Half-title page, prepared by publisher

Publishers' page

Full title page, prepared by publisher

Copyright page, prepared by publisher

Contents

Chapter 1

Artificial Intelligence for Materials

Debra J. Audus, Kamal Choudhary, Brian L. DeCost, A. Gilad Kusne, Francesca Tavazza, James A. Warren

Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899

The application of artificial intelligence (AI) methods to materials research and development (MR&D) is poised to radically reshape how materials are discovered, designed, and deployed into manufactured products. Materials underpin modern life, and advances in this space have the potential to markedly increase the quality of human life, address pressing environmental issues, and provide new, enabling, technologies that can help people realize their potential. This chapter delves into the many ways that AI is currently being applied to accelerate MR&D, the implications of this revolution, and the new frontiers that are now being opened for exploration.

1.1 Introduction

Modern materials research and development is the driving force behind new materials that transform our lives ranging from advances in battery technology to the nanoparticle technology behind messenger ribonucleic acid (mRNA) vaccines. It tackles grand challenges such as clean energy and a circular economy, as well as more mundane, but important advances such as stronger glass for cell phones and warmer, less bulky clothing. It includes brand new products never seen before, as well as incremental advances. However, it is also a slow process. For example, it took over half a century from when poly(lactic acid) was first discovered until it was regularly used in products such as compostable cups. To reduce the barriers, the Materials Genome Initiative (MGI) was established in 2011 in the United States where the goal is to bring materials to market twice as fast and at a fraction of the cost by making data accessible, leading a culture shift, training the next generation, and integrating experiment, computation, and theory [1]. Related efforts include the Findability, Accessibility, Interoperability, and Reusability (FAIR) Data Infrastructure (Europe, 2018), which encompasses the Novel Materials Discovery (NOMAD) data repository (Europe, 2014) and Material Research by Information Integration Initiative (Japan, launched 2015, ended 2020).

Since the initial launching of the MGI, artificial intelligence (AI), including machine learning (ML), has led to numerous, rapid advances in society including virtual assistants, image recognition, recommender systems, language translation, etc. It quickly became obvious that these concepts could also play a similarly transformative role in the materials domain. AI in material science starts to appear in comprehensive reports [2] a mere five years since the initial MGI strategy [1], which included the foundations such as accessible data but not AI explicitly.

To understand the power of AI in the materials domain, consider recent advances in the protein structure prediction problem. Specifically, this is an over 50 year old grand challenge to predict the structure of a protein based on its sequence of amino acids that has spurred an enormous amount of research and, notably, a competition called Critical Assessment of protein Structure Prediction (CASP) that occurs every two years since 1994 to benchmark global advances. In 2020, AlphaFold, [3] an AI entry from DeepMind, achieved an average error on the order of the size of a carbon atom and almost a factor of 3 less than the next closest entry. The reason why it is significant is not only that it made astonishing progress on a grand

challenge, but that it can often predict accurate structures with uncertainty in minutes as opposed to costly, time-consuming experiments that require specialized equipment and that both the model and a database of results was made available [4]. This subsequently enabled research in antibiotic resistance, investigations of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) biology, and enzymes to break down plastics. However, there is still more progress to be made. Experiments will still be necessary for outliers, the architecture of AlphaFold is complex, and both why it works and the science behind it is still unknown.

AI can also catalyze similar advances within the more traditional materials domain. Below we highlight the rapid advances, outstanding challenges and vision for the future in the field of materials AI; we also refer the reader to an excellent perspective [5]. In Section 1.2 we introduce the reader to AI for materials including the different types of tasks, models and approaches that are commonly used. Particular attention is paid to recent advances. One of the largest barriers for widespread adoption of AI in materials is often the need for large, curated datasets. Fortunately, AI can be used to tackle this problem. Section 1.3 discusses how AI can be used for generation of data with a focus on simulation data, as well as unstructured text. As discussed in Section 1.4, AI can, in principle, plan and execute scientific experiments and simulations completely independently or with humans in the loop. Yet, for AI in materials to reach its full potential, it is not only the data scarcity problem that needs to be solved. Other challenges include the need for benchmarks, uncertainty quantification, interpretability, the need for semantic search and further intertwining of experiment, simulation, theory and data in the spirit of the MGI as discussed in Section 1.5. As detailed in Section 1.6, it is an exciting time in the field of materials and AI—there are both lots of exciting advances and lots of interesting problems still to solve. For readers new to the field, we recommend REsource for Materials Informatics (REMI) [6] and learning resources therein, as well as a best practices guide [7].

1.2 AI allows you to make a material model

The transformative value of adding AI approaches to the collection of tools already available to material scientists results from ML's capacity to accelerate, optimize, correlate, and discover many of the linkages between how a material is processed, the internal structures that are formed, and the resultant properties. This capacity reflects ML's ability to produce material

models that, given specific inputs, quickly and reliably predict searched-for outputs.

