# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 03-04-2023 | Final Report | 1-Jul-2020 - 30-Jun-2022 |

**4. TITLE AND SUBTITLE**
Final Report: Workshop on Data-Driven Design of Heterogeneous Materials

**5a. CONTRACT NUMBER**
W911NF-20-1-0060

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
611102

**6. AUTHORS**

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**
University of Chicago
5801 South Ellis Avenue

Chicago, IL 60637 -5418

**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)**
76955-SM-CF.1

**12. DISTRIBUTION AVAILIBILITY 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 contrued 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 | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Andrew Ferguson |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER 773-702-3018 |

# RPPR Final Report
as of 11-Apr-2023

Agency Code: 21XD

Proposal Number: 76955SMCF **Agreement Number: W911NF-20-1-0060**
**INVESTIGATOR(S):**


**Name:** Andrew Ferguson
**Email:** andrewferguson@uchicago.edu
**Phone Number:** 7737023018
**Principal:** Y

**Name:** Juan de Pablo
**Email:** depablo@uchicago.edu
**Phone Number:** 7737027791
**Principal:** N

Organization: **University of Chicago**
Address: 5801 South Ellis Avenue, Chicago, IL 606375418
Country: USA
DUNS Number: 005421136 EIN: 362177139


**Report Date:** 30-Sep-2022 Date Received: 03-Apr-2023
**Final Report** for Period Beginning 01-Jul-2020 and Ending 30-Jun-2022
**Title:** Workshop on Data-Driven Design of Heterogeneous Materials
**Begin Performance Period:** 01-Jul-2020 **End Performance Period:** 30-Jun-2022
**Report Term:** 0-Other


Submitted By: Andrew Ferguson Email: andrewferguson@uchicago.edu
Phone: (773) 702-3018



**STEM Degrees:** 0 **STEM Participants:**

**Major Goals:** Despite significant investments in various subdomains supporting materials data and design science, there is a sizable gap in developing and integrating the solutions needed to solve the larger-scale and more complex materials design problems relevant to the Army. For example, multi-functional, reconfigurable, and autonomous soft materials systems, and heterogeneous and/or composite structural materials. The complexity and scale of these challenges mandates a need for sustained multidisciplinary effort from the materials science, chemistry, physics, mechanics, mathematics, and computer science communities, and the principled integration and deployment of data-driven design (i.e., machine learning and artificial intelligence) to accelerate materials discovery and development. The proposed joint workshop will bring together data and computer scientists with scientists working in both the structural/composite hard materials and the soft functional materials communities to identify critically pressing problems and funding gaps in the data-driven design of heterogeneous materials and prescribe potential paths forward.

**Accomplishments:** During the workshop, we were fortunate to have many of the most renowned experts in material science and artificial intelligence (AI) present at UChicago. The two-day workshop has been divided into presentations of the experts in the morning and a guided discussion about the most pressing questions in the afternoon. The first day focused on soft matter materials, while the second day has the topic on hard material science. The attached report summarizes the workshop detail in the individual sections.

**Training Opportunities:** Nothing to Report

**Results Dissemination:** Nothing to Report

**Honors and Awards:** Nothing to Report

**Protocol Activity Status:**

**Technology Transfer:** Nothing to Report

# RPPR Final Report
as of 11-Apr-2023

**PARTICIPANTS:**

**Participant Type:** PD/PI
**Participant:** Andrew Ferguson
**Person Months Worked:** 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Participant Type:** Co PD/PI
**Participant:** Juan de Pablo
**Person Months Worked:** 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

**Partners**

,

I certify that the information in the report is complete and accurate:
Signature: Andrew Ferguson
Signature Date: 4/3/23 8:09PM

# Workshop on data-driven design of heterogeneous materials @UChicago: Final Report

During the workshop, we were fortunate to have many of the most renowned experts in material science and artificial intelligence (AI) present at UChicago. The two-day workshop has been divided into presentations of the experts in the morning and a guided discussion about the most pressing questions in the afternoon. The first day focused on soft matter materials, while the second day has the topic on hard material science. This report summarizes the workshop detail in the individual sections.

