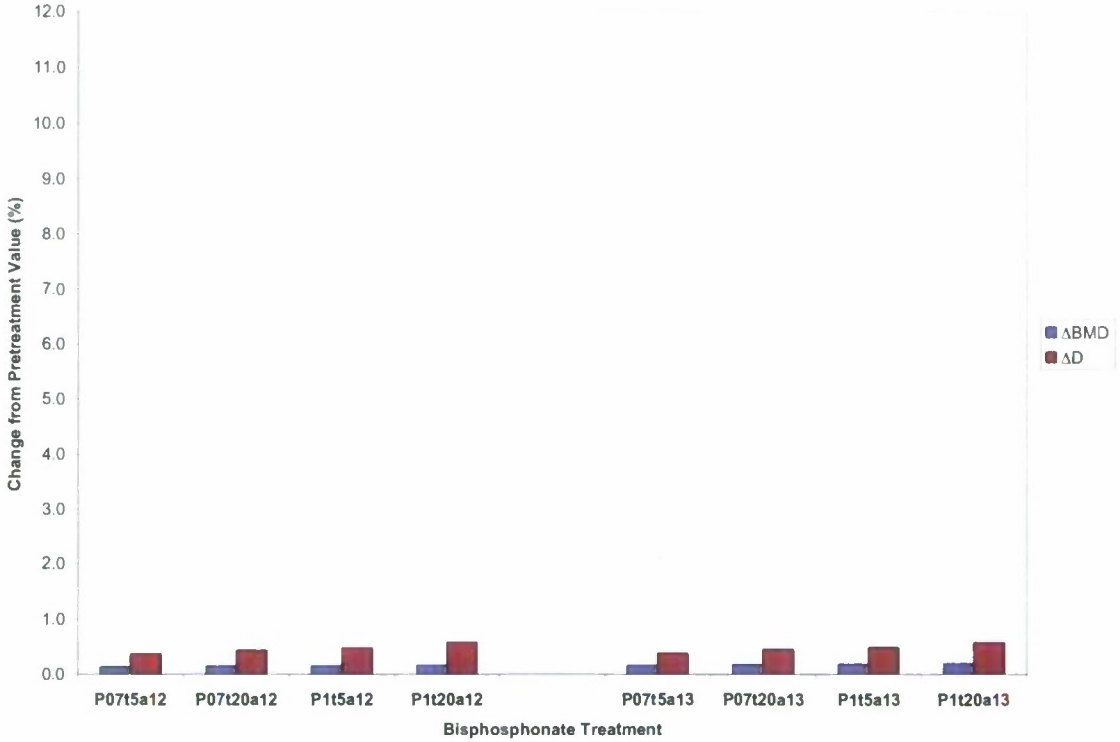


Figure 3.8. Predicted preflight increase in BMD and damage (D) due to 7-day preflight treatment.

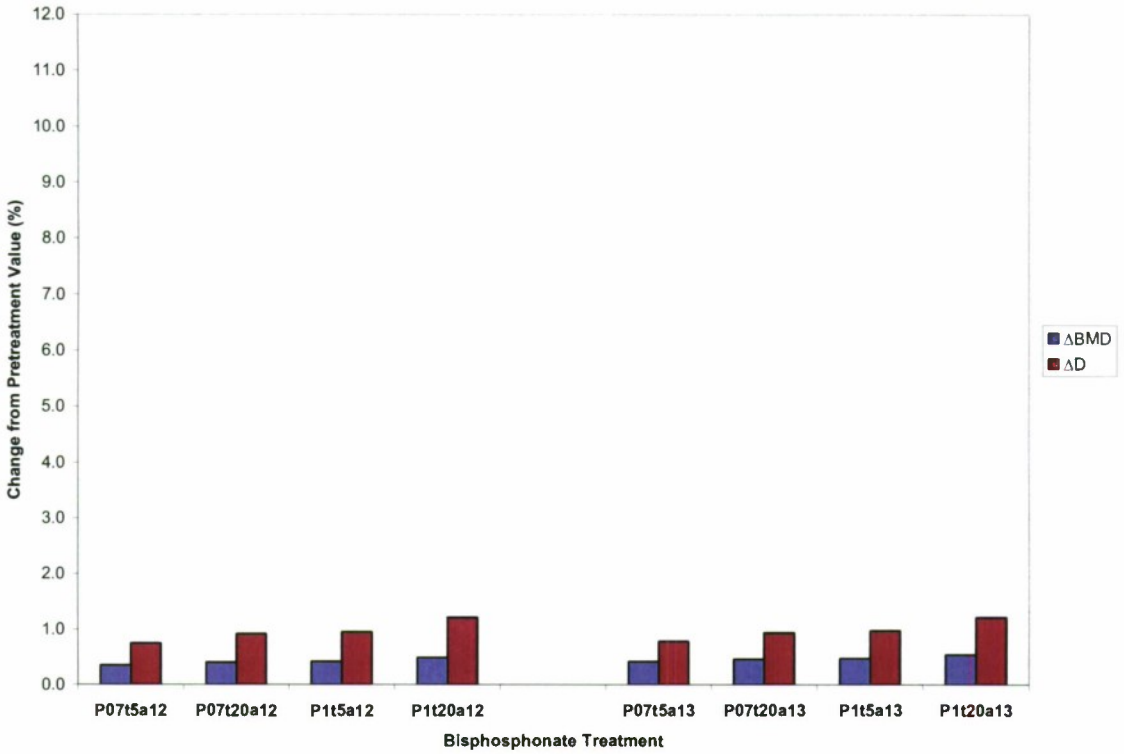


Figure 3.9. Predicted preflight increase in BMD and damage (D) due to 14-day preflight treatment.

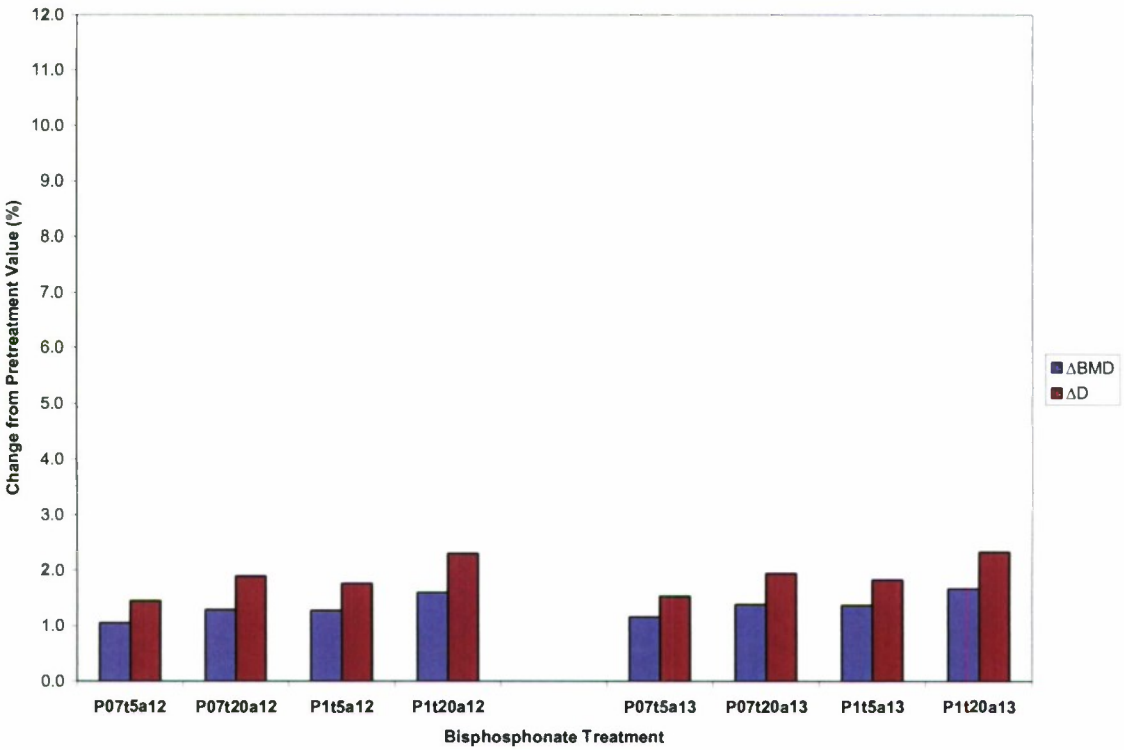


Figure 3.10. Predicted preflight increase in BMD and damage (D) due to 30-day preflight treatment.

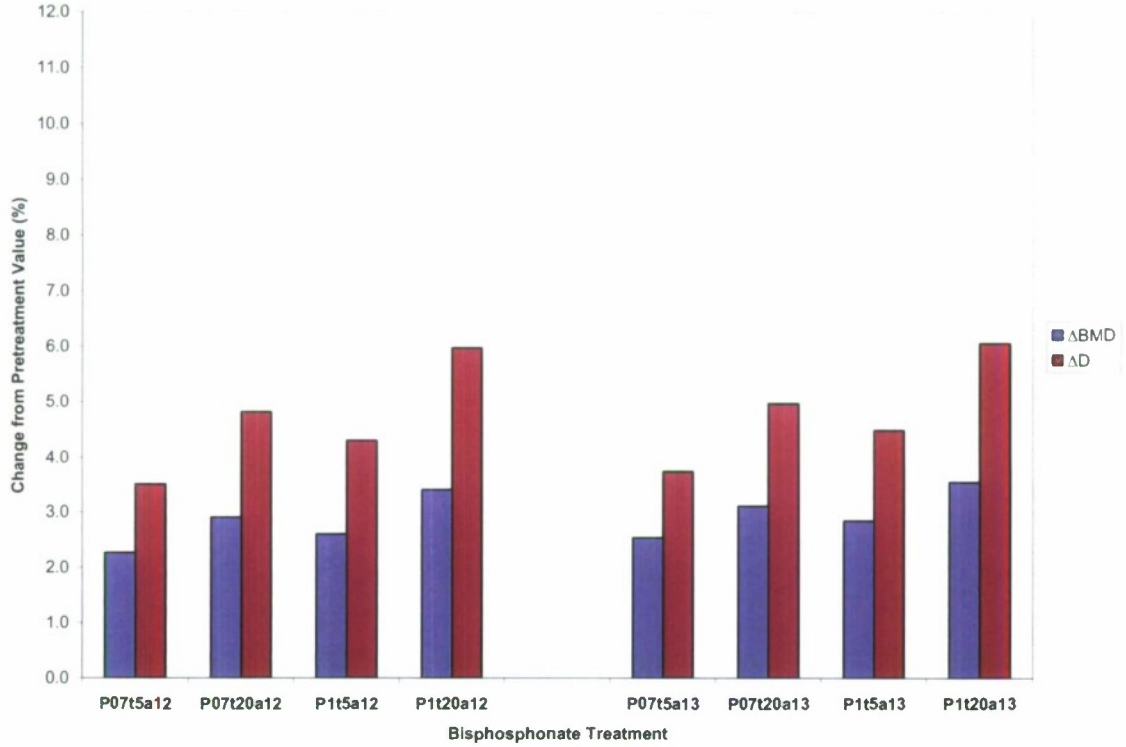


Figure 3.11. Predicted preflight increase in BMD and damage (D) due to 90-day preflight treatment.

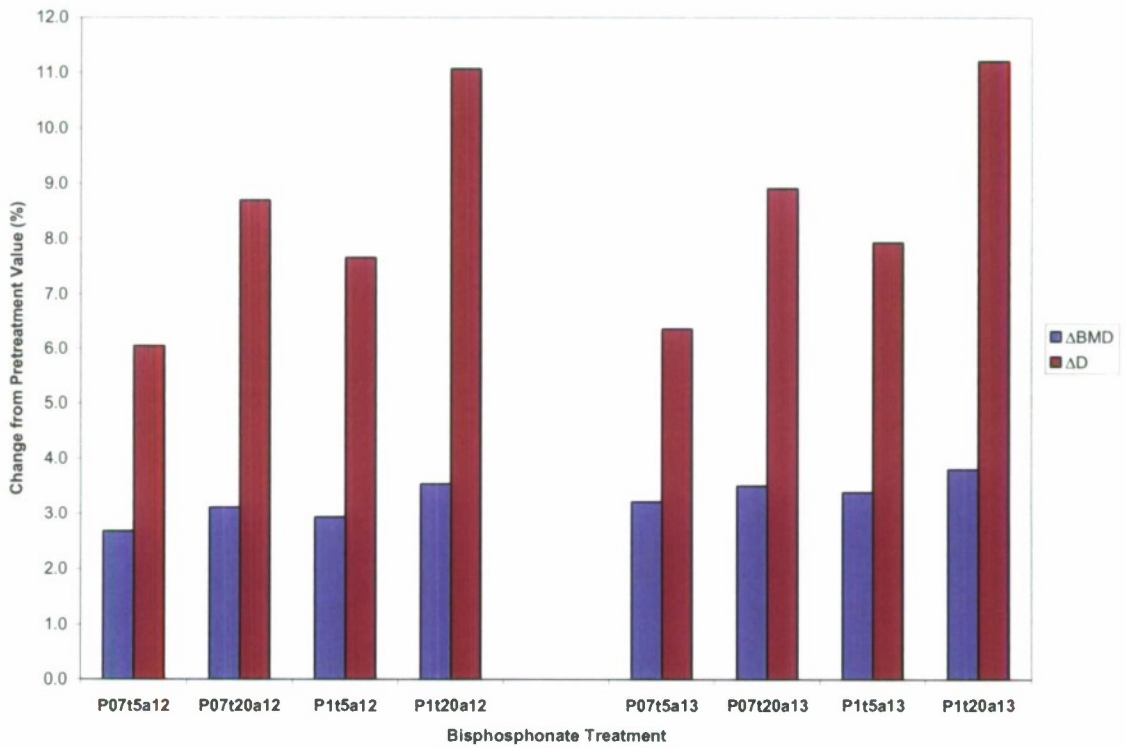


Figure 3.12. Predicted preflight increase in BMD and damage (D) due to 180-day preflight treatment.

Table 3.1. Ratio of percent preflight increase of damage (D) to BMD.

	%ΔD:%ΔBMD for Number of Days of Preflight Treatment				
	7	14	30	90	180
P07t5a12	2.83	2.08	1.37	1.55	2.25
P07t20a12	2.91	2.30	1.48	1.66	2.80
P1t5a12	3.17	2.29	1.39	1.66	2.62
P1t20a12	3.39	2.49	1.44	1.76	3.14
P07t5a13	2.44	1.88	1.32	1.47	1.98
P07t20a13	2.52	2.04	1.42	1.60	2.55
P1t5a13	2.72	2.04	1.34	1.58	2.35
P1t20a13	2.92	2.26	1.40	1.71	2.96

3.4 Varying Onset of Bisphosphonate Treatment for Spaceflight

For longer durations in space, the model predicted pretreatment periods to become less effective at influencing the end-of-flight values for BMD and damage (Figure 3.7, 3.13, and 3.14). As duration in space increased, the model predicted end-of-flight values to be influenced more by the effects of spaceflight than the effects of preflight treatment.

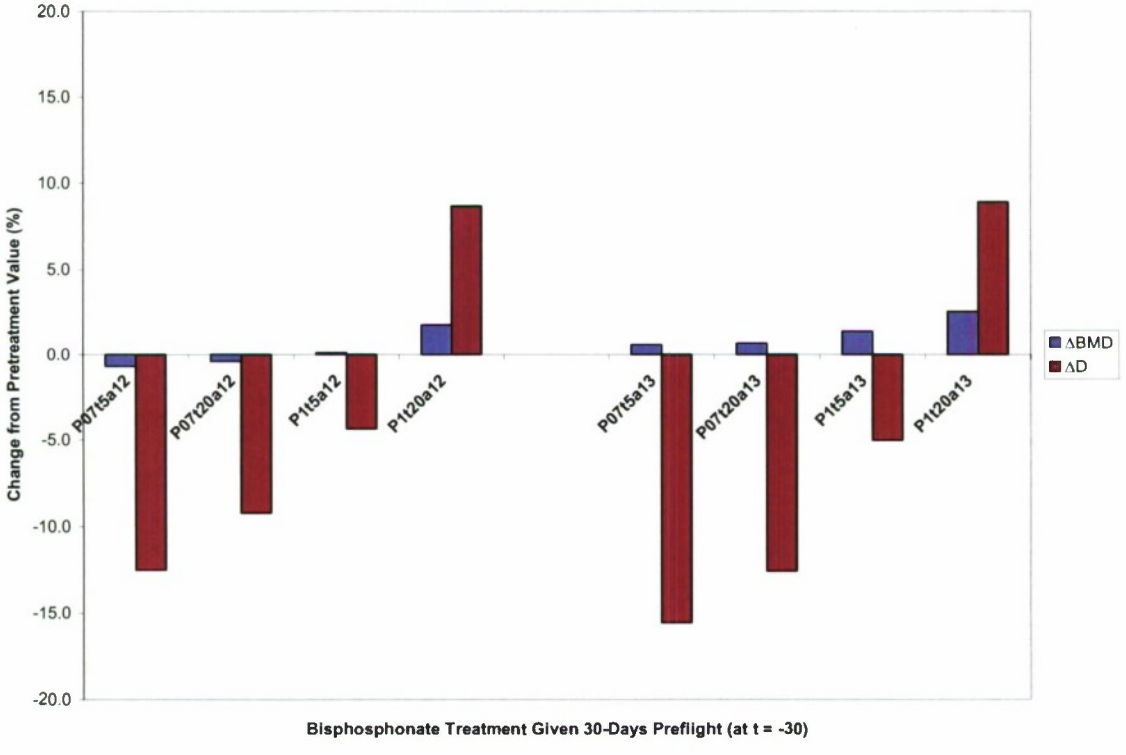


Figure 3.13. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight due to 30-day preflight treatment.

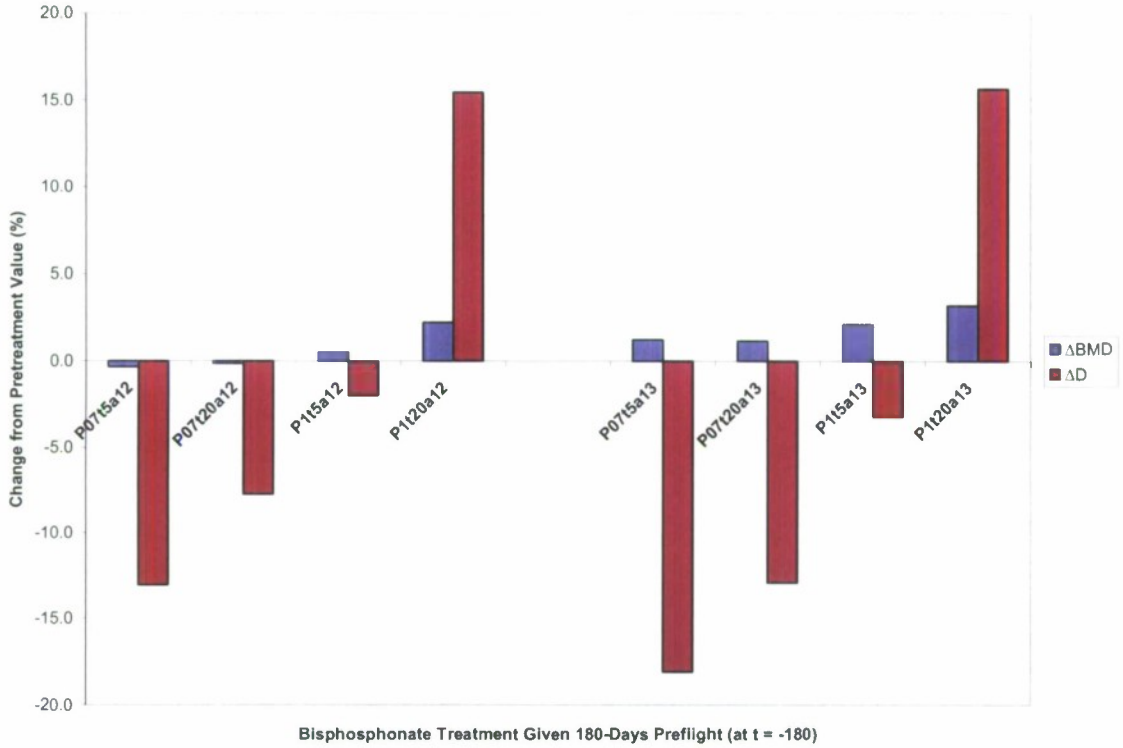


Figure 3.14. Predicted percent changes in BMD and damage (D) at end of 365-day spaceflight due to 180-day preflight treatment.

Alternately, pretreatment phases highly influenced the model's predictions of BMD and damage for shorter durations of spaceflight. For nearly all bisphosphonate potencies simulated, the addition of preflight treatment caused increased BMD and damage accumulation at the end of 10-day spaceflight. For 90-day spaceflight, a pretreatment period of 30 days predicted increases in BMD without significantly increasing damage, and the same occurred with a 90-day pretreatment for 180-day spaceflight. Also, adding preflight treatment for therapies with high suppression predicted further increases to the already large gains in BMD and damage that occurred without pretreatment.

3.5 One-Year Postflight, Posttreatment Recovery

The model predicted treatments with low levels of remodeling suppression given initially at the beginning of spaceflight to result in the highest BMD and lowest

microcrack density accumulation 1-year after returning to Earth (which is also 1-year posttreatment). Treatments with high suppression, though they resulted in higher BMD at the end of flight, were predicted to generate the lowest BMD of all the applied treatments. Predicted 1-year postflight values for BMD and damage for treatments with intermediate levels of remodeling suppression were between the two extremes (Figures 3.15 and 3.16). Note that bisphosphonate treatment begins where the x- and y-axes meet for the line graphs that display timelines of preflight through 1-year postflight predicted values.

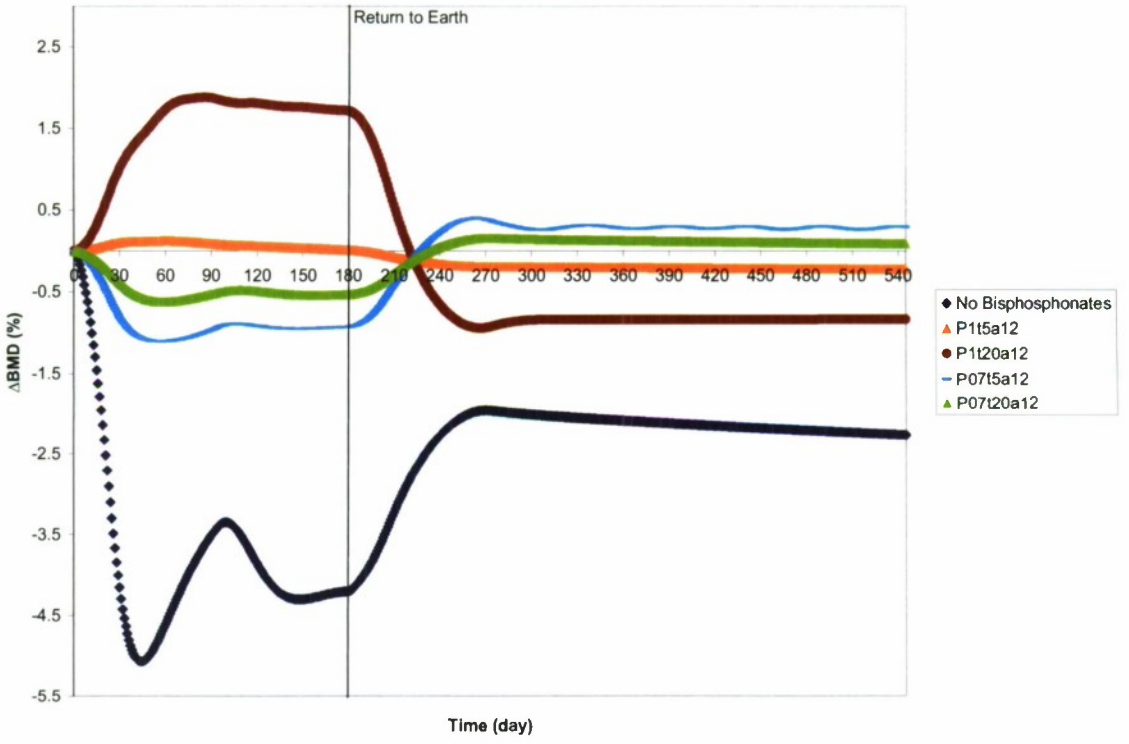


Figure 3.15. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight.

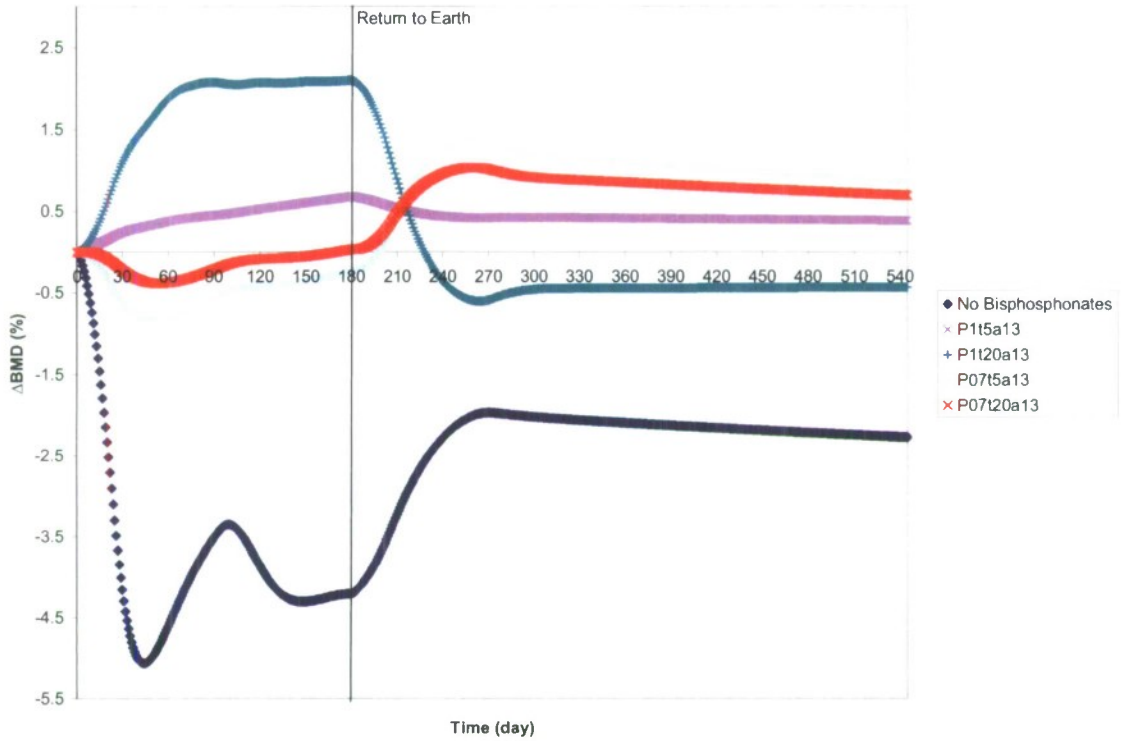


Figure 3.16. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight.

Upon return to Earth, predicted damage accumulation increased for treatments given at the beginning of spaceflight that created bone balances of 1.2 (Figure 3.17). This was the opposite case for bone balances of 1.3, where the model predicted further decreases in damage post-flight and post-treatment (Figure 3.18), except in the case of 365-day spaceflight where predicted damage loss was substantial during flight (Figure 3.19). Prediction results for treatments with high levels of remodeling suppression did not follow these trends, as they always led to decreases in damage during postflight recovery.

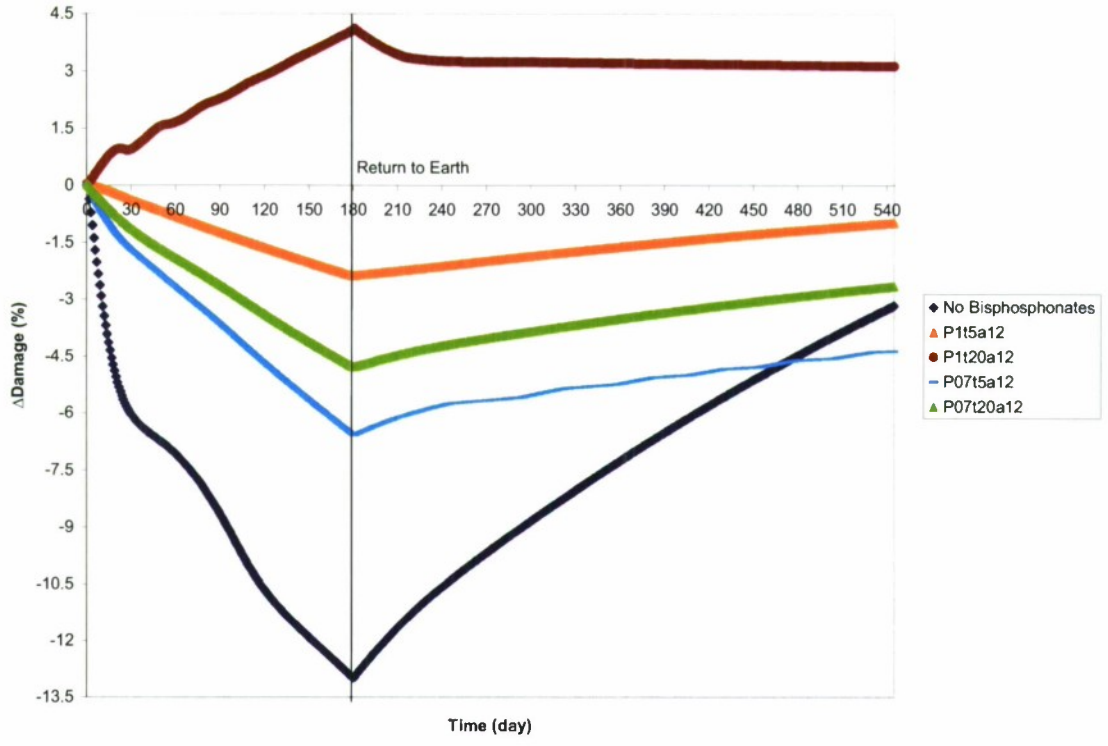


Figure 3.17. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

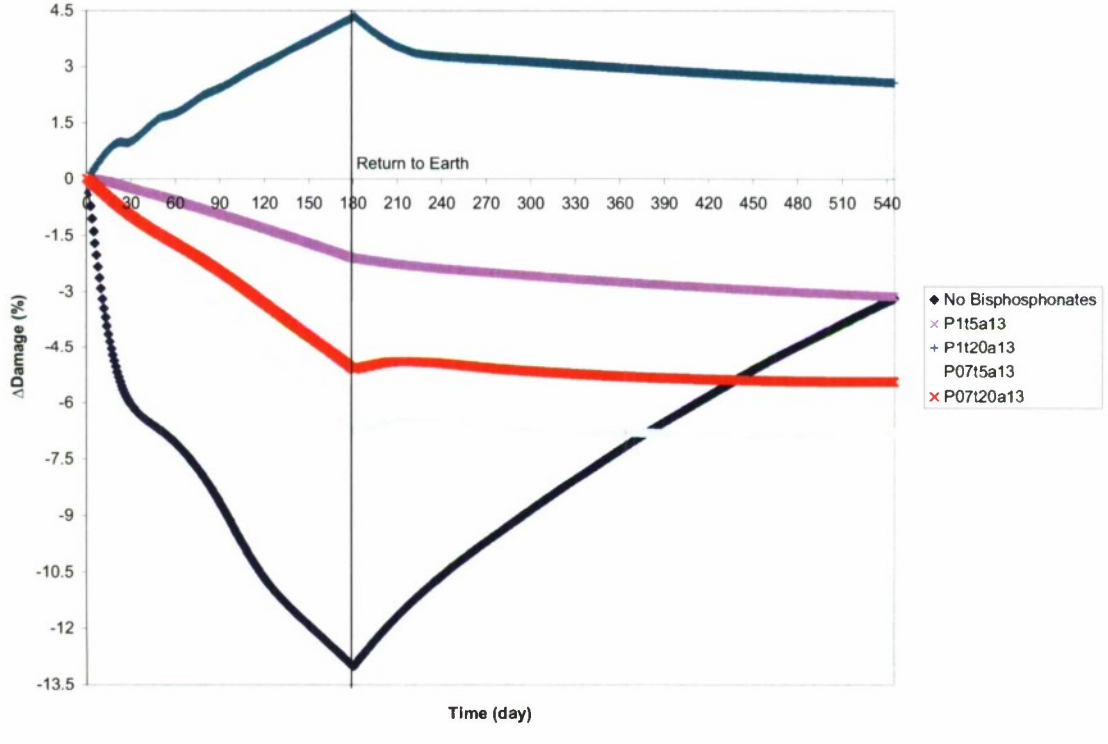


Figure 3.18. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

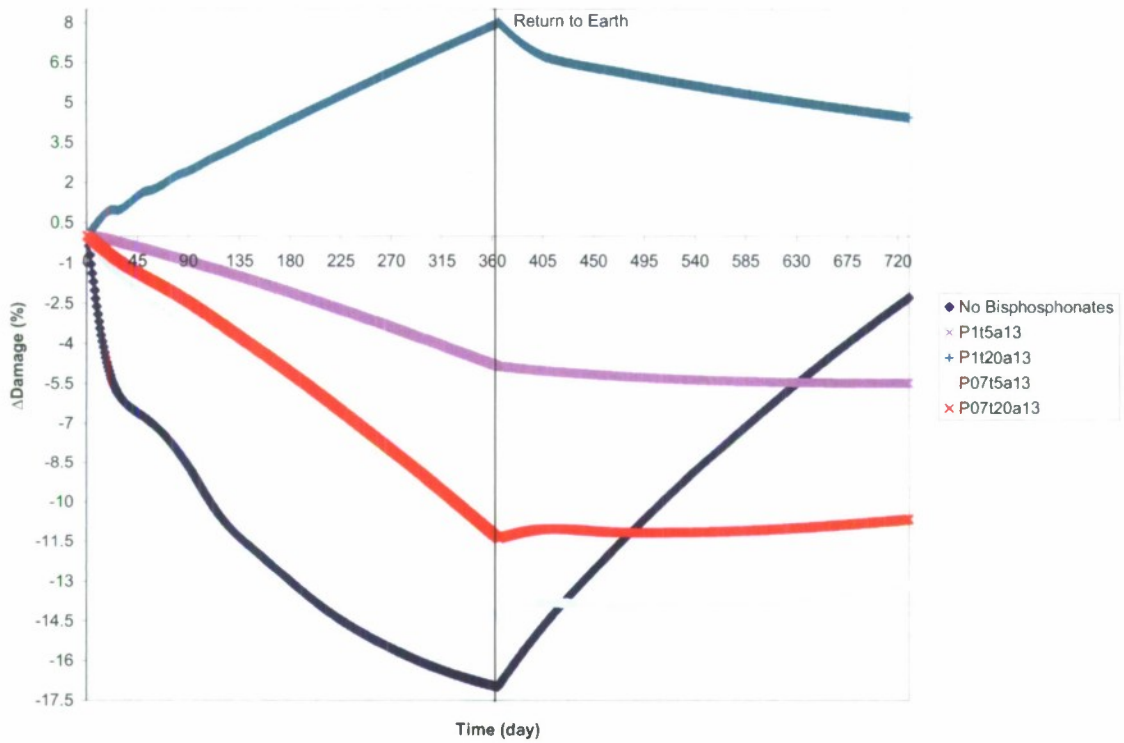


Figure 3.19. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

The predieted gains in BMD due to preflight treatment appear to have inereased the ability to recover from spaeeflight as predicted BMD and damage nearly reaeh pretreatment values (Figures 3.20 and 3.21 for bisphosphonates given 90 days preflight. See appendiees B, C, D, and E for treatment results for 10-, 90-, 180-, and 365-day spaceflights, respectively.). Most of the predicted values for BMD and damage either were near or reaehed equilibrium after 1-year of postflight reeovery. The model predicted 1-year postflight recovery values for BMD and damage to be both above and below pretreatment values, depending on the treatment.

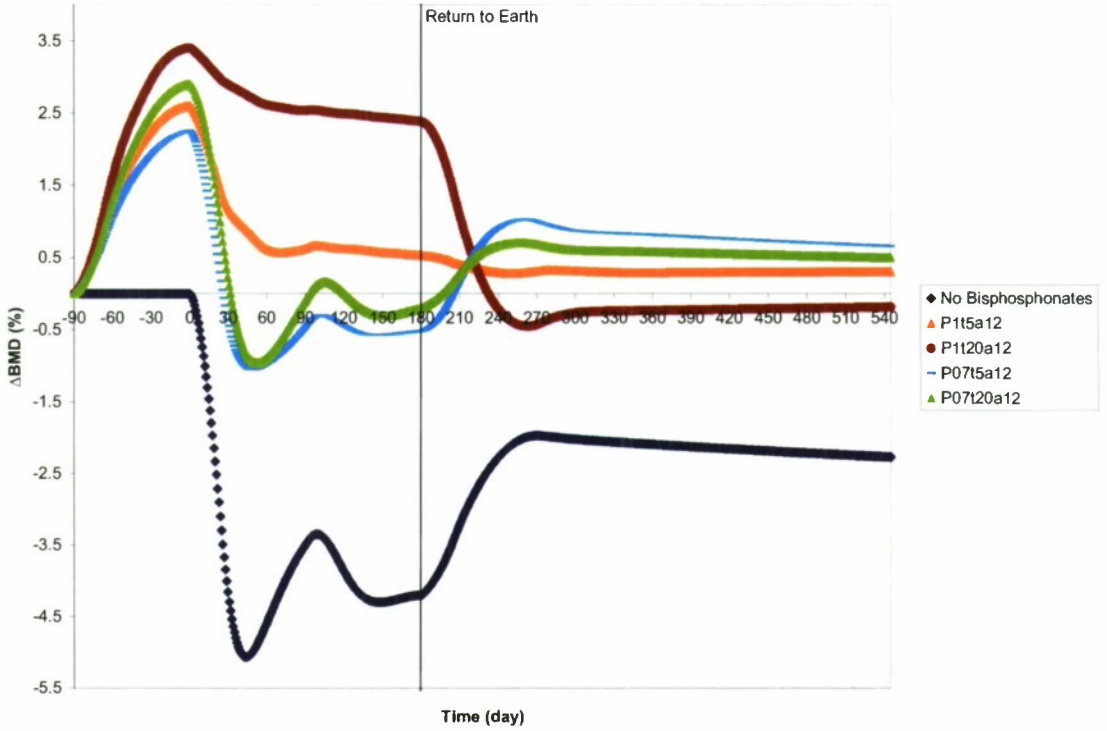


Figure 3.20. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

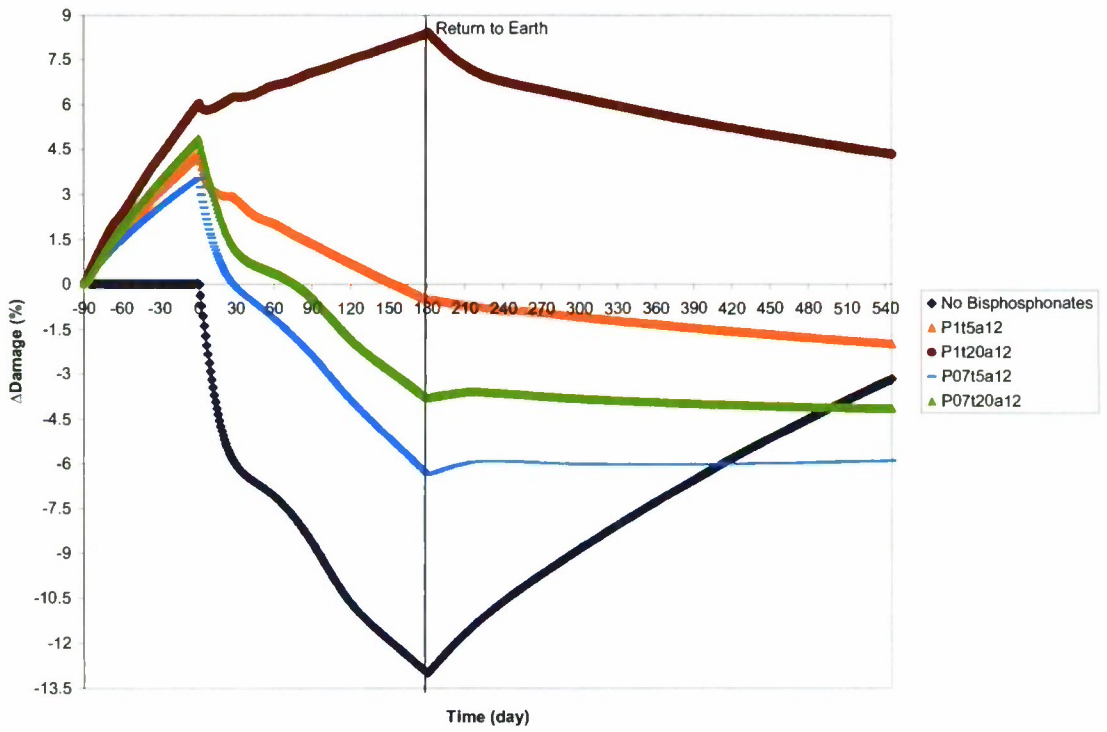


Figure 3.21. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

CHAPTER 4: DISCUSSION

The computer model developed here combines previous bone remodeling and bisphosphonate algorithms plus spaceflight data obtained from experimental studies in literature in order to better understand the adverse effects of microgravity on bone and predict potential treatments for space explorers. The model predicted reduced risk of fracture by increasing bone quantity and either increasing or only slightly reducing bone quality for treatments (1) with low to intermediate suppression of remodeling activation and (2) that create higher bone balance ratios. The simulation also predicted significant changes to BMD and damage upon return to Earth as the remodeling response readjusted to higher stress conditions. For treatments highly suppressing remodeling activation, these predicted postflight changes included decreased BMD and increased damage accumulation. Low levels of remodeling suppression led the model to predict substantial increases in BMD and small increases in damage postflight. Postflight changes were minimal for treatments with intermediate suppression.

