

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 this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 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) 10-12-2021		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 14-Sep-2015 - 30-Sep-2021	
4. TITLE AND SUBTITLE Final Report: Active Context			5a. CONTRACT NUMBER W911NF-15-1-0544		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER		
6. AUTHORS			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Harvard Medical School 25 Shattuck Street Boston, MA 02115 -6027			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 67825-CS-DRP.93		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Peter Sorger
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 617-432-6901

RPPR Final Report

as of 20-Dec-2021

Agency Code: 21XD

Proposal Number: 67825CSDRP

Agreement Number: W911NF-15-1-0544

INVESTIGATOR(S):

Name: Peter K. Sorger

Email: Peter_Sorger@hms.harvard.edu

Phone Number: 6174326901

Principal: Y

Organization: **Harvard Medical School**

Address: 25 Shattuck Street, Boston, MA 021156027

Country: USA

DUNS Number: 047006379

EIN: 042103580

Report Date: 31-Oct-2021

Date Received: 10-Dec-2021

Final Report for Period Beginning 14-Sep-2015 and Ending 30-Sep-2021

Title: Active Context

Begin Performance Period: 14-Sep-2015

End Performance Period: 30-Sep-2021

Report Term: 0-Other

Submitted By: Cheryl Hutt

Email: cheryl_hutt@hms.harvard.edu

Phone: (617) 432-6072

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 2

STEM Participants: 1

Major Goals: The primary goal of our project was to build a human-machine communication system for collaborative creation of explanatory mechanistic models in biomedicine. We aimed to build a system leveraging modeling and simulation, as well as large knowledge bases (most of which are themselves machine-built) to provide an open-ended, natural language-driven platform where human and machine play equally important, yet complementary roles. The system aims to achieve major advances in both human-machine communication, and the understanding of complex molecular systems and disease. In addition, we also aimed to develop algorithms and software for reasoning about complex mechanisms operating in the natural world, explaining large-scale data, assisting humans in generating actionable, model-based hypotheses and testing these hypotheses empirically. We tackled this challenge by developing and combining fully automated algorithms for large-scale data interpretation and hypothesis generation with focused, dialogue-driven human-machine communication. Our integrated dialogue systems are designed to accelerate laboratory-based research and to assist molecular tumor boards in reasoning about anti-cancer therapies.

Accomplishments: This report encompasses work between September 2015 and September 2021 in the DARPA Communicating with Computers (CwC) program by the Sorger (HMS), Fontana (HMS), Pratt (UCSD), Harmer (Lyon) and Demir (OHSU) teams. The centerpiece of our work is the development of Bob with Bioagents, a fully integrated and publicly deployed human-machine dialogue system for biocuration. The system integrates natural language interpretation, goal recognition, context-aware problem solving, domain-specific reasoning agents, and natural language generation as well as visualization. It is available with multiple front-ends that focus on different aspects of human-machine collaboration: building model hypotheses of biochemical systems, interpreting experimental data, or advising molecular tumor boards. Second, we developed the CLARE system, an integrated human-machine dialogue system that uses a simplified communication model and software architecture and can be more flexibly extended and deployed, and having the ability to learn new interpretations during runtime.

In addition to dialogue systems, we developed a variety of standalone tools that enable rich human-machine communication or contribute to machine-assisted biological discovery more generally. These include the continued development of the INDRA knowledge assembly system as well as the INDRA Database which monitors and processes all new biomedical literature published each day and makes statements extracted from literature available in a queryable form. It also includes the Biofactoid framework which allows authors to annotate their manuscripts through a web interface; the GeneWalk context-dependent gene function prediction tool; the Gilda named entity grounding system, and several other tools described in more detail in this report.

At multiple stages during our project, our human-machine dialogue system was evaluated by third-party experts

RPPR Final Report as of 20-Dec-2021

from the MITRE team and by domain expert medical school researchers. Evaluations first focused on pre-defined tasks to be completed through dialogue and progressed towards increasingly open-ended scientific questions. The final user evaluation conducted in 2020 asked researchers to analyze data in relation to a real ongoing scientific project and revealed that users were able to construct novel hypotheses or ideas for planning new experiments by working with our system.

Finally, we applied the technologies developed in the course of this project to multiple biological discovery projects. These include the development of an automatically assembled kinase knowledge base in support of the NIH Illuminating the Druggable Genome program and hazard assessment of chemical compounds. We also used large-scale automatically assembled models developed in this project for the automated explanation of gene-gene correlations in The Cancer Dependency Map revealing several novel hypotheses and leading to the design of new experiments around the KIRREL1 gene.

Training Opportunities: A significant portion of work in this project was performed by postdoctoral trainees and graduate students. Postdoctoral fellows are expected to take aspects of this project with them when they establish their own laboratories in industry and academia. This is an essential component of training in a biomedical research environment. There were all together 6 postdoctoral fellows and 2 students funded on this project.

Results Dissemination: The results of our work, and the tools we create to enable this work, are freely distributed to the scientific community as presentations, software and research articles. Our openly accessible software repositories are listed under "Products." We have deployed the Bob with Bioagents system as a web-based service and publicized it within our research groups and to collaborators. The system is publicly available at <http://www.dialogue.bio>. We have also deployed the CLARE system within multiple Slack workspaces including the Harvard Program in Therapeutic Science Slack workspace, and the COVID-19 Disease Maps workspace. CLARE has gotten usage from dozens of scientists all together serving hundreds of interactions. We have also deployed the INDRA DB as a publicly available, standalone service at <https://db.indra.bio>; similarly, the INDRA Network Search is publicly available at <https://network.indra.bio>. Our work on the Dark Kinase Knowledge Base is available at kinome.org. The Biofactoid project to which both the OHSU and HMS team have contributed is available at <https://biofactoid.org/> and is beginning to build up a user base.

Honors and Awards: Benjamin Gyori was awarded the Best Poster award for his presentation at the SysMod (Systems Modeling) special interest group meeting at the ISMB (Intelligent Systems for Molecular Biology) Conference taking place July 2016 in Orlando, Florida.

Benjamin Gyori was selected as a DARPA Riser in 2018 and presented a poster about "Model-building machines monitor, understand, and explain complex systems at scale" at DARPA's 60th anniversary event (D60).

Protocol Activity Status:

Technology Transfer: We released all code in the project into the open source, allowing technologies to be used by a wide variety of academic and industrial users. Licenses are permissive to both academic and commercial use facilitating technology transfer.

PARTICIPANTS:

Participant Type: PD/PI

Participant: Peter Sorger

Person Months Worked: 5.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Benjamin Gyori

Person Months Worked: 5.00

Project Contribution:

National Academy Member: N

Funding Support:

RPPR Final Report
as of 20-Dec-2021

Participant Type: Staff Scientist (doctoral level)

Participant: Artem Sokolov

Person Months Worked: 6.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Catherine Luria

Person Months Worked: 2.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Patrick Greene

Person Months Worked: 3.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Juha Ruokone

Person Months Worked: 10.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Diana Kolusheva

Person Months Worked: 6.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Samuel Newton Prasad Bunga

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Nienke Moret

Person Months Worked: 4.00

Project Contribution:

National Academy Member: N

Funding Support:

RPPR Final Report
as of 20-Dec-2021

Participant Type: Other Professional

Participant: Klas Karis

Person Months Worked: 4.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Albert Steppi

Person Months Worked: 5.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: John Ata Bachman

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Jeremy Muhlich

Person Months Worked: 10.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Undergraduate Student

Participant: William Lawrence Gerecke

Person Months Worked: 3.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Gregory Joseph Baker

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Mohammed AlQuraishi

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

RPPR Final Report
as of 20-Dec-2021

Participant: Nazim Bouatta

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Lily Chylek

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Joseph Cunningham

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Luca Gerosa

Person Months Worked: 2.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Isabel Latorre

Person Months Worked: 2.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

Participant: Daniel Milstein

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Robert Sheehan

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Thiagarajan Subramanian

Person Months Worked: 1.00

Funding Support:

RPPR Final Report
as of 20-Dec-2021

Project Contribution:
National Academy Member: N

Participant Type: Co-Investigator
Participant: Russ Harmer
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Co-Investigator
Participant: Walter Fontana
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Graduate Student (research assistant)
Participant: Levgeniia Oshurko
Person Months Worked: 3.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Other Professional
Participant: Julia Somers
Person Months Worked: 8.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Metin Can Siper
Person Months Worked: 11.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Hannah Manning
Person Months Worked: 10.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Aaron Gary
Person Months Worked: 1.00
Project Contribution:
National Academy Member: N

Funding Support:

RPPR Final Report
as of 20-Dec-2021

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Yves Le Cornec
Person Months Worked: 11.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Benjamin Gyori
Person Months Worked: 5.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Joseph Eastbrook
Person Months Worked: 7.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: Michal Grzadkowski
Person Months Worked: 2.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Faculty
Participant: Ozgun Babur
Person Months Worked: 4.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)
Participant: David Servillo
Person Months Worked: 12.00
Project Contribution:
National Academy Member: N

Funding Support:

Participant Type: Other Professional
Participant: David Welker
Person Months Worked: 6.00
Project Contribution:
National Academy Member: N

Funding Support:

RPPR Final Report
as of 20-Dec-2021

Participant Type: Other Professional

Participant: Ziv David Scully

Person Months Worked: 1.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Kartik Subramanian

Person Months Worked: 4.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Vladimir Rynkov

Person Months Worked: 15.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Co-Investigator

Participant: Dexter Pratt

Person Months Worked: 6.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Staff Scientist (doctoral level)

Participant: Funda Durupinar

Person Months Worked: 13.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Co-Investigator

Participant: Emek Demier

Person Months Worked: 11.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Pierre Andre Bernard Boutillier

Person Months Worked: 3.00

Project Contribution:

National Academy Member: N

Funding Support:

Participant Type: Other Professional

RPPR Final Report

as of 20-Dec-2021

Participant: Mutaamba-Byona Wa Maasha

Person Months Worked: 11.00

Project Contribution:

National Academy Member: N

Funding Support:

ARTICLES:

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: Electronic Proceedings in Theoretical Computer Science

Publication Identifier Type: DOI

Publication Identifier: 10.4204/EPTCS.204.5

Volume: 204

Issue:

First Page #: 47

Date Submitted: 8/31/16 12:00AM

Date Published: 3/1/16 5:00AM

Publication Location:

Article Title: A knowledge representation meta-model for rule-based modelling of signalling networks

Authors: Adrien Basso-Blandin, Walter Fontana, Russ Harmer

Keywords: artificial intelligence, signaling networks

Abstract: The study of signaling pathways and their deregulation in disease states is a complex task. These systems involve many parts and processes but are studied piecewise. Literature and data are consequently fragmented, which makes it difficult to build explanatory models. The rule-based approach to modelling can represent highly combinatorial systems where proteins are capable of simultaneous interactions and are controlled by post-translational modifications. However, it requires highly detailed information about the precise conditions for each interaction. In this paper, we introduce a graph-based meta-model, attuned to the representation of signaling networks, which allows for flexible representation of knowledge at various levels of granularity. Our approach provides a basis for gradual aggregation of fragmented biological knowledge extracted from the literature into an instance of the meta-model from which we can define an automated translation into executable Kappa programs.

Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors

Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: Molecular Systems Biology

Publication Identifier Type: DOI

Publication Identifier: 10.15252/msb.20177651

Volume: 13

Issue: 11

First Page #:

Date Submitted: 10/22/18 12:00AM

Date Published: 11/1/17 4:00AM

Publication Location:

Article Title: From word models to executable models of signaling networks using automated assembly

Authors: Benjamin M. Gyori, John A. Bachman, Kartik Subramanian, Jeremy L. Muhlich, Lucian Galescu, Peter K

Keywords: computational modeling, natural language processing, signaling pathways

Abstract: Word models describing molecular mechanisms are a common currency in biomedical communication but are of limited use in predicting the behavior of complex biological networks. We present an approach to building computational models directly from natural language using automated assembly. Molecular mechanisms described in simple English are read by natural language processing algorithms, converted into an intermediate representation and assembled into executable or network models. We demonstrate the use of our software implementation, INDRA, to model three biological processes of increasing scope: (i) p53 dynamics in response to DNA damage, (ii) adaptive drug resistance in melanoma, and (iii) the RAS signaling pathway. The use of natural language for modeling makes routine tasks more efficient for modeling practitioners and increases the accessibility and transparency of models for the broader biology community.

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info

Acknowledged Federal Support: Y

RPPR Final Report as of 20-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Unknown
Publication Identifier Type: DOI **Publication Identifier:** 10.1093/bioinformatics/btz289
Volume: 2019 **Issue:** btz289 **First Page #:** 1
Date Submitted: 10/9/19 12:00AM **Date Published:** 5/9/19 1:56PM
Publication Location: Oxford, England
Article Title: INDRA-IPM: Interactive pathway modeling using natural language with automated assembly
Authors: Petar V. Todorov, Benjamin M. Gyori, John A. Bachman, Peter K. Sorger
Keywords: pathway, network, visualization, natural language
Abstract: INDRA-IPM is a web-based pathway map modeling tool that allows natural language-based construction with automated assembly and visualization. INDRA-IPM also enables contextualization with expression profiles, and export into various common modeling and graphical standards.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: BCM Bioinformatics
Publication Identifier Type: DOI **Publication Identifier:** <https://doi.org/10.1186/s12859-018-2211-5>
Volume: **Issue:** **First Page #:** **Date Published:** 6/28/18 8:00AM
Date Submitted: 10/22/18 12:00AM
Publication Location: **Article Title:** FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining
Authors: John A. Bachman, Benjamin M. Gyori, Peter K. Sorger
Keywords: Text mining, Protein families, Grounding, Named entity linking, Named entity recognition, Biocuration, Event extraction, Natural language processing
Abstract: For automated reading of scientific publications to extract useful information about molecular mechanisms it is critical that biological entities be correctly associated with uniform identifiers, a process known as named entity linking or "grounding." In a task involving automated reading of 215,000 articles we found that grounding was disproportionately inaccurate for multi-protein families (e.g., "AKT") and complexes with multiple subunits (e.g. "NF-?B"). To address this problem we constructed FamPlex, a manually curated resource defining identifiers for protein families and complexes as they are commonly encountered in biomedical text. Evaluation of REACH extractions on a test corpus of ~54,000 articles showed that FamPlex significantly increased grounding accuracy for families and complexes (from 15% to 71%). FamPlex is available at <https://github.com/sorgerlab/famplex> under the Creative Commons CC0 license and has been integrated into the TRIPS/DRUM and REACH reading systems.
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: Y

RPPR Final Report

as of 20-Dec-2021

Publication Type: Journal Article

Peer Reviewed: N

Publication Status: 1-Published

Journal: bioRxiv

Publication Identifier Type: DOI

Publication Identifier: <https://doi.org/10.1101/2020.04.02.022277>

Volume: Issue:

First Page #:

Date Submitted: 8/30/21 12:00AM

Date Published: 6/30/20 8:00AM

Publication Location:

Article Title: Exploring the understudied human kinome for research and therapeutic opportunities

Authors: Nienke Moret, Changchang Liu, Benjamin M Gyori, John A Bachman, Albert Steppi, Rahil Taujale, Liang

Keywords: kinase, human kinome, kinase inhibitors, drug discovery, cancer, cheminformatics

Abstract: The functions of protein kinases have been widely studied and many kinase inhibitors have been developed into FDA-approved therapeutics. A substantial fraction of the human kinome is nonetheless understudied. In this perspective, members of the NIH Understudied Kinome Consortium mine publicly available databases to assess the functionality of these understudied kinases as well as their potential to be therapeutic targets for drug discovery campaigns. We start with a re-analysis of the kinome as a whole and describe criteria for creating an inclusive set of 710 kinase domains as well as a curated set of 557 protein kinase like (PKL) domains. We define an understudied ('dark') kinome by quantifying the public knowledge on each kinase with a PKL domain using an automatic reading machine. We find a substantial number are essential in the Cancer Dependency Map and differentially expressed or mutated in disease databases such as The Cancer Genome Atlas. Based on this and other data, it seems

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

Publication Type: Journal Article

Peer Reviewed: N

Publication Status: 1-Published

Journal: bioRxiv

Publication Identifier Type: DOI

Publication Identifier: <https://doi.org/10.1101/755579>

Volume: 22 Issue: 55

First Page #:

Date Submitted: 8/27/21 12:00AM

Date Published: 2/2/21 5:00AM

Publication Location:

Article Title: GeneWalk identifies relevant gene functions for a biological context using network representation learning

Authors: Robert Ietswaart, Benjamin M. Gyori, John A. Bachman, Peter K. Sorger, L. Stirling Churchman

Keywords: GeneWalk, functional analysis, differential expression, machine learning, network representation learning, INDRA (Integrated Network and Dynamical Reasoning Assembler), Pathway Commons, GO enrichment, gene set enrichment analysis, next-generation sequencing, RNA-seq, NET-seq

Abstract: A bottleneck in high-throughput genomics experiments is identifying the most important genes and their relevant functions from a list of gene hits. Existing methods such as Gene Ontology (GO) enrichment analysis provide insight at the gene set level. For individual genes, GO annotations are static and biological context can only be added by manual literature searches. Here we introduce GeneWalk (github.com/churchmanlab/genewalk), a method to identify individual genes and relevant functions under particular experimental conditions. After automatic assembly of an experiment-specific gene regulatory network, GeneWalk quantifies the similarity between vector representations of each gene and its GO annotations through representation learning, yielding annotation significance scores that reflect their functional relevance for the experimental context. GeneWalk performs functional analysis to convert a gene list into data-driven hypotheses accelerating interpretation of genomics experiments.

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

RPPR Final Report as of 20-Dec-2021

Publication Type: Journal Article

Peer Reviewed: N

Publication Status: 1-Published

Journal: bioRxiv preprint

Publication Identifier Type: DOI

Publication Identifier: <https://doi.org/10.1101/2021.03.10.382333>

Volume: Issue:

First Page #:

Date Submitted: 8/30/21 12:00AM

Date Published: 3/11/21 5:00AM

Publication Location:

Article Title: Capturing scientific knowledge in computable form

Authors: Jeffrey Wong, Max Franz, Metin Can Siper, Dylan Fong, Funda Durupinar, Christian Dallago, Augustin L

Keywords: systems biology, biocuration, text mining

Abstract: Technological advances in computing provide major opportunities to accelerate scientific discovery. The wide availability of structured knowledge would allow us to take full advantage of these by enabling efficient human-computer interaction. Traditionally, biological knowledge is captured in publications and knowledge bases, however, the information in articles is not directly accessible to computers, and knowledge bases are constrained by finite resources available for manual curation. To accelerate knowledge capture and communication and to keep pace with the rapid growth of scientific reports, we developed the Biofactoid (biofactoid.org) software suite, which crowdsources structured knowledge in articles from authors. Biofactoid is a web-based system that lets scientists draw a network of interactions between genes, their products, and chemical compounds and employs smart-automation to translate user input into a structured language using the expressive power of a formal ontology.

