Machine learning in materials science: From explainable predictions to autonomous design

Pilania, Ghanshyam

To be published in: Computational Materials Science

DOI to publisher's version: 10.1016/j.commatsci.2021.110360

Machine Learning in Materials Science: From Explainable Predictions to Autonomous Design

Ghanshyam Pilania[a]

[a]*Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA*

Abstract

The advent of big data and algorithmic developments in the field of machine learning (and artificial intelligence, in general) have greatly impacted the entire spectrum of physical sciences, including materials science. Materials data, measured or computed, combined with various techniques of machine learning have been employed to address a myriad of challenging problems, such as, development of efficient and predictive surrogate models for a range of materials properties, screening and down-selection of novel candidate materials for targeted applications, new methodologies to improve and further expedite molecular and atomistic simulations, with likely many more important developments to come in the foreseeable future. While the applications thus far have provided a glimpse of the true potential data-enabled routes have to offer, it has also become clear that further progress in this direction hinges on our ability to understand, explain and rationalize findings of a machine learning model in light of the domain-knowledge. This focused review provides an overview of the main areas where machine learning has been widely and successfully used in materials science. Subsequently, a brief discussion of several techniques that have been helpful in extracting physically-meaningful insights, causal relationships and design-centric knowledge from materials data is provided. Finally, we identify some of the imminent opportunities and challenges that materials community faces in this exciting and rapidly growing field.

Keywords: Materials informatics, statistical learning, physics-informed learning, domain-specific learning, materials discovery

1. Introduction

We are living in the age of "big data" and information. The amount of data generated, shared, processed and stored around the planet on a daily basis is unprecedented. The summation of all this data is collectively called the global datasphere. Based on the latest estimates, it is predicted that that the global datasphere will grow exponentially to 175 zettabytes by 2025 (zettabyte is 1021 bytes) [1]. The sheer volume, production speed and heterogeneous nature of this data naturally demands for new and efficient methods of analysis to unearth hidden patterns, trends and insights in this vast sea of information. This ever growing need to analyze and make sense out of the big data has been a primary driving force behind the state-of-the-art machine learning (ML) methods and algorithms.

ML broadly refers to the use of algorithms and computer systems that can learn to perform a task given just the relevant data, do not require any explicit programming specific to the task, and get better with experience (i.e., the available past data). ML models can be supervised, semi-supervised or unsupervised, depending on the type of available training data. In supervised learning, the training data consist of sets of input and associated output values. In other words, labelled training samples are required. On the other hand, if the training dataset contains unlabeled samples, unsupervised learning can be used in ordered to identify trends and patterns in the data. Semisupervised learning can be used for large datasets with partially missing labels. In the past decade, tools and applications built on ML have found a widespread use in applications as diverse as, transportation, communications, healthcare, business intelligence and strategy, social networking, and industrial research [2]. This paradigm shift is brought about by a confluence of sustainable growth in computing power, the aforementioned data revolution, multiple algorithmic breakthroughs and design and deployment of hardware customized to boost the performance of ML algorithms. Moreover, a self-reinforcing and synergistic growth of computing, data, algorithms and co-designed software and hardware–the components forming the ML ecosystem–has further helped in expediting the pace of development in each individual area, while benefitting from and driven by the progress made in the other fields of the ecosystem.

The resounding success of big data and ML methods in tasks, such as, image and speech recognition [3, 4], language translation [5, 6] as well as the superhuman performance achieved by artificial intelligence (AI) based algorithms in games of chess [7], Go [8, 9], poker [10, 11] and Jeopardy [12] has also been reflected in their wide spread adoption in physical sciences. More specifically in materials science and related fields use of ML-based methods has led to several developments pertaining to the design and development of new materials and a better understanding of the existing ones. Novel ML-based routes for mapping potential energy surfaces and forcefields have allowed for atomistic simulations of molecules and solids reaching beyond the tradeoffs of accuracy, speed, time- and length scales possible within the traditional molecular dynamics simulations. More recently, the focus has been on using active learning (or adaptive design) to enable autonomous robot-assisted

development of functional materials with prespecified properties. Given a target chemical space and constraints related to available resources, development time and a property wish list, these efforts have focused on harnessing the power of optimal learning concepts within the context of efficient experimental design. Finally, going beyond supervised learning, use of natural language processing techniques to automatically extract and synthesize materials science knowledge present in the published literature in form of information-dense word embeddings to capture complex materials science concepts remains a very exciting and potentially transformative new area of research.

A wide range of ML algorithms, vastly varying in terms of complexity and transparency, have been employed for data-enabled materials design. On one side of the spectrum lie, for instance, tree-based classification and regression methods that are completely transparent when it comes to explaining the model predictions. The other extreme is occupied by deep neural networks and ensemble-based methods, which allow for little insights and reasoning into the inner workings of the models leading to the final results. Since a vast majority of studies in the field of materials informatics have focused on developing ML-based surrogate models of structure-property-processing relationships, the primary emphasis has always been on achieving a predictive accuracy as high as possible. This quest for improved performance naturally creates a bias for employing more complex, and therefore, less transparent and poorly-explainable models. However, eventually an integration of the knowledge mined from and assimilation of the discoveries made with statistical pattern recognition techniques into materials science demands for a deeper analysis and a better understanding of the findings in light of the domain-knowledge. To expedite the pace of progress and potential impact ML methods can bear on materials development, AI algorithms must be tasked with generation of understanding that explains the obtained results, as we now uniquely task human intuition. The need for explainable models in hard sciences, including materials science, has recently led to a surge of activity in the field of explainable AI (frequently referred to as XAI) [13, 14].

In this contribution, after reviewing various recent applications of ML in materials science and related fields, we focus on a selected set of ML tools and techniques that have been employed in the past both to automatically extract physically meaningful knowledge from the data and to better rationalize existence of causal relationships in the identified patterns. Using selected examples from recent studies, we emphasize that integration of relevant domain knowledge is a crucial step in devising an ML strategy and that this becomes even more critical when dealing with small training datasets. Finally, we identify and discuss key challenges and opportunities faced by the potent and quickly growing field of ML-enabled materials design. Throughout the review, it is assumed that the reader is familiar with the basic nomenclature and standard methods of ML applied to the field of materials informatics. A familiarity with the best practices of statistical learning is also assumed and, therefore, these topics are not covered here but could be found elsewhere [15, 16, 17].

2. Use of Machine Learning in Materials Science

Use of ML to address design and development challenges in materials science and other related fields is an actively growing area, which has seen a rapid growth in the last decade. The amount of scientific research published in this field has exhibited a sustained exponential growth since 2014, with the number of contributions approximating doubling in every one and half years [15]. Therefore, an exhaustive survey of the entire spectrum of this research is beyond the scope of this review, however, we refer interested readers to a number of excellent reviews where a significant portion of these recent developments and applications have been covered and discussed [15, 22, 16, 23, 24, 25, 26]. Below we present a selected set of key areas where informatics-based methods have proved particularly promising and widely applicable. These are also depicted graphically in Fig. 1. By choosing specific examples in each of these areas, we highlight how ML is helping to progress the field by reducing barriers in materials design via addressing challenges related to materials modeling, synthesis and characterization.

2.1. Efficient and predictive surrogate models

A vast majority of recent work in the field falls within this category where ML-based surrogate models provide an alternative data-enabled route to establish desired processing-structure-property-performance linkages within the target chemical space. Relying on easily accessible and carefully devised numerical representations, frequently referred to as features or descriptors, ML algorithms are used to develop validated mappings that connect problem-relevant aspects of materials' composition, structure, morphology, processing etc. to the target property or performance criteria, while largely bypassing traditional time and resource-intensive experimental and computational routes (see Fig. 1a). The selection of an appropriate descriptor is one of the most crucial aspect of the entire surrogate model building exercise, which often relies heavily on domain-specific expertise. Further, best practices of statistical learning, such as, appropriate and unbiased selection of the training data representative of the underlying true data distribution, use of cross-validation for the model hyper parameter selection, testing on unseen data are required to ensure a truly predictive optimal learning model. Once developed, validated and rigorously tested to be predictive within a given domain of applicability, the true value of such models lies in their remarkable speed compared to the traditional property prediction or measurement routes. As a result, ML-based surrogate models are particularly well suited for high throughput screening efforts where one targets to identify molecules or compounds with a one or more properties in a pre-specified range. If a subset of properties exhibits conflicting trends or inverse relationships, looking for "optimal" compounds corresponds to finding chemistries falling on or near the underlying Pareto front—providing the best achievable tradeoffs among the conflicting responses.

ML algorithms have been employed to identify potential nonlinear multivariate relationships for a wide variety of materials

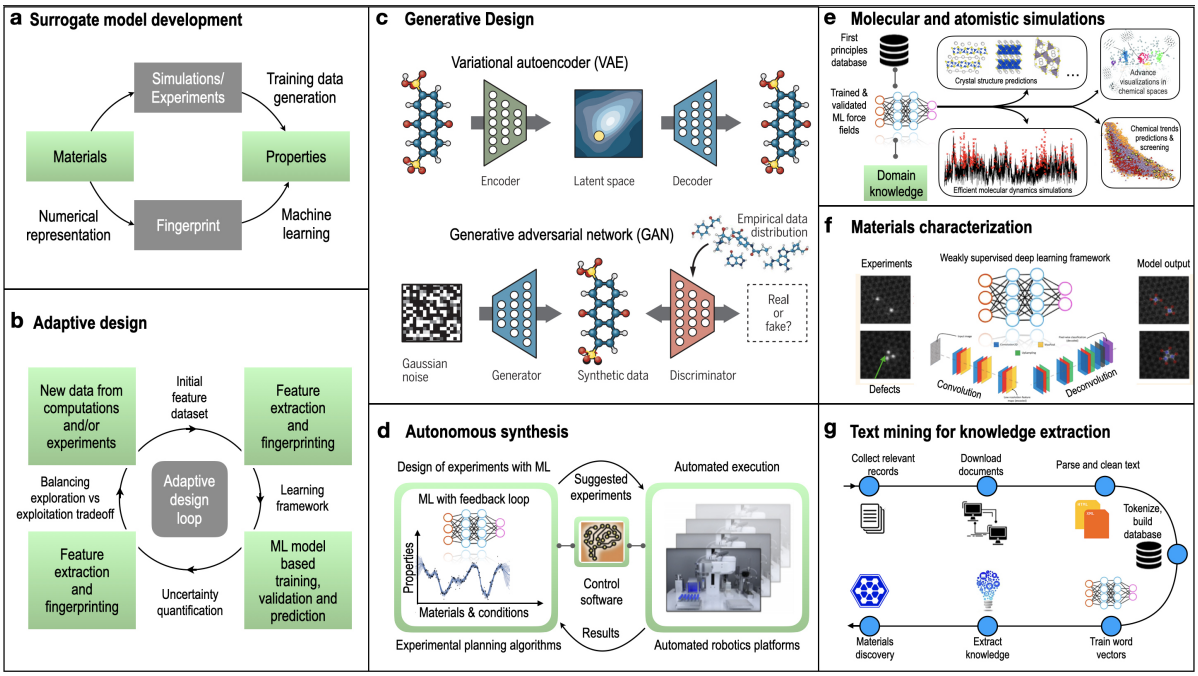


Figure 1: Key application areas of ML in materials science are highlighted and further discussed in the text. (a) Surrogate model development for efficient materials property predictions. (b) Iterative framework for adaptive design and active learning. Adapted from Ref. [16]. (c) Generative materials design using variational autoencoders (VAEs) and generative adversarial networks (GANs). Adapted from Ref. [18]. (d) ML-enabled autonomous materials synthesis via combining design of experiment algorithms with automated robotic platforms. Adapted from Ref. [19]. (e) Use of ML-based force fields to address a range of atomistic materials simulation problems. (f) Deep learning for accurate characterization of atomic-scale materials imaging data. Adapted from Ref. [20]. (g) Use of natural language processing and ML to automatically extract scientific knowledge and insights from scientific texts. Adapted from Ref. [21].