The model was developed to match the 4.2 percent loss in BMD over 180 days in space as seen on the International Space Station [3,5,6]. The model's greatest predicted BMD loss in untreated, trabecular vertebral bone was 5.01 percent for a 365-day spaceflight. This is less than half the highest loss (-10.8 percent) seen in Russian cosmonauts on Salyut missions lasting 5 to 7 months [5]. Though the model does not match the results from these older missions, it is likely due to advancements in technology, physical preparedness of subjects, and onboard exercise routines that were developed for missions to the International Space Station.

Full recovery for space explorers returning from the ISS took from 1 to 3 years to complete [5]. The model developed here predicted that full recovery to preflight BMD values may never be attained without treatment. With bisphosphonate treatment, the model predicted complete recovery to occur; some treatments even resulted in higher BMD values than existed preflight. The model suggests that bisphosphonate treatment, combined with exercise, may be the solution that NASA and other space exploration programs desire to combat bone deterioration in space.

The predicted remodeling response of untreated bone to environmental changes was non-linear, as most BMD was lost or gained early on in the transitions from Earth to space and space to Earth. For 10 days in space, the model predicted more mineral to be lost while readjusting to Earth's gravity than lost during spaceflight. This has yet to be examined experimentally, but certainly the model provided insight into a possible trend that may have gone unnoticed. Knowing when fracture risk is maximized may provide for better postflight recovery programs so that fracture can be avoided.

The model predicted bisphosphonate treatment to be beneficial for all durations of spaceflight, not just for longer duration missions. In many instances, preflight treatments were shown to reduce the fracture risk upon return to Earth. Longer simulated flight durations required longer preflight treatments to provide similar effects to those with shorter flights and shorter pretreatments. The problem with this is that as treatment on Earth is lengthened, damage accumulation increases to such an extent that it could actually cause an increase in fracture risk before entering space. Based on the model's predictions of damage increase, pretreatment periods longer than 30 days may put the subjects at risk. During pretreatment, the subjects are still on Earth where higher stresses

cause greater increases in damage as compared to space. They may also be exposed to even higher stresses due to exercises in preparation for the mission. These exercises, combined with brittler bones, could lead to a fracture before flight and put a hold on the mission.

Most interestingly, the model predicted treatments with high suppression of remodeling activation to have the highest gain in BMD at the end of flight, and the lowest BMD values 1-year postflight. These treatments almost completely inhibited remodeling, causing large amounts of damage to accumulate. Though BMD was much higher, the quality of bone was poor. Upon return to earth and discontinuation of treatment, bone remodeling was no longer inhibited and responded to the high amount of damage. Damage and bone were removed at a much greater rate than bone formation occurred, causing quality of bone to increase, but quantity to decrease.

Alternately, the model predicted treatments with low suppression of activation frequency to have the lowest BMD at the end of spaceflight, but they had the highest BMD and lowest amount of damage 1-year postflight. These treatments allowed a fair amount of remodeling to continue in space, but limited it enough so that bone loss was kept to a minimal amount. A majority of the predicted bone loss came from lost damage, so upon return to earth damage accumulation did not activate a remodeling response. These treatments had the largest postflight gains in BMD because the response was mostly due to loading in which bone was added to meet the strength required to support the subject.

Treatments with intermediate suppression and bone balance ratios of 1.3 were optimal for both end-of-flight and 1-year postflight. The model predicted these treatments

to lower risk of fracture both upon return to Earth and after 1 year of recovery. Lower levels of suppression allowed just enough resorption to remove a good amount of fatigue microdamage and increase bone quality, while higher bone balances appropriated more formation than resorption, increasing bone mass.

Since the simulation of bisphosphonate treatment in this model was short in comparison with other models, it is difficult to determine if limits of BMD growth were reached. Predicted BMD gains were non-linear and fluctuated throughout the simulation; in contrast, previous models that did not account for damage and disuse stimuli showed only permanent, linear gain in mass [32,33,34].

Similar to Lacy's model of trabecular bone turnover [34], activation frequency was found to have the greatest effect on bone mass. Unloading in microgravity caused a disuse response in untreated bone, increasing BMU activation and leading to the resorption of large amounts of bone. The model by Hernandez et al. [33] also showed similar results in that the initial gain in bone mass was dependent upon the level of remodeling suppression and bone balance ratio. Heaney's model [32] exhibited an initial gain dependent upon the pretreatment remodeling parameters; however this model did not examine variations in these parameters before treatment.

It is difficult to compare the accumulation of damage of this model to others since it also simulates microgravity. The model by Nyman et al. [23] predicted small gains in microdamage initially that were proportional to activation frequency suppression. In contrast, the model developed here predicted losses in microdamage for treatments with low and intermediate levels of suppression and gains in damage accumulation when activation frequency was highly suppressed.

The limitations of this model occur where assumptions have been made due to the lack of available information. First, the damage rate coefficient, k_D, was kept constant throughout the simulation, though it is likely to change in space or for various bone balances. Other coefficients, too, such as the damage rate exponent or the activation frequency coefficients, are likely to be altered in space and would benefit from further study of bone remodeling in microgravity. Second, the predicted postflight results are limited by the fact that they are based on the same stress applied preflight even though postflight recovery programs enable space explorers to ease back into full loading. This high postflight stress would cause overpredictions of both BMD and damage. Also, a bone balance ratio of 1.0 was instantly applied upon return to Earth, when it is more likely that the ratio would slowly ease back down. Third, the simulation applies a constant stress derived from bone loss to a section of bone rather than deriving the actual strain and applying it to a finite element model. Using a finite element model would create a more accurate remodeling response with more precise loading conditions and allow detailed analysis of the effects of specific exercises on maintaining bone mass. Lastly, the model does not take into account effects of spaceflight on blood flow, drug metabolism, tissue binding, drug elimination, fluid shear stress, or changes in hormone levels [2,4]. Many of these affect the efficacy of the drug itself. Changes to fluid shear stress in a microgravity environment could affect the mechanosensory ability of osteoclasts to sense signals indicating bone loading and would lead to further loss of bone even under heavy exercise [4]. Also, although the model tracks changes to the populations of osteoblasts and osteoclasts, it only accounts for changes due to the

remodeling response and not due to the physiological adaptations that may occur in microgravity [2].

Changes in hormone levels or physiological alterations to the populations of osteoblasts or osteoclasts could significantly alter the remodeling response in space and the response to bisphosphonate treatment [4]. With reduced levels of PTH, IGF-1, and growth hormone [4], it would be likely for remodeling formation and resorption rates to be altered in microgravity. With so little bone formation occurring, it could also be possible that active osteoclasts may significantly outnumber osteoblasts and lead to a slower recovery response upon return to earth.

It is clear that the model would significantly benefit from further studies on spaceflight. Though the model has to overcome the many unknown variables of bone remodeling, bisphosphonates, and microgravity, it has shown the ability to provide potential trends for future studies. As new data and information becomes available, the model's accuracy can only be improved and could eventually be a tool used for predicting effects of other treatments as well.

SUMMARY OF CONCLUSIONS

- The model predicted bisphosphonates reduced fracture risk by increasing bone quantity and either increasing or only slightly reducing bone quality for treatments:
    - (1) with low to intermediate suppression of remodeling activation
    - (2) that create higher bone balance ratios
- Most changes to BMD occurred early on when adjusting to new environments
- Predicted BMD loss was fairly consistent with data from missions to the International Space Station
- Preflight treatments were shown to reduce the risk of fracture for all durations of spaceflight
- Longer preflight treatment periods may put the subject at risk due to increased microdamage accumulation
- Overall, the model suggested that bisphosphonate treatment, combined with existing exercise programs, may be the solution to combat bone deterioration in space

REFERENCES

[1] Payne MWC, Williams DR, Trudel G. Space flight rehabilitation. Am J of Phys Med Rehabil 2007;86:583-91

[2] Shapiro, JR. Microgravity and drug effects on bone. J. Musculoskelet Neuronal Interact 2006;6:322-23.

[3] Lang T, LeBlanc A, Evans H, Lu Y, Genant H, Yu A. Cortical and trabecular bone mineral loss from the spine and hip in Long-Duration Spaceflight. J Bone Miner Res 2004;19:1006-12.

[4] Bloomfield, SA. Summary - bone in microgravity environments: "Houston, we have a problem." J Musculoskelet Neuronal Interact 2006;6:329-30.

[5] LeBlanc AD, Spector ER, Evans HJ, Sibonga JD. Skeletal responses to space flight and the bed rest analog: a review. J Musculoskelet Neuronal Interact 2007;7:33-47.

[6] Iwamoto J, Takeda T, Sato Y. Interventions to prevent bone loss in astronauts during space flight. Keio J Med 2005;54:55-9.

[7] Martin RB, Burr DB, Sharkey NA. Skeletal Tissue Mechanics. New York: Springer; 1998.

[8] Gunaratne G. A theoretical analysis of vibrational modes aimed at their use as measures of bone damage. ISSO Annual Report 2005:105-7,126.

[9] Epstein S. The roles of bone mineral density, bone turnover, and other properties in reducing fracture risk during antiresorptive therapy. Mayo Clin Proc 2005;80:379-88.

[10] Parfitt AM. Bone age, mineral density, and fatigue damage. Calcified Tissue International 1993;53(Supplement 1):S82-6.

[11] Osteoblasts and Osteoclasts. 2004. Children's Hospital Boston. 1 June 2008. <http://www.childrenshospital.org/cfapps/research/data_admin/Site31/mainpageS31P1sublevel2.html>

[12] Bentolila V, Boyce TM, Fyhrie DP, Drumb R, Skerry TM, Schaffler, MB. Intracortical remodeling in adult rat long bones after fatigue loading. Bone 1998;23:275-81.

[13] Burr DB, Martin RB, Schaffler MB, Radin EL. Bone remodeling in response to in vivo fatigue microdamage. J Biomech 1985;18:189-200.

[14] Burr DB, Martin RB. Calculating the probability that microcracks initiate resorption spaces. J Biomech 1993;26:613-6.

[15] Mori S, Burr DB. Inereased intraeortieal remodeling following fatigue damage. Bone 1993;14:103-9.

[16] Li XJ, Jee WSS, Chow SY, Woodbury DM. Adaptation of eaneellous bone to aging immobilization in the rat: a single photon absorptiometry and histomophometry study. The Anatomieal Reeord 1990;227:12-24.

[17] Chaffler MB, LI XJ. Immobilization indueed bone loss: quantitive histologieal studies of eortieal bone resorption. Transaetions of the 36th Annual Meeting of the Orthopaedie Researeh Soeiety 1990;15:187.

[18] Hazelwood SJ, Martin RB, Rashid MM, Rodrigo, JJ. A meehanistie model for internal bone remodeling exhibits different dynamie responses in disuse and overload. J Biomeeh 2001;34:299-308.

[19] Cullen DM, Smith RT, Akhter MP. Bone-loading response varies with strain magnitude and eycle number. J Appl Physiol 2001;91:1971-6.

[20] Lin JH. Bisphosphonates: a review of their pharmaeokinetie properties. Bone 1996;19:80-100.

[21] Rogers MJ, Watts DJ, Russell RG. Overview of bisphosphonates. Caneer 1997;80:1652-60.

[22] Sato M, Grasser W, Endo N, Akins R, Simons H, Thompson DD, et al. Bisphosphonate aetion. Alendronate loealization in rat bone and effeets on osteoelast ultrastrueture. J Clin Invest 1991;88:2095-105.

[23] Nyman JS, Yeh OC, Hazelwood SJ, Martin RB. A theoretieal analysis of long-term bisphosphonate effeets on trabeeular bone volume and mierodamage. Bone 2004;35:296-305.

[24] Fleish H. Bisphosphonates: meehanisms of aetion. Endoer Rev 1998;19:80-100.

[25] Kimmel DB. Meehanism of aetion, pharmaeokinetie and pharmacodynamic profile, and elinieal applieations of nitrogen-eontaining bisphosphonates. J Dent Res 2007;86:1022-33.

[26] Rodan GA, Fleiseh HA. Bisphosphonates: meehanisms of aetion. J Clin Invest 1996;97:2692-6.

[27] Lieberman Ua, Weiss Sr, Broll J, Minne HW, Quan H, Bell NH, Rodrigues-Portales J, Downs RD Jr, Dequeker J, Favus M. Effeet of oral alendronate on bone mineral density and the ineidenee of fraetures in postmenopausal osteoporosis. The Alendronate Phase III Osteoporosis Treatment Study Group. N Engl J Med 1995;333(22):1437-43.

[28] Tonino RP, Meunier PJ, Emkey R, Rodriguez-Portales JA, Menkes CJ, Wasnich RD, Bone HG, Santora AC, WU m, Desai R, Ross PD. Skeletal benefits of alendronate: 7-year treatment of postmenopausal osteoporotic women. Phase III Osteoporosis Treatment Study Group. J Clin Endocrinol Metab 2000;85:109-15.

[29] Mashiba T, Turner CH, Hirano T, Forwood MR, Johnston CC, Burr DB. Effects of suppressed bone turnover by bisphosphonates on microdamage accumulation and biomechanical properties in clinically relevant skeletal sites in beagles. Bone 2001;28:524-31.

[30] Carter DR, Fyhrie DP, Whalen RT. Trabecular bone density and loading history: regulation of connective tissue biology by mechanical energy. J Biomech 1987;20:785-94.

[31] Huiskes R, Weinans H, Grootenboer HJ, Dalstra M, Fudala B, Slooff TJ. Adaptive bone-remodeling theory applied to prosthetic-design analysis. J Biomech 1987;20:1135-50.

[32] Heaney RP, Yates AJ, Santora II AC. Bisphosphonate effects and the bone remodeling transient. J Bone Miner Res 1997;12:6-15.

[33] Hernandez CJ, Beaupré GS, Marcus R, Carter DR. A theoretical analysis of the contributions of remodeling space, mineralization, and bone balance to changes in bone mineral density during alendronate treatment. Bone 2001;29:511-6.

[34] Lacy ME, Bevan JA, Boyce RW, Geddes AD. Antiresorptive drugs and trabecular bone turnover: validation and testing of a computer model. Calcif Tissue Int 1994;54:179-85.

[35] Currey JD. The effect of porosity and mineral content on the Young's modulus of elasticity of compact bone. J Biomech 1988;21:131-9.

[36] Rho JY, Ashman RB, Turner CH. Young's modulus of trabecular and cortical bone material: ultrasonic and microtensile measurements. J Biomech 1993;26:111-9.

[37] Turner CH, Anne V, Pidaparti RMV. A uniform strain criterion for trabecular bone adaptation: do continuum-level strain gradients drive adaptation? J Biomech 1997;30:555-63.

[38] Hart RT, Davy DT. Theories of bone modeling and remodeling. In: Cowin SC, editor. Bone Mechanics. Boca Raton, FL: CRC Press; 1989, pp.253-77.

[39] Martin RB. The usefulness of mathematical models for bone remodeling. Yearbook of Physical Anthropology 1985;28:227-36.

[40] Parfitt AM. The physiologic and clinical significance of bone histomorphometric data. In: Recker RR, editor. Bone Histomorphemtry: Techniques and Interpretation. Boca Raton, FL: CRC Press; 1983, pp.143-223.

[41] Frost HM. On rho, a marrow mediator, and estrogen: their roles in bone strength and "mass" in human females, osteopenias, and osteoporoses – insights from a new paradigm. J Bone Miner Metab 1998;16:113-23.

[42] Martin RB. Porosity and specific surface of bone. Critical Reviews in Biomedical Engineering 1984;10:179-222.

[43] Frost HM. Tetracycline-based histological analysis of bone remodeling. Calcif Tissue Res 1969;3:211-37.

[44] Martin RB. Mathematical model for repair of fatigue damage and stress fracture in osteonal bone. J Orthop Res 1995;13:309-16.

[45] Chappard D, Minaire P, Privat C, Berard E, Mendoza-Sarmiento J, Tournebise H, Basle MF, Audran M, Rebel A, Picot C, Gaud C. Effects of tiludronate on bone loss in paraplegic patients. J Bone Miner Res 1995;10:112-8.

[46] Rapillard L, Charlebois M, Zysset K. Compressive fatigue behavior of human vertebral trabecular bone. J Biomech 2006;39:2133-9.

[47] Whalen RT, Carter DR, Steele CR. Influence of physical activity on the regulation of bone density. J Biomech 1988;21:825-37.

[48] Schaffler MB, Choi K, Milgrom C. Aging and matrix microdamage accumulation in human compact bone. Bone 1995;17:521-25.

[49] Brockstedt H, Kassem M, Eriksen EF, Mosekilde L, Melsen F. Age- and sex-related changes in iliac cortical bone mass and remodeling. Bone 1993;14:681-91.

[50] Hernandez CJ, Keaveny TM. A biomechanical perspective on bone quality. Bone 2006;39:1173-81.

[51] Morgan EF, Yeh OC, Keaveny TM. Damage in trabecular bone at small strains. Eur J Morphol 2005;42:13-21.

[52] Ammann P, Rizzoli R. Bone strength and its determinants. Osteoporos Int 2003;14:S13-8.

APPENDIX A: MATLAB CODE

```
%Bone Mass Preservation and Fracture Risk Assessment
%with Bisphosphonate Therapy During Spaceflight
%Simulation of Bisphosphonates acting on Bone Remodeling in Space
%Chris Gardina 6-10-08
%
clear
clc

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Remodeling parameters
Rc = 0.095;                             % Cement line radius
Rh = 0.020;                             % Haversian canal radius
Tr = 25;                                % Resorption period
Tv = 5;                                 % Reversal period
Tf = 64;                                % Formation period
trab1 = 0.2;                            % Porosity partition for cortical
to trabecular bone: Change resorption area & Disuse
trab2 = 0.097267787;                    % Porosity partition for cortical
to trabecular bone: Change in stiffness constants
phi0 = 0.0000000001875;                 % Equilibrium stimulus
%por0 = 0.2;                             % Initial trabecular porosity
por0 = 0.04432132964;                   % Initial cortical porosity
sa0 = ((((28.8*por0-101)*por0+134)*por0-93.9)*por0+32.1)*por0;  %
Initial surface area
phc=0.04432132964;                      % Adjusts resorption rate to match
apposition rate for cortical bone (i.e.,
                                        % assumes resorption process
includes void spaces)

% Activation frequency conditions
% Describes Ac.f versus disuse and Ac.f versus damage curves
Acfmax1 = 0.5;                          % Maximum Ac.f due to damage
Acfmax2 = 0.5;                          % Maximum Ac.f due to disuse
samax = 4.1905;                         % Normalizes specific area
Acfdis0 = 0.0;                          % Equilibrium Ac.f for disuse
Acfdam0 = 0.0224693284;                 % Equilibrium Ac.f for damage
Acf0 = (Acfdam0 + Acfdis0)*sa0/samax;   % Equilibrium Ac.f
kb = 65000000000;
kc = phi0/2;

% Damage conditions
fs = 5;                                 % Damage repair factor
d0 = 0.03662944;                        % Equilibrium damage
if por0 <= trab1
   kd = d0*Acf0*(pi*Rc^2)*fs/phi0;      % Equilibrium damage constant
else
   kd = 0.5*d0*Acf0*(pi*Rc^2)*fs/phi0;
end

kr = -1.6;

% Mechanical conditions
Area = 100;                             % Cross-sectional area of bone
```

```
rll = 3000;                             % Frequency of loading in no. of
cycles per day
q = 4;                                  % exponent of mechanical stimulus
change = 0;                             % percent change in force
chgper = 1;                             % time period of force change
days = 7300;                            % Time length of initial value
Setup
flight = 180;                           % Time length of spaceflight
postflight = 7300;                       % Time length of return to earth
dt = 1;                                 % Time step; One day

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
%Loop through values of tau, Pmax, and bone balance
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
for x=1:3
   if (x == 1)
        Pmax=0;
   end
   if (x == 2)
        Pmax=0.7;
   end
   if (x == 3)
        Pmax = 1;
   end
   for y=1:2
        if (y == 1)
            tau=5;
        end
        if (y == 2)
            tau=20;
        end
        for z=1:2
            if (z == 1)
                area_r_ratio=5./6;              %bone balance 1.2
            end
            if (z == 2)
                area_r_ratio=10./13;            %bone balance 1.3
            end
            if Pmax == 0
                area_r_ratio=1;                 %no drug therapy,
bone balance 1
            end
            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
            % Prior conditions before start of simulation

            for t=1:(Tr+Tv+Tf)
                por(t) = por0;                                  % porosity
                if por(t) <= trab2
                    modulus(t) = 23440*(1-por(t))^5.74;     %Stiffness
of bone if cortical
                else
                    modulus(t) = 14927*(1-por(t))^1.33;     %Stiffness
of bone if trabecular
```

```
                end
                Phi(t) = phi0;                                  %
mechanical stimulus
                strain(t) = (phi0/rl1)^0.25;                    % principal
strain
                stress(t) = modulus(t)*strain(t);               % principal
stress
                Force(t) = stress(t)*Area;                      % force on
bone
                SA(t) = sa0/samax;                              %
normalizes Ac.f by available surface area
                Df(t) = kd*phi0;                                % damage
formation rate
                Dr(t) = d0*Acf0*(pi*Rc^2)*fs;                   % damage
removal rate
                if por(t) > trab1
                    Dr(t) = 0.5*Dr(t);
                end
                D(t) = d0;                                      % damage
                NfBMU(t) = Acf0*Tf;                             % No. of
refillng BMUs
                NrBMU(t) = Acf0*Tr;                             % No. of
resorbing BMUs
                if por(t) <= trab1
                    Qf(t) = pi*(Rc^2-Rh^2)/Tf;                  % mineral
apposition rate
                    Qr(t) = pi*Rc^2/Tr;                         %
resorption rate
                    QrNr(t) = (1-phc)*Qr(t)*NrBMU(t);           % amount
bone removed
                    QfNf(t) = Qf(t)*NfBMU(t);                   % amount of
bone added
                else
                    Qf(t) = 0.5*pi*Rc^2/Tf;
                    Qr(t) = 0.5*pi*Rc^2/Tr;
                    QrNr(t) = Qr(t)*NrBMU(t);
                    QfNf(t) = Qf(t)*NfBMU(t);
                end
                Qnet(t) = QfNf(t)-QrNr(t);                      %
difference between mineral added and mineral removed
                Acfdam(t) = Acfdam0;                            % Ac.f due
to damage
                Acfdis(t) = Acfdis0;                            % Ac.f due
to disuse
                Acf(t) = (Acfdam(t) + Acfdis(t))*SA(t);         % Total
Ac.f
                TIME(t) = t;
            end
            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
            %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%
            %%%%%%%Adding in Constants for Bisphonate
Treatment%%%%%%%%%%%%%%%%

          % Pmax=0.7;
           % Pmax=1;
           % tau=5;
```

```
            % tau=20;
             %area_r_ratio=1;                          %bone balance 1
               % area_r_ratio=5./6;                     %bone balance 1.2
               % area_r_ratio=10./13;                   %bone balance 1.3
             preflight_treat = 0;                     %Days of Bisphosphonate
therapy before spaceflight
             %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
             %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%
             % Simulation
             for t=(Tr+Tv+Tf+1):(days+flight+postflight)
                 TIME(t) = t;

                 % Mechanical conditions upduate
                 if por(t-1) <= trab2
                     modulus(t) = 23440*(1-por(t-1))^5.74;    %Stiffness
of bone if cortical
                 else
                     modulus(t) = 14927*(1-por(t-1))^1.33;    %Stiffness
of bone if trabecular
                 end
                 if (t > (days)) && (t < (days + flight + 1))
                     Force(t) = 89.09;                    %Force in
spaceflight to match 0.7%BMD loss per month
                         %Force(t) = 85.47;                        %Force
in spaceflight to match 1.0%BMD loss per month
                 else
                     Force(t) = 100;                          %Force
needed to determine initial values
                 end
                 stress(t) = Force(t)/Area;                   %Stress on
bone
                 strain(t) = stress(t)/modulus(t);            %Strain on
bone

                 % Porosity update
                     Phi(t) = (abs(strain(t))^q)*rl1;         %Calculate
current stimulus

                     if por(t-1) <= 0.2
                         area_f = pi*(Rc^2-
Rh^2);                          %Formation area of cortical bone (Cement
line to Haversian canal)
                         if (t > (days - preflight_treat)) && (t < (days
+ flight + 1))       %Bisphosphate treatment period

area_r=area_r_ratio*pi*Rc^2;                %Resorption area if
cortical & on Bisphosphonates
                         else
                             area_r =
pi*Rc^2;                                %Resorption area if cortical
                         end
                     else
```

```
                    area_f =
0.5*pi*Rc^2;                            % Formation area of trabecular
bone (half of osteon area w/o pore)
                    if (t > (days - preflight_treat)) && (t < (days
+ flight + 1))       %Bisphosphate treatment period
                        area_r =
area_r_ratio*0.5*pi*Rc^2;          %Resorption area if trabecular & on
Bisphosphonates
                    else
                        area_r =
0.5*pi*Rc^2;                            %Resorption area if trabecular
                    end
                end
                %  Calculate change in damage level and update
                Df(t) = kd*Phi(t);                              %Damage
formation rate
                Dr(t) = D(t-1)*Acf(t-1)*area_r*fs;              %Damage
removal rate
                D(t) = D(t-1) + (Df(t) - Dr(t))*dt;             %Damage
update

                %  Calculate demand for new BMUs (Ac.f)
                Acfdam(t) = (Acfdam0*Acfmax1)/(Acfdam0+(Acfmax1-
Acfdam0)*exp(kr*Acfmax1*(D(t-1)-d0)/d0));  %Damage stimulus
                if D(t) <= d0
                    Acfdam(t) = Acfdam0*D(t-1)/d0;
                end
                sa = ((((28.8*por(t-1)-101)*por(t-1)+134)*por(t-1)-
93.9)*por(t-1)+32.1)*por(t-1);          %Surface area
                Acfdis(t) = 0;  % No demand for additional
remodeling if not in disuse
                if Phi(t) < phi0
                    Acfdis(t) = Acfmax2/(1+exp(kb*(Phi(t)-
kc)));    % Demand for remodeling when bone is in disuse
                end
                SA(t)=sa/samax;
                Acf(t) = (Acfdam(t) +
Acfdis(t))*SA(t);         %Update Ac.f

                % Calculate daily amount of bone removed per
resorbing BMU
                % Include less refilling on trabecular surfaces in
disuse
                ac = area_r;                            % Resorption
area
                ab = area_f;                            % Formation area

                Qf(t) = ab/Tf;
                Qr(t) = ac/Tr;

                if por(t-1) > trab1
                    if Phi(t) < phi0
                        Qf(t) = (0.5 + 0.5*Phi(t)/phi0)*Qf(t);  %
For trabecular bone, formation rate decreases in disuse
                    end
                end
```

```
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    %%%%%%%%%%% bisphosphonate potency effect
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    if (t > (days - preflight_treat)) && (t <
(days+flight+1))          %Bisphosphate treatment period
                        P=Pmax*(1-exp(-1*tau*NrBMU(t-1)));
                            %   P=1; %completely suppressed
                        Acf(t) = (1-P)*(Acfdam(t) +
Acfdis(t))*SA(t);    %Bisphosphonate suppressed activation frequency
                    end
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    % Calculate number of refilling BMUs for current
day
                    % Calculate number of resorbing BMUs for current
day
                     NfBMU(t) = NfBMU(t-1) + (Acf(t-Tr-Tv) - Acf(t-Tr-
Tv-Tf))*dt;
                     NrBMU(t) = NrBMU(t-1) + (Acf(t) - Acf(t-Tr))*dt;

                    % Calculate net amount of bone added per day
                    QfNf(t) = NfBMU(t-1)*Qf(t-1) + (Acf(t-Tr-Tv)*Qf(t-
Tr-Tv) - Acf(t-Tr-Tv-Tf)*Qf(t-Tr-Tv-Tf))*dt;
                    QrNr(t) = NrBMU(t-1)*Qr(t-1) + (Acf(t)*Qr(t) -
Acf(t-Tr)*Qr(t-Tr))*dt;
                    Qnet(t) = QfNf(t)-QrNr(t);
                    if por(t-1) <= 0.20
                        Qnet(t) = QfNf(t) - (1-phc)*QrNr(t);
                    end

                % Calculate change in porosity
                   por(t) = (por(t-1) - Qnet(t))*dt;
                   BMD(t) = 2*(1-por(t));

                   uplim = 0.99;
                   if por(t) >= uplim;
                      por(t) = uplim;
                   end
            end
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    %Format data in charts
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    if (Pmax == 0)&&(tau == 5)&&(area_r_ratio==1)
                        for t=1:(days+flight+postflight)
                            DataP0a1(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP0a1.txt DataP0a1 -ascii -double
                    end
                    if (Pmax == 0.7)&&(tau == 5)&&(area_r_ratio==1)
                        for t=1:(days+flight+postflight)
                            DataP07t5a1(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t5a1.txt DataP07t5a1 -ascii -double
                    end
```

```
                    if (Pmax == 0.7)&&(tau == 5)&&(area_r_ratio==5./6)
                        for t=1:(days+flight+postflight)
                            DataP07t5a12(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t5a12.txt DataP07t5a12 -ascii -
double
                    end
                    if (Pmax == 0.7)&&(tau == 5)&&(area_r_ratio==10./13)
                        for t=1:(days+flight+postflight)
                            DataP07t5a13(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t5a13.txt DataP07t5a13 -ascii -
double
                    end
                    if (Pmax == 0.7)&&(tau == 20)&&(area_r_ratio==1)
                        for t=1:(days+flight+postflight)
                            DataP07t20a1(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t20a1.txt DataP07t20a1 -ascii -
double
                    end
                    if (Pmax == 0.7)&&(tau == 20)&&(area_r_ratio==5./6)
                        for t=1:(days+flight+postflight)
                            DataP07t20a12(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t20a12.txt DataP07t20a12 -ascii -
double
                    end
                    if (Pmax == 0.7)&&(tau ==
20)&&(area_r_ratio==10./13)
                        for t=1:(days+flight+postflight)
                            DataP07t20a13(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP07t20a13.txt DataP07t20a13 -ascii -
double
                    end
                    if (Pmax == 1)&&(tau == 5)&&(area_r_ratio==1)
                        for t=1:(days+flight+postflight)
                            DataP1t5a1(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t5a1.txt DataP1t5a1 -ascii -double
                    end
                    if (Pmax == 1)&&(tau == 5)&&(area_r_ratio==5./6)
                        for t=1:(days+flight+postflight)
                            DataP1t5a12(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t5a12.txt DataP1t5a12 -ascii -double
                    end
                    if (Pmax == 1)&&(tau == 5)&&(area_r_ratio==10./13)
                        for t=1:(days+flight+postflight)
```

```
                            DataP1t5a13(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t5a13.txt DataP1t5a13 -ascii -double
                    end
                    if (Pmax == 1)&&(tau == 20)&&(area_r_ratio==1)
                        for t=1:(days+flight+postflight)
                            DataP1t20a1(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t20a1.txt DataP1t20a1 -ascii -double
                    end
                    if (Pmax == 1)&&(tau == 20)&&(area_r_ratio==5./6)
                        for t=1:(days+flight+postflight)
                            DataP1t20a12(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t20a12.txt DataP1t20a12 -ascii -
double
                    end
                    if (Pmax == 1)&&(tau == 20)&&(area_r_ratio==10./13)
                        for t=1:(days+flight+postflight)
                            DataP1t20a13(t,:) =
[TIME(t),Df(t),Dr(t),D(t),BMD(t)];
                        end
                        save DataP1t20a13.txt DataP1t20a13 -ascii -
double
                    end
                     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                    %End of formating data in charts
                    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        end
    end
end
```

APPENDIX B: FIGURES (10-DAY SPACEFLIGHT)

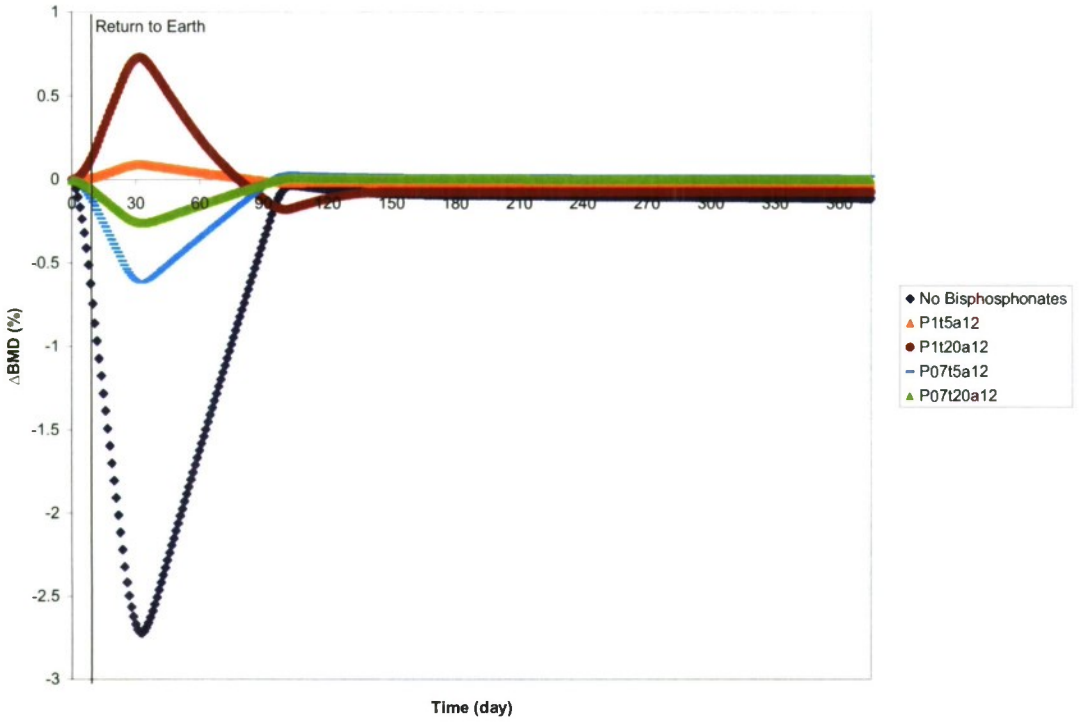


Figure B1. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 10-day spaceflight.

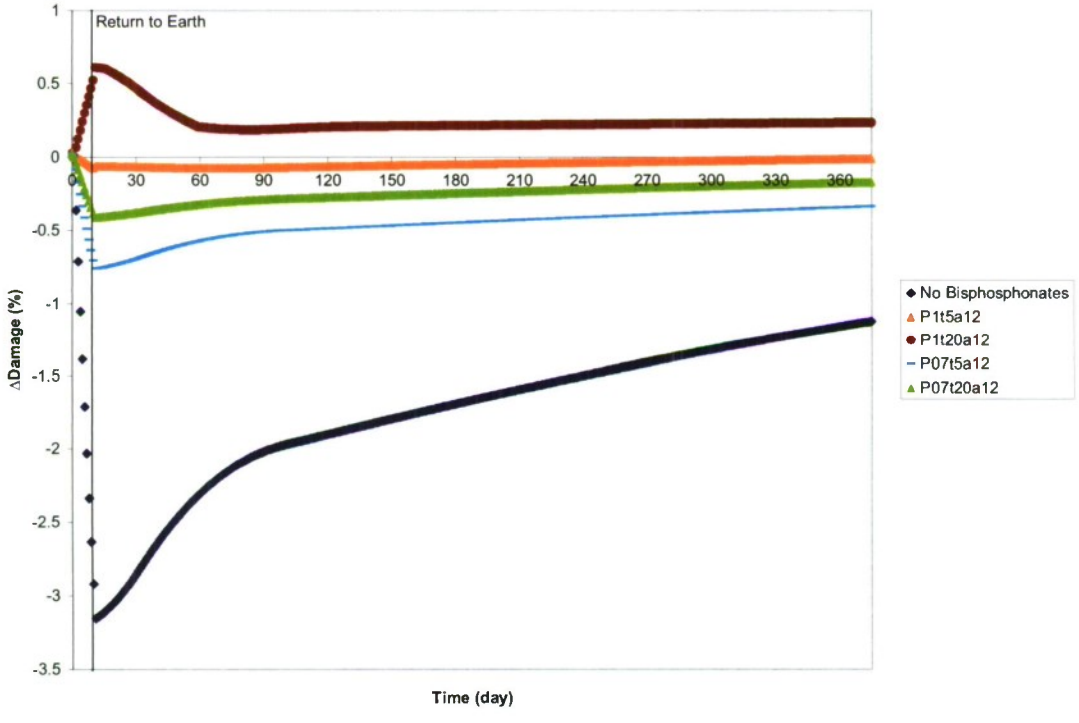


Figure B2. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

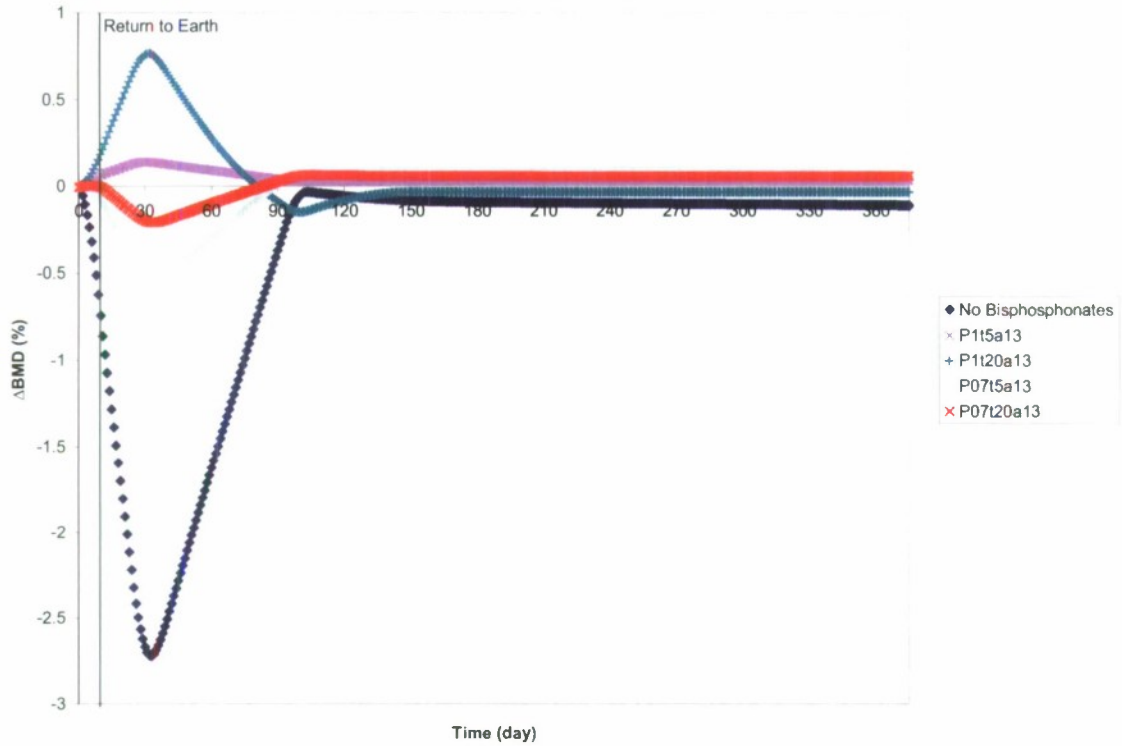


Figure B3. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 10-day spaceflight.

