CellPress

Matter

Matter of Opinion

Teaching machine learning to materials scientists: Lessons from hosting tutorials and competitions

Shijing Sun,[1] Keith Brown,[2] and A. Gilad Kusne[3,4,*]

The growing field of data-driven materials research poses a challenge to educators: teaching machine learning to materials scientists. We share our recent experiences and lessons learnt from organizing educational sessions at the fall 2021 meeting of the Materials Research Society.

In parallel to advancements in technology and computational power, machine learning has catapulted to the forefront of scientific discovery, encompassing all forms of matter. Contemporary materials science is ripe with examples of advances powered by machine learning methods, from data mining the literature, to robot scientists, to rapid image processing that leverages massive databases. Given these examples, it is increasingly clear that familiarity with machine learning methods is becoming a mandatory tool in the belt of a modern materials researcher. Selecting the right tool for the right job, however, requires a nominal amount of knowledge and training. To give an analogy to tasks in our everyday life, one does not want to use a sledgehammer to crack a nut.

While formal curricula evolve to meet this need, many established researchers—from graduate students to professionals—are already in the field and in need of rapid training to give them the foundation needed to get started. Interestingly, we live in an age with ubiquitous online and freely available resources; however, not everyone thrives in such environments that require high self-motivation and an ability to debug the myriad small technical hurdles that might emerge. A major way to overcome this hurdle is to learn in a group with live or immediate instruction, including mini-courses and bootcamps. These are at their best when they get students to engage with the material and to try solving problems for themselves. Our overarching motivation was to tap into this potential for peer learning and learning-by-doing by combining a tutorial with a competition. While the "basics" can be conveyed in the tutorial segment, the competition aspect encourages more creative approaches and rewards application of the information gleaned from the tutorial. The students could immediately translate the skills they learned to their own specific research problem. Through this paired tutorial and competition, we could learn how to instruct the next generation of materials scientists while showing them how machine learning could directly impact their work.

To promote learning-by-doing with direct feedback, we organized a competition during the Tutorial Day at the fall meeting of the Materials Research Society in 2021 in Boston, MA (Figure 1). The Tutorial Day coincided with conference sessions in the area of applied machine learning and artificial intelligence in materials science held in the later part of the week. Many of the participants attended the conference sessions, reinforcing both what they learned as well as being exposed to cutting-edge data-driven materials research.

The competition itself consisted of two challenges focusing on applying machine learning techniques to (1) help understand patterns in data and (2) optimally guide experiment design. Our goal was to provide instruction, but we were also interested to see that the effective participant-produced solutions to these challenges taught us two important metalessons about how machine learning can interface with materials science.

Prior to submerging the students in code, we reasoned that it was important to provide an engaging overview of topics that we have found highly valuable to practical materials science. Thus, the beginning of the tutorial consisted of a series of interactive lectures, including an introduction to machine learning, an overview of deep learning, and finally an introduction to Gaussian processes and sequential Bayesian experimental design. These topics set up the main tasks that would form the basis of the later competition, namely regression and experimental selection. The hands-on tutorial provided the software building blocks, which attendees could assemble and modify into challenge solutions. Importantly, students were encouraged to work along with the instructors through a comprehensive online platform, thus gently getting the students over the often overlooked but critical barrier associated with

[1]Toyota Research Institute, Los Altos, CA 94022, USA

[2]Department of Mechanical Engineering, Boston University, Boston, MA 02215, USA

[3]Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

[4]Department of Materials Science & Engineering, University of Maryland, College Park, MD 20742, USA

