ANNUAL REVIEWS

Annual Review of Materials Research

Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches

James E. Saal,[1] Anton O. Oliynyk,[2] and Bryce Meredig[1]

[1]Citrine Informatics, Redwood City, California 94063, USA; email: bryce@citrine.io

[2]Department of Chemistry and Biochemistry, Manhattan College, Riverdale, New York 10471, USA

Annu. Rev. Mater. Res. 2020. 50:49–69

First published as a Review in Advance on May 18, 2020

The *Annual Review of Materials Research* is online at matsci.annualreviews.org

https://doi.org/10.1146/annurev-matsci-090319-010954

Keywords

machine learning, materials discovery, materials informatics

Abstract

The rapidly growing interest in machine learning (ML) for materials discovery has resulted in a large body of published work. However, only a small fraction of these publications includes confirmation of ML predictions, either via experiment or via physics-based simulations. In this review, we first identify the core components common to materials informatics discovery pipelines, such as training data, choice of ML algorithm, and measurement of model performance. Then we discuss some prominent examples of validated ML-driven materials discovery across a wide variety of materials classes, with special attention to methodological considerations and advances. Across these case studies, we identify several common themes, such as the use of domain knowledge to inform ML models.

1. INTRODUCTION

Over the past decade, machine learning (ML) has emerged as a powerful tool to accelerate materials development (1–5). Academic, government, and commercial entities are broadly deploying ML in service of materials discovery. Publication activity in ML for materials is growing exponentially even as a fraction of all materials research (**Figure 1**). Despite an increasing body of literature on data-driven materials research, which we refer to as materials informatics, only a fraction of published studies culminate in predictions that are subsequently validated by an experiment, either in the laboratory or as a "virtual" experiment via physics-based simulation. A trained ML model is merely a means to an end, and the utility of materials informatics is fully realized only when ML predictions are confirmed. In this review, we (*a*) describe the key components of a materials informatics discovery pipeline; (*b*) highlight recent works that describe validation of materials informatics predictions, as summarized in **Table 1**; and (*c*) note some materials discovery–specific considerations for ML. We begin by describing a typical materials informatics pipeline in more detail.

2. THE MATERIALS INFORMATICS DISCOVERY PIPELINE

In this section, we discuss critical components of a materials informatics pipeline common to validated ML studies. These standard steps are summarized in the generalized pipeline of **Figure 2**. The pipeline begins with establishing a materials data set for training, as well as a set of materials descriptors to extend the data with available physical information. This data set is then used to train an ML model, which is used to make a prediction of novel materials for validation.

2.1. Training Data

Data of sufficient quality and quantity are an essential prerequisite for the successful application of ML methods to materials problems. Large companies in the technology industry, such as Google,

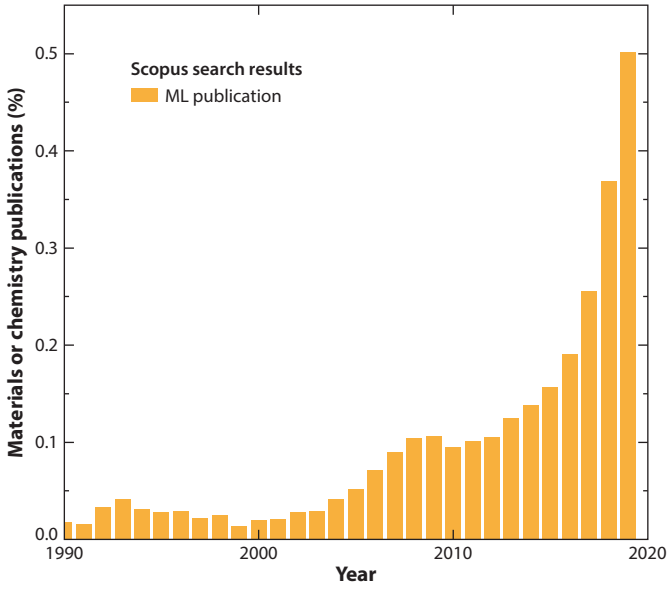

Figure 1

Share of materials and chemistry publications referencing machine learning (ML) as a function of time. The data are normalized to account for the overall exponential growth (6) in scientific publications over time, illustrating the relative growth of ML-related work.

Table 1 Summary of validated machine learning (ML) predictions

Reference	Materials class	Application	Predicted properties	Initial training data set	ML algorithm	Design space	Number of candidates evaluated	Year[a]
70	Inorganic ternary solids ($A_xB_yC_z$)	Stable composition prediction	Formation energy	15,000 density functional theory (DFT) calculations	Random forest	1.6 million enumerated compositions	Nine DFT calculations	2013
64	Molecules	Organic light-emitting diodes	Several underlying properties contributing to external quantum efficiency (EQE)	DFT calculations on 40,000 randomly selected candidates	Neural network	1.6 million molecules constructed from fragments	Four experiments	2015
90	Polymers	Dielectrics	Electronic dielectric constant, ionic dielectric constant, and band gap	Newly generated DFT data (284 records)	Kernel ridge regression and genetic algorithm	~156,000 enumerated four-, six-, and eight-block polymers	28 DFT calculations	2015
78	Organically templated metal oxides	Hydrothermal synthesis	Reaction success	Labeled reactions from internal data set (3,955 records)	Support vector machine	1,680 commercially available diamines	34 experiments	2015
91	Inorganic binary solids (AB)	Structure discovery	Stability	Pearson's Crystal Data (PCD) database (92) (706 records)	Support vector machine	2,926 possible binary combinations	One experiment	2016
65	Inorganic ternary solids (AB_2C Heusler)	Structure discovery	Stability	PCD (92) (1,948 records)	Random forest	>400,000 AB_2C compositions	21 experiments	2016
68	Metal alloys	Shape memory alloys	Endothermic peak temperature	Newly synthesized alloys (53 records)	Polynomial regression	1,652,470 compositions	One experiment	2016
19	Perovskites	Ferroelectric	Ferroelectric Curie temperature and perovskite stability	Literature review (167 stability and 117 Curie temperature records)	Support vector classifier and linear regression	61,506 enumerated compositions	10 experiments	2017
93	Inorganic ternary solids (ABC)	Structure discovery	Polymorphism	PCD (92) (1,556 records)	Support vector machine	98,769 ABC compositions	One experiment	2017
7	Bulk metallic glasses	Structural applications	Glass formability	Landolt–Börnstein (79) (6,780 records)	Random forest	Enumerated grids in 2,024 ternary systems	Four experiments	2017
94	Ternary ionic solids ($A_aB_bX_x$)	Structure discovery	Stability	Inorganic Crystal Structure Database (ICSD) (95)	Tucker decomposition recommender system	7,405,200 enumerated compositions	27 DFT calculations	2017
77	Metal–organic frameworks (MOFs)	Hydrogen storage	Hydrogen-deliverable capacity	Grand canonical Monte Carlo simulations (1,000 records)	Least absolute shrinkage and selection operator (LASSO)	54,776 MOF structures from Cambridge Structural Database (CSD) (9)	One experiment	2018

(*Continued*)

Table 1 (*Continued*)