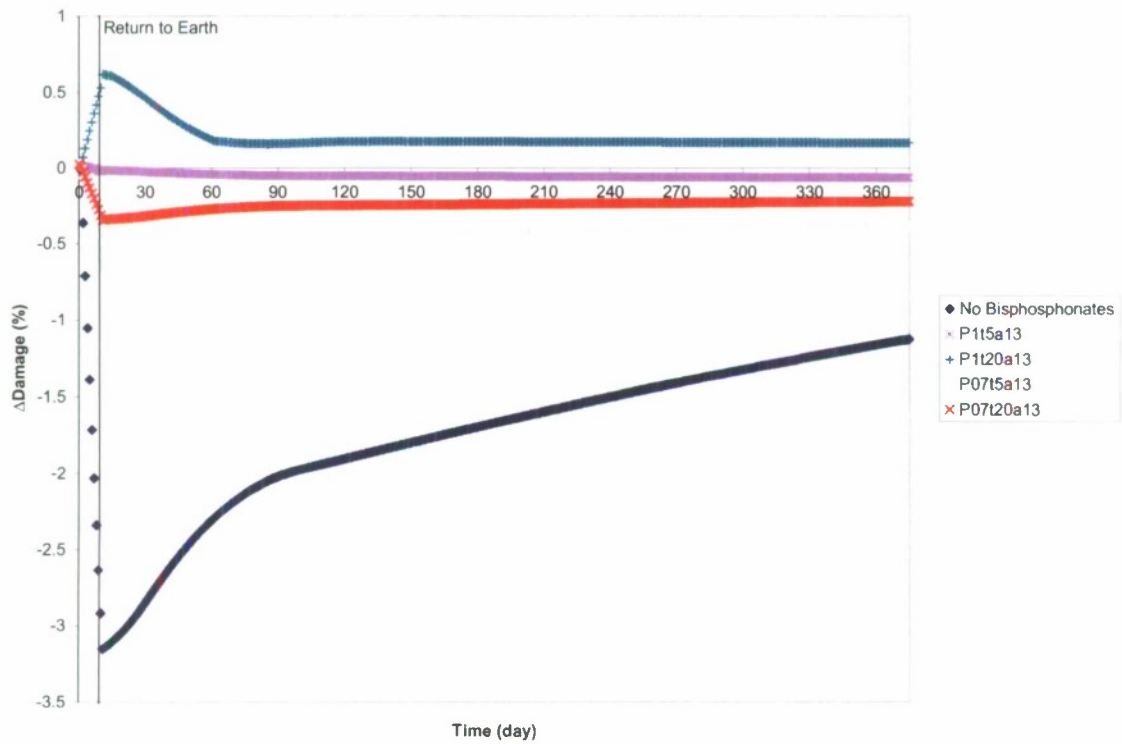


Figure B4. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

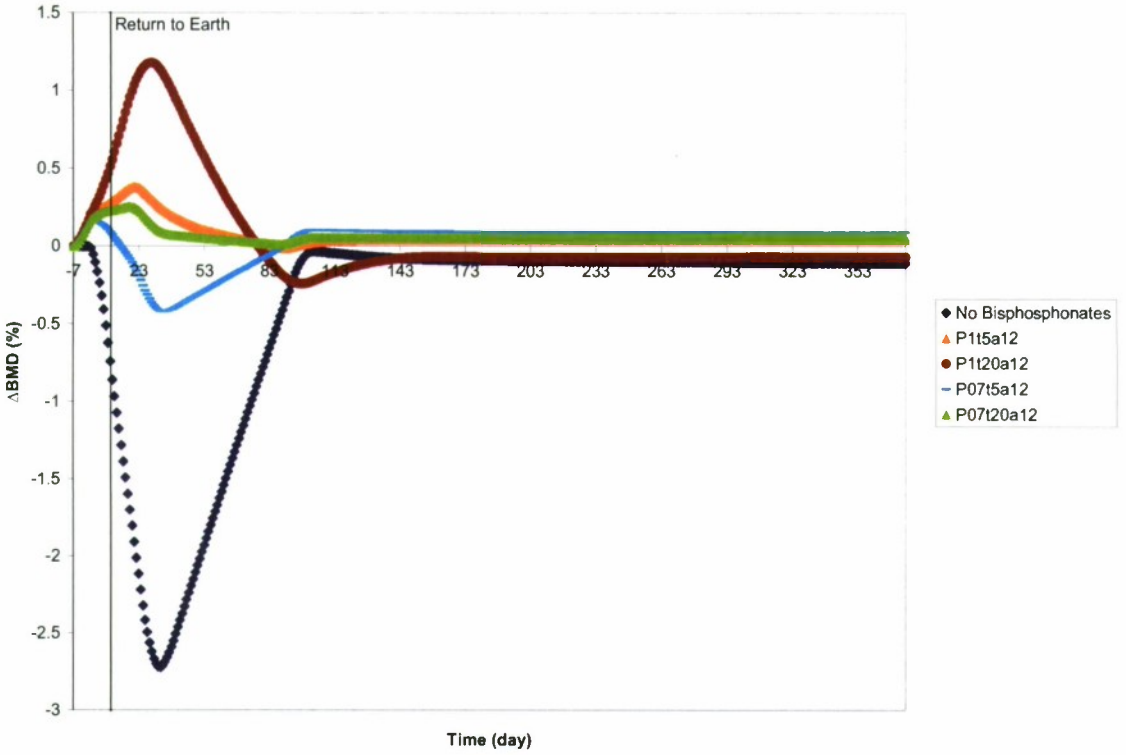


Figure B5. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

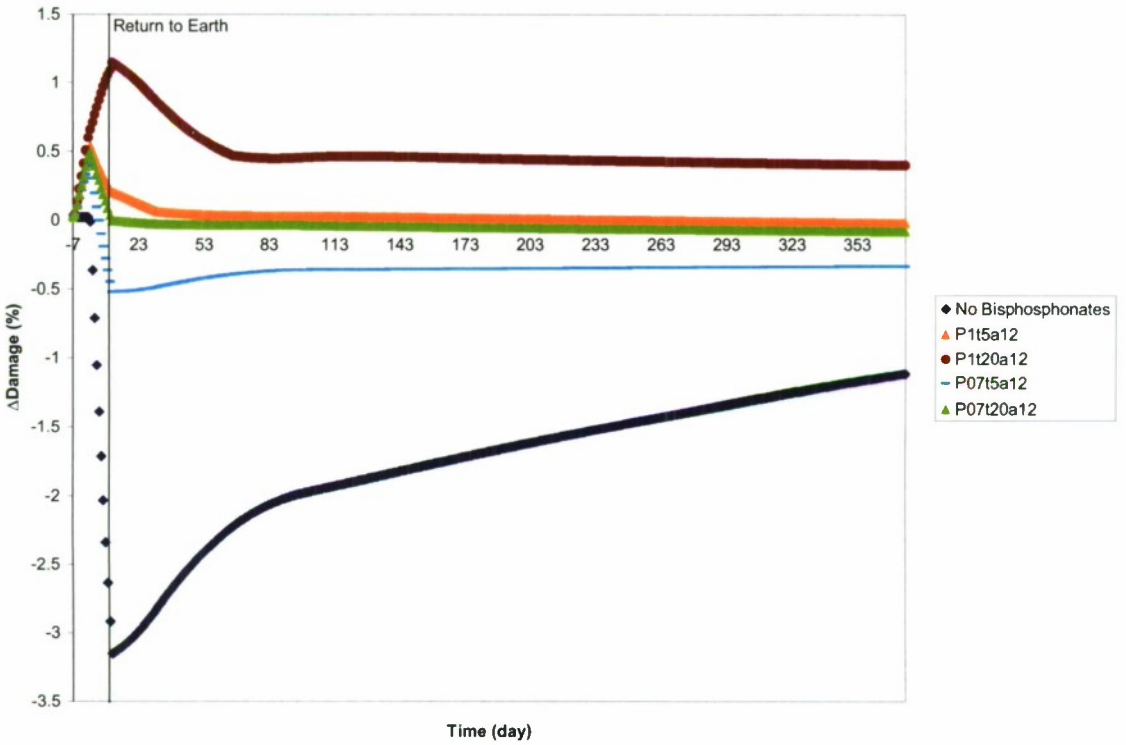


Figure B6. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

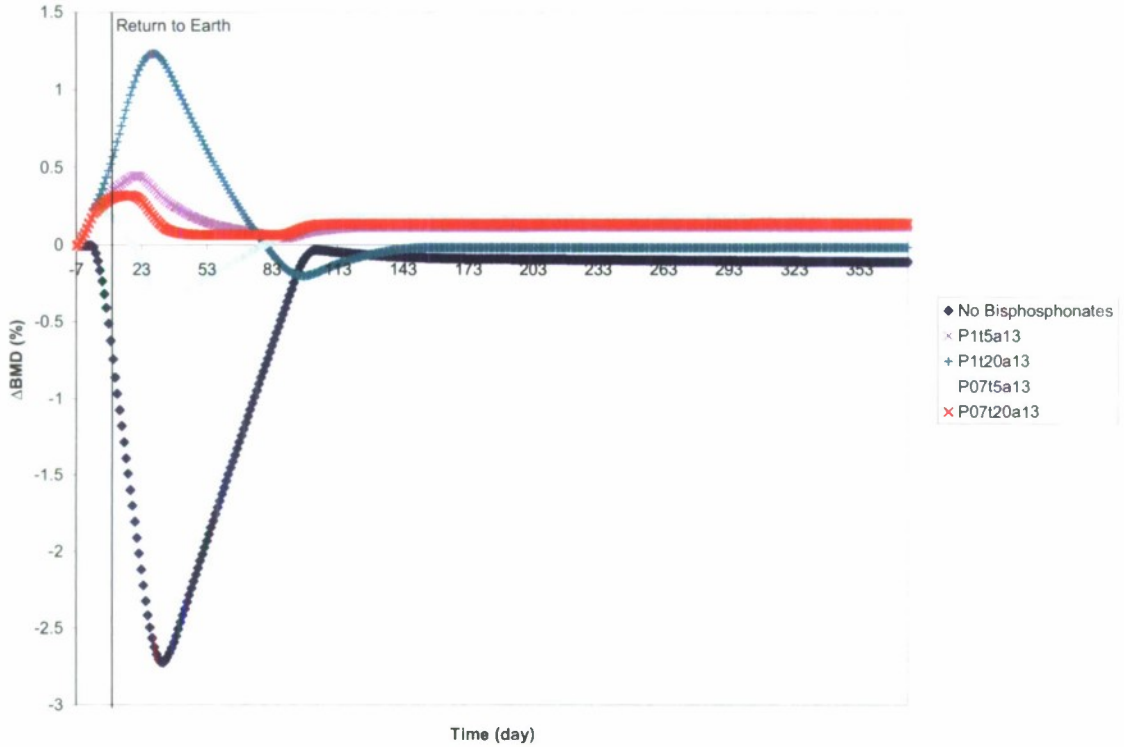


Figure B7. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

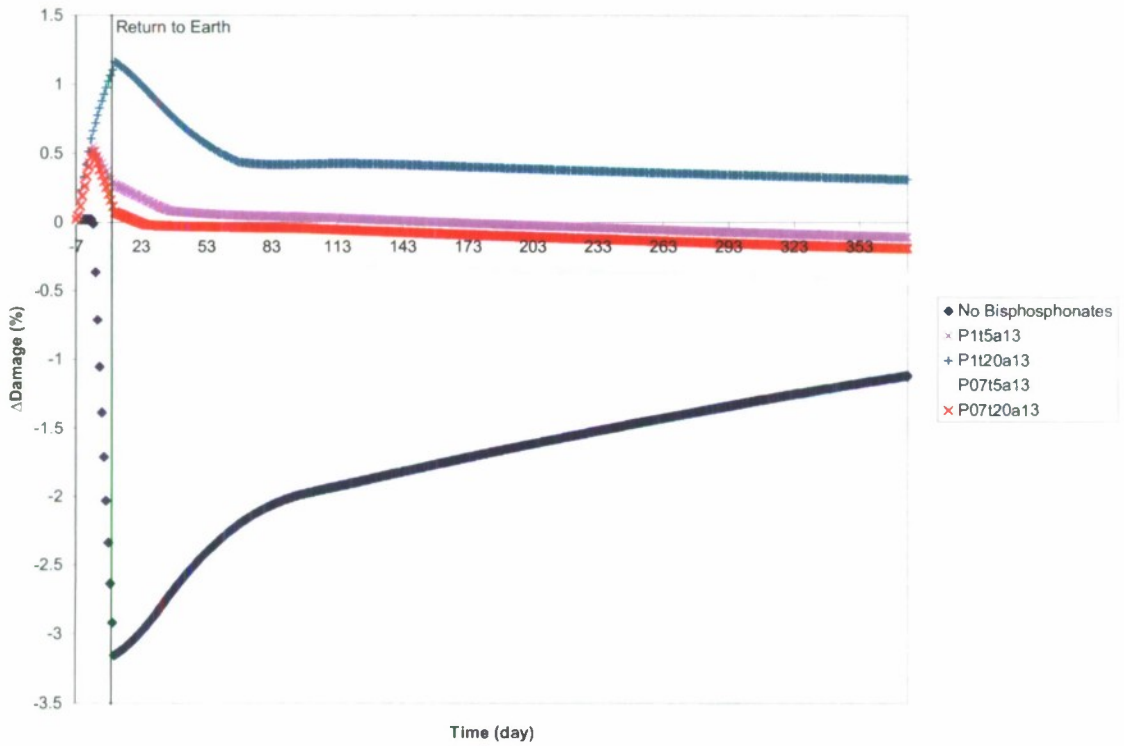


Figure B8. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

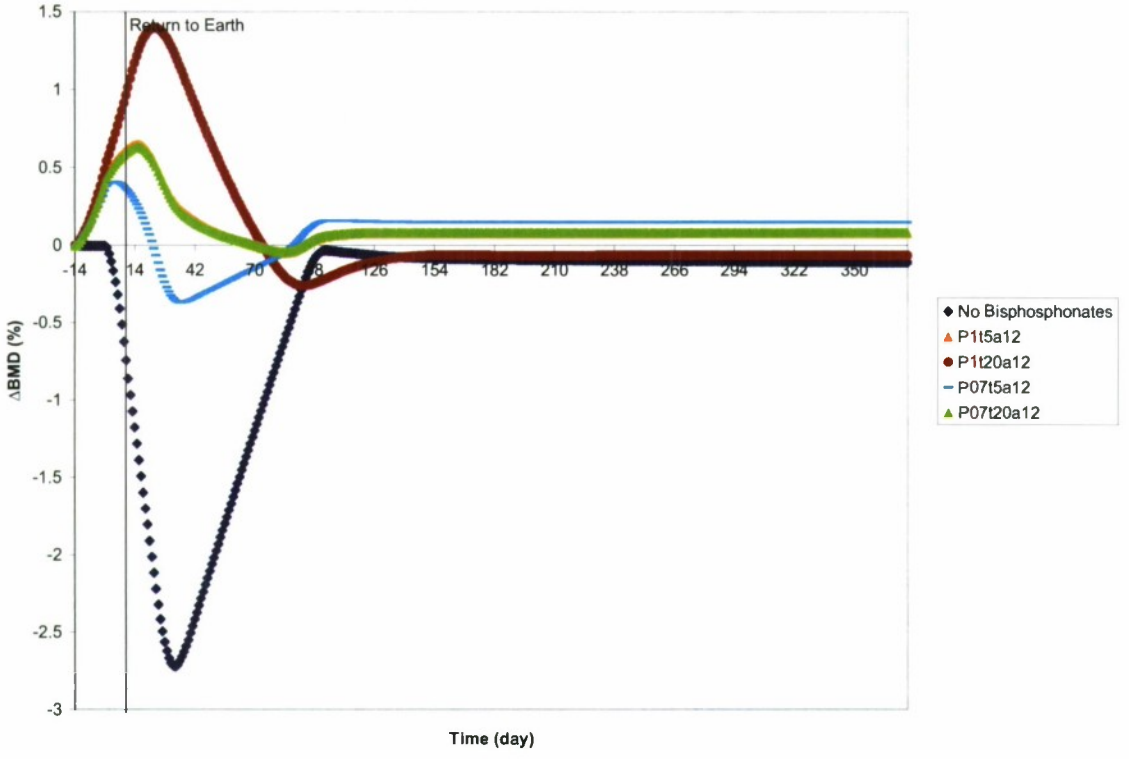


Figure B9. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

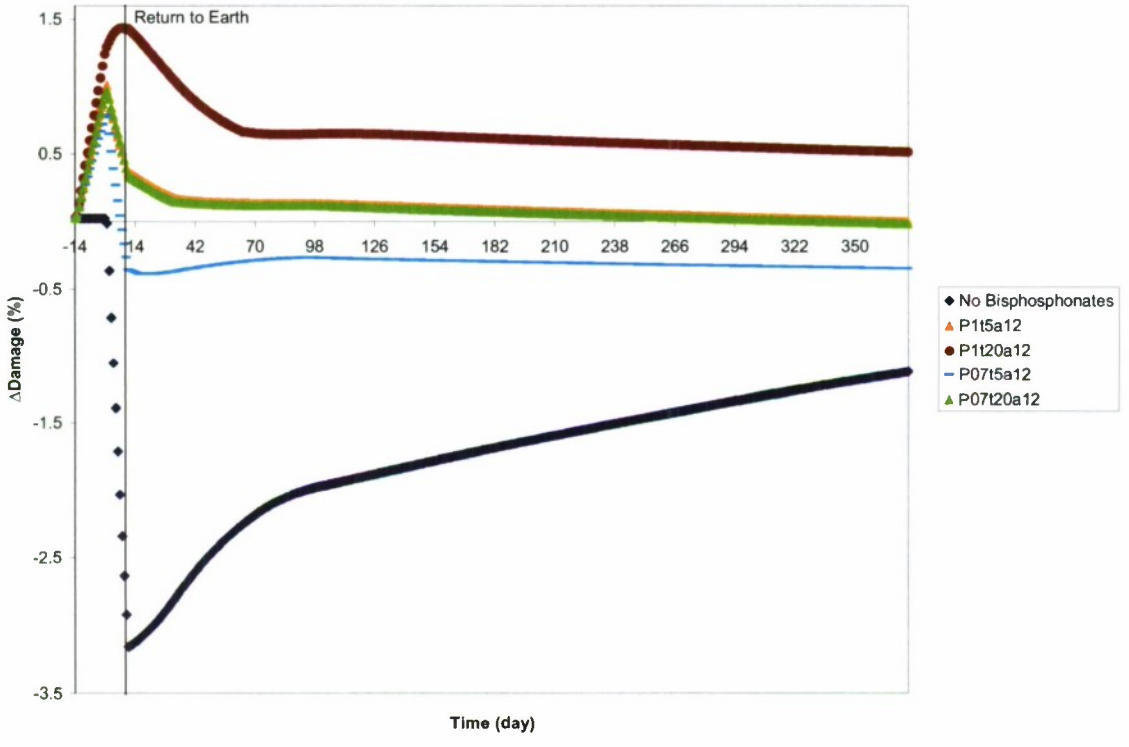


Figure B10. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

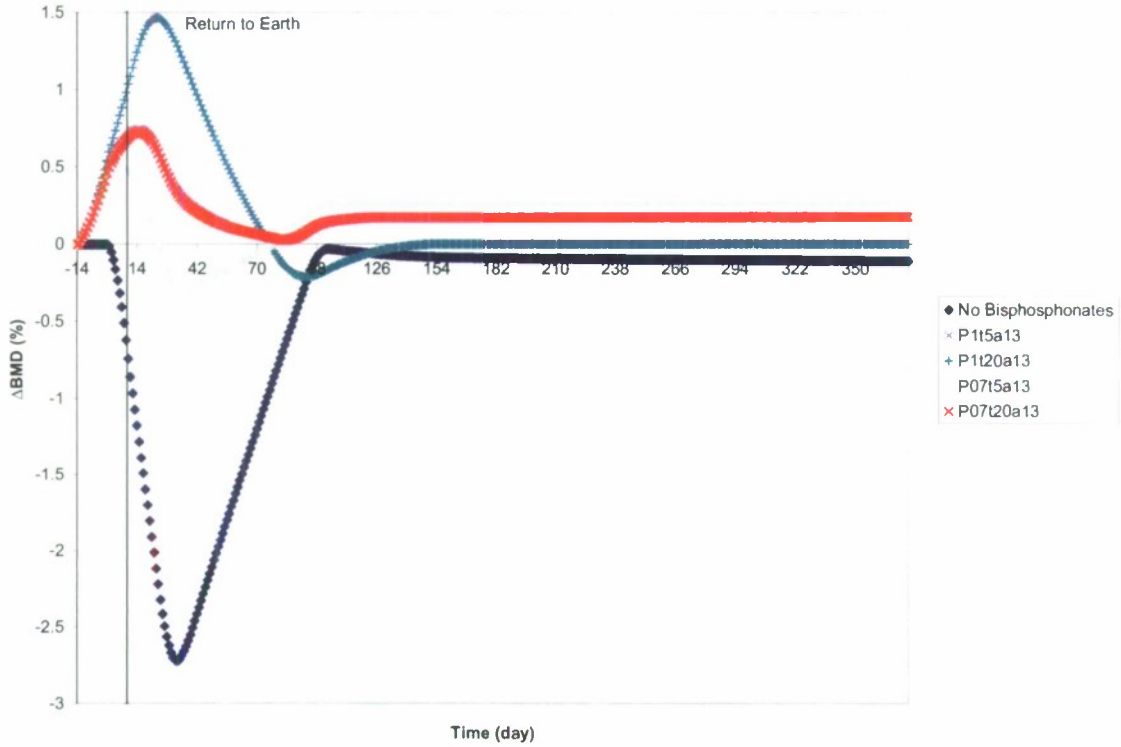


Figure B11. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatmcnt return to Earth from 10-day spaceflight.