Distribution Statement: 4-Distribution authorized to the Department of Defense and U.S. DoD contractors only
Acknowledged Federal Support: Y

Publication Type: Journal Article

Peer Reviewed: Y

Publication Status: 1-Published

Journal: Genome Biology

Publication Identifier Type: DOI

Publication Identifier: 10.1186/s13059-021-02264-8

Volume: 22 Issue: 55

First Page #:

Date Submitted: 12/10/21 12:00AM

Date Published: 2/2/21 5:00AM

Publication Location:

Article Title: GeneWalk identifies relevant gene functions for a biological context using network representation learning

Authors: Robert Letswaart, Benjamin M. Gyori, John A. Bachman, Peter K. Sorger, L. Stirling Churchman

Keywords: GeneWalk, Functional analysis, Differential expression, Machine learning, Network representation learning, INDRA (Integrated Network and Dynamical Reasoning Assembler), Pathway Commons, GO enrichment, Gene set enrichment analysis, Next-generation sequencing, RNA-seq, NET-seq

Abstract: A bottleneck in high-throughput functional genomics experiments is identifying the most important genes and their relevant functions from a list of gene hits. Gene Ontology (GO) enrichment methods provide insight at the gene set level. Here, we introduce GeneWalk (github.com/churchmanlab/genewalk) that identifies individual genes and their relevant functions critical for the experimental setting under examination. After the automatic assembly of an experiment-specific gene regulatory network, GeneWalk uses representation learning to quantify the similarity between vector representations of each gene and its GO annotations, yielding annotation significance scores that reflect the experimental context. By performing gene- and condition-specific functional analysis, GeneWalk converts a list of genes into data-driven hypotheses.

Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: Y

RPPR Final Report as of 20-Dec-2021

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: Bioinformatics

Publication Identifier Type: DOI

Publication Identifier: 10.1093/bioinformatics/btz289

Volume: 35

Issue: 21

First Page #: 4501

Date Submitted: 12/10/21 12:00AM

Date Published: 11/1/19 4:00AM

Publication Location:

Article Title: INDRA-IPM: interactive pathway modeling using natural language with automated assembly

Authors: Petar V Todorov, Benjamin M Gyori, John A Bachman, Peter K Sorger

Keywords: Pathway modeling, INDRA

Abstract: Summary INDRA-IPM (Interactive Pathway Map) is a web-based pathway map modeling tool that combines natural language processing with automated model assembly and visualization. INDRA-IPM contextualizes models with expression data and exports them to standard formats. Availability and implementation INDRA-IPM is available at: <http://pathwaymap.indra.bio>. Source code is available at http://github.com/sorgerlab/indra_pathway_map. The underlying web service API is available at <http://api.indra.bio:8000>. Supplementary information Supplementary data are available at Bioinformatics online.

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published

Journal: IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Publication Identifier Type: DOI

Publication Identifier: 10.1109/TCBB.2019.2906164

Volume:

Issue:

First Page #:

Date Submitted: 12/10/21 12:00AM

Date Published: 3/19/19 4:00AM

Publication Location:

Article Title: Bio-Curation for Cellular Signalling: The KAMI Project

Authors: Russ Harmer, Yves-Stan Le Cornec, Sebastien Legare, Eugenia Oshurko

Keywords: Knowledge representation, bio-curation, graph rewriting

Abstract: The general question of what constitutes bio-curation for rule-based modelling of cellular signalling is posed. A general approach to the problem is presented, based on rewriting in hierarchies of graphs, together with a specific instantiation of the methodology that addresses our particular bio-curation problem. The current state of the ongoing development of the KAMI bio-curation tool, based on this approach, is outlined along with our plans for future development.

Distribution Statement: 1-Approved for public release; distribution is unlimited.

Acknowledged Federal Support: Y

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 4-Under Review

Journal: bioRxiv preprint

Publication Identifier Type: DOI

Publication Identifier: 10.1101/2021.09.10.459803

Volume:

Issue:

First Page #:

Date Submitted: 12/10/21 12:00AM

Date Published:

Publication Location:

Article Title: Gilda: biomedical entity text normalization with machine-learned disambiguation as a service

Authors: Benjamin M. Gyori, Charles Tapley Hoyt, Albert Steppi

Keywords: Named entity normalization

Abstract: Summary Gilda is a software tool and web service which implements a scored string matching algorithm for names and synonyms across entries in biomedical ontologies covering genes, proteins (and their families and complexes), small molecules, biological processes and diseases. Gilda integrates machine-learned disambiguation models to choose between ambiguous strings given relevant surrounding text as context, and supports species-prioritization in case of ambiguity. Availability The Gilda web service is available at <http://grounding.indra.bio> with source code, documentation and tutorials are available via <https://github.com/indralab/gilda>.

Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info

Acknowledged Federal Support: Y

RPPR Final Report
as of 20-Dec-2021

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: Annual International Conference on Intelligent Systems for Molecular Biology
Date Received: 31-Aug-2016 Conference Date: 08-Jul-2016 Date Published: 08-Jul-2016
Conference Location: Orlando, Florida
Paper Title: Automated assembly of mechanistic pathway models from natural language, literature and databases
Authors: Benjamin Gyori, John Bachman, Kartik Subramanian, Jeremy Muhlich, Daniel Carlin, Dexter Pratt, Peter
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: International Conference on Systems Biology of Human Disease
Date Received: 31-Aug-2016 Conference Date: 14-Jun-2016 Date Published: 14-Jun-2016
Conference Location: Cambridge, Massachusetts
Paper Title: Automated assembly of mechanistic pathway models from natural language, literature and databases
Authors: Benjamin Gyori, John Bachman, Kartik Subramanian, Jeremy Muhlich, Daniel Carlin, Dexter Pratt, Peter
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: Harvard Program in Therapeutic Science Annual Symposium
Date Received: 05-Sep-2017 Conference Date: 13-Dec-2016 Date Published:
Conference Location: Boston, MA
Paper Title: Our future lab mate: Bob - an artificial intelligence dialogue system for biology
Authors: Benjamin M. Gyori, John A. Bachman, Lucian Galescu, Jeff Rye, Scott Friedman, David McDonald, Mar
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: 12th Annual q-bio Conference
Date Received: 22-Oct-2018 Conference Date: 26-Jun-2018 Date Published:
Conference Location: Houston, TX
Paper Title: Using the automated building of computational models to understand cardiotoxic drug responses
Authors: Robert P. Sheehan, Huan Wang, Benjamin M. Gyori, John A. Bachman, Peter K. Sorger
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: 26th Intelligent Systems for Molecular Biology (ISMB) Conference
Date Received: 22-Oct-2018 Conference Date: 06-Jul-2018 Date Published:
Conference Location: Chicago, IL
Paper Title: A natural language dialogue system for model-driven explanation of biological experiments
Authors: Benjamin M. Gyori, John A. Bachman, Patrick Greene, Funda Durupinar, Xue Zhang, James Allen, Luci
Acknowledged Federal Support: **Y**

WEBSITES:

RPPR Final Report as of 20-Dec-2021

URL: <http://www.home.ndexbio.org/>

Date Received: 31-Aug-2016

Title: The Data Network Exchange

Description: The NDEx Project provides an open-source framework where scientists and organizations can share, store, manipulate, and publish biological network knowledge.

URL: <http://www.ndexbio.org/#/network/50e3dff7-133e-11e6-a039-06603eb7f303>

Date Received: 31-Aug-2016

Title: The Ras Machine

Description: The RAS Machine reads new literature on RAS daily, extracts and assembles mechanisms into a computational model using INDRA (github.com/sorgerlab/indra)

URL: <http://www.rasmodel.org>

Date Received: 31-Aug-2016

Title: The Ras Executable Model

Description: The Ras Executable Model (REM) is a model of the signaling pathway of the Ras family proteins (K-RAS, N-RAS, H-RAS), including their regulators and effectors, at a biochemical level of detail. It is a rule-based model written in PySB.

URL: <http://pathwaymap.indra.bio>

Date Received: 09-Oct-2019

Title: INDRA Interactive Pathway Map (INDRA-IPM)

Description: Website 1

URL: <https://db.indra.bio>

Date Received: 09-Oct-2019

Title: INDRA Database

Description: Website 2

URL: <http://54.84.114.146/>

Date Received: 09-Oct-2019

Title: Bob with Bioagents portal

Description: Website 3

URL: <http://daialogue.bio>

Date Received: 28-Oct-2020

Title: Bob with Bioagents portal

Description: Bob with Bioagents portal

URL: <https://darkkinome.org/>

Date Received: 28-Oct-2020

Title: Dark Kinase Knowledge Base

Description: Dark Kinase Knowledge Base

URL: <http://grounding.indra.bio>

Date Received: 28-Oct-2020

Title: Gilda grounding service

Description: Gilda grounding service

URL: <https://biofactoid.org>

Date Received: 28-Oct-2020

Title: Biofactoid

Description: Biofactoid

URL: https://github.com/bgyori/cwc_integ_service

Date Received: 30-Aug-2021

Title: CwC service

Description: This repository contains all the functionalities needed to host the integrated CwC system as a service, and includes a website, an application server, a Docker container session management system, and utilities for exporting logs from dialogue sessions.

URL: <https://github.com/bgyori/clare>

Date Received: 30-Aug-2021

Title: CLARE website

Description: A dialogue system which exposes Bioagents capabilities through multiple front ends, and implements context-aware dialogue management driven by multiple levels of natural language interpretation.

RPPR Final Report as of 20-Dec-2021

URL: <https://github.com/indralab/pybiopax>

Date Received: 30-Aug-2021

Title: PyBioPAX website

Description: The PyBioPAX package implements the BioPAX Level 3 object model as a set of Python classes. This is made available as a standalone software package and allows using BioPAX content with INDRA and Bioagents natively in Python.

URL: <https://github.com/sorgerlab/bioagents>

Date Received: 30-Aug-2021

Title: Bioagents website

Description: Biological reasoning agents that are the basis of domain-specific reasoning in the Bob with Bioagent system.

URL: <https://github.com/bgyori/pykqml>

Date Received: 30-Aug-2021

Title: PyKQML website

Description: A package to read, write and process KQML messages in Python. KQML is the basis of communication between agents in the Bob with Bioagents architecture.

URL: <https://github.com/sorgerlab/indra>

Date Received: 30-Aug-2021

Title: INDRA Github

Description: The Integrated Network and Dynamical Reasoning Assembler (INDRA) is an automated model assembly system which interfaces with NLP systems and pathway databases to collect information and assembled them into a variety of executable and causal modeling formalisms. INDRA is the basis of many Bioagents functionalities as well as automated scientific discovery approaches.

URL: https://github.com/indralab/indra_db

Date Received: 30-Aug-2021

Title: INDRA DB website

Description: The INDRA Database is a framework for creating, maintaining, and accessing a database of literature content, reading system outputs, and INDRA Statements on Amazon Web Services. This database provides a comprehensive collection of mechanistic information used by multiple Bioagents.

URL: https://github.com/indralab/indra_network_search

Date Received: 30-Aug-2021

Title: INDRA Network Search

Description: The INDRA Network Search is a web service that exposes the assembled knowledge in the INDRA DB for network path searches.

URL: <https://github.com/indralab/opaque>

Date Received: 30-Aug-2021

Title: Opaque website

Description: Opaque (Outlier Prevalence Analysis for Quantification of Unknown Entities) is an open-world machine learning tool. It allows for interval estimation of the proportion of datapoints in a real-world dataset that are outside of the closed world of a training set under certain circumstances.

URL: <https://github.com/indralab/adeft>

Date Received: 30-Aug-2021

Title: Adeft website

Description: Adeft (Acromine based Disambiguation of Entities from Text context) is a software for building models to disambiguate acronyms and other abbreviations of biological terms in the scientific literature. It is able to identify possible longform expansions for shortforms in a text corpus. Adeft also allows users to build disambiguation models to disambiguate shortforms based on their text context.

URL: <https://github.com/labsyspharm/kinome-website>

Date Received: 30-Aug-2021

Title: Kinome.org Github

Description: This repository contains the code for the kinome.org portal which integrates quantifiers of knowledge about each kinase from INDRA, as well as links to kinase networks generated by INDRA.

URL: https://github.com/IDG-Kinase/indra_analysis

Date Received: 30-Aug-2021

Title: INDRA kinase networks

Description: This repository contains the scripts to assemble kinase-specific networks and quantify the amount of knowledge about each kinase using INDRA.

RPPR Final Report as of 20-Dec-2021

URL: <https://github.com/churchmanlab/genewalk>

Date Received: 30-Aug-2021

Title: GeneWalk Github

Description: GeneWalk determines for individual genes the functions that are relevant in a particular biological context and experimental condition. GeneWalk quantifies the similarity between vector representations of a gene and annotated GO terms through representation learning with random walks on a condition-specific gene regulatory network, including ones assembled by INDRA

URL: https://github.com/indralab/go_networks

Date Received: 30-Aug-2021

Title: INDRA GO networks

Description: The INDRA GO networks repository provides the scripts to generate networks based on INDRA Statements for genes annotated with a given GO term. Networks are automatically assembled, styled, and uploaded to NDEx. They are then made available via the iQuery service for gene set enrichment analysis (<https://iquery.ndexbio.org/>).

URL: https://github.com/bernice493/INDRA_hazard

Date Received: 30-Aug-2021

Title: INDRA hazard identification

Description: This repository contains the scripts to use Reach and INDRA to assemble networks around compounds of interest from articles describing their effects on carcinogenesis

URL: https://github.com/bernice493/INDRA_hazard

Date Received: 30-Aug-2021

Title: INDRA hazard

Description: : This repository contains the scripts to use Reach and INDRA to assemble networks around compounds of interest from articles describing their effects on carcinogenesis.

URL: <https://kinome.org>

Date Received: 27-Aug-2021

Title: Kinome.org portal

Description: Kinome.org portal

URL: <https://sorgerlab.github.io/famplex/>

Date Received: 27-Aug-2021

Title: FamPlex website

Description: FamPlex website

URL: <https://network.indra.bio>

Date Received: 27-Aug-2021

Title: INDRA network search

Description: INDRA network search

Partners

,

During this project we took part in several collaborations to advance our transition goals. We interacted with the NIH

I certify that the information in the report is complete and accurate:

Signature: Benjamin Gyori

Signature Date: 12/10/21 4:57PM

CwC: Final Report

Human-machine collaboration and automated hypothesis generation to accelerate scientific discovery

Peter Sorger PhD Principal Investigator; Benjamin Gyori PhD Project Lead

12/2/2021

Agreement Number W911NF-15-1-0544



Summary of challenge

Collaboration between humans and machines in which machines are not mere receivers of instructions but active collaborators—able to synthesize information, develop unanticipated responses, and harness several modes of communication including natural language—has the potential to dramatically advance basic biological science as well as healthcare. To date, two-way sharing of complex ideas in a collaborative context has proven difficult to achieve, particularly for highly technical topics where reasoning involves a combination of evidence and theory. Prior knowledge is often essential in the interpretation of technical material, but can be difficult to precisely delineate and define. We are tackling this challenge by developing and combining fully automated algorithms for large-scale data interpretation and hypothesis generation with focused, dialogue-driven human-machine communication. Our system is designed to accelerate laboratory-based research and to assist molecular tumor boards in reasoning about anti-cancer therapies.

Executive summary

This report encompasses work between September 2015 and September 2021 in the DARPA Communicating with Computers (CwC) program by the Sorger (HMS), Fontana (HMS), Pratt (UCSD), Harmer (Lyon) and Demir (OHSU) teams. The centerpiece of our work is the development of Bob with Bioagents, a fully integrated and publicly deployed human-machine dialogue system for biocuration. The system integrates natural language interpretation, goal recognition, context-aware problem solving, domain-specific reasoning agents, and natural language generation as well as visualization. It is available with multiple front-ends that focus on different aspects of human-machine collaboration: building model hypotheses of biochemical systems, interpreting experimental data, or advising molecular tumor boards. Second, we developed the CLARE system, an integrated human-machine dialogue system that uses a simplified communication model and software architecture and can be more flexibly extended and deployed, and having the ability to learn new interpretations during runtime.

In addition to dialogue systems, we developed a variety of standalone tools that enable rich human-machine communication or contribute to machine-assisted biological discovery more generally. These include the continued development of the INDRA knowledge assembly system as well as the INDRA Database which monitors and processes all new biomedical literature published each day and makes statements extracted from literature available in a queryable form. It also includes the Biofactoid framework which allows authors to annotate their manuscripts through a web interface; the GeneWalk context-dependent gene function prediction tool; the Gilda named entity grounding system, and several other tools described in more detail in this report.

At multiple stages during our project, our human-machine dialogue system was evaluated by third-party experts from the MITRE team and by domain expert medical school researchers. Evaluations first focused on pre-defined tasks to be completed through dialogue and progressed towards increasingly open-ended scientific questions. The final user evaluation conducted in 2020 asked researchers to analyze data in relation to a real ongoing scientific project and revealed that users were able to construct novel hypotheses or ideas for planning new experiments by working with our system.

Finally, we applied the technologies developed in the course of this project to multiple biological discovery projects. These include the development of an automatically assembled kinase knowledge base in support of the NIH Illuminating the Druggable Genome program and hazard assessment of chemical compounds. We also used large-scale automatically assembled models developed in this project for the automated explanation of gene-gene correlations in The Cancer Dependency Map revealing several novel hypotheses and leading to the design of new experiments around the KIRREL1 gene.

Table of Contents

<i>Summary of challenge</i>	2
<i>Executive summary</i>	2
<i>Scientific progress and accomplishments</i>	4
1. Human-machine collaboration through dialogue	4
1.1. The Bob with Bioagents dialogue system	4
1.2. The CLARE dialogue system	10
2. Component technologies	15
2.1 INDRA	15
2.2. INDRA Database	16
2.3. PyBioPAX	18
2.4. Gilda	19
2.5. INDRA Interactive Pathway Map	20
2.6. Kappa and KAMI	21
3. Applications	22
3.1. Automated explanation of cancer gene dependencies at scale	22
3.2. Automated finding of surprising, significant, and actionable hypotheses: a case study of KIRREL1 function	24
3.3 Biofactoid	27
3.4. GeneWalk	28
3.5. IQuery: network-based gene set enrichment	29
3.6. Illuminating the Druggable Genome: Dark Kinases	30
3.7. Using INDRA for chemical hazard identification	32
3.8. COVID-19	33
3.9. Molecular Tumor Board	34

Scientific progress and accomplishments

1. Human-machine collaboration through dialogue

1.1. The Bob with Bioagents dialogue system

1.1.1. The integrated system

The Bob with Bioagents (BwB) system brings together components from a consortium of teams: in addition to the HMS team, the OHSO, IHMC, SIFT and Tufts teams developed different parts of the system. The integrated system delivers the following key capabilities:

- Working implementation of a generic **language understanding** system
- Working implementation of generic **collaborative problem solving and planning**
- Working implementations of **biological problem solving agents (Bioagents)**, which are generic for their specific sub-tasks
- These are integrated into a working **communication-for-biocuration system**

BwB is a natural language-based computational partner to help develop hypotheses and make decisions about complex biological systems and the experiments used to explore them. The system consists of three types of components that communicate with each other via messaging: (1) natural language understanding and dialogue management (developed by IHMC, yellow), (2) planning and execution, as well as natural language generation (developed by SIFT, green), and (3) domain-specific reasoning and visualization (blue, developed by HMS, OHSU and Tufts) (Figure 1).

