Conference Report: 2018 Materials and Data Science Hackathon (MATDAT18)

Andrew L. Ferguson, Institute for Molecular Engineering, University of Chicago
Tim Mueller, Department of Materials Science and Engineering, Johns Hopkins University
Sanguthevar Rajasekaran, Department of Computer Science & Engineering, University of Connecticut
Brian J. Reich, Department of Statistics, North Carolina State University

Synopsis

The National Science Foundation (NSF) 2018 Materials and Data Science Hackathon (MATDAT18) took place at the Residence Inn Alexandria Old Town/Duke Street, Alexandria, VA over the period May 30 – June 1, 2018. This three-day collaborative "hackathon" or "datathon" brought together teams of materials scientists and data scientists to collaboratively engage materials science problems using data science tools. The materials scientists brought a diversity of problems ranging from inorganic material bandgap prediction to acceleration of *ab initio* molecular dynamics to quantification of aneurysm risk from blood hydrodynamics. The data scientists contributed tools and expertise in areas such as deep learning, Gaussian process regression, and sequential learning with which to engage these problems. Participants lived and worked together, collaboratively "hacked" for several hours per day, delivered introductory, midpoint, and final presentations and were exposed to presentations and informal interactions with NSF personnel. Social events were organized to facilitate interactions between teams. The primary outcomes of the event were to seed new collaborations between materials and data scientists and generate preliminary results. A separate competitive process enabled participants to apply for exploratory funding to continue work commenced at the hackathon. Anonymously surveyed participants reported a high level of satisfaction with the event, with 100% of respondents indicating that their team will continue to work together into the future and 91% reporting intent to submit a white paper for exploratory funding.

Objectives

The exponential increase in available computing power has made it possible to generate and analyze large amounts of materials data. Initiatives such as the Materials Project, OQMD, AFlowLib, and NOMAD have created publicly accessible databases containing the structure and properties of tens of thousands of materials, and individual research groups are generating large data sets for more specific materials research problems. One of the leading challenges in materials science and engineering is determining how to best make use of this abundance of materials data to accelerate the development of new understanding and novel technologies. Despite the considerable progress that has been made in the application of data science to materials science in recent years, there is still a fundamental problem in that most experts in materials science and engineering are not experts in data science, and vice versa. Thus, it is difficult for materials researchers to effectively make use of leading data science techniques, and data scientists have limited insight into how they can apply their knowledge to problems in materials science and engineering in the most impactful ways.

The first objective of the 2018 Materials and Data Science Hackathon (MATDAT18) was to assemble new interdisciplinary teams – each composed of materials researchers and data scientists – to work together in applying advanced data science methods to address important and challenging problems in materials science and engineering. Success in this goal will seed new collaborations and generate preliminary data for future funding opportunities. A second aim was in forging connections and promoting cross-fertilization between the materials and data science communities. The hackathon provides for close interactions between participants wherein materials researchers are exposed to cutting-edge statistics and machine-learning techniques, and data scientists are motivated to develop new methods to analyze novel data streams produced by the materials community.

Organization and Solicitation

Funding for the hackathon was provided by a grant from the National Science Foundation. The organizing committee for the hackathon consisted of two materials scientists, Andrew Ferguson (University of Chicago) and Tim Mueller (Johns Hopkins University), and two data scientists, Sanguthevar Rajasekaran (University of Connecticut) and Brian Reich (North Carolina State University). The distribution of expertise, interests, and professional affiliations among the committee members facilitated outreach efforts to the various communities of researchers who might participate in the hackathon and ensured the necessary level of expertise to evaluate the variety of proposals that were received. The committee collaboratively authored a proposal for the conference that was peer-reviewed and funded through the NSF/DMR/CMMT program. The organizing committee coordinated among themselves and with NSF stakeholders through email and scheduled videoconferencing calls.