## Day 1: Soft Materials Contributions

Prof. Risi Kondor from the University of Chicago started the workshop with a presentation about "Equivariant Neural Networks for Molecules and Forcefields". The concept behind the new type of neural network has been envisioned first in 2016 and integrates symmetries at the level of the layers of neural networks. This concept has since been brought to maturity and is being integrated into common software tools like PyTorch. Prof. Kondor highlights the benefit of this approach for simulation applications. The physically known symmetries, like invariance under Galilei-transformation, can be included in the network architecture and do not need to be trained or artificially enforced. This aligns well with the common observation of the workshop, that AI methods become useful if they are combined with domain knowledge. This was followed by Prof. Sharon Glotzer from the University of Michigan presenting her group's research on the self-assembly of the colloidal system. The field has advanced such that the desired structure can be programmed into the individual particles. As the main challenge remains the production of industrial scales, which includes an estimation of tolerances for the individual particles. Her work employs AI techniques primarily to assist the established Monte-Carlo (MC) and molecular dynamics (MD) simulations for example for pathway optimization. Prof. Glotzer also emphasized her group's tool Signac, which automates simulation workflows to help organize data collection, important for subsequent machine learning (ML) projects. Prof. Oleg Gang from Columbia University presented "Programming nanoscale architectures and transformations" which discusses experimental aspects of the self-assembly of nanoparticles. Especially introducing DNA into the system can help tailor specific interactions. He also presented the experimental techniques to examine the final structure, which poses the challenge of combining different fidelity level data. A higher resolution is more cost-intensive compared to lower-cost techniques. AI techniques can be used to combine the two, and ML algorithms can be trained to learn the output of the high-cost technique with the input from the low-cost techniques. Prof. Juan de Pablo followed with "AI-enabled design of polymeric materials" and introduced the advances in applying AI techniques to polymer science. One significant challenge is to represent macromolecules in a latent space for ML techniques. Using graph-convolutional networks that employ common chemical motifs that are repeated in the molecules. Other steps that are required to advance polymer science with ML are production and collection of data, automated simulations that eliminate the human interaction of polymer simulations are key for this. This includes also the automated coarse-graining of the molecules, up to levels that enable full rheological characterization in silico. Prof. Nick Jackson from the University of Illinois Urbana-Champaign expands on this topic. He especially emphasizes how quantum-mechanic (QM) properties are necessary to answer today's most pressing questions. As a result, it is important to integrate QM descriptors into coarse-grained descriptions of molecules, especially macro-molecules. He presents his work that currently can incorporate targeted properties but also poses the challenge to develop a generic coarse-grained model for this purpose, similar to the MARTINI model for classical coarse-graining. Prof. Safa Jamali from the Northeastern University presented his work

about rheological informed neural networks, where he can split the work of predicting rheological curves by training a neural network to learn the underlying physical parameters, while a second neural network is used to integrate the ODEs or PDEs to yield the final rheological measurement. This workflow allows the prediction of rheological characteristics which were unobtainable otherwise. Prof. Nicolas Kotov from Michigan State University presented his work on the interplay between nano-particles and proteins. The geometrical interplay between self-assembled nano-particles enables potential medical applications. For ML techniques this can be encoded as graphs, which in turn can be used to train neural network models. Day 1 was closed by Prof. Peter Voorhees from the Northwestern University discussing additive manufacturing of metals. Especially, understanding the local melting processing caused by lasers heating the material. The resulting grain structure is different than conventional manufacturing, resulting in differing fatigue characteristics. Conventional, physical models can fully describe and understand the challenge; however, the computational and storage demands prohibit upscaling the technique to analyze and tailor the manufacturing process for each application. The inception of an ML technique that can be trained to predict this processing with less computational resources is required for industrial uptake.

## Day 1: Soft Materials Discussions

The first question addressed was how best to collect and store data so that it can be preserved for future use. This issue encompasses two distinct areas. First how we do preserve data that is currently being generated, and second, what should we do with legacy data that already exists in publication. The latter question is not easily solved. This legacy data rarely has proper metadata attached and is seldom in a format where it may be easily extracted. While it may be tempting to ignore this data, it likely contains prior measurements of materials and formulations that are unlikely to be repeated in the future, and thus must be preserved in some way. Additionally, the questionable legality of data scraping in some cases may further increase the difficulty of putting this legacy data in a convenient form.