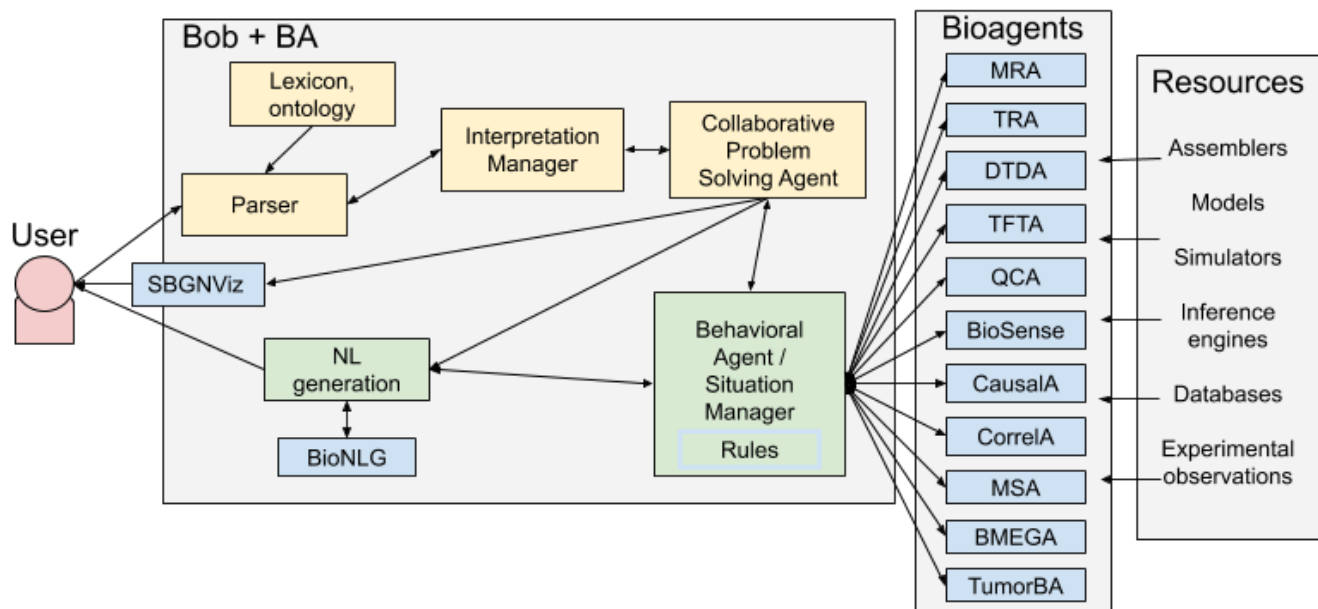


Figure 1. The architecture of BwB consists of agents that communicate with each other via messaging. Agents in yellow are responsible for natural language understanding and dialogue management, while agents in green provide a framework for defining the domain-specific execution of tasks and natural language generation. Domain-specific problem-solving capabilities and access to resources are implemented by Bioagents (shown in blue) along with other domain-specific components.

Natural language understanding is provided by a combination of a parser, which has access to lexicons and an ontology (covering general English language and specialized concepts relevant for biology), and an interpretation manager, which prioritizes the user's intent based on the meaning of an utterance and the dialogue state. The Collaborative Problem-Solving Agent (CPSA) handles the internal creation and

evolution of goals and subgoals during the dialogue, which serves as the basis for execution and problem-solving by the system.

Goals recognized by the CPSA (ranging from simple questions to be answered to overarching goals such as building a model of a given type of cancer) are passed for evaluation and execution to the Behavioral Agent (BA). The BA is aware of the domain-specific capabilities of the system and can decide whether the goal should be pursued (in which case it orchestrates an execution plan) or abandoned. The generic BA framework (Figure 1, green) takes its domain-specific behavior from a set of rules that recognize patterns in a representation of the current dialogue state and the incoming goal to decide what actions to take.

All of the domain-specific problem-solving capabilities of the system are implemented via dedicated reasoning agents called Bioagents (Figure 1, blue; available at <https://github.com/sorgerlab/bioagents>). These agents take requests from the BA to solve given problems and return solutions and suggestions to the BA. The BA communicates the state and result of problem-solving to the CPSA, and the system then responds to the user via natural language.

The system also displays multiple views of its model and reasoning visually through a multi-tabbed graphical interface called SBGNViz. A snapshot of an example dialogue with BwB on this visual platform is shown in Figure 2.

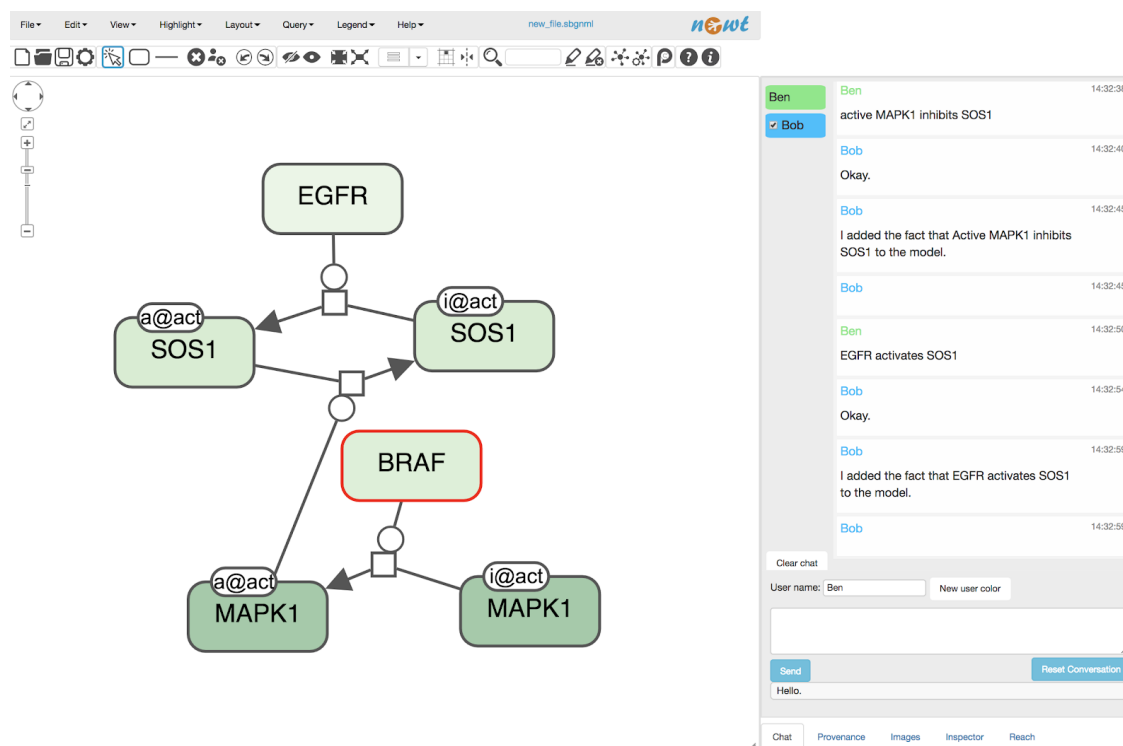


Figure 2. The integrated BwB system integrated with the SBGNViz interface, conducting a dialogue with a user about biochemical mechanisms governing a molecular system.

1.1.2. Bioagents: domain specific problem solving

Our teams developed a number of agents that plug into the dialogue system and expose a task-oriented interface to solve a given set of domain-specific problems. Here we provide a brief description of each Bioagent and its capabilities. Overall, agents cover a broad range of capabilities providing

access to mechanistic knowledge and data (MSA, DTDA, Tumor Board Agent, BMEG Agent, Biosense Agent) and model-based reasoning (MRA, TRA, QCA, CausalA).

Mechanistic Reasoning Agent (MRA)

The MRA is responsible for turning descriptions (i.e., prior knowledge or novel hypotheses) of biological systems expressed by the user during dialogue into mathematical models. To achieve this, it wraps the Integrated Network and Dynamical Reasoning Assembler (INDRA) system which is responsible for the actual conversion of interpretations of mechanistic descriptions into model elements, and combines this with a memory data structure that represents different stages of a model over the course of the dialogue. The MRA's memory data structure allows for incremental modeling (i.e., adding a new element to an existing model), removing specific elements or undoing previous actions, describing the content of the current model, and upon request, providing the content of the model to the BA or other Bioagents.

Importantly, the MRA has the ability to register a context-specific user goal (e.g., to explain "How does EGFR result in ERK activation in BT20 cells?"), and monitor whether the model structurally satisfies the goal. To achieve this, the MRA performs automated model diagnosis and can offer suggestions to the user both for completing the model based on background knowledge and to improve the model by adding missing assumptions.

Temporal Reasoning Agent (TRA)

The TRA provides dynamical simulation and analysis for models assembled by the MRA. The mechanistic statements constituting the model are first compiled into a rule-based formalism and then converted either to a set of ordinary differential equations (ODEs) using BioNetGen or into a network-free simulation model using Kappa. The former allows for deterministic, continuous-time/state simulation while the latter uses stochastic, continuous-time but discrete-state simulation. The Kappa simulation uses software developed in the course of this project by the HMS-Fontana and ENS-Lyon teams.

Each model simulation can be specified with a readout (e.g., the amount of phosphorylated ERK) and optionally a perturbation condition (e.g., increase the amount of Vemurafenib by 10-fold). In addition, the TRA can automatically analyze the outcome of the simulation and determine whether a given qualitative temporal behavior (specified as a Bounded-time Linear Temporal Logic (BLTL) property) is satisfied. Some of the supported temporal patterns include: always high/low, high/low at some point in time, eventually high/low, or transient (i.e., high at some point, then eventually low). Overall, this allows for determining whether e.g., "If we increase the amount of Vemurafenib by 10-fold, will the amount of phosphorylated ERK be transient?).

Mechanism Search Agent (MSA)

One of the important achievements of this project is the creation of the INDRA DB which makes available knowledge extracted automatically from all available biomedical literature in a queryable form. The MSA connects the dialogue system to the INDRA DB by exposing an appropriate interface for accessing DB content. The MSA supports the following query patterns:

- Neighborhood, e.g., "What does BRAF interact with?"
- From source, e.g., "What does KRAS activate?"
- To target, e.g., "What regulates MDM2?"
- Binary directed, e.g., "How does EGFR regulate SOS1?"
- Binary undirected, e.g., "How do EGFR and GRB2 interact?"
- Common upstreams/downstreams, e.g., "What are the common upstreams of ERK and AKT?"

For each of these query patterns, the MSA supports constraining the type of interaction/regulation (activation, inhibition, binding, phosphorylation, etc.).

The response from the MSA to each query is a ranked list of entities and statements (along with supporting evidence grounded in the scientific literature) that can be turned into a natural language response and also interactively browsable web pages to which the dialogue system can link.

Drug-Target-Disease Agent (DTDA)

The DTDA makes available resources for reasoning over drugs and their targets, and relationships between targets and diseases. The DTDA cross-references and combines multiple data sets including drug-binding assay data from the Small Molecule Suite and cancer genomics data derived from The Cancer Genome Atlas data, available via cBioPortal.

Qualitative Causal-path Agent (QCA) and Causality Agent (CausalA)

Both the QCA (developed by the UCSD team) and the CausalA (developed by the OHSU team) aim to find causal paths between a source and a target entity. The QCA can take any causal network on the Network Data Exchange (NDEX) assembled by INDRA and return ranked paths between a given source and target in that network. In contrast, the CausalA can itself propose source-target pairs of interest, derived from high-throughput experimental data between which an observed correlation can be mechanistically explained. It then uses path finding on a network derived from PathwayCommons to return a possible explanation. In some circumstances, it can also surface questions to the user if a correlation cannot be explained based on the built-in network.

BMEG Agent (BMEGA)

The BMEGA provides an interface to the Biomedical Evidence Graph, a knowledge graph that combines multiple experimental datasets in a graph data structure. The BMEGA exposes mutation frequency statistics from cancer genomics data and can propose drugs that are statistically shown to be effective in cancer cell lines in which that mutation occurs. It can also generate visualizations of these statistics to complement natural language responses.

BioSense Agent

The BioSense Agent serves multiple purposes, first, it has an important technical role in converting representations produced by the TRIPS interpretation components into INDRA Agent / Statement representations that are shared between the BA and Bioagents. Second, it exposes capabilities related to ontological properties of entities: return types (e.g., small molecule, kinase, transcription factor) for entities of interest, provide longer descriptions for entities, and determine members of protein families and complexes.

1.1.3. Front-ends

We developed two complementary front-ends for the Bob with Bioagents system. The CLiC interface is text-oriented and allows the user to upload and interact with their own data. The SBGNViz interface contains a large and interactive model canvas and thus allows for more detailed visualization in addition to the textual dialogue.

CLiC interface

The CLiC interface contains a full-width chat box on the top half of the interface putting textual dialogue in primary focus. On the bottom half, multiple tabs allow for examination of additional visual (static images) and structured (tables) information generated by the system during dialogue. Importantly, the CLiC interface allows the user to upload two types of data: a list of genes obtained from an experiment or a gene expression matrix obtained from multiple experimental conditions. The interface integrates the Maayan lab's clustergrammer tool (<https://github.com/MaayanLab/clustergrammer>) to display easy-

to-navigate heat maps of gene expression data, and allows the user to select subsets of the data for discussion. Whether starting with a list of uploaded genes, or selecting genes from the gene expression data, the system the user can refer to the genes in dialogue. Overall, this allows the user to discuss their data set of interest in the context of all available background knowledge provided by the dialogue system.



Figure 3. The integrated BwB system using the CLiC interface. In the bottom, the data upload and clustergrammer view for gene expression data is shown.

SBGNViz interface

The SBGNViz interface (see Figure 2 above), developed by the OHSU team, consists of a large and interactive model canvas next to a chat box containing the textual dialogue. The model canvas can accommodate various model-related and static or interactive visual content on multiple tabs; new tabs open dynamically during dialogue to display model analysis or background data. Similarly, the chat box area consists of multiple tabs and can display dynamically generated HTML content (tables, etc.) to provide provenance for responses.

1.1.4. Public deployment

Deploying the integrated system publicly required substantial engineering effort. First, we created a repository that integrates a total of 43 codebases from the developer teams, all needed to run the dialogue system. Next, we developed a containerization scheme for setting up all the dependencies required by these repositories (e.g., Python, Ruby, Java, C/C++, Lisp, node.JS are all required for running various components), cloning the integrated repository, and setting up launch scripts to start the system. Having containerized the system, it became possible to run multiple instances of the system on a single machine, each exposing a different port and thus allowing independent dialogue sessions for multiple users. Consequently, we developed a web application which can manage user sessions and can spin up containers on demand and assign one to each active user. The application also has to monitor sessions for activity and shut down containers that are idle for a given amount of time.

Another important aspect of public deployment is user identification, collecting user feedback, and saving session logs. When launching a dialogue session, the user enters their name and email address, and (in both user interfaces) can submit feedback at any point in the dialogue. User information, feedback, as well as all the logs from the dialogue are saved in a database and persisted on AWS S3. We also developed a separate log browser service that allows developer teams to browse the dialogue logs and identify issues that need to be fixed to improve the system.

The system is publicly available at www.dialogue.bio where users can launch sessions on-demand using either the SBGNViz or the CLiC interface (see Figure 4). The website also makes available tutorials and demos.

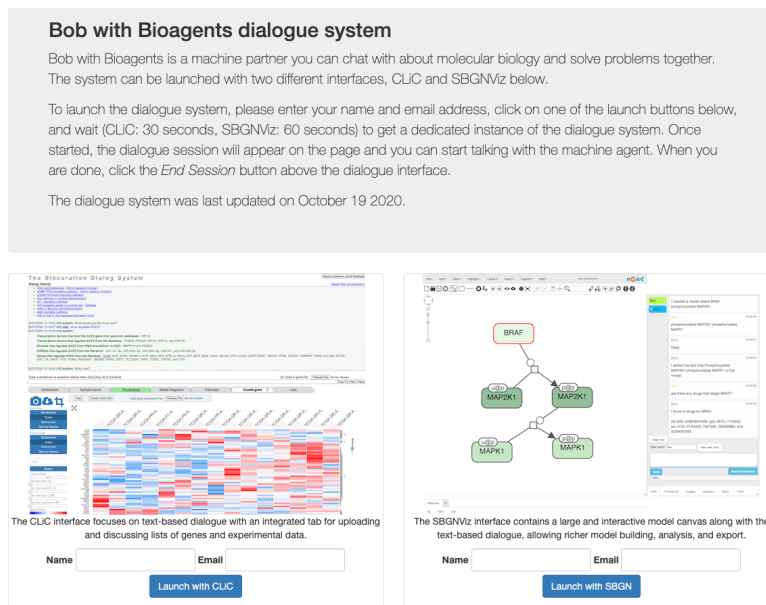


Figure 4. The *dialogue.bio* web interface for launching dialogue sessions with Bob with Bioagents on demand.

1.1.5. System evaluation

The repeated empirical testing and evaluation of the system was crucial in our development process. We set up a suite of automated unit tests (for each component of the system) and a set of integration tests to systematically test the system on end-to-end dialogues. These tests ensured that the system remained operational despite the distributed and asynchronous development process across teams.

We also worked closely with the MITRE team to perform technical evaluation of the system. Evaluations typically involved feeding a series of questions to the system, and observing the system's responses for technical correctness. Another approach used by the MITRE team involved solving pre-defined tasks with the system and evaluating its responsiveness and technical correctness. Notably, MITRE introduced a "hallmarks" framework to evaluate the system along several properties and helped us develop specific aspects of the system that required improvement.

During the course of the program, we organized multiple user evaluations with medical school scientists not involved in system development. We conceived and conducted a total of four, increasingly sophisticated user evaluations between 2017 and 2020. At a high level, the goals of these evaluations were as follows:

- 2017 April: solve pre-defined model building tasks through dialogue

- 2017 October: solve a pre-defined model building task, and construct explanations to two pre-defined experimental observation through dialogue
- 2019 January: construct explanations to three experimental observations in different modes (users randomized across tasks/modes): alone through dialogue, with a “wizard” to call on for help if needed but otherwise working alone through dialogue, or with any other available tool outside the dialogue system
- 2020 July: bring your own real experimental data and scientific questions and try to make progress on these with the dialogue system

Details of these evaluations and their results can be found in our annual reports. Here we highlight key results of the final, 2020 July evaluation. This evaluation went significantly beyond our prior evaluations by asking scientists to work on their own ongoing research and discuss real experimental data. To this end, we recruited four scientists from Harvard Medical School and one from Tufts Medical School, and asked them to use the system (each user launched the publicly available system at dialogue.bio on their own) for 45-60 mins to discuss their individual research. Users worked on a diverse set of research topics (cancer, Alzheimer’s, COVID-19, etc.) and biological entity types (proteins, drugs, micro-RNAs, processes, etc.), demonstrating that the system is capable of covering a broad set of use cases. Specifically, the users worked on the following research questions:

- Analyze differentially expressed genes obtained from SARS-CoV-2-infected lung samples from primates (HMS Postdoc)
- Investigate the role of ULK1 in Alzheimer’s disease and possible therapeutics (HMS Research Scientist)
- Find mechanisms explaining the upregulation of specific genes in drug-resistant melanoma cells (HMS Postdoc).
- Learn about deubiquitinating enzymes as druggable targets for cancer (HMS graduate student)
- Learn about micro-RNA regulation of bone density (Tufts graduate student)

After talking with the system, each user filled out a survey that included open-ended questions about their experience. Some highlights from our analysis of the logs, screen recordings and surveys are:

- All the users used real data from ongoing research projects and
 - Uploaded one or more differentially expressed gene lists (4 users)
 - Uploaded gene-expression matrix (1 user)
 - Used the system to explain specific perturbation-readout effects observed in experiments (1 user)
- All the users reported having learned something new
- Multiple users highlighted the system’s unique ability to refer to prior results and refer to lists of entities sequentially
- Multiple users highlighted the system’s ability to find previously (to them) unknown relations and point to relevant literature
- Multiple users reported finding specific new hypotheses to pursue next
- Two users returned later to run more dialogues about their research

1.2. The CLARE dialogue system

The key motivation behind building the CLARE system was to create an alternative dialogue architecture to expose Bioagents capabilities that is easier to develop, extend, and deploy as compared to Bob with Bioagents which is built on the TRIPS platform. Further, we sought to support communication features that are difficult to engineer in the TRIPS platform, such as fully integrated

coreference resolution (i.e., support for follow-up questions) or relevant suggestions in cases where the system did not understand the user.