*Correspondence: aaron.kusne@nist.gov

https://doi.org/10.1016/j.matt.2022.04.019

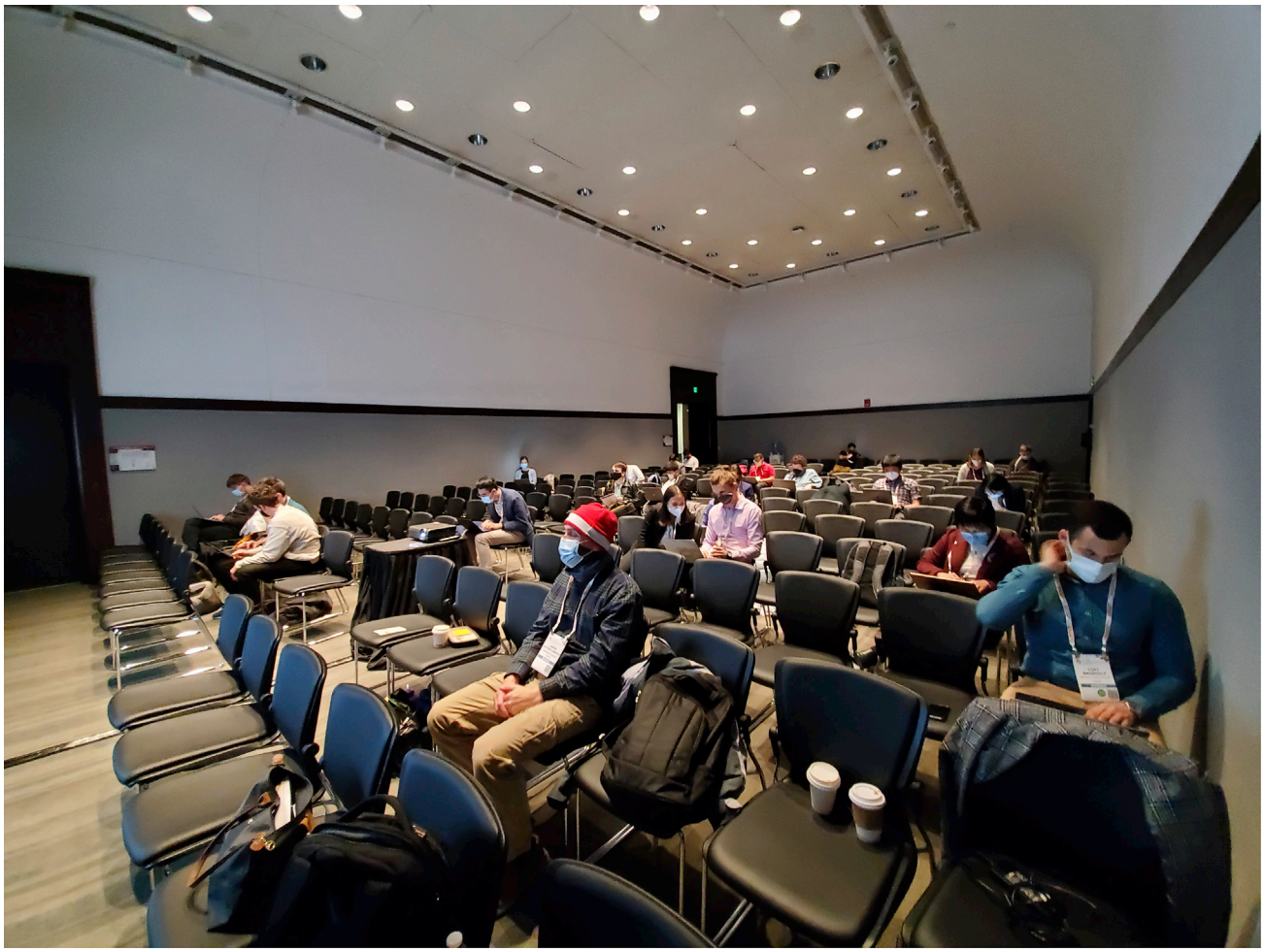

Figure 1. Photo taken during the data science competition

getting the software running. The comprehensive platform also allows instruction to scale to larger numbers of students without the issues of managing numerous computer environments.

When considering the topics and formats of the interactive challenges, we looked at how machine learning is commonly used in materials research. In the emerging field of materials informatics, one can apply statistical tools to extract correlations between materials variables and hence predict properties using machine learning. For Challenge I, participants were asked to develop a regression model to predict properties of new materials. Submissions were evaluated based on their root-mean-square error from the ground truth, and results were made available immediately through the Kaggle platform. Two of the top three submissions, both from graduate students majoring in materials with prior machine learning experience, made good use of the tutorial materials on Gaussian processes and succeeded in the competition through thoughtful selections and combinations of the kernels. Interestingly, the second-best submission, which we received from an industry researcher specialized in materials informatics, utilized a linear interpolation model, which one can consider to be less sophisticated than the methods we had been discussing. Using a less sophisticated model and obtaining high-performing results is not a bug, it is a feature! Indeed, this example reflects on the subtlety of model selection in materials science and highlights that more complicated machine learning models do not necessarily provide more accurate answers. This understanding led us to:

Metalesson 1: Complexity is not itself a virtue—models can be effective as long as they capture the complexity of the data.