Advertising for the hackathon was done through personal contacts, mass emails, a web site (matdat18.wordpress.ncsu.edu), and announcements at various meetings, including:

- The Fall 2017 meeting of the Materials Research Society
- The 2017 NSF EFRI-2DARE/DMREF-2D/MIP Grantees Meeting
- The 2017 NSF Nanoscale Science and Engineering Grantees Conference
- The TMS 2018 Annual Meeting & Exhibition
- IEEE International Conference on Data Mining (ICDM) 2017
- IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2017
- Email list for the IEEE International Conference on Computational Advances in Bio and medical Science (ICCABS)
- Email list for the IEEE International Conference on Big Data
- Northeast Big Data Innovation Hub
- AIChE 2017 Annual Meeting
- Aspen Center for Physics 2018 Winter Conference: Data-driven Discovery and Design in Soft and Biological Materials

- APS 2018 March Meeting GSOFT Short Course: Machine Learning and Data Science in Soft Matter
- ASA Section on Physical and Engineering Statistics
- ASA Section on Statistical Learning and Data Science

Additional advertising was provided by the journal Molecular Systems Design & Engineering, both on the web and via Twitter, and through Calphad.org.

Each hackathon team typically consisted of two materials researchers and two data scientists. To assemble the interdisciplinary teams who competed in the hackathon, the solicitation of the hackathon proceeded in two stages. In the first, materials researchers were asked to submit descriptions of problems in materials science and engineering that could potentially be addressed through the application of data science methods. In these descriptions the materials researchers included a brief description of the data set, its availability, and the project objectives. The materials proposals were screened by the organizing committee and those that were determined to be responsive to the call were put on a publicly available web site that was advertised to the data science community. In this stage, 26 applications were received, of which 21 were determined to be responsive.

In the second stage of the solicitation, data scientists were asked to describe how they proposed to address up to three of the materials science problems. Proposals from 20 different data science teams were received. Many teams submitted proposals for more than one materials problem, resulting in a total of 34 proposals. From these proposals, the organizing committee selected 14 pairings of materials and data teams that were believed to have the greatest chance for a successful collaboration by combining a compelling materials science problem with appropriate data science tools. Of those selected, 12 teams consisting of a total of 38 researchers were able to participate in the hackathon. The teams consisted of professors, postdocs, students, industry researchers, and researchers from government labs. Of the participants, 28 were from universities, 5 from government labs, and 4 from industry. Most of the attendees came from the United States (34), with the remainder from Nigeria (2), Denmark (1), and Sweden (1). A photograph of the conference organizers and participants is presented in **Figure 1**.

Figure 1. MATDAT18 organizers and participants.

Project Topics and Teams

A total of 12 teams participated in the hackathon. A listing of the team members and project titles are provided in **Table 1**. The first day of collaboration began primarily with the materials teams describing the data and objectives to data teams, and the data teams exploring the data and making analysis plans. Most of the teams had been in e-mail contact before the meeting which made this process easier but establishing common language and defining roles remained a challenge. After becoming acquainted with each other and the problem, the teams began exploring more sophisticated analyses and refining the scope of the project as dictated by preliminary results. Because the event was spread over three days, most teams had sufficient time to try several approaches and identify shortcomings of the current data and the most promising avenues for future research.

Table 1. Teams and project topics.

Team	Topic	Materials Scientists	Data Scientists
1	Computational discovery of novel organic metals and narrow-gap semiconductors with generative models	Bart Olsthoorn and Matthias Geilhufe (Condensed Matter, Statistical and Biological Physics, NORDITA)	Stanislav Borysov (Management Engineering, Technical University of Denmark) Ranjan Srivastava (Chemical & Biomolecular Engineering, University of Connecticut)
2	Characterizing protein hydrophobicity using high dimensional descriptors	Nicholas Rego (Chemical and Biomolecular Engineering, University of Pennsylvania)	Victor Osamor and Emmanuel Adetiba (Department of Computer and Information Sciences, Covenant University)