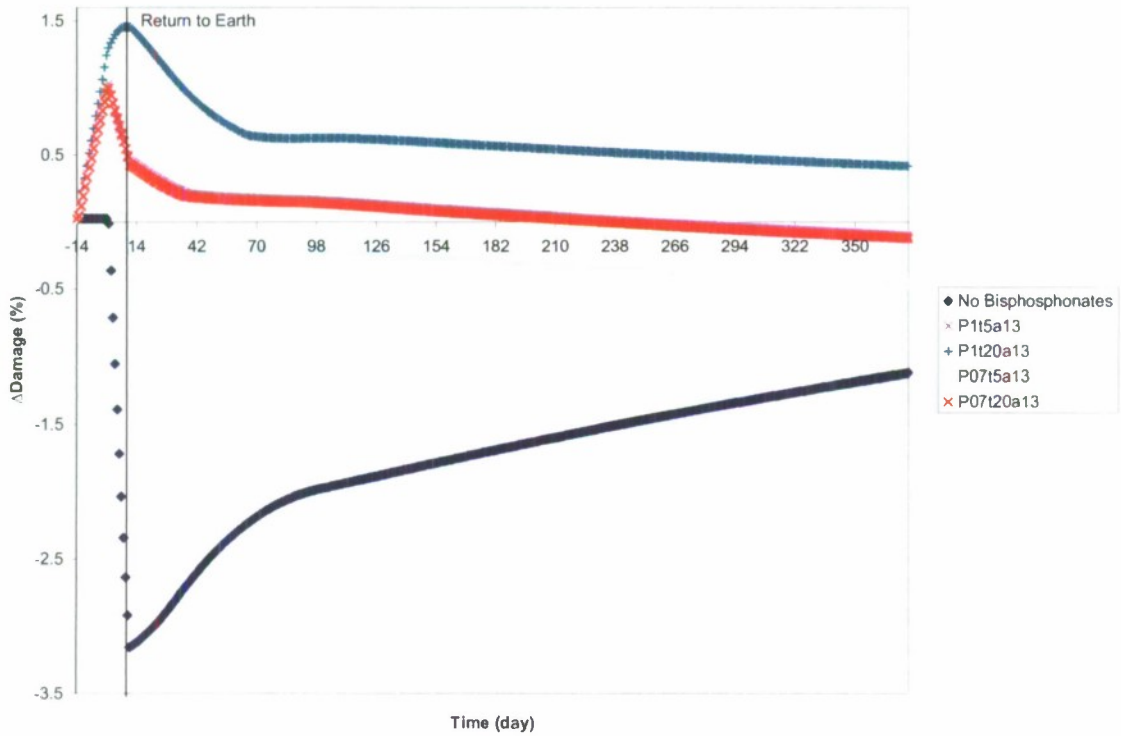


Figure B12. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

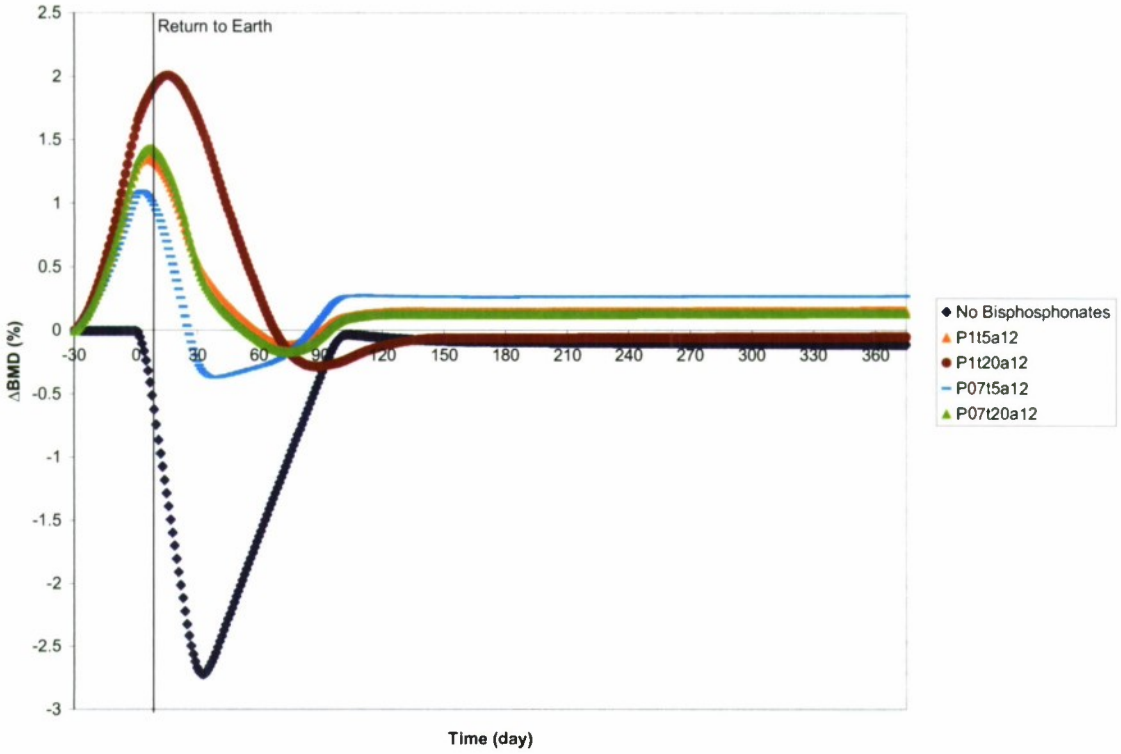


Figure B13. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

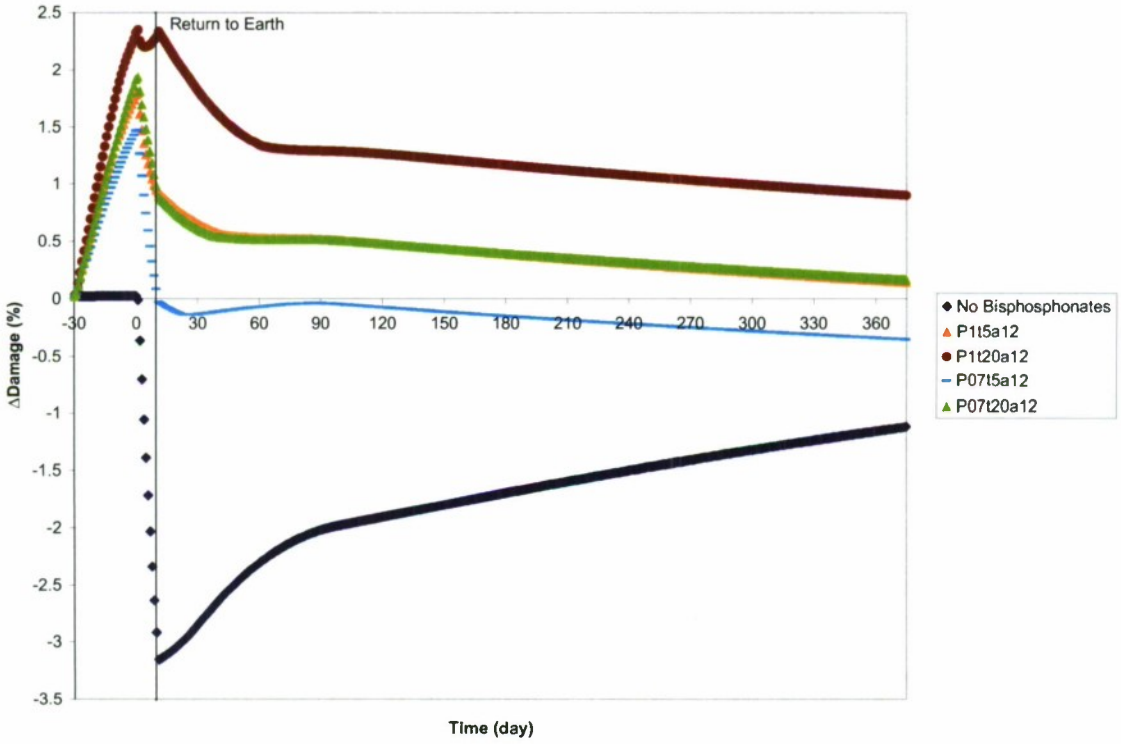


Figure B14. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

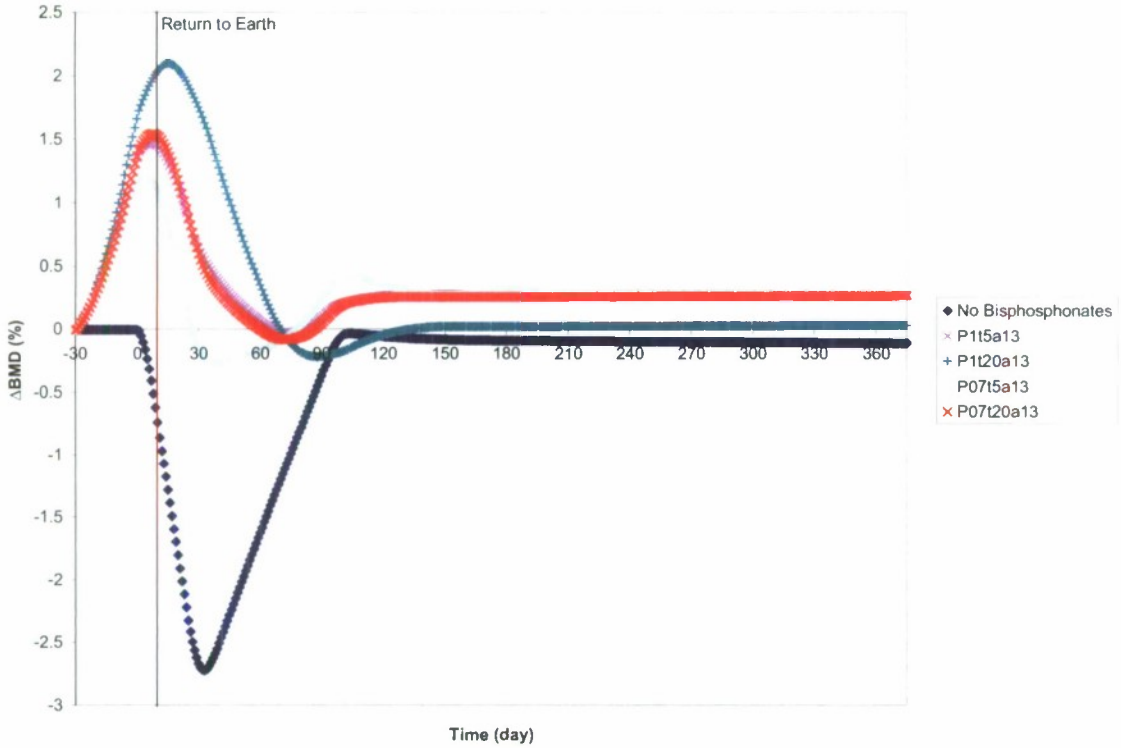


Figure B15. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

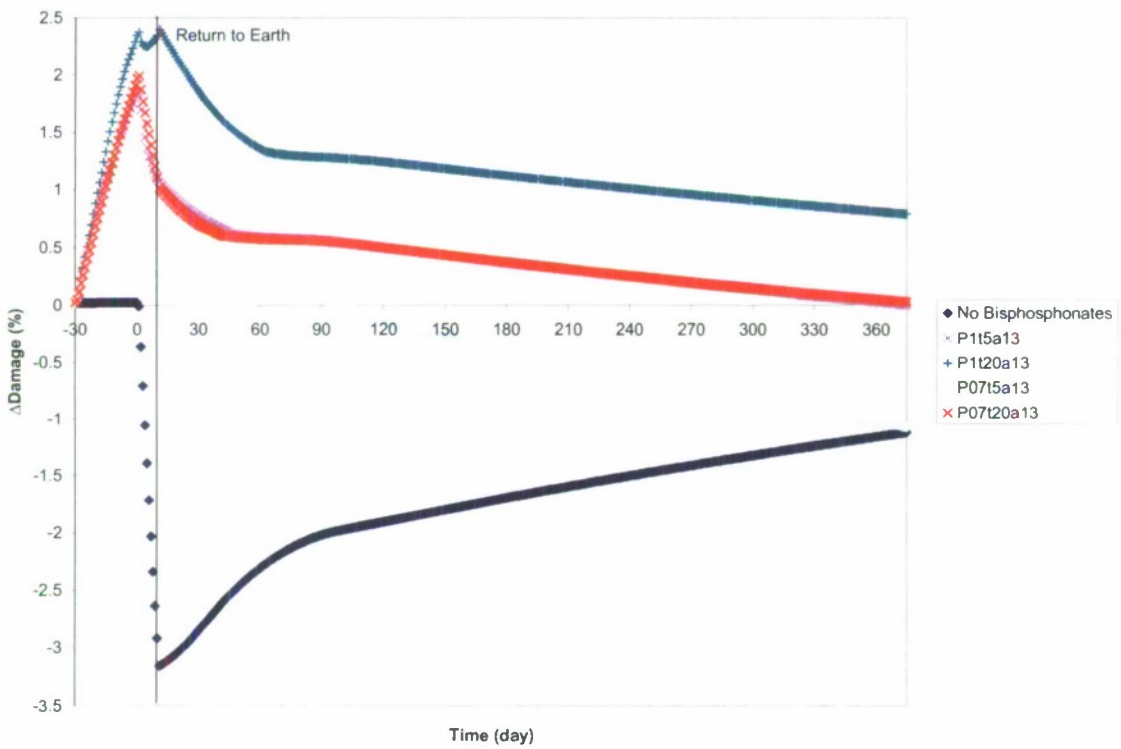


Figure B16. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

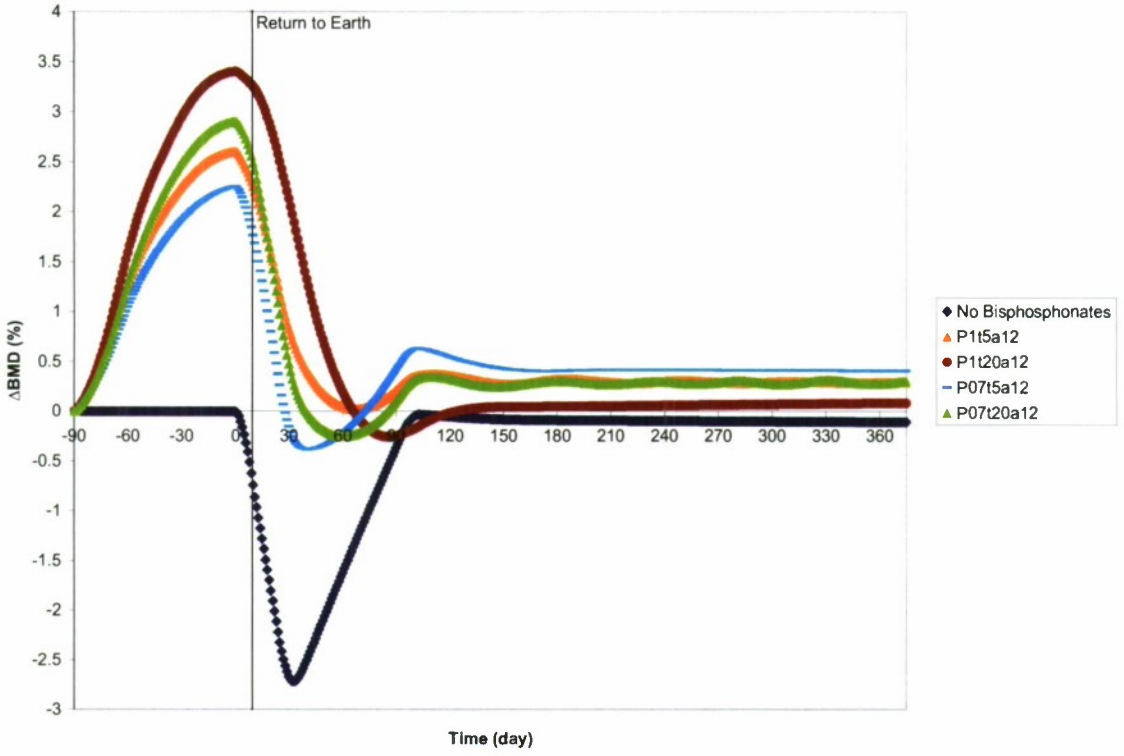


Figure B17. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

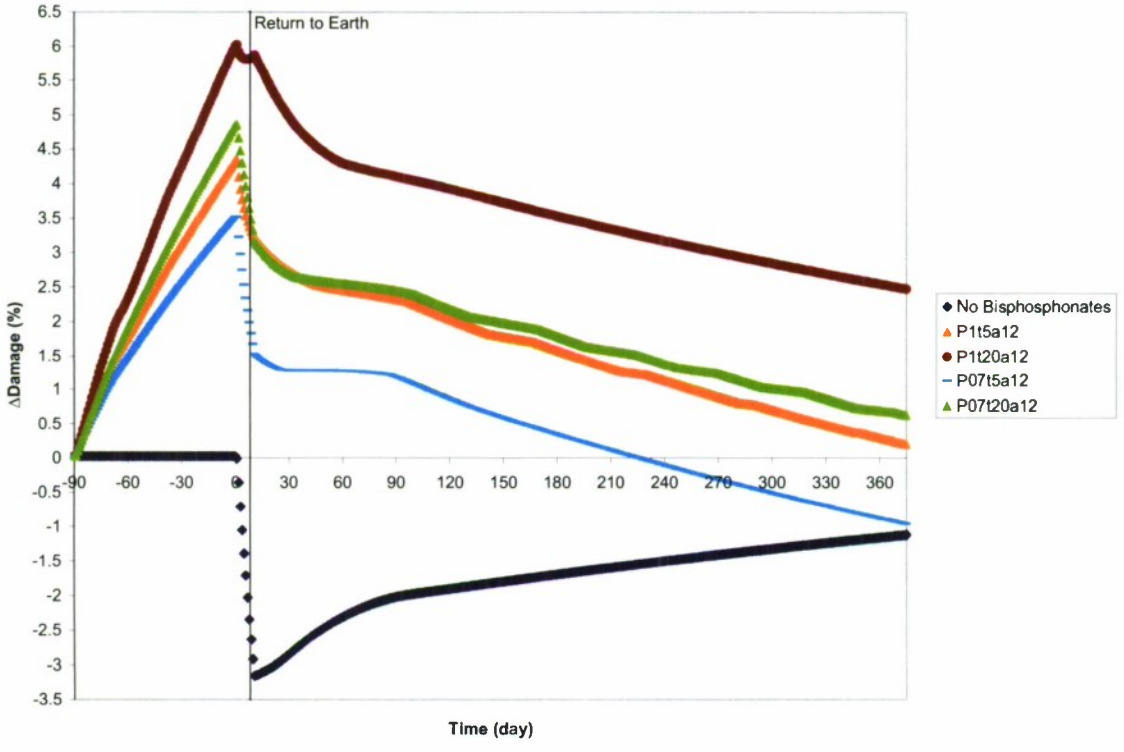


Figure B18. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

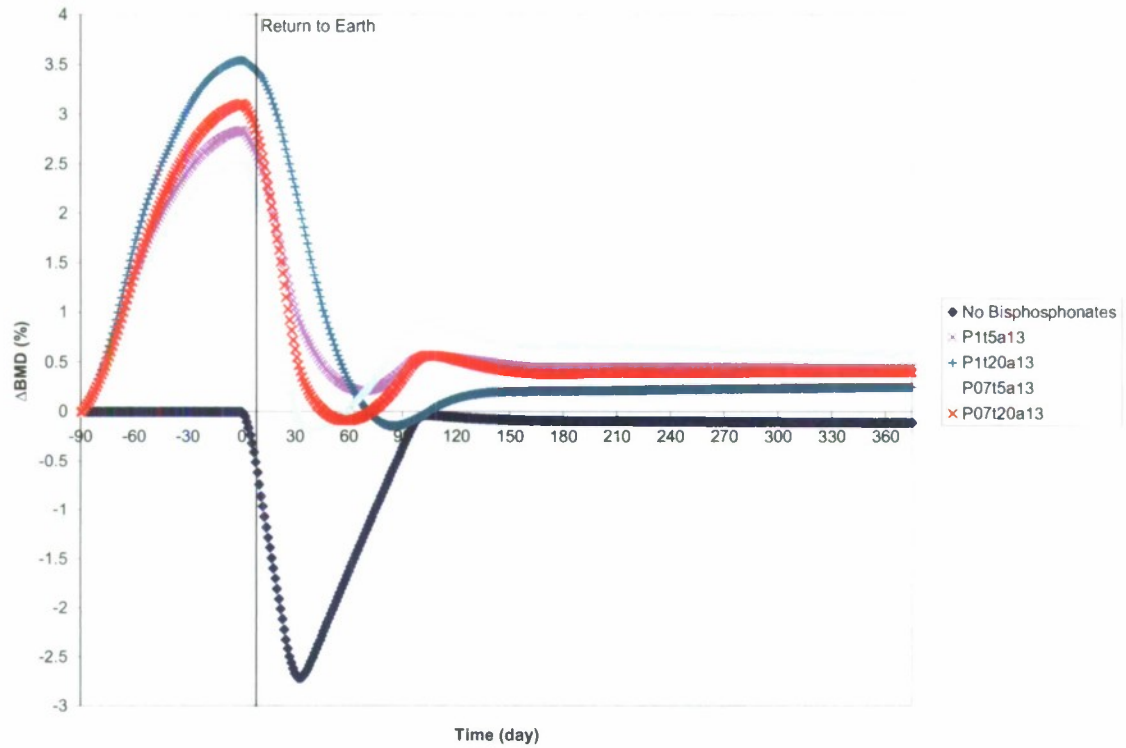


Figure B19. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

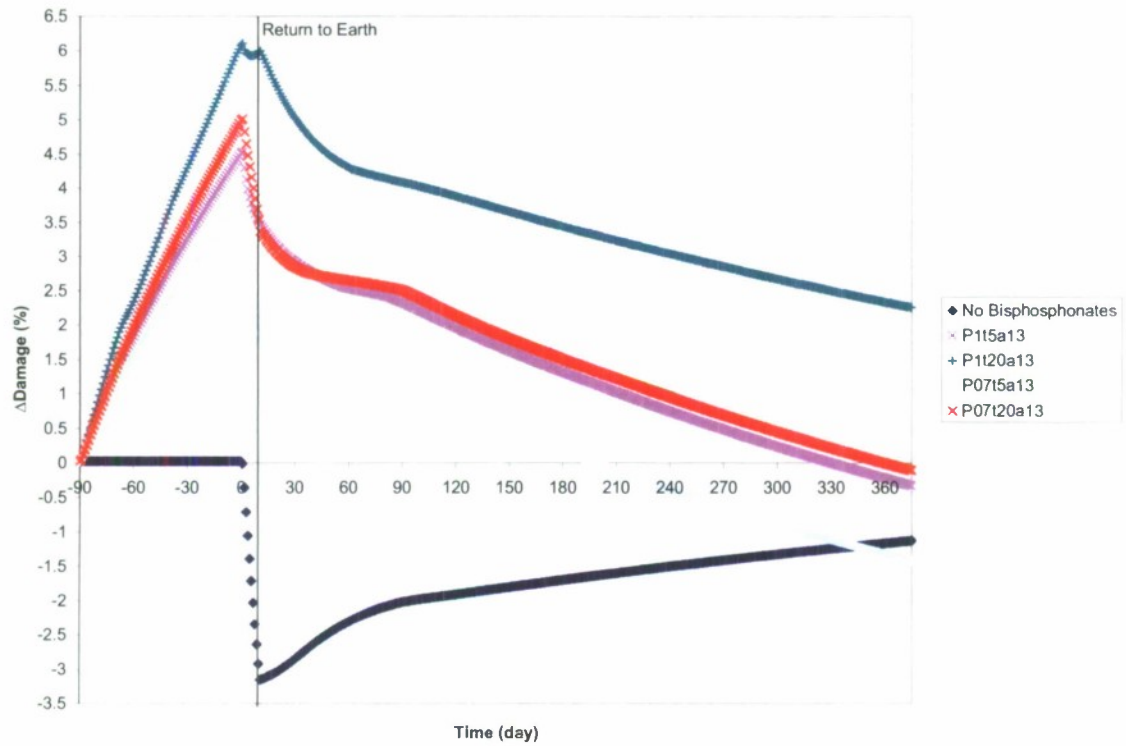


Figure B20. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

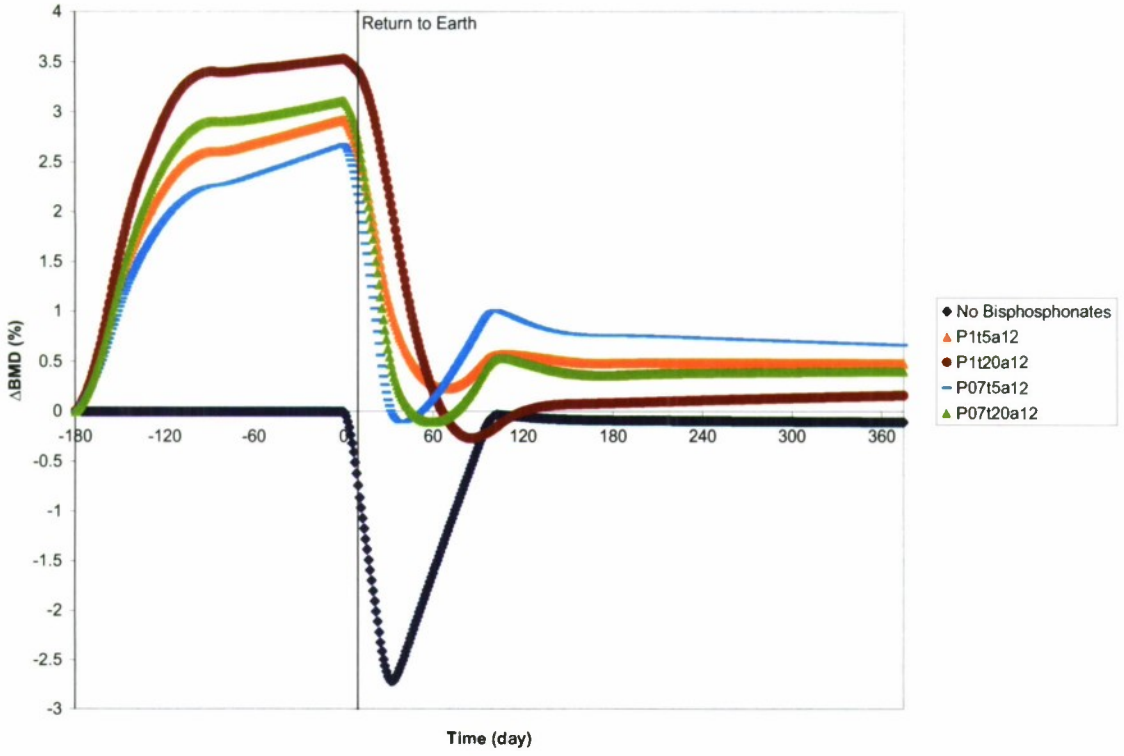


Figure B21. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

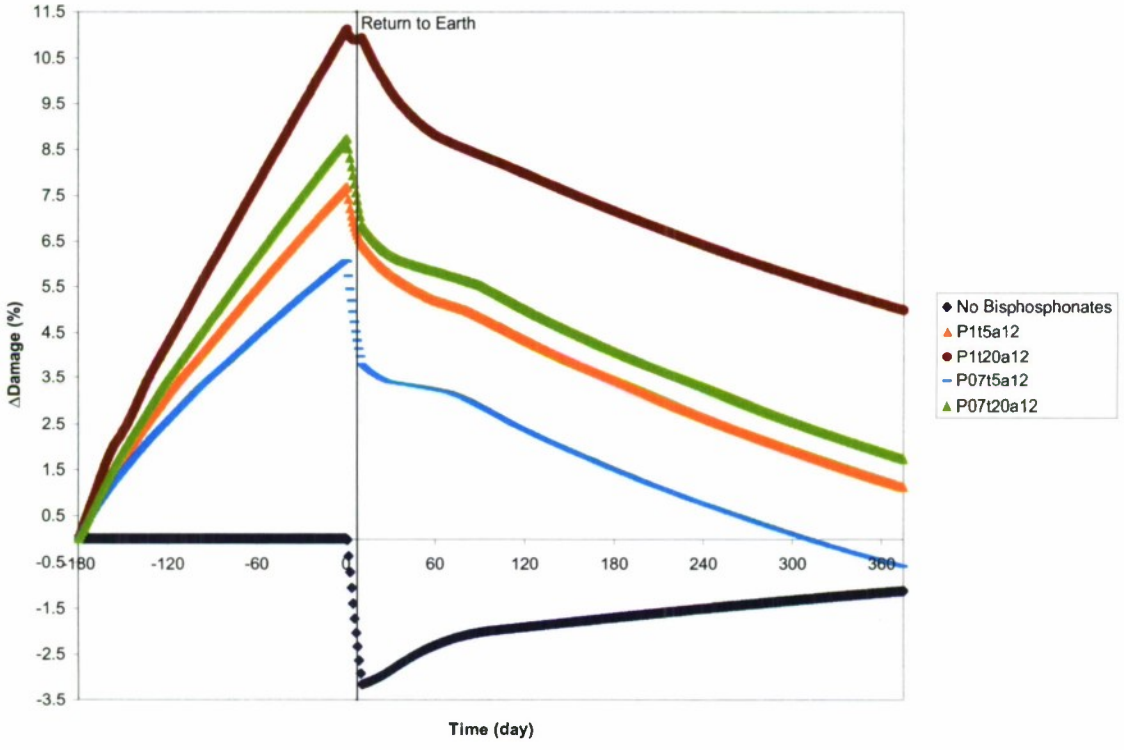


Figure B22. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

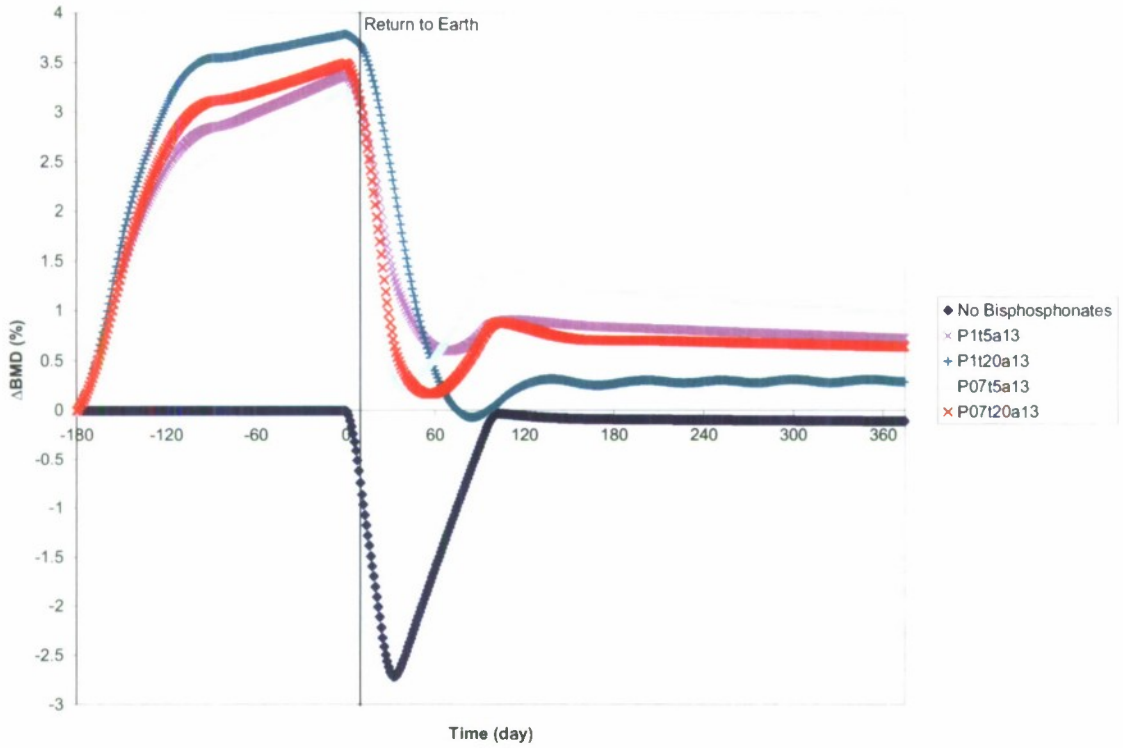


Figure B23. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 10-day spaceflight.

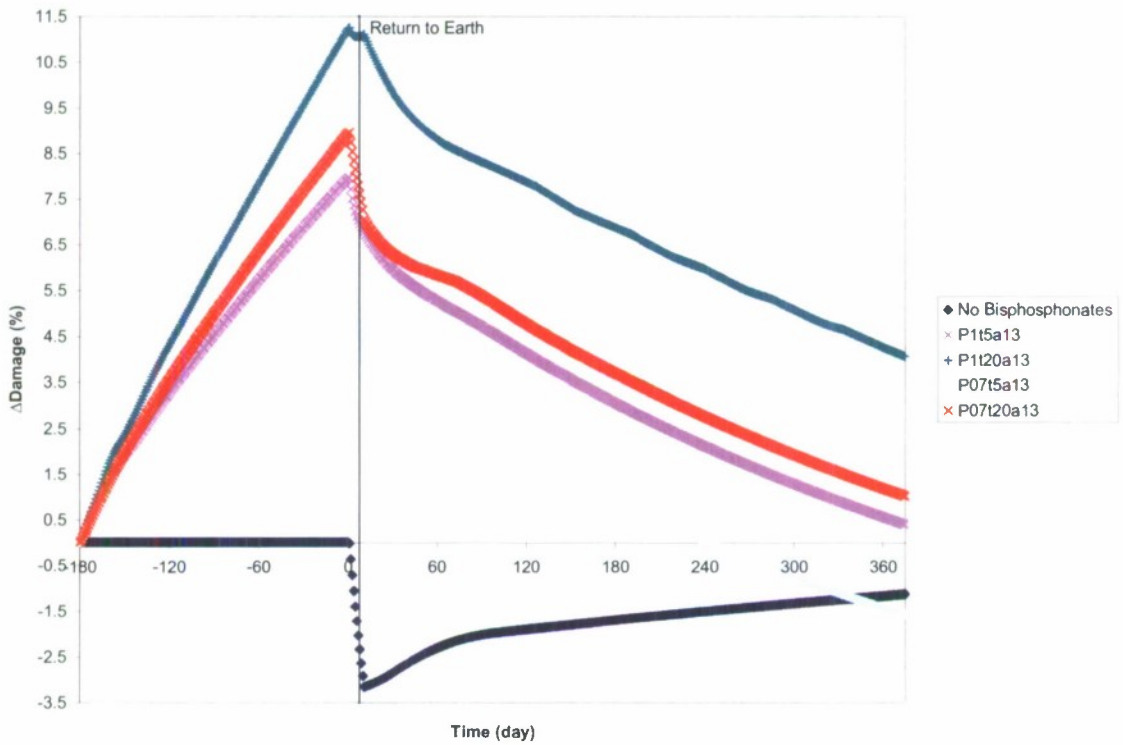


Figure B24. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 10-day spaceflight.

APPENDIX C: FIGURES (90-DAY SPACEFLIGHT)

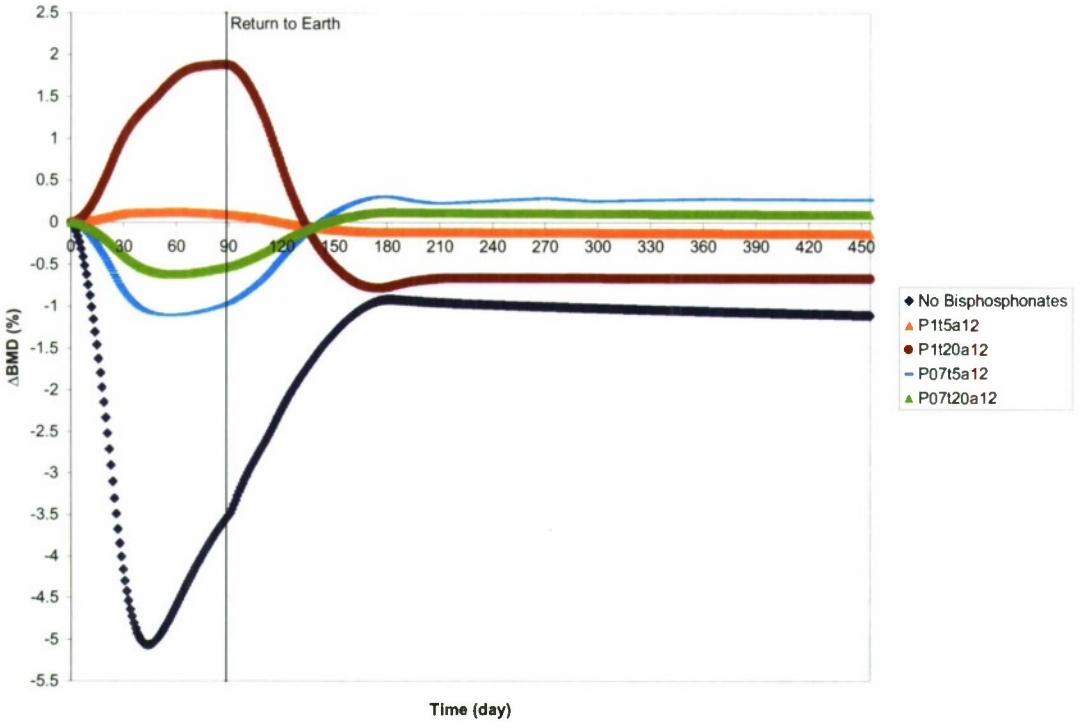


Figure C1. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 90-day spaceflight.

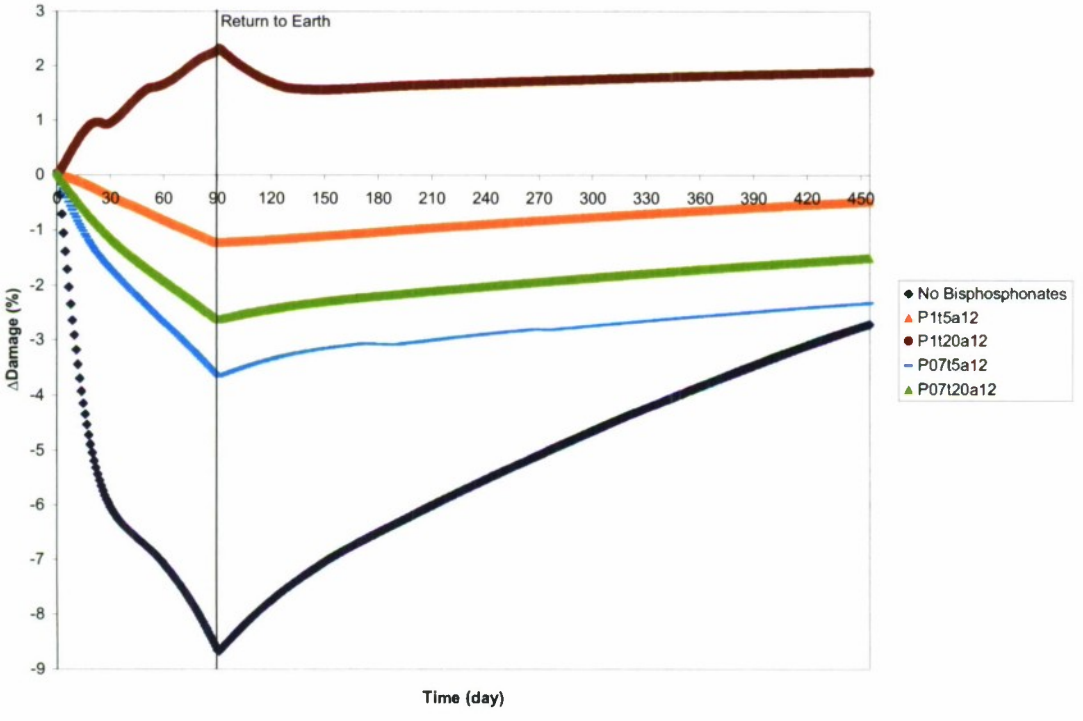


Figure C2. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

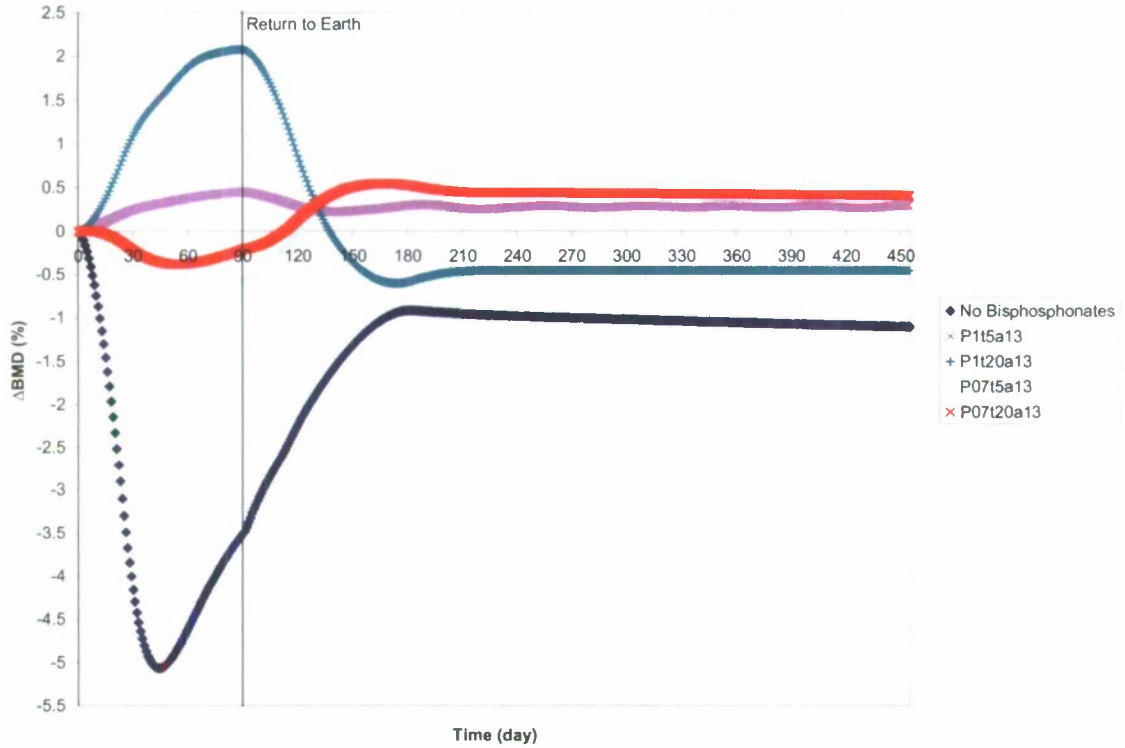


Figure C3. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 90-day spaceflight.