The second challenge was also motivated by a common hurdle that must be overcome in materials research. Despite the increasing interest in machine-learning-guided materials discovery, a

key challenge in data-driven materials research is data scarcity. Experiments are slow and complicated in many fields, making a brute-force combinatorial screening expensive and impractical. Further, many parameter spaces in materials exhibit local extrema that can trap optimization, while trends leading to the global extrema may be highly localized. We therefore designed a small data problem with a complex parameter space as Challenge II of the competition. Within a design space of 1,000 simulated x-y data pairs, which we used to represent 1,000 hypothetical materials and their associated property, participants were invited to set up a sequential learning framework that is able to identify the optimal materials with minimal sampling. It was clear from the submissions that it didn't take participants long to realize that this was a stochastic process. The number of experiments needed to locate the optimal material depended on how close each new sampling was to the actual optimum point. Through this example, we created a scenario for participants to reflect on the advances and current limitations of applying machine learning to perform experimental design. In a winning submission from an experienced graduate student majoring in materials science, a Bayesian optimization algorithm was employed to identify the minimal samples needed in this challenge. To ensure the robustness of the model, the participant set the Bayesian optimizer to repeat itself twenty-five times, where an average of all output was taken as the final answer. Seeing the interplay between luck and robustness at play in these solutions led us to understand:

Metalesson 2: Luck is an unavoidable feature of data-driven experimental selection, and successful strategies must both leverage and protect against this.

In addition to the metalessons highlighted above, we highlight the following key lessons for future teaching in scientific machine learning:

1. Introductory sessions focusing on a programming language, e.g., Python, are essential for a successful class, as this allows every student to implement algorithms independently.
2. Preparing an intuitive coding framework beforehand and offering clear data preprocessing guidance will save significant time in class. We found that most questions we got during the competitions were on data uploading, saving, and formatting.
3. Emerging online platforms such as Collaboratory* and Jupyterlab make code sharing convenient and allow collaborative practices.

With the rapidly increasing demand to empower next-generation scientists with both materials and data science expertise, our curricula are also constantly evolving. We found formats that highlight learning-by-doing, such as small-scale tutorials and in-class competitions, are effective for materials scientists to gain practical experience dealing with challenges involving different levels of data complexity and can shine light on overarching but subtle metalessons. Simple frameworks discussed in class are also foundations for a bottom-up approach to build complex models in their own work. To date, there have been several online resources and workshops providing educational materials. We envision the application of data-driven tools in materials research will become more common in the coming years, which promotes closer collaborations in the multidisciplinary fields of materials informatics, scientific machine learning, and digital manufacturing. Our intent is to continue to iterate on these tutorials, offer competitions to provide hands-on experience, and learn from how the students teach their machines to learn. For example, data-science-related tutorials are currently being planned at the Materials Research Meeting in spring 2022, including one titled DS00 that is planned to be a spiritual successor to the tutorial discussed here. This one-day tutorial is a spinoff of the annual five-day Machine Learning for Materials Research bootcamp,[1] now on its seventh year. Information about future tutorials and competitions can be found on the Resource for Materials Informatics website.[2]

DECLARATION OF INTERESTS

The aforementioned Tutorial Day at the fall 2021 meeting of the Materials Research Society in Boston, MA was sponsored in part by *Matter*/Cell Press.

Certain commercial equipment, instruments, or materials are identified in this report in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

1. University of Maryland Nanocenter (2022). 7th annual machine learning for materials research bootcamp. https://www.nanocenter.umd.edu/events/mlmr/.

2. NIST (2022). REMI: REsource for materials informatics. https://pages.nist.gov/remi/.