3	Dilute solute diffusion	Benjamin Afflerbach (Materials Science & Engineering, University of Wisconsin – Madison)	Lay Wai Kong (Intel Corporation)
4	Development of a data-driven method to optimize ReaxFF force field	Mert Sengul (Materials Science and Engineering, Pennsylvania State University)	Tirthankar Dasgupta and Ying Hung (Statistics and Biostatistics, Rutgers University)
5	Predicting band edge positions of perovskite photocatalysts for water-splitting application	Yihuang Xiong and Weinan Chen (Materials Science and Engineering, Pennsylvania State University)	Hua Wei and Wenbo Guo (Information Science and Technology, Pennsylvania State University)
6	Mitigating hazards posed by stretchable electronic circuits: Liquid metal embrittlement by exposure of engineering alloys to eutectic gallium indium	Victoria Miller (Materials Science and Engineering, North Carolina State University)	Carena Church (Citrine Informatics)
7	Machine learning for structure-performance relationships in organic semiconducting devices	Evan Miller and Matthew Jones (Materials Science and Engineering, Boise State)	Bryan Stanfill (Applied Statistics and Computational Modelling, Pacific Northwest National Lab)
8	Unsupervised classification of nanostructured thin films	Wesley Tatum (Materials Science and Engineering, University of Washington)	Patrick O'Neil and Diego Torrejon (Spaceflight Industries)
9	Finding predictive descriptors for singlet fission: Revealing fundamental physics in data	Xingyu Liu and Noa Marom (Materials Science and Engineering, Carnegie Mellon)	Laura Wendelberger and Brian Reich (Statistics, North Carolina State University) Matthew Spellings (Chemical Engineering, University of Michigan) Bradley Dice (Physics, University of Michigan)
10	Data-driven analysis of nanoscale chemical structure and electrical function	Jessica Kong (Chemistry, University of Washington)	Karl Pazdernik and Sarah Reehl (Applied Statistics and Computational Modelling, Pacific Northwest National Lab)

11	High fidelity universal prediction of bandgaps in inorganic materials	Bharat Medasani (Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory)	Sumit Kumar Jha and Sunny Raj (Computer Science, University of Central Florida)
12	Quantifying rupture risk of brain aneurysms by combining morphological descriptors and blood flow data from large-scale lattice Boltzmann simulations	Mehrdad Yousefi and Ulf D. Schiller (Materials Science and Engineering, Clemson University)	Benjamin Erichson and George Stepaniants (Applied Mathematics, University of Washington)

Hackathon Format and Schedule

The hackathon was scheduled to take place in Alexandria, VA in order to be proximate to the NSF headquarters and facilitate participation and attendance of NSF stakeholders. The Marriott Residence Inn was selected as a venue for the availability of conference facilities, capacity to accommodate all participants within a room block, and favorable group rates. It was key to the success of the event that all participants stayed in the same hotel so as to support the strongly collaborative and interactive nature of the event. A large conference room arranged with presentation facilities and shared round tables was reserved for the hacking, although some teams chose to work in nearby common areas of the hotel. Hacking proceeded during the allocated time periods in the schedule and also during the evenings. The team messaging application Slack was used to facilitate communication and file sharing between the organizers and participants, and within the teams themselves. A number of teams continued to use this mode of communication beyond the close of the hackathon providing continuity and a relatively frictionless mode of communication. The hackathon schedule was assembled in line with best practices and historical experience to provide large blocks of uninterrupted time for collaborative hacking interspersed with breaks, meals, social events, and short talks. Formal presentations from NSF program officials exposed participants to existing and upcoming funding opportunities, and one-slide / two-minute "lightning presentations" from the teams on each of the three days provided introductory, midpoint, and final updates on team progress. The lightning talks were valuable in providing structure to the event, imposing accountability upon the teams, and in illuminating possibilities for collaborative interaction between teams. The event schedule was as follows:

Wednesday 5/30

9:00 — Introductions and orientation
9:15 — Presentation: "MATDAT18: Welcome and comments"
9:30 — Lightning intro presentations
10:00 — Hacking!
12:30 — Lunch

1:30 — Presentation: "Good practices for interdisciplinary research"
2:00 — Hacking (coffee at 3:00)!
5:00 — Social hour sponsored by Citrine