1.2.1. System architecture

CLARE leverages domain-specific (back-end) problem-solving capabilities already implemented as Bioagents, but with a simpler dialogue-management approach compared to the agent architecture of Bob with Bioagents.

CLARE implements a number of Capabilities that are each triggered during dialogue when a given set of associated Templates are matched. These Templates can be defined either as simple regular expression patterns on user utterances, or as graph patterns on TRIPS interpretation graphs. In this system, TRIPS is simply used as a third-party NLP service rather than an underlying communication platform, which takes natural language as input and returns an interpretation graph as output. In general, CLARE supports multiple levels of Capability-matching by multiple levels of complexity on NLP, ranging from Template-based matches using regular expressions to semantic interpretation by TRIPS, with syntactic dependency parsing as a (yet unused) intermediate layer. Overall, this architecture allowed us to implement the CLARE dialogue system as a single Python package, thereby leading to an easily extensible, testable, and deployable system. We describe the basic building blocks of CLARE below:

- **Capabilities** describe a set of pre-conditions represented as Templates, an Action that is to be performed if one of the Templates matches, and a Response function that is applied to the result of the Action to say or display information to the user.
- **Templates** can be defined as surface regular expressions applied directly to the natural language input from the user, or as graph patterns applied to the interpretation of the user input as produced by one of the TRIPS variants (Bob, CWMS, Musica, etc.). Each Template can define a list of examples, in which each example is a sentence matched by the Template. This makes automated testing of Capabilities easier, since it allows tests to aggregate all examples defined in the Templates for Capability testing.
 - **Regular expression Templates** (ReTemplates) contain a pattern string, which matches zero or more groups in the text, and a list of match names, which are keywords assigned to each group in the match.
 - **Graph Templates** assume that a given text has been interpreted by TRIPS, and provide a way to match patterns and extract arguments from the interpretation graph, which are represented as a KQMLGraph object (built on a standard NetworkX DiGraph). The KQMLGraph API exposes several methods to traverse the graph, typically starting with `get_root()`. From the root, one can traverse the graph with two sets of methods. One set always returns a response, but can return 'None' if no matching pattern is found, making it necessary to explicitly handle 'None' responses at each pattern-matching step. The other set raises a 'PatternNotFound' error when a matching pattern is not found, resulting in cleaner Template definitions with a single try/except block to determine an overall match.
- Each Capability needs to define a *do* method that takes the return value for associated Templates as input, performs an Action, and returns a data structure representing the results of that **Action**. Typically, a specific Bioagent is invoked via its Python API to perform a task and return the result.
- Each Capability defines a *respond* method to take the return value produced by the do action and generate a **Response**. The Response can be as simple as a string sentence, but can also contain visualizations.
- Each Capability can require values from the global **Context**, and can also store values in the global Context. In the example below, a Capability defines the `need_context` attribute to expose

that it needs the model entry in the global Context to perform its Action. The do action of each Capability can also set an entry in the global Context.

A simplified diagram of the CLARE system's architecture is shown in Figure 5.

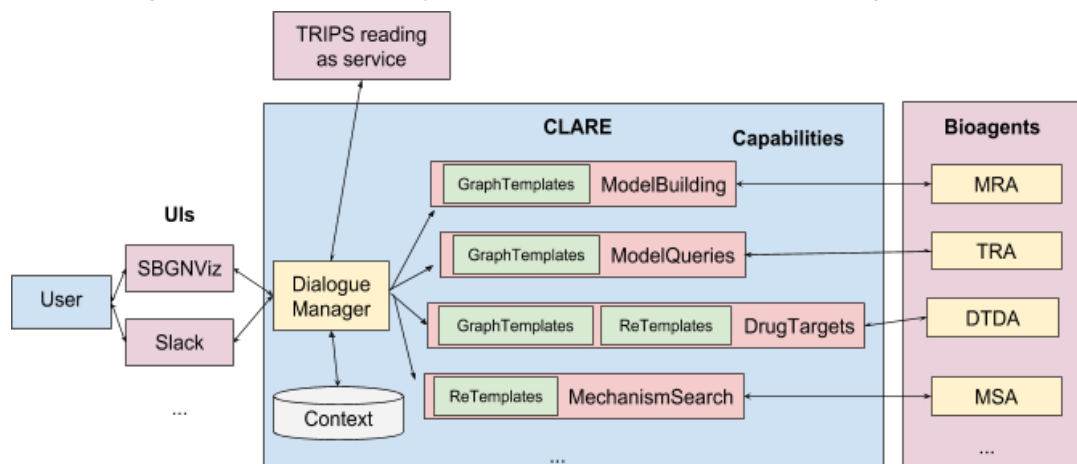


Figure 5. Simplified architecture of the CLARE dialogue system.

A key advantage of this architecture is that it runs a single integrated process for the entire system while being able to instantiate user-specific instances of the Dialogue Manager within this process. This allows for the stateless components of the system to be singular while the stateful components can be user/session-specific, resulting in optimal resource usage. These properties make the CLARE system significantly easier to deploy and scale up to a large number of simultaneous sessions – one of the key limitations of the Bob with Bioagents system.

The session management strategy for handling multiple user sessions is shown in Figure 6.

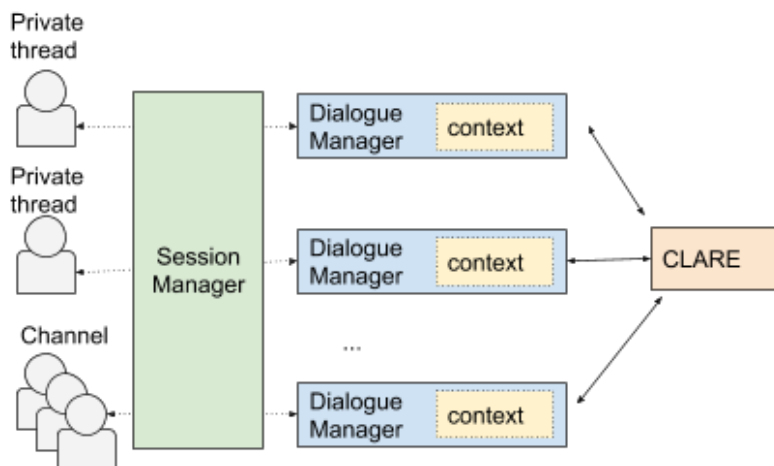


Figure 6. Session management in CLARE. A Session Manager is responsible for creating and assigning Dialogue Managers corresponding to various dialogue sessions. Each Dialogue Manager has its own contextual memory making it user-specific.

1.2.2. Learning new capabilities

An important limitation of natural language dialogue system is their ability to adapt to variability in language from users. For instance, the system might be able to interpret and correctly determine how to answer “what is the mutation frequency of BRAF in melanoma” but not “how frequently is BRAF

mutated in melanoma?”. The reason behind this is that the machine-constructed interpretation graph behind the two sentences is different, and one of the structures may not be handled by the dialogue system’s logic responsible for choosing the next action. Adapting to such variations would normally require going into the dialogue system’s software code and changing its logic.

In the context of the CLARE system, we aimed to address these challenges by implementing the AutoPattern functionality to automate the learning of patterns to recognize interpretation graphs based on annotated example sentences. For example, for the sentence “how frequently is BRAF mutated in melanoma?” it is sufficient to annotate that in this sentence “BRAF” has the semantic role of the “gene” and “melanoma” has the semantic role of the “disease” for calling on the FindMutationFrequency capability. The system can then internally construct the interpretation graph for the sentence, find the nodes corresponding to “BRAF” and “melanoma”, and derive the graph pattern that describes their relative position. From then on, the system can recognize other sentences with a matching interpretation graph structure such as “with what frequency is EGFR mutated in glioblastoma?”.

We refactored CLARE to take AutoPatterns provided in a simple human-readable and -editable JSON file, making it possible to improve the system without any coding. We also exposed AutoPattern construction during dialogue runtime as shown in Figure 7. Here, a new AutoPattern is defined as part of the dialogue and thereafter the system can interpret and correctly answer a question that it previously did not understand.

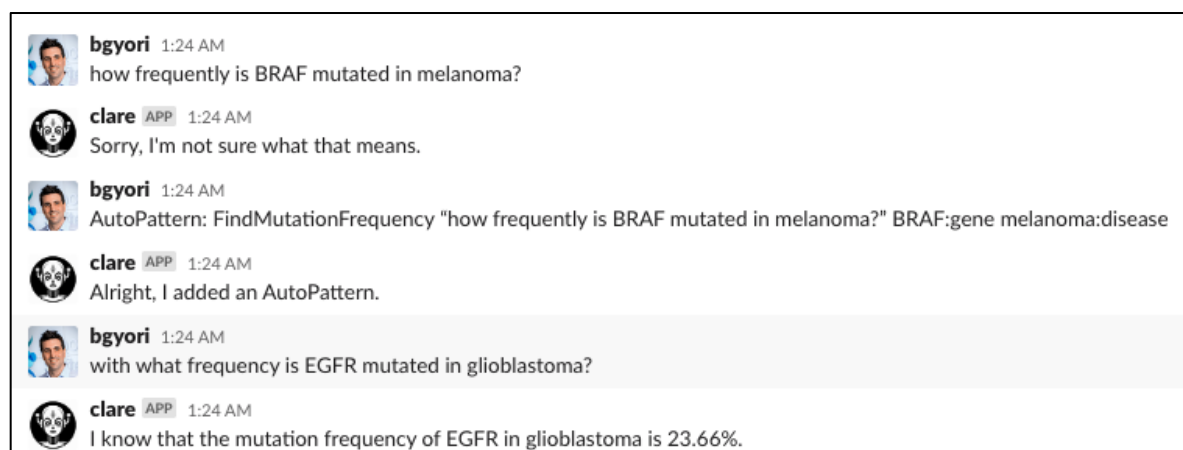


Figure 7. Teaching the CLARE system during runtime using AutoPatterns.

1.2.3. Public deployment

We integrated CLARE with multiple front-ends and deployed it in several ways. First, we implemented a Slack adapted for CLARE which allows it to be installed as an application in Slack workspaces. There, it can either be messaged directly or added to public channels. Since 2019, we added CLARE to a total of 8 Slack workspaces:

- Harvard Program in Therapeutic Science
- DARPA Panacea program
- DARPA COVID literature-mining collaboration
- Harris lab at UMass Medical School
- Ideker lab at UCSD
- Quantitative Biomedical Research Center at UTSW
- CoronaWhy project
- COVID-19 Disease Map project

The start of the COVID-19 pandemic in early 2020 led to an important integration with the COVID-19 Disease Map community. In Figure 8, we quantified the number of questions asked by different users in this community. We found that a total of 17 users interacted with CLARE with 11 of the users asking at least 10 questions and the top user asking more than 150 questions. By helping curators find mechanistic knowledge from the COVID-19 literature, this integration has contributed to a major publication by the COVID-19 Disease Map community: <https://www.embopress.org/doi/full/10.15252/msb.202110387>.

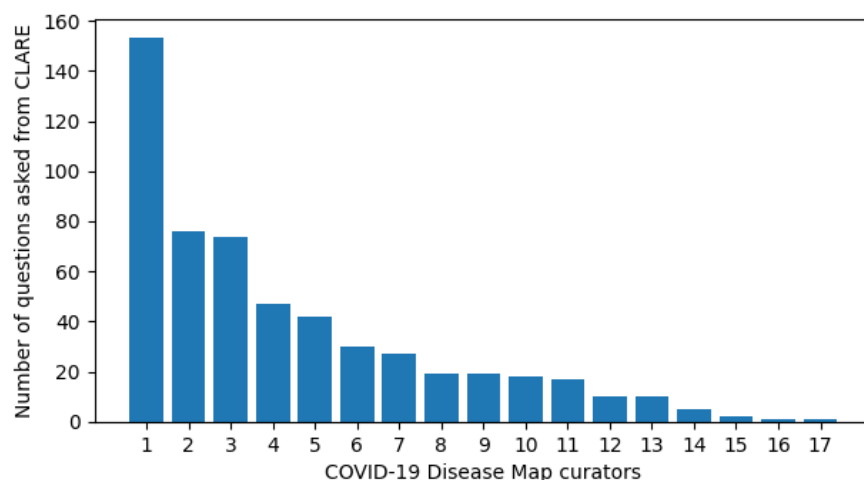


Figure 8. Usage statistics of CLARE in the COVID-19 Disease Map Slack workspace.

We also integrated CLARE with the Ecosystem of Machine-maintained Models with Automated Analysis (EMMAA, emmaa.indra.bio). In this context, CLARE can conduct dialogue with respect to any of the self-updating EMMAA models; answers are based on the specific model's content. Figure 9 shows an example dialogue using the COVID-19 EMMAA model.

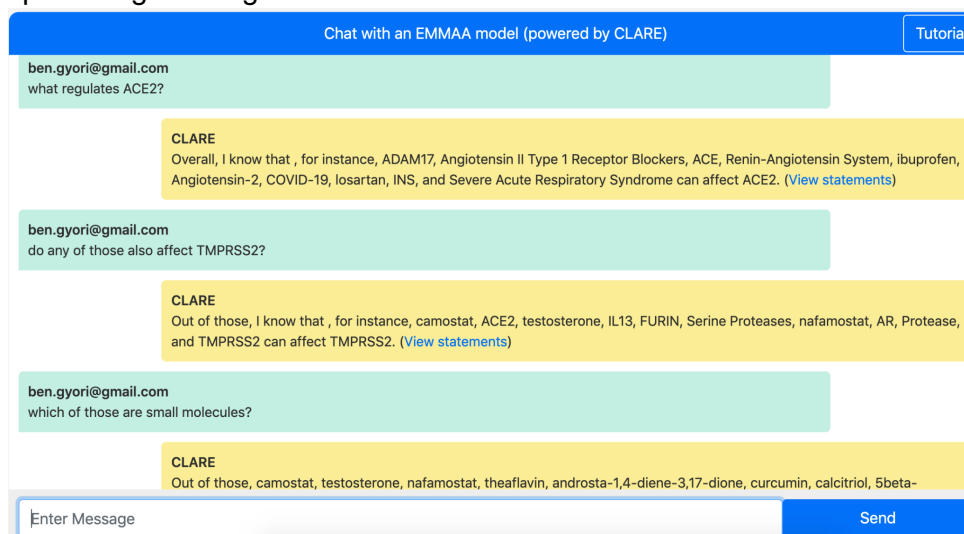


Figure 9. Example dialogue with CLARE on the EMMAA website. Here, CLARE answers questions based on knowledge captured in the EMMAA COVID-19 model.

Finally, we integrated CLARE with the SBGNViz interface that we introduced as one of the UIs for the Bob with Bioagents system. Here, CLARE makes use of the large model canvas to display relevant interactions. This integration is used for the Tumor Board use case which we describe in the Applications section.

1.2.4. System evaluation

Similar to Bob with Bioagents, CLARE was evaluated multiple times over the course of our project. Evaluations included both task-oriented tests where representatives of the MITRE team attempted to solve specific problems via human-machine dialogue, and hallmarks-oriented evaluations in which a set of pre-defined sentences were used in the dialogue to assess the communicative capabilities of each system. Here we present details of the task-based evaluation of the CLARE system for which we made the CLARE system available via the SBGNViz interface. The following two tasks were attempted via dialogue:

- Three of the top upregulated genes in SDHB null pheochromocytoma (as compared to normal adrenal tissue) are FN1, MMP3, and SERPINE1. Identify a drug that would likely inhibit the expression of all three of these genes simultaneously.
- Inhibition of the PI3K signaling pathway has been observed after treatment with Niclosamide or Pyrvinium Pamoate. Identify a mechanism that explains this observation.

Solving both tasks involved talking with the system about potential mechanisms that could result in the observed outcome, then formulating a hypothesis (as a model), and finally confirming, via model simulation, that the model reproduces the observation as expected. The MITRE team performed a “stoplight” analysis in which each round of dialogue was scored according to whether the dialogue – and thus, solving the task – progressed productively or got stuck. In both dialogues, the information discovery phase proceeded successfully, while parts of model building in the first task as well as simulation using natural language proved problematic. In some cases, questions used language that the system could not readily interpret, e.g., “Is the amount of PI3K ever high if the amount of niclosamide increased 10 fold?” (many of the relevant interpretations have since been generalized to cover the problematic phrases). However, ultimately, the model could reproduce the desired observation, as revealed by asking “does Niclosamide decrease PI3K in the model?” at the end of the dialogue.

2. Component technologies

In this section, we describe technologies that we developed over the course of this project that have utility as standalone software and can be integrated or applied in settings distinct from their role in the integrated human-machine dialogue system described in the previous section.

2.1 INDRA

INDRA is a knowledge- and model-assembly system that can combine mechanisms extracted from the scientific literature, pathway databases, and natural language expert input. During this project, we made numerous extensions to modules of INDRA:

- Adding new knowledge sources to INDRA including text mining systems and pathway databases
- Improving ontological resources in INDRA, for example those for the normalization of chemical entities between ontologies such as ChEBI, PubChem and ChEMBL
- Automated assembly of PySB, SBGN, and Cytoscape JS models

We also made significant developments in the INDRA Belief Engine which is responsible for assessing evidence and assigning belief scores to statements. We generated a large corpus of curation data (statements curated as correct or incorrect with respect to the supporting evidence) and then developed a number of machine learning approaches to predict probability of correctness given a set of statement features. We have trained these models on our corpus of 1,330 curated statements sampled from text mining a corpus of 580,000 articles with five readers (Reach, Sparser, Medscan, RLIMS-P, and TRIPS). Precision-recall curves resulting from fitting five different model types are shown in Figure 10.

Though performance measures vary with the sets of features used, the way that the dataset is weighted, and whether information from curated pathway databases is included, the relative performance of different classifiers was consistent. Log-logistic regression and random forests consistently performed best in terms of AUPRC, with results of 0.926 and 0.913, respectively. We also found that augmenting the feature set to include not only the counts of evidence from various sources but also the statement type (e.g., Activation, Phosphorylation, Complex, etc.) and the total number of unique supporting articles (PMIDs) invariably improved performance for all classifiers tested.