Reference	Materials class	Application	Predicted properties	Initial training data set	ML algorithm	Design space	Number of candidates evaluated	Year[a]
63	Small molecules	Organic light-emitting diodes	Max light-absorbing wavelengths; triplet (T_1) energy levels	Random sample from PubChem (12) (50,000 records)	Deep neural network	40,000 randomly generated simplified molecular-input line-entry system (SMILES) strings	Three experiments	2018
96	Hybrid organic–inorganic perovskites (ABX_3)	Photovoltaics	Band gap	Literature review (212 records)	Gradient boosting regression	5,158 enumerated compositions	Six DFT calculations	2018
31	Inorganic solids	Superhard materials	Bulk modulus and shear modulus	Materials Project (28) (3,248 records)	Support vector machine	118,287 compounds from PCD (92)	Two experiments	2018
75	Polymers	Cement plasticizer	Slump	Commercial data (seven records)	LASSO regression	Maximization of analytical expression for slump	One experiment	2018
20	Ni-rich cathode materials ($LiNi_xCo_{1-x-y}Mn_{1-x-y-z}O_2$)	Batteries	Initial capacity, cycle life, and amount of residual Li	Literature review (330 records)	Extremely randomized tree and adaptive boosting	50,000 randomly generated candidate syntheses	Five experiments	2018
17	High-entropy alloys	Structural applications	Hardness	Literature review (82 records)	Canonical correlation analysis and genetic algorithm	4.6 billion enumerated compositions	Seven experiments	2018
80	Layered semiconductor metamaterials	Thermal radiator	Figure of merit from emissivity spectra	Iterative Bayesian optimization on 42,000 groups of structures	Gaussian process regression	>8 billion candidates	Three experiments	2018
97	Polymers	Thermoplastics	Thermal conductivity (λ), several other properties	PoLyInfo (98), QM9 (99) (38,310 total properties; only 28 λ values)	Neural networks	Monte Carlo generative model (iqspr)	Three experiments	2018
32	Inorganic phosphor host materials	Solid-state lighting	Debye temperature and band gap	Materials Project (28) (2,610 records)	Support vector machine	>300,000 compounds from PCD (92)	One experiment	2018
82	Li-containing inorganic solids	Li-ion battery electrolytes	Ionic conductivity	Literature review and ICSD (95) (40 records)	Logistic regression	12,831 compounds from Materials Project (28)	Four DFT calculations	2019
22	High-entropy alloys ($Al_xCo_yCr_z$-$Cu_uFe_vNi_w$)	Structural applications	Hardness	Literature review and new experimental data (155 records)	Support vector machine	1,895,147 compositions	42 experiments	2019

[a]Year the work was submitted for publication.

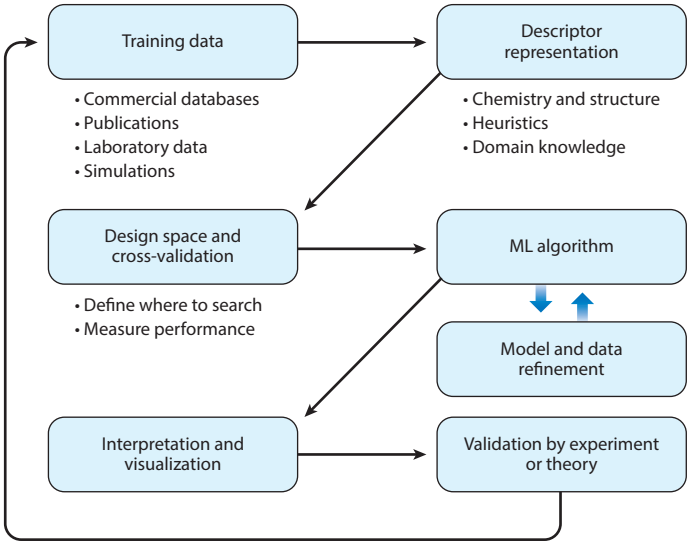

Figure 2

Schematic of a materials informatics discovery pipeline. The pipeline begins with input property data, which require a descriptor (i.e., vector) representation suitable for machine learning (ML). Model training and refinement comprise an iterative process, resulting in a model with known predictive performance that may be used to evaluate candidate materials. Data gathered from experiments or simulations should be used to further improve models, as described by Ren et al. (7).

Facebook, and Amazon, are able to train ML models on up to 10^{11} examples (8). However, data are far more scarce in materials science, and data sets for materials discovery typically include 10^{1}–10^{4} training examples, as seen in **Table 1**.

Training data quality is associated with reliability and consistency. Reliability issues can result from unreported sources of variance (e.g., poorly calibrated or aged instruments) or human factors (e.g., typos that arise during manual data entry). Data consistency is related to the method that was used to generate the data. Inconsistency can be experimental (for example, thermal conductivity can be measured with different techniques, giving slightly different results) or computational (different simulation input choices can give different results).

2.1.1. Challenges in training data aggregation. The source of training data for each materials problem must be considered on a case-by-case basis due to the lack of a centralized, homogeneous database of all materials measurements. Such authoritative databases are more common in the biological and pharmaceutical sciences; examples include the Cambridge Crystal Structure Database (9), ChemBank (10), GenBank (11), and PubChem (12). The materials science community has fewer data resources, although the Materials Data Facility (13), the Novel Materials Discovery (NOMAD) Repository (14), and Open Citrination (15) have emerged to provide free-of-charge materials data services to the research community. However, for most materials problems, relevant experimental data are scattered across publications in the literature (16), requiring researchers to manually extract and structure training data sets (17–22). While manually constructed data sets are highly time and resource intensive to build, they benefit from expert curation, which can be essential to providing the context needed for successful ML models.

2.1.2. Density functional theory as a source of training data. Among successful materials informatics use cases, density functional theory (DFT) (23, 24) is a widely used source of training data and descriptors. DFT, as a means of generating large-scale materials data, has become more accessible over time (25, 26) as computational power has increased and the software has improved. The construction of high-throughput DFT (HT-DFT) databases (27–29) has been a key contributor to many early informatics successes (30). Advancing beyond relatively simple calculations of thermodynamic stability, more computationally costly HT-DFT data sets are enabling data-driven solutions to materials problems, such as DFT-calculated elastic tensors to predict superhard materials (31) and discovery of novel phosphors for solid-state lighting (32).

HT-DFT databases offer an attractive supplement to experimental databases for use in informatics as they can be resource efficient to build, comprehensive across chemistries, well structured, and internally consistent. For instance, identifying the DFT-calculable Debye temperature as a reliable indicator for the complex materials property of photoluminescent quantum yield enabled the use of an HT-DFT elastic tensor data set to predict novel light-emitting diode (LED) materials (32). DFT, when coupled with active learning, can also be a rapid engine for exploring a design space and optimizing properties (33).

2.2. Descriptor Representation

Any successful application of ML to materials discovery relies on a suitable choice of representation (34). Representation refers to how a material is encoded in a machine-readable format, typically as a fixed-length vector of descriptors (also referred to as features or input variables). For instance, Ni_3Al could be represented as simply the combination of character strings in the composition, the atomic coordinates of Ni and Al in the $L1_2$ crystal structure, or a scanning electron micrograph of the microstructure. Representation is an opportunity to inject known physical information into an ML prediction problem. For example, that Ni crystallizes in the face-centered cubic structure and has a melting point of 1,455°C can be directly exploited in differentiating Ni from other elements when training ML models. The choice of representation can have a large effect on ML model performance, as observed by Askerka et al. (35) for the thermodynamic stability of double perovskites (**Figure 3**).

In the majority of case studies in **Table 1**, researchers used simple empirical descriptors associated with the elemental compositions of materials under investigation. The `magpie` (40) and `matminer` (37) packages are widely used sources of these chemical features, which include physical concepts such as atomic size, electronegativity, and electron configuration. The selection of a high-quality representation is likely to be important for success in using ML for materials discovery.

Across many materials problems, the optimal choice of representation is nonobvious. For example, recent research (41, 42) has developed representations for the concept of a grain boundary, which may be intuitively clear to a human scientist but is not straightforwardly captured as a vector or matrix object suitable for linear algebra operations. Likewise, new atomic-scale representations for small molecules (43, 44), periodic systems (e.g., crystal structures) (39, 45), and combinations thereof (34, 46) are being actively developed.