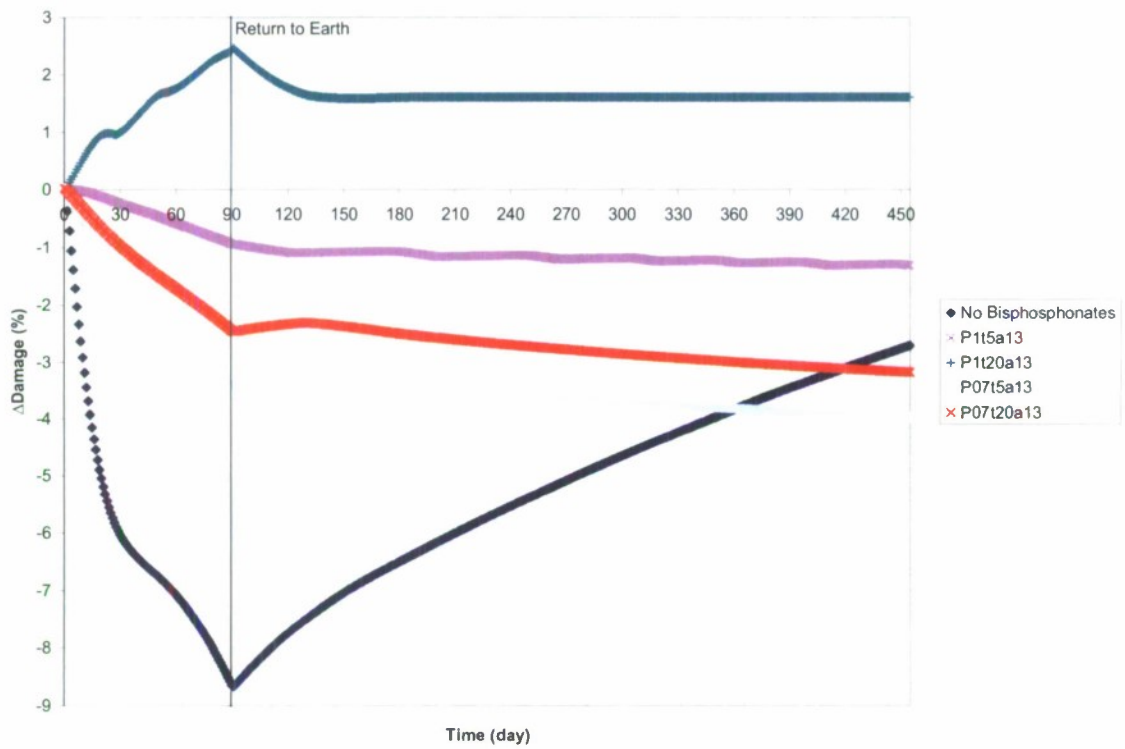


Figure C4. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

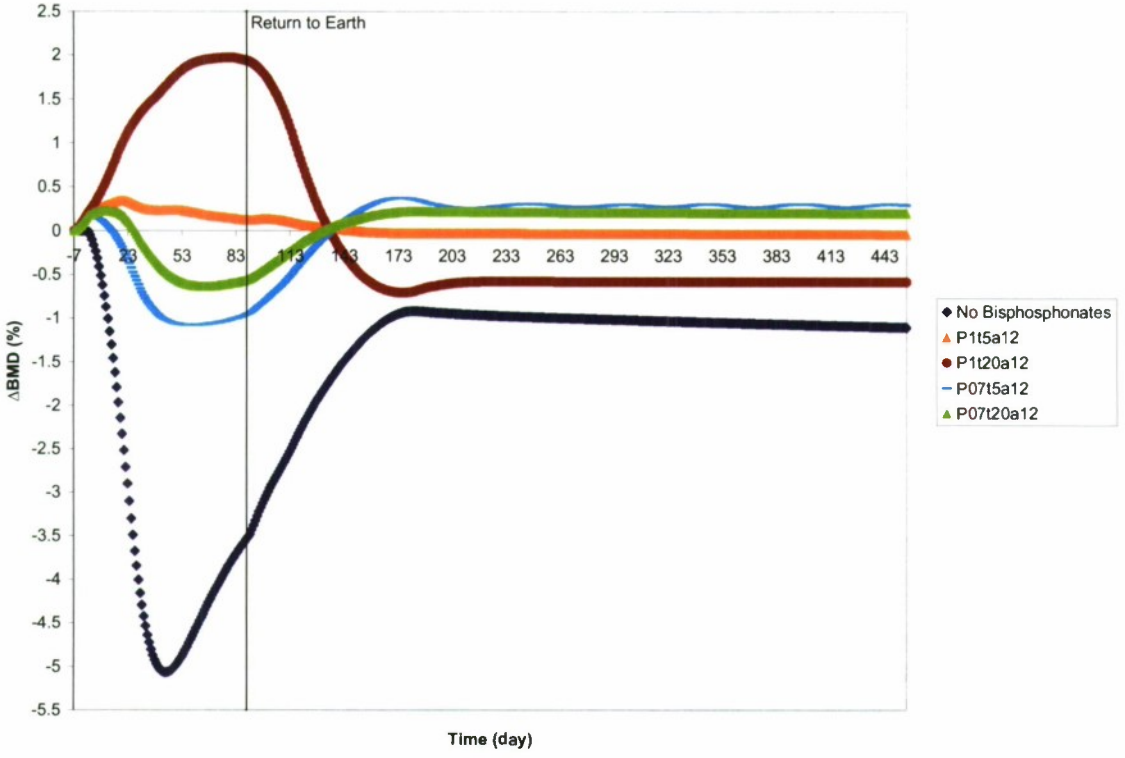


Figure C5. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

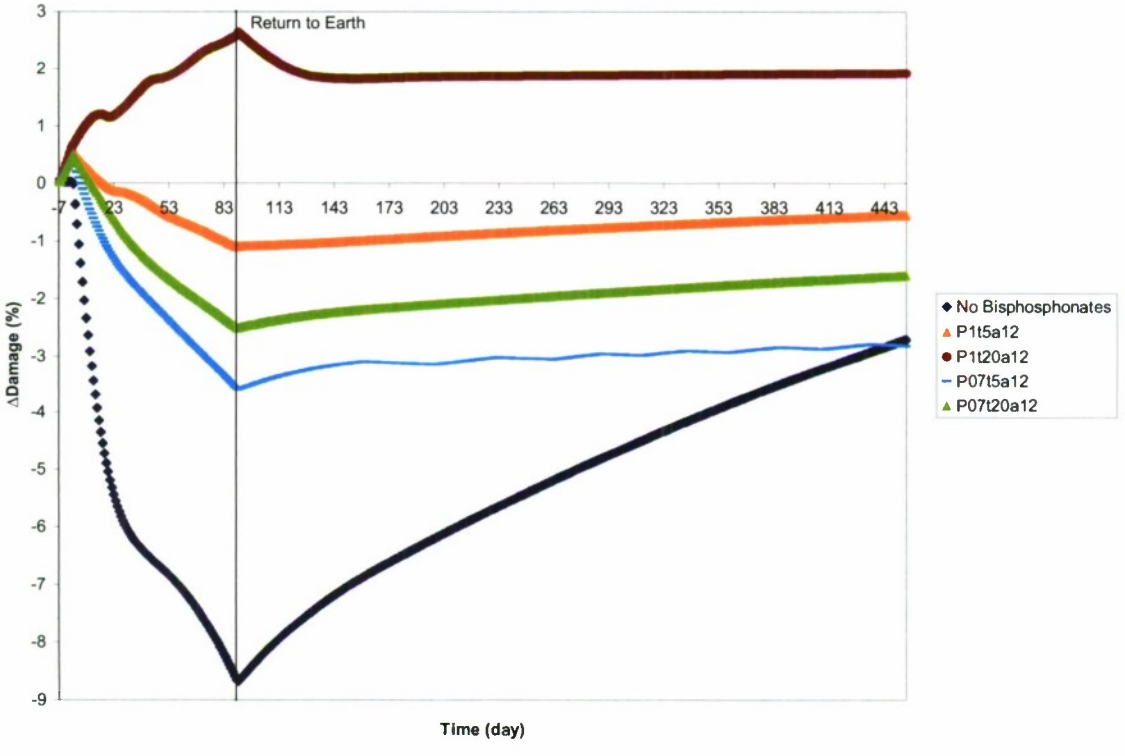


Figure C6. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

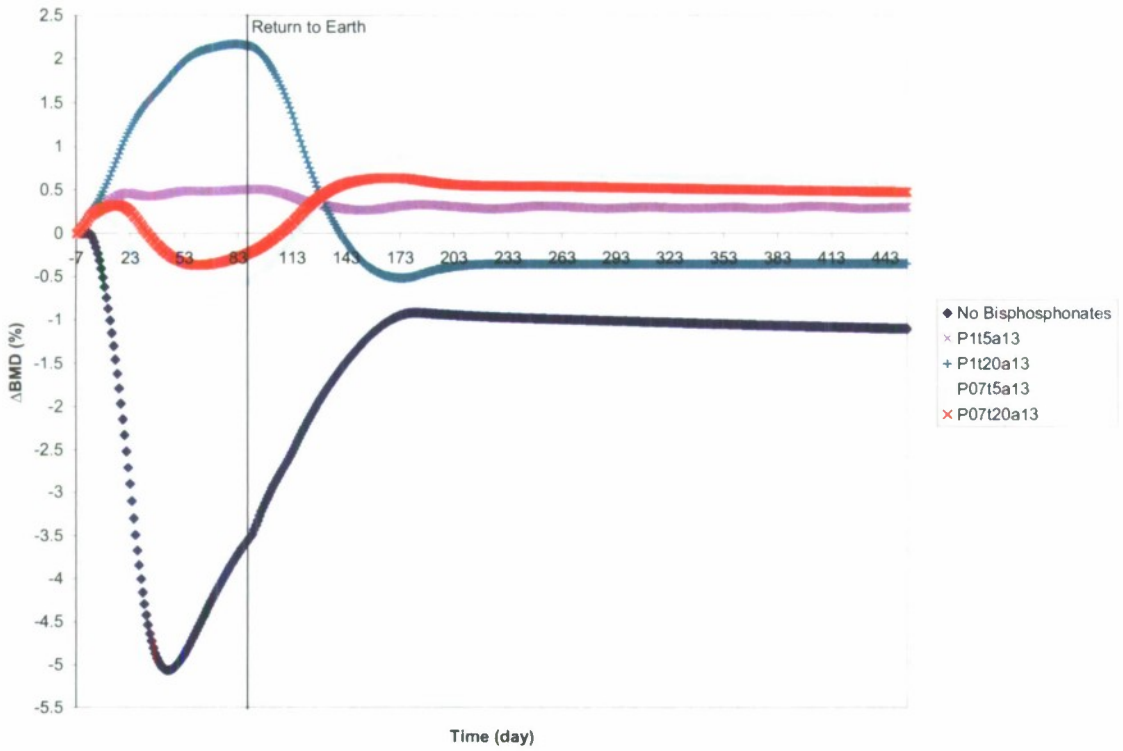


Figure C7. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

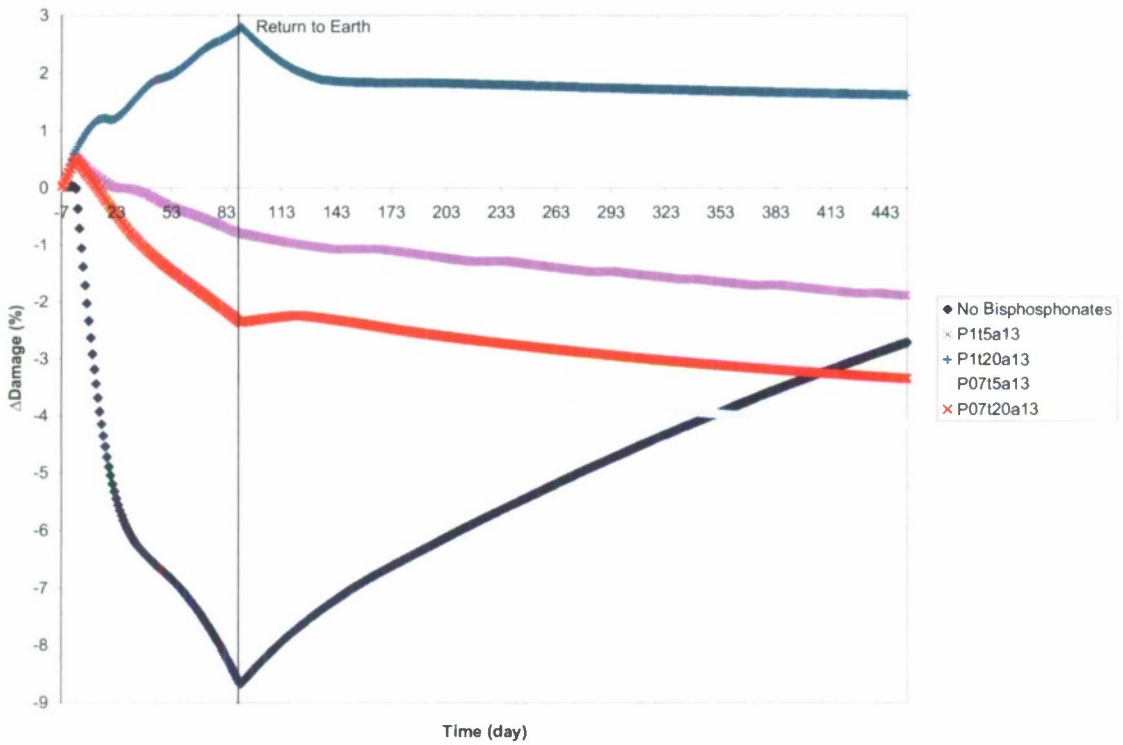


Figure C8. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

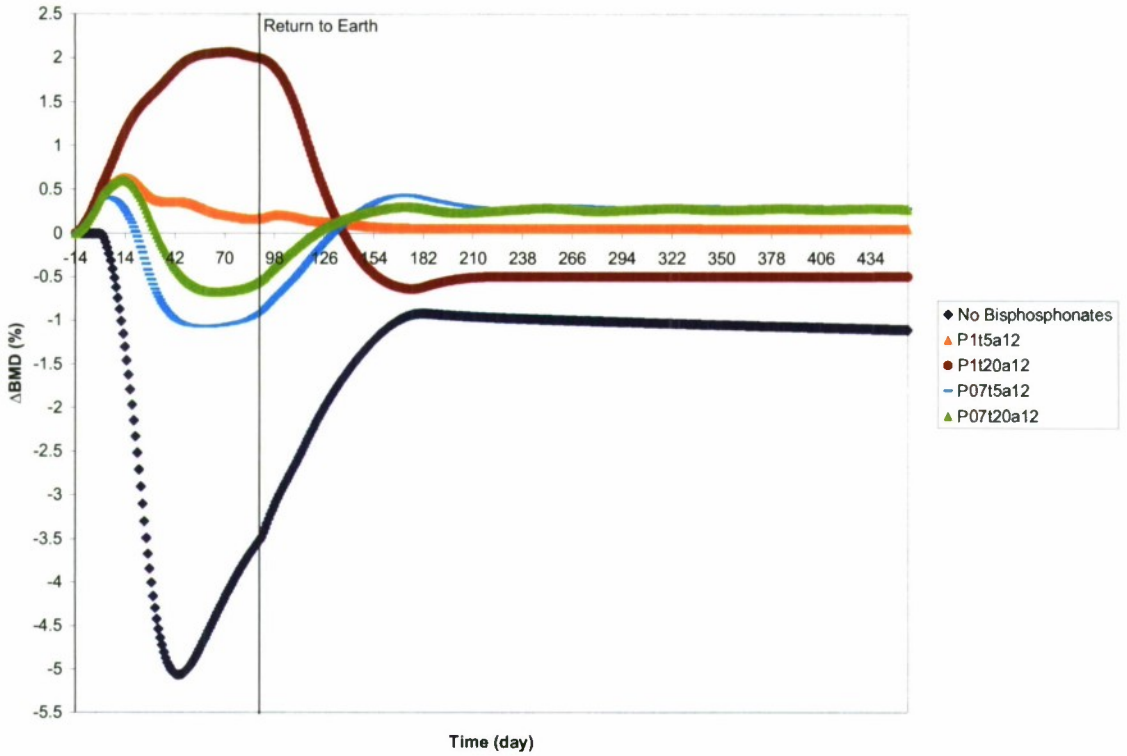


Figure C9. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

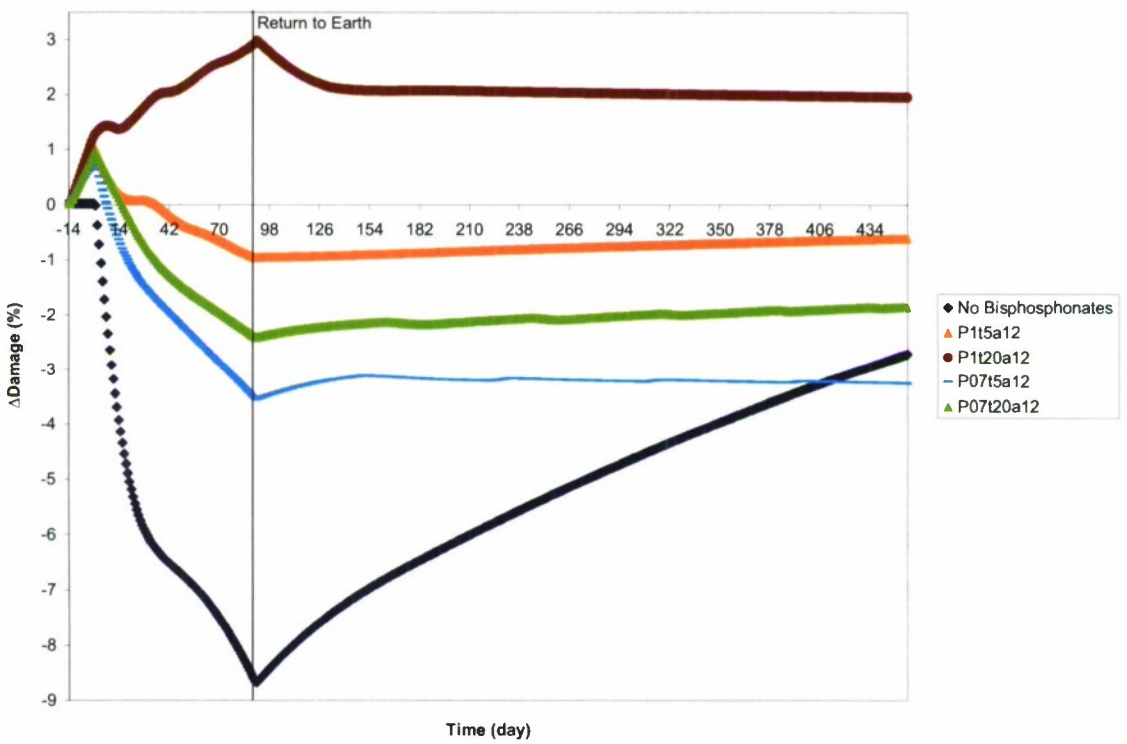


Figure C10. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

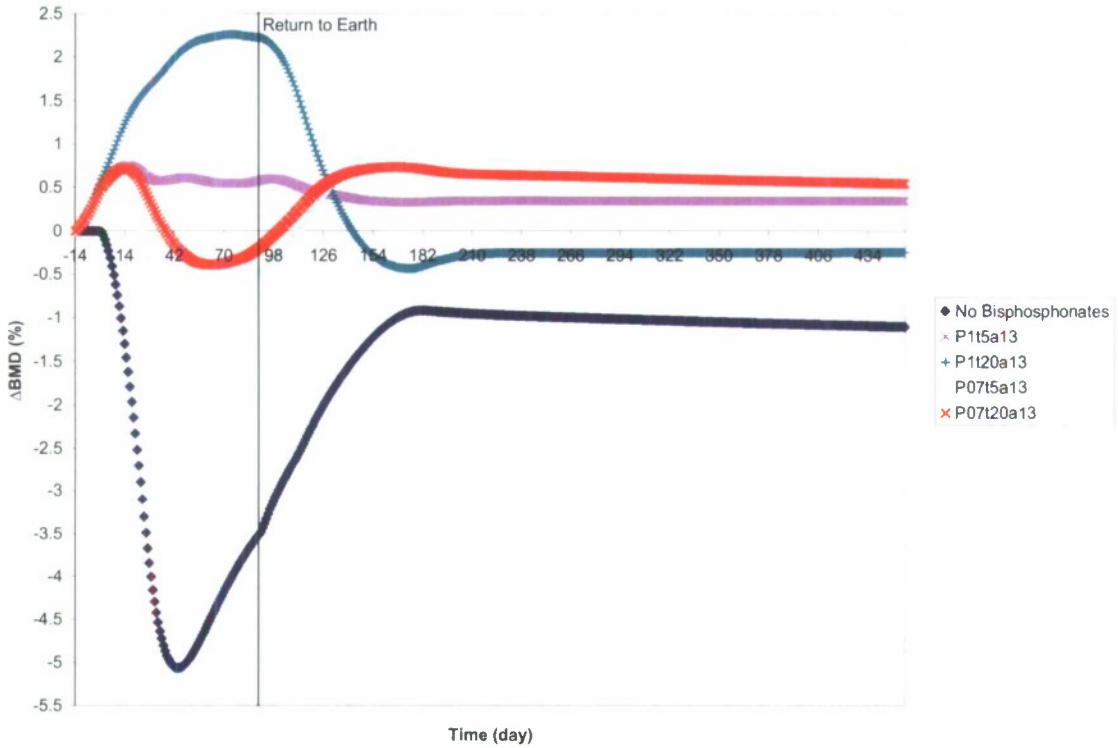


Figure C11. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

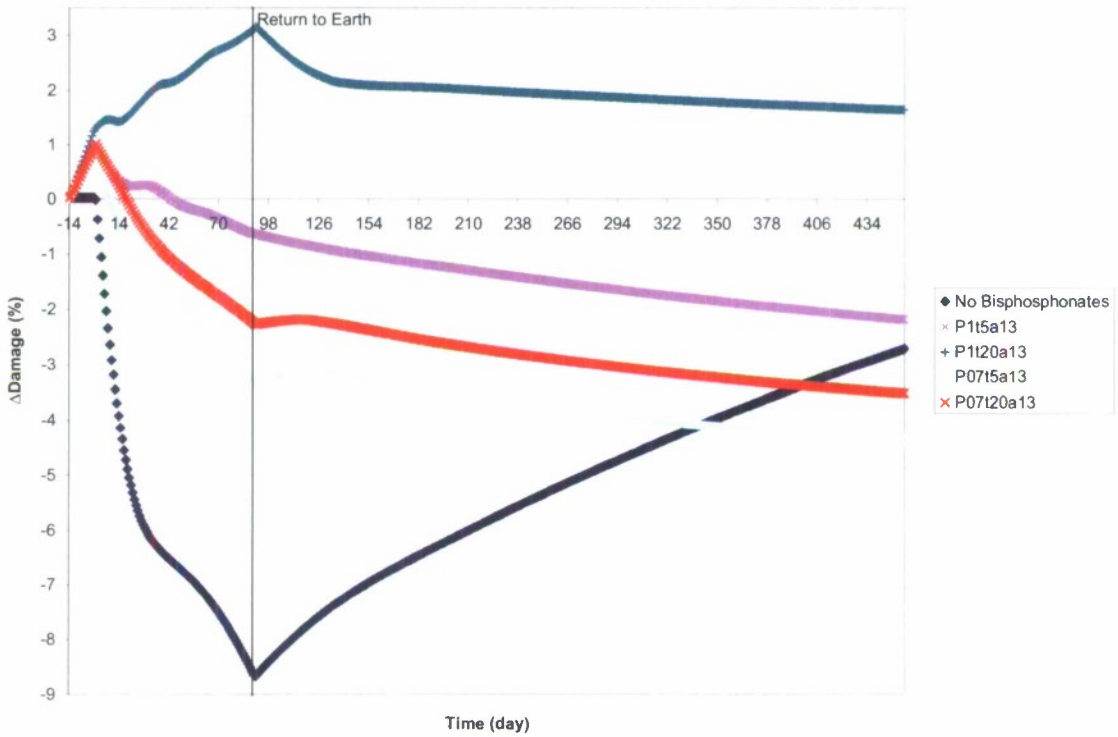


Figure C12. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

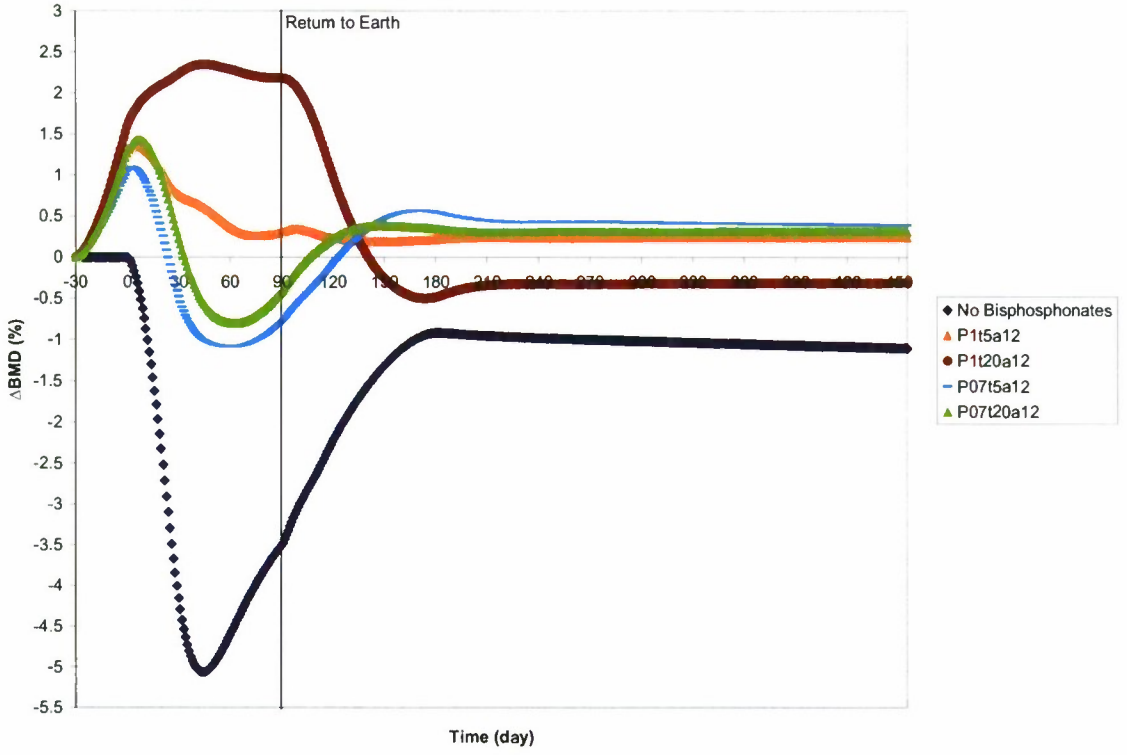


Figure C13. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

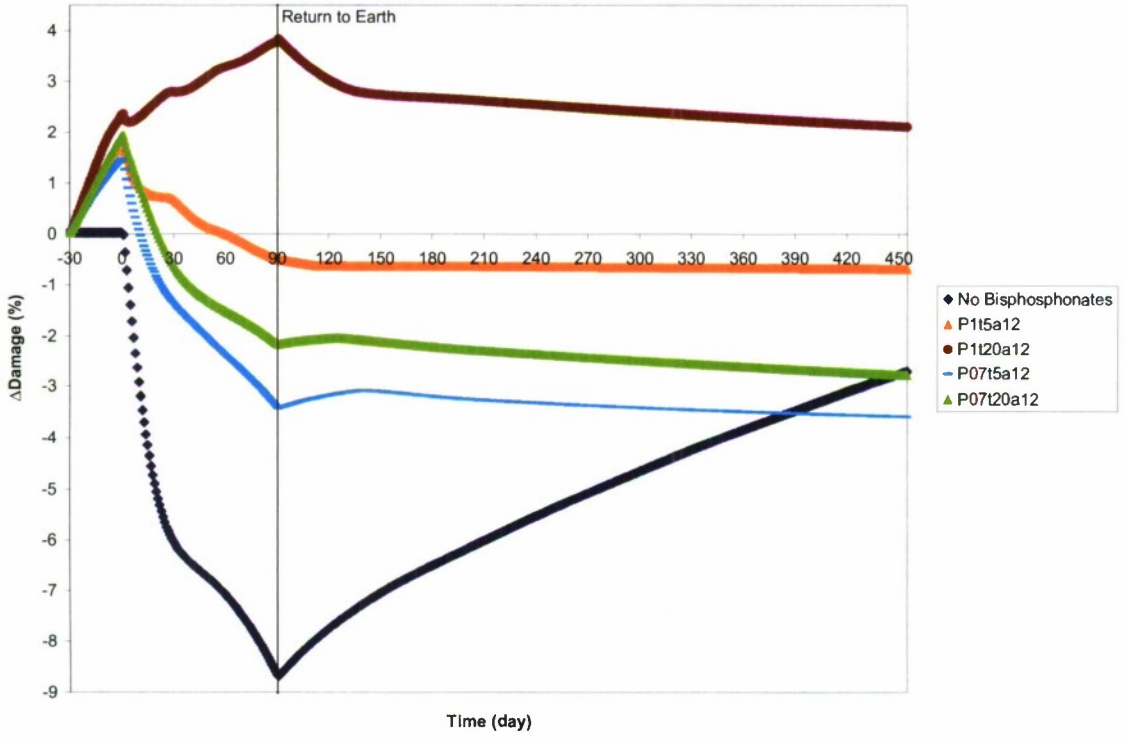


Figure C14. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

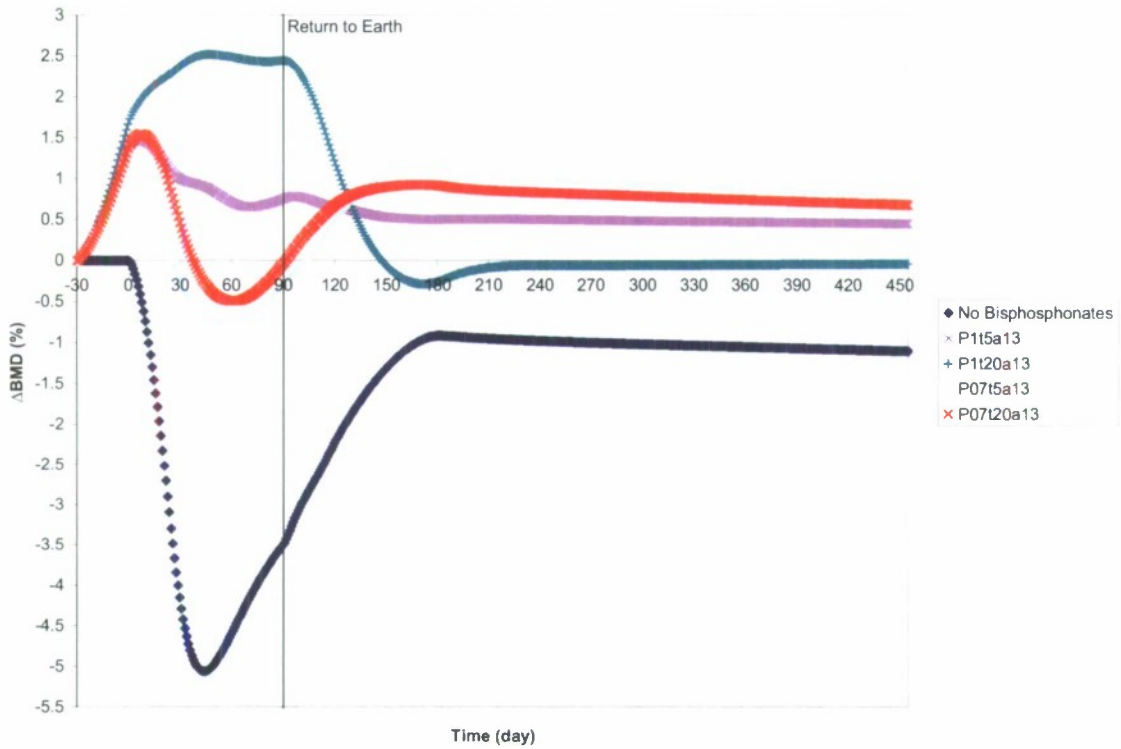


Figure C15. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

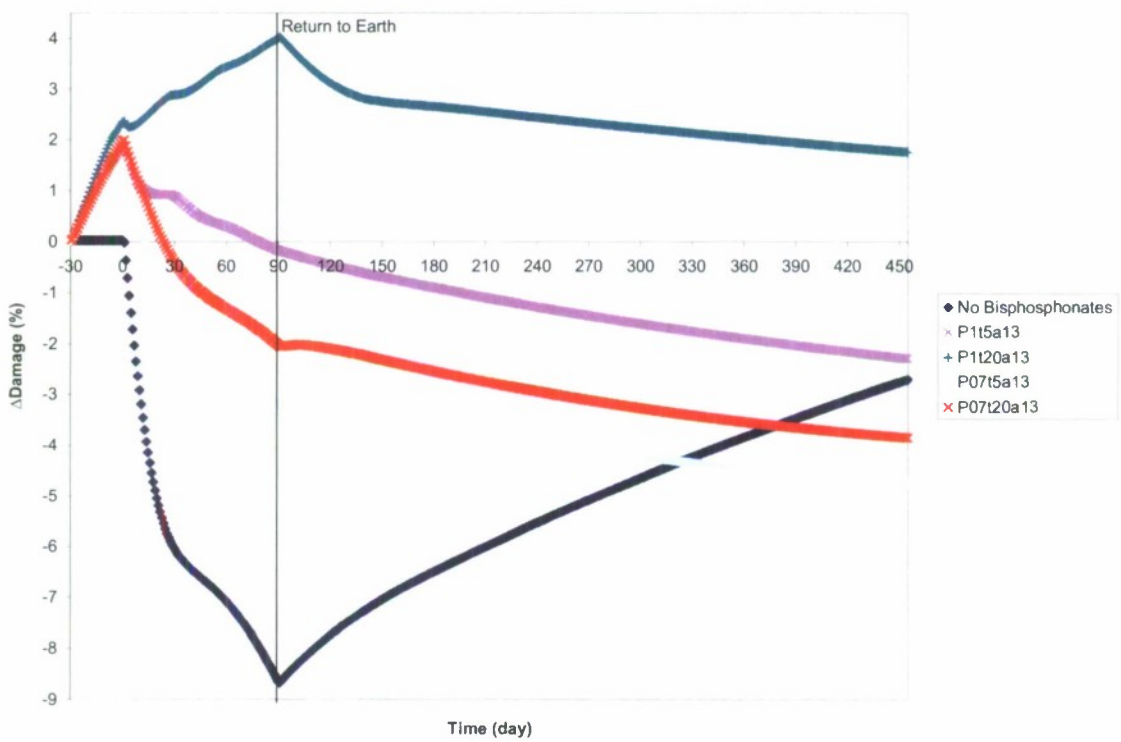


Figure C16. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

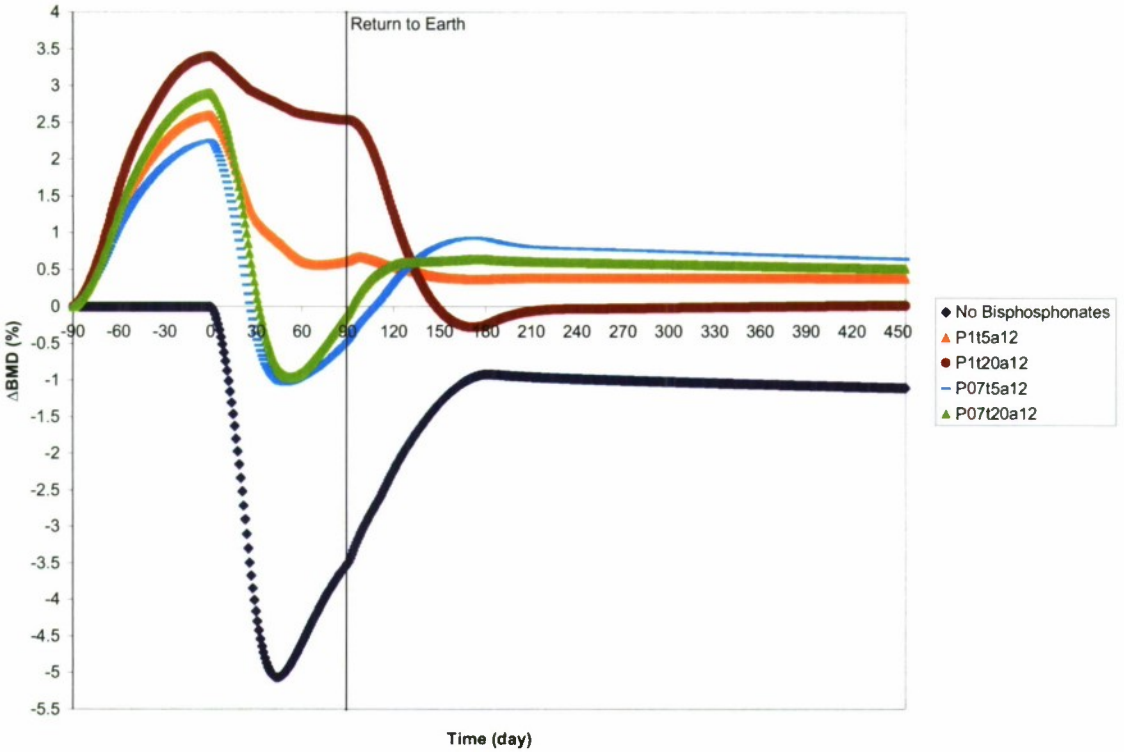


Figure C17. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

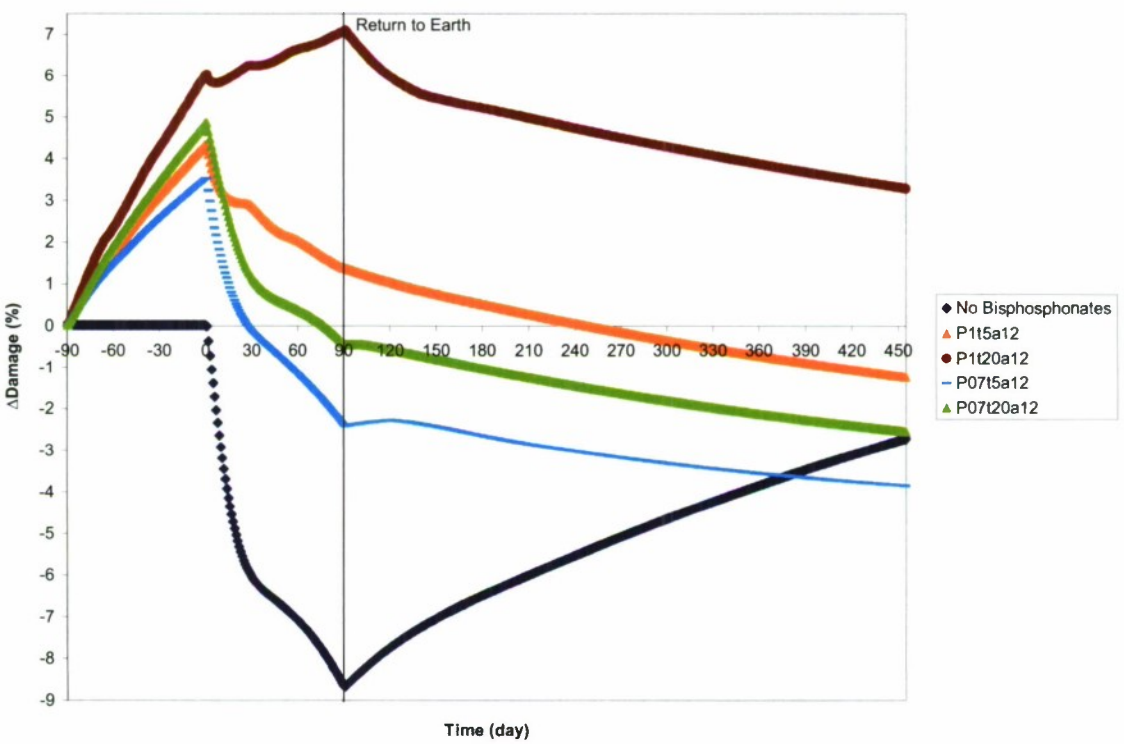


Figure C18. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

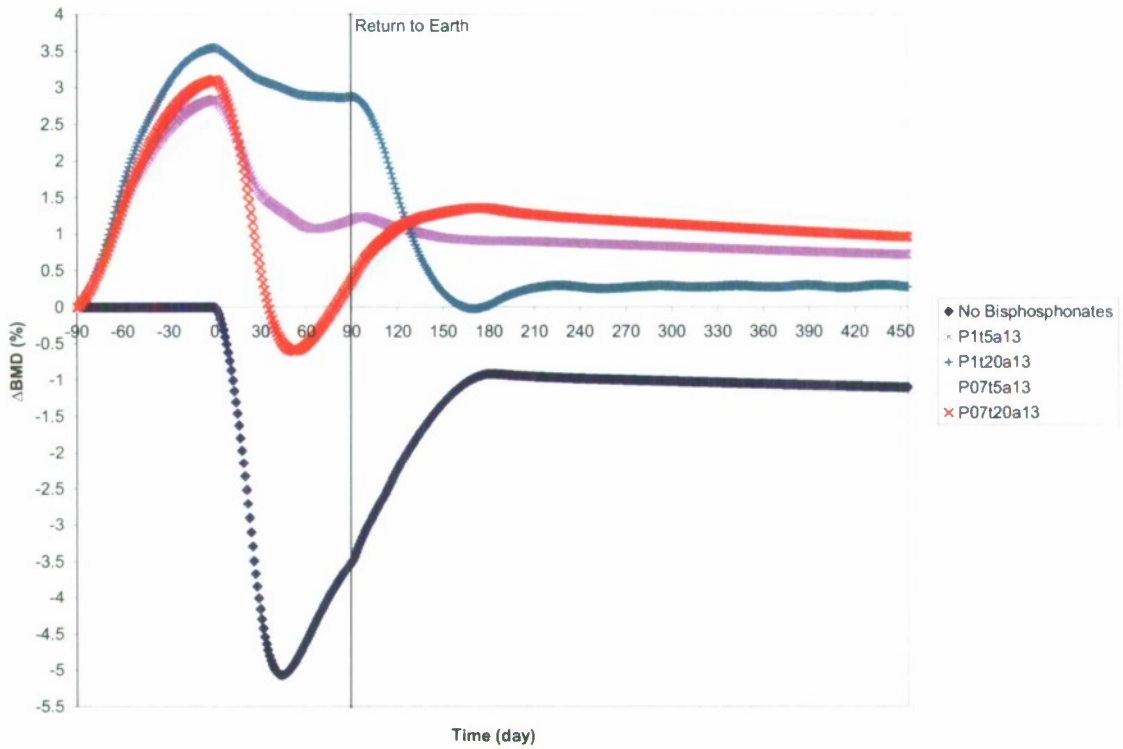


Figure C19. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

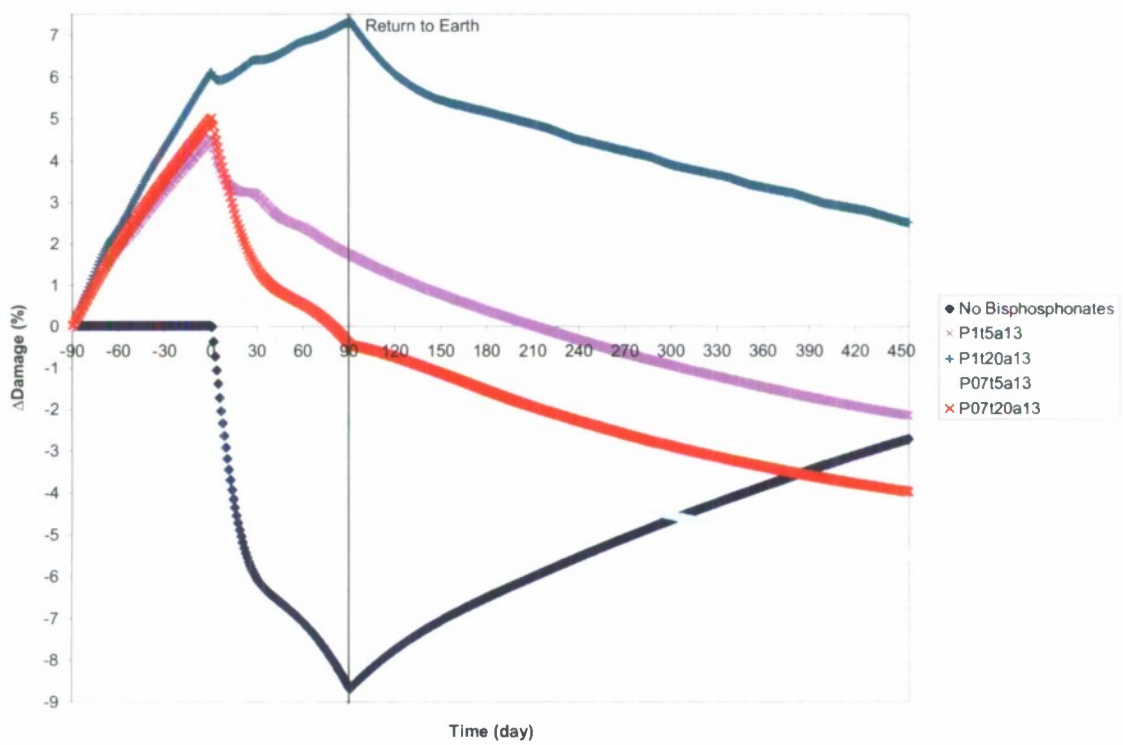


Figure C20. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

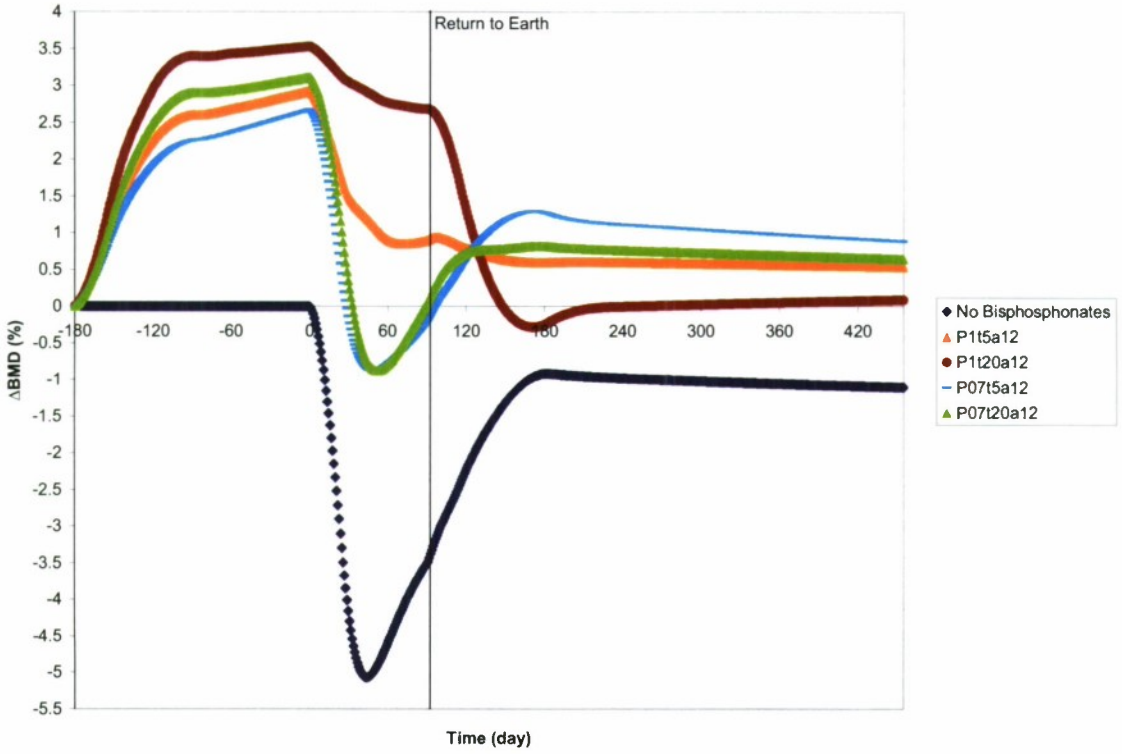


Figure C21. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

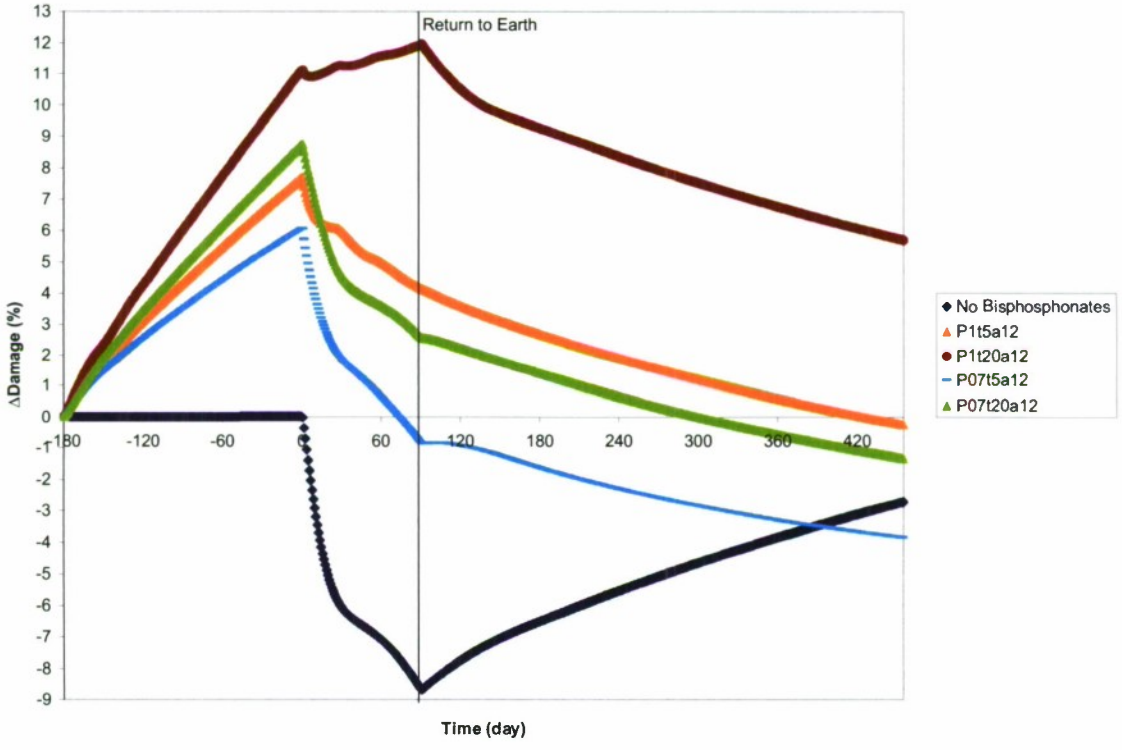


Figure C22. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

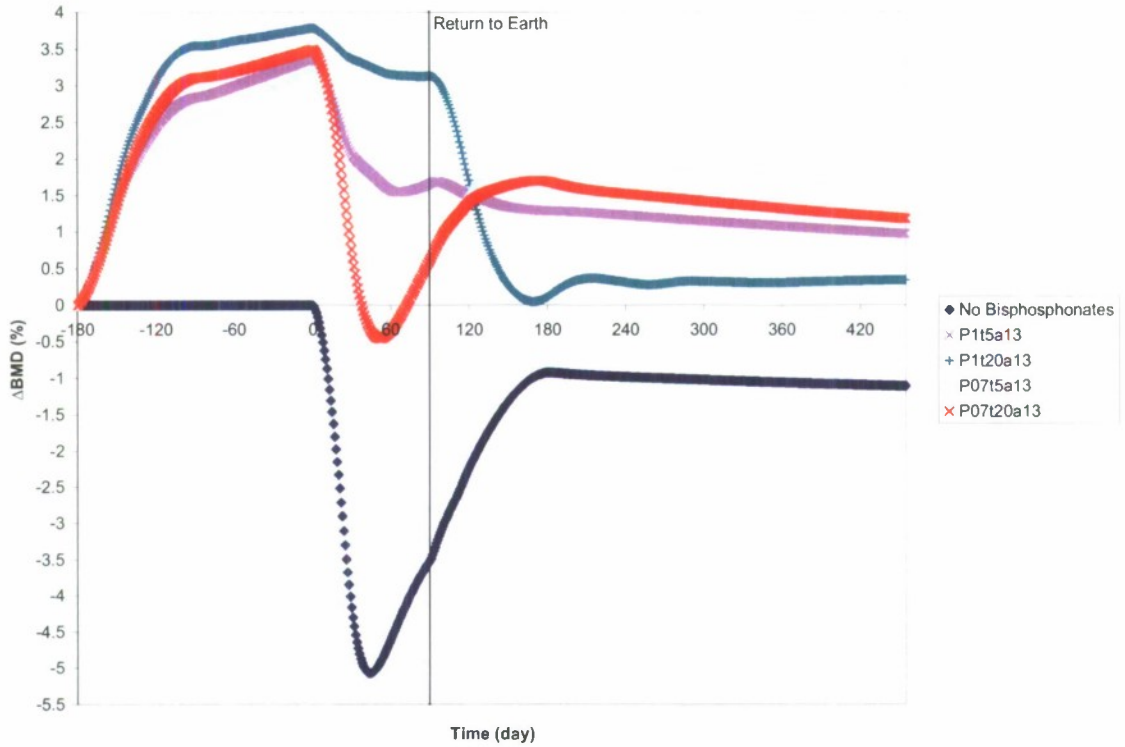


Figure C23. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 90-day spaceflight.

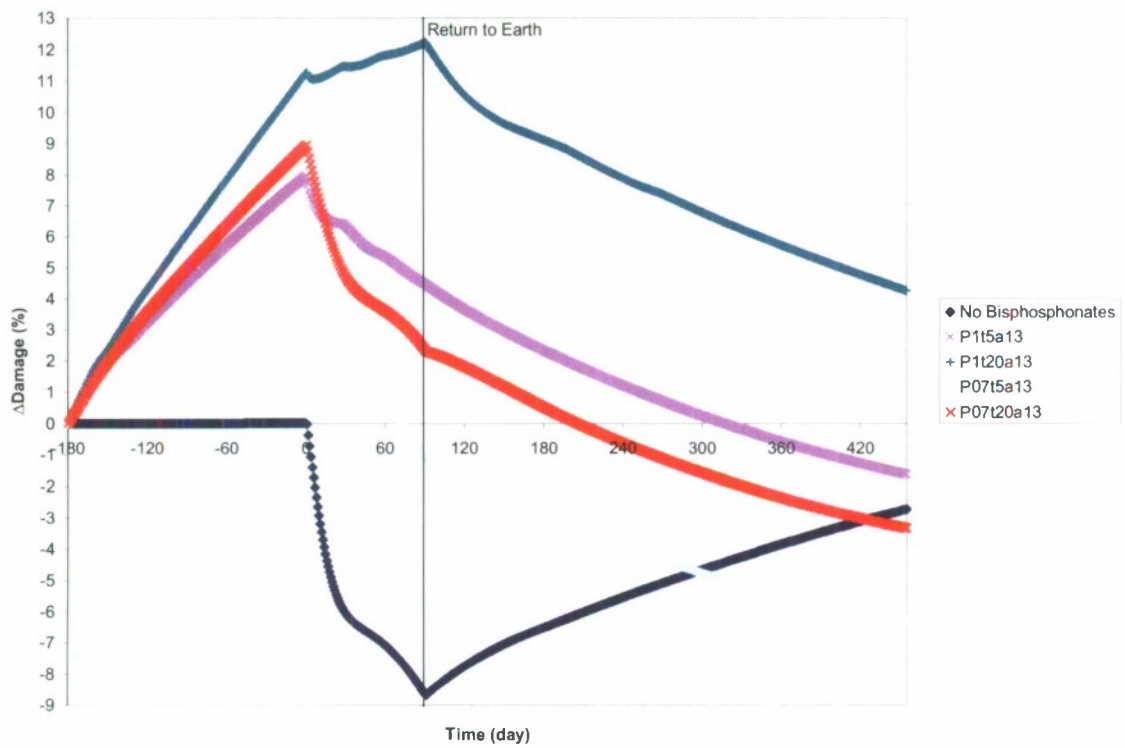


Figure C24. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 90-day spaceflight.

APPENDIX D: FIGURES (180-DAY SPACEFLIGHT)

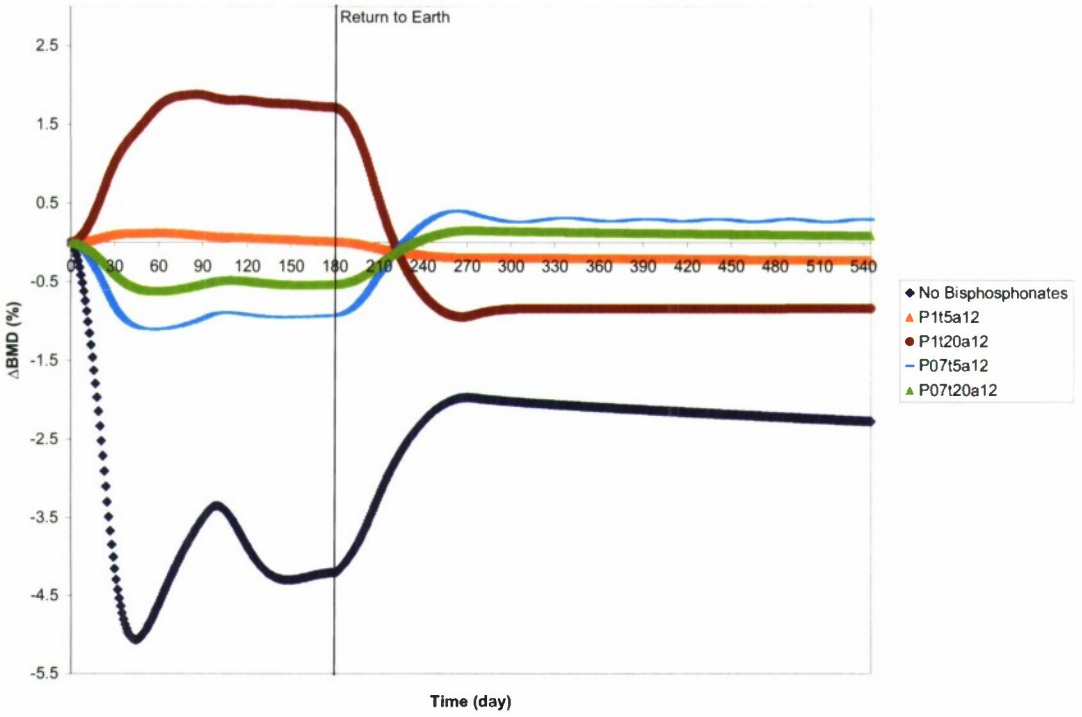


Figure D1. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight.

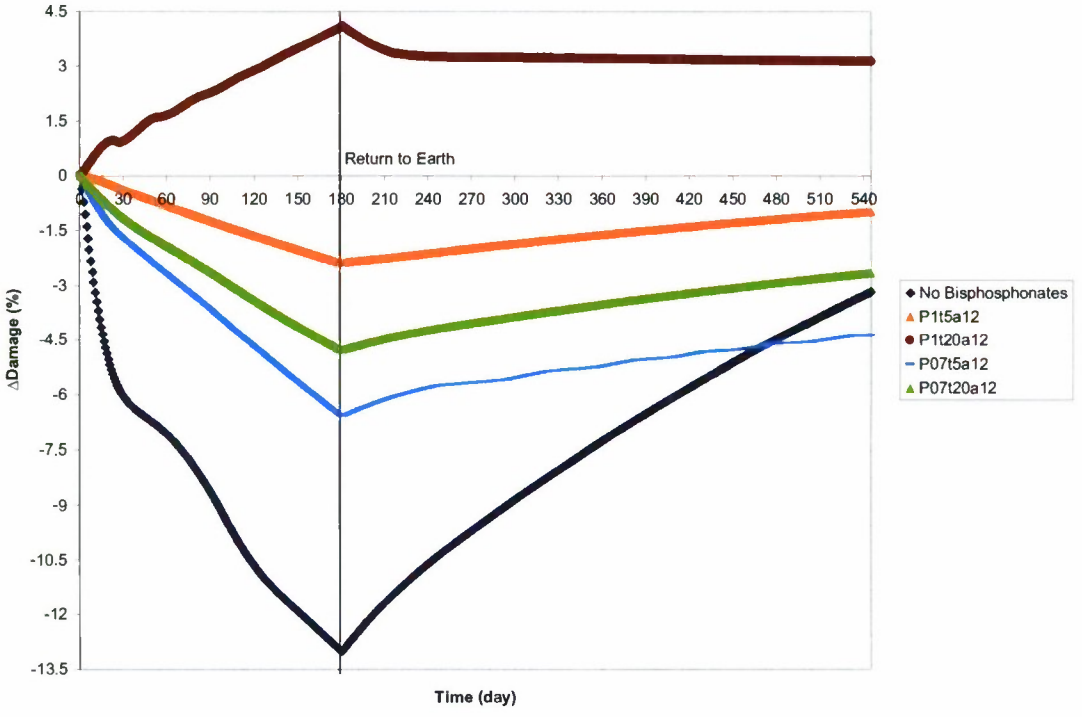


Figure D2. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

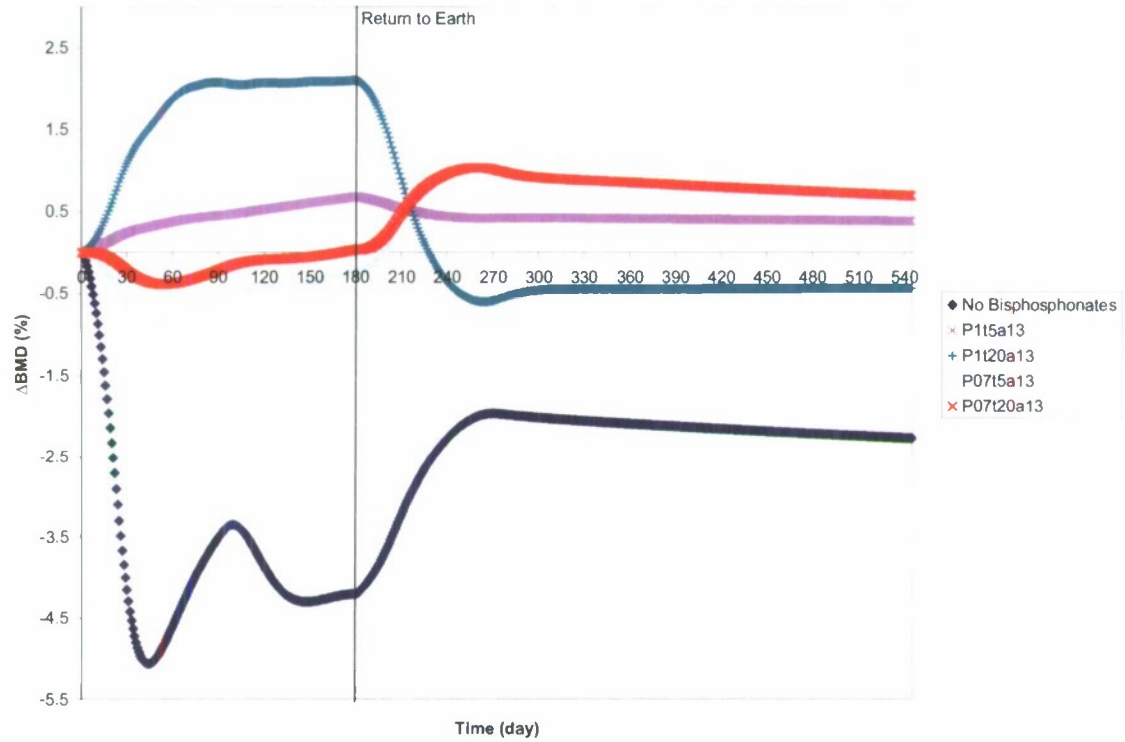


Figure D3. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 180-day spaceflight.