classes spanning over metals and alloys, ceramics and composites, polymers, two-dimensional materials, organic-inorganic hybrids and multicomponent heteroanionic compounds [23, 27, 28]. The applications cover varied length scales, e.g., electronic-, atomic- and meso-scales [16]. Successful attempts for materials property predictions using ML include estimation of energetics [29, 30, 31, 32, 33], phase stability and cation/anion ordering [34, 35, 36, 37, 38, 39, 40], defect energetics [41, 42, 43, 44], bandgaps [45, 46, 47, 48, 49], melting and glass transition temperatures [38, 50, 51], mechanical and elastic properties [52, 53, 54], thermal conductivity [55], dielectric properties [56, 57], tendency for crystallization [58], catalytic activity [59, 60] and radiation damage resistance [61].

2.2. *Materials design and discovery*

Further building on the primary strength of allowing fast yet accurate predictions of materials properties, ML-based surrogate models can be employed in various ways to enable materials design and discovery. In a most straight forward approach, a developed model can be used to make predictions on the entire set of combinatorially-enumerated compounds falling within the domain of applicability of the model. Even more excitingly, multiple property prediction models can be integrated as a part of a hierarchical down-selection pipeline to screen materials based on increasingly complex and stringent criteria employed at each of the subsequent stages [62, 63, 64]. Another approach is to "invert" the forward materials-to-properties prediction route via employing an optimization routine such as evolutionary algorithms, simulated annealing, minima-hopping, or swarm optimization-based routines [63, 65]. Contrary to the direct brute-force enumeration approach that rely on virtual screening of candidate materials from a pre-defined set of possibilities, the optimization-based inversion route focuses on directly predicting a set of materials that satisfy certain pre-specified target objectives, leading to a more general approach to materials discovery. In addition to the enumeration, multi-step screening and optimization-based inversion routes, more sophisticated approaches are being explored by the community to further expedite materials development, as discussed below.

2.2.1. *Active learning*

The ML-based surrogate models discussed above can allow for a quick identification of candidates with tailored properties for further validation via experimental synthesis or more elaborate domain-knowledge-based computations. However, such an approach is inherently passive and does not allow for any control over the prediction errors resulting from the size and quality of the training dataset. Therefore, given a ML model, selection of candidates to perform next experiments or computations on such that the generated data when fed back into the current model leads to the maximum expected improvement (measure in terms of either improving the model or identifying materials with properties falling within or close to the desired range) is a key challenge to achieve optimal experimental de-

sign. In recent years, active-learning algorithms that exploit Bayesian optimization frameworks have been developed to effectively address this challenge [66, 67, 68].

As shown schematically in Fig. 1b, active learning adopts an iterative procedure where predictions using the current ML model are used to guide to the data collection effort in a batch mode to further improve the model [69]. The approach heavily relies on the use of model predictions and uncertainties together with a judiciously selected acquisition or utility function that prioritizes the decision-making process on unseen data. More specifically, the adaptive design loop employs a ML model to achieve a target objective with the smallest possible number of measurements or computations. This is achieved by balancing the exploitation-exploration trade-off during the model development. At any given stage, one can perform the next computation/measurement on the candidate predicted to have the property closest to the desired value (i.e., model exploitation) or try to further improve the model by selecting a material where the predictions are worst in quality (i.e., with largest predictive uncertainties). By choosing the latter, one allows for exploration of less-sampled portions of the design space, leading to an improved model with reduced uncertainties as well as improved likelihood of meeting the objective upon exploitation. A number of recent materials design and discovery efforts have demonstrated the power and utility active learning methods in applications as diverse as design of shape memory alloys with improved thermal hysteresis [70, 71] to identifying Pb-free piezoelectric material with the largest measured electrostrain [72] and from optimizing GaN light emitting diode structures [73] to finding high glass transition temperature polymers [74].

2.2.2. Generative design

In conventional screening and discovery efforts, including past active-learning based efforts, the exploration space is defined generally by a set of candidates that either already exist as a part of a known database or can be systematically enumerated. In contrast, deep-learning-based generative models focus on building a continuous materials vector space, often referred to as latent space. Once the information embedded in the materials training dataset is mapped onto the latent space, it can be used to generate new data points on demand. Furthermore, by building a parallel mapping between the latent space and a property of interest, new materials with the property in a target range can be generated to enable inverse design [18, 75, 76]. In this respect, generative models are a class of deep learning methods that seek to model the underlying probability distribution of both structure and property mapped over a non-linear latent space. The materials generated using these models can be very diverse and considerably distinct, in terms of the functionality they exhibit, from the known materials in the training data. This is because the underlying structure-property relationships are frequently nonlinear in nature for complex functional materials. As a result, the generative design approach presents a higher potential for discovery and novel materials design compared to conventional high throughput virtual screening efforts that are typically limited by the existing materials databases [77].

Schematically illustrated in Fig. 1c, the variational autoencoders (VAEs) [78, 79, 80, 81] and generative adversarial networks (GANs) [82, 83, 84, 85, 86] have recently emerged as the two most popular methods in deep-learning-based generative models. A VAE setup consists of two deep neural networks, namely, the encoder and the decoder. The encoder nonlinearly projects the target chemical space onto a low dimensional latent space, and the decoder implements the inverse mapping allowing for generation of materials corresponding to the specific regions in the latent space. In contrast, a GAN uses a pair of networks—the generator and the discriminator—to learn the underlying materials data distributions implicitly. The generator tries to emulate the real data distribution while the discriminator is tasked to distinguish the generated synthetic (or fake) data from the real data. The overall training process is built around the generator trying to maximize the probability of the discriminator making an error, while the discriminator getting better at catching the fake data.

While a number of exciting studies utilizing the generative power of VAEs and GANs to identify molecules with desired properties have recently been reported [75, 82, 83, 84, 87, 88, 89, 90, 91], applications of these methods to solids have been rather limited owing to the additional challenges associated with representing materials with periodic boundary conditions. Although a number of suitable representations built on composition and configurational details or graph-based encodings exist for solids and have been demonstrated to predict several key properties, as discussed above in Sec. 2.1, most of these representations are not invertible. That is given a representation, the composition and crystal structure details of the material cannot be uniquely identified. On the other hand, any successful domain-specific application requires that the features generated from the latent space should be invertible back to a realistic crystal structure. To address this issue, 3-dimensional voxel image representations have been put forward with some success [76, 92, 93]. However, this route faces challenges associated with images not being translational-, rotational-, and supercell-invariant as well as relatively poor efficiency due to memory intensive nature of the representations leading to longer training times. More recently, a crystal representation inspired by the "point cloud" [94, 95, 96] method (where objects are considered as a set of points and vectors with three-dimensional coordinates) was suggested by Kim et al. [97] to represent the crystal structure as a set of atomic coordinates and cell parameters. The new representation was used with a GAN to generate and explore new crystal structures within the Mg-Mn-O ternary system to find a promising photoanode material for water splitting. Moreover, this inversion-free representation was shown to be more efficient by a factor of ~400 compared to the previously reported image-based representations.

2.2.3. Autonomous synthesis

In the previous sections, we have discussed how ML-based surrogate modeling, active learning and deep learning generative models are being used to expedite chemical space explorations and enable inverse design. The power of ML when combined with automated robotic platforms has led to even more

exciting opportunities in autonomous synthesis and self-driving laboratories [19, 98]. Here it is important to note the distinction between automated versus autonomous systems. While the former refers to robotic platforms that can handle repetitive tasks in a high throughput manner, the latter points specifically to intelligent systems that can adapt appropriately to new information, as and when it becomes available, with little human intervention. In this regard, when compared to automated systems, autonomous systems are very dynamic in nature and can adjust on-the-fly to available information in order to achieve optimal experimental design. As depicted graphically in Fig. 1d, the ability to employ ML algorithms as experiment planner to avoid marginally informative experiments in lieu of the most informative next experiments lies at the heart of the high efficiency boost gained with an autonomous discovery process. These gains in experimental efficiency can be as high as an order of magnitude over conventional high throughput screening approaches [99].

Some of the early studies reporting autonomous materials synthesis targeted unsupervised growth of carbon nanotubes and production of Bose–Einstein condensates [100, 101]. Since then a number of other applications, including discovery of chemical reactions [102, 103], crystallization of giant self-assembled polyoxometalate clusters [104], assembly of layered superlattices [105], synthesis of perovskite quantum dots with tuned bandgaps, quantum yield and composition polydispersity [106], and optimization of synthesis conditions for the formation of high quality organic-inorganic hybrid halide perovskites single crystals [107] have been successfully demonstrated. In addition, open source portable, modular and versatile software packages, such as ChemOS [108], are under active development to enable remote control of self-driving laboratories, provide access to distributed computing resources, and integrate cutting-edge ML methods in a seamless manner. In addition to the three core components, namely the automation hardware, compute resources and ML algorithms, integration of additional auxiliary features such as image and speech recognition, access to on-demand distributed cloud computing resources, improved graphical user interface and web interfaces is expected to both improve their user-friendliness and enrich their capabilities in imminent future [19].

