ADVANCED ENERGY MATERIALS
www.advenergymat.de

A Critical Review of Machine Learning of Energy Materials

Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong**

Machine learning (ML) is rapidly revolutionizing many fields and is starting to change landscapes for physics and chemistry. With its ability to solve complex tasks autonomously, ML is being exploited as a radically new way to help find material correlations, understand materials chemistry, and accelerate the discovery of materials. Here, an in-depth review of the application of ML to energy materials, including rechargeable alkali-ion batteries, photovoltaics, catalysts, thermoelectrics, piezoelectrics, and superconductors, is presented. A conceptual framework is first provided for ML in materials science, with a broad overview of different ML techniques as well as best practices. This is followed by a critical discussion of how ML is applied in energy materials. This review is concluded with the perspectives on major challenges and opportunities in this exciting field.

1. Introduction

Machine learning (ML) is the branch of artificial intelligence that deals with the development of algorithms and models that can automatically learn patterns from data and perform tasks without explicit instructions. While ML models and algorithms have been known since the 1950s, it is only in the recent decade that the systematic generation and curation of data on unprecedented scales—coupled with exponential increases in computing power—that ML has begun to break new frontiers across many fields, including biology,[1] physics,[2,3] and chemistry.[4,5] ML is especially suited for exploratory tasks that feature combinatorially or exponentially complex solutions. This is exemplified by the recent triumph of AlphaGo in solving the problem of Go, which has an estimated 10^{170} potential outcomes.[6]

This ability of ML to generalize from a set of training data to explore unknown spaces makes it a tantalizing panacea to many challenges in materials science.[7,8] Take, for example, the problem of novel materials discovery. To date, there are about 10^6 crystalline materials and 10^9 molecules explored either computationally or experimentally,[9–16] a minuscule fraction of the universe of possible crystals and molecules (e.g., it is estimated that there are 10^{60} possible small organic molecules alone). While accurate first-principles computational techniques such as density functional theory (DFT)[17,18] have brought about a revolutionary leap in our ability to predict properties and design materials in silico,[19–22] high computational costs and poor scaling still limit their effectiveness in exploring unconstrained chemical spaces and/or complex real-world materials. For instance, high-throughput DFT screening works typically limit the search space to hundreds or, at best, thousands of materials, while DFT simulations of materials are mostly limited to typically less than 1000 atoms, i.e., bulk crystals and isolated molecules. ML therefore offers a solution to the materials exploration problem, making predictions of new materials or properties from existing data, which in turn can drive the generation of more data that can be used to further refine the ML models.

Here, we will provide an in-depth, critical review of ML-guided design and discovery of energy materials, a field where a novel material with superior performance (e.g., higher energy density, higher energy conversion efficiency, etc.) can have a transformative impact on the urgent global problem of climate change. This review is structured along the steps in a typical workflow for materials ML model building, as shown in **Figure 1**. The next four sections will provide a concise overview of ML concepts designed to give the reader an appreciation of state-of-the-art techniques as well as resources for building ML models for materials. Section 6 reviews the actual application of ML techniques to the discovery and design of various classes of energy materials, from energy storage (e.g., batteries, fuel cells, etc.) to energy conversion (e.g., thermoelectrics, catalysis, etc.). The final section outlines our perspectives on various challenges and opportunities in ML for energy materials design.

2. Goal/Target Identification

The first step in any ML project is to identify the goals and prediction targets of the ML models, typically relying on the domain knowledge of experts. This step is arguably the most important as the choice of target must be potentially learnable from available information, e.g., crystal/molecule structure and composition, elemental information, experimentally measured quantities or images, etc., and is unambiguously defined. The wrong choice of prediction target can lead to models that are either nongeneralizable or have spuriously high errors. Experimental and computational sources of materials data often have well-known uncertainties or errors. For example, stability is a practical criterion that cannot be ignored in most materials design problems.[23–26] Computationally, the thermodynamic stability is typically estimated using either the 0 K DFT formation energy E_f or the energy above convex

Dr. C. Chen, Y. Zuo, W. Ye, Dr. X. Li, Dr. Z. Deng, Prof. S. P. Ong
Department of NanoEngineering
University of California San Diego
9500 Gilman Dr, Mail Code 0448, La Jolla, CA 92093-0448, USA
E-mail: chc273@eng.ucsd.edu; ongsp@eng.ucsd.edu

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aenm.201903242.

DOI: 10.1002/aenm.201903242

hull E_{hull}.[23] The E_{hull} is generally a difficult target property for ML.[27] It is highly sensitive to competing phases in the phase diagram; incompleteness of the phase diagram or the presence of artificially stabilized phases due to DFT errors can lead to large errors in E_{hull}. Furthermore, the E_{hull} is lower bounded at zero. E_f is a better regression target, though the choice of the reference states is also important. Often, the elemental ground states are used as the reference states for computing E_f. Unfortunately, this quantity is plagued by well-known DFT errors associated with incomplete cancellation of errors especially when redox reactions are involved.[28] By choosing reference states that have the same oxidation state as the final compound, e.g., binary oxides, such errors can be minimized.[27] It should be noted that E_{hull} can always be obtained from E_f with the existence of a predictive E_f model, and stability classification can then be obtained with the application of a suitable threshold, e.g., a strict threshold of 0 to identify phases on the hull, or more commonly, some positive threshold to account for potential metastability and uncertainty in DFT and ML predictions.

The nature of the materials problem can also influence the choice of model (see Section 5). For example, one key consideration is whether the target problem is one of classification or regression. A classification task aims at learning the mapping between inputs and categorical targets. In materials science, many properties can be seen as categorical, for example, metal versus nonmetals, superconductor versus non-superconductors, relative stability between polymorphs (e.g., the cubic, tetragonal, and orthorhombic forms of perovskites), etc. A regression task, on the other hand, aims to learn the mapping between inputs and numerical target values, e.g., formation energies, bandgaps, conductivity, etc. It should be noted that with the use of proper thresholds, regression tasks can be converted to classification tasks.[29,30]

3. Data Collection

In the second step, training data—experimental, or more commonly, computed—is collected. Data can come from publicly available sources or be self-generated, e.g., by carrying out experiments or high-throughput computations using various software platforms. It is critical that the data be of both sufficient quantity and high quality. The required quantity of data depends on the choice of ML model (see Section 5), but a general rule of thumb is that at least 50 data points are necessary for a reasonable ML model, with certain models, e.g., neural networks, requiring much larger quantities. The quality of the data is determined by its coverage of the chemical-property space of interest as well as the uncertainty associated with the data. For example, if the goal of ML models is to predict materials properties across the periodic table, then the collected data with a limited number of elements is unlikely to form good data distribution for this purpose. In general, the collected data should represent future unseen data and have the same distributions if possible. Data uncertainty, on the other hand, can come from many sources, including experimental errors from measurements or computational errors from unsatisfactory approximations.

Chi Chen is an assistant project scientist in Materials Virtual Lab at University of California San Diego. He received a B.E. from the University of Science and Technology in China and Ph.D. from the Hong Kong University of Science and Technology in 2016. His research interests include materials informatics and machine learning model developments for accelerating material simulations.

Shyue Ping Ong is an associate professor in the Department of NanoEngineering at the University of California, San Diego. He leads the Materials Virtual Lab, a research group focused on the interdisciplinary application of high-throughput first-principles computations and machine learning to the study and design of materials. He received his M.Sc. from Cambridge University and Ph.D. in materials science engineering at Massachusetts Institute of Technology in 2011.

Data quantity and quality is perhaps the central challenge in the application of ML in materials science.[31] Experimental crystal/molecular structure databases, such as the Pauling File Database,[14] Inorganic Crystal Structure Database (ICSD),[32] Pearson Crystal Data,[33] Cambridge Structural Database,[16] Crystal Open Database,[34] CRYSTMET,[15] Protein Data Bank (PDB),[35] ZINC database,[36] GDB databases,[37–40] PubChem,[41] etc., have been steadily built up over the past few decades. For a comprehensive review of crystallographic databases, please see ref. [42]. Similarly, there are several well-established compilations of measured thermodynamic properties and general materials physical chemistry data.[43–46] However, many existing experimental data repositories are still either too small or too heterogeneous (e.g., different experimental conditions or measurement techniques) for high quality ML models. Human bias can also affect data diversity.[47] Furthermore, the vast majority of databases are commercial products requiring a license, and programmatic application programming interfaces (APIs) for large-scale data access are rarely implemented. A large fraction of experimental data are only available in journal publications, though recent successes in text mining offer a potential solution to this conundrum.[48–53] Finally, major efforts are underway in high-throughput/combinatorial experiments that can generate large experimental materials database with diverse properties.[31]

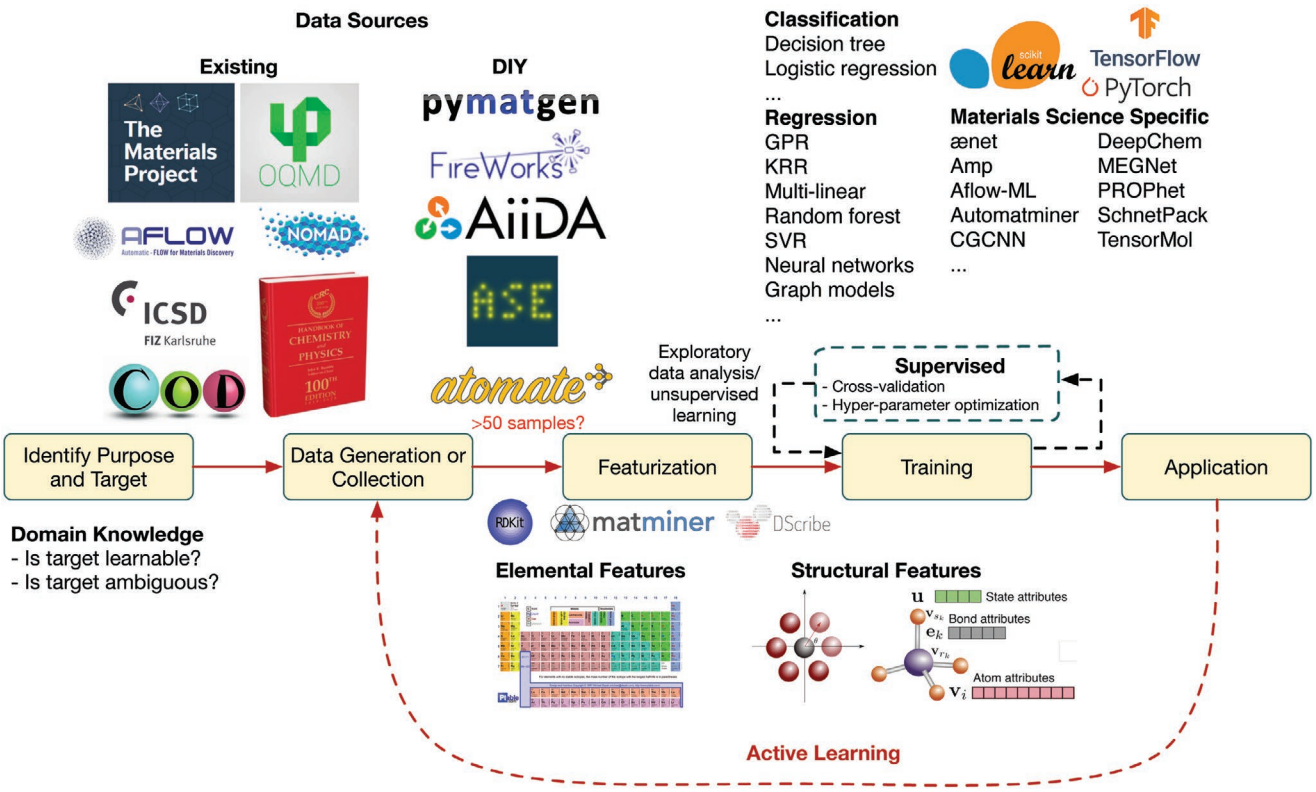

Figure 1. A workflow for constructing ML models. Five major steps are involved in this workflow, starting from identification of purpose, to data collection, featurization, model building, and eventually application. Various open source databases and model packages have enabled a much easier experience of model construction.