The process of making ML models is a complex one, involving a series of steps, each one critical to the overall success. The process starts with data collection, often requiring a large, or relatively large, set of data. Luckily, in the last few years many material properties databases have been developed and continue to be developed, containing either computational or experimental data [8]. The next step is data curation, where any incorrect data are identified and eliminated, as well as areas of missing data are searched for and, if possible, corrected. The third step focuses on representing and preprocessing the data to facilitate machine learning use and maximize performance. This step may include data rescaling, normalization, binarization, and other data operations. A particular focus for supervised learning analysis is featurization, i.e., identifying which material properties need to be passed to the ML model as inputs. It's a key step, as the better the feature selection, the more effective the model. Often, domain knowledge is used in determining such features, but that is not the only approach. Other strategies include dimension reduction methods such as Principal Component Analysis (PCA), where high-dimensional data is projected into low-dimensional properties-relevant features, or describing the material using graph features, where atoms are nodes and bonds are edges, as done in graph neural network. The next critical choice, which is tied to the featurization method, is the choice of the ML algorithm. A vast gamut of possibilities are available, and are currently used, in material science. The choice depends on the type of application. Materials property prediction, in particular, has seen wide use of of ML, requires supervised learning, where the dataset has to contain both input and output data. Specifically, the prediction of continuous-value physical quantities like electronic energy gaps or stress distributions, is called regression, while establishing discrete-value labeled groups is performing classification, as, for instance, when determining if a material is metallic or not. Conventional algorithms (decision tress, kernel methods, etc.) as well as deep learning methods are used to these ends. An example of a different use of supervised algorithms is Symbolic Regression, where ML is used to determine the best mathematical formula to express correlations between inputs and outputs, instead of evaluating a physical property. Regression and classification often focus on learning a surrogate ML function to map input to output. The function is then used for predictions. Symbolic regression is used to find an interpretable model. A variation in the possible applications is the use of

supervised ML to determine, or optimize, not specific physical quantities, but parameters in functions, as it is the case when determining classical force fields [9]. Examples of all such application will be discussed later in this section. While supervised learning is the AI type most used in materials science, it's not the only one. Examples of unsupervised learning can be found, for instance, in automatic terminology extraction (natural language processing [10]) for materials science, aimed at identifying information on material properties, microstructures or process parameters currently disperse in unstructured sources. As the name says, unsupervised learning detects correlations/patterns between inputs with minimal external guidance and without being given any form of outputs. Phase mapping is an another example where unsupervised learning has been used. In addition to supervised and unsupervised learning, material science's applications are also making use of semi-supervised learning [11], transfer learning [12], and representation learning [13]. The last steps in the making of an ML model are the validation and uncertainty quantification. A certain amount of the input data is always set aside for validation purposes, i.e. never enters in the training. This data is then used to quantify the performance of the model though the evaluation of statistical quantities like the mean average error (MAE) or the root mean square error (RMSE). Quantifying uncertainty for predictions is a significant challenge on its own. While probabilistic models natively output uncertainty, others may require additional processes such as varying parameters (and hyperparameters) of the machine learning model. Additionally, uncertainty can be quantified over collections of predictions using statistics measures such as the number of true positives, true negatives, etc.. Uncertainty analysis is important for quantifying a model's performance, though it is commonly not included in published research.

Focusing on most current applications of supervised learning in material science, it is notable that one of its key strengths is its applicability to a variety of length scales. In the case of property prediction applications, for instance, ML has often been used at the atomistic level to substitute for high-fidelity, high-computational cost density functional theory (DFT) calculations. Specifically, many groups have taken advantages of the availability of vast and open sources DFT databases and modelled material properties like stability, exfoliability, elastic behavior, electronic energy gaps, and optical properties, for ideal crystalline [14], molecular, and polymeric materials [15]. At the continuum level, it has been extensively used to model fracture propagation and other failure mechanisms

[16]. In these applications, ML has taken the place of finite element (FEM) and Finite-Discrete Element (FDEM) modeling. Even relatively new fields, like additive manufacturing, have made extensive use of ML techniques, often applied to the investigation of defect-related material behavior. For instance, the effect of surface roughness or of defect location, size, and morphology on the fatigue life of a laser melted alloys, such as those found in additive manufacturing, have been probed using deep learning [16]. Multiscale investigations have been pursued as well, as in the case of Hsu et al. [17], where a machine-learning approach connecting molecular simulation into a physics-based data-driven multiscale model was utilized to predict fracture processes and toughness values. Within the polymers domain [18], a major challenge that continues to progress is representation of the polymer. However, with correct choices, exciting advances such as determining better materials for gas separation [19] are in the works.