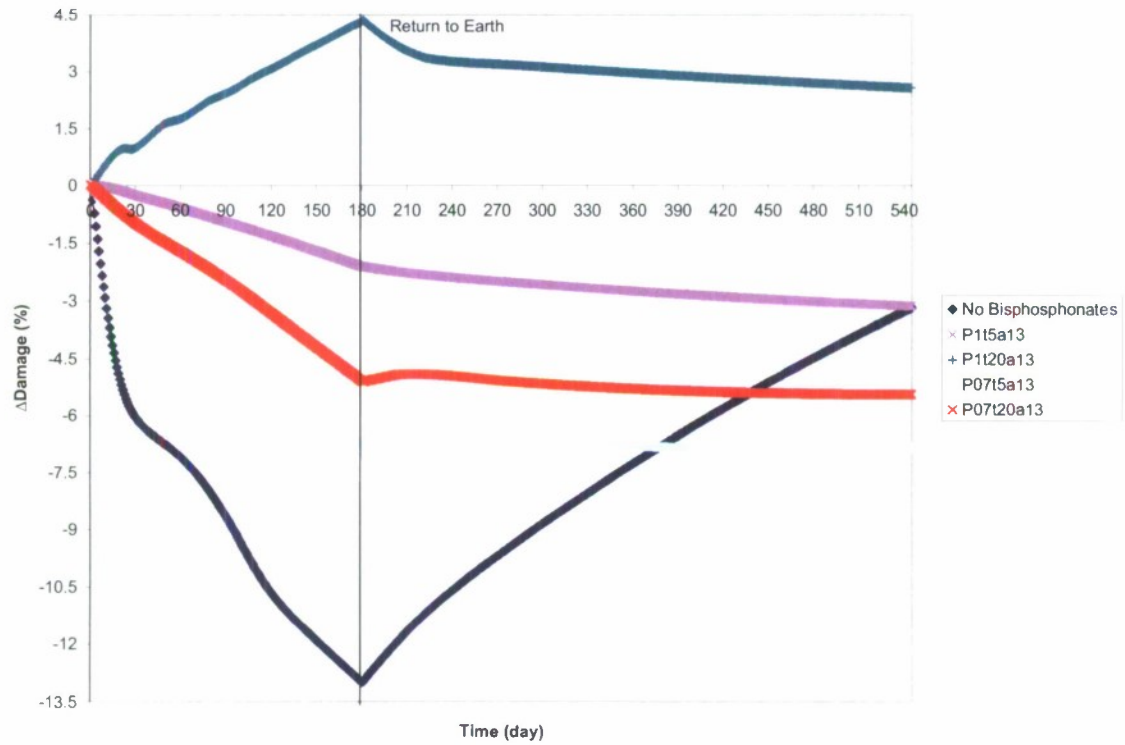


Figure D4. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

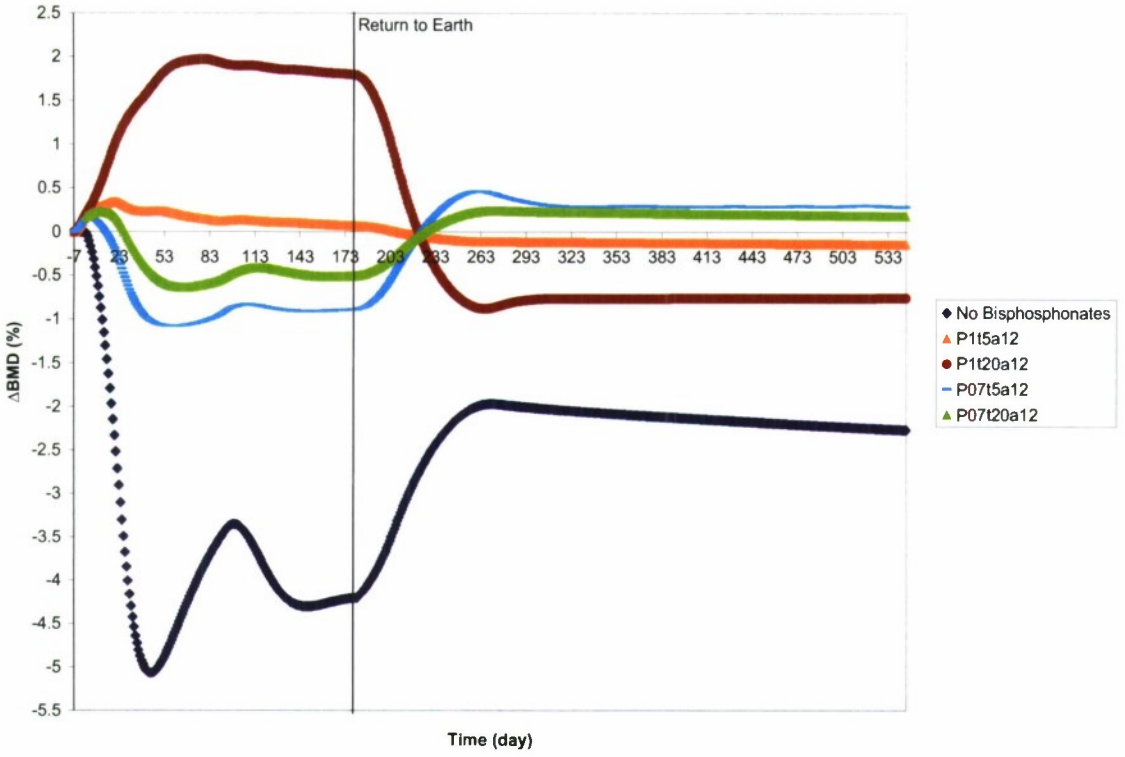


Figure D5. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

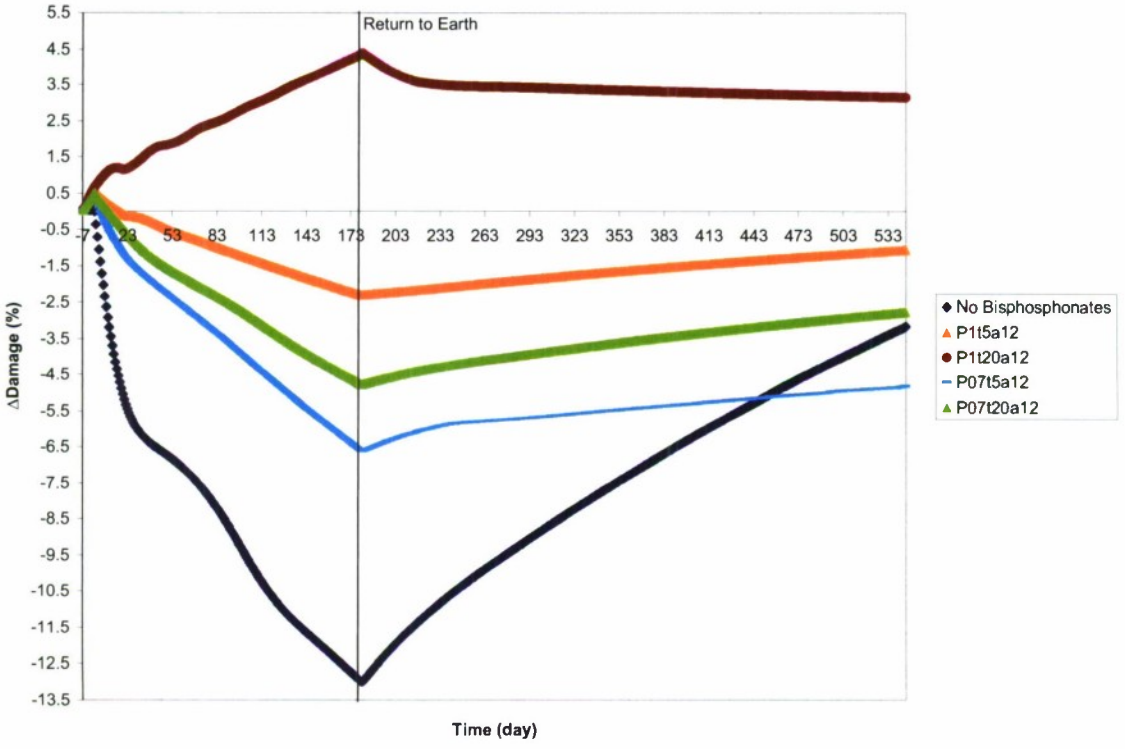


Figure D6. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

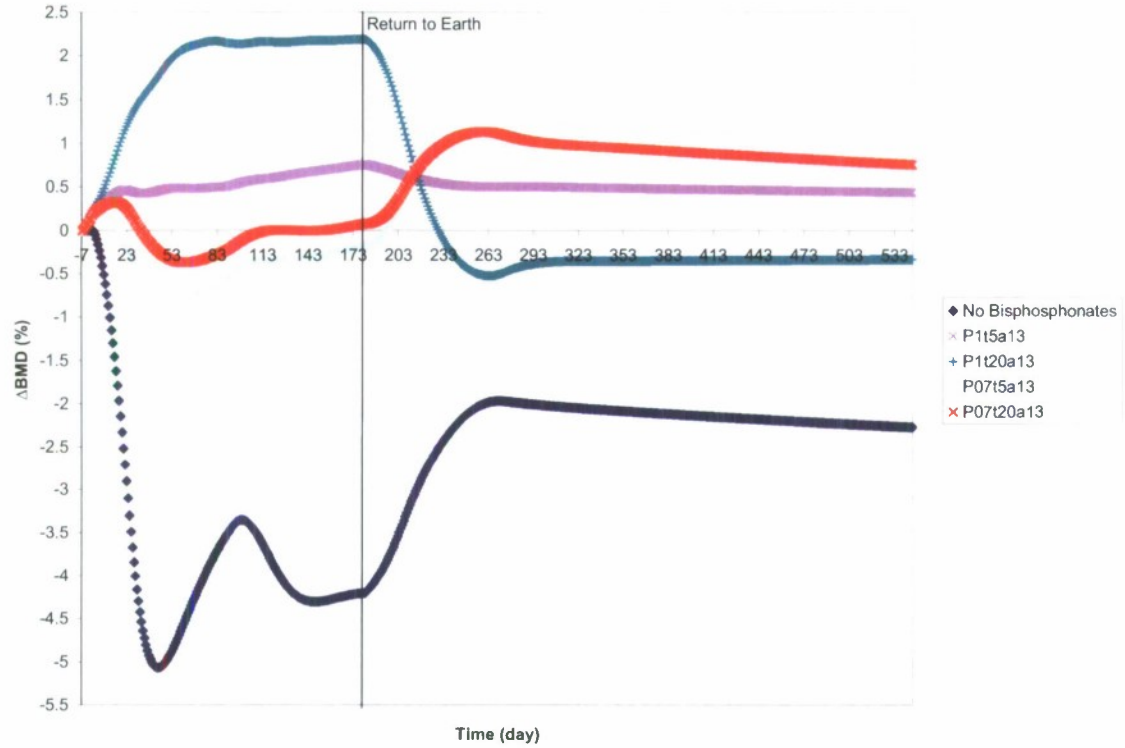


Figure D7. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

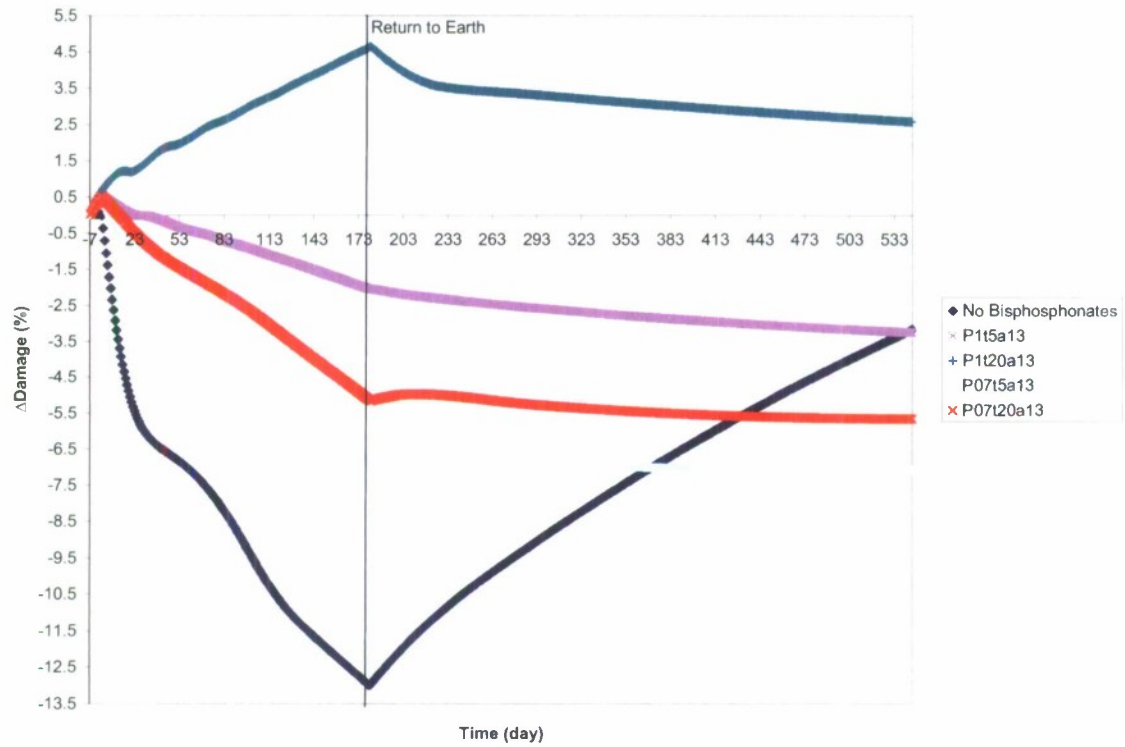


Figure D8. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

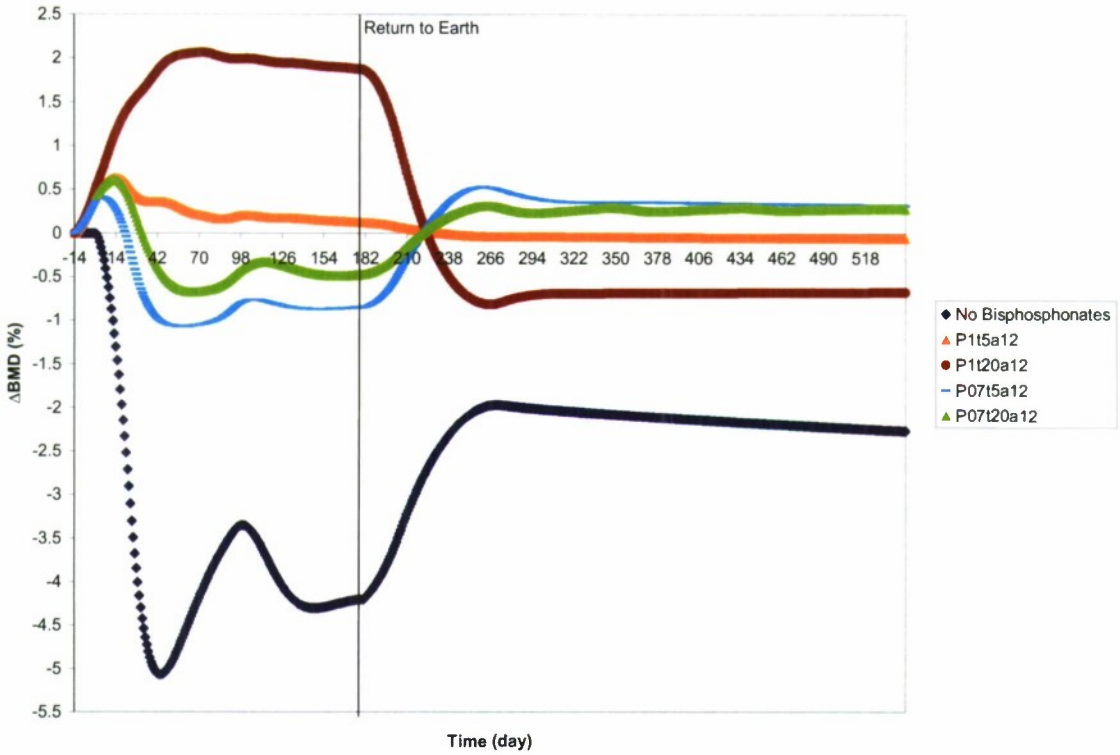


Figure D9. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

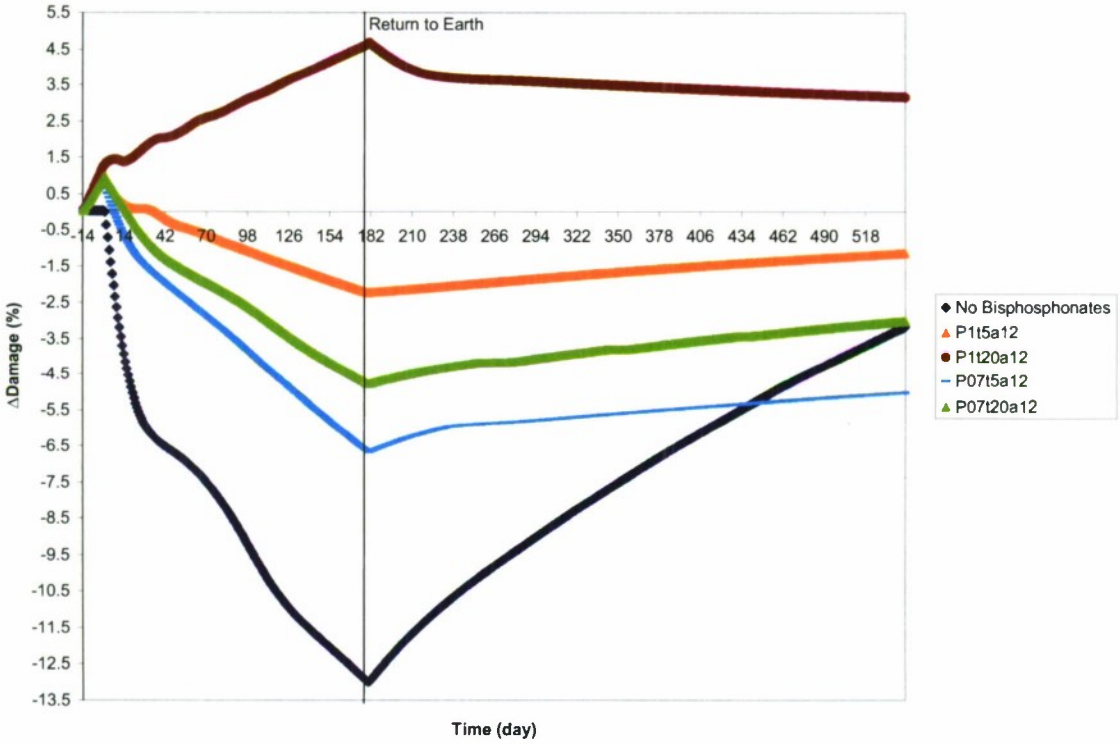


Figure D10. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

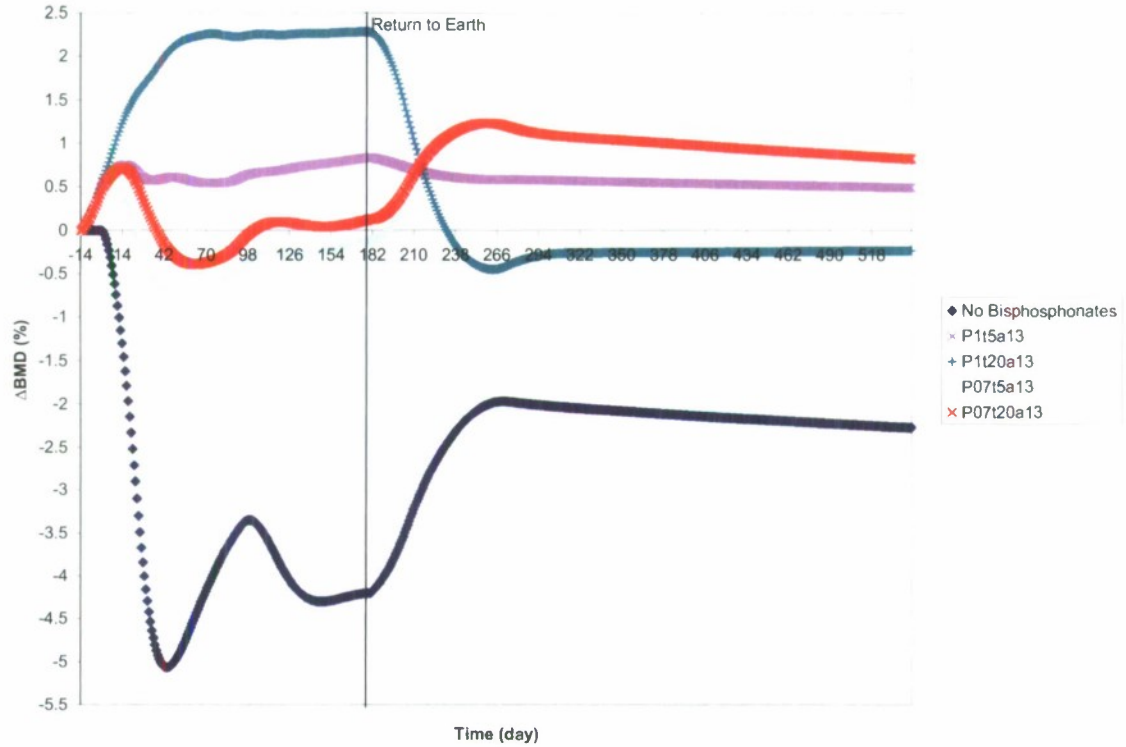


Figure D11. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

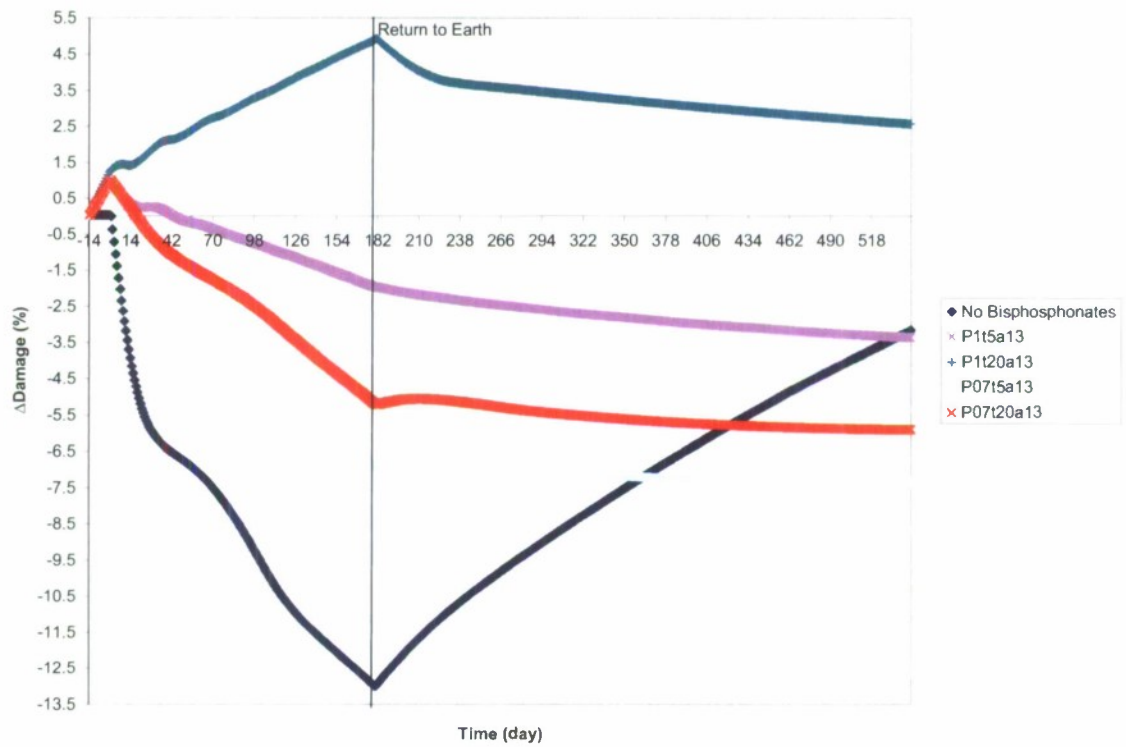


Figure D12. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

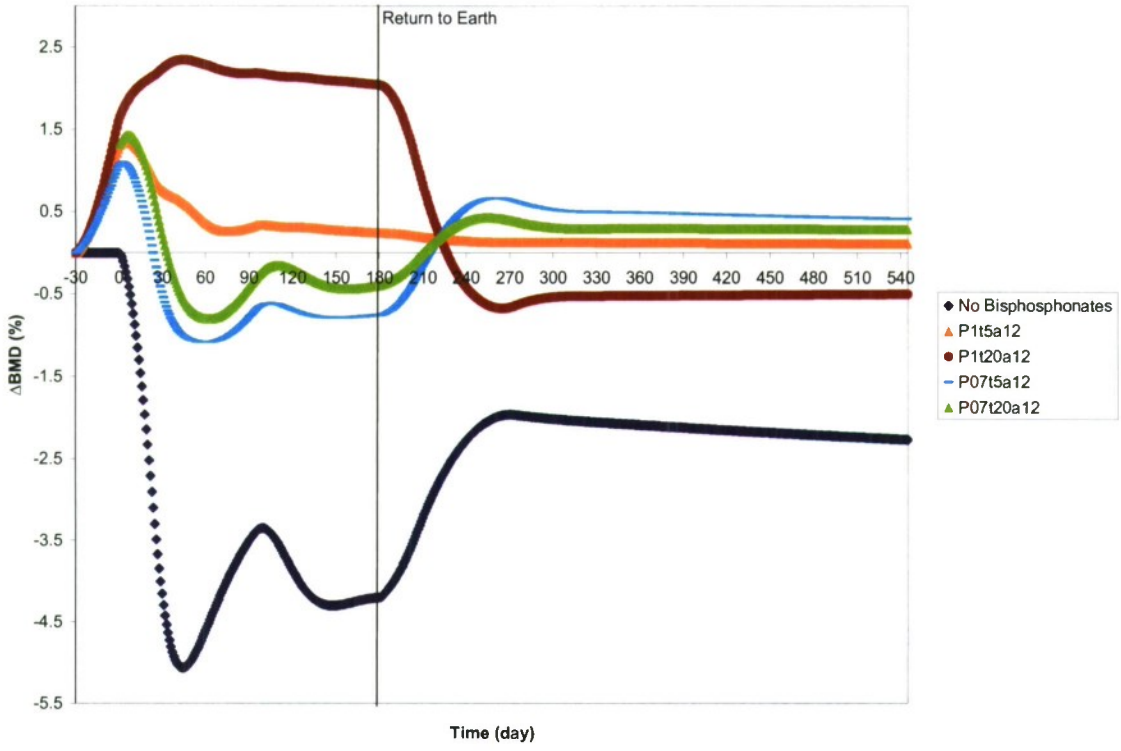


Figure D13. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

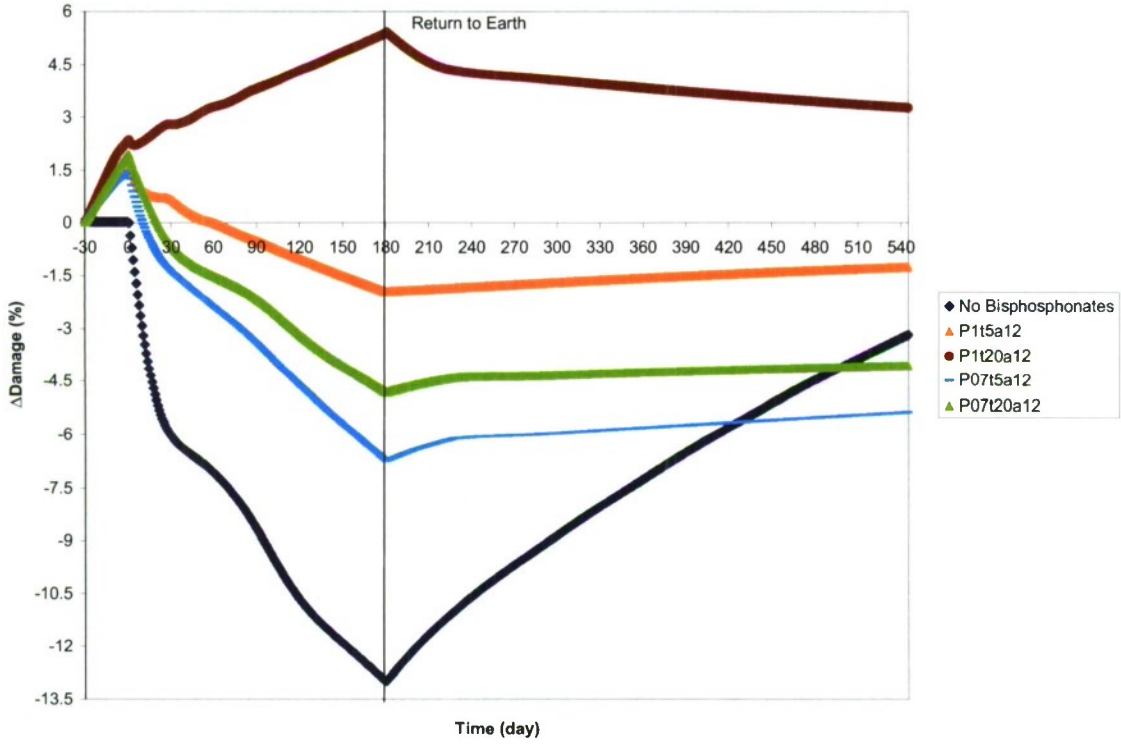


Figure D14. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

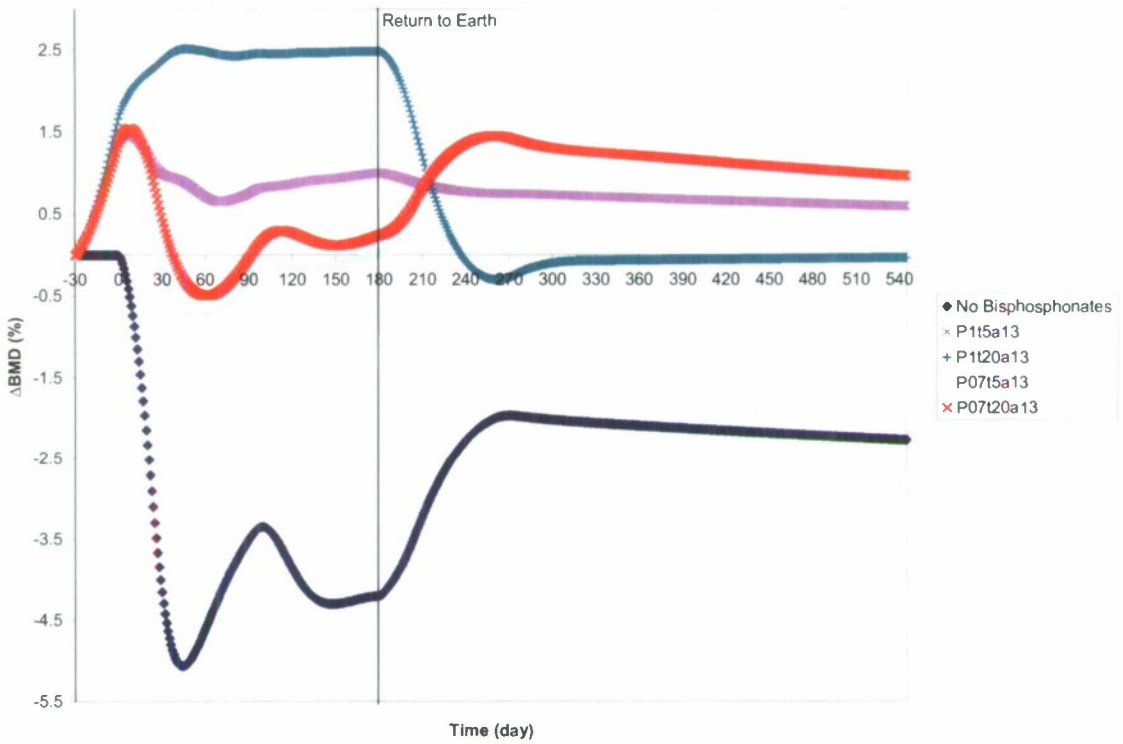


Figure D15. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

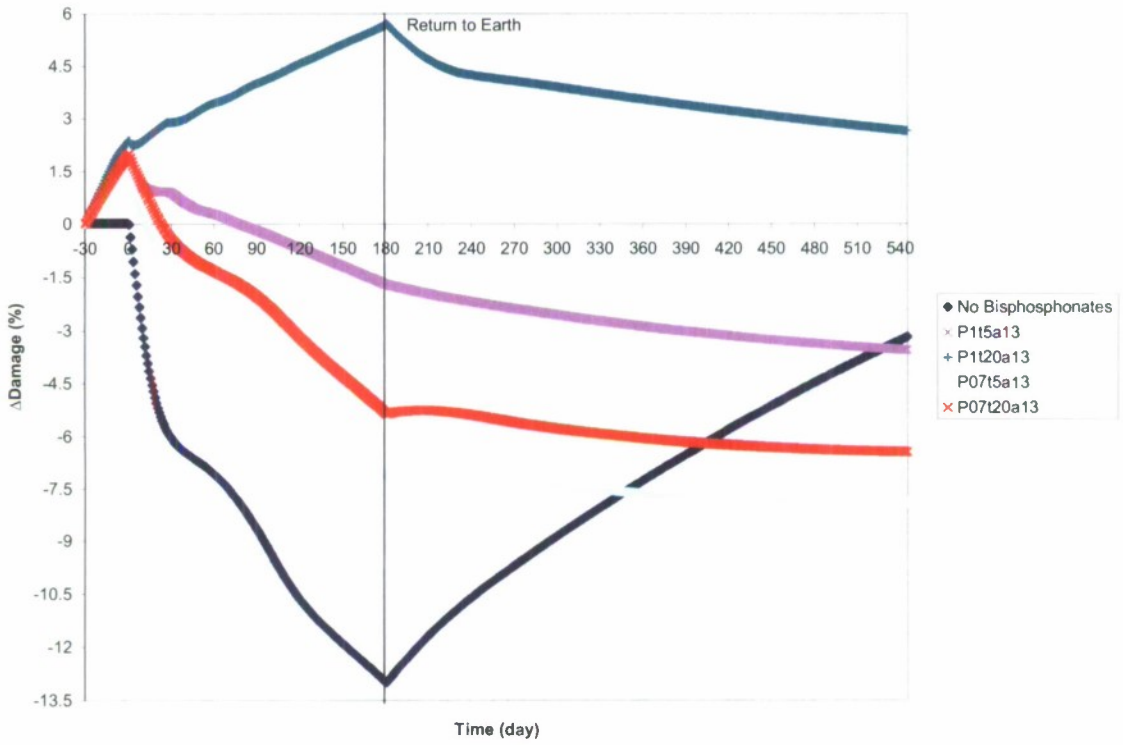


Figure D16. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

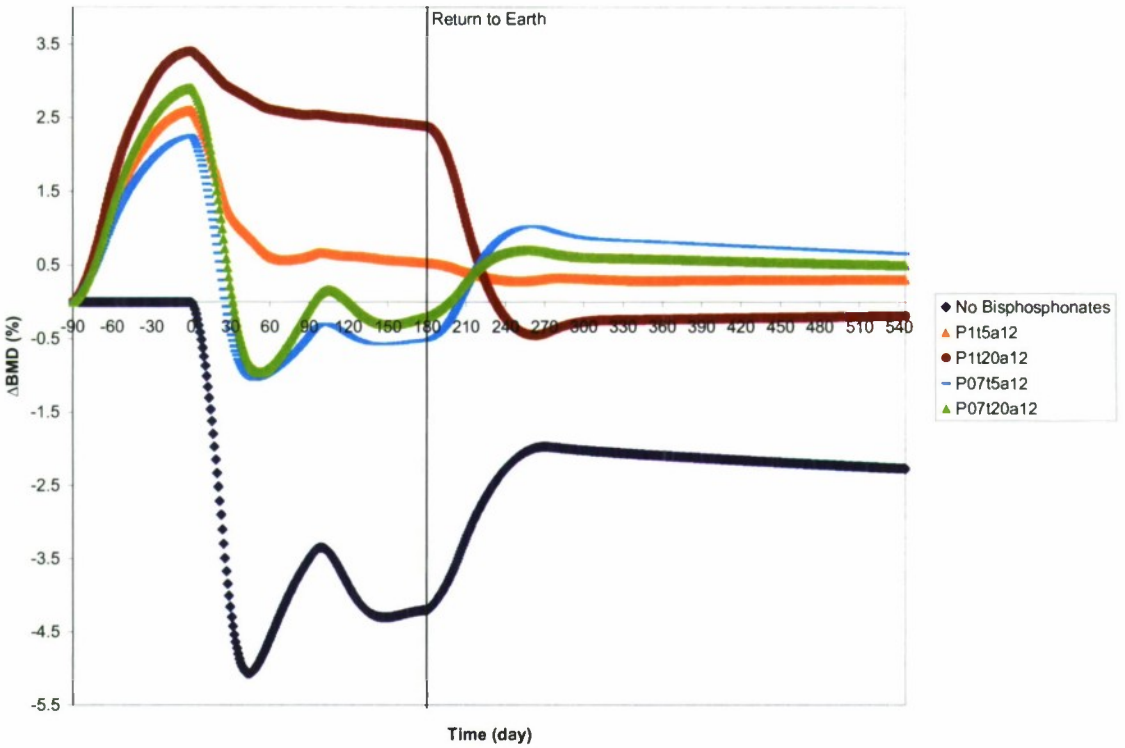


Figure D17. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

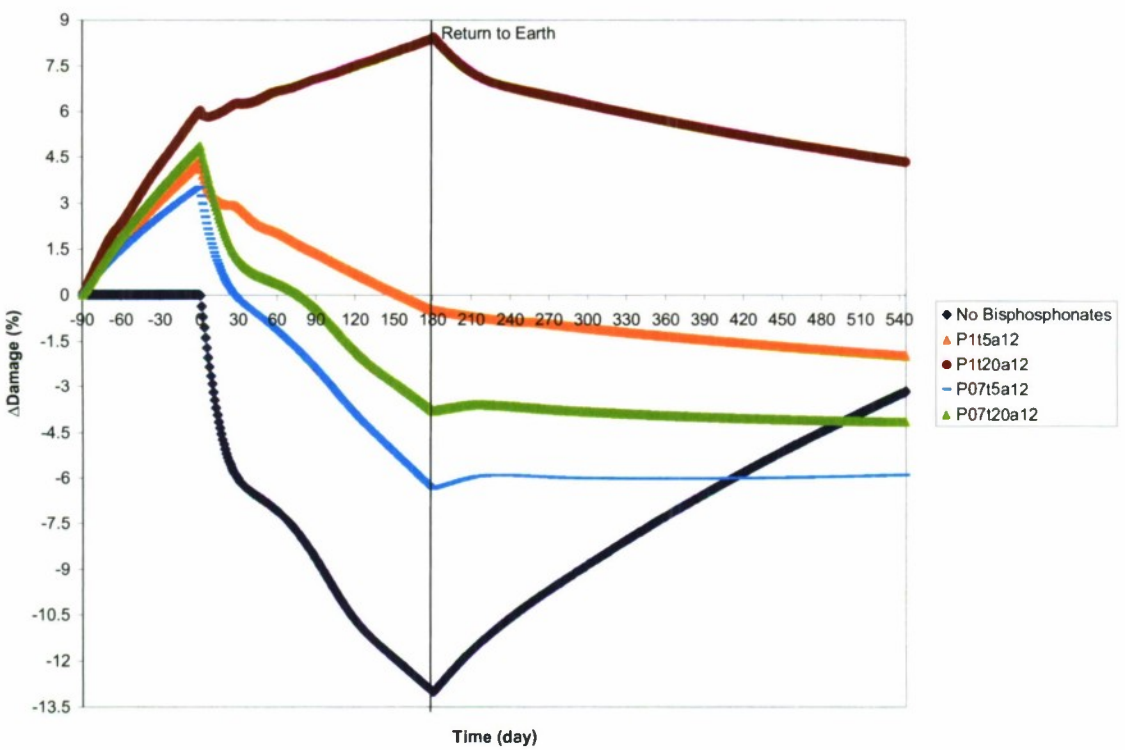


Figure D18. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

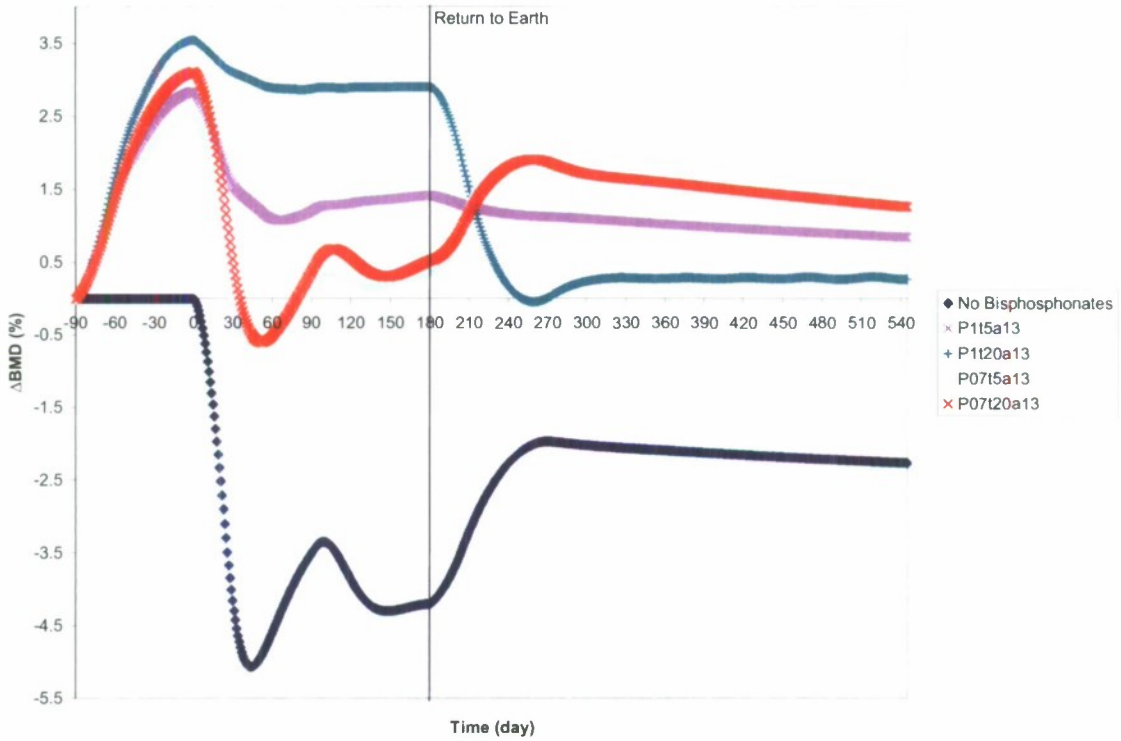


Figure D19. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

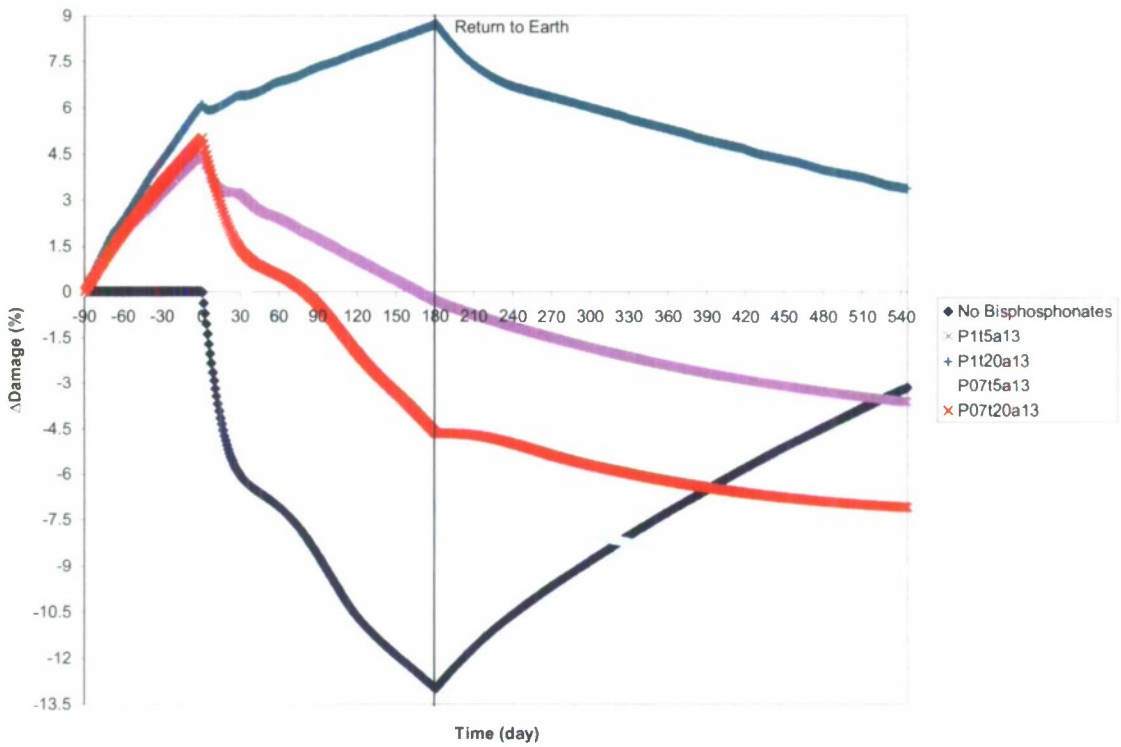


Figure D20. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