It is therefore not surprising that many ML works have turned to computed data sources. Computations, particularly those based on DFT and other ab initio techniques, are more easily scaled across diverse chemical spaces than experiments. Moreover, recent efforts under the Materials Genome Initiative have led to the proliferation of large, public databases of computed materials properties. Examples of general-purpose databases with high chemical diversity (typically based on large experimental databases of all known inorganic materials such as the ICSD as well as a subset of generated hypothetical structures) and property diversity (relaxed structures, energies, electronic structure properties, etc.) include the Materials Project,[10] AFLOWLIB,[54] the computational materials repository,[55] open quantum materials data (OQMD),[56] novel materials discovery (NOMAD) repository,[57] AiiDA,[58] JAVIS-DFT,[59] CatApp,[60] etc. For example, the Materials Project,[10] one of the most popular computed data sources for ML works, currently hosts ≈133 000 DFT-relaxed crystal structures, at the time of writing, with energies and electronic structure properties such as the bandgap available for the majority of materials and other properties such as elastic constants, piezoelectric coefficients, etc., available for a subset of materials. **Figure 2** shows the historical trend in the number of crystals with various properties computed. Several of these databases, e.g., Materials Project and AFLOWLIB, also have well-defined APIs for rapid data access—a key requirement for efficient ML construction.[61,62] In addition, there are many useful specialized databases such as the Harvard Clean Energy Project (CEP)[63] and the NREL materials database[64] that focus on energy applications that have been used for several ML works. For a more complete list of materials database and the corresponding tools, please see refs. [65,66].

While existing computed databases serve as excellent starting points for ML model building, augmenting the data

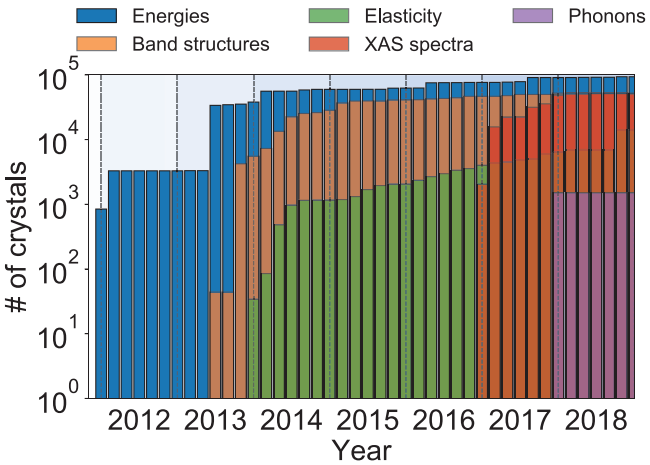

Figure 2. Number of crystal structures with various properties computed in the Materials Project since inception.

through additional computations is unavoidable in many instances, e.g., when there is insufficient data in a particular property of interest or poor coverage in a particular chemistry or when extremely high consistency in data is required. Fortunately, many of the large open databases also have open-source software platforms to facilitate the conduct of high-throughput computations. Examples include the Materials Project's software suite comprising the Python Materials Genomics (pymatgen) materials analysis library,[67] FireWorks[68] and Atomate scientific workflow packages,[69] AFLOW,[9] and AiiDA.[58]

4. Featurization

Featurization refers to the process in which the training data is transformed into numerical values (typically in the form of vectors or tensors) that distinguishes between different materials. These numerical values have been referred to as features, descriptors, or fingerprints in the literature. In this review, we will use the general term "features." The choice of features depends on the goals of the ML model (see Section 2) and is frequently the step that involves the most human intervention and has a major impact on model performance.

Good feature sets provide just sufficient resolution for the prediction space. While this definition is clearly problem-dependent, a typical requirement in many materials ML problems is that the feature set needs to provide a unique, i.e., one-to-one, correspondence with the crystal/molecular composition and/or structure. At the same time, feature sets should also not be excessive in size for efficient training and predictions and to avoid overfitting, especially in materials science where data sets tend to be small. In particular, redundant and/or highly correlated features should be avoided where possible. For example, the atomic number of an element together with its group and period numbers in the periodic table would constitute redundant information and a choice should be made between them, while many elemental properties are highly correlated, e.g., melting/boiling points with elastic constants, both being related to the cohesive energy. Related to feature set size is the requirement that the feature set can be efficiently constructed from available data and/or computed ones. ML models trained on relaxed crystal/molecular structures and/or electronic structure[29,70,71] from first-principles computations are limited in this aspect unless the target property is so expensive that the initial first-principles computation is a negligible part of the cost and/or the property is not highly sensitive to errors between first-principles and experimental structures.

It is not possible to comprehensively enumerate the feature sets used in materials ML models. Here, we will limit our discussion to the description of crystals/molecules, which is the input for a large proportion of materials ML problems from surrogate property prediction to interatomic potentials (IAPs). One possible classification of crystal/molecular features is whether the crystal or molecular structure, i.e., the positions of the individual atoms and the bonds between them, are included. Composition-based features, which excludes structural information, has been extensively used in materials science even prior to the current resurgence of interest in ML. Typically, such features are derived from the known properties of the constituent elements, e.g., the atomic number, electronegativity, atomic radii, electronic structure, etc. For example, Ward et al.[72] has shown that using elemental physical properties as descriptors for structure yields reasonably good performance in predicting various properties, including glass-forming ability and bandgaps. However, composition-based features by definition are unable to distinguish between crystal polymorphs and molecular isomers/conformers. For example, diamond and graphite, which have very different physical and chemical properties, would be indistinguishable in a composition-based ML model. As such, composition-based featurization should be used only in instances where the structural degrees of freedom are constrained, e.g., in problems where only a particular structural prototype, such as perovskite or garnets, is of interest,[27,73] or an assumption is made that only the ground state polymorph is of interest.[74,75]

For most problems, a feature set that describes the full crystal/molecular structure is desired. The development of crystal/molecular structural features remains an active area of research. Nevertheless, there are well-established guidelines. For example, crystal/molecular structural features must be invariant to translation, rotation, and permutation of homonuclear atoms,[76,77] unless these invariances are imposed within the ML model itself (e.g., convolutional neural networks (CNNs) are frequently used to address translational invariance in images.[78]) In addition, certain applications may impose additional constraints, e.g., differentiability of features is a typical requirement for ML-IAPs. The Coulomb matrix, which encodes the Coulomb interactions between all pairs of atoms,[79] is an example of a molecular structure featurization. Other distance-based features include London matrix,[80] histograms of distance, angle, dihedral (HDAD),[81] and molecular atomic radial angular distribution (MARAD).[81] For molecules, intermediate representations exist that encode connectivity between molecular fragments, such as the commonly used simplified molecular-input line-entry system (SMILES),[82] extended-connectivity fingerprint (ECFP),[83] bag-of-bonds,[84] bonding angular ML (BAML),[80] etc. These representations can distinguish between isomers, but not conformers. Another common strategy is a bottom-up approach, whereby features are constructed from the local environment of each atom and combined at the crystal level. Such descriptors include atom-centered symmetry functions (ACSF),[85] bispectrum coefficients,[86] smooth overlap of atomic positions (SOAP),[76] moment tensors,[87] classical force-field-inspired descriptors (CFID),[88] etc. They benefit from the locality of target properties, e.g., energy can be divided into atomic energy. Such assumptions can still be valid in molecules and thus these descriptors have also been applied to molecular structures.[89] Less obvious structure-based features may take existing properties or extract computational results using full structural information for the investigated materials. For example, the d-band center descriptor[70,90] for metals, as well as the related oxygen p-band center descriptor for oxides,[71] has been used extensively to describe catalytic activities.

Finally, graph-based featurization has gained substantial interest in recent years. Graphs, which are natural representations for atoms (nodes) and the bonds between them (edges), have been used for molecules for many decades[91] and have recently been applied to ML in crystals, achieving state-of-the-art

performance in predicting the formation energies, bandgaps, as well as metal/insulator classification.[92–94]

As noted at the beginning of this section, proper feature selection is critical to the performance and generalizability of the ML model. The selection of features can be domain-knowledge-driven or data-driven. The former relies on the application of physical and chemical intuition to select appropriate features for the ML problem. For example, the electronegativity and atomic radii are commonly used features in many ML models[27,95,96] due to their prominence in well-established rules such as the Pauling's five rules[97] and the Goldschmidt tolerance factor.[98] While undoubtedly more efficient features can generally lead to more interpretable models, the domain-knowledge-driven approach introduces bias into the feature selection process, which may result in nonoptimal performance and blind spots to new insights. In contrast, the data-driven approach starts from an initial large set of candidate features and down-selects a subset of features. This down-selection process can be automatic, e.g., using L_1 or L_0 regularization (least absolute shrinkage and selection operator, LASSO),[99–101] feature importance,[102,103] genetic algorithms,[104] etc. However, a drawback of data-driven feature selection is that the selected features do not imply causality with respect to the target and will be highly dependent on the chosen hyperparameters of the model.[105] Yet another approach is to use dimension reduction algorithms to "synthesize" new and low-dimensional features from the original features. The principal component analysis (PCA)[106] is widely used in this context.[92,107,108] It works well if the data in high dimension is intrinsically low-dimensional, e.g., a plane in 3D. While PCA works on linear projection, the manifold learning is able to capture nonlinear relationships. For example, the t-distributed stochastic neighbor embedding (t-SNE)[109] method learns low-dimensional representations such that the local distance between data points is roughly preserved and has been applied in visualizing the elemental embedding vector trained from materials property prediction models,[94] structural similarity of perovskites,[110] word embeddings in text mining,[52] electronic fingerprints,[111] etc.

Owing to the explosion in interest in materials ML, there has been a proliferation of open-source software tools to facilitate featurization of crystals/molecules. A noncomprehensive list includes RDKit,[112] Dscribe,[113] Matminer,[114] Materials Agnostic Platform for Informatics and Exploration (Magpie),[72] MatErials Graph Network (MEGNet),[94] etc. Interested readers may wish to explore these tools for their ML projects.

5. Model Selection and Training

5.1. Model Categories

There are three main categories of ML—supervised learning, unsupervised learning and reinforcement learning. By far the most common type of ML in materials science is supervised learning, where a model learns the functional mapping between input features (e.g., crystal/molecular structure and composition) and output labels/values (e.g., properties such as energies, bandgaps, etc.) using example input–output pairs.[7,66,115,116] The goal of such ML models is typically to bypass expensive and time-consuming experiments or first-principles computations. They have been used to provide guidance to experimental design, i.e., the next areas to explore for a potential "blockbuster" material for an application[117–121] as well as rapid computational screening of chemical spaces.[30,122–125] ML-IAPs are also a subcategory of supervised materials ML models where the target is to predict the energies, forces and stresses, i.e., the potential energy surface, for a given atomic configuration.