2.3. Molecular and atomistic simulations

Quantum mechanical and classical force field based atomistic simulation methods play a powerful role in modeling and understanding materials behavior and properties via accurate studies of a diverse range of phenomena including thermal and mass transport, phase transformations, chemical reactions, mechanical behavior, materials degradation and failure [109, 110, 111]. From fundamental laws governing interatomic interactions, molecular dynamics [112] (and related atomistic methods) can be used to follow in time the classical equations of motion to enable highly accurate predictions of materials behavior with full atomistic detail. However, the quantum mechanical methods and classical simulations vastly differ in the accuracy (and the concomitant computational cost) of how well they capture details of the interatomic interactions. Quantum mechanics-based methods, such as density functional theory (DFT), are versatile and offer the capability to accurately model a range of chemistries and chemical environments. However, these methods remain computationally very demanding; limiting both the length and time scales of phenomena (to nanometers and picoseconds, respectively) [113] Semi-empirical methods capture the essence of the interatomic interactions in a coarse-grained manner (via parameterized analytical functional forms), and are thus an inexpensive solution to the materials simulation problem [114, 115, 116]. Nevertheless, their applicability is severely restricted to the specific chemistries and chemical environments considered during parameterization, and accuracy cannot be guaranteed for properties not explicitly targeted by the fit. Therefore, one of the goals for ML algorithms in this arena is to help develop potentials (referred to as ML potentials or ML forcefields) that can achieve accuracies approaching to quantum mechanical methods at the cost of semi-empirical methods to accomplish a variety of tasks (for instance, see Fig. 1e).

The last decade has witnessed a tremendous amount of activity and successes in the field of data-driven atomistic simulations and, in particular, in the area of ML forcefields development. Unlike the traditional semi-empirical methods utilizing domain-knowledge-based specific functional forms or rigid parameterizations adopted in semi-empirical methods, ML methods use past accumulated data to make interpolative predictions of the energy and forces in the chemical space of interest. A major challenge in this direction has been the development of configurational representations for ML that respect required symmetry and invariance constraints (e.g., translational, rotational, and exchange of like atoms), capture the details of potential energy surfaces at sub-atomic-level resolution and are "smoothly-varying" (i.e., continuous and differentiable) with respect to small variations in atomic positions. Several local-atomic environment fingerprinting schemes with varying cost accuracy tradeoffs have been proposed, including those based on symmetry functions [117, 118, 119], Coulomb matrices [120, 121], bispectra of neighborhood atomic densities [122], smooth overlap of atomic positions (SOAP) [123, 124, 125], AGNI [126, 127, 128], momentum tensor potential [129] and others [130, 131]. These representations combined with well-established ML algorithms such as kernel ridge regression, Gaussian process regression or deep neural networks have been used widely to explore diverse materials energy landscapes. Transfer learning approach has been particularly promising in accessing cheap surrogate models for highly accurate and computationally demanding beyond-DFT-level energetics [132, 133]. Active learning strategies can also be very effective here in strategically acquiring training data that is uniformly spread out over the target configurational space in order develop robust and effective ML models [133, 134, 135]. In near future, ML-augmented molecular and materials simulations hold promise to significantly narrow the gap between the simulated experimentally-observed time and length scales, while providing high-fidelity predictions of the behavior of matter under varying environmental and processing conditions.

Another direction that has been explored on this fron-

tier deals with using ML to bypass explicit solutions of the Schrodinger's equation (or the Kohn-Sham equation with in the DFT framework) to come up with a much faster linearly-scaling data-enabled route to address the electronic structure problem for materials. A number of good ideas, including learning the kinetic energy functionals as well as learning density-potential and energy-density maps, have been proposed [136, 137, 138, 139, 140]. However, one can argue that the field is still a state of infancy, as most of these studies have dealt largely with toy problems and simple test cases. One notable exception in this direction was presented by Chandrashekhar et al. [141] showing how ML can be utilized to map an external potential (governed solely by the type and positions of nuclei) directly onto the corresponding electronic charge density and local density of states. The predicted density of states and charge density can in turn be utilized to obtain the total energy and other derived properties of the system. The demonstration of the proposed approach on realistic polymeric and metallic systems further shows tremendous promise of similar approaches in an attempt to integrate ML within the inner workings of DFT (and more broadly quantum mechanics).

2.4. Materials characterization

In addition to the modeling, simulations and synthesis, progress of atomically resolved imaging techniques has opened up new avenues for ML-based methods to aid in achieving rapid and quantitative characterization of functional matter under both static and dynamic conditions. Although traditionally characterization techniques have mainly been used to illustrate a material system's qualitative structure or behavior, improved resolution and multi-probe characterization options accessible in modern imaging tools offer much more quantitative and information-rich measurements. For instance, today real-space imaging techniques such as scanning transmission electron microscopy [142, 143], scanning tunneling microscopy [144, 145] and atomic force microscopy [146] permit direct imaging of atomic level structure and functional properties in complex multi-component and multi-phase materials. However, extraction of structure-property relationships from these truly large databases remains a formidable challenge beyond the scope of conventional data-analysis techniques that are based largely on manual inspection by a domain-knowledge expert. Taking advantage of big data sets available from state-of-the-art characterization techniques, recent ML efforts have focused on developing theory-guided mappings between the characterized atomic-level structure and measure the response surfaces (e.g., see Fig. 1f) [147, 20, 148].

In addition, ML-based efficient on-the-fly analysis of materials characterization data can help address workflow bottlenecks in imaging applications [149, 150]. For instance, electron backscatter diffraction (EBSD) technique is routinely employed to obtain three-dimensional spatially resolved crystallographic characterization of polycrystalline samples as large as 10 mm [151, 152] and provides orientation maps at about 200 nm spatial resolution and 0.5 deg crystal orientation resolution. While top-of-the-line commercially available orientation imaging microscopes can make these measurements at unprecedented speeds (one diffraction pattern measurement in less than 1ms) [153]. However, use of traditional indexing techniques for orientation reconstruction from highly noisy EBSD patterns remains a bottleneck, requiring much longer time scales. This has been a major limitation towards implementing efficient real-time orientation indexing required, for instance, to study in-situ microstructure evolution. A number of recent studies have shown that ML based methods are robust to experimentally measured image noise and can be used to index orientations as fast as the highest EBSD scanning rates [150, 154, 155]. Other notable examples of ML-aided characterization include classification of local chemical environments from X-ray absorption spectra [156], identification of two-dimensional heterostructures in optical microscopy [157], automated tuning of microscope controls, data acquisition and analysis [158], phase identification in Raman spectroscopy [159], automated image segmentation and image reconstruction for magnetic resonance imaging [160, 161, 162]. In future, merging the ML-extracted knowledge from materials characterization data with physics-informed models will enable a new paradigm of materials research where theoretical predictions and experimental observations go hand-in-hand at the microscopic levels.

2.5. Automated knowledge extraction from text

A significantly large amount of materials scientific knowledge today exists as text (manuscripts, reports, abstracts etc.), which is continuously growing at an unprecedented rate. However, due to absence of efficient algorithms that can directly extract correlations, connections and relationships from text inputs, this information rich resource remains largely untapped and the materials community has mainly relied on expert-curated and well-structured property databases for materials design and discovery efforts undertaken in the past. However, in the last decade a number of breakthroughs in natural language processing (NLP) have opened up exciting new avenues in materials science and related fields. Most remarkably, use of ML algorithms, such as Word2vec [163, 164] and GloVe [165] to construct high dimensional vector spaces (commonly referred to as embeddings) for words appearing in a text corpus such that their relative semantic and syntactic relationships are preserved has given rise to a generalized approach that can be used to mine scientific literature in a highly effective manner. These word embeddings can capture complex materials science concepts and structure-property relationships direct from text without any need for explicit domain knowledge insertion. These notions are graphically captured in Fig. 1g.

A practical demonstration using this approach to capture latent knowledge from materials science literature was put forward by Tshitoyan et al. [166], who collected and processed materials-related research from approximately 3.3 million scientific abstracts published between 1922 and 2018 in more than 1,000 journals. As a major finding, this study showed that information regarding future discoveries already exists, to a large extent, in past publications in a latent form and therefore such NLP models can potentially recommend new functional mate-

rials several years before their normal course of discovery timeline.

In the same vein as the previous study, using polymers as an example class of functional materials, Shetty and Ramprasad also confirmed that materials science knowledge can be automatically inferred from textual information contained in scientific texts [21]. Using a data set of nearly 0.5 million polymer papers, it was shown that vector representations trained for every word appearing in the accumulated text corpus were able to capture crucial materials knowledge in a completely unsupervised manner. Subsequently, ML-based temporal studies aimed at tracking popularity of various polymers for different applications were able to identify new polymers for novel applications based solely on the domain knowledge contained in mined database.

Another challenging problem that has been targeted with automated text mining via combining ML and NLP pertains to identifying realistic materials synthesis routes. In particular, Kim et al. [167] demonstrated use of ML methods to successfully predict the critical synthesis parameters needed to make targeted materials–titania nanotubes via hydrothermal methods in this particular case–where the training dataset was automatically compiled from tens of thousands of scholarly publications using NLP techniques. More importantly, the study also showed the capacity for transfer learning for the developed ML models by predicting synthesis outcomes on materials systems not included in the original training set.

As evident from the examples discussed in this section, ML-based methods and algorithms have found a wide range of applications within the field of materials design and development. The developed models largely rely on materials descriptors or features to numerically represent details of the problem, such as, chemical composition and configurational structure of the material, processing conditions and relevant environmental factors. The choice of an appropriate descriptor set is a crucial step of enormous importance. The choice of an initial set is typically based either solely on the underlying domain knowledge of the problem (*i.e.*, mechanistic details, well-established and physically-intuitive relationships, constitutive laws of physics and chemistry *etc.*) or on an unbiased selection using the available data starting from a very large set of combinatorial possibilities. One can argue that both the route have their own pros and cons. while the former approach is likely to result in models that are more amenable to physical interpretation, the latter harbors an increased potential for discoveries that are typically beyond the realm of conventional wisdom. Regardless, an exploratory analysis utilizing several approaches to the problem at hand during a ML model building exercise is always helpful. Eventually, our ability to not only generate transparent ML models, but also extract physical insights from these surrogates, while preserving the potential for discovery that is intrinsic to the data-enabled methods, will dictate the extent of the impact of materials informatics on the field.

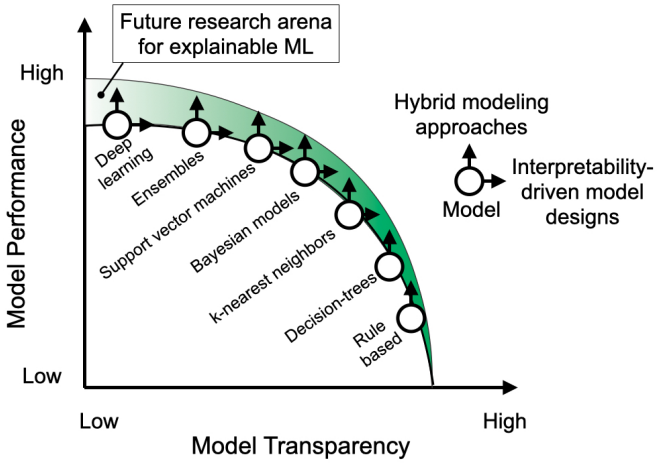


Figure 2: Schematic showing Trade-off between model performance and transparency. The area of potential future improvement due to improved explainable AI techniques and tools is highlighted in green with the improvement directions represented with arrows. Adapted from Ref. [14].