Including as many descriptors as possible is one way to ensure that no known physics is left out of a modeling exercise. On the other hand, many ML algorithms are susceptible to degraded performance in the presence of correlated and/or meaningless descriptors (47). The processes of dimensionality reduction [e.g., principal component analysis (48)] and feature selection aim to identify a reduced set of maximally informative and ideally uncorrelated descriptors for input to an ML model. For example, a metric called cluster resolution (49), which uses the relative positions

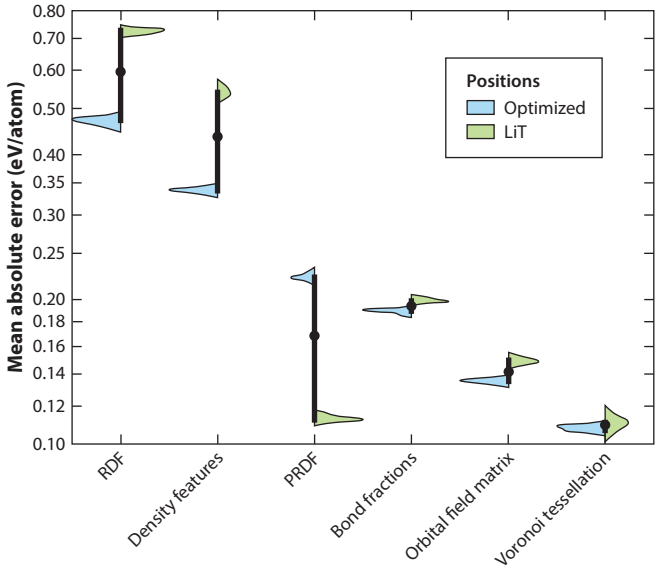

Figure 3

Askerka et al. (35) investigated the representation dependence of machine learning model predictions for double perovskite thermodynamic stability. Their learning-in-templates (LiT) approach assumes nominal atomic positions, which they compared to using optimized atomic positions. They benchmarked several crystal structure representations from the literature: radial distribution functions (RDF) and partial radial distribution functions (PRDF) (36), density features and bond fractions (37), the orbital field matrix (38), and Voronoi tessellations (39). Figure adapted with permission from Reference 35. Copyright 2019, American Chemical Society.

and geometries of clusters present in data sets to quantify the preservation of differences between various types of materials, was used to perform feature selection in several case studies listed in **Table 1**.

2.3. Design Space and Cross-Validation

Cross-validation (CV), which quantifies the performance of ML models, can be customized to a desired design space in order to obtain realistic estimates of model performance for the discovery application at hand. The concept of design space refers to the chemistries, experimental conditions, and processing routes one is interested in searching for new materials. A critical question is the relationship between the training data and the design space; generally speaking, the more different these two regimes are, the more challenging ML-driven discovery will be.

2.3.1. Measuring model performance via cross-validation. CV is the gold standard approach for quantifying the performance of ML models. In CV, models are trained on a subset of all available data and then used to predict values (for regression-type problems) or labels (for classification-type problems) for a held-out set of data for which the ground truth is known. Conceptually, CV is intended to probe an ML model's ability to generalize to unseen examples and embodies the idea that models should not be evaluated on data to which they were fit.

The most widely used form of CV is random k-fold, in which the available training data are randomly divided into k partitions (i.e., folds). Over k iterations, the ML model is trained on

$k - 1$ partitions and used to predict the held-out partition. Performance metrics such as root mean square error (RMSE) or Pearson r^2 may be computed across all held-out partitions. Among the case studies in **Table 1**, random k-fold CV is almost universally employed to compute these model accuracy statistics.

Interestingly, while random k-fold CV is widely used in both the ML research and materials informatics communities, materials discovery may pose challenges for traditional CV methods. In particular, because (*a*) materials data sets typically exhibit strong clustering and (*b*) materials discovery may involve extrapolation, random k-fold CV has a tendency to overestimate the performance of ML models (50). CV techniques such as leave-one-cluster-out (LOCO) (50) and leave-out-group (LOG) (51) have emerged to simulate materials discovery use cases. Drug discovery researchers have made similar observations (52, 53), and appropriately matching CV techniques to the problem at hand remains an essential step in scientific applications of ML (54).

2.3.2. Extrapolative design spaces. Supervised ML approaches generally assume that all training data, along with unseen examples we wish to predict, are independent of one another and drawn from the same underlying distribution; this is the so-called independent and identically distributed (i.i.d.) assumption (55). However, real-world materials discovery applications often strain the i.i.d. conditions. As mentioned in the preceding section, experiments may be clustered due to sampling bias. In particular, scientists tend to measure many small changes to a few successful parent materials and to heavily underreport failed materials. Further, we may be interested in designing materials with structures, chemistries, or properties very different from those reflected in our training data. We refer to these latter situations as extrapolation.

Based on this definition, extrapolative materials discovery may or may not actually violate the i.i.d. assumption. The key question is whether the physics associated with as-yet-undiscovered materials of interest is satisfactorily sampled in the training data. To take superconductors as an example, a training set of Bardeen–Cooper–Schrieffer materials does not at all sample the anomalous physics of the cuprates, and the i.i.d. assumption is catastrophically violated. Thus, as we would intuitively expect, an ML model trained on Bardeen–Cooper–Schrieffer superconductors has no predictive power whatsoever for the cuprates (50, 56). In this vein, none of the case studies in **Table 1** describe the discovery of entirely new physical regimes with ML, although ML-optimized sampling of chemical space can accelerate such discoveries beyond the pace of a purely random search (57).

To illustrate an alternative scenario, imagine that we focus specifically on the discovery of materials with exceptional (i.e., extreme) values of certain properties. In this case, ML may be able to extrapolate beyond the property ranges present in the training data, if the necessary physics is present in the training data. However, this type of prediction problem is difficult. As shown in **Figure 4**, an ML model for bulk modulus (B) based on the Materials Project elastic property database (58) shows good performance over a range of B from 0 to 400 GPa when standard k-fold CV is applied. However, when an ML model is trained only with compounds with $B < 300$ GPa, the model's accuracy under extrapolation to greater values of B suffers.

2.4. Machine Learning Algorithms

The choice of algorithm underlying trained models in materials informatics is one that is largely driven by the specific materials problem being addressed, i.e., the nature of the model inputs and the desired outputs. As seen in **Table 1**, various algorithms have been successfully employed for various problems. The popular random forest (RF) (59, 60) and (deep) neural network (NN) (61, 62) algorithms are illustrated conceptually in **Figure 5**.

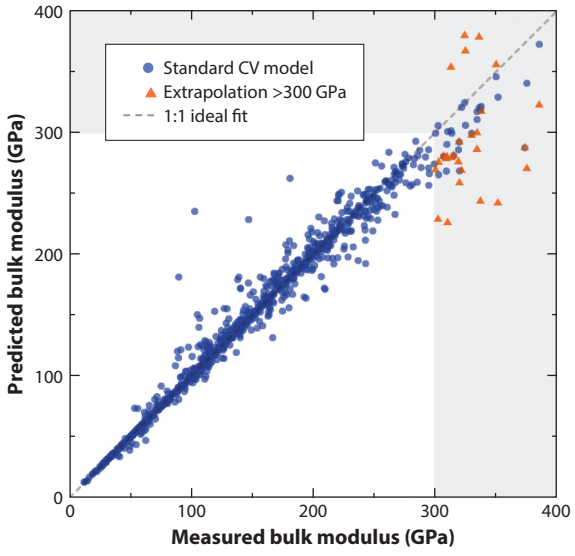

Figure 4

Example illustrating the effects of property extrapolation on ML model performance, using DFT-calculated mechanical property data from the Materials Project (31, 58). Blue circles show the results of typical CV across the entire data set, whereas orange triangles demonstrate the degradation in model performance that occurs when only materials with $B < 300$ GPa are used to train, and the remaining higher-B compounds are predicted. Abbreviations: CV, cross-validation; DFT, density functional theory; ML, machine learning.