In unsupervised learning, the goal is to identify patterns from data without input labels. In materials science, it has been applied to study the collective diffusion of ions[126–128] and visualize complex high-dimensional data.[129] Lastly, reinforcement learning mimics how humans learn by interacting with environments; the algorithm improves in its ability to perform certain tasks through feedback in the form of rewards or punishments. Although reinforcement learning is new to the field, related approaches such as active learning and Bayesian optimizations have already been used to train systematically improvable IAPs[130,131] or optimize the composition within a chemical space to achieve better performance.[132]

Here, we will provide a noncomprehensive enumeration and summary of common ML models necessary to understand their application in various energy materials problems; readers interested in a more comprehensive treatise on the topic are pointed to several excellent textbooks and recent reviews.[66,133–135]

5.1.1. Linear Models and Generalized Linear Models

The multilinear regression model is the simplest and arguably the most widely used model in materials science. Indeed, many single-variable-descriptor approaches such as the d-band[70,90] and p-band center descriptor[71] are essentially linear models. In a linear model, the importance of each variable is directly related to the model prediction, as follows

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} \tag{1}$$

where $\mathbf{X}$ is the feature matrix or design matrix with each row being a data point, $\mathbf{y}$ is the target property column vector, and $\boldsymbol{\beta}$ is the column linear coefficients vector. $\boldsymbol{\beta}$ is usually estimated using a least squares approach, whereby

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 \tag{2}$$

with $\| \cdot \|_2$ indicating L2 norm. Linear regression models typically assume that the errors follow a Gaussian distribution. In particular, this condition is violated if the target property is bounded in some way, e.g., the bandgap of a material is always non-negative. Generalized linear models attempt to address this issue by using a link function of the target. For example, the logistic regression model (a binary classification model despite the name) is a generalized linear model that uses the logit function $\mathbf{u} = \log(\mathbf{y}/(\mathbf{1}-\mathbf{y}))$ as the link function

$$\log \frac{\mathbf{y}}{\mathbf{1}-\mathbf{y}} = \mathbf{X}\,\boldsymbol{\beta} \tag{3}$$

As discussed in Section 4, selecting the right features is necessary to improve model generalizability and interpretability. Regularization is frequently carried out with linear models to perform feature selection. In these techniques, a penalty term based on the magnitude of the coefficients is added during model training. For example, in least squares estimation, the optimized coefficients satisfy the following

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \alpha_1 \left\| \boldsymbol{\beta} \right\|_1^2 + \alpha_2 \left\| \boldsymbol{\beta} \right\|_2^2 \right) \tag{4}$$

where $\|.\|_1$ and $\|.\|_2$ denote the L1 and L2 norms, respectively, and α_1 and α_2 are tunable hyperparameters. $\alpha_1 = 0$ yields the ridge regression model, and $\alpha_2 = 0$ yields the LASSO model. When both α_1 and α_2 are nonzero, the elastic net model is obtained. The ridge and LASSO regression models have found wide applications in materials ML to solve the small-data problem,[136] compute phonons using compressive sensing,[137] perform feature selection,[99] and avoid overfitting.[138]

5.1.2. Kernel-Based Models

In reality, most relationships between inputs and outputs are nonlinear. One straightforward way to extend the linear model is to use higher order polynomial expansion of the input features. For example, instead of using $y = \beta_0 + \beta_1 x$, one may use $y = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\beta}$, with the polynomial expansion basis $\boldsymbol{\phi}(\boldsymbol{x}) = [1, x, x^2, \ldots, x^{(m-1)}]^{\mathrm{T}}$, where m is the feature dimension. It can be easily shown that the optimized coefficient vector $\boldsymbol{\beta}$ can be expressed as a linear combination of the basis function valued at data features Equation (5), and the prediction of new inputs only requires the combination coefficients $\boldsymbol{\lambda}$ instead of the original linear coefficients $\boldsymbol{\beta}$, Equation (6).

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i \, \boldsymbol{\phi}(x^{(i)}) \tag{5}$$

$$y^* = \boldsymbol{\phi}(x^*)^{\mathrm{T}} \, \boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i \langle \boldsymbol{\phi}(x^*), \, \boldsymbol{\phi}(x^{(i)}) \rangle \tag{6}$$

where n is the data size, λ_i is the combination coefficients, and the $<\cdot, \cdot>$ is the inner product. The inner product can be generalized to other functions that compute the similarity between two inputs, i.e., the kernel. In the original model, m coefficients (β) need to be solved from n data points, and using the kernel trick, only n coefficients (λ) are needed. When the polynomial feature dimension m goes to infinite, the inner product kernel is equivalent to Gaussian kernel $k(x, x') = \exp(-|x - x'|^2/2\sigma^2)$.[139] The Gaussian kernel has been routinely applied in other kernel-based methods, including Bayesian linear regression, support vector machine/regression/classification (SVM/SVR/SVC), kernel ridge regression (KRR), Gaussian process regression (GPR),[139] etc. Kernel-based models have been widely used in materials ML, for example, in constructing Gaussian approximation potential (GAP),[86] predicting molecular properties,[81] adsorption of gases on alloy nanoparticles,[140] lithium conductivity in LISCON,[141] thermal conductivity in solids,[142] potential energy surfaces,[86] etc.

5.1.3. Tree-Based Models

The decision tree model is a rule-based model that classifies the target label by asking a series of yes or no questions on the input features. The decision tree model has been used to predict faults in steel,[143] predict the defect types in B2 intermetallics,[144] select dopants for ceria working as water splitting catalysts,[145] etc. To further enhance the prediction accuracy and robustness, the random forest (RF) model uses an ensemble of decision trees (forming a "forest") and then computes the average or majority votes as the final prediction result. The RF model has been shown to be extremely robust in predicting materials properties, such as bulk modulus[146] and thermal conductivity,[147] and in compound classification.[148]

5.1.4. Cluster Analysis

Cluster analysis is used for finding intrinsic data relationships and grouping objects that are similar within the group but not so across different groups. The cluster analysis in fact is similar to kernel-based methods, because the data similarity or kernel is key to their success. For example, the k-nearest neighbors (kNN) algorithm uses similarity or distance between input features and obtains the prediction results of new input using averaging or voting based on the target results of its k nearest neighbor data points. Cluster analysis excels when data insights are extracted from unlabeled data. For example, Chen et al.[126,127] have used a k-means clustering algorithm in the analysis of oxygen diffusion patterns, hopping statistics, and site occupancies in crystals. Such algorithms however require prior knowledge of the number of clusters, i.e., the k. To solve this issue, the same authors[128] developed a parameter-free density-based clustering approach for studying the fast lithium diffusion with a relatively flat potential energy surface. In a different study, Meredig and Wolverton[149] have used cluster analysis to group defect energies in the periodic table by a x-means method.

5.1.5. Deep Learning

Deep learning[150] is a class of methods that are gaining popularity recently due to its revolutionary performance in various tasks including speech recognition, object detection, drug discovery, and chemistry.[151] It is defined in terms of multiple layers of neurons that progressively extract features and multiple levels of abstractions from the inputs. Deep learning is flexible in the choice of the number of parameters. It can handle problems with different levels of complexity and is particularly suited for big data problems.[152–154] The other key use of deep learning is to learn the feature representations from the data instead of feature engineering the input data. The learned representations can often be transferred to other similar tasks and greatly enhance the performance of the model.[155]

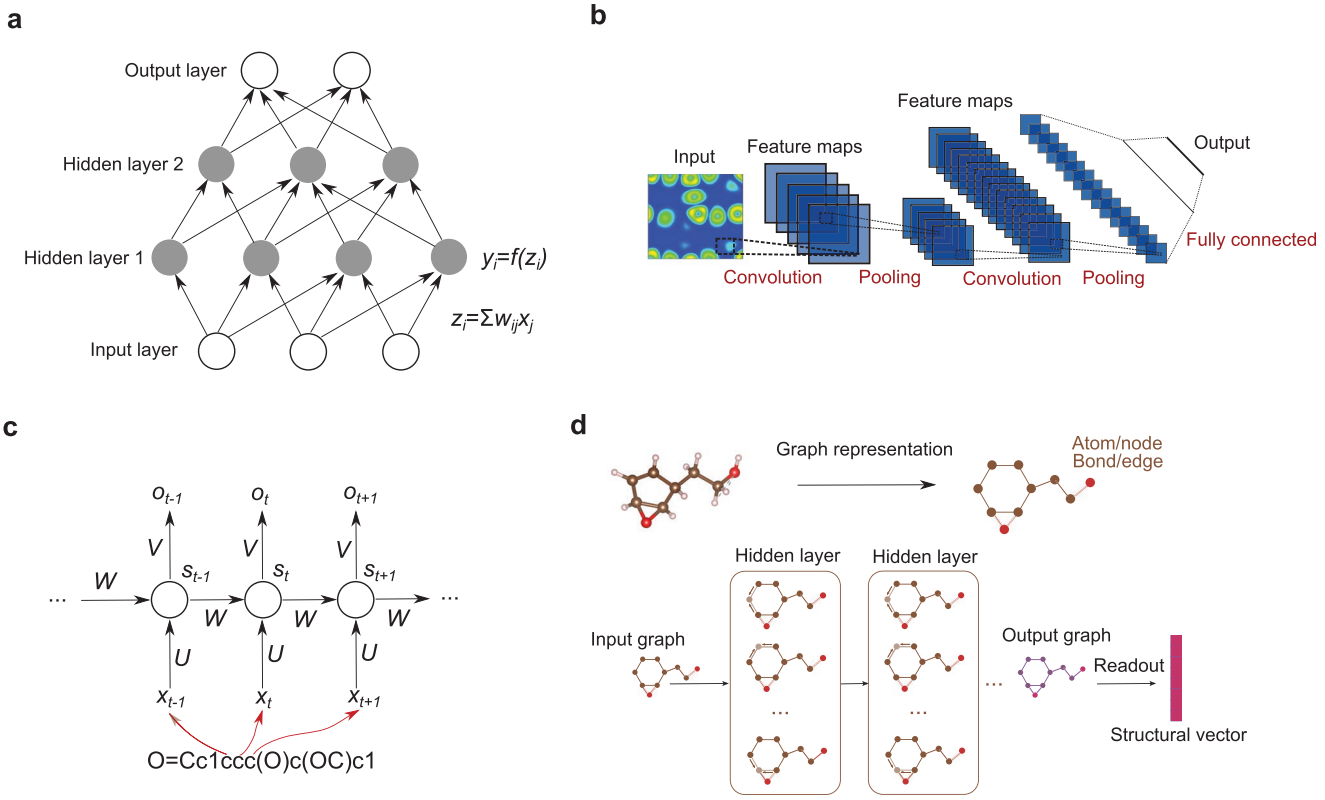

Figure 3. a) Deep learning architectures for multilayer perceptron model, b) convolutional neural networks, c) recurrent neural networks with example of processing SMILES representation of Vanillin, and d) graph representation of a molecule and the graph convolutional neural networks on the molecule graph.

The most conceptually simple and often-used deep learning model is the multilayer perceptron (MLP), otherwise known as the artificial neural network (**Figure 3**a). MLPs loosely model a biological brain where the computing units work as artificial neurons. An MLP is a layered architecture and in each layer the inputs undergo an affine transformation followed by a nonlinear activation function. Their ability to be universal function approximators[156] makes them extremely useful for materials property predictions,[27,157] if sufficient training data exist. MLPs have also been used in combination with local environment features to develop interatomic potentials.[85,158–160] CNNs (Figure 3b)[161] are rapidly gaining interest due to their recent feat of outperforming other ML algorithms by a considerable margin in image recognition.[162] In CNN, the same convolution filters are applied on all image patches that have the same size and eventually the filters output different feature maps. The locality of the convolution filters in fact share the same principal of locality of interactions in many materials. One naive approach of using CNN in materials science is therefore converting a material structure into an image, followed by training the CNN models on the converted data. However, this naive approach is not practical since CNN does not satisfy the requisite rotational and permutational invariances. Nevertheless, the model robustness can be improved by rotational data augmentation.[163,164] Instead of constraining the operation spatially as in CNN, the recurrent neural network (RNN) connects computing nodes in temporal sequence and all the steps share the same weights (Figure 3c). RNN works on sequential data and has been the model-of-choice for text and speech recognition. However, RNN is less common in materials science due to the lack of sequential data, except for molecules where the SMILES string-like representations can be readily fed into the RNN model.[165]

Deep learning models can be used with graph representation of molecules or crystals. In early graph CNN (GCNN) models (Figure 3d), information exchange is carried out between bonded atoms using deep learning models (typically MLPs) as function approximators. With more graph convolutional layers, one atom is able to "see" longer distances.[166–169] Modified graph models can also update bond information using information from atoms that form the bond.[167] More recently, graph networks further generalize the GCNN by introducing global attributes in addition to atom and bond attributes, and allowing information flow among all three levels of quantities.[94,170] Graph-based deep learning models have shown remarkable performance in molecular and crystal property predictions compared to other ML models.[81,92–94,171]

In addition to ML models, heuristics and advanced artificial intelligent algorithms can be particularly useful. Ensemble models combine prediction results from different models via, for example, majority voting in classification or averaging in regression. The aforementioned RF model belongs to this category, where many decision trees with different input features and model sizes are combined. A key assumption in