3. Physical Insights from Materials Learning

*3.1. **Performance-transparency tradeoffs***

In addition to deliver robust and accurate predictions, ML models in physical sciences are often required to provide new scientific understanding and physical insights directly from observational or simulated data. As a prerequisite to domain knowledge extraction via ML is explainability–the ability to rationalize individual predictions by examining inner workings of a transparent model and further interpreting the outcomes in combination with expert-knowledge. Therefore, a collection of interpretations for a transparent model when evaluated by a domain-knowledge expert leads to explainability. Within this context, transparency is largely confined to details of the employed ML model (i.e., details pertaining to the specific choices of model class, model complexity, learning algorithm employed, hyper parameters, initial constraints etc.), while *interpretability* combines both the input data as well as the ML model to make sense of the output. Going from *interpretability* to *explainability* requires involvement of human with a scientific understanding of the problem. In the quest to learn from learning machines or intelligible intelligence widely acceptable concepts of transparency, interpretability and explainability have recently emerged as the core elements of utmost importance that are deemed necessary to enable scientific outcomes from ML endeavors [168].

The aforementioned notions are directly connected to model complexity. Simple, and therefore transparent ML models are highly amenable to interpretations and explanations, however, generally suffer from relatively poor accuracy and reliability as compared to more complex "black-box" type models. Therefore, similar to the well-known bias-variance tradeoff that are provoked to prevent overfitting while building a robust predictive ML model, balancing of a performance (reliability and ac-

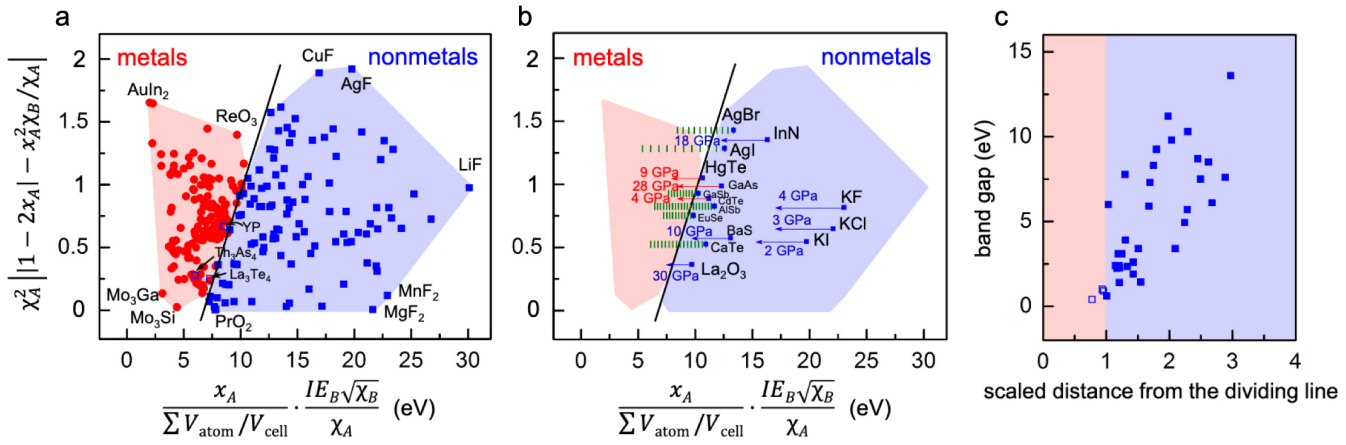


Figure 3: An example depicting SISSO classification performance in separating metals from insulator. (a) A near-perfect classification of metal/nonmetal for 299 binary A_xB_y-type materials. Symbols χ, IE and x represent Pauling electronegativity, ionization energy and atomic composition, respectively. V_{atom}/V_{cell} represents packing factor. Red circles, blue squares, and open blue squares represent metals, non-metals, and the three erroneously characterized non-metals, respectively. (b) Representation of pressure induced insulator to metal transitions (red arrows) and materials that remain insulators upon compression (blue arrows). Computational predictions at step of 1 GPa are shown with green bars. (c) Correlation between the band gap of the non-metals and the scaled coordinate from the dividing line. Adapted from Ref. [169], with permissions.

curacy) versus transparency tradeoff needs to be carefully considered for explainable ML models.

3.2. *Hybrid and local-learning approaches for improved transparency*

Given the common scenarios where model performance closely accompanies model complexity, model transparency (and therefore, interpretability and explainability) exhibits a downwards slope that has largely remained unavoidable in the past. This situation is graphically represented in Fig. 2, where deep neural networks are at one extreme offering excellent performance but little transparency [14]. The other extreme of high transparency with relatively lower performance is occupied by decision trees and rule-based algorithms that are completely interpretable. However, going beyond traditional single-model frameworks, more sophisticated hybrid methods have been suggested to simultaneously improve model transparency and performance [170, 171, 172, 173]. For instance, Kailkhura et al. [170] recently presented an approach that first transforms a regression problem into a multi-class classification problem on a sub-sampled training data to balance the distribution of the least represented material classes. Subsequently, smaller and simpler models for the different classes are trained to gain better understanding of different subdomain-specific regimes. This domain-specific learning enabled a rationale generator component to the framework which can provide both model-level and decision-level explanations. This led to improvements in the overall transparency and explainability of the model as compared to the conventional approach of training just one regression model for the entire dataset. Finally, a transfer learning technique harnessing correlations between multiple properties was employed to compensate for the model performance reduction as a result of improved transparency. In a different study, Sutton et al. [171] presented a sub-group discovery based new approach to identify domains of applicability of ML models and showed that the domain-specific learning is not only crucial for a deeper understanding and improved interpretability, but can also significantly improve prediction performance for certain domains. The idea of fitting local domain-specific model to gain improved understanding of otherwise opaque ML models lies at the heart of the local interpretable model-agnostic explanations (LIME) algorithm [172] that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model. In future, further developments in the direction of transparency-preserving hybrid modeling approaches and focus on interpretability-driven new model developments are going to further expand these frontiers, highlighted in green in Fig. 2.

3.3. *Causality- and consistency-based validations*

An explainable model further opens doors for devising testable hypothesis or more stringent validation tests for specific predictions to address their consistency, generalizability and causality. A compelling example demonstration in this direction was presented by Ouyang et al. using sure independent screening and sparsifying operator (SISSO) method based on the compressed sensing technique [169]. Note that this method allows for efficient exploration of vast descriptor spaces—with the number of unique descriptors typically reaching up to several billions–to identify transparent analytical descriptor-property relationships and has been widely applied to address a diverse set of materials design and discovery problems [34, 169, 174, 175, 176]. Ouyang et al. applied a SISSO-based approach to learn an accurate, transparent and predictive metal-insulator classification model for binary AxBy-type compounds [169]. Simple two-dimensional analytical descriptors found by SISSO led to almost perfect classification (with 99.0% accuracy) of metal versus nonmetal chemistries for a set of to-

tal 299 compounds (see Fig. 3a). More interestingly, to conclusively show that the discovered descriptors indeed bore a causal relationship with the metallic or insulating behavior exhibited by the materials, the model was employed to rediscover the available pressure-induced insulator to metal transitions with a number of chemistries that were known to undergo such a transition laying consistently near the classification boundary, as shown in Fig. 3b. Furthermore, the model was able to make additional prediction of yet unknown transition candidates, ripe for experimental validation. As an additional evidence of an underlying causal relationship was provided by depicting a qualitative yet clear trend between the experimental band gap of the insulators versus the scaled distance from the dividing line (Fig. 3c). In a similar spirit of finding accurate symbolic expressions that match data from an unknown function, Udrescu et al. [177] developed a recursive multidimensional symbolic regression algorithm, named AI Feynman, and demonstrated rediscovery of a set of 100 hand-picked equations from the Feynman Lectures on Physics [178, 179, 180]. These contributions suggest that compressed sensing and symbolic regression-based techniques, combined with appropriately identified domain knowledge-based constraints can be enormously helpful in gaining physical insights from materials data.

3.4. ***Informatics-enhanced design maps***

Efficient interpolation ability of ML algorithms in high-dimensional spaces can be harnesses in development of informatics-enhanced design maps that are much more informative and information rich as compared to traditional methods that have largely employed two-dimensional maps. As an example, Fig. 4 compares a traditional tolerance factor versus octahedral factor structure map often invoked to identify formable perovskite oxides [181]. Indeed, the pair of geometrical descriptors shows a remarkable predictive power and all the known compounds that has been successfully synthesized in a perovskite crystal structure tend to cluster in this plot as depicted in Fig. 4a. A shortcoming of such an approach, however, could be that the descriptor pair is solely based on size effects (i.e., coordination environment dependent Shannon's ionic radii [182]) and completely ignores aspects of local bonding interactions, such as, ionicity versus covalency, relative electronegatively differences between different cations etc. which might also play an important role in dictating formability in perovskites. Although one can argue that some of these aspects are implicitly accounted for in the relative atomic and ionic size trends, the ability to explicitly incorporate additional relevant factors that might play a role might significantly improve such conventional maps in terms of their predictive power. For instance, Fig. 4b shows an analogous plot which is generated by a random forest ML model that was trained and validated on a much larger set of descriptors, including octahedral and tolerance factors as well as electronegativities, ionization potentials, electron affinities, orbital-dependent pseudo potential radii of the cations. Once the model has been trained and validated, it can be used to make probabilistic estimates of perovskite formability in the entire multi-dimensional input feature space and these predictions can be projected back on to a two-dimensional plot for the two classic geometric factors, while integrating out or marginalizing all the other feature dimensions, as shown in Fig. 4b. One might argue that the Fig. 4b is more informative since it implicitly contains trends reflected by the entire set of descriptors that were used to train the model and not just the tolerance and octahedral factors, as in the case of Fig. 4a. Moreover, the informatics-based route allows for generation of analogous plots for any pair of features drawn from the original input feature set. Here we note that a closely related approach is readily available in tree-based ensemble models known as partial dependence plots [183]. Although we have focused on a relatively simple example, it is not hard to imagine much more complex situations where such an approach can be applied. In complex materials design problems, the ability to construct such design maps to explore and rationalize intricate trends and tradeoffs among key design variables can be enormously helpful.