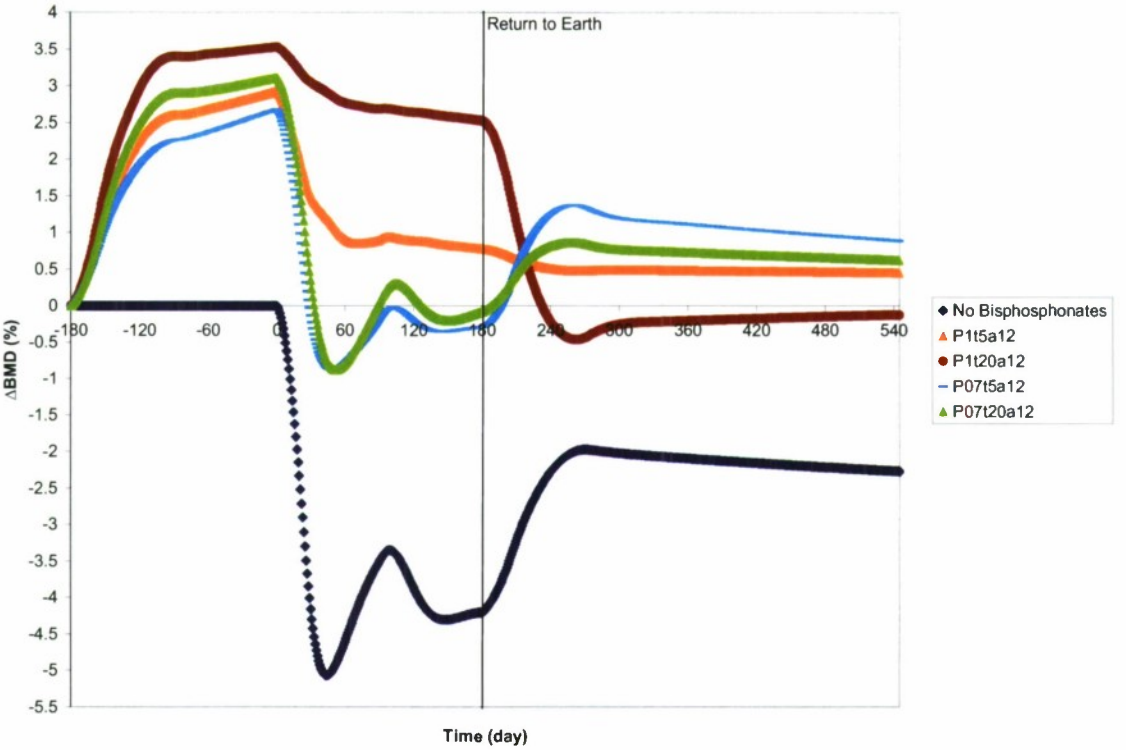


Figure D21. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

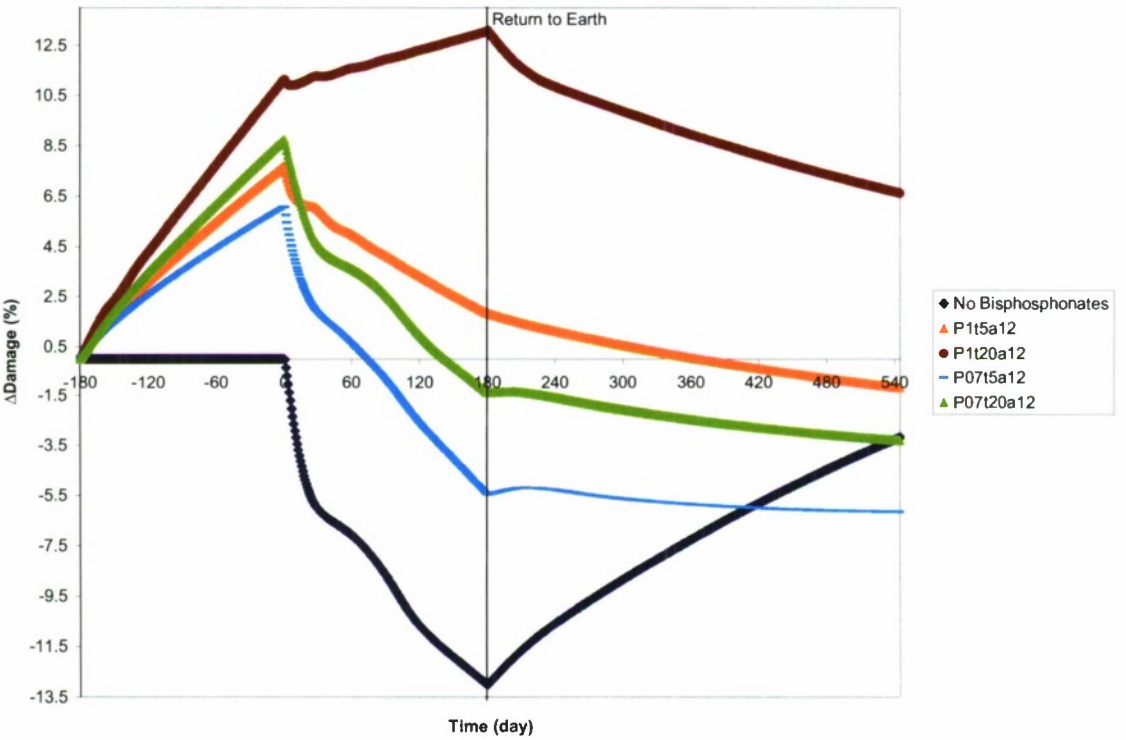


Figure D22. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

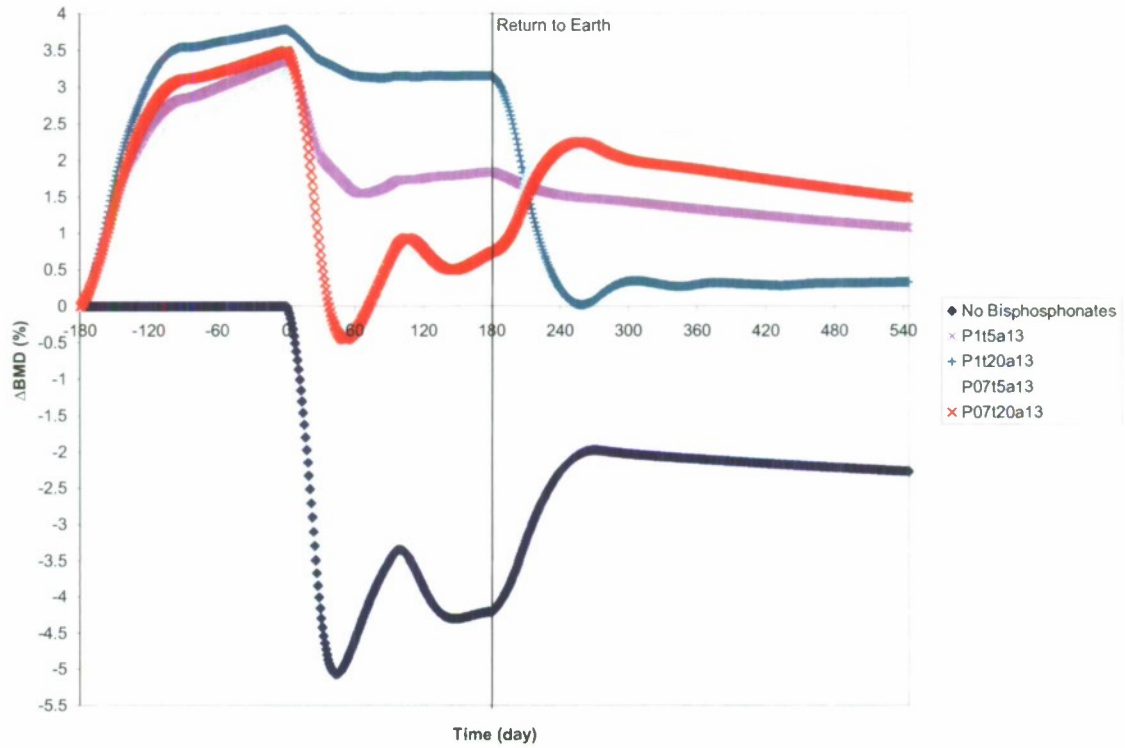


Figure D23. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 180-day spaceflight.

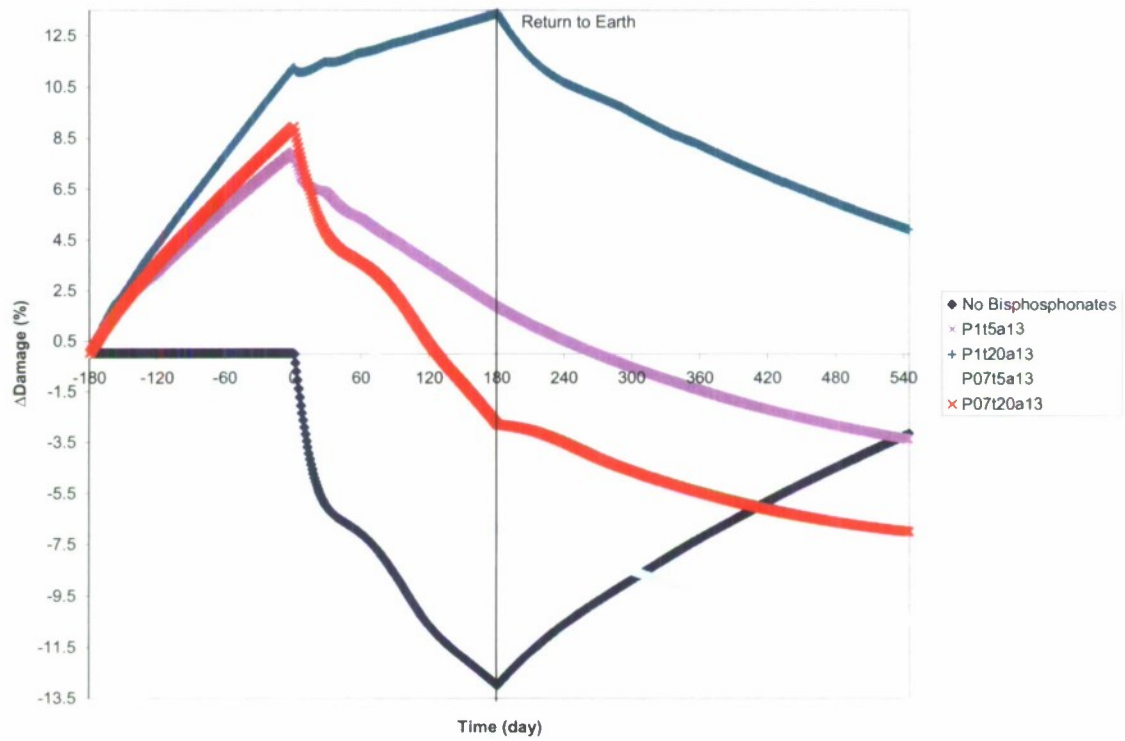


Figure D24. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 180-day spaceflight.

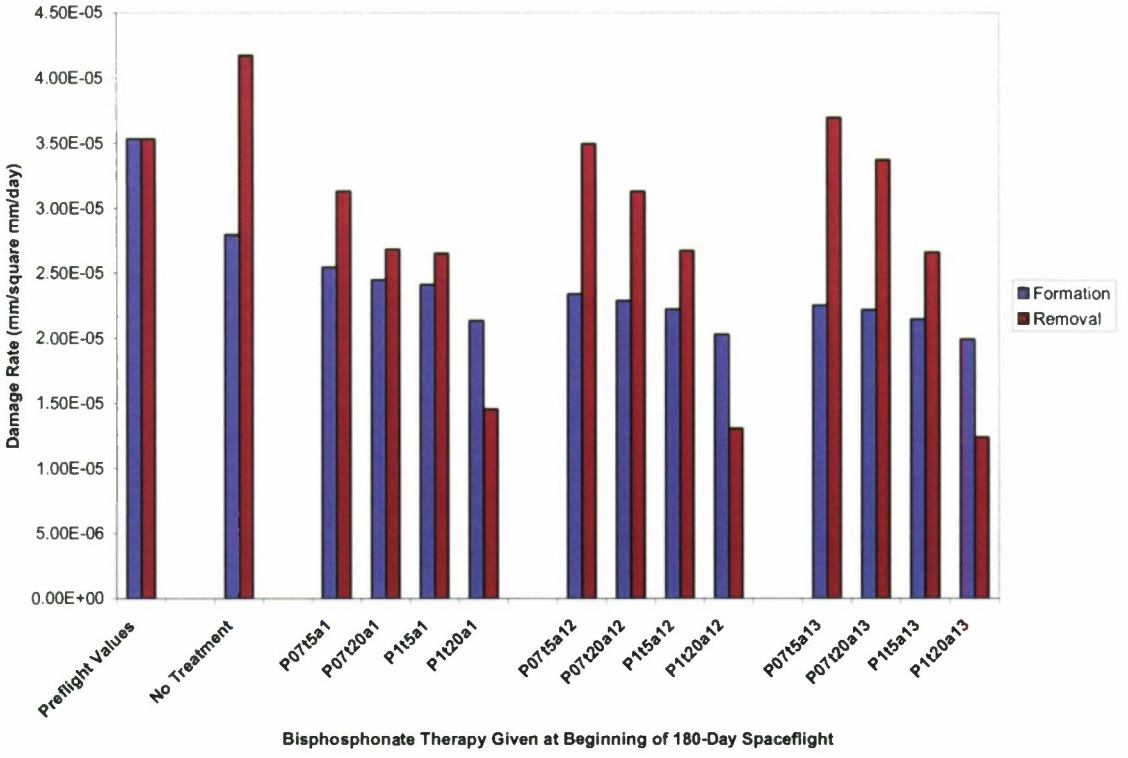


Figure D25. Predicted effects of 180-day spaceflight with and without bisphosphonate treatment on damage formation and removal rates. Note that these are end-of-flight values.

APPENDIX E: FIGURES (365-DAY SPACEFLIGHT)

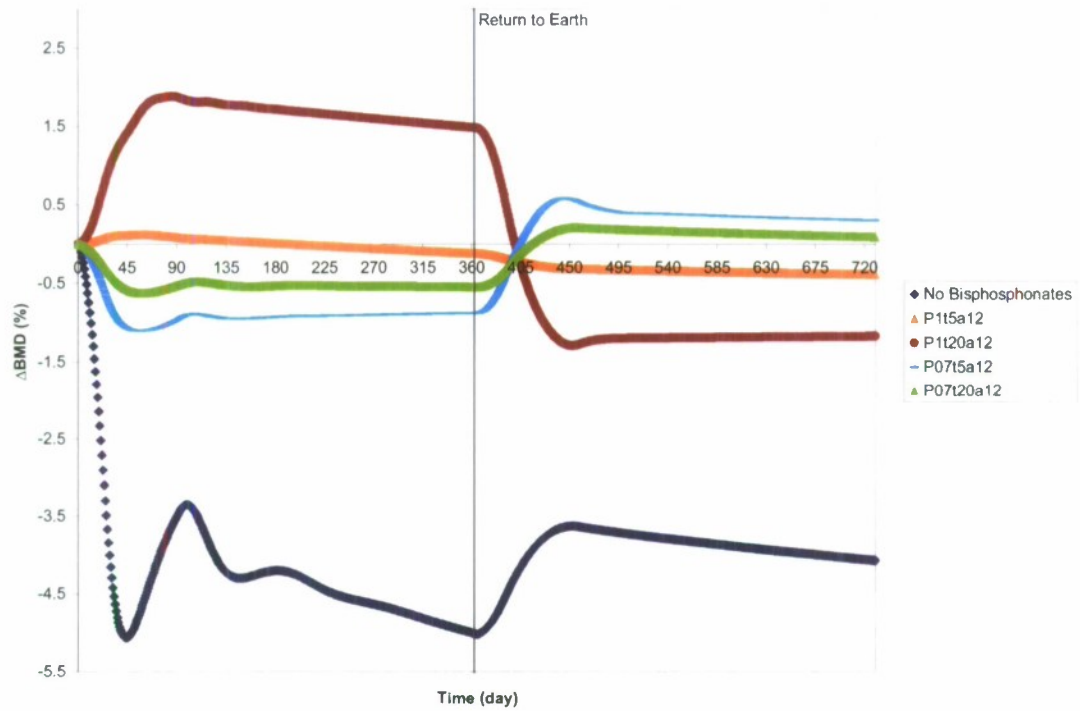


Figure E1. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 365-day spaceflight.

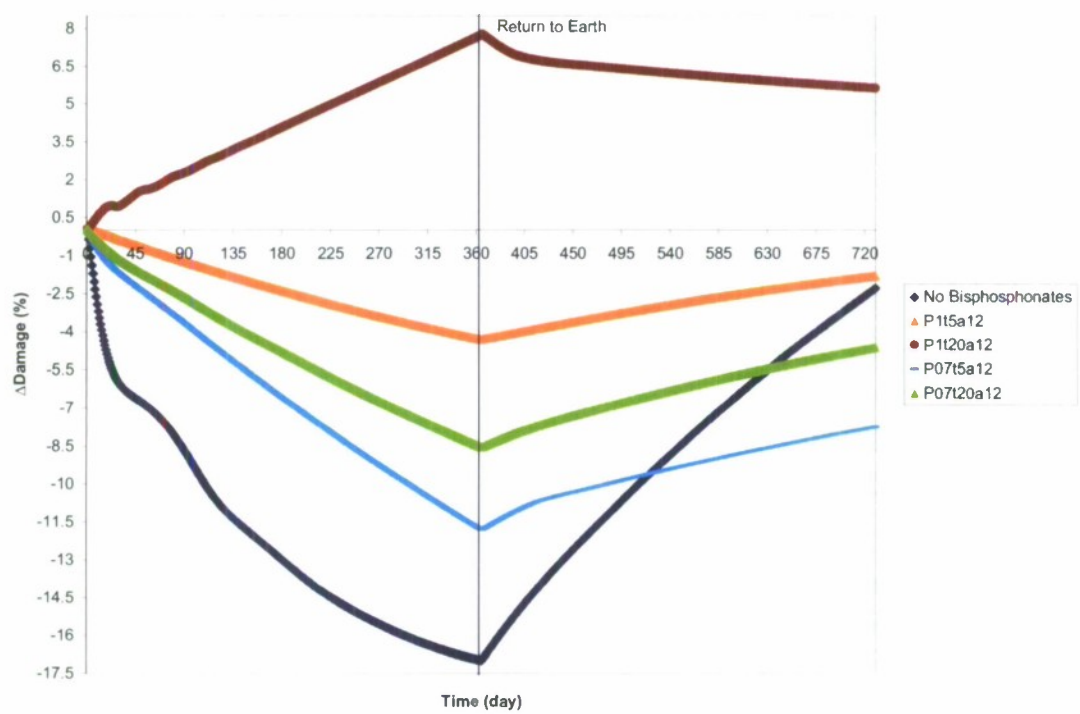


Figure E2. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

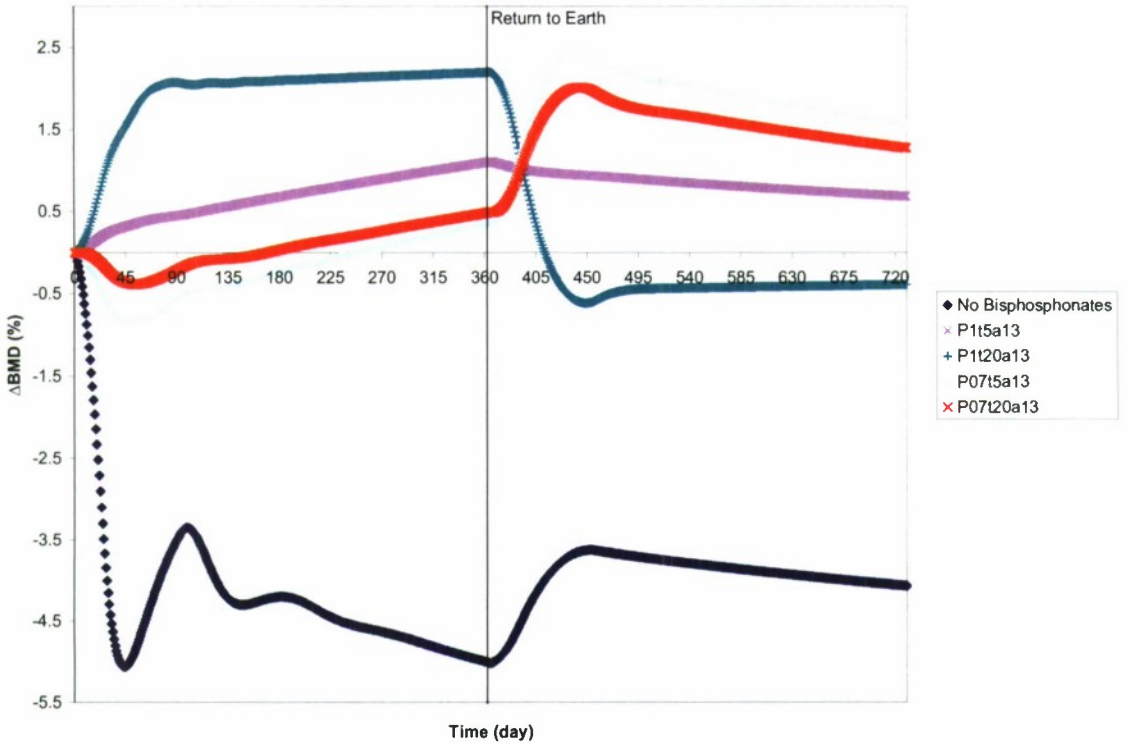


Figure E3. Predicted bisphosphonate effects on BMD and posttreatment return to Earth from 365-day spaceflight.

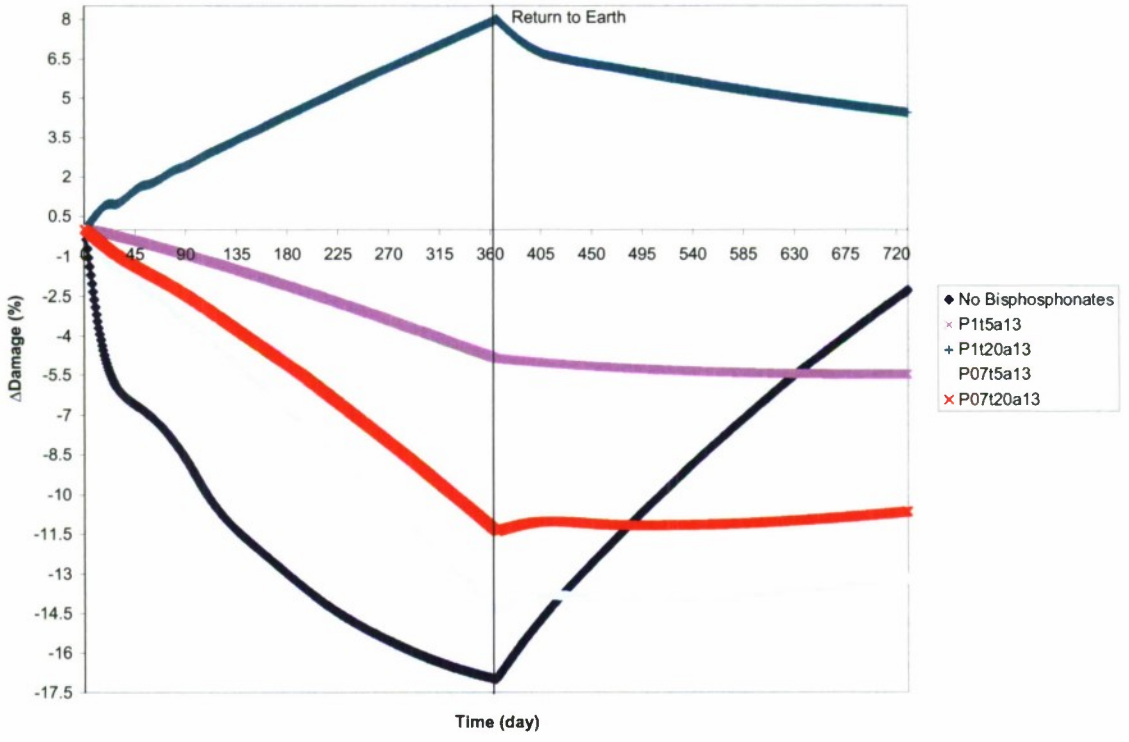


Figure E4. Predicted bisphosphonate effects on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

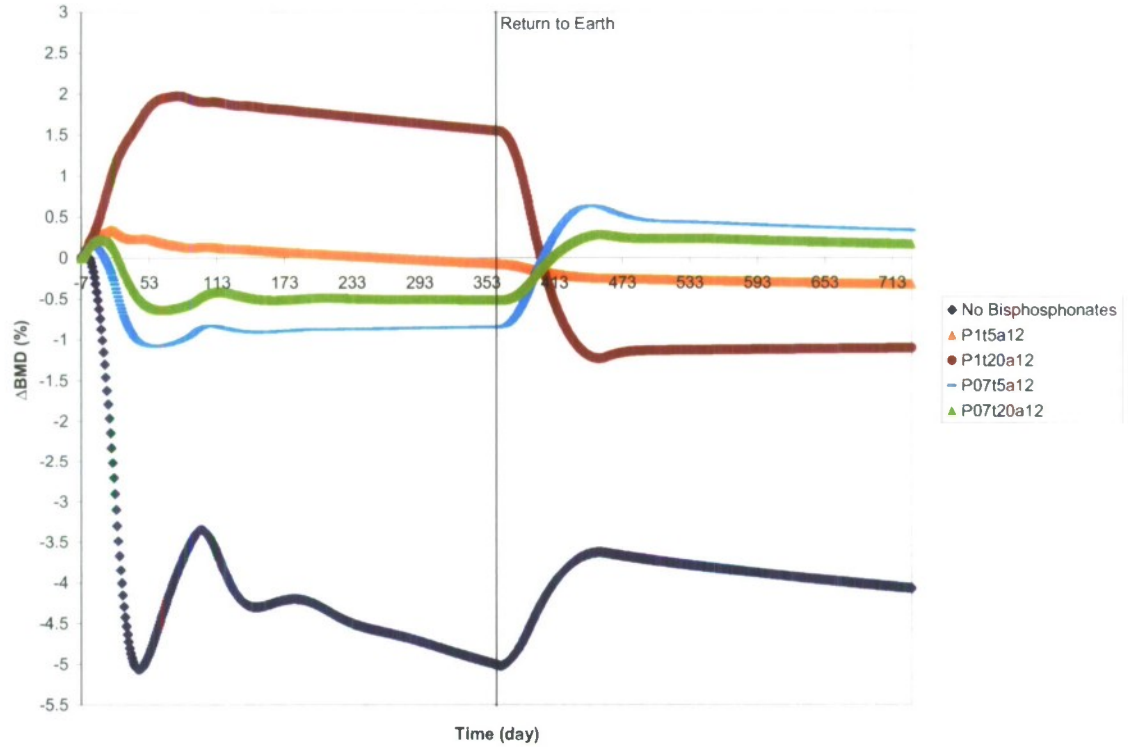


Figure E5. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

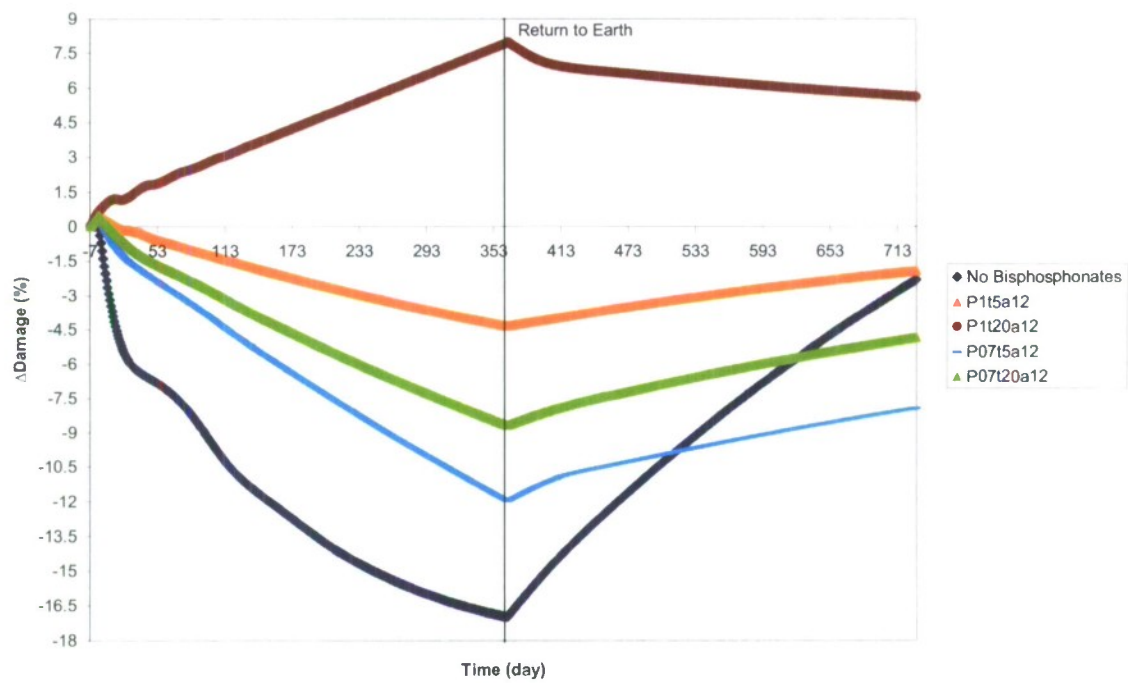


Figure E6. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

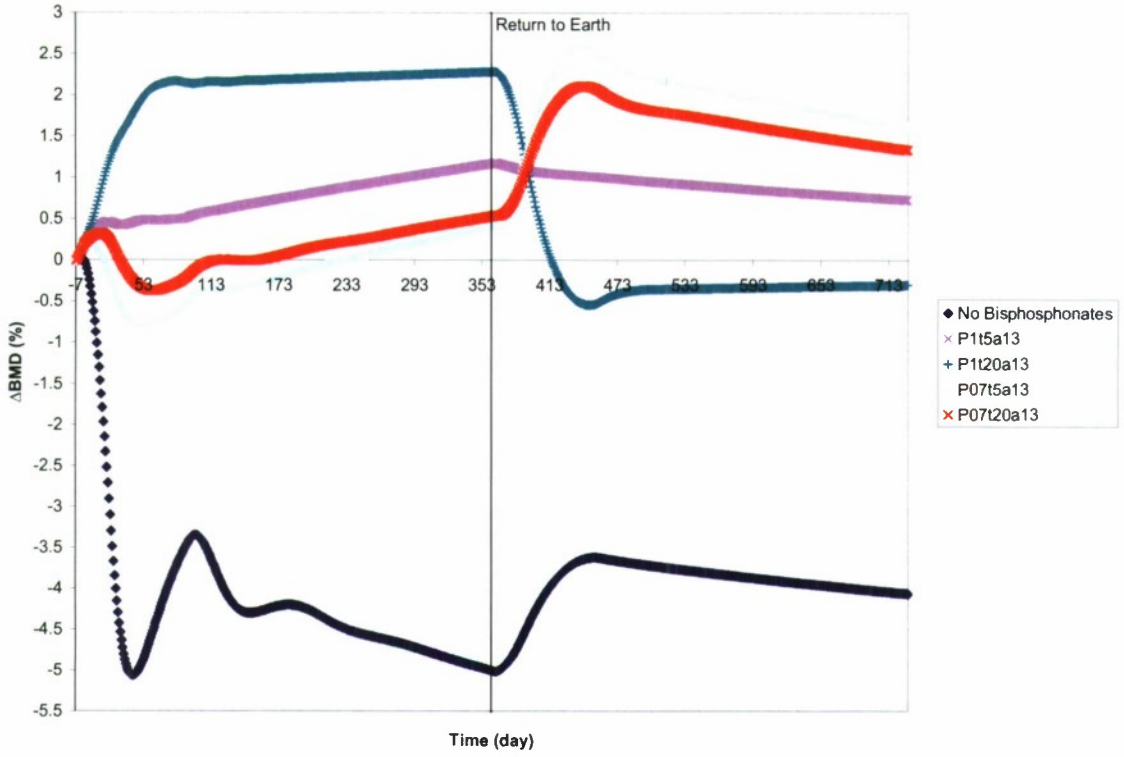


Figure E7. Predicted bisphosphonate effects beginning 7 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

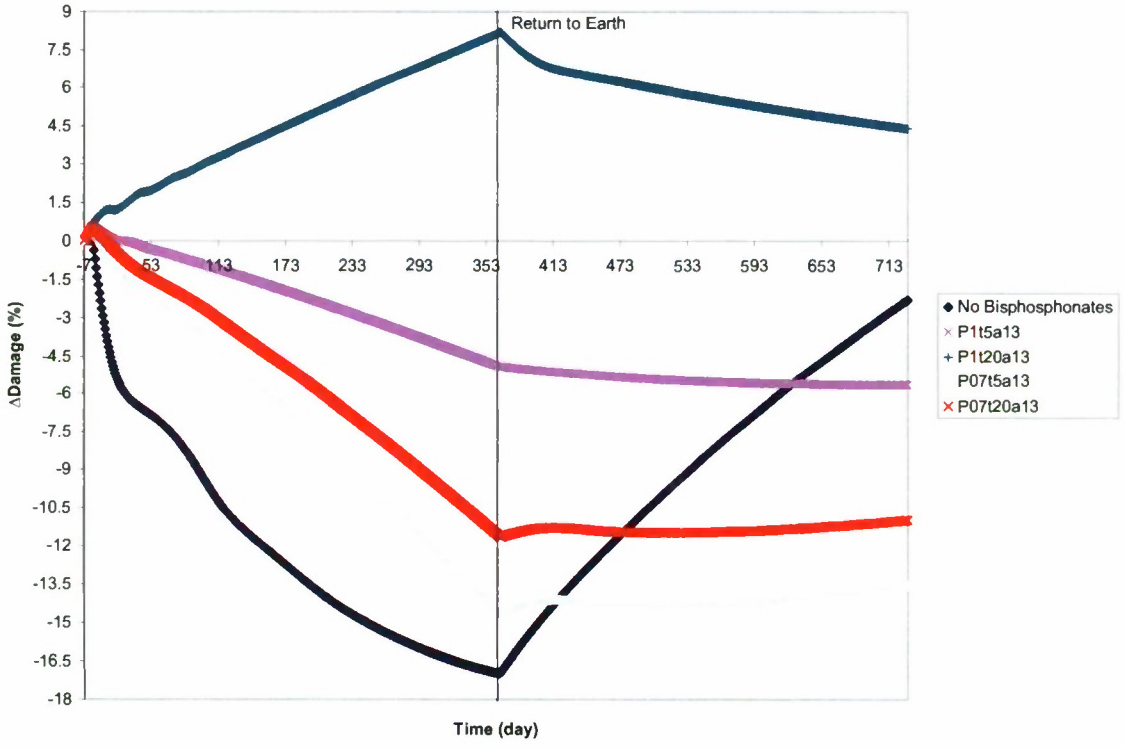


Figure E8. Predicted bisphosphonate effects beginning 7 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