A NN consists of an interconnected set of logical gates, called neurons, to transform a series of input data into an output decision. NN models lend themselves to pattern matching when training data sets are substantial and complex. In the informatics use cases reviewed here, such training data sets consist of data generated from high-throughput physics-based simulations (63, 64). The series of logical decisions made by the NN enable reproduction of complex abrupt changes in output

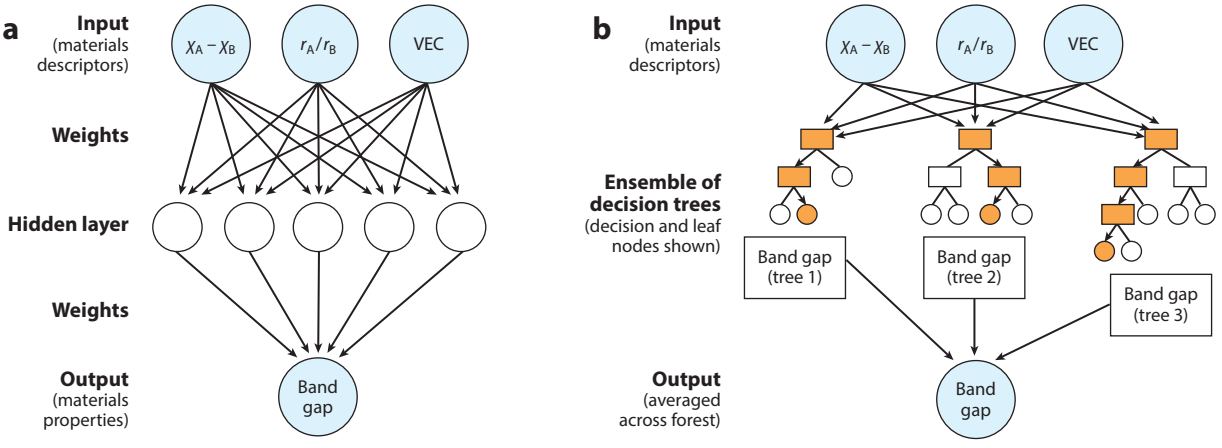

Figure 5

Schematic representation of (*a*) neural network and (*b*) random forest machine learning algorithms. In this hypothetical example, the models predict the band gap of an AB compound based on three features: the electronegativity difference between A and B ($\chi_A - \chi_B$), the atomic radius ratio between A and B (r_A/r_B), and the valence electron concentration (VEC) of the AB compound.

space given inputs, such as how a minor change in a simplified molecular-input line-entry system (SMILES) string can radically alter the rate constant (64).

The RF algorithm trains multiple decision tree models on randomized bootstrapped subsets of the training data. Model predictions are then made based on the collected predictions of the decision trees. This results in a model robust to overfitting, which works well for the smaller data set sizes common in materials science, such as the ~2,000 Heusler structures used to predict stability (65) and the ~6,800 compositions used to predict glass formability (7). Two of the main advantages of the RF algorithm are robustness to sparse data and ease of performing feature selection.

Comparison of performance between these and other algorithms is a subject of study in the material informatics literature (66) and is an important consideration when selecting one for a design problem. For instance, Gómez-Bombarelli et al. (64) found that the NN algorithm dramatically outperformed linear regression for prediction of molecular organic light-emitting diode (OLED) performance when the training data set size grew large, which is expected for molecular design spaces. However, for continuous composition-dependent design spaces (such as inorganic solids) and smaller training data sets, algorithm choice becomes less significant for model performance. Iwasaki et al. (67) and Wen et al. (22) observed little change in RMSE between NN and RF models for inorganic solid property prediction. Xue et al. (68) report similar prediction errors between various regression techniques, with polynomial regression being lower. However, in the context of discovery and design of experiments, RMSE and other simple error metrics are not necessarily the ideal performance criteria. Ultimately, trained models are a means to an end, and measuring the ability of a model to predict high-performance candidates is preferred. Gómez-Bombarelli et al. (64), for instance, quantify this as the fraction of molecules in the test set correctly ranked in the top 5% of all molecules.

3. CONFIRMED PREDICTIONS WITH METHODOLOGICAL HIGHLIGHTS

Table 1 presents a selection of 23 applications of ML to materials discovery wherein the predictions from ML were subsequently confirmed by experiment or simulation. In the following subsections, we highlight some of the key aspects of these case studies that are likely to have contributed to verifiable success.

3.1. Augmenting Domain Expertise with Machine Learning

After his defeat by IBM's Deep Blue in 1997, chess champion Garry Kasparov concluded that great complementarity exists between human and computer intelligence. In 1998, Kasparov created advanced chess, a format in which humans play with the aid of a chess computer (69). An analogous pairing of domain expertise and ML has led to several successes in materials design.

This concept of combining existing expert knowledge with ML to exploit complementarity can be powerful. While an unaided domain expert may simply miss an underlying trend in a large data set, ML can readily surface these patterns. Conversely, while an ML model is capable of generating massive numbers of candidate materials, a domain expert can sensibly prioritize materials for synthesis from the suggested list, taking into account, e.g., compatibility with laboratory equipment and economic factors that may be difficult to formally incorporate into ML models. Feedback from domain experts regarding predictive failures can also drive the subsequent refinement of ML models.

In an early example of the application of ML to materials discovery, Meredig et al. (70) demonstrated that models trained on DFT calculations could very accurately predict the thermodynamic

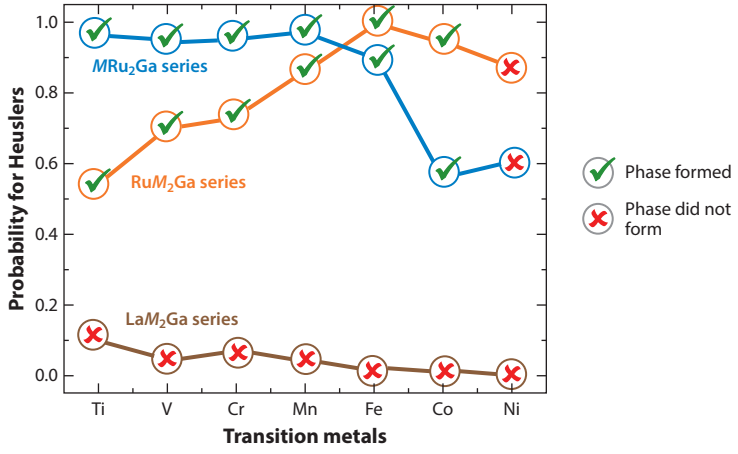

Figure 6

Summary of the machine learning (ML)–driven discovery results of Oliynyk et al. (65). Out of 14 ML-predicted high-probability Heusler compounds, 12 were experimentally confirmed; 7 predicted negatives were also verified in the laboratory. Adapted with permission from Reference 65. Copyright 2016, American Chemical Society.

stability of new compounds. Interestingly, they found that a model that combined pure ML with a domain knowledge–derived heuristic outperformed either ML or the heuristic individually. The ML model used only chemical composition as input; the heuristic used information from binary phase diagrams to predict the stability of ternary phases. With ML and the heuristic together, Meredig et al. identified nine materials as candidates for discovery, and confirmed using DFT calculations that eight of these compounds were more energetically stable than any combinations of known materials (70).

The work of Oliynyk et al. (65) offers another instance of extending domain knowledge with ML. A common question that arises when ML is applied to materials discovery is whether the predictions from ML are "nonobvious," and in this paper, Oliynyk et al. directly compare a pure electron-counting heuristic [akin to the 18-electron rule for half-Heuslers (71)] to the results of using electron counts *plus* many other descriptors in an ML model. Oliynyk et al. selected for synthesis several compounds whose likelihood of Heusler formation was predicted to be low by electron counting but high according to the ML model. The compounds $TiRu_2Ga$, VRu_2Ga, $CrRu_2Ga$, $RuTi_2Ga$, RuV_2Ga, $RuCr_2Ga$, and $RuMn_2Ga$ were experimentally confirmed to crystallize in the Heusler structure (65), in spite of very low predicted probabilities from the electron count–only approach. **Figure 6** summarizes the discovery results of Oliynyk et al. This example illustrates how heuristics often used to identify "usual suspect" materials may miss interesting candidates that a more elaborate ML approach can successfully identify.

3.2. Combining Machine Learning with Physics-Based Simulations

Widely used physics-based simulations, such as DFT and molecular dynamics, are natural complements to ML for three primary reasons. First, as ML (especially in materials) is often starved for precious training data, physical simulations can generate realistic training examples in abundance. Second, ML algorithms have no innate knowledge of physics, and concurrent use of simulations can guard against unphysical or pathological behavior in ML models. Third, ML in the context of transfer learning (i.e., using an ML model trained on a larger data set to assist in a related

prediction problem that has fewer training data available) can build predictive connections between simulation and laboratory experimentation (72).