Thursday 5/31
9:00 — Lightning midpoint reports
10:00 — Hacking!
12:30 — Lunch
1:30 — Presentation: "More data at DMR: DMREF and beyond"
2:00 — Hacking (coffee at 3:00)!
6:00 — Social hour

Friday 6/1
9:00 — Final hacking!
10:00 — Lightning final reports
12:00 — NSF program officer panel
1:00 — Wrap and close

Outcomes, Perceptions, and Reflection

The intermediate and final presentations made it clear that the teams were able to accomplish a lot over the course of the three-day hackathon. In general, the teams tried multiple data science approaches to solve the materials science problem of interest and typically obtained promising preliminary results. Results generally appeared strong enough to warrant continued collaboration, application for follow-on funding, and ultimately publication in archival journals. We provide below final summaries provided by selected teams upon conclusion of the hackathon to illustrate the objectives, approaches, and outcomes of particular projects.

Team 4: Development of a data-driven method to predict ReaxFF force field parameters
Mert Sengul (Materials Science and Engineering, Pennsylvania State University)
Tirthankar Dasgupta and Ying Hung (Statistics and Biostatistics, Rutgers University)

The ReaxFF is a reactive force field capable of simulating bond formation/breaking along with dynamics of large molecular systems at elevated temperatures and pressures for long simulation times. It is widely used in the materials science community, producing around 700 publications in literature. The smallest ReaxFF force field parameter set is composed of around 300 parameters that must be optimized before application to different physical systems. Given the popularity of ReaxFF, its performance and usability involve quality and convenience of the optimization algorithm that is challenging due to high dimensionality and complex interactions. The data is generated through complex simulation models based on Newtonian mechanics. Our objective is to address this problem through application of a systematic data-driven framework that consists of efficient design for simulating combinations of FF parameters, fast statistical surrogate modes based on Gaussian process, and efficient global optimization approaches. In our

preliminary study during the Hackathon, we implemented tools like Latin hypercube designs, Gaussian process models, and the expected improvement procedure to develop an efficient global optimization of small groups of FF parameters and will be working on scaling up the framework to high-dimensional settings.

Team 6: Mitigating hazards posed by stretchable electronic circuits: Liquid metal embrittlement by exposure of engineering alloys to eutectic gallium indium
Victoria Miller (Materials Science and Engineering, North Carolina State University)
Carena Church (Citrine Informatics)

Liquid metal bearing electronics are a potentially transformative technology for stretchable electronics and reconfigurable antennas. However, liquid metals can catastrophically degrade the mechanical properties of the solid metals they contact, i.e. the liquid metal will embrittle the solid. There are no existing methods to predict whether a given liquid metal will embrittle a given solid metal. A combination of data mined from the literature and preliminary experimental results were used to train machine learning models on the Citrination platform. The highest performing model was used for sequential learning (SL), a data-driven optimal experimental design framework that narrows the alloy space to be experimentally probed. The first iteration of SL identified alpha Ti alloys as a promising candidate for mechanistic investigation of embrittlement; they have already been ordered and will be tested within a week of the Hackathon.

Team 7: Machine learning for structure-performance relationships in organic semiconducting devices
Evan Miller and Matthew Jones (Materials Science and Engineering, Boise State)
Bryan Stanfill (Applied Statistics and Computational Modelling, Pacific Northwest National Lab)

Organic electronic devices are becoming increasing promising alternatives to their inorganic counterparts, due in part to inexpensive device fabrication and fast return-on-investment. The efficiency of these devices is strongly dependent on the molecular morphology, which describes the nanoscale structure resulting from the self-assembly of molecules. Morphology is strongly influenced by materials choices, chemistry, and processing conditions resulting in a vast phase space that necessitates computational methods to explore. Currently, determining the electronic efficacy of organic materials is a computationally intensive process, requiring of the order 10,000 slow quantum chemical or semi-empirical calculations for a single morphology. Our goal is therefore to explore machine learning algorithms to model important electronic features in a fraction of the computational runtime in order to permit a large sweep of the organic phase space. We tried several linear and non-linear machine learning techniques, including support vector machines, artificial neural networks, and random forests to model the electronic coupling between a variety of molecules, with data generated previously using our open-source MorphCT software package

(https://doi.org/10.5281/zenodo.1243843). We have found that a random forest method provides the best agreement to the calculated data, with a correlation coefficient of 98.7% for our test polymer system. The average errors on the important electronic properties are within the prediction uncertainty of the quantum chemical methods, suggesting that our machine learning methodology could successfully replace the more computationally expensive techniques in our current simulation pipeline.