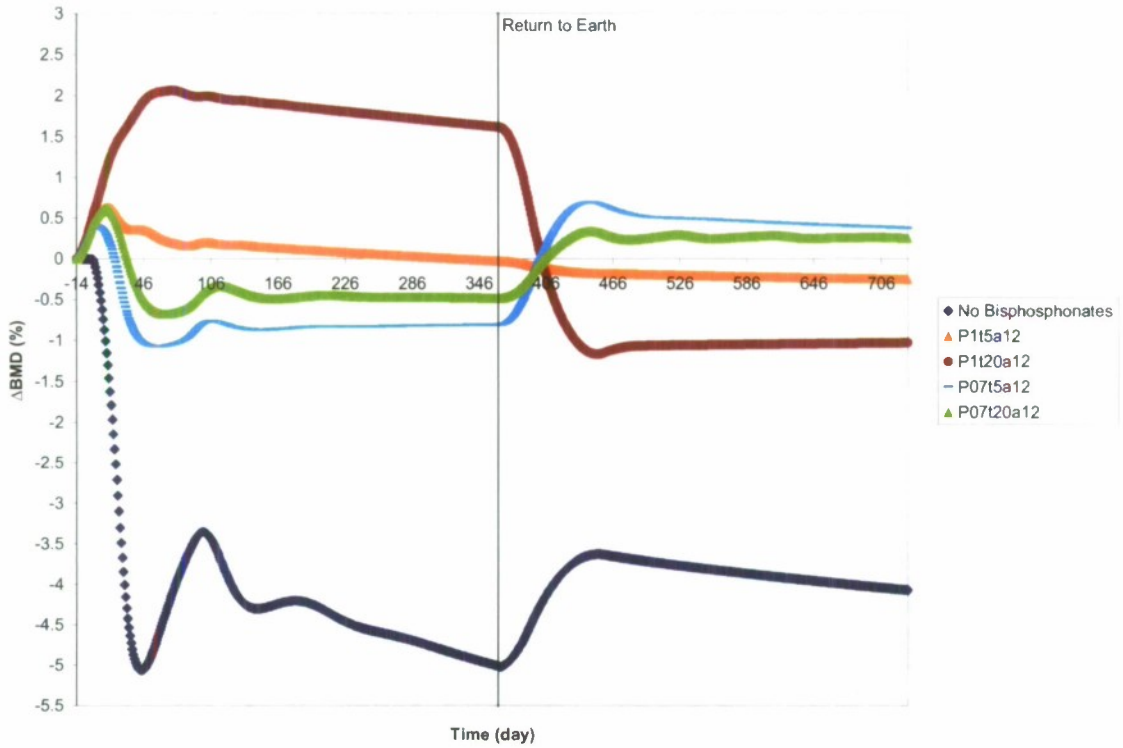


Figure E9. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

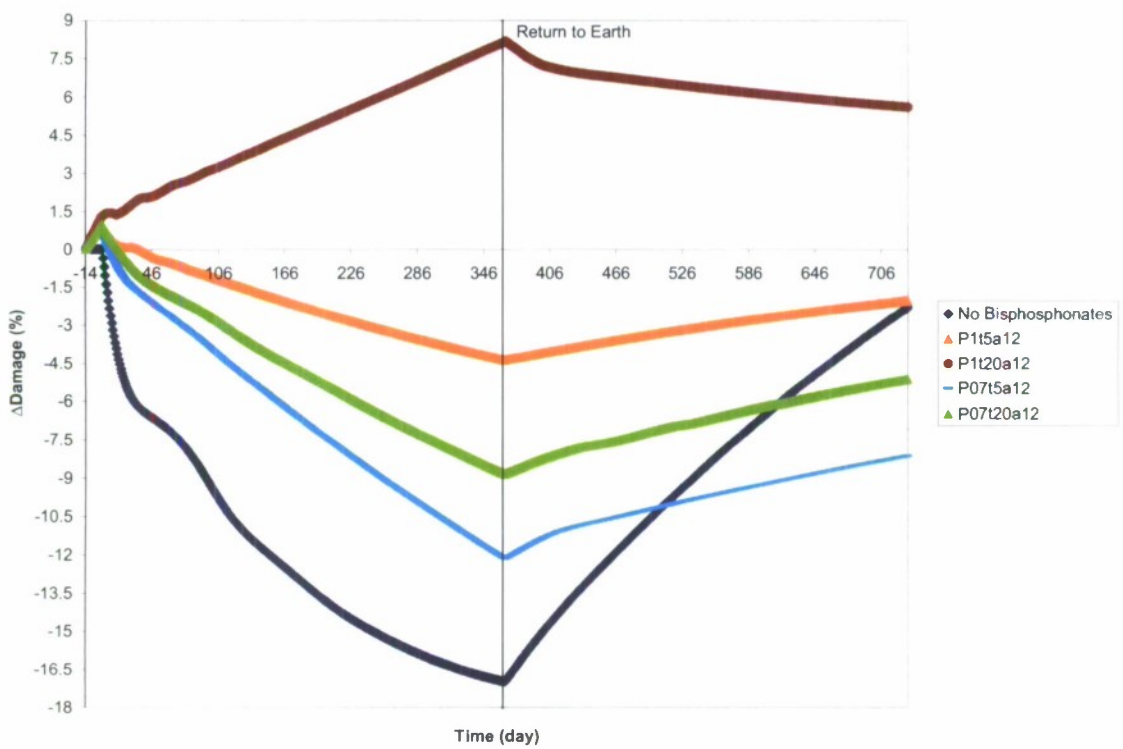


Figure E10. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

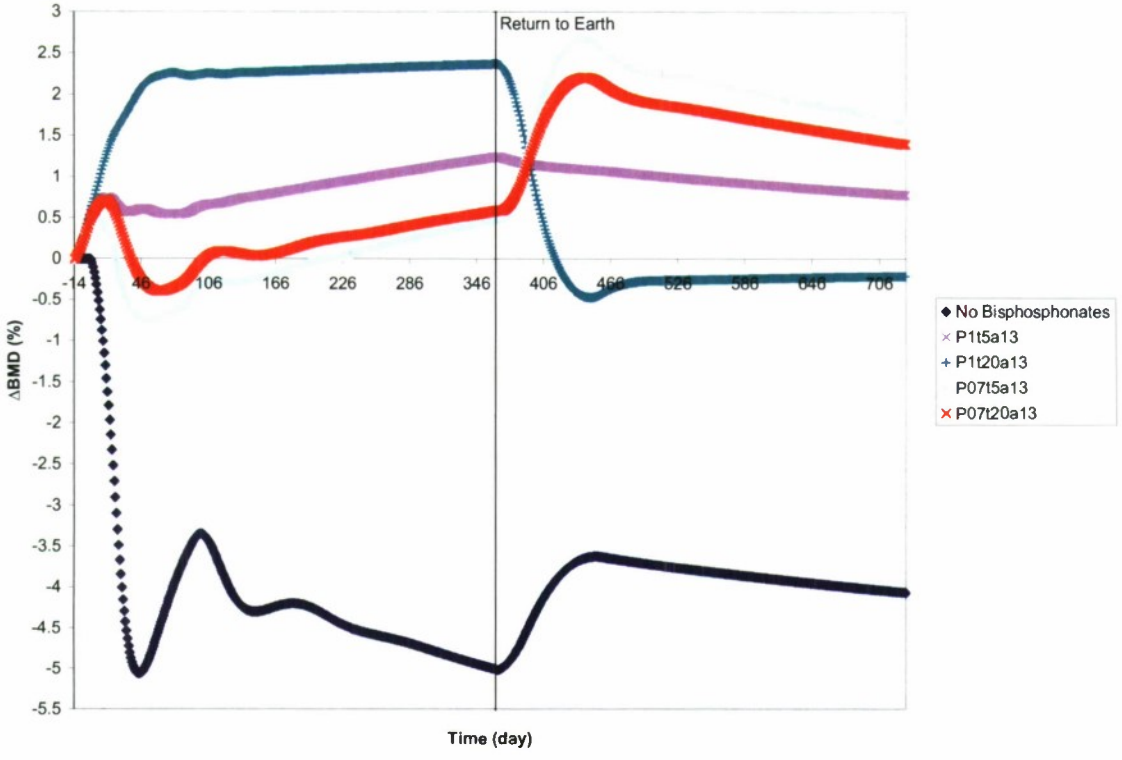


Figure E11. Predicted bisphosphonate effects beginning 14 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

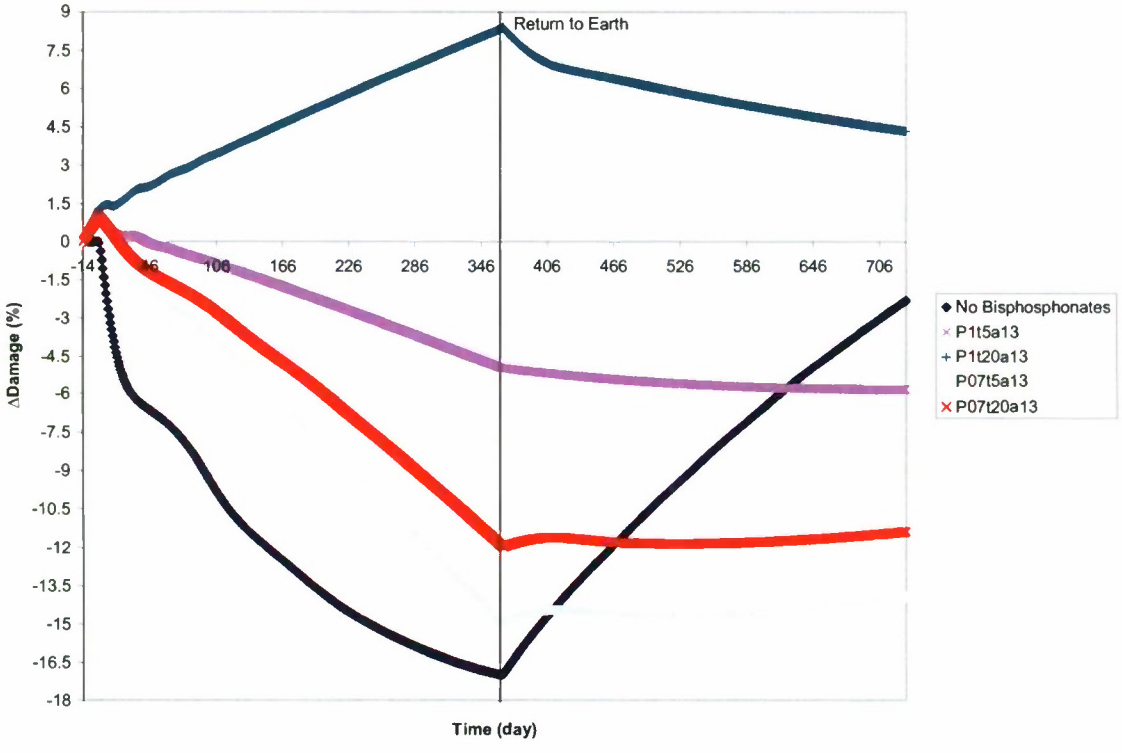


Figure E12. Predicted bisphosphonate effects beginning 14 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

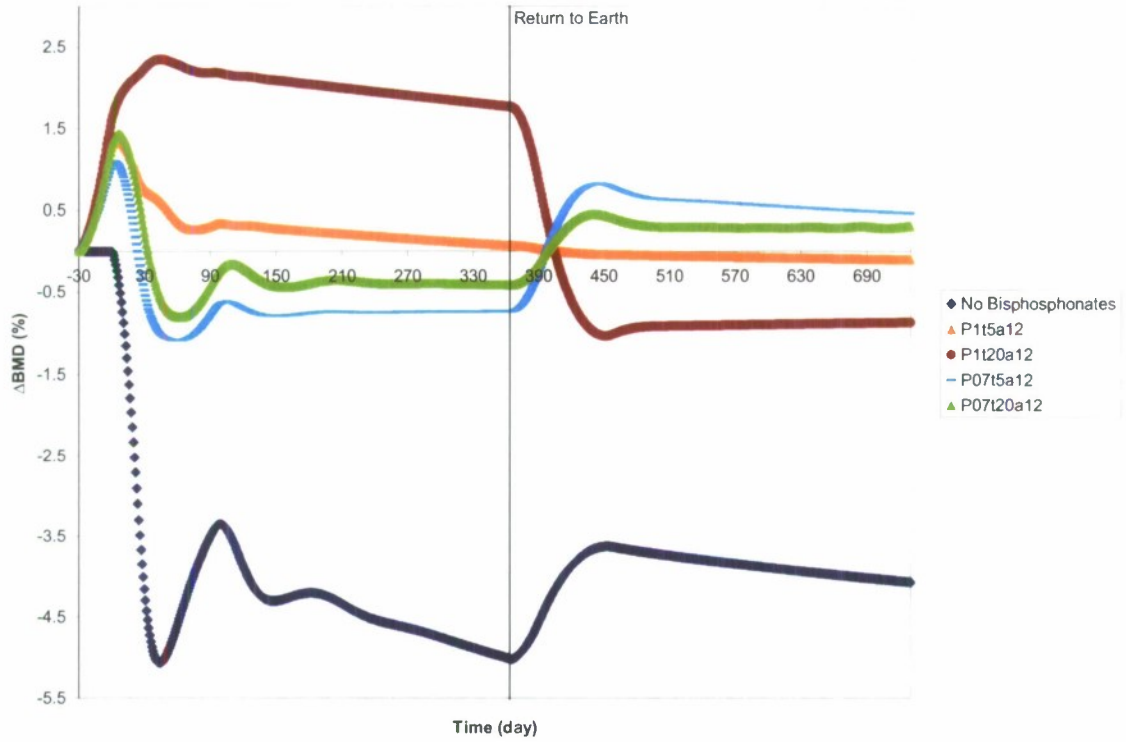


Figure E13. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

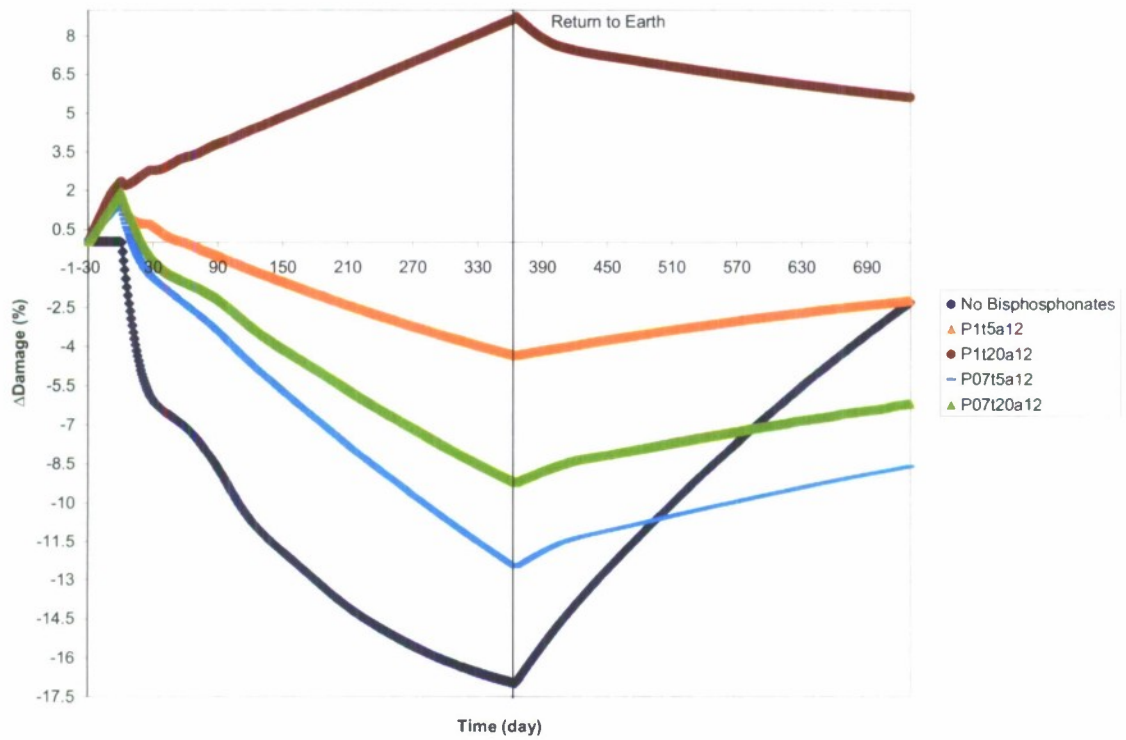


Figure E14. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

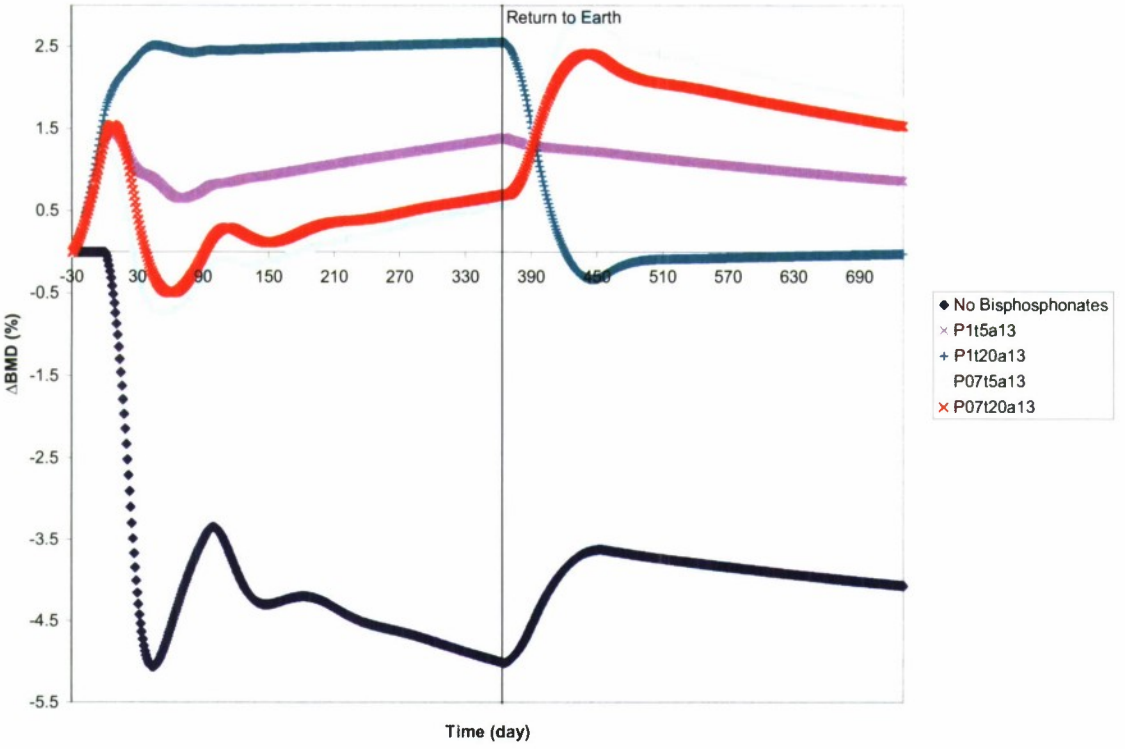


Figure E15. Predicted bisphosphonate effects beginning 30 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

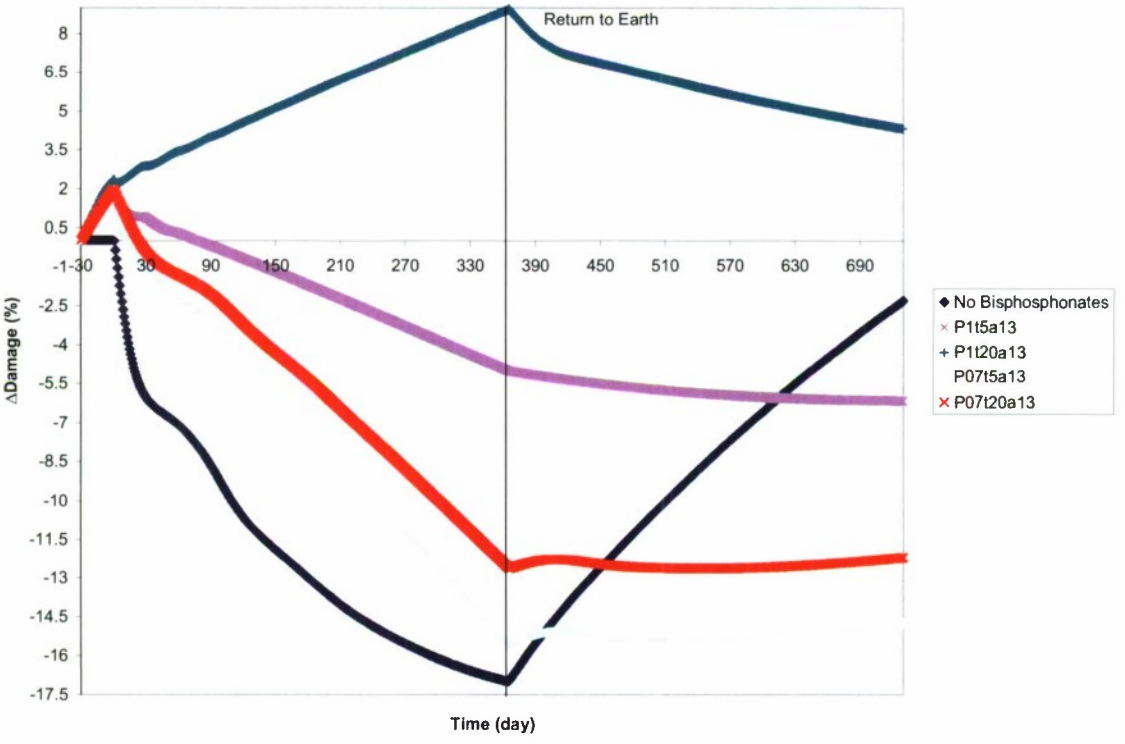


Figure E16. Predicted bisphosphonate effects beginning 30 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

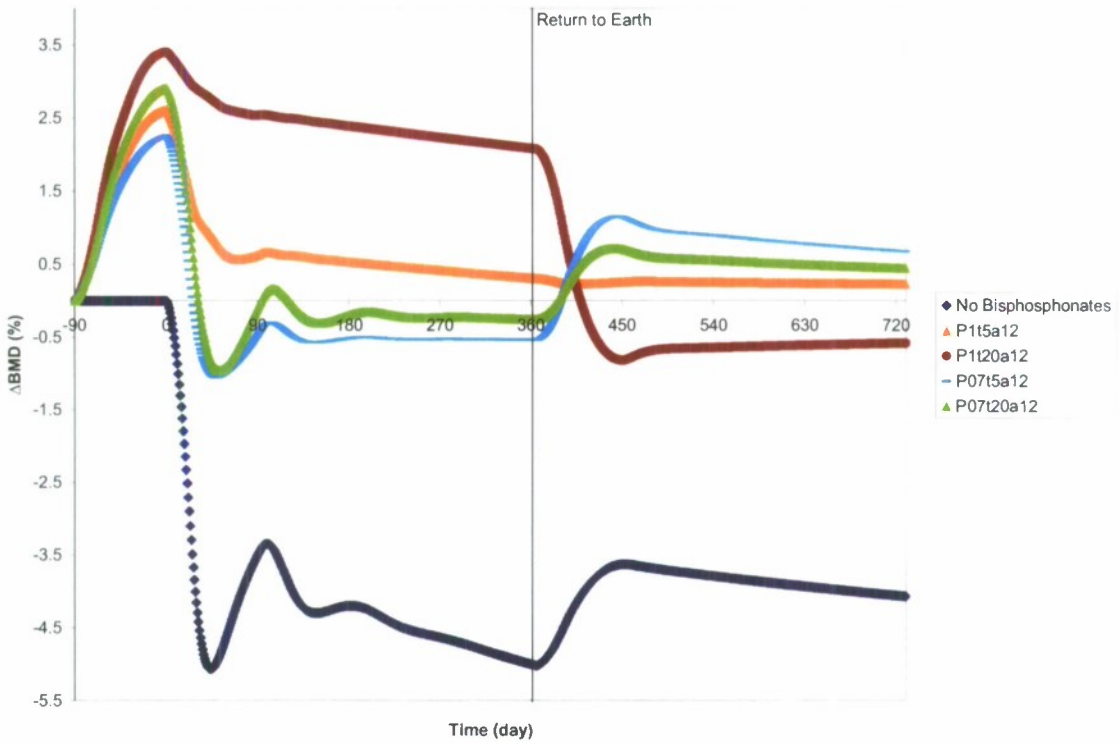


Figure E17. Predieted bisphosphonate effeets beginning 90 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

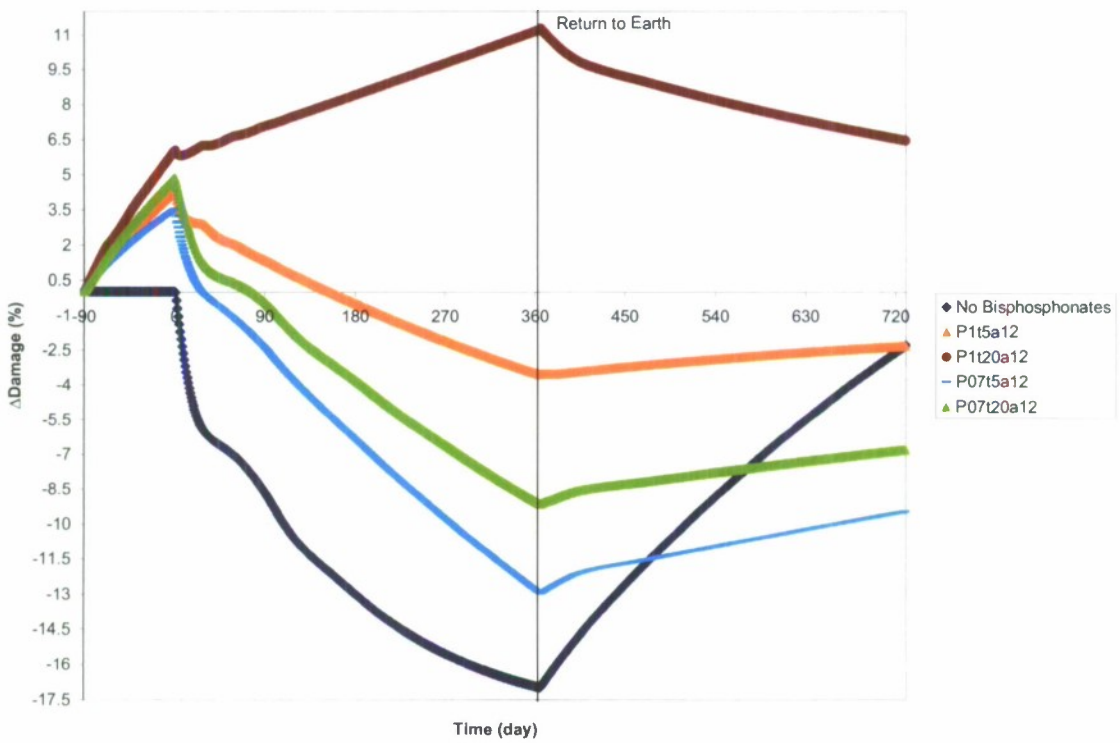


Figure E18. Predicted bisphosphonate effeets beginning 90 days preflight on damage aeeumulation (D) and posttreatment return to Earth from 365-day spaeeflight.

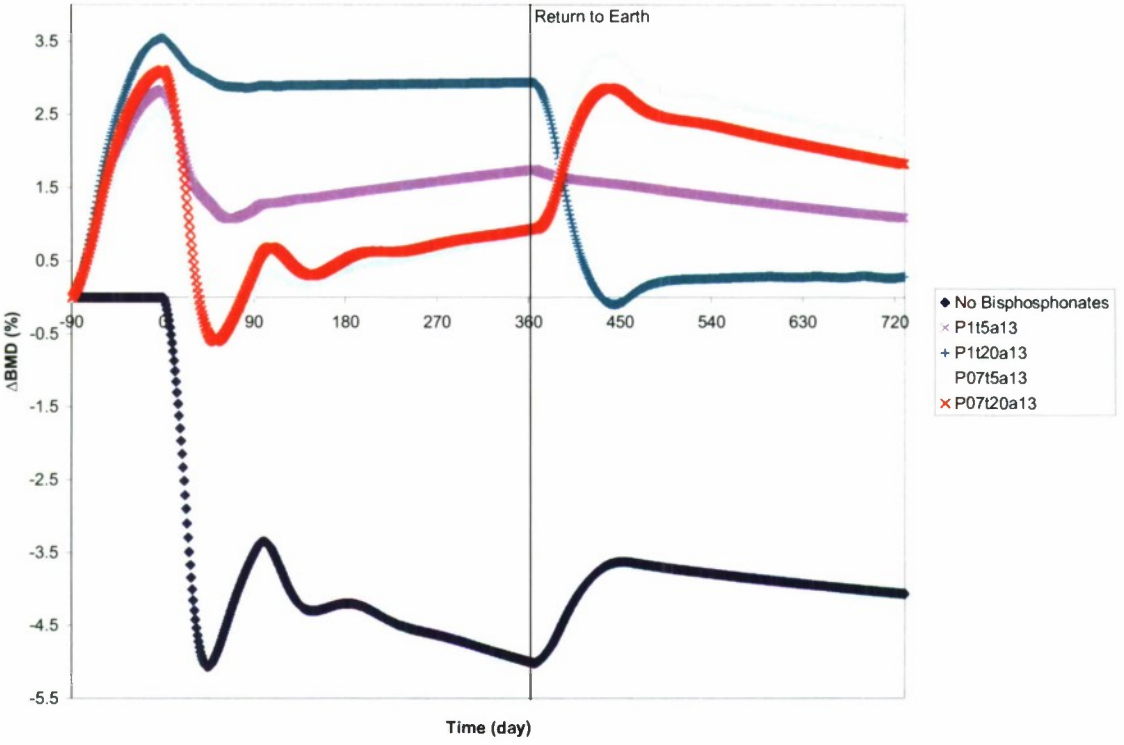


Figure E19. Predicted bisphosphonate effects beginning 90 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

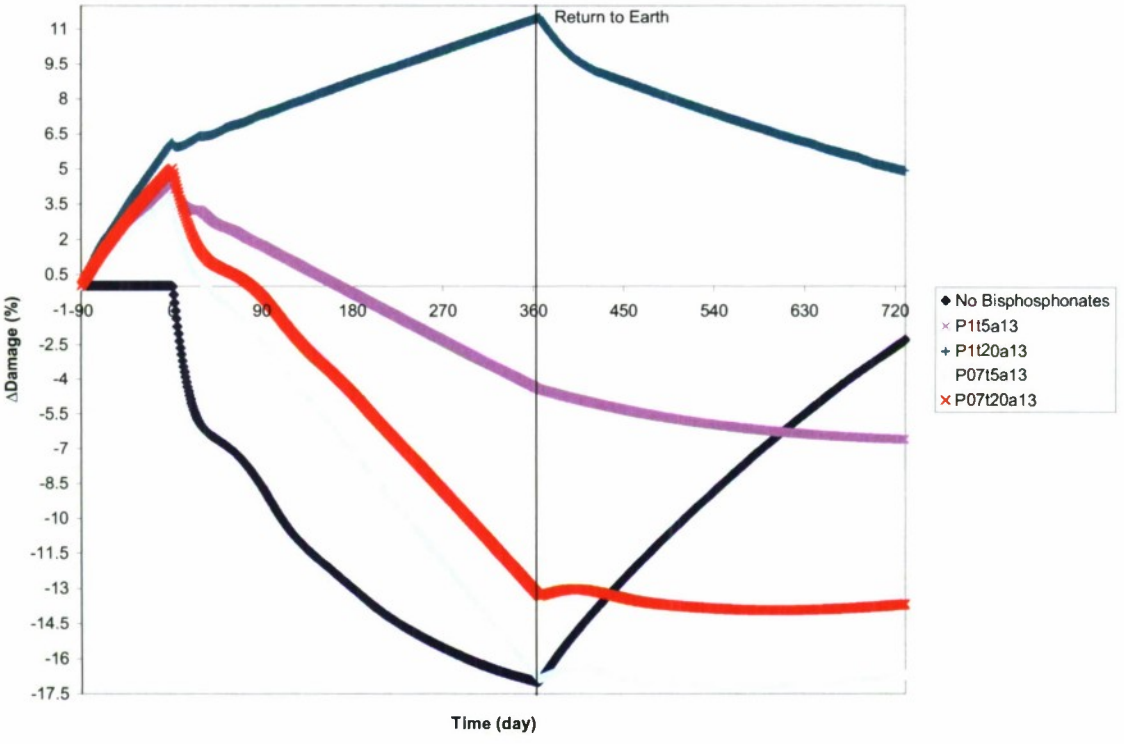


Figure E20. Predicted bisphosphonate effects beginning 90 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

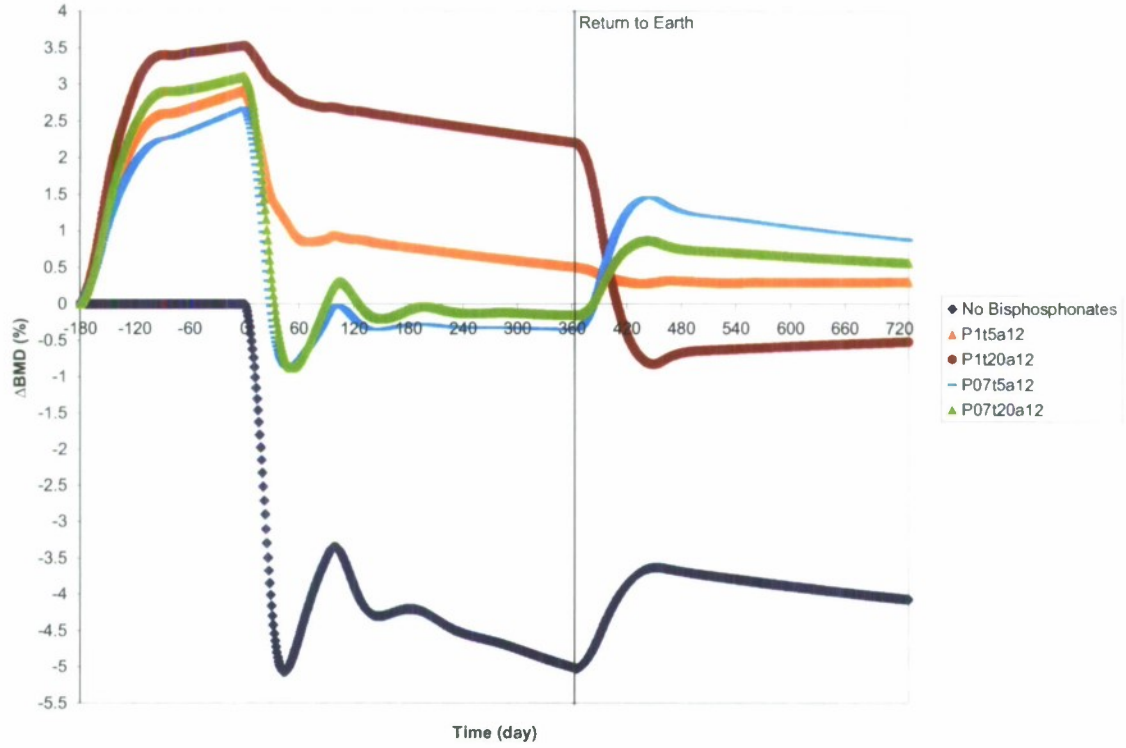


Figure E21. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

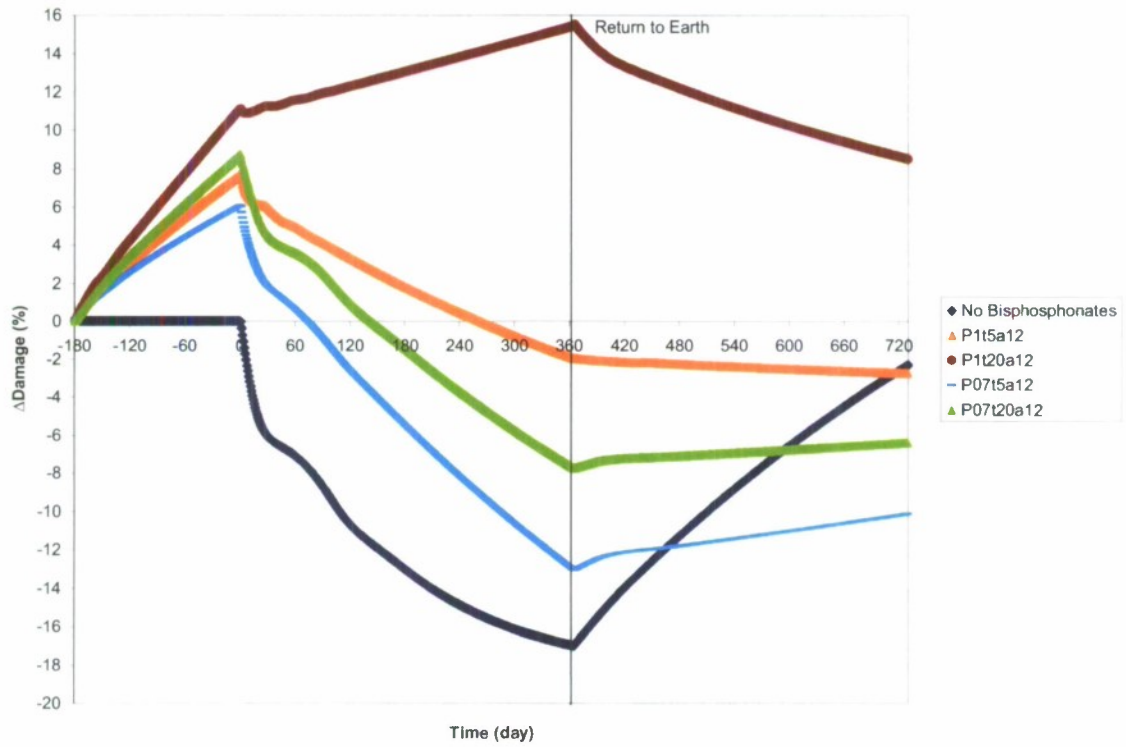


Figure E22. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.

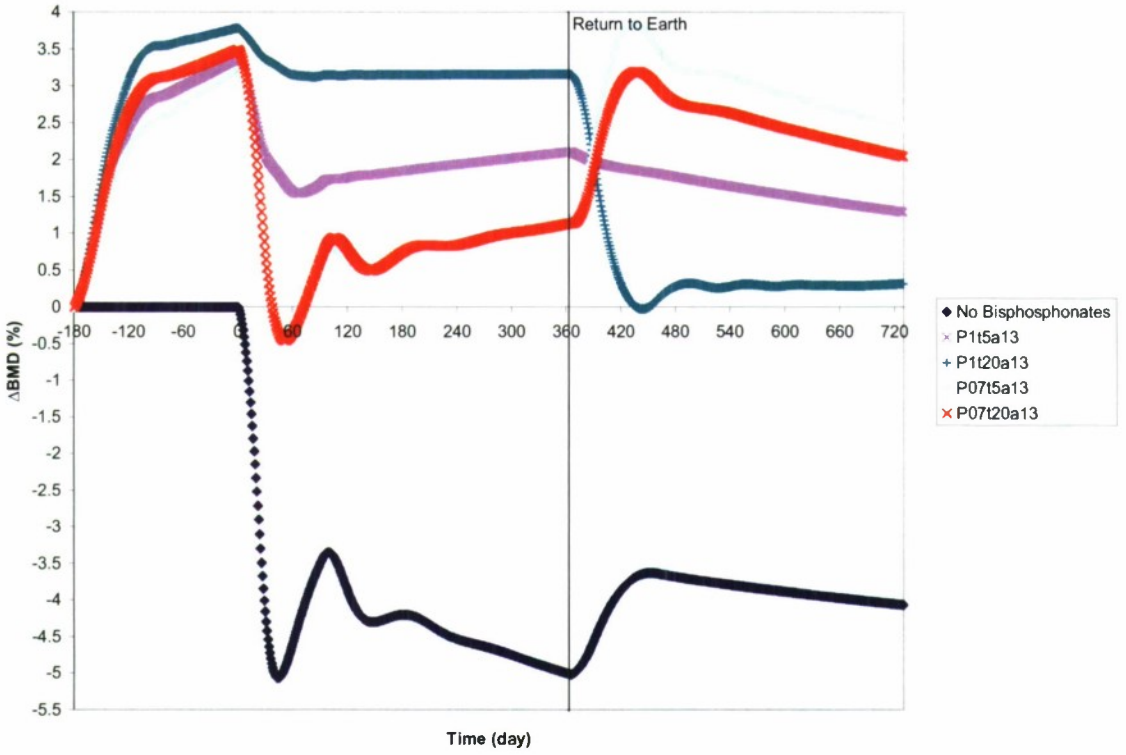


Figure E23. Predicted bisphosphonate effects beginning 180 days preflight on BMD and posttreatment return to Earth from 365-day spaceflight.

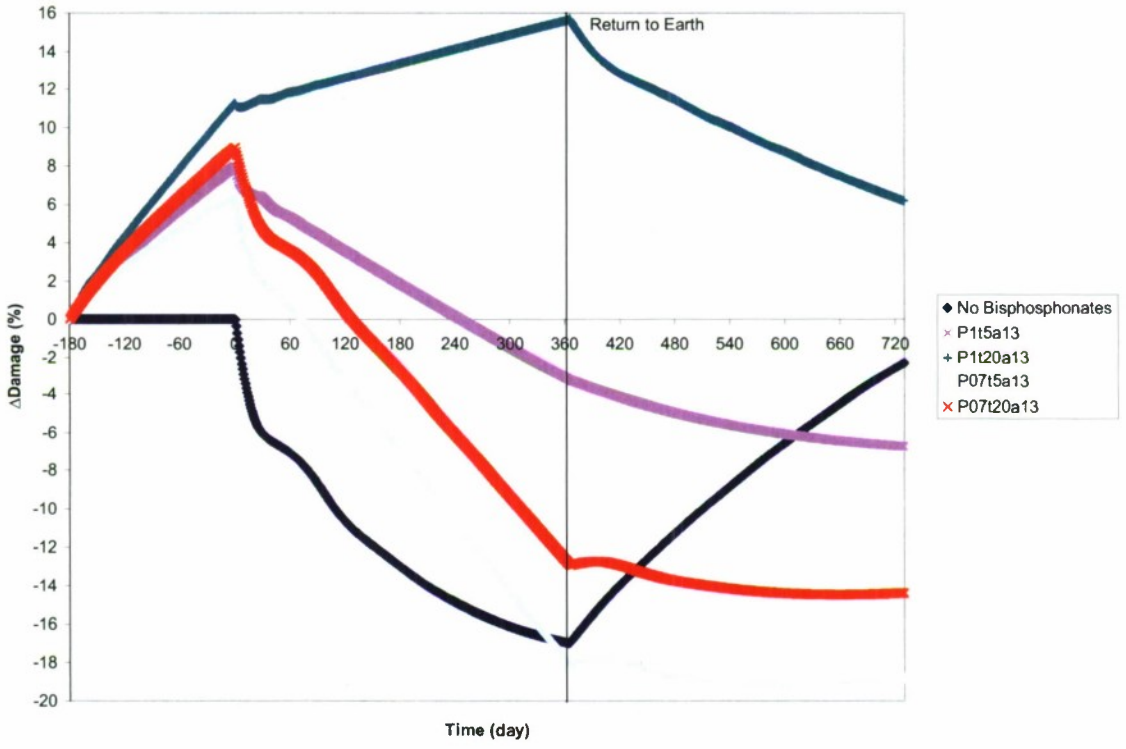


Figure E24. Predicted bisphosphonate effects beginning 180 days preflight on damage accumulation (D) and posttreatment return to Earth from 365-day spaceflight.