Work by Mansouri Tehrani et al. (31) illustrates these points. In their effort to discover new earth-abundant superhard materials, these researchers trained ML models on approximately 2,500 DFT-calculated bulk (B) and shear (G) moduli (73) in the Materials Project, generated consistently in a high-throughput fashion. Over 100,000 candidates were then screened for very large values of these elastic moduli. The training data were limited mainly to binary compounds due to the computational cost of DFT mechanical property calculations on more complex unit cells. Despite this restriction, the employed ML model successfully extrapolated into ternary and quaternary systems. Mansouri Tehrani et al.'s approach was informed by the fact that DFT mechanical property data are much more abundant than experimental data and by the observation that Vickers hardness (H_V) values tend to correlate with B and G (74). This ML-driven screening process surfaced $Mo_{0.9}W_{1.1}BC$ and $Re_{0.5}W_{0.5}C$ as promising candidates, the highest predicted bulk modulus values of quaternary and ternary phases, respectively. Subsequent experiments confirmed that both materials exhibited superhard behavior (taken to be $H_V > 40$ GPa) at low load (31).

Besides the DFT-based approach exemplified by Mansouri Tehrani et al. (31), other physics-based or mechanistic models can be used as critical components in an informatics pipeline. Menon et al. (75) utilize a suite of physicochemical models to transform an initially simple molecular candidate space, based on mixtures of only seven commercial polymers, into a complex space of physical parameters which determine how these mixtures plasticize cement. In this way, domain expertise and prior knowledge are encoded in these physical models, enabling a predictive model [in this case, least absolute shrinkage and selection operator (LASSO) (76)] from a more limited data set. Another example by Bucior et al. (77) generates 3D H_2 adsorption energy profiles for known metal–organic framework (MOF) structures using grand canonical Monte Carlo and Lennard-Jones plus Coulombic potentials. These 3D profiles were transformed into 1D histograms of adsorption energy, with the histogram bin heights used as the MOF descriptors for LASSO training. Interestingly, ignoring spatial dimensions of adsorption behavior did not diminish the model's ability to predict MOF performance.

3.3. Active Learning and Iterative Discovery

As is often the case in materials informatics problems, one starts with a limited quantity of training data on which to build a model in search of new materials in a large design space. Consequently, the initial model will likely be inadequate to describe the entirety of the design space. In such instances, iterative active learning approaches are useful to efficiently sample the design space for additional training data to collect. Active learning involves the use of a model's prediction uncertainties to identify the ideal candidate experiments to achieve a goal, either to broaden the model's applicability in the design space and improve accuracy (exploration) or to identify the highest-performing candidates (exploitation)—or some combination thereof. Upon execution of these suggested experiments, the resulting data are used to retrain and enhance the underlying ML model, such that the candidates in the next iteration are likelier to represent improvements. Active learning, when applied to materials data sets, has been shown to significantly reduce the number of experiments needed to identify the highest-performing material when compared to random guessing (57).

Such an iterative approach was used by Ren et al. (7) to efficiently develop a model for bulk metallic glass (BMG) formability in ternary metal systems and identify novel BMG-forming systems. The authors started with a training data set of 6,780 experimental reports, covering 293 ternary systems. The authors noted that most of these data come from only 38 ternary systems

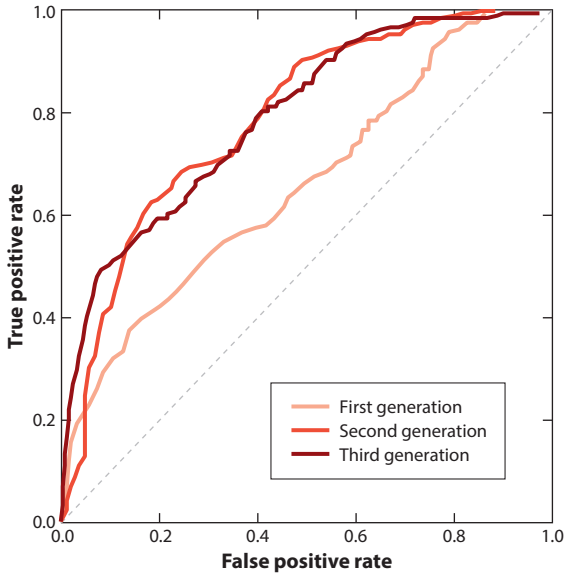

Figure 7

Receiver operating characteristic (ROC) curves illustrate a model's trade-off between true positives and false positives, where greater area under the ROC (AUROC) curve indicates superior performance. Ren et al. (7) iteratively improved their machine learning (ML) models of bulk metallic glass (BMG) formability. The first-generation model contained only BMG data derived from the Landolt–Börnstein handbook (79). The second and third generations introduced experimental data obtained in testing ML predictions. Adapted from Reference 7. Copyright The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. Distributed under a Creative Commons (CC BY-NC) license.

and are biased toward systems with known BMGs (i.e., a bias toward positive results). A classifier model for whether a ternary composition will form a BMG was trained on these data (first generation; **Figure 7**). The Co-V-Zr system was identified as a ternary with many novel BMG-forming compositions and experimentally validated by a combinatorial thin-film synthesis approach, which revealed a large glass region. These results were added to the training data, and the resulting model (second generation) has significantly superior performance to the first, as shown by the receiver operating characteristic curve in **Figure 7**. Importantly, the authors attribute this fact to the considerable number of negative training data added, improving the data set quality for learning. Raccuglia et al. (78) made a similar observation. For this reason, in the second round of model-selected experiments, a system predicted to *not* form BMGs (Fe-Ti-Nb) was synthesized alongside two that do (Co-Ti-Zr and Co-Fe-Zr), and the results agreed well with the second-generation model. Adding these new systems to the training data, a third-generation model exhibited further improvement over the second, particularly at low false-positive rate (i.e., true prediction of glass-forming systems).

3.4. Identifying Materials Exhibiting Property Extrema

Because ML is often used in pattern-matching applications, it is particularly challenging to surface materials with property values that lie outside the range of any initially known training materials. However, in an example of successful property value extrapolation, Xue et al. (68) found that a simple polynomial containing up to quadratic terms in three features (valence electron

count, atomic radius, and electronegativity) could successfully extrapolate to much higher values of shape-memory alloy transition temperatures T_p than were present in the training data. The training data generated by the researchers contained T_p values no greater than approximately 100°C, but the model accurately predicted values extracted from the published literature that ranged up to nearly 300°C. Further, an experiment on a material whose predicted T_p was 189.56°C was found experimentally to have a transformation temperature of 182.89°C.

The case studies in **Table 1** contain several additional examples of using ML to design materials with property values outside the range of the initially available training data. In rare cases, these discoveries can be accomplished in a single shot, as was done by Xue et al. (68). In many cases, an iterative active learning strategy must be employed. Rickman et al. (17) and Wen et al. (22) designed high-entropy alloys with measured hardness values greater than the alloys in their respective training sets. Sakurai et al. (80) used Bayesian optimization to develop a layered metamaterial thermal radiator with a Q-factor higher than previously observed. Thus, with appropriate training data, and often also with the aid of active learning, ML can enable discovery of materials with extreme property values.

3.5. Designing for Multiple Properties

Materials discovery applications often do not involve optimizing only a single property of interest. Rather, real-world problems are inherently multiobjective, which is more complex than single-variable optimization. Most works take one of two approaches to this challenge: (*a*) passing materials candidates through a number of property screens, with the hope that a tractable list of candidates satisfies all of the screening criteria, or (*b*) applying scalarizing functions to transform a multiobjective design into a surrogate single-objective problem. The first approach assumes one can satisfactorily evaluate all properties for a preenumerated list of candidate materials; the second approach involves iteratively evaluating individual candidates or batches, and retraining ML models at each iteration.