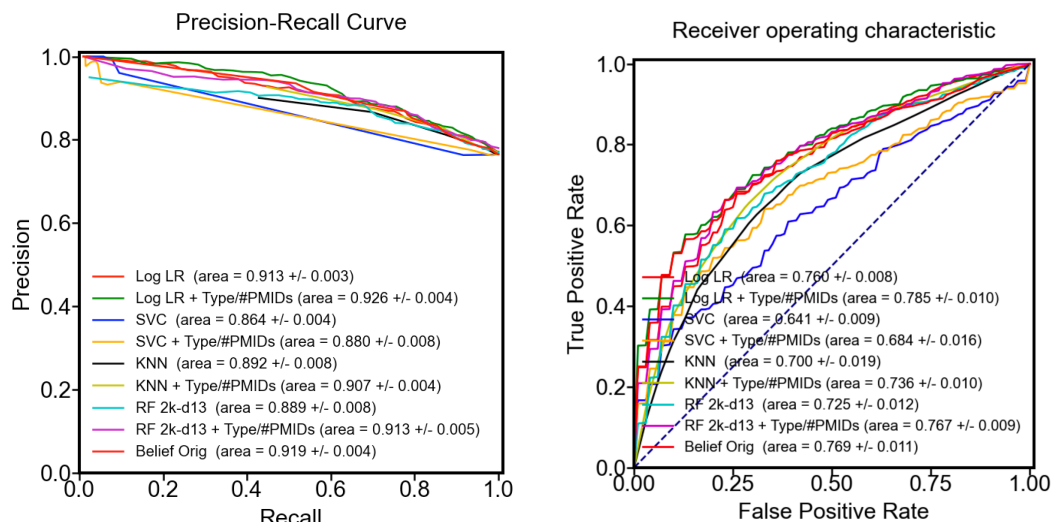


Figure 10: Performance of different classifiers in prediction statement correctness from assembled evidences. LR: logistic regression; RF 2k-d13: random forest classifier with 2,000 trees and a maximum depth of 13; SVC: support vector classifier; KNN: K-nearest neighbors classifier; Type/#PMIDs: indicates that the classifier uses not only evidence counts but also information on statement type and the number of unique supporting PMIDs. $\pm$ values indicate the standard errors of the mean areas under the curve taken across 10 stratified cross-validation folds.

To address a potential lack of curation data as INDRA is extended to new sources, we also developed a **Hybrid Belief Scorer** that can combine machine-learned models with explicit probabilistic models based on prior assumptions about a given source.

2.2. INDRA Database

One of the important features of the dialogue systems is the ability to expose mechanistic knowledge extracted automatically from the scientific literature. Our goal was to create the INDRA Database (DB), a standalone resource that can be queried as a service to obtain such statements.

2.2.1. Architecture and content

We implemented a software infrastructure to run reading systems at scale (using cloud computing resources on Amazon Web Services) on all accessible scientific literature and to obtain and process new literature as it is published. The outputs of these reading systems are stored and processed into INDRA Statements. The Statements then go through processing and assembly in which compatible Statements and their evidences are aggregated and refinement relationships between Statements are reconstructed. The assembled Statements, their Evidences, and Agent arguments (representing proteins, small molecules, etc.) are then stored in a relational database using a structure optimized for querying by Agent grounding. Figure 11 shows the overall concept behind the INDRA DB.

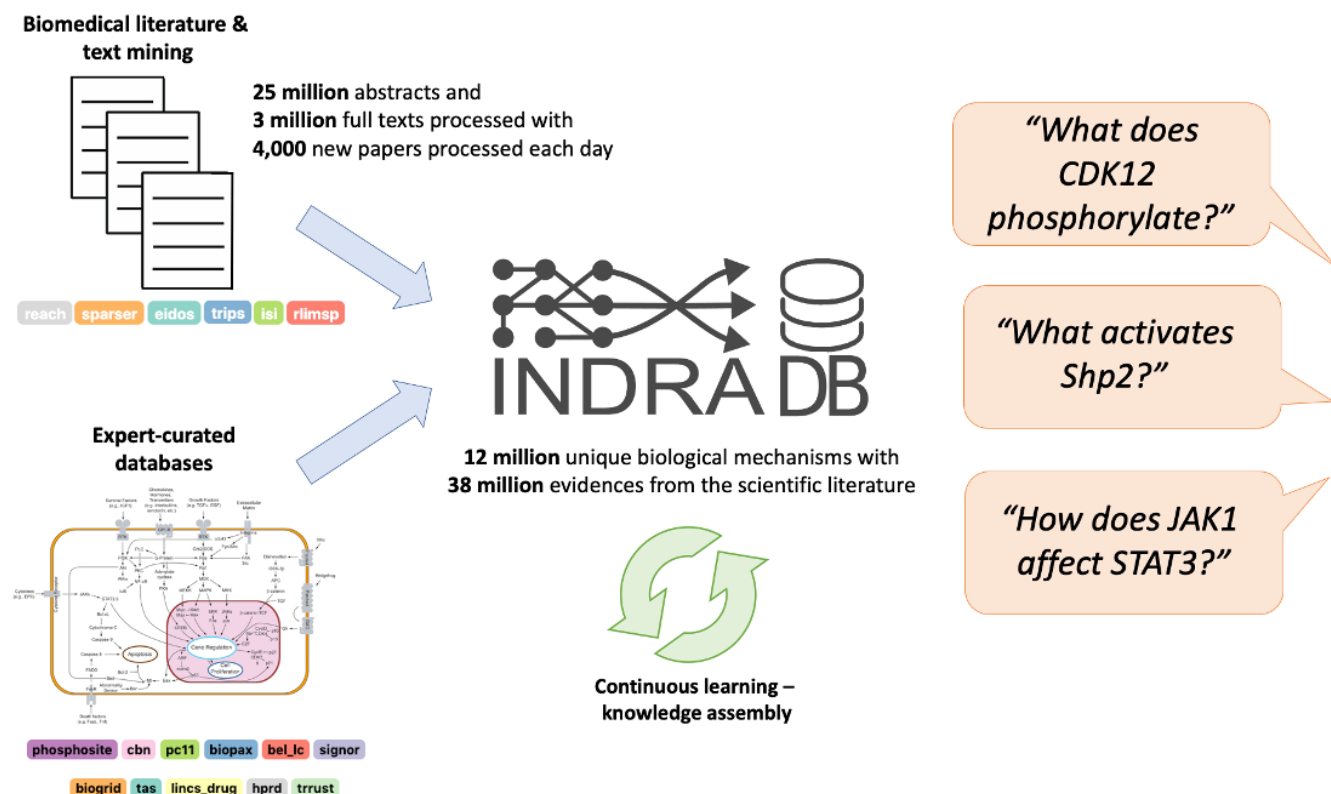


Figure 11. A high-level conceptual overview of the INDRA DB.

To date, the INDRA DB has processed over 25 million abstracts and 3 million full text articles, with extractions from around 4 thousand papers added to the DB each day. In addition, the DB integrated the content of 11 structured databases, including all widely used molecular pathway knowledge bases. The assembled Statement set represents 12 million unique mechanisms out of which around 8 million contain fully grounded (i.e., identified with respect to an ontology) Agents. These Statements are supported by 38 million pieces of evidence, each referring to a specific publication, evidence sentence, and further metadata. To our knowledge, this is the largest such resource that currently exists.

2.2.2. Querying the INDRA DB

We implemented a REST API service and a web UI front-end around the INDRA DB to allow queries, available at <https://db.indra.bio>. Queries to the DB can be constrained in several ways: by the agents to search for, the type of Statement, specific papers, as well as Medical Subject Heading (MeSH) terms. The MeSH constraint option is particularly significant as it allows for selection of statements and their evidences specific to a given context (e.g., breast cancer or diabetes).

The UI for the INDRA DB displays responses to queries in an interactive way. Groups of Statements are organized into headings (e.g., statements between ACE2 and TMPRSS2 appear under a single heading under which statements of a specific type between ACE2 and TMPRSS2 are listed). A set of badges showing evidence counts from different sources is displayed at each level, allowing users to quickly grasp how much evidence there is for a given statement and the sources of those evidences. Under the lowest-level heading, specific evidences are displayed in separate rows. Each evidence provides a sentence from the original publication supporting it and links out to that publication.

Importantly, the INDRA DB UI also allows for curation of Statements for correctness. Next to each evidence, there is a pencil icon that can be clicked to reveal a simple curation form in which users can select an error type and submit the curation. Curations are stored in the INDRA DB and can be used to filter results in downstream analysis. The INDRA DB statement browser UI is shown in Figure 12.

Results
I found statements that are from papers with MeSH ID C000657245, are not only from medscan, and have an agent where HGNC=11876 with role=OBJECT

ERG affects TMPRSS2
ACE2 affects TMPRSS2
camostat affects TMPRSS2 3
Camostat inhibits TMPRSS2 3
Camostat inhibits TMPRSS2 3

reach However, the mechanism of inhibition of **TMPRSS2** by **Camostat** mesylate, per se, has not been clearly elucidated. 32469279

eidos Next, using the established assay in 384-well format, we tested the inhibition of **TMPRSS2** by **camostat**, nafamostat and gabexate (Figure 3). 32596694

reach **Camostat** mesylate (a serine protease inhibitor), well documented for the treatment of cancer, pancreatitis and liver fibrosis, has recently been shown to inhibit **TMPRSS2** and entry of SARS-CoV-2 virus into the cells (Hoffmann et al., 2020). 32469279

eidos It was reported the ten-fold reduction of SARS-CoV titers in the Calu-3 cells with the inhibition of **TMPRSS2** by **camostat** (Shulla et al. 2011). 32546018

Select error type... Optional description (240 chars) Submit

Prior Curations

7/4/2020, 5:18:20 PM	ben.gyori@gmail.com	correct	No text given.	EMMAA
----------------------	---------------------	---------	----------------	-------

Our Sources:

Databases

- Causal Bionet
- PathwayCommons
- CTD
- VirHostNet
- Phosphosite Plus
- HPRD
- Target Affinity Spectrum
- Phospho.ELM
- Signor
- BioGRID
- DrugBank
- TRRUST
- BEL

Reading

- ISI/AMR
- TRIPL/DRUM
- Sparsen
- REACH
- HLIMS-P
- Eidos
- MedScan

Figure 12. INDRA DB Statement browser UI.

2.3. PyBioPAX

BioPAX is a widely used exchange format for pathway information. However, due to its generality and ability to represent complex biochemical mechanisms, it is challenging to use without a dedicated software library. PaxTools is a Java package facilitating the use of BioPAX models, however, PaxTools cannot be used directly in a Python setting.

We therefore developed PyBioPAX, a Python software package to process and manipulate BioPAX models. PyBioPAX implements the BioPAX Level 3 object model as a set of Python classes, and implements a BioPAX OWL processor to deserialize BioPAX content from OWL files or strings into these objects. Once a BioPAX model and all its linked elements are deserialized into Python objects, they can be traversed and modified in memory. PyBioPAX supports serialization of BioPAX Models into OWL/XML files compatible with other tools in the BioPAX ecosystem.

PyBioPAX implements the BioPAX OWL semantics whereby object attributes can be subtyped (e.g., "display name" is a subtype of "name") using Python property attributes and getter/setter functions. It also supports exposure of "inverse links" between objects; for example, a BioPAX Xref object, which represents a cross-reference exposes a list of `xref\_of` links back to the objects of which it is a cross-reference. Again, the coherence of these links at the level of a BioPAX Model is guaranteed through the use of Python property attributes. The inverse links contribute to the efficient traversal of BioPAX Models by allowing to link from e.g., one participant of a reaction to the reaction itself and its other participants.

PyBioPAX also implements a client to the PathwayCommons web service. This client makes available three different graph query types: paths-from-to, paths-between, and neighborhood to extract subsets of knowledge aggregated from structured sources in PathwayCommons (which include e.g., Reactome, KEGG, and NCI-PID) as BioPAX models. PyBioPAX further provides web service clients for processing

BioPAX content from other pathway databases including NetPath, HumanCyc, BioCyc, MetaCyc, and EcoCyc.

PyBioPAX is available as a package on PyPI with the source code available at <https://github.com/indralab/pybiopax>, and documentation available at <https://pybiopax.readthedocs.io/>. The repository also contains an interactive Jupyter notebook tutorial. PyBioPAX has been integrated into INDRA and serves as the primary entry point for processing BioPAX content into INDRA Statements through the traversal of a BioPAX Model.

2.4. Gilda

Gilda is a software tool and web service that implements a scored string-matching algorithm for names and synonyms across entries in biomedical ontologies covering genes, proteins (and their families and complexes), small molecules, biological processes and diseases. This capability is useful as a standalone tool, but is also a key component of the CLARE dialogue system, responsible for grounding named entities mentioned during the dialogue.

Gilda integrates over 1.8 million names and synonyms from 9 ontologies and lexical resources for the purpose of name matching. In addition to standard resources such as UniProt, Gilda integrates FamPlex, a resource providing lexicalizations for protein families and complexes (shown to be a key missing component in named entity normalization systems) as well as over two thousand lexicalizations for entities that were manually curated using a frequency-ranked list of strings mined from the literature that did not appear in other standard resources.

For a given entity string, the grounding process first generates relevant variants (e.g., by spelling out Greek letters) and then identifies possible matches for each variant. This involves canonicalizing the variant string and then searching the resource table for the same canonicalized text name. This lookup is efficient since the service loads the resource table in memory, indexed by canonicalized text names in a hash map. A string comparison algorithm then compares the original (i.e., not yet canonicalized) entity string with each matched entry to assign a score. Gilda also integrates machine-learned disambiguation models to choose between ambiguous strings given relevant surrounding text as context, and supports species-prioritization in case of ambiguity.

We benchmarked Gilda on the BioCreative VI training corpus. This corpus contains 86,574 entity texts extracted from open-access full-text publications that were manually grounded by curators to a given entity type and namespace / ID pair. Table 1 shows Gilda's performance on this corpus without enabling non-human gene grounding (this mode is most relevant for our dialogue use cases). We see that precision and recall are very high for human genes (PR: 0.949, RC: 0.883).

Entity Type	Precision	Exists Correct PR	Recall	Exists Correct RC
Cell types/Cell lines	0.733	0.762	0.527	0.548
Cellular Component	0.602	0.627	0.443	0.461
Human Gene	0.949	0.965	0.883	0.897
Nonhuman Gene	0.002	0.002	0.001	0.001
Small Molecule	0.7	0.771	0.571	0.629
Taxon	0.312	0.809	0.233	0.605
Tissue/Organ	0.585	0.585	0.363	0.363
miRNA	0.358	0.358	0.282	0.282
Total	0.524	0.586	0.412	0.461

Table 1. Gilda grounding performance in “human only” mode on the BioCreative VI benchmark.

The Gilda web service is available at <http://grounding.indra.bio> with source code, documentation and tutorials are available via <https://github.com/indralab/gilda>.

Our preprint on Gilda is available at: <https://www.biorxiv.org/content/10.1101/2021.09.10.459803v1>

2.5. INDRA Interactive Pathway Map

The INDRA Interactive Pathway Map (INDRA-IPM) is a web-based pathway modeling tool that builds on the capabilities of INDRA to construct and edit pathway maps in natural language and display the results in familiar graphical formats. INDRA-IPM allows models to be exported in several different standard exchange formats, thereby enabling the use of existing tools for causal inference, visualization and kinetic modeling. We also make the capabilities of IPM available as a web service to facilitate use by other software.

INDRA-IPM provides an interface to enter English language text describing mechanisms used to generate a pathway map. This description is processed by one or more natural language processing (NLP) systems; users can choose between the REACH and TRIPS NLP systems. A pathway map is then generated automatically from the NLP output and visualized dynamically. Users can iteratively update and extend the pathway map by editing the underlying natural language.

The pathway map is represented as a directed graph with nodes corresponding to molecular entities (genes/proteins, families, complexes and small molecules) and edges representing mechanistic relationships among them. The graph is displayed using CytoscapeJS with a two-stage hierarchical layout procedure designed to reduce visual complexity. INDRA-IPM groups nodes with identical incoming and outgoing edges and aggregates them into a single bounding box with collapsed edges (e.g., RASGRF and SOS in Figure 13). Nodes representing protein families or complexes (e.g., the Sprouty family, SPRY in Figure 13), are recognized using the FamPlex ontology and represented by a single node subdivided to show the genes in the family as slices.

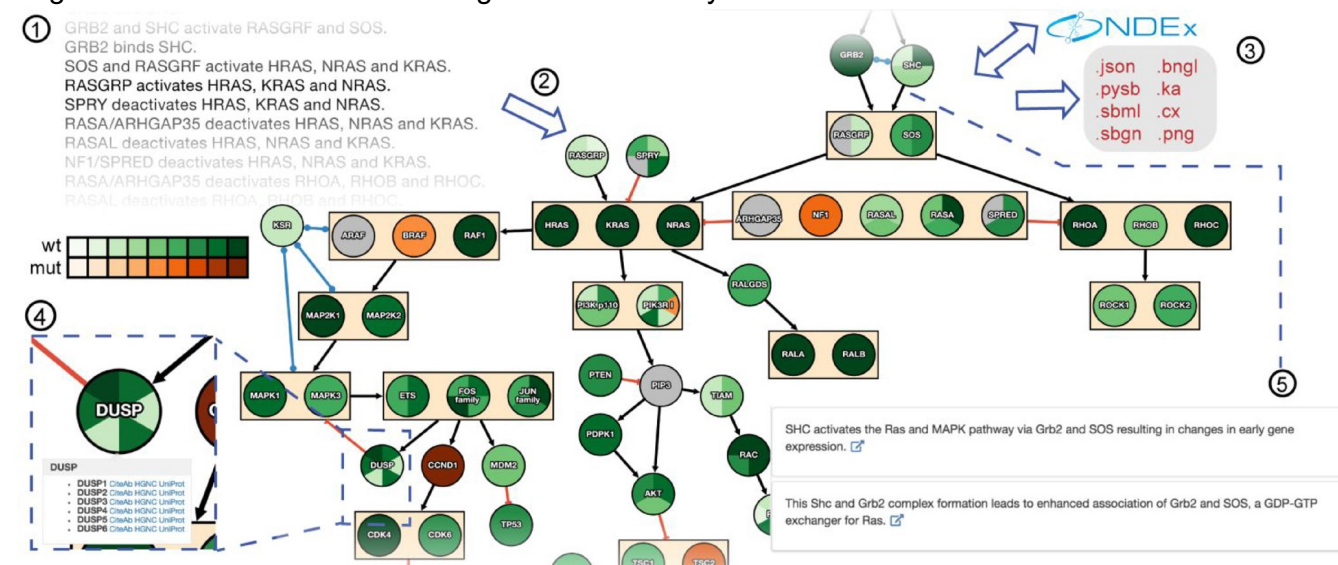


Figure 13. Pathway maps are assembled from natural language descriptions of mechanisms (1). INDRA-IPM renders pathway maps as graphs with node and edge grouping and color determined by mutational status and expression level (2). Pathways can be stored and shared on NDEx and exported and downloaded in many standard formats (3). Node tooltips provide links to online databases having information on genes/protein in the pathway and also to antibodies against proteins in the node (4). Literature-based evidence for a given interaction can be accessed by clicking on an edge.