Beyond property prediction, ML models have been extremely effective in model optimization and material discovery. One example of model optimization is the development of neural network-based classical force fields computations. Because of their numerically very accurate representation of high-dimensional potential-energy surfaces, they can provide energies and forces many orders of magnitude faster than ab-initio calculations, thus enabling molecular dynamics simulations of large systems. Transferability may, though, be a problem, as ML doesn't extrapolate well. Including information on the physical nature of the interatomic bonding into a neural network potential is one possible solution to such a shortcoming, as proposed by the developers of the Physically Informed Neural Networks (PINN) potential [20]. A different example of model optimization is the use of ML to develop exchange and correlation functionals for density functional theory (DFT) [21]. DFT is the ab-initio approach most frequently used when studying the electronic structure of materials at the atomic scale. While exact in principle, various approximations are required to perform DFT simulations in practice, among which the form of the exchange-correlation functional is, possibly, the most important, as it captures the electron–electron interaction. Traditionally built exchange-correlation functionals take a long time to be developed and, occasionally, have extremely complex analytical forms and are problematic in their accuracy and transferability for certain classes of materials. ML-built exchange-correlation functionals have the potentiality to improve over the traditional ones, while being effective for solids and molecules alike.

1.3 Application modalities

Well-curated datasets are necessary precursors to a supervised machine learning model. In materials informatics, data often includes atomic structures and chemical formulae, images, spectra, point values, and texts such as peer-reviewed articles as shown in Fig. 1.1. There have been numerous efforts for organizing the datasets. A complete list of such efforts can be found elsewhere [22–24]. In the following sections we provide a few well-known example applications of the above modalities. Atomic structure data

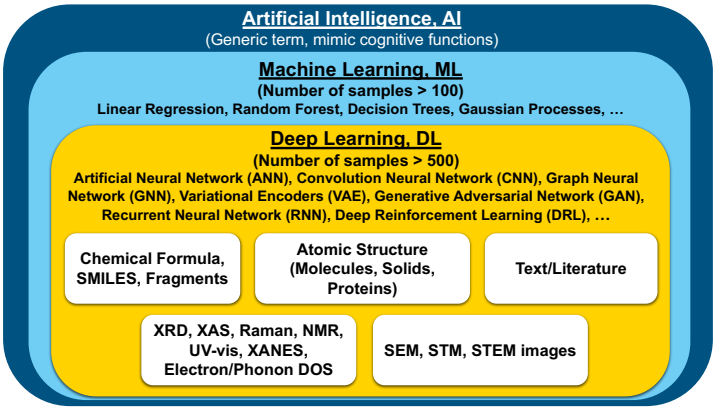

Fig. 1.1 Schematic showing an overview of Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) methods and its applications in materials science and engineering (Reprinted according to the terms of the CC-BY license [[23]]).

is usually obtained from quantum and classical mechanical methods such as density functional theory and molecular dynamics simulations. Recently, there has been a huge upsurge in specialty DFT based datasets due to growing computational power. The physio-chemical information in atomistic structure data have been used in machine learning tasks through the use of Euclidean hand-crafted descriptors and non-Euclidean graph neural network frameworks. An example of converting atomic structure into graph is shown in Fig. 1.2. Some of the common material property models generated using atomistic data are that for formation energy, total energy, bandgap, modulus of elasticity and so on. Such developed models have been used for materials screening purposes such as for alloy, energy and quantum materials design. More details about such applications can be

found elsewhere [23, 25].

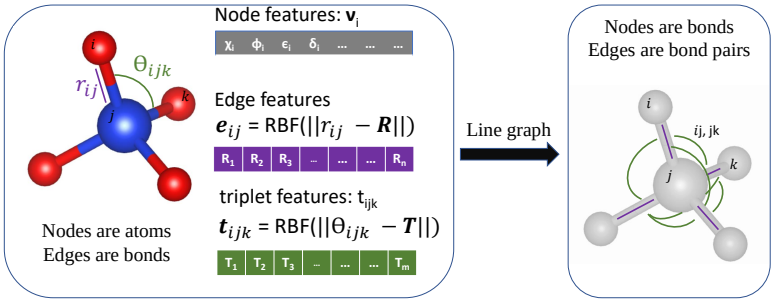

Fig. 1.2 Schematic representations of an atomic structure as a graph using Atomistic Line Graph Neural Network (ALIGNN) (Reprinted according to the terms of the CC-BY license [[26]]).