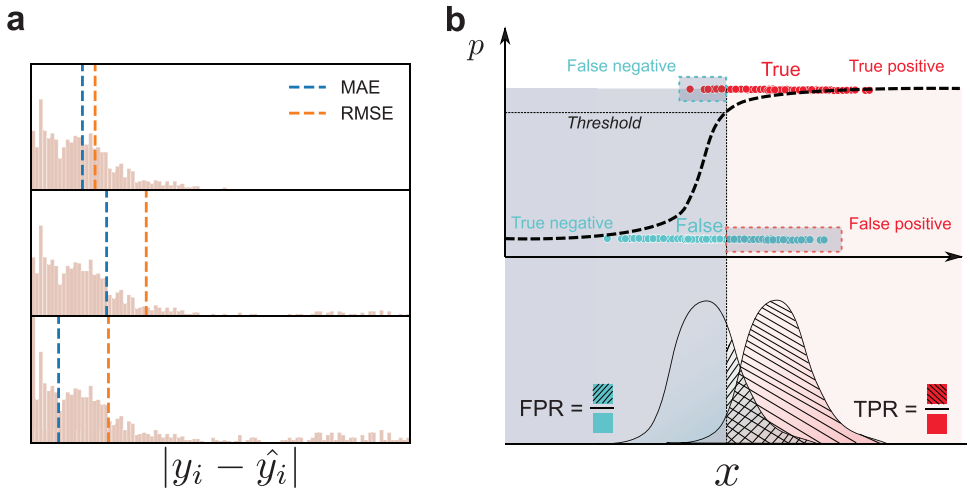

Figure 4. a) Regression metrics mean absolute error (MAE) and root mean squared error (RMSE). The RMSE value is always larger than MAE (top), but is more sensitive to large error data as seen by larger increase in RMSE when high error data points are added (middle). However, it is possible to increase RMSE and reduce MAE at the same time by proper error distribution comparing the top and bottom subplots. b) Binary classification with true negative, false negative, true positive, and false positive definition. A classifier predicts a probability (p) for each sample and a threshold value in the range (0, 1) is applied to the probability to determine confusion matrix. The false positive rate (FPR) and true positive rate (TPR) decreases with increasing threshold value.

ensemble models is that individual models or weak learners are independent and can capture different aspects of the data. Most importantly, all of them should be better than random guesses, e.g., binary classification accuracy greater than 50%. These models are generally more accurate and have low model variances. The ensemble methods have been applied in the construction of accurate ML models for predicting atomic local environment from K-edge X-ray near-edge structure (XANES).[172] or in various tree-based ensemble methods.[146–148]

Evolutionary models borrow the concepts from biological evolution. An evolutionary algorithm typically starts with a randomly generated population and then the fitness of each individual is evaluated. In each subsequent generation, the individuals with the highest fitness will be chosen to give offspring via crossover or mutation, until the best candidates are found. Evolutionary methods have been applied in automatically searching for new crystal structures[173] using the USPEX[174] and XtalOpt[175] codes. The CALYPSO software shares the same aim with USPEX but uses a different global optimization algorithm named particle swarm optimization.[176]

5.2. Model Loss, Metrics, and Training

During training, the weights/parameters of the model are adjusted iteratively to minimize the model loss, in response to training data.

For regression models, common model loss or prediction error metric are the mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE), defined as follows

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=0}^{N}\left|y_i - \hat{y}_i\right| \tag{7}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=0}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{9}$$

where N is the number of data points, and y_i and $\hat{y}_i$ are the actual and model-predicted values, respectively. The MSE/RMSE tends to give more weight to larger errors (**Figure 4**a) and therefore is more appropriate when larger errors are especially undesirable, e.g., in an interatomic potential. One common confusion is that MAE and MSE/RMSE are not monotonic with each other, and it is possible that one increases at the decrease of the other. Normally the loss function needs to be differentiable.

For classification models, a surrogate loss function to the accuracy metric is typically applied. For example, the cross-entropy loss is defined as

$$\ell(y,\hat{y}) = \sum_{i=1}^{n}(-y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})) \tag{10}$$

In a binary classification problem, the predicted labels and ground truth labels form a 2 × 2 confusion matrix (see **Table 1**), where a true positive (TP) refers to the correct report of the presence of a condition (e.g., predicted superionic conductor that turns out to be one) and a false positive (FP) is the incorrect report of the presence of a condition (e.g., predicted

Table 1. Confusion matrix for binary classifier.

	Predicted true	Predicted false
Actual true	True positive (TP)	False negative (FN)
Actual false	False positive (FP)	True negative (TN)

superionic conductor is actually a poor conductor). Similar definitions can be reached for false negative (FN) and true negative (TN). Common metrics for model performance include

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{11}$$

$$\text{False positive rate (FPR)} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{12}$$

$$\text{Recall, true positive rate (TPR)} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$\text{Specificity or selectivity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{14}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{15}$$

The direct prediction from a classifier is a continuous probability from 0 to 1. To predict a true/false label, a threshold in the range (0, 1) needs to be applied to the final probability and the threshold value impacts the distributions of the true/false labels (Figure 4b). Ideally, the model shall have a high precision and a high recall (small FN and FP). In reality, FN and FP compete with each other under different threshold values. To make a single metric from the confusion metric, the F_1 score is defined as the harmonic mean between precision and recall, i.e.

$$F_1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

With different thresholds, the FPR and TPR will also change. The relationship between FPR and TPR forms the receiver operating characteristic (ROC) curve and its area under curve (AUC) is typically used as a performance measure for binary classifiers. A classifier that performs no better than random guessing has an AUC of 0.5, while a perfect classifier has an AUC = 1. The Jaccard coefficient[177] is also frequently used as a metric and is defined as the intersection over union for the predicted and true classes set. For example, the Jaccard coefficient for the true label is (TP/(TP + FP + FN)). The average Jaccard coefficient for the binary problem is therefore (TP/(TP + FP + FN) + TN/(TN + FP + FN))/2.

In general classification problems, the model output can be a single label, or several labels simultaneously in a multilabel classifier (e.g., predicting the formability probability and the electronic insulating probability at the same time). Sometimes training a model with multitasks can simultaneously facilitate information exchange between tasks and thus enhance the model performance in individual task.[178] For each label, the possible categories (classes) can be more than two, in which the classifier becomes a multiclass classifier. Using conventional classification algorithms, such as logistic regression and linear SVC, the multiclass output can be achieved by training one model for each class using one-versus-all approaches. In a n-class classification problem, the confusion matrix becomes $n \times n$ and metrics such as accuracy, precision, recall, and Jaccard coefficients can still be defined similarly to the binary problem.

5.3. Model Selection

The "no free lunch" theorem[179] states that no learning algorithms are better than others if the model performance is averaged uniformly on all problems. This suggests that there is no single "best" ML approach for all materials problems, and in practice, the selection of the model has to fit the data and the prior knowledge and assumptions of a particular problem. Therefore, an exploratory data analysis step is suggested before model selection to map out the feature space and the feature correlation with the target values. During this step, unsupervised learning algorithms, for example, PCA, cluster analysis, and manifold learning, may further help the visualization. This step will be vital in helping understand the data distribution and intrinsic feature properties.

In practice, striking a good balance between model predictive power and simplicity can be challenging. The learning curve analysis is particularly useful in this context where training data is gradually increased and the model training errors and the hold-out validation errors are recorded. In common settings, the training errors will gradually increase with training data size, while it is the opposite for the validation errors. Ideally, we expect these two errors to converge to a relatively low error (low bias) and the differences between training and validation errors to be small (low variance). Underfitting occurs when the training errors converge to a relatively high value, suggesting a large bias and increasing model complexity may solve the problem. On the other hand, overfitting occurs when the model contains too many trainable parameters relative to the amount of training data, e.g., a polynomial function of power n can always fit n data points perfectly, but would generalize poorly with unseen data. Typical symptoms of overfitting are a large gap between training and validation errors and high model variance, i.e., different splits of the training data give models with very different prediction results. In particular, the data limitations in many materials domains make materials ML models susceptible to overfitting. Cross-validation (CV), where the data is split several times and the performance of the model is averaged across splits, is therefore important to ensure that any constructed ML models are reasonably general. One common CV strategy is the K-fold CV where the data are split equally into K non-overlapping folds and each time $K-1$ folds are used as train set and the remaining fold works as validation. It should be noted, however, that CV can also lead to overconfidence in the predictive power of ML models. Many materials problems involve the search for materials that possess rare combinations of properties or extraordinary properties, while such materials may not exist in the available data space. Hence a highly accurate ML model model trained on K-fold CV may not generalize well to novel material classes.[180]

In addition to the learnable parameters in the model, another set of parameters that are set before the model training is called the hyperparameters. The hyperparameters define model configuration and architecture. A typical model building practice is to split the data into training, validation and test data set.

The training data set is used in the learning process to optimize the model parameters. A grid search is often performed on the hyperparameters. Out of the available models, the one with the lowest errors on the validation set is chosen. Then the chosen model is evaluated on the test data set to obtain the true model performance. To make the maximum use of the data, the model with optimized hyperparameters may be refitted on the train-validation combined data and then evaluated on the test data set.[181]

The absolute performance of an ML model in terms of accuracy metrics is not the only or even the predominant consideration. In materials science, like in other scientific disciplines, substantial value is placed on the interpretability and simplicity of the ML models. Simpler, interpretable models provide insights that can guide future materials design. They are also less prone to overfitting and can be computationally less expensive (important in certain applications such as ML-IAPs).The ease and process in which insights can be extracted depends on the choice of ML models. For example, the coefficients in simple linear regression model as well as the logistic regression can provide an indication of the relative importance of different features. Feature importance analysis in tree-based models has been routinely used to assess the important physical parameters that are related to the predictive targets.[102,182–184] For deep learning-based models, visualizing the hidden layer activations has been found to give interpretable chemical intuition and accurate mapping of structural space.[107,185] While these methods are dependent on the specific choice of models, one can choose to measure the feature importance by assessing the model performance degradation upon removing certain features.[153]

5.4. Software Libraries

Fortunately, there are already many tools and software packages to aid in the development of ML models for materials science. Besides general purpose ML tools such as scikit-learn,[186] tensorflow,[187] and Pytorch,[188] there has been an explosion of customized open-source ML software libraries for materials science. A nonexhaustive list includes AutoMatminer,[189] PROPhet[190] for general materials ML; amp,[191] ænet,[192] and ANI[158] for developing neural network potentials; CGCNN,[93] MEGNet,[94] and SchnetPack[193] are graph-based deep learning model packages for accurate crystal and/or molecule property modeling.

6. Application

In this section, we will critically review recent applications of ML models to the design and discovery of energy materials. Before delving into the specific application domains, it is useful to provide a broad overview of the capabilities that ML models enable. A fundamental goal of all ML models is to provide cheap and reasonably accurate predictions that substitute for more expensive computational, experimental, or human-driven techniques. In doing so, ML models enable:

1. Accelerated discovery of novel materials by providing rapid predictions of properties and novel materials. As we shall see in the subsequent sections, a large number of materials ML works have prediction of materials stability (formation energies, energy above hull) as the primary goal or at least, a subgoal, along with key application-specific metrics (e.g., battery voltages and ionic conductivity, catalytic adsorption energies and activities, bandgaps, etc.). Also of interest are ML models that attempt to predict properties that are highly expensive to obtain via DFT calculations or experiments, i.e., elastic properties, phonons, etc. ML can even yield novel chemical insights that enable the development of improved structure prediction algorithms that generate better "guesses" for novel structures/compositions.
2. Accurate simulations of complex materials at larger length/time scales. Most real-world materials are not perfect bulk crystals or isolated molecules. Linear-scaling ML-IAPs have greatly enhanced our ability to perform dynamical simulations of complex materials systems—polycrystals, liquids, interfaces, etc.—while retaining close to DFT accuracy. Besides accuracy and scaling, arguably the greatest advantage of ML-IAPs is their potential for automated reproduction across different systems and can even be learned on-the-fly coupling with ab-initio molecular dynamics (AIMD).[194]
3. Enhanced characterization and interpretation. Finally, an application of ML models that has received far less attention in other reviews is their potential to enhance experimental characterization. The interpretation and labeling of experimental images (e.g., scanning transmission electron microscopy (STEM)[195]) and spectra (e.g., X-ray diffraction (XRD),[196] X-ray absorption near-edge structure (XANES),[172,197] nuclear magnetic resonance (NMR),[198] etc.) are today still mostly painstakingly carried out by humans. Adaptions of ML advances in computer vision and speech recognition can provide invaluable tools to accelerate this process.

6.1. Rechargeable Alkali-Ion Batteries

The rechargeable lithium-ion battery (LIB)[199–203] has proved to be a disruptive energy storage technology that powers our current digital age and is a leading candidate to power our electrified transportation. It is therefore no surprise that LIBs, as well as its analogues based on Na and other alkali ions, have been the subject of intense research, especially ML-driven design and discovery of materials.