Finally, we note that a great deal of research has lately gone into the development of explainable deep learning techniques and the efforts have been reviewed in a number of surveys [14, 184, 185, 186, 187]. While it is impractical to delve deep into this large body of work, we note here in passing that, at a higher level, explainable deep learning methods largely fall in three broad categories, namely, visualization, model distillation and intrinsic methods [187]. As the nomenclature suggests, visualization methods rely scientific visualization to single out key characteristics of an input that strongly influence the output to generate an explanation. Model distillation approach resorts to a separate, "glass-box" ML model that is trained to mimic the input-output behavior of original "black-box" model but in a more transparent manner by identifying specific decision rules that lead to the final output. Intrinsic methods, on the other hand, have an explanation system integrated within by design and therefore can balance the transparency-performance tradeoff on-the-fly by jointly optimize both model performance and some quality of the explanations produced. In future, as materials datasets grow larger, these techniques will play roles of increasing importance in materials design problems.

4. Challenges and Opportunities Ahead

As briefly alluded to above, over the past decade the field of materials informatics has grown exponentially. While the early phase of this growth was largely focused on developing a deeper understanding of ML model development itself with a primary focus on testing efficacy and efficiency of the data-enabled approaches in materials development. In this phase, studies emphasized on addressing basic questions such as: "How different statistical learning methods work?"; "What are their potential strengths and weaknesses?"; "How does one select appropriate method for a given problem?"; "What are some best practices of statistical learning that one should follow for developing and validating a predictive model?" etc. These efforts have culminated in democratization of the process of training a model on a materials data with the ability of several open source ML

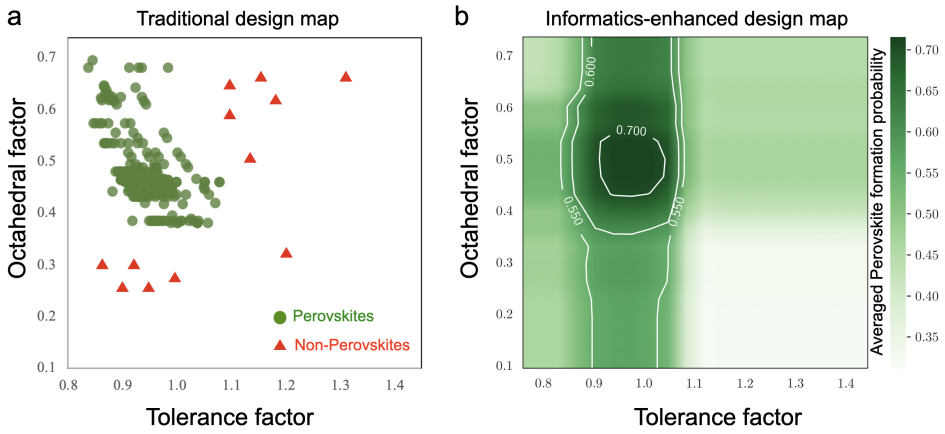


Figure 4: A comparison of structure maps between the tolerance and octahedral factors for perovskite formability. (a) Conventional structure map with a scatter plot. Perovskite formability region is given by a convex hull encompassing the known examples (green circles). (b) Informatics-enhanced structure map with the same set of variables, explicitly accounting for the probability of formation.

packages and repositories for model development and dissemination. Now that the field has matured into an established discipline from a specialized area of research, the research focus has shifted to a number of more general materials-science-specific issues that the community is currently grappling with.

Although ML problems are frequently referred to as "big data" problems, datasets used in materials design and discovery problems are generally relatively small, barring certain cases dealing with small molecules or imaging data for materials characterization. A large fraction of materials data available in open sources materials databases comes from first principles computations with a major emphasis on the ground state atomic structures and energetics. Availability of high-quality data on most functional properties generated via direct experimental measurements is rather limited. On the other hand, an informatics effort that targets to discover a new functional material with a desired set of properties usually requires a dataset with several compounds in a target compositional and configurational space (i.e., for given chemistries and crystal structures) with entries on multiple properties, spanned over a range of processing conditions. Such datasets are extremely difficult to populate starting from publicly available materials data. In fact, accumulation and curation of an initial high-quality dataset remains a highly laborious and time-consuming step for most of the materials design efforts today. To address this data scarcity problem going forward, development and wide-spread use of data-mining and NLP-based high throughput data acquisition techniques and advanced methods to extract data directly from graphics that permit a much faster and semi-automated extraction of materials datasets from past literature is a crucial next step.

In past, materials development has largely been led by chemical intuition guided explorations. Moreover, results on failed experiments are rarely reported in the peer-reviewed literature. As a result, in addition to being sparce, the available data distributions can be highly imbalanced and skewed, violating one of the central assumptions of most standard ML methods requiring uniformly sampled and balanced training data. Furthermore, data coming from different sources can have varying levels of noise. A robust model development in such situations demands for more advanced analysis going beyond standard predictive-accuracy-centric "testing on unseen data" approach. Advance methods for rigorous uncertainly quantification, establishment of domain of applicability, effectively correcting class imbalance problems and skewed data distributions are just beginning to find inroads into materials informatics [170, 171, 188].

In addition to being sparce and skewed, materials data can be highly heterogeneous and can be generated at varying levels of fidelities. Furthermore, due to the underlying cost-accuracy tradeoffs in both experimental and computational techniques available for data acquisition, a larger amount of data is generally available at a lower fidelity level. Development and use of advance algorithms that allow for effective integration of information coming in from varying fidelity sources, while explicitly accounting for different noise levels in the different data segments, to make predictions at the highest level of fidelity (i.e., at the highest predictive accuracy and lowest uncertainty levels) is highly desirable in materials informatics [189]. Such algorithms also provide a means to address the data-scarcity problem as they have been proven effective in learning-from-small-datasets scenarios [190].

In addition to the aforementioned challenges that largely concern with the amount and quality of available training data, effective approaches that enable integration of domain-knowledge with ML could be transformative. In this direction, both the ability to put in domain knowledge into a ML model as well as extract new physical insights from an explainable ML model should be considered. On one hand ML algorithms that can directly integrate available mechanistic understanding and known domain knowledge (in terms of physical laws and well-established principles, constraints such as boundary conditions, asymptotic limits, smoothness criteria, symmetries, in-

variances, and other problem-specific knowledge obtained from theory and simulations) to train more efficiently with smaller datasets are required [191]. On the other hand, breakthroughs in terms of implementing hybrid and locally-interpretable models (as discussed in Sec. 3) to explain ML predictions are sought. Going beyond standard statistical validation techniques, more stringent domain-specific validation criteria, using either direct experimentation or rigorously validated first principles computational methods, are required to establish that the identified correlations indeed represent causal relationships with truly predictive power.

Finally, to facilitate documentation, dissemination and effective use of highly multi-scale, multidimensional and heterogeneous nature of materials data, it is desirable to develop new file formats and data structures that are flexible enough to handle this level of complexity. Encouraging documentation of not just the data, but also the relevant metadata–providing a much richer context for the primary data–across the community would be increasingly helpful going forward as text to knowledge mining methods progress. Infrastructure development for not just sharing the data, but the developed ML models themselves would be desired to address one of the most important issue of reproducibility. Some efforts in these directions are currently in progress [192, 193, 194, 195]. Going forward, cultivating a culture that encourages publishing results from failed experiments and adoption of publication file formats that enable by-design an efficient data extraction via text mining will open new avenues for information-rich materials datasets. As ML methods become increasingly popular and more widely used within the materials community, addressing these challenges becomes critical for expediting the pace of progress.

5. Conclusions

ML and data-enabled methods represent the advent of a new paradigm in materials science. As a result, the way in which materials design and discovery has traditionally been pursued in the field is poised to change in profound ways. Starting from a niche area, within a short period of time materials informatics has already been established as a full-blown mature discipline. ML algorithms are already aiding in efficient materials property predictions, materials design and discovery, as well as different components of experimental design, dealing with identification, and the organization and prioritization of next experiments. Going forward, a number of crucial challenges pertaining to accessibility and quality of data as well as regarding integrating domain-knowledge into ML models (beyond the means of feature selection and feature engineering) and extracting novel insights out of the trained models need to be addressed. Upon success, materials science in coming decades will be defined by our ability to learn from *learning machines*.

CRediT authorship contribution statement Ghanshyam Pilania: Conceptualization, Writing – reviewing and editing.

Acknowledgements I would like to acknowledge the many fellow researchers, collaborators and mentors whom I had a chance to work with in the last few years. I also acknowledge the financial support from Laboratory Directed Research and Development (LDRD) program of the Los Alamos National Laboratory (LANL) via multiple projects (#20190001DR and # 20190043DR). LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

References

[1] D. R.-J. G. Rydning, The digitization of the world from edge to core, Framingham: International Data Corporation (2018).
[2] P. Larrañaga, D. Atienza, J. Diaz-Rozo, A. Ogbechie, C. E. Puerto-Santana, C. Bielza, Industrial applications of machine learning, CRC Press, 2018.
[3] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.
[4] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, Vol. 1, MIT press Cambridge, 2016.
[5] C. Manning, H. Schutze, Foundations of statistical natural language processing, MIT press, 1999.
[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.
[7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Mastering the game of go without human knowledge, nature 550 (7676) (2017) 354–359, publisher: Nature Publishing Group.
[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, Mastering the game of go with deep neural networks and tree search, nature 529 (7587) (2016) 484–489, publisher: Nature Publishing Group.
[9] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, Mastering chess and shogi by self-play with a general reinforcement learning algorithm, arXiv preprint arXiv:1712.01815 (2017).
[10] M. Moravčík, M. Schmid, N. Burch, V. Lisy, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, M. Bowling, Deepstack: Expert-level artificial intelligence in heads-up no-limit poker, Science 356 (6337) (2017) 508–513, publisher: American Association for the Advancement of Science.
[11] N. Brown, T. Sandholm, Superhuman AI for heads-up no-limit poker: Libratus beats top professionals, Science 359 (6374) (2018) 418–424, publisher: American Association for the Advancement of Science.
[12] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, Building watson: An overview of the DeepQA project, AI magazine 31 (3) (2010) 59–79.
[13] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, publisher: IEEE.
[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115, publisher: Elsevier.
[15] D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science, Annual Review of Materials Research 50, publisher: Annual Reviews (2020).
[16] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Computational Materials 3 (1) (2017) 1–13, publisher: Nature Publishing Group.
[17] T. Mueller, A. G. Kusne, R. Ramprasad, Machine learning in materials science: Recent progress and emerging applications, Reviews in Computational Chemistry 29 (2016) 186–273, publisher: Wiley Online Library.