Team 8: Unsupervised classification of nanostructured thin films
Wesley Tatum (Materials Science and Engineering, University of Washington)
Patrick O'Neil and Diego Torrejon (Spaceflight Industries)

Thin films of semiconducting materials will enable stretchable and flexible electronic devices, but these thin films are currently stochastic and inconsistent in their properties and morphologies because processing and chemical conditions influence the mixing and domain size of the different components. By using atomic force microscopy (AFM), a cheap and quick technique, it is possible to spatially resolve and quantify these different domains based on differences in their mechanical properties, which are strongly correlated to their electronic performance. For this project, a library of AFM images has been curated, which includes poly(3-hexylthiophene) that has been processed in different ways (e.g. annealing time and temperature, thin film vs nanowire), as well as thin film mixtures of PTB7-th and PC71BM. To analyze these samples, several semantic segmentation methods from the fields of machine learning and topological data analysis are employed. Among these, a Gaussian mixture model utilizing machine learned local geometric features proved effective. From the segmentation, probability distributions describing the mechanical properties of each semantic segment can be obtained, allowing the accurate classification of the various phase domains present in each sample.

Team 9: Finding predictive descriptors for singlet fission: Revealing fundamental physics in data
Xingyu Liu and Noa Marom (Materials Science and Engineering, Carnegie Mellon)
Laura Wendelberger and Brian Reich (Statistics, North Carolina State University)
Matthew Spellings (Chemical Engineering, University of Michigan)
Bradley Dice (Physics, University of Michigan)

Singlet fission is a rare phenomenon observed in organic molecular crystals that increases the Shockley-Queisser efficiency limit from 33% to 47%. However, the prohibitively high cost to precisely calculate the thermodynamic driving force hinders screening of large datasets for singlet fission candidates. Our objective in the hackathon was to use machine learning methods to build a predictive model for the results of the high-fidelity evaluation of a material's performance to optimize the selection of future experiments. A LASSO technique is utilized for selection of cheminformatic variables to estimate the target property in order to narrow the field for DFT candidates. We then perform linear regression on the 16 DFT features. The machine learning results suggest

a useful workflow where low-fidelity cheminformatic data can be used to guide a series of further simulations, thereby accelerating the materials discovery process.

Team 10: Data-driven Analysis of Correlations between Chemical Structure and Electrical Function on the Nanoscale
Jessica Kong (Chemistry, Washington)
Karl Pazdernik and Sarah Reehl (Applied Statistics and Computational Modelling, Pacific Northwest National Lab)

The goal of this project is to develop and understand the relationship between chemical composition and electrical function via pixel-to-pixel analysis of multimodal atomic force microscopy images. The materials objectives are to determine the spectra of pure components within a material such that fractional abundances can be determined and estimate the electrical properties of a material based on its chemical composition. The data comprised hyperspectral photoinduced force infrared and conductive atomic force microscopy images. We applied nonnegative matrix factorization (NMF) to obtain component spectra and random forest regression, and convolutional neural network to predict electrical current from hyperspectral infrared information. With NMF, we obtained spectra that were closer to component spectra than with principal component analysis. The best predictions of electrical current are obtained by regressing onto each spectroscopic dimension of the hyperspectral data with random forests.

Exploratory funding opportunities were provided through NSF EAGER grants through the NSF/DMR/CMMT program. White papers were solicited from interested teams and a subset of these applications invited to submit full EAGER proposals through an independent competitive peer-review process. The white papers were divided into two categories: Type 1 and Type 2. Type 1 was intended to fund teams that worked together in the hackathon, and Type 2 intended to support new teams.