6.1.1. Diffusion Properties

A rechargeable alkali-ion battery is an electrochemical device that operates by reversibly shuttling alkali ions between two electrodes through an electrolyte. Facile alkali ion conduction across all components and their interfaces is therefore a requirement for batteries that can achieve high rate capability and power.[204] However, this property is among the most difficult and expensive to determine via standard DFT-based approaches. Nudged elastic band (NEB) calculations of the alkali migration barriers are notoriously difficult and expensive to converge, while AIMD simulations are too expensive to extract reliable diffusion

statistics except for extremely high ionic conductivity superionic conductors.[205] While this limitation of first-principles methods presents significant opportunities for ML-driven acceleration, we shall see that it also presents substantial challenges in terms of the availability of training data.

Using high-throughput computations and ML, Jalem et al. have studied vacancy-mediated Li ion migration and the Li hopping energies in olivine $LiMXO_4$ (main group M^{2+}–X^{5+}, M^{3+}–X^{4+})[206] and tavorite $LiMTO_4F$ (M^{3+}–T^{5+}, M^{2+}–T^{6+})[123] cathodes. In olivine $LiMXO_4$, the authors revealed that the degree of M octahedron distortion increases with ionic size of the M cation, which in turn results in an energy-penalizing local lattice distortion around the migration pathway and a higher migration barrier (see **Figure 5**a). Based on these observations, the authors extracted 42 structural descriptors including structural parameters, Born effective charges of cations and intra/interpolyhedron parameters from DFT-relaxed structures and developed a partial least-squares (PLS) model to correlate the structural descriptors and NEB Li^+ hopping energies for 15 M–X pairs. The difference between DFT and PLS-predicted results was within 35 meV (see Figure 5b). In addition, the three most important descriptors identified via variable importance in the projection (VIP)[207] were associated with topology of M octahedron, namely, quadratic elongation of M octahedron, bond angle variance of M octahedron, and Li–O–M angle at edge sharing, respectively. Similar results and insights have been obtained by the same authors using neural networks and a causal index (CI) approach to extract important features[208] for tavorite $LiMTO_4F$ (M^{3+}–T^{5+}, M^{2+}–T^{6+})[123] (Figure 5c,d).

In recent years, all-solid-state alkali-ion batteries (SSABs) have experienced a resurgence of interest as a potentially safer and more energy dense alternative to traditional LIBs. The enabling component in SSABs is the superionic conductor solid electrolyte, which must have very high ionic conductivity and ideally, electrochemical stability against the electrodes. Fujimura et al.[141] has used SVR models to predict the ionic conductivity at 373 K σ_{373} of the LISICON-type superionic conductors with formula γ-$Li_{8-c}A_aB_bO_4$ (A = Zn, Mg, Al, Ga, P, As; B = Si, Ge). The training data comprises the diffusivity from AIMD simulations at 1600 K, D_{1600}, as well as the order-disorder phase transition temperature, T_{pc}, estimated by determining the temperature at which the DFT energies of the ordered and disordered phases are equal. Unsurprisingly, the model predicts that systems with high D_{1600} and low T_{pc} tend to have high σ_{373}. More recently, Sendek et al.[124] has developed a logistic regression model to predict superionic conductive behavior (0.1 mS cm^{-1} as threshold) in materials, utilizing features such as the average number of Li–Li neighbors for each Li, the average sublattice bond ionicity, the average anion coordination number in the framework, the average shortest Li-anion distances, and the average shortest Li–Li distances.

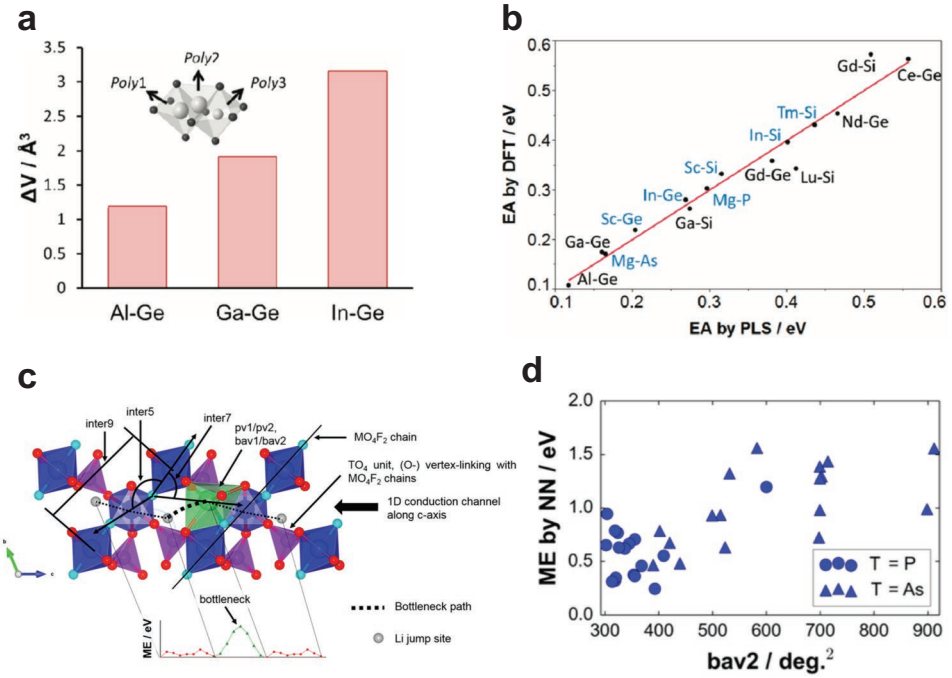

Figure 5. a) Local lattice distortion along migration pathway represented by volume difference ΔV of quasi path volumes between the transition state and initial state for M = Al, Ga, and In and X = Ge in M–X pairs. The quasi path volume is defined as the sum of the volumes of polyhedron Poly1, Poly2, and Poly3. b) Calculated hopping energies via NEB method versus predicted hopping energies via PLS model for different M–X pairs of olivine compositions. The blue data points represent experimental data in ICSD. Reproduced with permission.[206] Copyright 2012, American Chemical Society. c) Demonstration of structural features with high importance along the Li bottleneck pathway of tavorite structure. inter9: distance between the end member M cations of a MO_4F_2 chain; pv1: polyhedral volume of Li ion cage; bav2: bond angle variance of Li octahedron, indicating the degree of local lattice distortion. d) Neural network predicted migration energies versus bav2 for different covalent T cations. Reproduced with permission.[123] Copyright 2015, American Chemical Society. Reproduced with permission.[123,206] Copyright 2012 American Chemical Society and 2015 American Chemical Society.

This model was trained on 40 data points and then used in a screening workflow to screen an initial candidate list of 12 831 materials to 21 possible superionic conductors.

While the above examples show potentially interesting applications of ML to battery materials design, it is clear that data availability is a major limitation. None of the above works have more than 50 training data points (from DFT calculations), and in some instances, the small data sets are used in neural network models[122,123] that are notoriously data hungry. As such, it is difficult to establish confidence in the generalizability of these models to unseen data points. Moreover, a number of the models were developed for restricted structure types, e.g., olivine, tavorite, or LISICON, which substantially narrows the scope of their application. In fact, the total number of enumerated crystals in these structure types is sufficiently small that they are well-within capabilities of DFT calculations today, which limits the value of ML as a means for predictive exploration. Nevertheless, useful insights have been gained by identifying the key features contributing to diffusivity,[122,123] and these insights are more likely to be transferable to other structure types.

Using less expensive computational techniques allows the generation of a larger quantity of training data. For example, Nakayama et al.[209] trained PLS and GBR models on the Li migration energy in 400 Li-containing compounds computed using bond-valence force fields (BVFFs). While statistically more robust, it is unclear whether the source of the training data—BVFF calculations—are sufficiently accurate to yield useful predictions, i.e., data quantity may be sufficient, but data quality is in question. Furthermore, the value of such ML models in screening is not evident given that BVFFs are sufficiently inexpensive to run over thousands of potential candidates and indeed, have been applied in such a manner.[210]

Yet another alternative is to use adaptive learning approaches. In a recent extension to their previous work, Jalem et al.[211] have developed a Bayesian-driven approach to efficiently screen for fast-conducting Li- and Na-containing tavorites. The search space of 318 $AMXO_4Z$ tavorite covered all possible ionic substitutions of A, M, X, and Z sites. The initial sampled data set containing five randomly chosen compositions were used to train the surrogate Gaussian process model. The model posterior provided the predicted mean and standard deviation, which were then used in the acquisition function to find the next candidate for calculations. The maximization of acquisition function sets the balance between the exploitation and exploration and determines the next sampling composition in the search space for evaluation. The authors further incorporated an additive structure into the representation of feature space, decomposing the objection function into a sum of subsidiary functions with fewer dimensionally disjoint features because of the poor performance of Gaussian process caused by the high-dimensionality of feature space. In general, the additive Bayesian optimization showed the best performance and the ordinary Bayesian optimization surpassed the random search when the number of DFT evaluations larger than 20. In terms of finding the optimal composition, both additive Bayesian optimization and ordinary Bayesian optimization outperformed random search significantly.

Finally, another possible area to circumvent the need for expensive data generation is to use unsupervised learning or related techniques that do not require excessive amount of target data. Recently, Zhang et al.[212] have developed a modified XRD (mXRD) feature to describe the anion lattice of Li-containing compounds. Their results have shown that compounds that have similar conductivity tend to be close in mXRD feature space. The developed methods are able to identify 16 compounds with conductivities of 10^{-4} to 10^{-1} S cm^{-1}.

6.1.2. Mechanical Properties

Unlike diffusion properties, predicting elastic properties presents a more tractable problem given the availability of precomputed DFT data (≈13 000 elastic tensors at the time of writing) in the Materials Project.[213] Furthermore, substantial successes have already been demonstrated in the prediction of elastic moduli using graph-based deep learning methods.[93,94] Ahmad et al.[214] have leveraged on the CGCNN framework,[93] gradient boosting regression (GBR), and KRR to screen desirable solid electrolytes and interfaces for suppressing dendrite initiation in contact with Li metal anode. To achieve stable electrodeposition (i.e., suppression of dendrite formation), the interface needs to be stabilized with suitable solid electrolyte as well as particular orientations of Li metal and electrolyte forming the interface. Hence, the screening task was composed of two parts: isotropic screening for candidate solid electrolytes and anisotropic screening for candidate orientations of both Li metal and electrolytes. The authors used stability parameter χ, a quantified representation of dendrite initiation, as the criteria for screening. χ is a function of shear modulus G_s, Poisson's ratio ν_s, and molar volume ratio V_M of a solid electrolyte. By training CGCNN models using 2041 structures with moduli from the Materials Project,[10] the authors predicted the shear moduli and stability parameters for 12 950 Li-containing compounds. The screening results however indicated that none of the materials could be stabilized without the aid of surface tension but some candidates had desirably low stability parameters and high critical wavelength of roughening λ_{crit}. For anisotropic contact, an interface was characterized by the directions of Li metal normal to it and the low-index facets of the solid electrolyte. During the screening process, the authors trained GBR and KRR using the DFT elastic tensors for 2401 unique surfaces of 482 electrolyte materials obtained by applying rotating axes and transformation rules[215] to full elastic tensors from Materials Project,[10,213] and then predicted the elastic tensors of 548 cubic crystal structures with each having 3 unique surfaces as well as the stability parameters corresponding to 4 distinct directions of Li metal for each surface. The models predicted twenty dendrite-suppressing interfaces formed by four solid electrolytes (Li_2WS_4-$P\bar{4}2m$, $LiBH_4$-$P\bar{1}$, LiOH-$P4/nmm$, and Li_2WS_4-$I\bar{4}2m$) and Li.