Some of the numerous, common imaging techniques (microscopy) for materials science include optical microscopy (OM), scanning electron microscopy (SEM), scanning probe microscopy (SPM), which includes scanning tunneling microscopy (STM) or atomic force microscopy (AFM), and transmission electron microscopy (TEM) and its variants, such as scanning transmission electron microscopy (STEM). As applications of imaging techniques have become widespread, there is often a large amount of generated data which can be difficult to analyze manually. ML techniques are useful for automating and aiding manual work in tasks such as image classification of lattices, detecting defects, resolution enhancement, microstructure learning, image reconstruction and making physics based models using image data. A detailed review of ML for materials image data can be found elsewhere [23, 27]. To give a few example works, in ref. [28] the authors developed datasets, and machine models for analyzing defects in graphene and FeTe systems. In ref. [29], the authors developed semantic segmentation models for steel that can be used for accurately detecting damages. In ref. [30] a computational STM image dataset was developed along with ML models to classify 2D materials lattice with up to 90 % accuracy. In Ref. [31], the authors used a convolutional neural net (CNN) that was pretrained on non-scientific images to perform microstructure reconstruction.

Spectral data such as X-ray and neutron diffraction (XRD, ND), X-ray absorption near edge structure (XANES), electron energy loss (EELS), infrared and Raman (IR/Raman), ultraviolet and visible (UV-VIS) provide great insight into structural, composition and dynamic behavior of materials. XRD is one of the widely used method for developing large datasets for materials and training machine learning models. A detailed overview on such a topic can be found in refs. such as [23, 32]. To give a few example applications, Park et al. [33] used CNN for predicting structural properties of materials with up to 95 % accuracy. Dong et al. introduced parameter quantification network (PQ-Net) to extract physio-chemical information from XRD data. Timoshenko et al. [34] used a neural network for XANES to predict coordination number of metal clusters. Liu et al [35] used SVM and CNN methods to predict Raman spectra using a Raman spectra database. Fung et al. [36] used CNNs to predicting electronic density of states obtained from DFT.

The text data for materials is obtained from academic journal articles, preprint repositories such as arXiv, and crowd-sourced resources like Wikipedia etc. Compared to other applications, datasets for NLP are often scarce due to copyright restrictions and paywalls. Nevertheless, there have been substantial research in this field due to rapid developments in the field of sequence modeling. Such techniques are being used for automatically extracting relevant material property data from text to guide theoretical and experimental work such as material synthesis, magnetic phase transition temperature calibration, designing batteries and solar-cells etc. [37–39].

1.4 Autonomous science

Advances in technology often require the discovery and development of novel advanced materials. As a result, the search for advanced materials continually expands to materials of greater and greater complexity in both number of chemical components and processing steps. However, with each new element or processing step, the number of potential materials to investigate grows exponentially (e.g., if 10 experiments are used to investigate the impact of one variable, 10^N experiments will be required to investigate the impact of N variables). Manual investigations rapidly become infeasible. High-throughput techniques [40] provide a short-term solution, making it possible to synthesize thousands of unique materials in a few hours and characterize these materials in rapid succession. Machine learning can then be used to reduce data analysis from weeks and months to seconds. Nev-

ertheless, high-throughput methods do not provide the exponential scaling needed.

The last few years have seen the rise of autonomous materials research systems [41] – where machine learning controls experiment design, execution (both physical and *in silico*), and analysis in a closed loop. These systems use active learning – the machine learning field focused on optimal experiment design, to ensure that each subsequent experiment maximizes knowledge gained toward user defined goals. Active learning allows users to balance two high-level goals: exploration and exploitation. Exploration seeks to gain global knowledge of the target material system, while exploitation uses past knowledge to hone-in on possible optimal materials. Active learning on its own has been used to build research recommendation engines to advise scientists in the lab and *in silico* [42]. Autonomous systems combine the automated experiment and simulation tools of high-throughput science with active learning, guiding each study in the lab or *in silico* through the vast materials search space, accelerating the discovery and study of advanced materials.

The advent of autonomous systems does not mean that prior physical knowledge is thrown out. In fact, autonomous system performance improves by incorporating prior physical knowledge found in laws, heuristics, and databases. Scientific machine learning [2] is the AI subfield that focuses on incorporating prior physical knowledge into AI and ML. The goal is to design models and learning algorithms so that their inductive bias reflects physical principles. This promotes physically realistic predictions and provides opportunities for physically meaningful model interpretations. For instance, the Gibbs phase rule, along with more general rules such as a material can only be composed of a positive amount of constituent phases, and materials of similar synthesis are more likely to have similar properties, have been used to improve phase map data analysis and prediction [43].

Autonomous measurement systems take the first step toward autonomous materials exploration and discovery. These systems guide each subsequent measurement experiment to obtain maximal information for the target sample. Parameters that are optimized include where on the sample to measure next, the temperature of the material, and measurement specific parameters such as beam wavelength. For example, autonomous X-ray measurements were used to guide beam location to accelerate nano-particle density determination [44] and autonomous X-ray measurements were used to select subsequent sample compositions in the search for novel materials, resulting in a 10x acceleration toward the discovery of the new, best-in-class

phase change memory material [45].