One major improvement for currently produced data is to ensure that it is easily searchable. To further improve upon this, publication of and making available data should be structured to benefit the PI who generates the data, which is not always the case currently. Furthermore, the keeping of this data will likely have substantial costs attached, not just due to storage, which is relatively cheap, but in the need for administrators to manage this data. Alongside the data from experiments that worked, it is vital to ensure that dark data, i.e., negative results are also shared and made available such that prior research is not repeated. However, at present, there is very little gain in publishing or making this data available. The obvious solution is a stick approach in which data is forced to be included via the review process any time someone wishes to publish. However, even doing so leaves open a question of data quality and reliability.

Vital to the question of data publishing is the question of how this data should be formatted. While metadata is vital for presenting a complete dataset, metadata alone does not provide a complete understanding of the data, and frequently lacks context, such as the actual workflow and why it was chosen. Compounding the formatting problem, there is a wide array of diverse solutions being proposed for data storage, all with incompatible independent formats and standards. The most obvious solution is to set out a pre-agreed format, but reaching a consensus on such matters has proved difficult.

The answers to many of these questions, it was pointed out, likely lie in the historical record. There exists already both crystal structure and protein databanks that are well established and

widely utilized. From these (and other potential efforts) it may be possible to distill a path forward to an openly available data bank for soft materials. Finally, with these efforts, the question was asked, but not answered, whether small single PI, or large multigroup efforts are more effective for establishing such databases.

The next area which was discussed was: Do we have the necessary tools to study heterogeneous material? This area proves difficult because the materials exhibit wildly different properties, and there is a potential that a new solution is needed for each new system of interest. Additionally, anytime the time and length scales are changed so does the resolution, and so there is some loss of model fidelity when mapping from one simulation scheme to another, which should be minimized.

On the industrial side, it was stated that these models do not need to be perfect. Instead, ideally, they need to just be able to point towards a new direction of study. Thus far the human element has been necessary to determine a path forward. There is a fundamental desire that these ML models be capable of providing an understanding of the data. For obvious reasons, if this is true, explainable AI is necessary going forward. It should also be mentioned that what is desired is to use ML to find new, novel materials, rather than just providing small incremental improvements. However, it is unclear whether or not this is truly a possibility. One example of this is superconducting materials, where high- and low-temperature superconductors each are the result of distinctly different underlying physical mechanisms, and an AI trained on one type would never discover any of the other types.

## Day 2: Hard Materials Contributions

Prof. Chris Wolverton (Northwestern) opened the workshop on the second day by presenting on “network theory and machine learning in materials discovery”. Broadly, Prof. Wolverton discussed a database (OQMD) for DFT, machine learning of material properties, and applying ideas from network theory to materials science. OQMD (Open Quantum Materials Database, oqmd.org) is an open-source and searchable large-scale database with DFT calculated properties created using structures primarily from ICSD (Inorganic Crystal Structure Database) for ~1M inorganic compounds to computationally screen materials. To develop ML models for materials discovery using either OQMD or other databases, Prof. Wolverton then discussed the need for including structure information of materials through either Voronoi tessellation or graphs. In addition to materials discovery, ML models can be applied to guide the data collection (DFT calculations) aspect of building the database. Prof. Wolverton then discussed the analysis of a 100-dimensional phase diagram network with phases as nodes and linkage between them as edges to study topological features of materials discovery over time and for developing relational AI to predict whether a material should be synthesized. On a topic of common interest in the workshop, Prof. Dane Morgan (UW Madison) then talked about “uncertainty quantification in materials property predictions with ML models”. This includes a discussion on the need for FAIR ML models and databases. Currently, there is limited reuse of ML models in the MS&E community but there is progress in generating FAIR models and databases such as Foundry by Prof. Ian Foster. Prof. Morgan then discussed the limitations and challenges of estimating uncertainty on ML model predictions such as the estimation of variance from ensemble models using bootstrapping. To address the issue, Prof. Morgan discussed a calibration method for improving the accuracy of the uncertainty measurement (error on the error) with applications in a wide range of MS&E problems. The calibration method improves the accuracy of uncertainty (standard deviation) by assuming that the true standard deviation is linearly related to the standard deviation estimated using bootstrapping. The proportional parameters are evaluated by a log-likelihood optimization