Corresponding evidence sentences drawn from these publications are then shown with links out to the PubMed entry in which the sentences are found (5)

To leverage automated assembly for diverse modeling tasks, INDRA-IPM exports models as SBML, SBGN, BNGL, Kappa, PySB and CX. These formats are widely used in computational biology for modeling, simulating and visualizing pathways. Users also have the option of storing maps on the Network Data Exchange (NDEx) where they can be shared and reloaded into INDRA-IPM using a persistent URL. More details on these formats are available in the Supplementary Material.

Clicking on an edge in a pathway map retrieves support for that interaction by querying a database of interactions aggregated by INDRA. This database includes information gathered from reading literature at scale and information found in curated knowledge bases such as Pathway Commons and the BEL Large Corpus. Users are therefore able to access literature support for relationships specified in natural language descriptions.

INDRA-IPM is available at: <http://pathwaymap.indra.bio>. Source code is available at http://github.com/sorgerlab/indra_pathway_map. The underlying web service API is available at <http://api.indra.bio:8000>.

2.6. Kappa and KAMI

As part of this project, the HMS-Fontana and Lyon-Harmer teams developed a number of standalone software tools revolving around the detailed mechanistic modeling of biological pathways.

The Fontana team significantly extended the Kappa modeling framework (<https://github.com/Kappa-Dev/KappaTools>), notably, by implementing a server-client architecture for the simulator, and making the simulator interactive (simulations can be paused and restarted, conditions changed, etc. during runtime). Further, the KaSa static analysis module developed as part of Kappa is used in our dialogue systems for automated model diagnosis. The Kappy Python package provides an adapter around Kappa binaries (both for simulation and static analysis) and enables integration with Python modules which is important for our dialogue systems.

The key contribution of the Harmer team to the project was the creation of KAMI and KAMISudio. KAMI is a graph-based representation for detailed protein interaction mechanisms. As opposed to commonly used PPI representations, KAMI explicitly accounts for protein domains and sites, conditions of binding via those domains and sites, as well as their modification states. KAMI can also draw on structured knowledge about protein sites and domains, and combine this with expert curation.

In collaboration with the HMS-Sorger team, the Harmer group tackled the integration of KAMI with INDRA. The pipeline was extended to identify which agents (and their regions and/or key residues) mentioned in an INDRA Statement already exist in KAMI's action graph. From this, a graph rewriting rule can be determined, and applied to update the action graph with new agents (and regions and/or key residues). Other work on the KAMI-to-INDRA translation process incorporated definitions of proteins and the subsequent automatic determination of whether or not a KAMI nugget applies to a particular protein. A significant part of this pipeline has been included in an updated and generalized KAMI-to-Kappa translation process.

The Harmer team's work also included development of the Agent Anatomizer that extracts details about the functional structure of proteins — domains, disease-related mutations, splice variants, etc. — from UniProt, PFAM and related databases and constructs the appropriate representations in KAMI. This started as a stand-alone project but was subsequently integrated into KAMI where it provides useful additional knowledge for the accumulating update process described above. The Anatomizer was

further extended to allow it to use InterPro and thus, full coverage of reviewed UniProt isoforms. An offline mode was built for the Anatomizer to gather agent domains quickly when building large; this also required scripts for the performing the automatic update (every two months) of InterPro data.

KAMI is described in R. Harmer et al., "Bio-Curation for Cellular Signalling: The KAMI Project," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 5, pp. 1562-1573, 2019.

Software available at: <https://github.com/Kappa-Dev/ReGraph>

The usability of KAMI was improved by implementing an enhanced graphical user interface. As part of this effort, the team implemented tabs for different types of graphs in the interface (e.g., nuggets, splice variants, etc); re-worked the import of a saved hierarchy and merged it with the current one; optimized "nugget unfolding", which generates multiple nuggets out of a nugget with conflicts; and added the ability to define "models": sets of nuggets that can be selected at the same time when exporting to Kappa. This work culminated in KAMI Studio, a web-based bio-curation environment for modeling cellular signaling based on KAMI. KAMI Studio is described in Harmer, R., & Oshurko, E. (2019). *KAMISStudio: An Environment for Biocuration of Cellular Signalling Knowledge*. In *International Conference on Computational Methods in Systems Biology* (pp. 322-328). Springer, Cham. Software available at: <https://github.com/Kappa-Dev/KAMISStudio>.

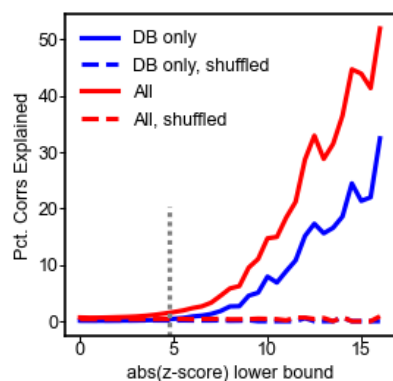
3. Applications

Over the course of this project, we applied our tools to a number of scientific applications. Some of these applications made use of the integrated dialogue systems directly or benefited from a component technology (such as INDRA or the INDRA DB) we developed for this project.

3.1. Automated explanation of cancer gene dependencies at scale

One of the key metrics for our goal of implementing algorithms for automatically explaining high-throughput experimental data is to apply the approach at scale, and **quantify the improvement offered by our method compared to a baseline control**. We evaluated our explanations of gene dependency correlations in the Dependency Map, comparing the use of our full INDRA corpus (including both text mining results and pathway databases) with relations from curated databases only.

We used a benchmark set of ~830,000 INDRA statements derived from processing a targeted corpus of 580,000 articles and calculated the proportion of correlations involving genes with direct causal links to each other. As a control, we shuffled the edges of both the full network and the database-only network and performed the same analysis to determine the range of correlations where the presence of an explanation for a correlation was above what would be expected by random chance. The results, shown in Figure 14, highlight that using the full corpus explained a much greater proportion of



correlations than the database-only corpus, despite the fact that the text mining results were drawn from only 580,000 articles, a small fraction of what is currently available in PubMed. Both corpora also explained significantly more correlations than would be expected at random, using a very conservative edge permutation approach that preserves the degree distribution of nodes in the network.

Figure 14. Explanation of gene dependency correlations with a combination of text mining and curated databases or curated databases only. Dotted lines show the proportion of correlations explained with edge-permuted networks. The proportion of correlations explainable via direct A-B interactions is shown as a function of the combined CRISPR and RNAi z-score.

As a second approach to show the value of our automatically assembled knowledge, we used our **network as a prior to control the false discovery rate in the hypothesis testing** of DepMap correlations and thereby identify hidden signals in the data. We found that the use of the network prior can identify statistically significant correlations that would otherwise be lost in the noise due to the false discovery corrections required when examining all correlations. For this analysis we needed to determine the statistical significance of each correlation under different sets of assumptions. To do this we first obtain raw p-values for the Pearson correlations among gene pairs in the CRISPR and RNAi datasets. As a form of meta-analysis, we sought to combine these p-values to identify the correlations that were most robust. Interestingly, we found that because the widely-used Fisher's method ignores the direction of effects between the two datasets, a gene pair that is positively correlated in CRISPR and negatively correlated in RNAi results in a very strong combined p-value despite the contradictory direction of effects.

Because of this issue we instead used Stouffer's method, an alternative approach whereby p-values are converted into z-scores which can be given a sign based on the direction of the original effect (negative z-score for negative correlation, positive z-score for positive correlation). These z-scores are then averaged according to Stouffer's formula and then converted back into a final combined p-value. Averaging the signed z-scores results in weaker p-values when the correlations are in conflicting directions. Having computed the combined p-values, we used different approaches to correcting for multiple hypothesis testing (Bonferroni, Benjamini-Hochberg and Benjamini-Yekutieli) with a false discovery rate of 0.05 and examined the gene pairs that were determined to be statistically significant with and without the use of our network prior. In all cases we ignored the correlations strictly among mitochondrial genes for the reasons described in last period's report.

The results show that though the prior from our corpus of 580,000 articles is relatively small relative to the total number of correlations (267 thousand vs 122 million), it can nevertheless identify thousands of relationships that would not otherwise be statistically distinguishable from noise (Table 2). Of these, the most meaningful numbers are those obtained after applying the Benjamini-Yekutieli (B-Y) procedure, which correctly controls the false discovery rate even when the p-values being tested are not statistically independent.

	No prior	INDRA prior	Only from INDRA
Total correlations	121,778,711	121,778,711	N/A
Number of comparisons	121,778,711	266,877	N/A
Correlations with uncorrected $p < 0.05$	25,211,483	68,662	N/A
After Bonferroni	86,948	4,774	1,840
After Benjamini-Hochberg	7,349,009	33,481	6,520
After Benjamini-Yekutieli	1,279,296	13,425	4,232

Table 2. Statistically significant gene dependency correlations with and without a prior knowledge network used to restrict the number of hypotheses tested.

Of the 13,425 correlations that were significant at a false discovery rate of 0.05 after the B-Y procedure, 4,232 were obtained *only* when using the INDRA prior, that is, they were not obtained after correcting for all ~122M comparisons in the full dataset. Notably, a large majority of these (2,917) were obtained only from machine reading, sometimes from many papers and evidence sentences. For example, this approach highlighted the role of the kinase ROCK2 in promoting the acetyltransferase activity of EP300

([link](#)), the direct interaction between Parkinson's disease gene PARK7 and the dopamine transporter SLC6A3 ([link](#)), and many others.

3.2. Automated finding of surprising, significant, and actionable hypotheses: a case study of KIRREL1 function

As part of our work to transition our computational tools to directly supporting empirical lab-based scientific discoveries, we have used our large-scale knowledge network to identify correlations in the Cancer Dependency Map that are both significant, surprising, and actionable. As a result of this analysis, we have **identified a potentially novel role for KIRREL1**, a gene that is best known for its role in the kidney slit diaphragm, in regulating the Hippo cell growth signaling pathway, and have begun work to confirm this experimentally.

3.2.1. Approach to automatically finding hypotheses at scale

A key part of our approach is converting the high-level scientific priorities like “significant, surprising, and actionable” into concrete criteria that can be used to filter observations in datasets like the Dependency Map. We have defined *significant* in terms of phenotypic (and possibly disease) relevance, focusing on explaining correlations among genes known to be key regulators of cellular processes or diseases. In the DepMap, this translates to focusing on essential or selectively essential genes where knockout or knockdown leads to a significant change in cell growth among at least a subset of cell lines.

Correlations that are *interesting or surprising* are those that involve genes about which very little is known, or genes known for functions distinct from those implied by the correlation. The way we have chosen to approach this is two-fold. First, we use direct links in our INDRA network to exclude genes that have already been described in the literature as having an interaction that could explain the correlation. The presence of such a direct link implies that articles have already appeared describing the relationship, making it “unsurprising.” Second, we use catalogs of biochemical pathways (in this case from Reactome) to determine whether the two genes are known to be related at a macro-level. If the correlated genes have been curated to be part of at least one shared pathway in Reactome, their correlation is considered unsurprising—even if the genes haven't been curated as directly interacting with each other. Because participating in a shared pathway is often sufficient for a human curator to gain a general appreciation of why two genes might be correlated, we classify those genes that are strongly correlated genes but have *no* pathways in common as “surprising.”

Finally, correlations that are *actionable* are those for which INDRA identifies at least one *indirect* link between the two genes. Each indirect link can be a common downstream target of the correlated genes or an intermediate link in a linear pathway. Together these intermediates form a set of causal hypotheses that could potentially explain the correlation via mechanistic, “micro-level” information.

Using these three criteria, we have filtered the ~300 million correlations in the DepMap to those that are 1) strong (absolute z-score > 5 standard deviations), 2) significant (genes showing phenotypic impact of knockout in at least some cell lines), 3) surprising (not explained by a direct link in INDRA or by a common Reactome pathway) and 4) actionable (have at least one candidate intermediate link in the INDRA network). Of the 7,045 strong correlations in the data, this analysis results in 612 high-value correlations.

As part of our transition work, we are beginning the process of **experimentally following up several correlations** involving the Hippo signaling pathway, particularly the strong correlation between NF2, a gene associated with rare pediatric nerve tumors, and KIRREL1, a transmembrane protein known for its role in forming cell junctions in the kidney.

3.2.2. Proposing a novel function for KIRREL1 in the Hippo pathway

NF2 has a correlation of 0.64 with KIRREL1, second only to its known binding partner AMOTL2. The slope of the relationship is steeper among the subset of central nervous system cell lines, indicating a greater phenotypic effect of knockout in this context. The correlation is surprising because NF2 and KIRREL1 have no direct link in INDRA nor any Reactome pathways or GO terms in common. Of the ~150 articles on KIRREL1, all focus on the essential role of KIRREL1 in the kidney, where it maintains a selectively permeable barrier as part of a structure called the slit diaphragm. At a mechanistic level, KIRREL1 is known to bind the protein Nephrin, and to be phosphorylated by Src family kinases leading to binding of the well-studied scaffold protein GRB2.

KIRREL1's own correlations in the Dependency Map tell a seemingly unrelated story: of the top nine most strongly correlated genes with KIRREL1, eight, including NF2 itself, are known to be involved in the Hippo signaling pathway, a tumor suppressor pathway that prevents cell growth when cells are already in contact with other cells. Many of the remaining correlations involve genes known to be involved in RHO or RAC signaling which govern the actin cytoskeleton and cell motility. Interestingly, using the INDRA network search we have found that proteomics datasets indexed in BioGrid show that both KIRREL1 and NF2 bind key Hippo signaling proteins LATS1 and SAV1 (Figure 15).

Common Targets (KIRREL1→Z←NF2)		Targets: 7
Target (Z)	Support	
LATS1	KIRREL1 → LATS1	Support: 2
	KIRREL1, Complex, LATS1	biogrid: 1
	NF2 → LATS1	Support: 6
	NF2, Complex, LATS1	reach: 2
	NF2, IncreaseAmount, LATS1	reach: 1
	NF2, Inhibition, LATS1	sparser: 1 reach: 3 medscan: 1
	NF2, Phosphorylation, LATS1	reach: 2
SAV1	KIRREL1 → SAV1	Support: 2
	KIRREL1, Complex, SAV1	biogrid: 1
	NF2 → SAV1	Support: 2
	NF2, Complex, SAV1	sparser: 1 reach: 2

Figure 15. INDRA Network Search highlights LATS1 and SAV1, known Hippo pathway proteins, as links connecting KIRREL1 and NF2.

By examining KIRREL1's connections in the INDRA network and the associated literature evidence, we have converged on an **initial hypothesis for KIRREL1's apparent involvement in the Hippo signaling pathway** (Figure 16). Upon phosphorylation by SRC family kinase FYN, KIRREL1 binds GRB2; GRB2 recruits p190RhoGap (official name ARHGAP35), inhibiting RHO which is itself an inhibitor of the Hippo pathway, potentially via an interaction with LATS1.

Through a collaboration with experimentalists in the Lab of Systems Pharmacology, we are embarking on a series of **experiments to test this hypothesis**, first by showing that KIRREL1 loss leads to a decrease in Hippo pathway signaling (and increased cell proliferation), and then following up with additional experiments to validate the mechanistic details.

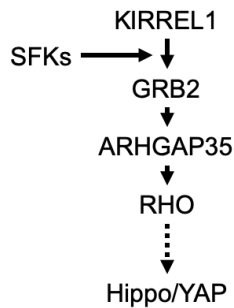


Figure 16: A hypothetical mechanism linking KIRREL1 to the Hippo signaling pathway via its known interactors. “SFKs” denotes Src family kinases.

3.2.3. A hypothesis for KIRREL1's role as a cancer biomarker

Most of the scientific literature describing KIRREL1 has focused on its role in forming the slit diaphragm in kidney podocytes. However, there are a handful of papers that have highlighted its role as a negative prognostic biomarker in a variety of cancer types, including skin, stomach, and pancreatic cancers. These reports linking KIRREL1 expression to aggressive cancer subtypes would appear to be in contradiction with KIRREL1's apparent function as a tumor suppressor in the Cancer Dependency Map data (i.e., knocking the gene out leads to increased growth in most cell types, suggesting its role is to restrain cell proliferation).

To investigate this paradox, we examined cell line co-expression data and found that the expression of KIRREL1, which we hypothesize to be a negative regulator of the Hippo pathway transcription factors YAP1, TEAD1, and WWTR1, has strong correlations with the expression of each of these genes. This suggests the possibility that KIRREL1 is not only an upstream negative regulator of YAP1/TEAD1/WWTR1 but also a downstream target, establishing a negative feedback loop controlling cell proliferation. Consistent with this hypothesis, we found reports of homeostatic negative feedback loops involving other negative regulators of the Hippo pathway.

KIRREL1 expression is also correlated with Vimentin expression and other transcriptional biomarkers of epithelial-to-mesenchymal transition (EMT), a shift in cell state associated with more aggressive and metastatic cancers. We therefore suspect that KIRREL1 is appearing to be a marker of more aggressive cancers not because it is driving growth, but is rather a marker of the homeostatic transcriptional program downstream of EMT-associated transcription factors such as YAP1. We therefore suspect that in investigating this gene we have the opportunity to highlight not only one of its underappreciated functions in the Hippo pathway but also add context to its role as a cancer biomarker.

3.2.4. Experimental plan

For our initial experiments we will assay YAP1 activity in the presence or absence (knockdown) of KIRREL1 using two methods, 1) single-cell imaging of YAP1 cellular localization, and 2) measurements of the transcript abundance of YAP1 target genes by RT-PCR. For the imaging experiments, we have validated the antibody that we will use to measure YAP1 localization in MCF10A cells. In a previous report we showed that the antibody was able to stain cells; since then we have confirmed that the staining pattern is consistent with what is expected for YAP1, with a higher nuclear to cytoplasmic ratio associated with lower cell density conditions.

For the transcriptional assays, we searched the literature for validated transcriptional targets of YAP1 and found several studies with relevant gene lists. Notably, the list of YAP1 targets includes a number of Hippo pathway proteins that are negative regulators of YAP1 activity, such as AMOTL2 and LATS2. Based on these reports we chose the following list of genes to measure by RT-PCR as indicators of YAP1 activity: CCN1, CCN2, LATS2, AMOTL2, and PTPN14. Because we also hypothesize that

KIRREL1 transcription is itself regulated by YAP1, we also included KIRREL1 in the list and have ordered primers for all six.

As assay controls for YAP1 activity, we have identified two drugs reported to inhibit Hippo pathway kinases and thereby activate YAP1. These are XMU-MP-1, an inhibitor of the kinases MST1/2, and also the recently described TRULI, an inhibitor of the LATS1/2 kinases, downstream of MST1/2. We have ordered these drugs and plan to use them as positive controls for the imaging-based and transcriptional assays of YAP1 activity, with the expectation that either drug should (as reported) lead to an increase of nuclear, transcriptionally active YAP1.

For all of these experiments, we will use the A375 melanoma cell line, which was previously reported to be strongly sensitized to BRAF or MEK inhibition in the context of KIRREL1 knockout. Our experimental collaborators within the Lab of Systems Pharmacology previously developed A375 cell lines stably expressing the genetic constructs to enable sgRNA-mediated activation (CRISPR-a) or inhibition (CRISPR-i) of the expression of target genes. We will therefore develop the guides to target KIRREL1 via CRISPR-i to downregulate KIRREL1 expression in A375s and measure the activity of YAP1 to determine whether KIRREL1 is a regulator of YAP1 activity.