Success in autonomous chemistry preceded materials science, with systems that optimize molecular mixtures for applications such as chemical reaction optimization [46]. Accordingly, the majority of autonomous materials systems focus on 'wet' synthesis such as optimizing quantum dots for emission energy [47]. More recently, closed-loop 'dry' materials synthesis and characterization was exemplified for thin-film resistance optimization [48].

While active learning can be used to guide a researcher in which material to investigate next across a large design space, often the lab systems are limited in the number of variables that can be automated and machine learning controlled. For instance sputtering systems are limited in the number of components that can be deposited at a time and control over measurement beam path may be locked out for the users. Advances in autonomous physical science will require innovations in physical systems that will allow for greater autonomous control of experiments in the lab.

Additionally, machine learning for autonomous systems could also gain great advantage from the large amount of prior knowledge of expert users. This requires the development of human-in-the-loop systems where the autonomous system is able to interact with and learn from expert users. This is also an ongoing area of research.

1.5 Challenges

At the time of writing, a robust discourse spanning the materials and chemistry fields has developed around constructing and deploying creative applications of ML methods to enable more efficient and scalable research and development. The value proposition and validity of ML-enabled scientific methodologies have been demonstrated many times over, and the education and training ecosystem in this subfield is rapidly maturing. Advancement of transformational ML applications in science now depends on addressing specific challenges in the scalability and generalization ability of ML systems, as well as their ease of integration into higher-level scientific workflows.

Availability of sufficiently large datasets for training ML systems is widely acknowledged as a challenge in the physical sciences, particularly where sample fabrication and characterization are time and cost intensive. Three general strategies are currently being pursued: investing in massive datasets though high throughput experimentation and computation,

deploying active learning and autonomous systems for targeted collection of high-value data, and improving the sample complexity of ML systems through algorithmic improvements. The preceding sections of this chapter outline current progress along each of these axes. Continued innovation critically depends on the fields' ability to quantify, communicate, and mitigate dataset and annotation bias.

Progress in ML is historically enabled by clearly defined tasks, with robust performance measures that guide innovation. Failing to account for dataset bias can lead to benchmarks that reward memorization [49], making clear assessments of methodological innovation difficult. At present, many of the large materials datasets used in applied ML research have been repurposed from materials discovery efforts not directly motivated by ML applications. The resulting forms of dataset bias are acknowledged in the materials literature, but further detailed quantification and research into mitigation strategies will enable the community to develop more productive ML systems. This bias assessment covers not only the data themselves, but extends to the framing of the ML task in the context of ultimate scientific goals. For example DFT formation energy is a popular atomistic prediction target used throughout the literature to benchmark model quality, but thermodynamic stability is a much more relevant (and challenging) target for materials discovery and design applications [50]. This kind of investment into clearly defined ML tasks and well-designed benchmark datasets will help the materials community engage more effectively around concrete problems with the broader AI community. The Open Catalyst Challenge [51] provides an excellent example of this kind of clear problem framing and effective coordination of the general ML community.

High-quality and well-characterized datasets provide a foundation for the field to develop model introspection and interrogation tools that build and preserve confidence for ML-based decision making in the sciences. These are of critical importance because scientific progress and discovery is often fundamentally at odds with a central assumption relied on by most ML systems: that data for which predictions are made is identically distributed to the training data. Machine learning systems, especially the overparameterized deep learning models popular in current research, can perform notoriously poorly when extrapolating with high confidence. The importance of principled uncertainty quantification for machine learning predictions is increasingly being emphasized in materials research, especially as more researchers adopt active learning methodologies that rely on well-calibrated measures of model confidence. More research is needed

to decompose predictive uncertainty across multiple model and training components, and to be able to propagate uncertainties of the underlying training data through the modeling pipeline as well.

Similarly, general-purpose ML model introspection and interpretability tools [52] are helping materials researchers to identify and explain trends in ML model behavior. Materials-specific adaptation of these methods, such as an additive modeling strategy to decompose the contribution of individual atoms in a crystal to a target property [53], can provide new ways of studying materials phenomena, and for understanding how ML models form predictions. Progress in this direction will depend on greater adoption of standard model introspection tools in the materials community, and an ongoing discourse of developing creative methods for critiquing ML models and predictions. These tools will be most effective in combination with ML systems designed from the ground up to incorporate and reflect physical principles, so that ML models and their predictions can directly be interpreted in terms of physical insight. [2]

Finally, there are important areas in which methodological innovation will substantially boost scientific progress. The subfield of scientific machine learning is rapidly developing creative modeling approaches that tailor the inductive biases of ML systems to reflect physical principles. For example, graph neural network architectures reflect well the local structure and permutation invariance in atomistic systems, and neural differential equations are enabling data-driven discovery of mechanistic models in mechanical and pharmacokinetic systems. There is a rich diversity of theoretical modeling approaches in materials science, and developing generalizable approaches for incorporating them into ML systems is an important and challenging task.