method. While significant work is still required to further improve uncertainty quantification, the calibration method seems to improve the accuracy of the uncertainty of ML models for several MS&E problems. Prof. Taylor Sparks from the University of Utah then presented “challenges and opportunities with featurization and algorithms for materials informatics”. This includes a discussion on the general characteristics desired for ideal features that allow invertible design, design of novel and interesting materials, predict structure-properties-performance, interpretability, quantification of uncertainty and that can work with small data sets. Compared to GPT-3 language models, the majority of materials problems don’t have sufficient data. For this, active learning frameworks can be applied to efficiently navigate the design space and discover novel functional materials by collecting quality data. For exploring unknown space, ideas from generative models such as Dali.E2 can probably be applied to materials discovery. To this end, Prof. Sparks highlighted CrabNet (Compositionally Restricted Attention-Based Network) from his group, which can predict material properties based on only chemical formulas. For discovering an optimal combination of elements to generate new material, Prof. Sparks also talked about the applicability of black-box optimization tools or genetic algorithms. In addition, for the discovery of novel materials, the usage of generative models was discussed. The last presentation of session 1 on day 2 was given by Dr. Samuel Schoenholz from Google Brain on “differentiable programming for materials design”. In this talk, Dr. Schoenholz described the automatic differentiation tool (autograd) implemented in JAX for efficiently computing exact derivatives. This is utilized in different fields including the end-to-end differentiable molecular dynamics MD software (JAX MD). Differentiable programming allows the incorporation of ML directly into simulations or training ML using physics-based simulations. JAX MD is efficient and written only in one language which could enable fast calculations of gradients in simulations. One easy application to envision is force field parameterization but other novel applications may be possible. Prof. Ian Foster (U.Chicago) opened session 2 of the workshop on day 2 by presenting “a collaborative data foundation for data-driven materials design”. In this talk, Prof. Foster focused on enabling collaboration between researchers and sharing FAIR data at scale using Materials Data Facility (MDF - https://materialsdatafacility.org). MDF allows sharing of datasets from simulations and/or experiments, ML models, and metadata from experiments. It generates a citable DOI that can be a valuable incentive, which is one of the key questions raised related to sharing experimental metadata in the workshop. The platform is mostly automated and currently, there are ~1000 users. When fully automated, the platform should in principle collect data, automatically process data into an appropriate format and predict or guide the design of new experiments for discovering new materials. Dr. Corrinne Lipscomb (3M) then presented “an industrial perspective on ML research and future challenges”. After presenting an overview of the focus research areas of 3M, Dr. Lipscomb discussed the importance of simulations or ML models (even low fidelity models) for reducing the industrial cost associated with experiments. Dr. Lipscomb also described the need for collecting targeted legacy data and metadata in experiments as they are typically influenced by various factors including the protocol used in experiments. To this end, 3M made some progress in building the necessary infrastructure and it appears that the industrial culture is transitioning to realize it. Then, Prof. Andrew L. Ferguson (U. Chicago) presented on “data-driven design of self-assembling optoelectronic peptides”. In this talk, Prof. Ferguson discussed an integrated experimental and computational framework using molecular simulations, machine learning, and spectroscopy-based experiments to design self-assembling optoelectronic peptides. By combining low-fidelity computational results with high-fidelity experimental trends, a multi-fidelity approach was developed using Gaussian Process Regression and Bayesian optimization (BO) to efficiently navigate the vast sequence design space of peptides for discovering novel optoelectronic peptides. This talk highlighted the importance of both low fidelity models and active learning frameworks for efficiently discovering materials that can also be applied in other areas of MS&E. The last talk of session 2 was given by Prof. Sergei Kalinin (UT Knoxville) on “machine learning for scanning probe and electron

microscopy for materials discovery". Prof. Kalinin highlighted some of the problems with BO that ignores physics. This was demonstrated by applying data from scanning probes and electron microscopy to discover materials and using human knowledge to carefully choose the prior. Prof. Kalinin also demonstrated the application of machine learning to generate hypotheses and design experiments. When realized, this can probably be incorporated into the database platform discussed by Prof. Foster for the completely automated discovery of novel functional materials. In addition, Prof. Kalinin emphasized the importance of physics or domain knowledge in developing ML models for automation of experiments for efficiently discovering new materials.

## Day 2: Hard Materials Discussions

Day two of the materials discussion focused on several key questions relevant to recent developments in machine learning and materials discovery that were detailed during the morning session.