[18] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (6400) (2018-07-27) 360–365, publisher: American Association for the Advancement of Science Section: Review. doi:10.1126/science.aat2663.
URL https://science.sciencemag.org/content/361/6400/360

[19] F. Häse, L. M. Roch, A. Aspuru-Guzik, Next-generation experimentation with self-driving laboratories, Trends in Chemistry 1 (3) (2019) 282–291, publisher: Elsevier.

[20] M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R. R. Unocic, R. Vasudevan, S. Jesse, S. V. Kalinin, Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations, ACS nano 11 (12) (2017) 12742–12752, publisher: ACS Publications.

[21] P. Shetty, R. Ramprasad, Automated knowledge extraction from polymer literature using natural language processing, Iscience 24 (1) (2021) 101922, publisher: Elsevier.

[22] R. Batra, L. Song, R. Ramprasad, Emerging materials intelligence ecosystems propelled by machine learning, Nature Reviews Materials (2020) 1–24Publisher: Nature Publishing Group.

[23] J. Schmidt, M. R. Marques, S. Botti, M. A. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Computational Materials 5 (1) (2019) 1–36, publisher: Nature Publishing Group.

[24] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555, publisher: Nature Publishing Group.

[25] L. Ward, C. Wolverton, Atomistic calculations and materials informatics: A review, Current Opinion in Solid State and Materials Science 21 (3) (2017) 167–176, publisher: Elsevier.

[26] A. Jain, G. Hautier, S. P. Ong, K. Persson, New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships, Journal of Materials Research 31 (8) (2016) 977–994, publisher: Cambridge University Press.

[27] G. Pilania, P. V. Balachandran, J. E. Gubernatis, T. Lookman, Data-based methods for materials design and discovery: Basic ideas and general methods, Synthesis Lectures on Materials and Optics 1 (1) (2020) 1–188, publisher: Morgan & Claypool Publishers.

[28] L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, R. Ramprasad, Polymer informatics: Current status and critical next steps, Materials Science and Engineering: R: Reports 144 (2021) 100595, publisher: Elsevier.

[29] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals, Physical review letters 117 (13) (2016) 135502, publisher: APS.

[30] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Physical Review B 89 (9) (2014) 094104, publisher: APS.

[31] A. M. Deml, R. O'Hayre, C. Wolverton, V. Stevanović, Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression, Physical Review B 93 (8) (2016) 085142, publisher: APS.

[32] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, How chemical composition alone can predict vibrational free energies and entropies of solids, Chemistry of Materials 29 (15) (2017) 6220–6227, publisher: ACS Publications.

[33] A. Talapatra, B. P. Uberuaga, C. R. Stanek, G. Pilania, A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides, Chemistry of MaterialsPublisher: ACS Publications.

[34] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Science advances 5 (2) (2019) eaav0693, publisher: American Association for the Advancement of Science.

[35] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, Frontiers in Materials 3 (2016) 19, publisher: Frontiers.

[36] G. Pilania, J. E. Gubernatis, T. Lookman, Classification of octet AB-type binary compounds using dynamical charges: a materials informatics perspective, Scientific reports 5 (2015) 17504, publisher: Nature Publishing Group.

[37] G. Pilania, P. V. Balachandran, J. E. Gubernatis, T. Lookman, Classification of ABO3 perovskite solids: a machine learning study, Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials 71 (5) (2015) 507–513, publisher: International Union of Crystallography.

[38] G. Pilania, J. E. Gubernatis, T. Lookman, Structure classification and melting temperature prediction in octet AB solids via machine learning, Physical Review B 91 (21) (2015) 214302, publisher: APS.

[39] G. Pilania, A. Ghosh, S. T. Hartman, R. Mishra, C. R. Stanek, B. P. Uberuaga, Anion order in oxysulfide perovskites: origins and implications, npj Computational Materials 6 (1) (2020) 1–11, publisher: Nature Publishing Group.

[40] G. Pilania, X.-Y. Liu, Machine learning properties of binary wurtzite superlattices, Journal of materials science 53 (9) (2018) 6652–6664, publisher: Springer.

[41] B. Medasani, A. Gamst, H. Ding, W. Chen, K. A. Persson, M. Asta, A. Canning, M. Haranczyk, Predicting defect behavior in b2 intermetallics by merging ab initio modeling and machine learning, npj Computational Materials 2 (1) (2016) 1–10, publisher: Nature Publishing Group.

[42] A. Mannodi-Kanakkithodi, M. Y. Toriyama, F. G. Sen, M. J. Davis, R. F. Klie, M. K. Chan, Machine-learned impurity level prediction for semiconductors: the example of cd-based chalcogenides, NPJ Computational Materials 6 (1) (2020) 1–14, publisher: Nature Publishing Group.

[43] V. Sharma, P. Kumar, P. Dev, G. Pilania, Machine learning substitutional defect formation energies in ABO3 perovskites, Journal of Applied Physics 128 (3) (2020) 034902, publisher: AIP Publishing LLC.

[44] R. Batra, G. Pilania, B. P. Uberuaga, R. Ramprasad, Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia, ACS applied materials & interfaces 11 (28) (2019) 24906–24918, publisher: ACS Publications.

[45] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, The journal of physical chemistry letters 9 (7) (2018) 1668–1673, publisher: ACS Publications.

[46] A. Mishra, S. Satsangi, A. C. Rajan, H. Mizuseki, K.-R. Lee, A. K. Singh, Accelerated data-driven accurate positioning of the band edges of MXenes, The journal of physical chemistry letters 10 (4) (2019) 780–785, publisher: ACS Publications.

[47] A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, A. K. Singh, Machine-learning-assisted accurate band gap predictions of functionalized MXene, Chemistry of Materials 30 (12) (2018) 4031–4038, publisher: ACS Publications.

[48] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Scientific reports 6 (2016) 19375, publisher: Nature Publishing Group.

[49] G. Pilania, J. E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Computational Materials Science 129 (2017) 156–163, publisher: Elsevier.

[50] G. Pilania, C. N. Iverson, T. Lookman, B. L. Marrone, Machine-learning-based predictive modeling of glass transition temperatures: A case of polyhydroxyalkanoate homopolymers and copolymers, Journal of Chemical Information and Modeling 59 (12) (2019) 5013–5025, publisher: ACS Publications.

[51] A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids, Physical Review B 89 (5) (2014) 054303.
URL https://journals.aps.org/prb/abstract/10.1103/PhysRevB.89.0

[52] M. De Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, A. Gamst, A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds, Scientific reports 6 (2016) 34256, publisher: Nature Publishing Group.

[53] S. Aryal, R. Sakidja, M. W. Barsoum, W.-Y. Ching, A genomic approach to the stability, elastic, and electronic properties of the MAX phases, physica status solidi (b) 251 (8) (2014) 1480–1497, publisher: Wiley Online Library.

[54] S. Chatterjee, M. Murugananth, H. Bhadeshia, δ TRIP steel, Materials

Science and Technology 23 (7) (2007) 819–827, publisher: Taylor & Francis.

[55] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization, Physical review letters 115 (20) (2015) 205901.
URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.115.205901

[56] C. Kim, G. Pilania, R. Ramprasad, From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown, Chemistry of Materials 28 (5) (2016) 1304–1311, publisher: ACS Publications.

[57] C. Kim, G. Pilania, R. Ramprasad, Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX3 perovskites, The Journal of Physical Chemistry C 120 (27) (2016) 14575–14580, publisher: ACS Publications.

[58] S. Venkatram, R. Batra, L. Chen, C. Kim, M. Shelton, R. Ramprasad, Predicting crystallization tendency of polymers using multifidelity information fusion and machine learning, The Journal of Physical Chemistry B 124 (28) (2020) 6046–6054, publisher: ACS Publications.

[59] M. Andersen, S. V. Levchenko, M. Scheffler, K. Reuter, Beyond scaling relations for the description of catalytic materials, ACS Catalysis 9 (4) (2019) 2752–2759, publisher: ACS Publications.

[60] B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan, W.-J. Yin, Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts, Nature communications 11 (1) (2020) 1–8, publisher: Nature Publishing Group.

[61] G. Pilania, K. R. Whittle, C. Jiang, R. W. Grimes, C. R. Stanek, K. E. Sickafus, B. P. Uberuaga, Using machine learning to identify factors that govern amorphization of irradiated pyrochlores, Chemistry of Materials 29 (6) (2017) 2574–2583.
URL http://pubs.acs.org/doi/abs/10.1021/acs.chemmater.6b04662

[62] V. Sharma, C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, Rational design of all organic polymer dielectrics, Nature communications 5 (1) (2014) 1–8, publisher: Nature Publishing Group.

[63] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad, Machine learning strategy for accelerated design of polymer dielectrics, Scientific reports 6 (2016) 20952, publisher: Nature Publishing Group.

[64] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Scientific reports 3 (1) (2013) 1–6, publisher: Nature Publishing Group.

[65] A. Zunger, Inverse design in search of materials with target functionalities, Nature Reviews Chemistry 2 (4) (2018) 1–16, publisher: Nature Publishing Group.

[66] W. B. Powell, The knowledge gradient for optimal learning, Wiley Encyclopedia of Operations Research and Management SciencePublisher: Wiley Online Library (2010).

[67] W. B. Powell, I. O. Ryzhov, Optimal learning, Vol. 841, John Wiley & Sons, 2012.

[68] I. O. Ryzhov, W. B. Powell, P. I. Frazier, The knowledge gradient algorithm for a general class of online learning problems, Operations Research 60 (1) (2012) 180–195, publisher: INFORMS.

[69] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, npj Computational Materials 5 (1) (2019) 1–17, publisher: Nature Publishing Group.

[70] D. Xue, D. Xue, R. Yuan, Y. Zhou, P. V. Balachandran, X. Ding, J. Sun, T. Lookman, An informatics approach to transformation temperatures of NiTi-based shape memory alloys, Acta Materialia 125 (2017) 532–541, publisher: Elsevier.

[71] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nature communications 7 (1) (2016) 1–9, publisher: Nature Publishing Group.

[72] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, Accelerated search for BaTiO3-based piezoelectrics with vertical morphotropic phase boundary using bayesian learning, Proceedings of the National Academy of Sciences 113 (47) (2016) 13301–13306, publisher: National Acad Sciences.

[73] B. Rouet-Leduc, K. Barros, T. Lookman, C. J. Humphreys, Optimisation of GaN LEDs and the reduction of efficiency droop using active machine learning, Scientific reports 6 (2016) 24862, publisher: Nature Publishing Group.

[74] C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, Active-learning and materials design: the example of high glass transition temperature polymers, MRS Communications 9 (3) (2019) 860–866, publisher: Cambridge University Press.

[75] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS central science 4 (2) (2018) 268–276, publisher: ACS Publications.