3.3 Biofactoid

The Biofactoid system allows authors to load meta-information about their publications, and then curate (using a simple graphical interface) key pathway interactions described in their paper. Biofactoid is accessible at <https://biofactoid.org/>. The collection of author-curated pathway elements constitutes a knowledge base that can be used for downstream analysis. Biofactoid is intended to become part of the publication process of multiple journals.

Both the HMS and OHSU teams contributed to Biofactoid as part of this project. One feature contributed by the HMS team aims to help people explore research papers that are related to the information that is represented in a given Biofactoid document. In a Biofactoid document each node represents a biological entity and each edge represents an interaction. We integrated the INDRA DB with Biofactoid to allow querying for INDRA statements corresponding to each interaction in the Biofactoid document. Each of the statements that is returned from the INDRA DB contains a list of evidences that are associated with PubMed articles. These are then surfaced as recommendations on the Biofactoid interface (Figure 17). Articles can either be sorted by recency or by similarity to the given Biofactoid document based on Semantic Search scores.

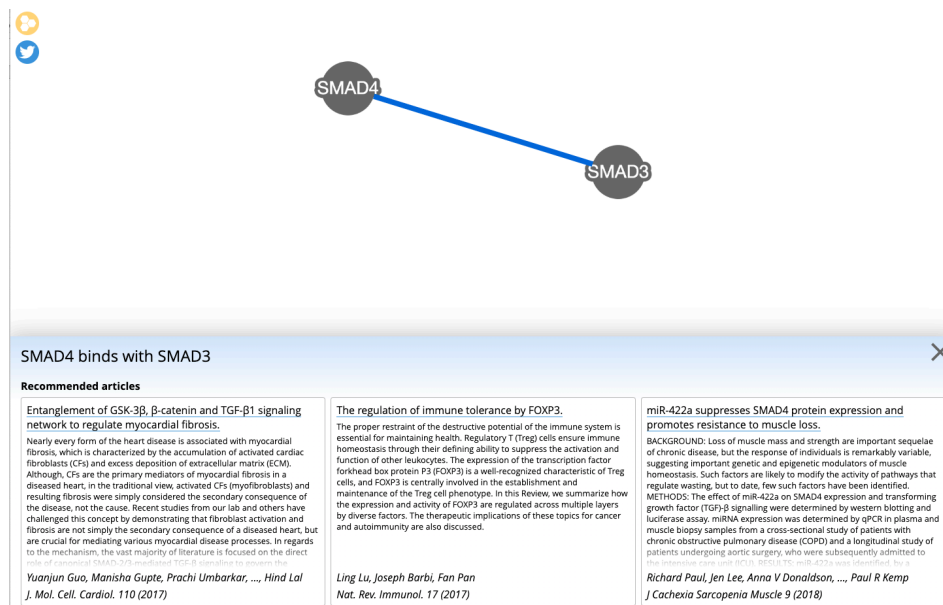


Figure 17. Recommended articles provided by the INDRA DB in the Biofactoid interface.

Going beyond article recommendations, we created an experimental branch of Biofactoid to integrate a CLARE chatbox into it. The interface monitors whenever a new entity is added to the existing Biofactoid model, which triggers CLARE to search for the interactions between the new entity and the existing ones from the literature. After completing the query CLARE lists the interactions that are found. Then, the user can ask CLARE to add any of these interactions to the Biofactoid model. Through the chat box, the user can utilize any other capabilities that CLARE provides as well. We are hoping to continue the development of this integration going forward. We also implemented a new integration between Biofactoid and INDRA that allows INDRA to pull factoids curated through the Biofactoid interface and use them as expert-curated knowledge in modeling and data analysis.

Our paper “Author-sourced capture of pathway knowledge in computable form using Biofactoid” has been accepted for publication in eLife.

3.4. GeneWalk

As part of a collaboration with the Churchman lab at the Department of Genetics, Harvard Medical School, we developed GeneWalk which uses INDRA-derived networks as a basis for determining context-specific gene functions using deep learning. Given a gene list derived from an experiment, GeneWalk takes a network of interactions/regulations between the genes and combines this with GO annotations and their hierarchy (Figure 18). It then uses the deep-walk algorithm to learn node embeddings for each of the genes, and then calculate measures of similarity between each gene and GO term. This results in a quantification of the context-specific associations to GO terms for each gene in the experiment. Our results show that using the prior network to infer these associations produces better results than network-free alternatives.

Our paper on GeneWalk is published in Genome Biology:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02264-8>, and was discussed in Harvard Medicine News: <https://hms.harvard.edu/news/panning-genetic-gold>.

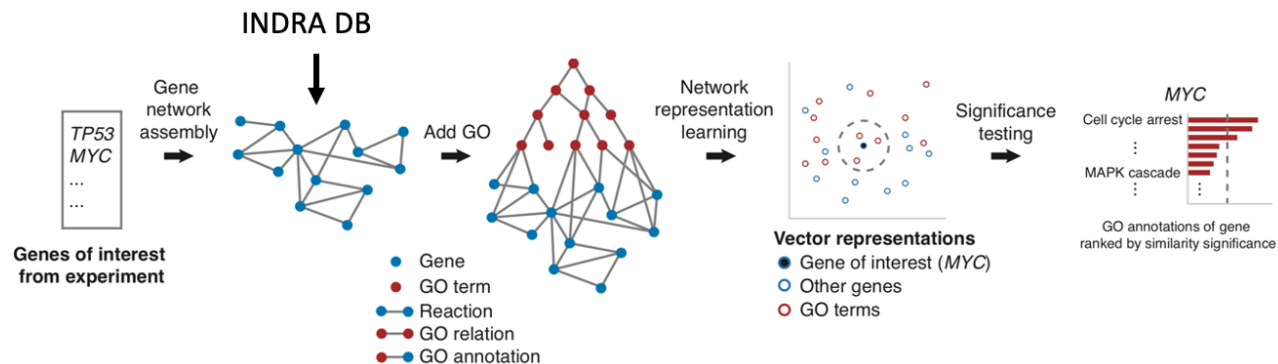


Figure 18. Conceptual overview of the GeneWalk algorithm which takes networks assembled by INDRA as input to derive context-specific gene functions.

3.5. IQuery: network-based gene set enrichment

IQuery is a web application developed primarily at UCSD for querying a curated subset of the online biological network repository NDEx. Given a set of query genes, IQuery will return NDEx pathways enriched for the query genes, networks representing the interactions between those genes and other proteins, and networks representing the associations between those genes and other biological or chemical entities. IQuery is designed such that new types of searches and analyses may be added in the future. We contributed to IQuery by creating INDRA-assembled networks of genes annotated with each Gene Ontology (GO) term (one network per GO term). The set of genes in each network is defined by GO's annotations, and each gene is represented as a network node. The relations between the genes are derived from INDRA DB and thus are assembled from processing the scientific literature and pathway databases. Figure 19 shows an example of one such network in the context of IQuery.

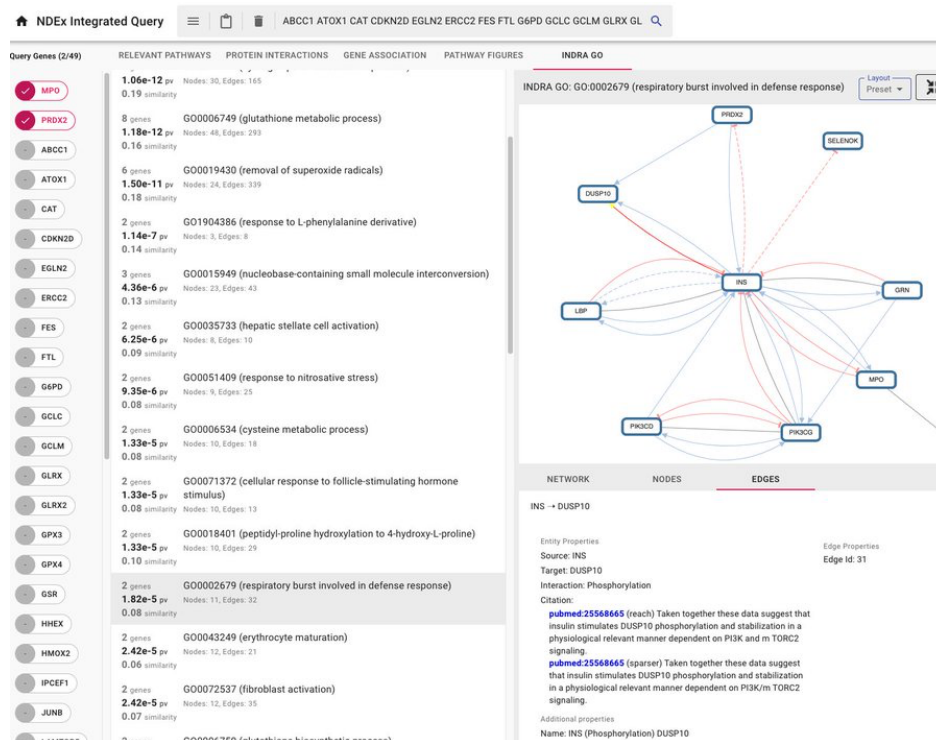


Figure 19. When searching IQuery with a set of genes, enrichment over the INDRA-generated GO networks is calculated and the top hits are listed on the INDRA GO tab.

3.6. Illuminating the Druggable Genome: Dark Kinases

We applied our dialogue systems and the underlying knowledge to the NIH Illuminating the Druggable Genome (IDG) program (<https://commonfund.nih.gov/idg>). IDG aims to shed light on the function of classes of proteins about which little information is known, but could be druggable and have therapeutic potential. One branch of the program focuses on “dark” kinases, and creating a knowledge base that allows exploring new therapeutic opportunities, and help with data interpretation. Our CwC technologies have contributed to this program in two ways: (1) provide structured knowledge about dark kinases assembled from pathway databases and the literature, and (2) exposing dialogue capabilities to allow scientists to ask questions about dark kinase mechanisms and build mechanistic hypotheses about their function.

We used knowledge captured in the INDRA DB to create a structured, browsable knowledge base of dark kinases. This is now integrated into the Dark Kinase Knowledge Base (DKKB) at <https://darkkinome.org/>. Figure 20 shows one of the dark kinase networks assembled from the INDRA DB.

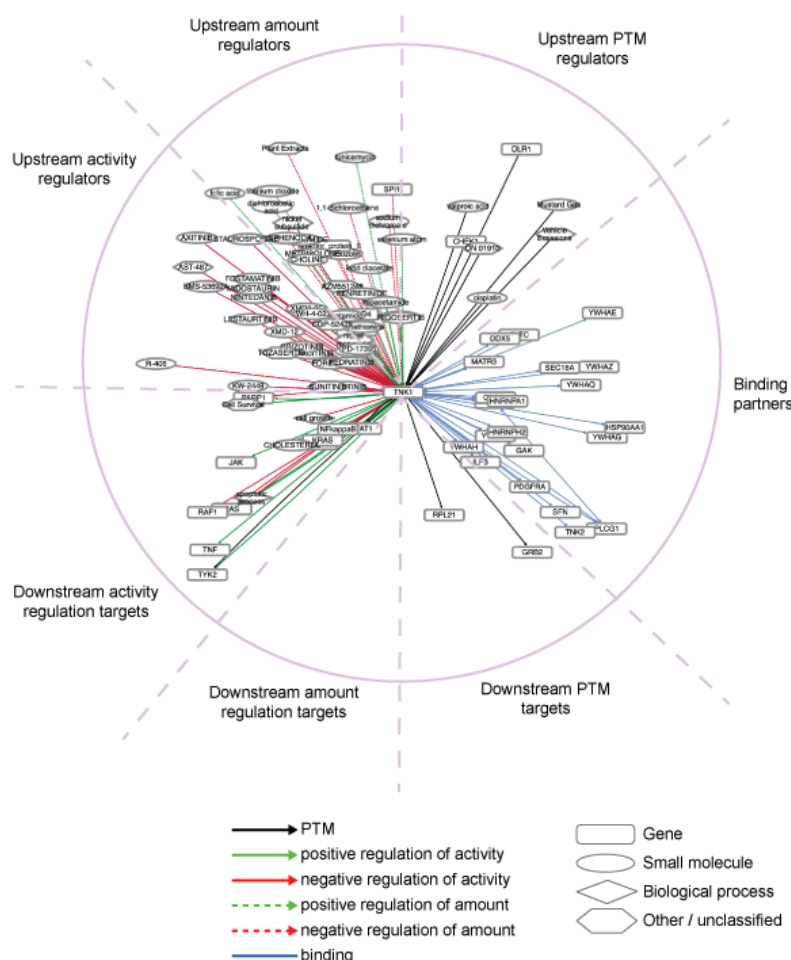


Figure 20. Annotated kinase network made available on NDEx.

Subsequently, we extended our analysis to all kinases and analyzed the number of distinct INDRA Statements to measure how widely the given kinase has been studied (Figure 21). We found this to be a useful metric for determining the extent to which a given kinase is understudied and identify ones that should be investigated further (red area on plots).

b. Knowledge about the kinome extracted by INDRA and TIN-X

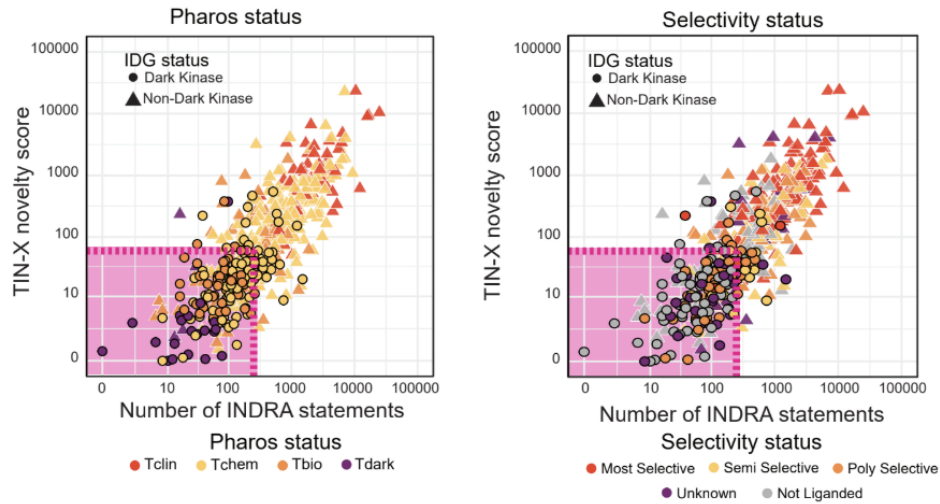


Figure 21. Number of INDRA Statements per kinase against an alternative novelty score (TIN-X) to measure the “darkness” of each kinase (dark kinases highlighted with red background).

In addition to creating the knowledge base, a computational scientist at HMS used the CLARE system to ask a series of questions about two dark kinases of interest (BRSK1 and NIM1K). Based on this dialogue, she was able to construct a model of mechanisms by which these two kinases affect cell cycle progression, which could be the key to their use as therapeutic targets. It is worth noting that this scientist could have used other tools (Googling, reading publications, searching Pathway Commons, etc.) to learn about potentially relevant information, but instead, chose to ask a series of NL questions from CLARE. Figure 22 shows a diagram of the hypothesis this scientist arrived on with the help of the dialogue.

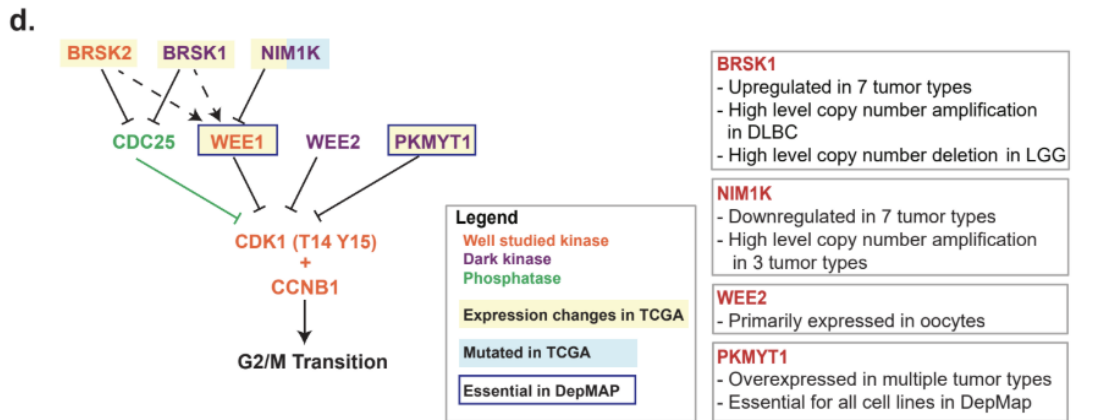


Figure 22. Mechanisms by which dark kinases BRSK1, NIM1K, WEE2, and PKMYT1 interact and control cell cycle progression. This hypothesis is partly based on natural language dialogue with the CLARE system.

The results have been written up as part of a manuscript titled “Exploring the understudied human kinome for research and therapeutic opportunities.”

(<https://www.biorxiv.org/content/10.1101/2020.04.02.022277v2>).

3.7. Using INDRA for chemical hazard identification

As part of a collaboration with the Netherlands Organisation for Applied Scientific Research (TNO), we applied INDRA to chemical hazard identification. Mechanistic data is increasingly used in hazard identification of compounds; however, the volume and variety of this data is large, making it challenging to efficiently collect and evaluate it. We investigated if evidence synthesis for hazard identification can be made more efficient, transparent and reproducible through an automated approach that combines large-scale machine reading of biomedical publications with network visualization. As a case study, we focused on the 10 “key characteristics” (KCs) of carcinogens (as defined by the International Agency for Research on Cancer), and for a set of chemicals of interest, evaluated the quantity of evidence available for its effect on these KCs. We first identified a set of publications relevant for each chemical, and used the Reach reading system coupled to INDRA to extract relations from text as INDRA Statements (a typical such relation between a chemical and a KC is “benzene activates DNA damage”). We then used INDRA to assembled a network around each chemical, and compared these to the evaluation of KCs for these chemicals by IARC. We generated a total of seventy networks (10 compounds, seven networks each). In general, larger networks with higher belief scores (a measure indicating more evidence supporting a statement) were obtained for those KCs that were evaluated by IARC to have strong evidence of being activated. The networks can be filtered in various ways allowing risk assessors to analyze the retrieved interactions in more detail. Our approach and the “classic” hazard identification approach followed by IARC are shown side-by-side in Figure 23.