6.1.3. Machine Learning Interatomic Potentials

Finally, ML has been used to developed ML-IAPs for dynamical simulations of battery materials. One of the earliest ML-IAPs is

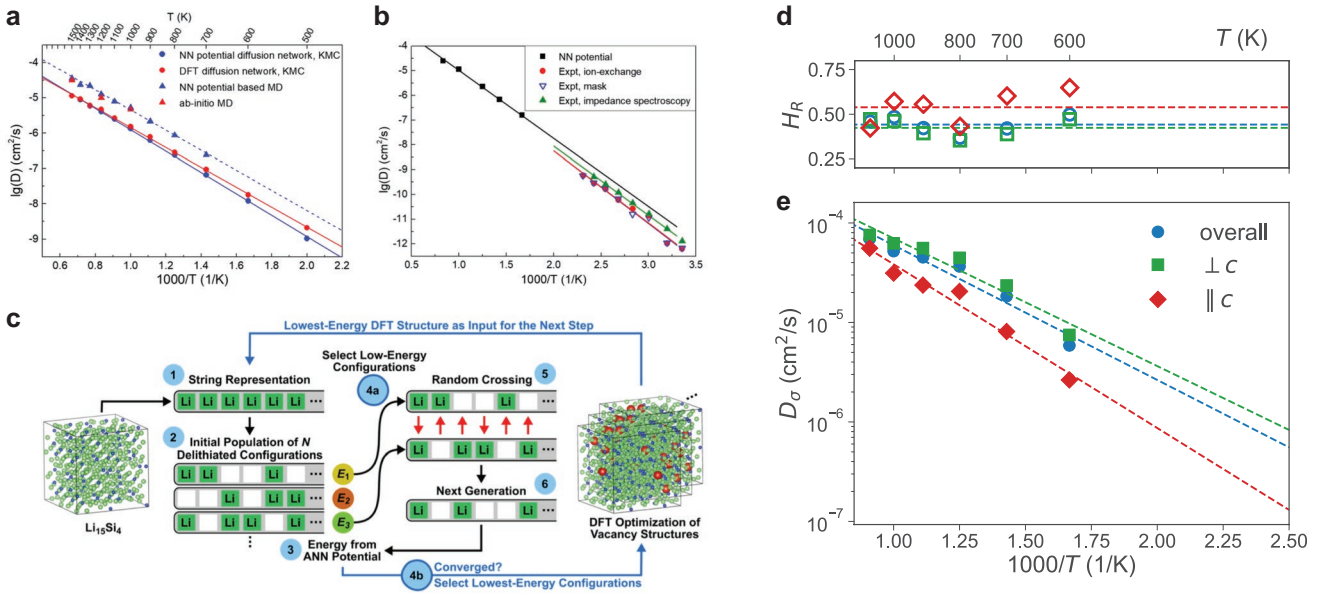

Figure 6. a) The Arrhenius plot of Li diffusion coefficients obtained from KMC and MD simulations with NNP and DFT, respectively. b) The Arrhenius plot of Li diffusion coefficients in amorphous Li_3PO_4 from large-scale MD simulations.[216] c) Schematic workflow of sampling strategy combining GA and specialized NNP. At each delithiation step, GA was used to identify the most stable Li/vacancy arrangement of that composition, with specialized NNP determining the energetics of different arrangements.[217] d) The predicted Haven ratio and e) the Arrhenius plot of Li diffusion diffusivity in α-Li_3N from eSNAP MD simulations.[131] Reproduced with permission.[216,217] Copyright 2017 American Institute of Physics and 2018 American Institute of Physics. d,e) Reprinted under the terms of the CC-BY license.[131] Copyright 2019, The Authors.

the neural network potentials (NNPs) of Behler and Parrinello[85] which expresses the potential energy surface as a function of atom-centered symmetry functions (ACSFs) representing radial and angular terms. Li et al.[216] has developed an NNP for amorphous Li_3PO_4, a prototypical solid electrolyte material for LIBs, using 38 592 DFT-calculated reference configurations. The Arrhenius plots of Li diffusion coefficients obtained from both kinetic Monte Carlo and molecular dynamics (MD) simulations with the NNP are in excellent agreement with those obtained from DFT simulations (**Figure 6**a). The authors have also applied the NNP to large-scale MD simulations of Li diffusion in 1006-atom amorphous Li_3PO_4 and showed that the predicted activation barriers and diffusivities are in excellent agreement with experimental measurements (Figure 6b).

The neural network model used in the NNP generally requires large data sets for optimal performance.[152] Artrith et al.[217] have proposed an iterative optimization strategy combining genetic algorithm and an NNP (schematic provided in Figure 6c), which only required ≈1000 DFT reference data. At each step, the NNP was used to identify the energetics of different configurations at specific composition of delithiated amorphous $Li_{15-x}Si_4$, based on the assumption that it was able to sample the near-ground-state Li/vacancy arrangements, and the genetic algorithm was applied in finding the optimal configuration. For every delithiation step, the 30 most promising configurations predicted by genetic algorithm were optimized by DFT method and the most stable configuration was used as the starting structure in the next delithiation step. The sampled structures by genetic algorithm were within the range of 100 meV per atom above the lowest energy structure for each composition, indicating its success in the determination of low-energy metastable amorphous structures of Li_xSi, which then allowed the phase diagram for amorphous Li_xSi to be constructed. The same authors have also developed NNP-type models incorporating compositional descriptors that are able to predict the energies of eleven-species cation-disordered lithium transition-metal (TM) oxides $LiMO_2$ (M = Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu) to within 3 meV per atom.[159]

The alternatives to the ACSF features are those based on a direct featurization of the local atomic neighbor density function.[76,218] The spectral neighbor analysis potential (SNAP) fits the potential energy surface to a linear or quadratic model of the coefficients of the bispectrum of local atomic density functions.[107,131,218–220] A particular challenge in IAP development for battery materials—many of which are ionic compounds—is the treatment of long-range electrostatics. Recently, Deng et al.[131] has augmented the linear SNAP approach with an electrostatic term and developed an eSNAP model for Li_3N, one of the earliest discovered lithium solid electrolytes with anisotropic Li diffusion mechanism in different crystallography orientations. The authors have demonstrated that the eSNAP model far outperforms the traditional Buckingham potential in predicting several key properties of Li_3N. Applying the eSNAP in large-scale simulations, the authors were able to compute the Haven ratio for Li_3N to excellent agreement with NMR experiments and show that the twist grain boundaries of Li_3N exhibit rapid Li diffusion even at room temperature (Figure 6d,e).

In addition to the studies on Li compounds, the modeling of dynamical intercalation of Li atoms into the electrodes (i.e., guest atoms in host frameworks) is also of great interest. Fujikake et al.[221] recently have introduced a newly developed GAP model by fitting the energy and force differences induced

by Li intercalation in the graphitic and amorphous carbon structures, one of the most commonly used anode in LIBs. The intercalation energy and force were used to construct the Li–C interactions in the addition to the GAP-modeled PES of pure element carbon[222] and an effective Li–Li interaction term was extracted separately to fit a two-body GAP pair potential, which accounts for long-range behavior. The authors have demonstrated that the GAP model is able to reproduce the energy profiles of the adsorption of a Li atom on high-symmetry sites of a graphene sheet and qualitatively identify the diffusion pathways of a Li atom through pristine graphite. They also showed that the radial distribution function and vibrational densities of states from GAP-MD simulations correspond well with DFT-MD results, albeit with a notably high Li intercalation energy MAE of 0.29 eV per atom.

6.1.4. Summary

From the above, it is evident that the majority of ML efforts in rechargeable lithium-ion batteries have been focused on alkali diffusion, e.g., ionic conductivity or migration barriers. While a critically important property for this application, diffusion properties are difficult to obtain reliably in large quantities—ab initio computation-based approaches such as AIMD and NEB are too computationally expensive and are applied mainly on small idealized cells, while experimental measurements are highly sensitive to synthesis and measurement conditions. Other important properties, such as voltages, elastic moduli, etc., are easier to obtain reliably via high-throughput computations. To date, there are relatively few ML works that target these[214,223] and other properties for the purposes of screening, which presents major opportunities for further exploration. More promisingly, ML-IAPs are emerging as a powerful new tool that enable long-time scale simulations of large systems at near-DFT accuracy, providing critical atomistic scale insights into the phase transformation pathways and diffusion processes in battery materials.

6.2. Photovoltaics

Solar energy is the most abundant clean energy source. Photovoltaics (PVs), which convert sunlight directly to electricity using semiconducting materials, is the most direct way of utilizing solar energy. The key performance metrics of a PV are its long-term stability and solar conversion efficiency (SCE), i.e., the percentage of energy in the form of sunlight that is converted into electricity, together with practical cost considerations. Although solar panels based on traditional semiconductors such as Si and GaAs have taken off commercially, there remains much interest in discovering new PV materials that are cheaper and have higher efficiencies. In particular, perovskite-based materials, hybrid organic–inorganic as well as inorganic, are being studied intensely at the present moment due to their high efficiencies and low fabrication and materials cost.[224] The SCEs of hybrid organic–inorganic perovskites (HOIPs) have risen rapidly from 3.8% in 2009[225] to over 22% in 2019.[226] The major limitation of current HOIPs is their instability, especially when exposed to moisture, light or heat.[227] Most ML studies of perovskites have therefore focused on predicting stability and the SCE.

6.2.1. Perovskite Stability

Perovskites with general formula ABX_3 where A and B are cations and X is the anion (usually oxide or halides), are among the most well-known crystal structures. Historically, the formability of perovskites is predicted using simple atomic radii arguments and descriptors such as the tolerance factor $t = \frac{r_A + r_B}{\sqrt{2}(r_B + r_O)}$, where r_A, r_B, and r_O are the ionic radii of A, B and O, respectively,[98] and the octahedral factor $\mu = r_B/r_O$.[228,229] Several ML works have begun to revisit these well-known descriptors in light of the availability of more experimental as well as computed data on perovskites. For example, Sun and Yin[230] have shown that $(\mu + t)^\eta$, where η is the atomic packing fraction, showed a better linear relationship with the thermodynamic stability of 138 perovskites. More recently, Bartel et al.[96] have proposed a new tolerance factor τ encompassing both halide and oxide perovskites after SISSO[101] feature selection as the following

$$\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\log(r_A/r_B)}\right) \tag{17}$$

where n_A is the oxidation state of A. Using the criteria of $\tau < 4.18$ for stable perovskites, this new tolerance factor can separate perovskites versus nonperovskite in an experimental data set of 576 ABX_3 with an accuracy of 92%.

Most ML works on halide perovskites have thus far focused on inorganic perovskites, due to the difficulty in modeling rotational disorder in organic cations such as methylammonium. Im et al.[103] have used DFT to compute the heat of formation and bandgaps of 540 Pb-free halide double perovskites ($A_2B(I)B(III)X_6$) and subsequently used the data to develop gradient boosting regression trees (GBRT) models. The best achieved RMSE on heat of formation is 21 meV per atom, and that of GGA bandgap is 0.223 eV. Li et al.[231] have worked on the same type of perovskites. The authors generated DFT-calculated decomposition energies of 354 halide double perovskites, and the data was used to train KRR models. The best model achieved RMSE of 34 meV per atom on test set using the ionic radii of constituent elements as descriptors, which is about 10 meV per atom lower than the RMSE achieved by only using tolerance factor τ[98] or revised tolerance factor $(\mu + \tau)^\eta$ proposed by Sun and Yin.[230] Using data of 185 experimentally known ABX_3 halide perovskites, Pilania et al.[232] have also developed SVC classifiers for perovskite formability using 11 structural features. The authors found that the best performing model with >92% accuracy in classifying perovskite formability requires only four of the features, namely the Shannon's ionic radii of A-, B-site atoms, tolerance factor, and the octahedral factor, indicating that the steric and geometric packing played a dominant role in deciding the stability of halide perovskites.

6.2.2. Solar Conversion Efficiency

The Shockley-Queisser limit states that single-junction solar cells with an optimal bandgap of 1.34 eV have a maximum SCE of 33%.[233,234] Therefore, the bandgap prediction or screening is often an initial step for the computational design of solar

cell materials. While DFT calculations based on standard semi-local exchange–correlation functionals such as the Perdew–Burke–Ernzerhof (PBE) functional[235] are well-known to underestimate the bandgap substantially,[236,237] the errors are relatively systematic and computed PBE bandgaps often form the training data for ML models.