[76] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, Inverse design of solid-state materials via a continuous representation, Matter 1 (5) (2019) 1370–1384, publisher: Elsevier.

[77] Q. Vanhaelen, Y.-C. Lin, A. Zhavoronkov, The advent of generative chemistry, ACS Medicinal Chemistry Letters 11 (8) (2020) 1496–1505, publisher: ACS Publications.

[78] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[79] C. Doersch, Tutorial on variational autoencoders, arXiv preprint arXiv:1606.05908 (2016).

[80] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, arXiv preprint arXiv:1401.4082 (2014).

[81] R. Batra, H. Dai, T. D. Huan, L. Chen, C. Kim, W. R. Gutekunst, L. Song, R. Ramprasad, Polymers for extreme conditions designed using syntax-directed variational autoencoders, Chemistry of MaterialsPublisher: ACS Publications (2020).

[82] E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, Reinforced adversarial neural computer for de novo molecular design, Journal of chemical information and modeling 58 (6) (2018) 1194–1204, publisher: ACS Publications.

[83] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC)Publisher: ChemRxiv (2017).

[84] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, arXiv preprint arXiv:1705.10843 (2017).

[85] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[86] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in neural information processing systems, 2016, pp. 2234–2242.

[87] N. De Cao, T. Kipf, MolGAN: An implicit generative model for small molecular graphs, arXiv preprint arXiv:1805.11973 (2018).

[88] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico, Molecular pharmaceutics 14 (9) (2017) 3098–3104, publisher: ACS Publications.

[89] T. Blaschke, M. Olivecrona, O. Engkvist, J. Bajorath, H. Chen, Application of generative autoencoder in de novo molecular design, Molecular informatics 37 (1) (2018) 1700123, publisher: Wiley Online Library.

[90] L. Yu, W. Zhang, J. Wang, Y. Yu, Seqgan: Sequence generative adversarial nets with policy gradient, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[91] J. Lim, S. Ryu, J. W. Kim, W. Y. Kim, Molecular generative model based on conditional variational autoencoder for de novo molecular design, Journal of cheminformatics 10 (1) (2018) 1–9, publisher: BioMed Central.

[92] J. Hoffmann, L. Maestrati, Y. Sawada, J. Tang, J. M. Sellier, Y. Bengio, Data-driven approach to encoding and decoding 3-d crystal structures, arXiv preprint arXiv:1909.00949 (2019).

[93] B. Kim, S. Lee, J. Kim, Inverse design of porous materials using artificial neural networks, Science advances 6 (1) (2020) eaax9324, publisher:

American Association for the Advancement of Science.
[94] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, Pointpillars: Fast encoders for object detection from point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12697–12705.
[95] J. Li, B. M. Chen, G. Hee Lee, So-net: Self-organizing network for point cloud analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9397–9406.
[96] W. Wu, Z. Qi, L. Fuxin, Pointconv: Deep convolutional networks on 3d point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9621–9630.
[97] S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik, Y. Jung, Generative adversarial networks for crystal structure prediction, arXiv preprint arXiv:2004.01396 (2020).
[98] B. P. MacLeod, F. G. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. Yunker, M. B. Rooney, J. R. Deeth, Self-driving laboratory for accelerated discovery of thin-film materials, Science Advances 6 (20) (2020) eaaz8867, publisher: American Association for the Advancement of Science.
[99] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, Accelerating the discovery of materials for clean energy in the era of smart automation, Nature Reviews Materials 3 (5) (2018) 5–20, publisher: Nature Publishing Group.
[100] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, Autonomy in materials research: a case study in carbon nanotube growth, npj Computational Materials 2 (1) (2016) 1–6, publisher: Nature Publishing Group.
[101] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. W. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. D. Quinlivan, P. Manju, C. C. Kuhn, Fast machine-learning online optimization of ultra-cold-atom experiments, Scientific reports 6 (1) (2016) 1–6, publisher: Nature Publishing Group.
[102] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, Controlling an organic synthesis robot with machine learning to search for new reactivity, Nature 559 (7714) (2018) 377–381, publisher: Nature Publishing Group.
[103] V. Dragone, V. Sans, A. B. Henson, J. M. Granda, L. Cronin, An autonomous organic reaction search engine for chemical reactivity, Nature communications 8 (1) (2017) 1–8, publisher: Nature Publishing Group.
[104] V. Duros, J. Grizou, W. Xuan, Z. Hosni, D.-L. Long, H. N. Miras, L. Cronin, Human versus robots in the discovery and crystallization of gigantic polyoxometalates, Angewandte Chemie International Edition 56 (36) (2017) 10815–10820, publisher: Wiley Online Library.
[105] S. Masubuchi, M. Morimoto, S. Morikawa, M. Onodera, Y. Asakawa, K. Watanabe, T. Taniguchi, T. Machida, Autonomous robotic searching and assembly of two-dimensional crystals to build van der waals superlattices, Nature communications 9 (1) (2018) 1–12, publisher: Nature Publishing Group.
[106] R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Han, K. G. Reyes, A. Amassian, M. Abolhasani, Artificial chemist: An autonomous quantum dot synthesis bot, Advanced Materials (2020) 2001626Publisher: Wiley Online Library.
[107] Z. Li, M. A. Najeeb, L. Alves, A. Z. Sherman, V. Shekar, P. Cruz Parrilla, I. M. Pendleton, W. Wang, P. W. Nega, M. Zeller, Robot-accelerated perovskite investigation and discovery, Chemistry of MaterialsPublisher: ACS Publications (2020).
[108] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. Yunker, J. E. Hein, A. Aspuru-Guzik, ChemOS: An orchestration software to democratize autonomous discovery, PLoS One 15 (4) (2020) e0229862, publisher: Public Library of Science San Francisco, CA USA.
[109] A. F. Voter, F. Montalenti, T. C. Germann, Extending the time scale in atomistic simulation of materials, Annual Review of Materials Research 32 (1) (2002) 321–346, publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.
[110] T. Frauenheim, G. Seifert, M. Elstner, T. Niehaus, C. Köhler, M. Amkreutz, M. Sternberg, Z. Hajnal, A. Di Carlo, S. Suhai, Atomistic simulations of complex materials: ground-state and excited-state properties, Journal of Physics: Condensed Matter 14 (11) (2002) 3015, publisher: IOP Publishing.
[111] V. Brázdová, D. R. Bowler, Atomistic computer simulations: a practical guide, John Wiley & Sons, 2013.
[112] D. C. Rapaport, The art of molecular dynamics simulation, Cambridge university press, 2004.
[113] D. Marx, J. Hutter, Ab initio molecular dynamics: basic theory and advanced methods, Cambridge University Press, 2009.
[114] F. H. Stillinger, T. A. Weber, Computer simulation of local order in condensed phases of silicon, Physical review B 31 (8) (1985) 5262, publisher: APS.
[115] M. S. Daw, M. I. Baskes, Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals, Physical Review B 29 (12) (1984) 6443, publisher: APS.
[116] E. B. Tadmor, R. E. Miller, Modeling materials: continuum, atomistic and multiscale techniques, Cambridge University Press, 2011.
[117] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Physical review letters 98 (14) (2007) 146401, publisher: APS.
[118] J. Behler, R. Martoňák, D. Donadio, M. Parrinello, Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential, Physical review letters 100 (18) (2008) 185501, publisher: APS.
[119] J. Behler, Representing potential energy surfaces by high-dimensional neural network potentials, Journal of Physics: Condensed Matter 26 (18) (2014) 183001, publisher: IOP Publishing.
[120] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Physical review letters 108 (5) (2012) 058301, publisher: APS.
[121] M. Rupp, Machine learning for quantum mechanics in a nutshell, International Journal of Quantum Chemistry 115 (16) (2015) 1058–1073, publisher: Wiley Online Library.
[122] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Physical review letters 104 (13) (2010) 136403, publisher: APS.
[123] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Physical Review B 87 (18) (2013) 184115, publisher: APS.
[124] W. J. Szlachta, A. P. Bartók, G. Csányi, Accuracy and transferability of gaussian approximation potential models for tungsten, Physical Review B 90 (10) (2014) 104108, publisher: APS.
[125] A. P. Bartók, G. Csányi, G aussian approximation potentials: A brief tutorial introduction, International Journal of Quantum Chemistry 115 (16) (2015) 1051–1057, publisher: Wiley Online Library.
[126] V. Botu, R. Batra, J. Chapman, R. Ramprasad, Machine learning force fields: construction, validation, and outlook, The Journal of Physical Chemistry C 121 (1) (2017) 511–522, publisher: ACS Publications.
[127] V. Botu, J. Chapman, R. Ramprasad, A study of adatom ripening on an al (1 1 1) surface with machine learning force fields, Computational Materials Science 129 (2017) 332–335, publisher: Elsevier.
[128] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, A universal strategy for the creation of machine learning-based atomistic force fields, NPJ Computational Materials 3 (1) (2017) 1–8, publisher: Nature Publishing Group.
[129] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Modeling & Simulation 14 (3) (2016) 1153–1173, publisher: SIAM.
[130] S. Jindal, S. Chiriki, S. S. Bulusu, Spherical harmonics based descriptor for neural network potentials: Structure and dynamics of au147 nanocluster, The Journal of Chemical Physics 146 (20) (2017) 204301, publisher: AIP Publishing LLC.
[131] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, Journal of Computational Physics 285 (2015) 316–330, publisher: Elsevier.
[132] J. S. Smith, O. Isayev, A. E. Roitberg, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, Chemical science 8 (4) (2017) 3192–3203, publisher: Royal Society of Chemistry.
[133] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, Nature communications 10 (1) (2019) 1–8, publisher: Nature Publishing Group.
[134] E. V. Podryabinkin, A. V. Shapeev, Active learning of linearly

parametrized interatomic potentials, Computational Materials Science 140 (2017) 171–180, publisher: Elsevier.

[135] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, Less is more: Sampling chemical space with active learning, The Journal of chemical physics 148 (24) (2018) 241733, publisher: AIP Publishing LLC.

[136] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, Finding density functionals with machine learning, Physical review letters 108 (25) (2012) 253002, publisher: APS.

[137] K. Yao, J. Parkhill, Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks, Journal of chemical theory and computation 12 (3) (2016) 1139–1147, publisher: ACS Publications.

[138] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, Bypassing the kohn-sham equations with machine learning, Nature communications 8 (1) (2017) 1–10, publisher: Nature Publishing Group.

[139] R. Nagai, R. Akashi, O. Sugino, Completing density functional theory by machine learning hidden messages from molecules, npj Computational Materials 6 (1) (2020) 1–8, publisher: Nature Publishing Group.

[140] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, K. Burke, Quantum chemical accuracy from density functional approximations via machine learning, Nature Communications 11 (1) (2020) 1–11, publisher: Nature Publishing Group.