The Li-ion battery electrolyte screening work of Sendek et al. (81), which led to a case study (82) in **Table 1**, illustrates the screening strategy. Sendek et al. trained an ML model on experimental ionic conductivity values for Li-ion conductors and screened a DFT-derived database (28) of candidate materials using this ML model along with a number of other criteria, as shown in **Figure 8**. This multiproperty screening approach, which involved tabulated elemental properties, heuristics, and DFT calculations in addition to ML, reduced 12,831 candidate electrolytes to 12. Subsequent DFT calculations confirmed materials in the Li-B-S system to have extraordinarily high ionic conductivity values (82).

Scalarizing functions are a second approach for multiproperty materials design, where multiple property requirements are combined into a single function. This strategy maps a multiobjective problem to a single-objective problem, which can then be solved with any standard optimization approach. For example, Häse et al. (83) developed Chimera, which takes as user input a prioritized list of all materials properties under consideration (e.g., a user could specify that thermodynamic stability matters more than high ionic conductivity) and outputs a single-objective function suitable for optimization. In measuring the success of scalarizing functions, studies tend to focus on the Pareto optimality of surfaced candidates (83, 84). A candidate is Pareto optimal (or, equivalently, lies on the Pareto frontier) if there exists no other candidate that is superior across every property dimension. Thus, the candidates on the Pareto frontier each offer a unique set of trade-offs across properties of interest. For instance, two superconducting materials could be Pareto optimal if one offers a high critical temperature T_c but subpar ductility (important for making wires of the material), while the other possesses superior ductility but a lower T_c. As it is impossible,

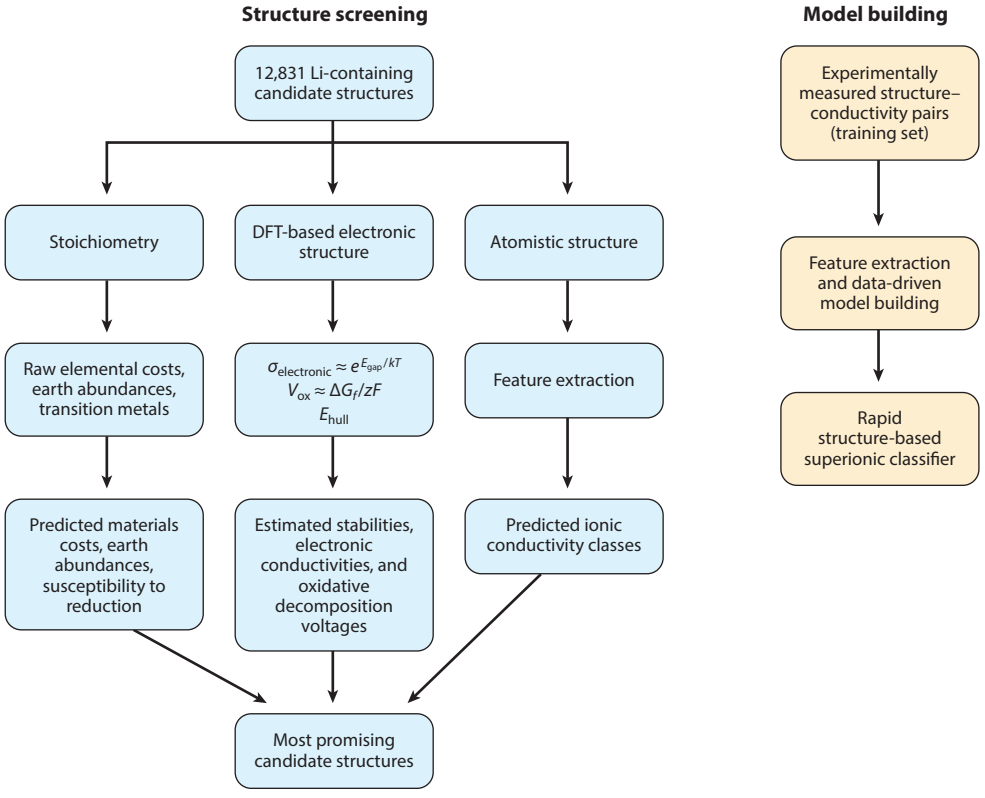

Figure 8

Multiproperty Li-ion battery electrolyte screening strategy of Sendek et al. (81). A combination of physics-based simulation, economic and practicality arguments, and machine learning modeling reduced a list of 12,831 candidate materials to 12. Abbreviation: DFT, density functional theory. Figure adapted with permission from the Royal Society of Chemistry, from "Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials," Sendek et al., *Energy Environ. Sci.* 10(1) 2017 (81); permission conveyed through Copyright Clearance Center, Inc.

without user input, to declare a winner among Pareto-optimal materials, the ability to efficiently surface as many materials on the Pareto frontier as possible is a sensible benchmark for scalarizing functions.

3.6. Generating Novel Materials

For some materials development challenges, particularly those involving organic small-molecule design, the design space is so massive that it cannot be exhaustively enumerated. In such cases, if novel molecules are desired, the design space will be populated by either (*a*) candidates generated de novo via ML without any prior structural input, perhaps using variational autoencoders (85), as illustrated in **Figure 9** (86), or generative adversarial networks (87); or (*b*) candidates generated from known constituent building blocks and filtered by rules or heuristics based on domain expertise. Of course, a third common route involves using a preexisting library of candidates, e.g., PubChem entries, but this strategy by definition will not yield any chemical novelty; it can only predict unknown properties of previously observed materials.

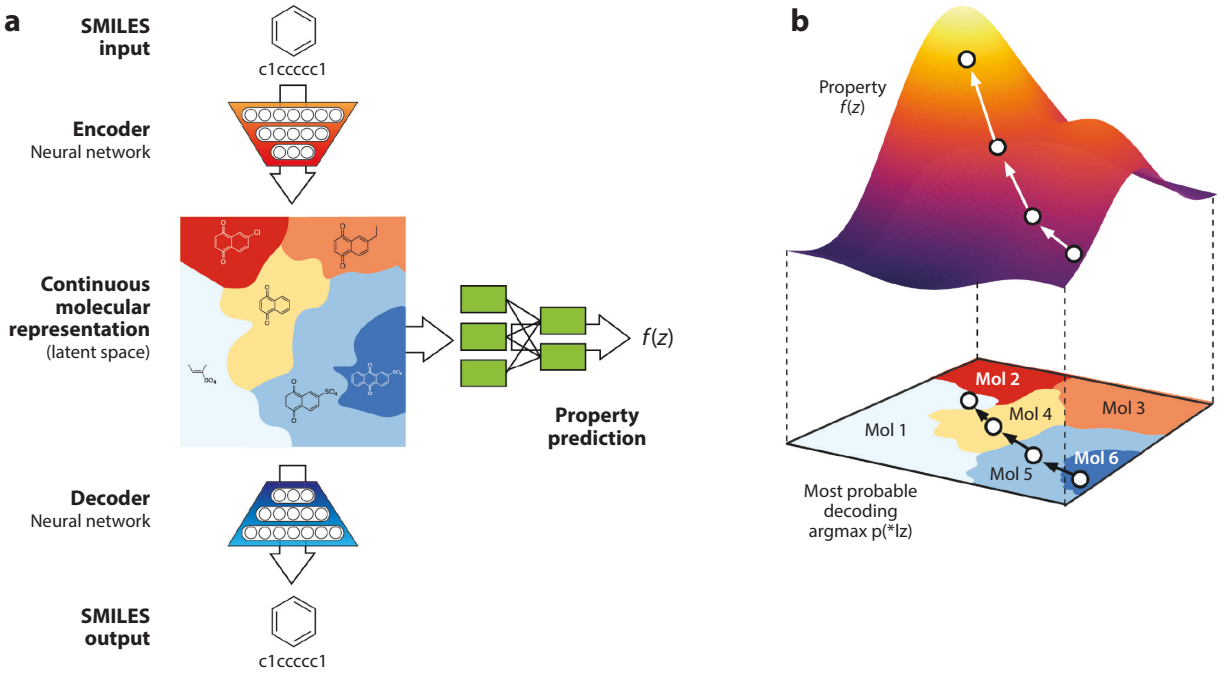

Figure 9

Gómez-Bombarelli et al. (86) utilized a variational autoencoder to encode molecular structures, as represented by SMILES strings, into a latent space that allows for continuous optimization of properties. A decoder then enables recovery of feasible molecules with human-interpretable SMILES representations. Abbreviation: SMILES, simplified molecular-input line-entry system. Adapted with permission from the American Chemical Society (ACS) from Reference 86 (**https://pubs.acs.org/doi/10.1021/acscentsci.7b00572**); further permissions related to the material excerpted should be directed to the ACS.