Allam et al.[238] have calculated the PBE bandgaps for 14 halide/oxides perovskite systems in layered Ruddlesden–Popper phase with the number of layers ranging from 2 to 5 and developed a neural network model to predict the bandgap. The selected descriptors included the inverse of the number of layers, ionic radii and oxidation state of atoms on each site. The results showed that the correlation, slope, and intercept between the neural network predicted bandgap and DFT PBE gap were 0.999, 0.993, and 0.0089 eV, respectively. Prediction of PBE bandgaps has also been attempted attempted by Lu et al.[121] for hybrid perovskites. Bandgaps of 212 hybrid perovskites were used to train a GBR model with selected 14 material features including structural factors (tolerance factor and octahedral factor) and elemental properties. The best model achieved an MSE of 0.085 eV. The model also pointed out that the tolerance factor was the biggest affecting factor on the bandgap and that B site properties played a bigger role than those of A and X sites. Using the trained model, the authors predicted 5158 unexplored possible hybrid perovskites and identified six orthorhombic lead-free hybrid perovskites as potential candidates for solar cell applications, including $C_2H_5OInBr_3$, $C_2H_6NInBr_3$, $NH_3NH_2InBr_3$, $C_2H_5OSnBr_3$, NH_4InBr_3, and $C_2H_5NSnBr_3$.

More accurate bandgaps can be obtained computationally using more expensive methods such as hybrid functionals (HSE06)[239,240] or GW calculations.[241] Agiorgousis et al.[242] have calculated 220 double chalcogenide perovskites (A2BB′X6) with the screened hybrid HSE06 functional, followed by training RF and SVM classifiers using this data set. The percentage error of classifying whether the bandgap of given perovskite falls between 0.7 to 2.0 eV was 13.80 % for RF, and 31.63% for SVM. The error could be further reduced to 13.28% and 16.7% for RF and SVM, respectively, by excluding A cation features. The authors explained that the bandgap variation range was only 0.05–0.3 eV when altering A site cation, which was much smaller compared to altering B and X site elements. This is due to the fact that the valence band maximum states are dominated by chalcogenide p states (X site) and conduction band minimum states are predominantly metal d states (B/B′ site). The authors then identified Ba_2AlNbS_6, Ba_2GaNbS_6, Ca_2GaNbS_6, Sr_2InNbS_6, and Ba_2SnHfS_6 as potential solar cell materials. A cheaper alternative to hybrid functionals is the GLLB-SC functional,[243,244] which allows larger data sets to be generated. Pilania et al.[245] have computed the GLLB-SC bandgaps for 1306 unique double perovskite oxides. A KRR model was trained on this data set with 16 selected features that included elemental properties and the best achieved RMSE was 0.36 eV. The same group[246] have more recently improved this model using an innovative data fusion approach using multi-fidelity modeling, where the HSE06 and PBE bandgaps are treated as high-fidelity and low-fidelity estimates, respectively. Using a two-level co-kriging model, the authors approximated the high-fidelity results by multiplying the low-fidelity results with a scaling factor plus an independent Gaussian process. The data set was comprised of 599 inorganic halide double perovskites with PBE bandgaps, and 250 of which were calculated with HSE06 as well. The authors found that MAE of HSE06 predictions can be reduced from 0.45 to 0.10 eV by varying the portion of low and high fidelity data in the training set (**Figure 7**).

The Shockley–Queisser relationship between bandgap and SCE[233] leads to many false positives (poor PVs with optimal bandgaps) and false negatives (good PVs with nonoptimal bandgaps). Yu and Zunger[247] have proposed a metric known as the "spectroscopic limited maximum efficiency" (SLME) to help initially screen potential PV materials. This metric considered several factors, including i) the existence of various energetic sequences of dipole-allowed, dipole-forbidden and indirect bandgaps, ii) the absorption shape near the threshold,

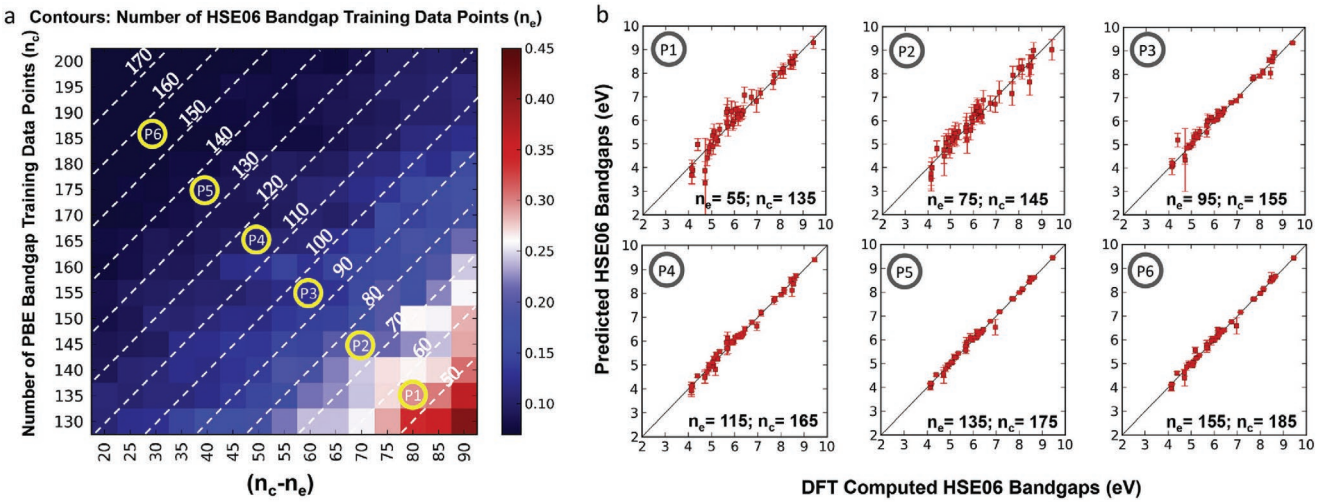

Figure 7. a) Contours of validation MAE of the multifidelity ML model as a function of number of low- (n_c) and high-fidelity (n_e) data points. b) Parity plots of predicted versus DFT-computed HSE06 bandgaps for selected combinations of n_c and n_e. Reproduced with permission.[246] Copyright 2017 Elsevier.

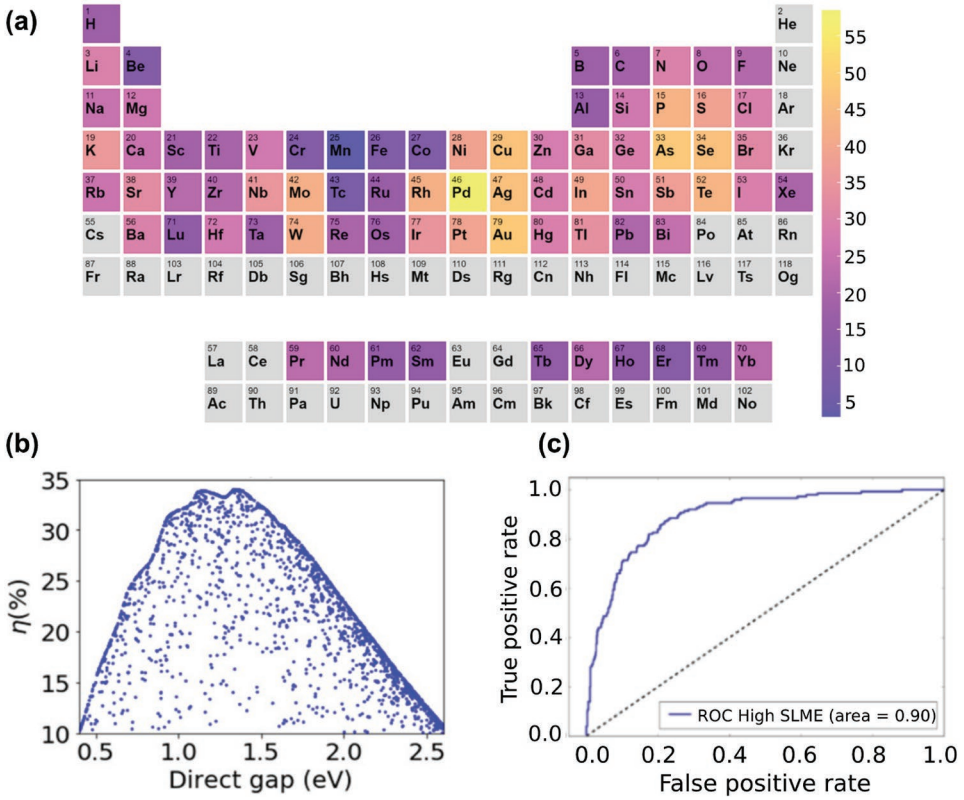

Figure 8. a) Element distribution for high-SLME materials, b) the SLME distribution with direct bandgap values, and c) ROC curves for a GBDT model. Adapted with permission.[120] Copyright 2019 American Chemical Society.

and iii) the dependence of radiative recombination losses on the energy separation between the minimum gap and dipole-allowed gap. Using 256 GW calculations, the authors were able to identify in this group high-SLME materials including almost all contemporary PV absorber materials. Choudhary et al.[120] have performed meta-GGA calculations of bandgaps using the Tran–Blaha-modified Becke–Johnson (TBmBJ) potential[248] for 12 881 out of ≈30 000 materials in the JAVIS-DFT database.[59] The SLME metric was calculated on 5097 nonmetallic materials and the elemental distribution for high-SLME materials is shown in **Figure 8**a. The distribution indicates that Pd is the most common element for high-SLME materials. The SLMEs with direct bandgap values are shown in Figure 8b, where a consistent volcano shape as original SLME results using GW[247] is obtained. A threshold of 10% was chosen to label high-SLME materials versus low-SLME materials, and the binarized data formed the training data set for ML modeling.[120] The authors then used CFID as structure features and compared various classification models in terms of the classification AUC. The AUC for decision tree models, RF, kNN, MLP, and gradient boosting decision trees (GBDT) are 0.67, 0.79, 0.77, 0.80, and 0.87, respectively. After hyperparameter tuning, the AUC for GBDT reached 0.90, as shown in Figure 8c. Using this classification model, the authors narrowed down target candidates from an initial pool of 1 193 972 structures from Aflow,[54] Materials Project,[10] OQMD,[56] and crystal open database,[34] to only 6342 with unique compositions. Future studies are necessary to confirm some of the predictions by experimental measurements.

6.2.3. Organic Photovoltaics

Organic PVs (OPVs) have also garnered much scientific and economic interest in the last decade. An OPV cell uses π-conjugated semiconducting organic molecules, oligomers, or polymers for light absorption and charge transport. Compared to their crystalline inorganic counterparts, OPVs have the advantages of light weight, low cost, flexibility, and facile fabrication. However, the biggest limitation of OPVs is the SCE, which is normally less than 10%, and the highest efficiency of 17.3% has been reached only recently.[249] The SCE of OPVs is heavily dictated by the optical gap of the acceptor, the energetic alignment of the lowest unoccupied molecular orbital (LUMO) of the acceptor and the highest occupied molecular orbital (HOMO) of the donor. The Scharber model[250] for calculating SCE from the frontier orbital energies has been widely used in the OPV field and has been optimized for [6,6]-phenyl-C61-butyric acid methyl ester (PCBM) acceptors. Large data sources for the frontier orbital energies exist from prior high-throughput databases, such as the Harvard Clean Energy Project Database[63] and the Harvard OPV data set (HOPV15).[251]

Pyzer-Knapp et al.[252] have developed MLP models to predict HOMOs, LUMOs, and SCE of molecules. The authors explored the possibility of replacing steps in a high-throughput virtual screening workflow with ML surrogate model predictions. Using 1024-bit Morgan circular fingerprints,[83] the surrogate MLP models were trained on calculations of 200 000 molecules from Harvard Clean Energy Project[63] and validated on