With respect to automated materials discovery and design, deep generative models have shown encouraging progress for molecular [54] and crystalline [55] systems. Important research directions include developing methods that go beyond structural prototype search, methods for using generative models for reaction engineering and synthesis planning, and developing a framework for connecting generative models to higher-level materials structures and models.

Curation of structured knowledge and, crucially, metadata will enable novel forms of ML application in materials science, such as semantic search, which incorporates specific domain knowledge into information retrieval. Automated literature mining to extract and link information to construct knowledge graphs is an active topic of research in the ML community. Spe-

cific materials science emphasis is needed to build on early work in chemical named entity recognition, information extraction from common graphical representations of measured materials properties, and extraction of material properties and synthesis protocols [56]. These tools and the resulting scientific knowledge graph and related infrastructure will create opportunities for integrating structured concepts into information retrieval, predictive modeling tools, and ML model interrogation and interpretability tools.

A final challenge area is seamless integration of ML into larger materials design systems. This covers both multiple interacting ML-based systems as well as conventional computational materials science infrastructure. The long-standing multiscale modeling challenge provides many opportunities, for example end-to-end learning in neural networks that parameterize physical simulations or perform coarse-graining to connect classical modeling components. Finally, rigorous incorporation of scientific domain knowledge may require going beyond machine learning to identify opportunities for applications of broader artificial intelligence methodology, such as search-based techniques and symbolic reasoning [57].

1.6 Conclusions and outlook

In this chapter, we have outlined the many ways that AI has the power to catalyze transformational advances in materials R&D. We have detailed the myriad ways that AI is, and will continue to, impact materials research, including in the predictions of materials properties, the influence of defects and dopants, and how materials fail. AI techniques can markedly accelerate the hunt for new materials with desirable properties, allowing for inverse design techniques that can zero in on the optimal processing techniques needed to obtain the desired materials performance.

Additionally, AI approaches can accelerate the characterization of materials systems, using the powerful image analysis/processing techniques that drove many of the notable breakthroughs in AI. Since materials can be characterized not just by images of their internal structure, but also by patterns resulting from techniques such as X-ray and electron scattering, image processing techniques are especially potent.

It was noted that with the maturation of data gathering and improved data infrastructures, the raw materials for AI will be in increasingly greater and greater supply. This positive development is compounded by the advent of AI-driven, autonomous laboratory setups, where closed loop experimental designs will enable the exploration of far wider processing and

compositional parameter variations than we previously practical, creating a virtuous feedback loop that could yield truly paradigm shifting levels of data and, ultimately, knowledge, about the existing and sought for materials.

We also spent some time addressing the challenges that must be overcome to realize the full potential of AI approaches. The required data (and metadata) infrastructures are still not there, and those that exist don't interoperate especially well. AI approaches have definite risks, and benchmarking efforts are needed to provide trust in these methods, and, where possible, uncertainty bounds provided. It is also desirable that AI methods incorporate the best available physical knowledge and insights, ensuring the algorithms respect known science and underlying symmetries.

We began our discussion by considering the success of the AlphaFold effort, which linked the fine structure of proteins (the amino acid sequence) to ultimate conformation. This type of linkage is not analogous to materials R&D, it *is* materials R&D, albeit for a very specific research question. Perhaps one of the great stories of the success of AlphaFold was the CASP competition, which drove the protein folding research community forward for 30 years, long before AI techniques became commonplace in either biomedical or materials research. There is an enormous opportunity for the materials R&D community to create new "grand challenge" problems that are of the same magnitude as the protein folding problem, and could result in materials that address many of the pressing concerns facing humanity today, including energy efficiency, climate change mitigation through carbon capture materials and cleaner energy production, enhanced recyclability and reduction in plastics pollution, and new biocompatible medical devices and implants. AI is poised to make these revolutions in the human condition achievable with a speed and expense that was, until very recently, unimaginable.

Bibliography

[1] J. P. Holdren *et al.*, "Materials genome initiative strategic plan," *Washington DC: Office of Science and Technology Policy*, vol. 6, 2014.