Chemical generation approaches were utilized in a pair of studies on OLED design. In the work of Gómez-Bombarelli et al. (64), 222 moieties and several design rules based on electronic structure and synthesizability were used to generate 1.6 million candidate molecules. The SMILES representations for these molecules were converted into the fixed-length vector extended-connectivity fingerprint (ECFP) representation and used as input into a NN trained on k_{TADF}, a time-dependent DFT-derived OLED performance metric. The NN predicted k_{TADF} for all candidate molecules, the resulting list was ranked, and DFT was performed on the top-ranked candidates to validate the prediction and add to the training data. After each iteration, the improved NN was used to identify top candidates, which were further downselected for experiment.

Kim et al. (63) dynamically generated candidate OLED structures through a pair of NNs. For their inverse design pipeline, they trained a deep NN on the DFT-predicted electronic properties of 50,000 randomly selected structures from the PubChem database. This deep NN then predicted the electronic properties of randomly generated ECFP vectors. For the top-performing candidates, a recurrent NN converted the ECFP vector into a SMILES string. If the SMILES string corresponded to a chemically valid structure, Kim et al. ran DFT on the structure, and recommended those structures exceeding performance targets for subsequent experimental investigation. The authors ran this pipeline until they had identified 1,500 molecules satisfying their design constraints. Approximately 10% existed in PubChem but were not present in the training set, suggesting that the pipeline is capable of reproducing molecules found by traditional chemistry expertise.

4. CONCLUSION

ML is beginning to exert a major impact on materials discovery. In **Table 1**, we highlighted a number of ML discovery case studies across a wide variety of materials applications. In areas ranging from small molecules to metal alloys, ML predictions have subsequently been confirmed by laboratory experiment or, in some cases, validated via physics-based simulations. In this review, we identified some strategies that have lent themselves to success in ML-aided materials discovery, such as explicitly combining domain knowledge with ML modeling. We also discussed the underlying components of a typical materials informatics discovery pipeline, such as high-quality input data and carefully selected materials representations.

Looking ahead, we anticipate important work in several areas of materials informatics. First, we must confront the issue of profound data scarcity in materials science; here, autonomous materials research (88), wherein ML-driven robotic synthesis and characterization apparatuses run in closed-loop fashion, could play a major role. Autonomy is particularly interesting because the time required for experimentation (i.e., data generation) is a severe bottleneck in the materials development process. Second, ML model interpretability, which enables researchers to gain greater physical insight from using ML models, will broaden the appeal and applicability of ML as a scientific tool. ML models are often considered black boxes (89) because the process by which inputs are transformed into outputs is opaque for many ML algorithms, but greater interpretability would ameliorate this limitation. Finally, we should be able to quantify the behavior of ML models during extrapolative materials discovery. In particular, design space optimization could maximize the likelihood that ML will find nonobvious, high-performing materials. If we can optimize where we are searching for discoveries, accelerate our rate of data generation, and enhance the transparency of ML models for scientific users, ML will be transformative for our ability to discover and design groundbreaking new materials.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. 2016. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* 41:399–409
2. Jain A, Hautier G, Ong SP, Persson K. 2016. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. *J. Mater. Res.* 31:977–94
3. Mueller T, Kusne AG, Ramprasad R. 2016. Machine learning in materials science: recent progress and emerging applications. *Rev. Comput. Chem.* 29:186–273
4. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. 2017. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* 3:54
5. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. 2018. Machine learning for molecular and materials science. *Nature* 559:547–55
6. Larsen P, Von Ins M. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics* 84:575–603
7. Ren F, Ward L, Williams T, Laws KJ, Wolverton C, et al. 2018. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* 4:eaaq1566
8. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs]
9. Allen FH. 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* 58:380–88

10. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, et al. 2007. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36:D351–59
11. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2008. GenBank. *Nucleic Acids Res.* 37:D26–31
12. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, et al. 2015. PubChem substance and compound databases. *Nucleic Acids Res.* 44:D1202–13
13. Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I. 2016. The materials data facility: data services to advance materials science research. *JOM* 68:2045–52
14. Draxl C, Scheffler M. 2018. NOMAD: the FAIR concept for big data–driven materials science. *MRS Bull.* 43:676–82
15. O'Mara J, Meredig B, Michel K. 2016. Materials data infrastructure: a case study of the Citrination platform to examine data import, storage, and access. *JOM* 68:2031–34
16. Seshadri R, Sparks TD. 2016. Perspective: interactive material property databases through aggregation of literature data. *APL Mater.* 4:053206
17. Rickman J, Chan H, Harmer M, Smeltzer J, Marvel C, et al. 2019. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nat. Commun.* 10:2618
18. Iwasaki Y, Takeuchi I, Stanev V, Kusne AG, Ishida M, et al. 2019. Machine-learning guided discovery of a new thermoelectric material. *Sci. Rep.* 9:2751
19. Balachandran PV, Kowalski B, Sehirlioglu A, Lookman T. 2018. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* 9:1668
20. Min K, Choi B, Park K, Cho E. 2018. Machine learning assisted optimization of electrochemical properties for Ni-rich cathode materials. *Sci. Rep.* 8:15778
21. Hatakeyama-Sato K, Tezuka T, Nishikitani Y, Nishide H, Oyaizu K. 2018. Synthesis of lithium-ion conducting polymers designed by machine learning–based prediction and screening. *Chem. Lett.* 48:130–32
22. Wen C, Zhang Y, Wang C, Xue D, Bai Y, et al. 2019. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* 170:109–17
23. Hohenberg P, Kohn W. 1964. Inhomogeneous electron gas. *Phys. Rev. B* 136:864–71
24. Kohn W, Sham LJ. 1965. Self-consistent equations including exchange and correlation effects. *Phys. Rev. A* 140:1133–38
25. Kresse G, Hafner J. 1993. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* 47:558–61
26. Hafner J. 2008. Ab-initio simulations of materials using VASP: density-functional theory and beyond. *J. Comput. Chem.* 29:2044–78
27. Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, et al. 2012. Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* 58:227–35
28. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1:011002
29. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. 2013. Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). *JOM* 65:1501–9
30. Ward L, Aykol M, Blaiszik B, Foster I, Meredig B, et al. 2018. Strategies for accelerating the adoption of materials informatics. *MRS Bull.* 43:683–89
31. Mansouri Tehrani A, Oliynyk AO, Parry M, Rizvi Z, Couper S, et al. 2018. Machine learning directed search for ultraincompressible, superhard materials. *J. Am. Chem. Soc.* 140:9844–53
32. Zhuo Y, Mansouri Tehrani A, Oliynyk AO, Duke AC, Brgoch J. 2018. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nat. Commun.* 9:4377
33. Bassman L, Rajak P, Kalia RK, Nakano A, Sha F, et al. 2018. Active learning for accelerated design of layered materials. *npj Comput. Mater.* 4:74
34. Huo H, Rupp M. 2017. Unified representation for machine learning of molecules and crystals. arXiv:1704.06439 [physics.chem-ph]
35. Askerka M, Li Z, Lempen M, Liu Y, Johnston A, et al. 2019. Learning-in-templates enables accelerated discovery and synthesis of new stable double perovskites. *J. Am. Chem. Soc.* 141:3682–90
36. Schütt K, Glawe H, Brockherde F, Sanna A, Müller K, Gross E. 2014. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* 89:205118