An anonymous and voluntary exit survey was administered to all participants upon conclusion of the hackathon to which 23 of the 38 participants responded. The survey was designed to assess participant perception and satisfaction with the event and collate feedback on what aspects of the event could be improved through a combination of numerical polls and short-form responses. The collated numerical responses are provided in **Table 2**, and a summary of written responses provided below.

Table 2. Summarized participant responses to numerical exit survey questions (n = 23).

Poll Question	**Response**
How was your overall experience of MATDAT18? (1-10 scale, 1 = terrible, 10 = excellent)	$\mu = 9.1$ $\sigma = 0.8$
How useful was MATDAT18 in enabling progress towards your objective? (1-10 scale, 1 = not at all useful, 10 = extremely useful)	$\mu = 8.7$ $\sigma = 1.2$
Will your team continue to work together after MATDAT18?	Yes (23) No (0)

Do you plan to submit an NSF EAGER white paper?	Yes (21) No (2)
Would you be interested in participating in MATDAT19 or MATDAT20?	Yes (23) No (0)

Participant Experiences and Perceptions

- We have learnt that unsupervised classifications of phase domains in AFM images enables quantitative relationships for thin film material properties and processing conditions. There were a lot of insights that occurred and collaborations that resulted from this hackathon.
- Getting to know data scientists who are interested in materials problems and seeding new collaborations, hearing about what other teams were doing was interesting, we made significant progress in a very short time.
- I learned new techniques I was previously unfamiliar with. Also, it reinforced my impression regarding the strength of random forests as an ML strategy.
- We ended up only using very basic data science tools, but I think it was still useful for our materials collaborators.
- We identified both unsupervised and supervised learning approaches that improved accuracy.
- Using machine learning we can save ~24 hours of computation time (on a high-performance computing cluster) per system, which will massively improve the throughput of our organic electronic materials phase sweep.
- The sequential learning approach identified a material candidate that is easily tested and would meet our criteria for "ideal candidates".
- First, we realized our current dataset is not large enough to go further, so we hacked around to get more chemical predictors. Second is we realized part of our structural dataset did not contribute much to approximating target value.
- We now have a method to initiate a better parameter combinations for optimization. And we started working on some procedures to make optimization easier.
- The material project was using a K-nearest neighbor approach. The data science team used a mixture of Gaussian to take into account all the mechanical properties. The new approach is more effective and faster.

Suggestions for Improvements

- A larger room or possibly multiple rooms could be beneficial to reduce the noise from the other teams. Perhaps the hackathon could be conducted in a university setting.
- One more day of the hackathon.
- Initial presentations from the teams could be longer (than the current 2 minutes). More details on the problem addressed could help the team members.
- An advance (slightly more detailed) schedule could help.
- A short seminar or structured discussion from a researcher about their research at the intersection of materials science and data science will be nice.
- Speakers could be provided with microphones.
- The panel discussion could probably be done on the second day afternoon.

- Knowledge across the groups has to be exchanged.
- The data science teams could present a poster of some of their current work in the evening.
- Color-coded name tags (to distinguish MAT and DAT participants).
- Better (and healthier) food.

Feedback indicated that attendees on both the materials and data sides enjoyed worthwhile and productive experiences at the hackathon and provided a number of constructive suggestions for how to improve the event. Encouragingly, 100% of respondents reported 8/10 satisfaction or higher, 100% indicated interest in attending a similar event in the future, and 91% intended to submit white papers for follow-on NSF EAGER funding. These results attest to the value of the hackathon model in bringing together materials and data scientists, advancing materials science research, and the desire for such events within the materials and data communities.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMR-1748198. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The organizers of MATDAT18 gratefully acknowledge the support of Molecular Systems Design & Engineering journal, a joint venture between the Royal Society of Chemistry and the Institution of Chemical Engineers, in publicizing and supporting the event, and of Citrine Informatics in sponsoring a social hour for hackathon participants.