[2] N. Baker, F. Alexander, T. Bremer, A. Hagberg, Y. Kevrekidis, H. Najm, M. Parashar, A. Patra, J. Sethian, S. Wild *et al.*, "Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence," USDOE Office of Science (SC), Washington, DC (United States), Tech. Rep., 2019.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[4] D. Hassabis, "Putting the power of alphafold into the world's hands," July 2021, [Online; posted 22-July-2021]. [Online]. Available: https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worlds-hands

[5] D. M. Dimiduk, E. A. Holm, and S. R. Niezgoda, "Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering," *Integrating Materials and Manufacturing Innovation*, vol. 7, no. 3, pp. 157–172, 2018.

[6] 2022, date accessed: 22 March 2022. [Online]. Available: https://pages.nist.gov/remi/

[7] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, and T. D. Sparks, "Machine learning for materials scientists: An introductory guide toward best practices," *Chemistry of Materials*, vol. 32, no. 12, pp. 4954–4965, 2020.

[8] Z. Song, X. Chen, F. Meng, G. Cheng, C. Wang, Z. Sun, and W.-J. Yin, "Machine learning in materials design: Algorithm and application," *Chinese Physics B*, 2020.

[9] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood *et al.*, "Performance and cost assessment of machine learning interatomic potentials," *The Journal of Physical Chemistry A*, vol. 124, no. 4, pp. 731–745, 2020.

[10] F. Olsson, "A literature survey of active machine learning in the context of

natural language processing," SICS, Tech. Rep. 2009:06, 2009.
[11] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, Nov. 2019.
[12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
[13] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
[14] K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. H. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone *et al.*, "The joint automated repository for various integrated simulations (jarvis) for data-driven materials design," *npj Computational Materials*, vol. 6, no. 1, pp. 1–13, 2020.
[15] H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty *et al.*, "Machine-learning predictions of polymer properties with polymer genome," *Journal of Applied Physics*, vol. 128, no. 17, p. 171104, 2020.
[16] H. Bao, S. Wu, Z. Wu, G. Kang, X. Peng, and P. J. Withers, "A machine-learning fatigue life prediction approach of additively manufactured metals," *Engineering Fracture Mechanics*, vol. 242, p. 107508, 2021.
[17] Y.-C. Hsu, C.-H. Yu, and M. J. Buehler, "Using deep learning to predict fracture patterns in crystalline solids," *Matter*, vol. 3, no. 1, pp. 197–211, 2020.
[18] L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, "Polymer informatics: Current status and critical next steps," *Materials Science and Engineering: R: Reports*, vol. 144, p. 100595, 2021.
[19] J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, "Designing exceptional gas-separation polymer membranes using machine learning," *Science advances*, vol. 6, no. 20, p. eaaz4301, 2020.
[20] G. P. P. Pun, R. Batra, R. Ramprasad, and Y. Mishin, "Physically informed artificial neural networks for atomistic modeling of materials," *Nature Communications*, vol. 10, no. 1, May 2019.
[21] S. Dick and M. Fernandez-Serra, "Machine learning accurate exchange and correlation functionals of the electronic density," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.
[22] C. Draxl and M. Scheffler, "Nomad: The fair concept for big data-driven materials science," *Mrs Bulletin*, vol. 43, no. 9, pp. 676–682, 2018.
[23] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. WooPark, A. Choudhary, A. Agrawal, S. J. Billinge *et al.*, "Recent advances and applications of deep learning methods in materials science," *arXiv preprint arXiv:2110.14820*, 2021.
[24] C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, Á. Fekete, A. Gopakumar, S. Gražulis, A. Merkys *et al.*, "Optimade: an api for exchanging materials data," *arXiv preprint*

arXiv:2103.02068, 2021.

[25] R. K. Vasudevan, K. Choudhary, A. Mehta, R. Smith, G. Kusne, F. Tavazza, L. Vlcek, M. Ziatdinov, S. V. Kalinin, and J. Hattrick-Simpers, "Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics," *MRS communications*, vol. 9, no. 3, pp. 821–838, 2019.

[26] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," *npj Computational Materials*, vol. 7, no. 1, pp. 1–8, 2021.

[27] M. Ge, F. Su, Z. Zhao, and D. Su, "Deep learning analysis on microscopic imaging in materials science," *Materials Today Nano*, vol. 11, p. 100087, 2020.

[28] S. Somnath, C. R. Smith, N. Laanait, R. K. Vasudevan, and S. Jesse, "Usid and pycroscopy–open source frameworks for storing and analyzing imaging and spectroscopy data," *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 220–221, 2019.

[29] B. L. DeCost, B. Lei, T. Francis, and E. A. Holm, "High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel," *Microscopy and Microanalysis*, vol. 25, no. 1, pp. 21–29, 2019.

[30] K. Choudhary, K. F. Garrity, C. Camp, S. V. Kalinin, R. Vasudevan, M. Ziatdinov, and F. Tavazza, "Computational scanning tunneling microscope image database," *Scientific data*, vol. 8, no. 1, pp. 1–9, 2021.