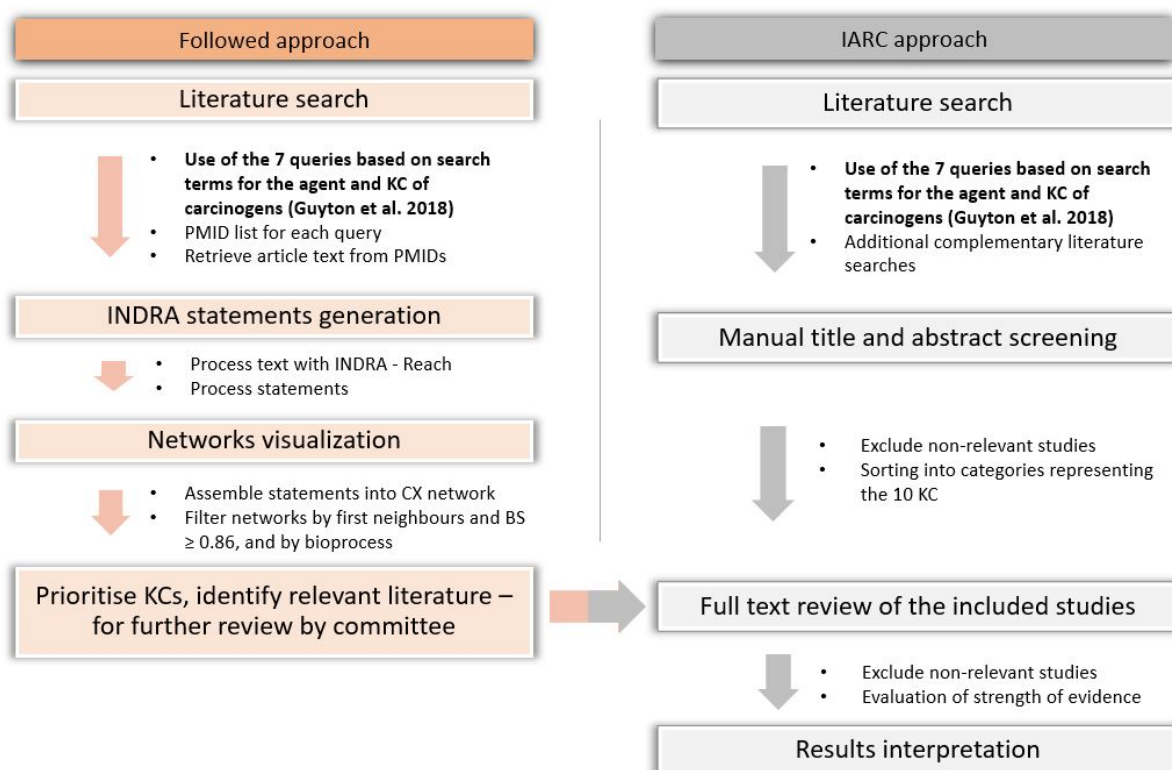


Figure 23. Workflow for partially automated chemical hazard identification using INDRA compared to the “classic” expert curation-based approach followed by IARC.

This manuscript about these results is currently under review in *Environmental Health Perspectives*. The scripts to generate the results are available at: https://github.com/bernice493/INDRA_hazard.

3.8. COVID-19

With the emergence of the COVID-19 pandemic in early 2020, we generalized our component technologies and integrated dialogue systems to support discovery relevant to SARS-CoV-2/COVID-19.

First, we applied the CLARE dialogue system to rapidly find potential drug candidates that could have a therapeutic effect on either SARS-CoV-2 infection, and/or the subsequent immune response. Using dialogue with CLARE, we were able to find multiple potential drug candidates, including, for instance, E-64 (a cathepsin inhibitor). Figure 24 illustrates this effort in which simultaneous human-machine and human-human dialogue is conducted on Slack to formulate hypotheses. The spreadsheet on the right of Figure 24 is curated in collaboration with Prof. Matt Might (<https://www.uab.edu/medicine/pmi/matt-might>) who is testing repurposable drug candidates in cells at the University of Alabama at Birmingham. The list contains drug candidates that we identified via dialogue with the CLARE system.

Experiments with lung organoids at UAB subsequently found that E-64d, one of the drugs we proposed was one of the top 5 highest confidence hits in the drugs screen against SARS-CoV-2.

bgryori 11:44 AM
@clare what affects ACE2?

clare APR 11:44 AM
Overall, I know that for instance, AGT, CBK1, ADAM17, ACE2, ACE, losartan, SLC33A1, INS, lipopolysaccharide, and HNF1A can affect ACE2.
You can browse the full set of 666 statements here: <https://s3.amazonaws.com/indrabot-results/612fce80-fb6a-43d3-a442-2f4e8609dc23.html>

bgryori 11:44 AM
@clare which of those also affects TMPRSS2?

clare APR 11:44 AM
Out of those, I know that for instance, AR, TMPRSS2, FGF23, FOXA1, NOG, E2, VEGFA, CDK, 3,3',5'-triiodo-L-thyronine, and 17alpha-ethynylestradiol can affect TMPRSS2.
You can browse the full set of 25 statements here: <https://s3.amazonaws.com/indrabot-results/75b9714d-b78d-4e0a-9802-80f2539caf27.html>

bgryori 11:44 AM
AR comes out very strongly on top
so presumably AR modulating drugs should be considered

johnbachman 11:47 AM
one thing to keep in mind is that a lot of the interactions of TMPRSS2 are described in the context of prostate cancer, where gene fusions between TMPRSS2 with ERG are very common
that's one reason is shows up with androgens a lot, which may be legit, but also may be a very different context

bgryori 11:48 AM
@clare what is NOG?

clare APR 11:48 AM
NOG is a protein with the following description: Inhibitor of bone morphogenetic proteins (BMP) signaling which is required for growth and patterning of the neural tube and somite. Essential for cartilage morphogenesis and joint formation. Inhibits chondrocyte differentiation through its interaction with GDF5 and, probably, GDF6 (PubMed:21976273, PubMed:26643732). Links: HGNC UP.

bgryori 11:48 AM
@clare what inhibits NOG?

clare APR 11:48 AM
Overall, I know that for instance, MAG, toluene, BMP, IGF1, toluene, FGF, BMP2, F2, FGF18, and FST can inhibit NOG.
You can browse the full set of 85 statements here: <https://s3.amazonaws.com/indrabot-results/a4bdd6ef-2461-4d16-80be-cc52aded0bf9.html>

bgryori 11:48 AM
@clare what is toluene?

johnbachman 11:48 AM
toluene is airplane glue, yikes

clare APR 11:48 AM
toluene is a small molecule. Links: CHEBI PUBCHEM, (edited)

COVID19 candidate drugs			
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive			
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1188015/			
A	B	C	D
Compound name	Help/Hurt	Mechanism(s) of Action	Effective concentration
1 quercetin	Help	ACE2-Spike interference (MD); Mpro inhibitor; estrogenic	
2 hydroxychloroquine	Help	breaks ACE2 glycosylation; alters endosomal pH	
3 hydroxychloroquine + azithromycin	Help	breaks ACE2 glycosylation; alters endosomal pH; azithro - ?	
4 chloroquine	Help	breaks ACE2 glycosylation; alters endosomal pH	
5 camostat	Help	inhibits TMPRSS2	
6 estrogen	Help	downregulates TMPRSS2	
7 bisphenol A	Help	downregulates TMPRSS2	
8 genistein	Help	downregulates TMPRSS2	
9 remdesivir	Help		
10 lopinavir-ritonavir	Help		
11 niclosamide	Help	inactivates viral protease post-infection / alters pH?	0.1 micromolar
12 ciclosenide	Help	inhibits Nsp15	
13 NSAIDS	Hurt?	unclear	
14 ACE inhibitors	Hurt?	upregulates ACE2	
15 losartan	Help? Hurt?	ARB (angiotensin receptor blocker). Conflicting evidence.	
16 MDL28170	Help	Cathepsin L inhibitor	EC50 2.5 nM
17 Vitamin D	Help	Inhibit Cathepsins by upregulating cystatins.	
18 E-64	Help	Cathepsin B and L inhibitor	
19 Odanacatib	Help	Cathepsin K inhibitor (may also target Cathepsin L)	

Figure 24. Using the CLARE dialogue system to query the INDRA DB to identify repurposable drug candidates to be tested in the lab.

Two scientists from Pacific Northwest National Laboratory used Bob with Bioagents for modeling COVID-19/SARS-CoV-2 mechanisms (Figure 25). In dialogue sessions so far, without any additional training or coordination, they were able to build elaborate models of how SARS-CoV-2 infection results in an immune response. As seen in Figure 20, the model connects SARS-CoV-2's binding to the human ACE2 protein with downstream molecular mechanisms culminating in IL6 production. Based on the logs, we found that building this model took less than 30 minutes by this scientist who we did not previously provide training to or coordinate with regarding the use of the system.

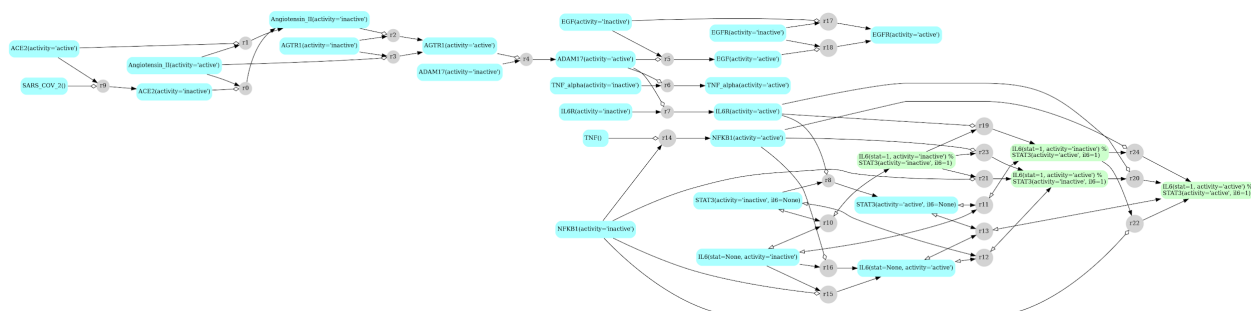


Figure 25. Model of SARS-CoV-2 molecular mechanisms built via human-machine dialogue using the Bob with Bioagents dialogue system (SBGNViz interface) by a PNNL scientist.

Scientists in the COVID-19 Disease Maps community have also been using CLARE via Slack to help with disease map curation (for quantitative usage statistics, see Section 1.2.3). We received positive feedback from scientists involved in this community about CLARE and INDRA. For instance, one scientist described that despite having worked for years on a specific protein, CLARE was able to highlight literature references in its answer that he did not previously know about. The comment (shown in Figure 26) also demonstrates how our system positively shapes the perception of text mining and human-machine dialogue in this community. (Note that in this example, Ben Gyorj asked the specific question from CLARE, however, the same scientist used CLARE in a separate chat session to follow up on a number of further questions).

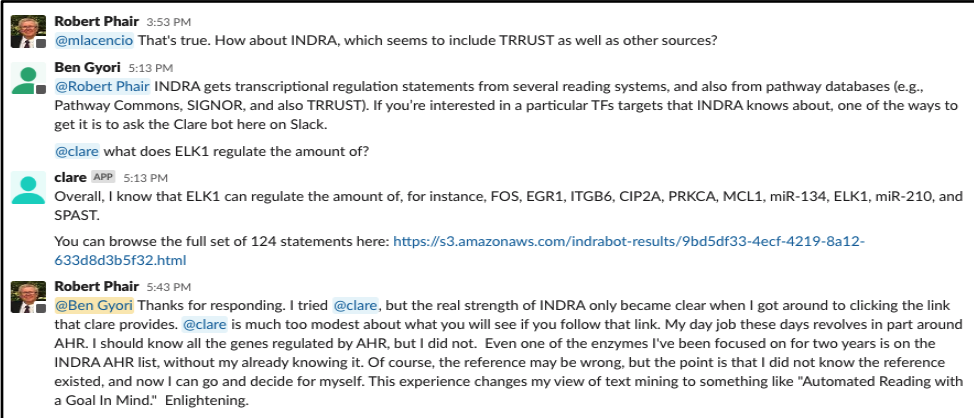


Figure 26. Comments on CLARE and INDRA from one COVID-19 Disease Maps scientist.

3.9. Molecular Tumor Board

Tumor boards consist of doctors and scientists with different specialties who meet to discuss cancer cases, especially in situations where standard of care does not lead to improvement. In some hospitals, tumor boards include a pathway analyst who can examine prior knowledge about cellular pathways in the context of given patient’s tumor (i.e., mutations, copy number alterations) and make a recommendation for potential therapeutic options.

The tumor board use case extends CLARE beyond supporting scientific discovery to support clinical decision making. At the core of CLARE’s tumor board capabilities is the ability to load and discuss a cancer patient’s molecular profile (Figure 27). The molecular profile, in its basic form, consists of a list of genetic variants derived from the patient’s tumor, which can include alterations such as point mutations, truncations, or copy number variations. To get real but publicly available data, we implemented a client for loading publicly accessible patient profiles from cBioPortal (these profiles are anonymized to prevent identification of the actual patient), serving as a source of real tumor data for the system. We also implemented the ability to load profiles of real patients discussed in the context of

OHSU tumor boards; in this case, the profiles are only used within the institution and cannot be publicly shared (for reasons of patient privacy).

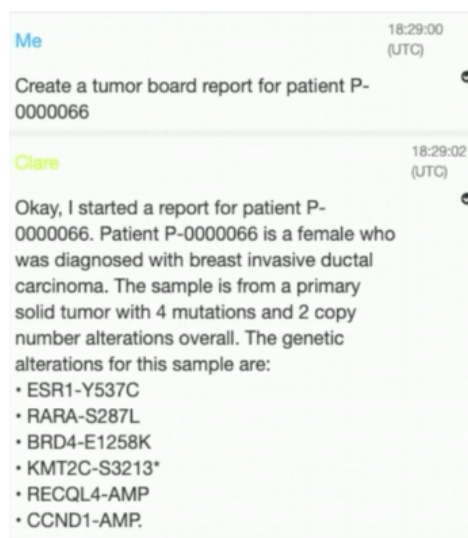


Figure 27. Creating a tumor board report for a patient profile loaded from cBioPortal. Basic clinical data about the patient as well as the tumor’s genetic alterations are summarized.

Once a report has been created, CLARE maintains it as context, and any subsequent references in the dialogue are resolved with respect to the patient being discussed. We implemented the following capabilities:

- List the most important interactors for neighbor genes in a tumor board report. (Example: what are the top 5 neighbor genes?)
- List the most important genes in a tumor board report. (Example: what are the top 5 genes?)
- List all important genes in a tumor board report. (Example: list all of the genes)
- Provide reason for why a given gene is important in the context of a tumor board report. (Example: why is TP53 important?)
- Provide description for the mutated genes from the tumor board report. (Example: tell me about mutated genes)
- Provide description for the copy number altered genes from the tumor board report. (Example: tell me about copy number altered genes)
- Describe and visualize how a gene and its important neighbors interacts in the context of a tumor board report. (Example: tell me about important neighbors of MDM2)
- List drug information for each variant of a tumor board report. (Example: tell me about variants)
- Provide information about drugs for a specific genetic variant. (Example: tell me about the variant BRAF-V600E)
- List the clinical trials that includes the specified drugs combination. (Example: are there any clinical trials attempting trametinib and dabrafenib?)

Collectively, these capabilities (along with all previous capabilities with which they integrate seamlessly) allow a tumor board pathway analyst to explore the relevant genetic variants in the tumor, identify their interactors and find druggable targets. Once druggable targets have been identified, the analyst can find specific drugs that inhibit those targets, and clinical trials in which a given combination of drugs is being evaluated. The analyst can then report their findings to the tumor board.

In order to highlight “top” altered genes and their interactors, we developed a set of scoring criteria for ranking genes. In addition to ranking genes using the scoring scheme, CLARE can also explain why a

given gene was highlighted as important. The following exchange (see also Figure 28) demonstrates this:

- Analyst: What are the top 10 genes?
- CLARE: Based on an analysis of interactions, EP300, ESR1, TP53, H3F3B, H3F3A, HIST1H3B, JUN, TNF, AR, and HIST1H3I are the most important genes in the tumor board report.
- Analyst: Are any of those transcription factors?
- CLARE: Out of those, ESR1, TP53, JUN, and AR are transcription factors.
- Analyst: Why is ESR1 important?
- CLARE: ESR1 is interesting because it contains a variant. It also interacts with mutated genes including CCND1, KMT2C, and RARA.

To further reflect on this dialogue snippet: the first answer of important genes is constructed by searching the neighborhood of the list of variant-containing genes using INDRA, scoring them, and producing the list of top hits. The analyst then uses a follow-up question with a co-reference (“are any of those...?”), handled by a general CLARE capability which allows filtering a list of entities by their type, to get a list of transcription factors. The analyst then follows up on ESR1 to find out why it was considered important and gets a clear, qualitative explanation.

Several interactions during a tumor board dialogue also involve visualization. While the previously described SBGN visualization can faithfully represent complex protein states and reaction-level details of mechanisms, it is not appropriate for generating a simple interaction network for the Tumor Board use case. We therefore implemented a Simple Interaction Format (SIF) network visualizer for SBGNViz, and extended CLARE to support this mode of network visualization. Figure 28 shows an example of a SIF network representing the neighborhood of ESR1 in the context of a tumor board report. The interactions constituting the network can also be browsed and curated as a set of INDRA Statements to which the system links out on the web. This allows reading the underlying evidence text, and linking out to the corresponding publications.

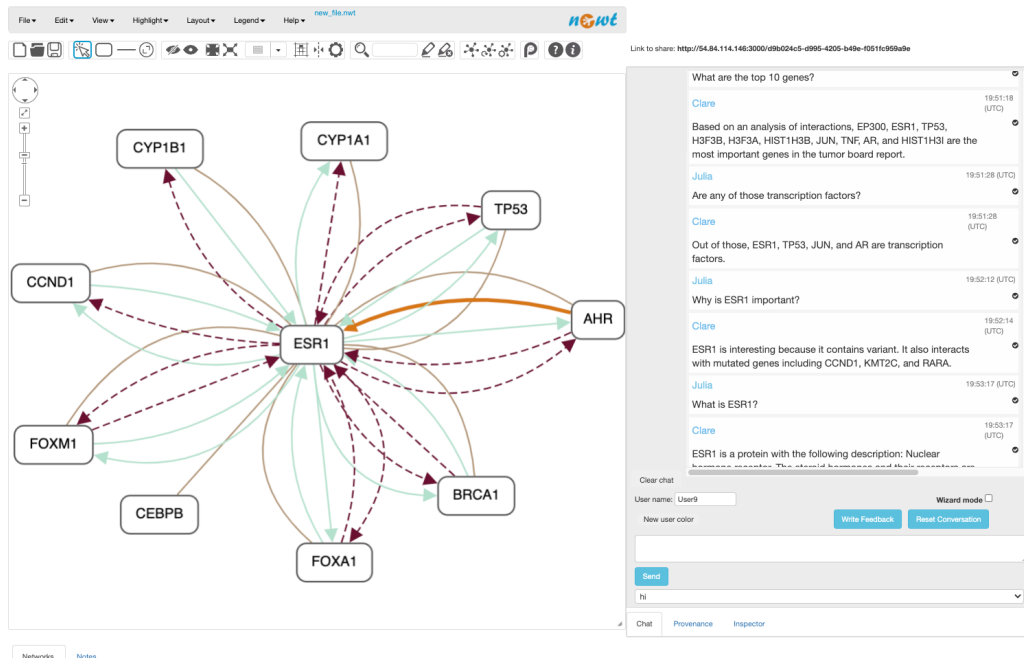


Figure 28. CLARE connected to the SBGNViz interface, showing a SIF network visualization of the most important neighbors of ESR1.

A demo video of the tumor board capability can be found here:

https://www.dropbox.com/s/gwa4jeekjubfyy/ClareTBA_Demo_Final.mp4?dl=0