The first question which was discussed was what advancements might be possible in physics and material science with increased use of differentiable programming. These differential programming techniques allow for automated differentiation of any computer function which in turn opens the door for gradient descent or other optimization techniques to be applied with relative ease. Immediate avenues for this tool that were mentioned were better fitting of experimental data, and a differentiable nudged elastic band method, both of which have been previously implemented. Of great interest was the possibility for differentiable DFT type calculations, however, it was noted that this area has thus far been confined almost entirely to toy problems. Other areas that could benefit from such calculations are free energy sampling methods in molecular dynamics, as well as improved optimization of hierarchical and coarse-grained simulations, but neither of these has yet been reported using differentiable programming. The next point which was discussed was how can we best integrate code and experiment, as well as better, disseminate these tools. On this note, there was one point of universal agreement. GitHub is entirely insufficient for sharing code amongst scientific users. Making code available on a repository only partially allows for reproducibility, as such details as the version used for data generation and the hardware the code is run on remains unknown. It was also noted that frequently sharing code between industry and academia can be difficult as the industry utilizes proprietary code that can't be shared. In regards to connecting code to an experiment, it was noted that this feat can prove quite challenging as well. Most lab tools do not natively include a python interface and so solutions must be hacked together. However, it appears this issue is slowly abating as more companies include such interfaces for their products. On one final point, it was mentioned that likely data from different sources (electron microscopes, photon sources, etc) may benefit from being segregated into multiple separate repositories where their respective communities can find them, as opposed to being heaped together.

Related to the above it was further discussed how high throughput experimentation and self-driving labs can enhance materials exploration and discovery. Generally, it was noted that such high throughput systems are becoming more common, for example, DOW uses high throughput for characterizing small molecules and catalysts. A key observation in these high throughput systems is there always exists a bottleneck in these systems whether it be synthesis, characterization, or analysis. Thus, a holistic approach that keeps in mind all three areas is necessary. However, such high throughput characterizations are frequently incomplete and only contain a single piece of data. Regarding analysis, it is likely fast but approximate methods, capable of indicating a generally forward direction in material space, are necessary to keep up with synthesis and characterization, due to the expense of retraining models.

It was also noted that combinatorial synthesis is vastly more difficult in soft materials than in hard materials. In soft materials capturing processing information is vital for reproducibility but not easily accomplished. Furthermore, in an R&D setting the relevant process information is not known a priori, and thus requires recording vast amounts of information and a complete process description. Even if this task is accomplished, there are still going to be hidden variables not listed in experimental logs.

The next question was the perennial issue of how best to drive connection between industry and academia. The key amongst the answers was that collaboration naturally encourages data sharing between both parties. Even so, open publishing and sharing between academia and industry can be difficult due to the protectiveness of data sets. One option to bypass this is so-called masked data sets. However, these are only partial solutions. These masked data sets while somewhat useful can make a physics-based understanding of the underlying phenomena quite difficult. Additionally, frequently finished industrial products are not good model systems, which complicates their study in academia. Rather than simply one or two-component mixtures, frequently these materials contain a wide array of additives which makes their study in a traditional academic setting difficult. Finally, it was noted that the key to driving further partnership is successful examples of collaboration where such success will drive the necessary cultural change.

The final question discussed was is it possible to make good statistical models with limited data. The essential answer to this question was that physical laws should be incorporated into these models. Doing so drastically reduces the amount of data required.

**ATTENDEES**

| | |
|---|---|
| Andrew Ferguson | UChicago |
| Chris Wolverton | Northwestern |
| Corinne Lipscomb | 3M |
| Dane Morgan | UW Madison |
| Evan Runnerstrom | ARO |
| Giulia Galli | UChicago |
| Heinrich Jaeger | UChicago |
| Ian Foster | UChicago |
| Joseph Palomba | DEVCOM Soldier Center |
| Joshua Mysona | UChicago |
| Juan de Pablo | UChicago |
| Klavs Jensen | MIT |
| Ludwig Schneider | UChicago |
| Mark Tschopp | ARL |
| Matt Guziewski | ARL |
| Michael Webb | Princeton |
| Nicholas Jackson | UIUC |
| Nicholas Kotov | U Michigan |
| Oleg Gang | Columbia |
| Peter Voorhees | Northwestern |
| Randi Christensen | 3M |
| Risi Kondor | UChicago |
| Safa Jamali | Northeastern |
| Samuel S. Schoenholz | Google Brain |
| Sergei Kalinin | Tennessee |
| Sharon Glotzer | U Michigan |
| Siva Dasetty | UChicago |
| Sukrit Mukhopadhyay | Dow |
| Taylor Sparks | Utah |
| Wenxiao Pan | UW Madison |
| William Klein | DARPA |