[31] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, "A transfer learning approach for microstructure reconstruction and structure-property predictions," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[32] Z. Chen, N. Andrejevic, N. C. Drucker, T. Nguyen, R. P. Xian, T. Smidt, Y. Wang, R. Ernstorfer, D. A. Tennant, M. Chan *et al.*, "Machine learning on neutron and x-ray scattering and spectroscopies," *Chemical Physics Reviews*, vol. 2, no. 3, p. 031301, 2021.

[33] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K.-S. Sohn, "Classification of crystal structure using a convolutional neural network," *IUCrJ*, vol. 4, no. 4, pp. 486–494, 2017.

[34] J. Timoshenko, D. Lu, Y. Lin, and A. I. Frenkel, "Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles," *The journal of physical chemistry letters*, vol. 8, no. 20, pp. 5091–5098, 2017.

[35] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, "Deep convolutional neural networks for raman spectrum recognition: a unified solution," *Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017.

[36] V. Fung, G. Hu, P. Ganesh, and B. G. Sumpter, "Machine learned features from density of states for accurate adsorption energy prediction," *Nature Communications*, vol. 12, no. 1, pp. 1–11, 2021.

[37] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," *Chemistry of Materials*, vol. 29, no. 21, pp. 9436–9444,

Oct. 2017.

[38] C. J. Court and J. M. Cole, "Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[39] S. Huang and J. M. Cole, "A database of battery materials auto-generated using chemdataextractor," *Scientific Data*, vol. 7, no. 1, pp. 1–13, 2020.

[40] M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, E. Campo, T. Chiang, S. Empedocles, J. M. Gregoire, A. G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J. V. Duren, and A. Zakutayev, "Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies," *Applied Physics Reviews*, vol. 4, no. 1, p. 011105, Mar. 2017.

[41] E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev, and B. Maruyama, "Autonomous experimentation systems for materials development: A community perspective," *Matter*, vol. 4, no. 9, pp. 2702–2726, Sep. 2021.

[42] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," *Nature Communications*, vol. 7, no. 1, Apr. 2016.

[43] S. Ament, M. Amsler, D. R. Sutherland, M.-C. Chang, D. Guevarra, A. B. Connolly, J. M. Gregoire, M. O. Thompson, C. P. Gomes, and R. B. van Dover, "Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams," *Science Advances*, vol. 7, no. 51, Dec. 2021.

[44] M. M. Noack, K. G. Yager, M. Fukuto, G. S. Doerk, R. Li, and J. A. Sethian, "A kriging-based approach to autonomous experimentation with applications to x-ray scattering," *Scientific Reports*, vol. 9, no. 1, Aug. 2019.

[45] A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, and I. Takeuchi, "On-the-fly closed-loop materials discovery via bayesian active learning," *Nature Communications*, vol. 11, no. 1, Nov. 2020.

[46] A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, and T. F. Jamison, "Reconfigurable system for automated optimization of diverse chemical reactions," *Science*, vol. 361, no. 6408, pp. 1220–1225, Sep. 2018.

[47] R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Han, K. G. Reyes, A. Amassian, and M. Abolhasani, "Artificial chemist: An autonomous quantum dot synthesis bot," *Advanced Materials*, vol. 32, no. 30, p. 2001626, Jun. 2020.

[48] R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando, and T. Hitosugi, "Autonomous materials synthesis by machine learning and robotics," *APL Materials*, vol. 8, no. 11, p. 111110, Nov. 2020.

[49] I. Wallach and A. Heifets, "Most ligand-based classification benchmarks reward memorization rather than generalization," *Journal of Chemical Infor-*

mation and Modeling, vol. 58, no. 5, pp. 916–932, Apr. 2018.

[50] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, "A critical examination of compound stability predictions from machine-learned formation energies," *npj Computational Materials*, vol. 6, no. 1, Jul. 2020.

[51] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, and Z. Ulissi, "Open catalyst 2020 (OC20) dataset and community challenges," *ACS Catalysis*, vol. 11, no. 10, pp. 6059–6072, May 2021.

[52] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 417–431.

[53] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Physical review letters*, vol. 120, no. 14, p. 145301, 2018.

[54] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, Jul. 2018.

[55] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, "Crystal diffusion variational autoencoder for periodic material generation," 2021.

[56] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, "Opportunities and challenges of text mining in materials research," *Iscience*, vol. 24, no. 3, 2021.

[57] D. Chen, Y. Bai, S. Ament, W. Zhao, D. Guevarra, L. Zhou, B. Selman, R. B. van Dover, J. M. Gregoire, and C. P. Gomes, "Automating crystal-structure phase mapping by combining deep learning with constraint reasoning," *Nature Machine Intelligence*, vol. 3, no. 9, pp. 812–822, 2021.