[141] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, R. Ramprasad, Solving the electronic structure problem with machine learning, npj Computational Materials 5 (1) (2019) 1–7, publisher: Nature Publishing Group.

[142] A. V. Crewe, Scanning electron microscopes: is high resolution possible?, Science 154 (3750) (1966) 729–738, publisher: American Association for the Advancement of Science.

[143] S. J. Pennycook, P. D. Nellist, Scanning transmission electron microscopy: imaging and analysis, Springer Science & Business Media, 2011.

[144] G. Binnig, H. Rohrer, C. Gerber, E. Weibel, 7$\times$ 7 reconstruction on si (111) resolved in real space, Physical review letters 50 (2) (1983) 120, publisher: APS.

[145] G. Binnig, H. Rohrer, C. Gerber, E. Weibel, Surface studies by scanning tunneling microscopy, Physical review letters 49 (1) (1982) 57, publisher: APS.

[146] C. Gerber, H. P. Lang, How the doors to the nanoworld were opened, Nature nanotechnology 1 (1) (2006) 3–5, publisher: Nature Publishing Group.

[147] S. V. Kalinin, B. G. Sumpter, R. K. Archibald, Big–deep–smart data in imaging for guiding materials design, Nature materials 14 (10) (2015) 973–980, publisher: Nature Publishing Group.

[148] J. L. Lansford, D. G. Vlachos, Infrared spectroscopy data-and physics-driven machine learning for characterizing surface microstructure of complex materials, Nature communications 11 (1) (2020) 1–12, publisher: Nature Publishing Group.

[149] M. J. Cherukara, Y. S. Nashed, R. J. Harder, Real-time coherent diffraction inversion using deep generative networks, Scientific reports 8 (1) (2018) 1–8, publisher: Nature Publishing Group.

[150] Y.-F. Shen, R. Pokharel, T. J. Nizolek, A. Kumar, T. Lookman, Convolutional neural network-based method for real-time orientation indexing of measured electron backscatter diffraction patterns, Acta Materialia 170 (2019) 118–131, publisher: Elsevier.

[151] R. A. Schwarzer, D. P. Field, B. L. Adams, M. Kumar, A. J. Schwartz, Present state of electron backscatter diffraction and prospective developments, in: Electron backscatter diffraction in materials science, Springer, 2009, pp. 1–20.

[152] S. I. Wright, M. M. Nowell, A review of in situ EBSD studies, in: Electron Backscatter Diffraction in Materials Science, Springer, 2009, pp. 329–337.

[153] T. B. Britton, J. Jiang, Y. Guo, A. Vilalta-Clemente, D. Wallis, L. N. Hansen, A. Winkelmann, A. J. Wilkinson, Tutorial: Crystal orientations and EBSD—or which way is up?, Materials Characterization 117 (2016) 113–126, publisher: Elsevier.

[154] R. Liu, A. Agrawal, W.-k. Liao, A. Choudhary, M. De Graef, Materials discovery: Understanding polycrystals from large-scale electron patterns, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 2261–2269.

[155] D. Jha, S. Singh, R. Al-Bahrani, W.-k. Liao, A. Choudhary, M. De Graef, A. Agrawal, Extracting grain orientations from ebsd patterns of polycrystalline materials using convolutional neural networks, Microscopy and Microanalysis 24 (5) (2018) 497–502, publisher: Cambridge University Press.

[156] M. R. Carbone, S. Yoo, M. Topsakal, D. Lu, Classification of local chemical environments from x-ray absorption spectra using supervised machine learning, Physical Review Materials 3 (3) (2019) 033604, publisher: APS.

[157] X. Lin, Z. Si, W. Fu, J. Yang, S. Guo, Y. Cao, J. Zhang, X. Wang, P. Liu, K. Jiang, Intelligent identification of two-dimensional nanostructures by machine-learning optical microscopy, Nano Research 11 (12) (2018) 6316–6324, publisher: Springer.

[158] C. C. Mody, Instrumental community: Probe microscopy and the path to nanotechnology, MIT Press, 2011.

[159] A. Cui, K. Jiang, M. Jiang, L. Shang, L. Zhu, Z. Hu, G. Xu, J. Chu, Decoding phases of matter by machine-learning raman spectroscopy, Physical Review Applied 12 (5) (2019) 054049, publisher: APS.

[160] A. Fakhry, T. Zeng, S. Ji, Residual deconvolutional networks for brain electron microscopy image segmentation, IEEE transactions on medical imaging 36 (2) (2016) 447–456, publisher: IEEE.

[161] T. M. Quan, D. G. Hildebrand, W.-K. Jeong, Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics, arXiv preprint arXiv:1612.05360 (2016).

[162] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, M. S. Rosen, Image reconstruction by domain-transform manifold learning, Nature 555 (7697) (2018) 487–492, publisher: Nature Publishing Group.

[163] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[164] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[165] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[166] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (7763) (2019) 95–98, publisher: Nature Publishing Group.

[167] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Materials synthesis insights from scientific literature via text extraction and machine learning, Chemistry of Materials 29 (21) (2017) 9436–9444, publisher: ACS Publications.

[168] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216, publisher: IEEE.

[169] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Physical Review Materials 2 (8) (2018) 083802, publisher: APS.

[170] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T. Y.-J. Han, Reliable and explainable machine-learning methods for accelerated material discovery, npj Computational Materials 5 (1) (2019) 1–9, publisher: Nature Publishing Group.

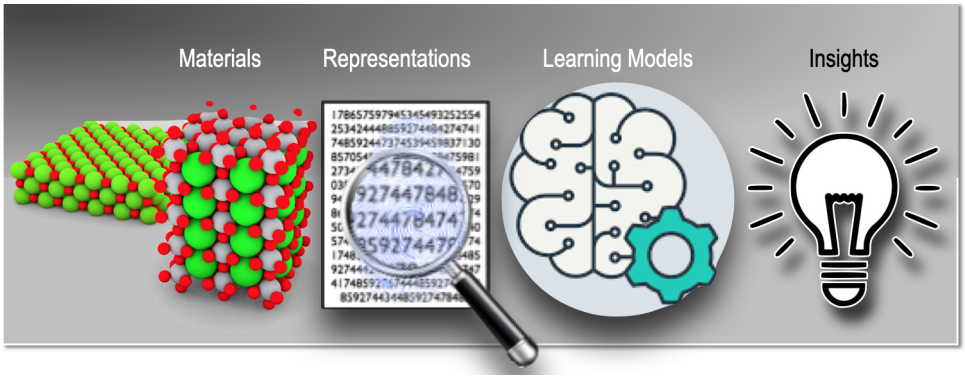
[171] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, M. Scheffler, Identifying domains of applicability of machine learning models for materials science, Nature communications 11 (1) (2020) 1–9, publisher: Nature Publishing Group.

[172] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[173] M. Haghighatlari, C.-Y. Shih, J. Hachmann, Thinking globally, acting locally: On the issue of training set imbalance and the case for local machine learning models in chemistry, preprint at ChemRxiv: https://doi.org/10.26434/chemrxiv 8796947 (2019) v2.

[174] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, L. M. Ghiringhelli, Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO, Journal of Physics: Materials 2 (2) (2019) 024002, publisher: IOP Publishing.

[175] S. R. Xie, G. R. Stewart, J. J. Hamlin, P. J. Hirschfeld, R. G. Hennig, Functional form of the superconducting critical temperature from machine learning, Physical Review B 100 (17) (2019) 174513, publisher: APS.
[176] G. Cao, R. Ouyang, L. M. Ghiringhelli, M. Scheffler, H. Liu, C. Carbogno, Z. Zhang, Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites, Physical Review Materials 4 (3) (2020) 034204, publisher: APS.
[177] S.-M. Udrescu, M. Tegmark, AI feynman: A physics-inspired method for symbolic regression, Science Advances 6 (16) (2020) eaay2631, publisher: American Association for the Advancement of Science.
[178] R. P. Feynman, R. B. Leighton, M. Sands, The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat, Vol. 1, Basic books, 2011.
[179] R. P. Feynman, R. B. Leighton, M. L. Sands, The Feynman Lectures on Physics vol 2 (Redwood City, CA, CA: Addison-Wesley, 1989.
[180] R. P. Feynman, R. B. Leighton, M. Sands, The Feynman Lectures on Physics, Volume III: Quantum Mechanics, vol. 3, Basic Books, 2010.
[181] O. Muller, R. Roy, The Major Ternary Structural Families, Springer Verlag (Berlin, Heidelberg, New York), 1974.
[182] R. D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta crystallographica section A: crystal physics, diffraction, theoretical and general crystallography 32 (5) (1976) 751–767, publisher: International Union of Crystallography.
[183] B. M. Greenwell, pdp: An r package for constructing partial dependence plots., R J. 9 (1) (2017) 421.
[184] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, in: Explainable AI: interpreting, explaining and visualizing deep learning, Springer, 2019, pp. 5–22.
[185] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical XAI, arXiv preprint arXiv:1907.07374 (2019).
[186] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832, publisher: Multidisciplinary Digital Publishing Institute.
[187] N. Xie, G. Ras, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, arXiv preprint arXiv:2004.14545 (2020).
[188] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (7601) (2016) 73–76, publisher: Nature Publishing Group.
[189] A. I. Forrester, A. Sóbester, A. J. Keane, Multi-fidelity optimization via surrogate modelling, Proceedings of the royal society a: mathematical, physical and engineering sciences 463 (2088) (2007) 3251–3269, publisher: The Royal Society London.
[190] Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science, Npj Computational Materials 4 (1) (2018) 1–8, publisher: Nature Publishing Group.
[191] G. Pilania, K. J. McClellan, C. R. Stanek, B. P. Uberuaga, Physics-informed machine learning for inorganic scintillator discovery, The Journal of chemical physics 148 (24) (2018) 241729, publisher: AIP Publishing LLC.
[192] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, M. Scheffler, Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats, NPJ Computational Materials 3 (1) (2017) 1–9, publisher: Nature Publishing Group.
[193] C. Draxl, M. Scheffler, Big data-driven materials science and its FAIR data infrastructure, Handbook of Materials Modeling: Methods: Theory and Modeling (2020) 49–73Publisher: Springer.
[194] C. Draxl, M. Scheffler, NOMAD: The FAIR concept for big data-driven materials science, Mrs Bulletin 43 (9) (2018) 676–682, publisher: Cambridge University Press.
[195] R. Chard, Z. Li, K. Chard, L. Ward, Y. Babuji, A. Woodard, S. Tuecke, B. Blaiszik, M. Franklin, I. Foster, DLHub: Model and data serving for science, in: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE, 2019, pp. 283–292.


Materials
Representations
Learning Models
Insights