37. Ward L, Dunn A, Faghaninia A, Zimmermann NE, Bajaj S, et al. 2018. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* 152:60–69
38. Faber FA, Christensen AS, Huang B, von Lilienfeld OA. 2018. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* 148:241717
39. Ward L, Liu R, Krishna A, Hegde VI, Agrawal A, et al. 2017. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* 96:024104
40. Ward L, Agrawal A, Choudhary A, Wolverton C. 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* 2:16028
41. Rosenbrock CW, Homer ER, Csányi G, Hart GL. 2017. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *npj Comput. Mater.* 3:29
42. Gomberg JA, Medford AJ, Kalidindi SR. 2017. Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning. *Acta Mater.* 133:100–8
43. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, ed. D Precup, YW Teh, pp. 1263–72. New York: ACM. **https://dl.acm.org/doi/10.5555/3305381.3305512**
44. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. 2018. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* 148:241722
45. Xie T, Grossman JC. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* 120:145301
46. Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, et al. 2017. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* 3:e1701816
47. Hall MA. 1999. *Correlation-based feature selection for machine learning*. PhD Thesis, Univ. Waikato, Hamilton, N. Z.
48. Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A* 374:20150202
49. Sinkov NA, Harynuk JJ. 2011. Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* 83:1079–87
50. Meredig B, Antono E, Church C, Hutchinson M, Ling J, et al. 2018. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* 3:819–25
51. Lu HJ, Zou N, Jacobs R, Afflerbach B, Lu XG, Morgan D. 2019. Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion. *Comput. Mater. Sci.* 169:109075
52. Sheridan RP. 2013. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* 53:783–90
53. Wallach I, Heifets A. 2018. Most ligand-based classification benchmarks reward memorization rather than generalization. *J. Chem. Inf. Model.* 58:916–32
54. Riley P. 2019. Three pitfalls to avoid in machine learning. *Nature* 572:27–29
55. Vapnik VN. 1999. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* 10:988–99
56. Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, et al. 2018. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* 4:29
57. Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. 2017. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integr. Mater. Manuf. Innov.* 6:207–17
58. De Jong M, Chen W, Angsten T, Jain A, Notestine R, et al. 2015. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* 2:150009
59. Breiman L. 2001. Random forests. *Mach. Learn.* 45:5–32
60. Wager S, Hastie T, Efron B. 2014. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* 15:1625–51
61. Krogh A, Vedelsby J. 1996. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 8* (*NIPS 1995*), ed. DS Touretzky, MC Mozer, ME Hasselmo, pp. 231–38. Cambridge, MA: MIT Press
62. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44

63. Kim K, Kang S, Yoo J, Kwon Y, Nam Y, et al. 2018. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.* 4:67
64. Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, et al. 2016. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* 15:1120–27
65. Oliynyk AO, Antono E, Sparks TD, Ghadbeigi L, Gaultois MW, et al. 2016. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* 28:7324–31
66. Agrawal A, Deshpande PD, Cecen A, Basavarsu GP, Choudhary AN, Kalidindi SR. 2014. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr. Mater. Manuf. Innov.* 3:90–108
67. Iwasaki Y, Sawada R, Stanev V, Ishida M, Kirihara A, et al. 2019. Materials development by interpretable machine learning. arXiv:1903.02175 [cond-mat.mtrl-sci]
68. Xue D, Xue D, Yuan R, Zhou Y, Balachandran PV, et al. 2017. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Mater.* 125:532–41
69. Hassabis D. 2017. Artificial intelligence: chess match of the century. *Nature* 544:413–14
70. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak J, et al. 2014. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* 89:094104
71. Zeier WG, Anand S, Huang L, He R, Zhang H, et al. 2017. Using the 18-electron rule to understand the nominal 19-electron half-Heusler NbCoSb with Nb vacancies. *Chem. Mater.* 29:1210–17
72. Hutchinson ML, Antono E, Gibbons BM, Paradiso S, Ling J, Meredig B. 2017. Overcoming data scarcity with transfer learning. arXiv:1711.05099 [cs.LG]
73. De S, Bartók AP, Csányi G, Ceriotti M. 2016. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 18:13754–69
74. Liu AY, Cohen ML. 1989. Prediction of new low compressibility solids. *Science* 245:841–42
75. Menon A, Childs CM, Poczós B, Washburn NR, Kurtis KE. 2019. Molecular engineering of superplasticizers for metakaolin-portland cement blends with hierarchical machine learning. *Adv. Theory Simul.* 2:1800164
76. Santosa F, Symes WW. 1986. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* 7:1307–30
77. Bucior BJ, Bobbitt NS, Islamoglu T, Goswami S, Gopalan A, et al. 2019. Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Mol. Syst. Des. Eng.* 4:162–74
78. Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, et al. 2016. Machine-learning-assisted materials discovery using failed experiments. *Nature* 533:73–76
79. Kawazoe Y, Yu J-Z, Tsai A-P, Masumoto T, eds. 1997. *Phase Diagrams and Physical Properties of Nonequilibrium Alloys*, Subvol. A: *Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys*. Landolt–Börnstein Numer. Data Funct. Relatsh. Sci. Technol. 37. Berlin/Heidelberg/New York: Springer
80. Sakurai A, Yada K, Simomura T, Ju S, Kashiwagi M, et al. 2019. Ultranarrow-band wavelength-selective thermal emission with aperiodic multilayered metamaterials designed by Bayesian optimization. *ACS Cent. Sci.* 5:319–26
81. Sendek AD, Yang Q, Cubuk ED, Duerloo KAN, Cui Y, Reed EJ. 2017. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* 10:306–20
82. Sendek AD, Antoniuk ER, Cubuk ED, Francisco BE, Buettner-Garrett J, et al. 2019. A new solid Li-ion electrolyte from the crystalline lithium-boron-sulfur system. *SSRN Electron. J.* **https://dx.doi.org/10.2139/ssrn.3404263**
83. Häse F, Roch LM, Aspuru-Guzik A. 2018. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* 9:7642–55
84. Solomou A, Zhao G, Boluki S, Joy JK, Qian X, et al. 2018. Multi-objective Bayesian materials discovery: application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. *Mater. Des.* 160:810–27
85. Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML]

86. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, et al. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4:268–76
87. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27* (*NIPS 2014*), ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger. San Diego: Neural Inf. Process. Syst.
88. Nikolaev P, Hooper D, Webber F, Rao R, Decker K, et al. 2016. Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* 2:16031
89. Holm EA. 2019. In defense of the black box. *Science* 364:26–27
90. Mannodi-Kanakkithodi A, Pilania G, Huan TD, Lookman T, Ramprasad R. 2016. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* 6:20952
91. Oliynyk AO, Adutwum LA, Harynuk JJ, Mar A. 2016. Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis. *Chem. Mater.* 28:6672–81
92. Villars P, ed. 2007. *Pearson's Crystal Data®: crystal structure database for inorganic compounds.* Database, ASM Int., Materials Park, OH
93. Oliynyk AO, Adutwum LA, Rudyk BW, Pisavadia H, Lotfi S, et al. 2017. Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases ABC. *J. Am. Chem. Soc.* 139:17870–81
94. Seko A, Hayashi H, Kashima H, Tanaka I. 2018. Matrix-and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Phys. Rev. Mater.* 2:013805
95. Belsky A, Hellenbrandt M, Karen VL, Luksch P. 2002. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. B* 58:364–69
96. Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. 2018. Accelerated discovery of stable lead-free hybrid organic–inorganic perovskites via machine learning. *Nat. Commun.* 9:3405
97. Wu S, Kondo Y, Kakimoto M, Yang B, Yamada H, et al. 2019. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* 5:66
98. Otsuka S, Kuwajima I, Hosoya J, Xu Y, Yamazaki M. 2011. PoLyInfo: polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pp. 22–29. Piscataway, NJ: IEEE
99. Ramakrishnan R, Dral PO, Rupp M, Von Lilienfeld OA. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1:140022

Annual Review of Materials Research

Volume 50, 2020

Contents

Data-Driven Discovery of Materials

Quantum Materials and Devices

Current Interest

Indexes

Errata

An online log of corrections to *Annual Review of Materials Research* articles may be found at http://www.annualreviews.org/errata/matsci