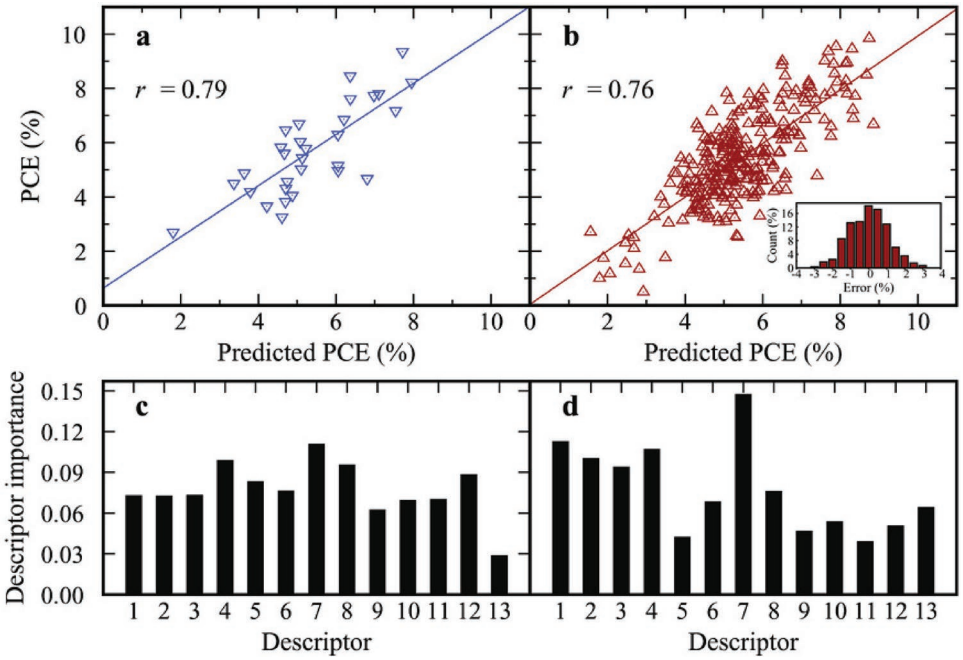

Figure 9. Prediction versus experimental SCE on a) test set (30 molecules) and b) all data points using leave-one-out CV technique. Inset shows the distribution of errors. c,d) The feature importance of GB and RF model, respectively. Adapted with permission.[182] Copyright 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

50 000 other molecules. The error on SCE was 0.28%, and the errors on HOMO and LUMO were 0.028 and 0.032 eV, respectively, well within chemical accuracy of 0.043 eV (1 kcal mol^{-1}), and were substantially lower than previously ML models using randomized Coulomb matrix features.[253] Only 0.8% of molecules were found to be promising OPVs with predicted SCE above 8%. A recent work by Padula et al.[254] further explored the factors that affect the SCE of molecules using linear and nonlinear ML. The authors found that using structural or electronic features alone yielded similar predictive results, while using them both with a KRR model led to increased performance.

While the Scharber model is widely used for computing SCE, Sahu et al.[182] have realized, from a data set of 280 small molecule OPV systems, that for all high-performance devices, frontier orbitals of donor molecules are nearly degenerated and therefore orbitals other than just HOMO and LUMO should be considered. The authors considered 13 quantum-mechanical descriptors including 1) number of unsaturated atoms in the main conjugation path of donor molecules(N^{D}_{atom}), 2) polarizability of donor molecules, 3) the energetic differences of LUMO and LUMO+1 of donor molecules (Δ_L), 4) the energetic differences of HOMO and HOMO–1 of donor molecules(Δ_H), 5) vertical ionization potential of donor molecules (IP(v)), 6) reorganization energy for holes in donor molecules (λ_h), 7) hole–electron binding energy in donor molecules(E_{bind}), 8) the energetic difference of LUMO of donor and LUMO of acceptor (E^{DA}_{LL}), 9) the energetic difference of HOMO of donor and LUMO of acceptor(E^{DA}_{HL}), 10) energy of the electronic transition to a singlet excited state with the largest oscillator strength(E_g), 11) change in dipole moment in going from the ground state to the first excited state for donor molecules(Δ_{ge}), 12) energy of the electronic transition to the lowest-lying triplet state(E_{T_1}), and 13) the energetic difference of LUMO and LUMO+1 of acceptors (Δ^{A}_{L}). They generated a data set of 280 experimental systems and trained several models including linear regression, kNN, artificial neural networks, RF, and gradient boosting (GB). The GB model was shown to outperform others with Pearson correlation coefficient r of 0.79 and RMSE of SCE being 1.07% on the test set. The feature importance analysis of the GB and RF models indicates that Δ_H is among one of the highly ranked features, indicating the necessity to consider more than just frontier orbitals (**Figure 9**). However, some of the descriptors can be expensive to obtain. Subsequently, Sahu et al.[255] have developed reduced-cost ML models that utilize the number of heteroatoms in place of the polarizability. Using GB and NN models trained using 300 newly reported small-molecule OPVs, 126 with predicted efficiencies larger than 8% were proposed out of 10 170 candidate molecules.

As noted earlier, a fundamental data limitation in the application of ML to PV materials (inorganic or organic) is the well-known underestimation of the bandgap with semilocal DFT functionals. Several ML efforts have therefore sought to close the gap between computational and experimental values. For example, Pyzer-Knapp et al.[256] have used a Bayesian approach with Morgan circular fingerprint (512-bit) features to eliminate the functional dependence of orbital energies. Similarly, Lopez et al.[257] have developed a GPR model to calibrate HOMO and LUMO between calculated and experimental values for 51 000 potential nonfullerene acceptors. The RMSE against experimental values of uncalibrated HOMO calculations was 0.28 eV and reduced to 0.17 eV after ML model calibration, and LUMO RMSE reduced from 0.45 to 0.26 eV, almost a factor of two in error reduction. The calibrated HOMO and LUMO were then used to calculate Scharber SCEs of potential molecules

and new nonfullerene acceptors diketopyrrolo-pyrroles and quinoidal thiophene derivatives were identified to be better than fullerene. Finally, Paul et al.[258] have used a transfer learning approach to improve agreement between predictions and experiments. In this approach, the weights of a customized ensemble deep neural network architecture trained on the computed Harvard CEP database were transferred to an experimental data set of only 243 molecules, substantially reducing the mean absolute percentage error (MAPE) from 2.782% to 1.513%.

6.2.4. Summary

Predicting stability, bandgaps, and SCE has been the main focus for ML in PV applications. For realistic energy materials applications, the stability and materials synthesizability should always be the first concern. Oddly, stability predictions in the PV field has been carried out mainly in target space groups or crystal families (e.g., perovskites), which enables the use of elemental/compositional features as descriptors. The use of elemental/compositional features has the advantage of simplicity, but ultimately, the synthesizability of a particular phase (structure and composition) is related to the energies of competing phases in the phase diagram. There has been no application of general purpose formation energy prediction models[93,94,259] that have already demonstrated relatively high accuracies across diverse chemical spaces for stability predictions in PV.

Ultimately, the current bandgap, and consequently, SCE, models in PV are constrained by well-known limitations of semilocal DFT. While more accurate methods such as the HSE functional[239,240] and the GW method[241] exist, they are computationally much more expensive. One possible future direction is to combine various sources of computational data by data fusion and multifidelity modeling. On the experimental side, high-throughput measurements of materials optical properties are providing a large quantity of data that may be used for ML purposes.[260] We are likely to see more efforts on experimental automation and high-throughput works in the near future.[261]

In addition to the bandgap, the prediction of full band structures has only been attempted on a limited set of materials, e.g., Si.[262] The full band structure provides much richer information for the material and will be a key quantity for future ML predictions. Other related properties, such as the carrier effective mass and dielectric response, are also to be explored using ML predictions to complement existing works.

6.3. Catalysts

Finding efficient and economically viable catalysts are key to furthering the adoption of many renewable energy technologies, including fuel cells and conversion of CO_2 to liquid fuels. Catalyst design generally follows the Sabatier principle, which states that the interaction between the reactant and the catalyst should be neither too strong nor too weak for optimal activity. This qualitative rule is typically represented as a "volcano" plot of activity as a function of binding energy between the reactant and catalyst[70,71,90,263] Despite significant progress in recent years, current catalysts are still either too expensive, e.g., precious metals, or need a high overpotential to drive the reaction. Finding a catalyst with low overpotential and cost becomes essential in catalyst materials design. ML methods have been used in catalyst design since the 1990s,[264–266] and are now a resurgence of interest and being applied more broadly to a large number of systems.[267–271]

6.3.1. Absorption Energies

One of the main computational tools for studying catalysis is the d-band model, which relates the d-band center of metal surfaces to the bonding formation and reactivity of transition metals;[70,90] a higher d state energy corresponds to more empty antibonding states and thus stronger bonding between the adsorbents and the surface.[272] However, obtaining the d-band center still requires performing the time-consuming DFT calculations, limiting its capability for large-scale materials screening. Takigawa et al.[183] have attempted to predict the d-band center using ML model on a data set containing 11 metals (Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Ir, Pt, and Au) and 110 metal pairs for secondary metals as surface impurity and overlay layer respectively. The d-band centers were obtained from the most close-packed surfaces of the corresponding metals ((111) for fcc, (001) for hcp, and (110) for bcc).[70] The authors used a set of nine elemental properties to represent each element type, including group number, bulk Wigner–Seitz radius, atomic number, atomic mass, period, electronegativity, ionization energy, enthalpy of fusion, and density at 25 °C. The feature vectors for structures were a simple concatenation of two types of elemental vectors. Such features were used in a GBR model that showed an RMSE of <0.5 eV in predicting the d-band center on the test data (75% of the whole data size). Based on the feature importance, the authors further reduced the 18 descriptors to just 6. While the achieved errors were around 10% of the d-band center range, it should be noted that these models cannot distinguish between surface types given that the descriptors contain no structural/surface information. Similarly, Meyer et al.[273] have predicted the energy of oxidative addition process between a transition metal complex and a substrate for C–C cross-coupling reactions and used this energy as a descriptor to estimate the activity of transition-metal complexes as homogeneous catalysts via a molecular volcano plot. With this descriptor and model, the authors performed screening of 18 062 homogeneous catalysts and identified 37 promising low-cost complexes that were derived from palladium and copper.

Catalysis is fundamentally a surface-driven phenomenon, and the type of surface termination and active sites have a large effect on catalytic activity. Ma et al.[274] have used artificial neural networks combined with atom projected electronic properties to predict the adsorption energies of CO on metal alloys. The authors noted that the d-band theory-based two-level models had a large error of 0.33 eV, which is not accurate enough for screening optimal catalysts. They went on to use electronic properties of clean surfaces, including filling, center, width, and kurtosis of d-states distributions and local Pauling electronegativity as primary features and host metal-depending physical constants as secondary features for the ML neural

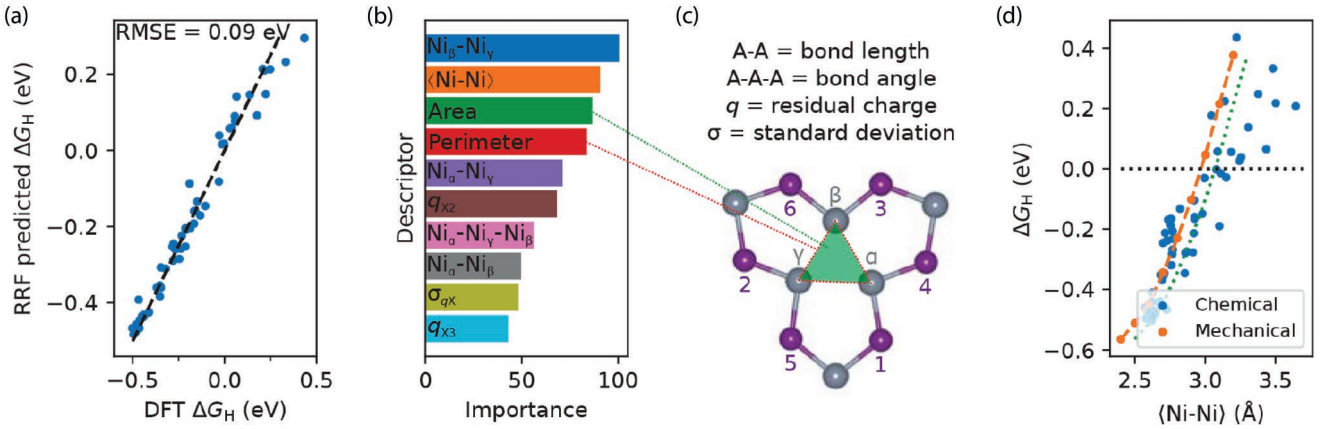

Figure 10. a) Parity plot of H binding free energy by RRF and DFT. b) Top 10 feature importance of descriptors obtained from RRF models. c) Geometry of Ni_3-hollow site and the descriptor visualization. d) The correlation between average Ni–Ni bond distance and the H binding free energy induced by chemical and mechanical pressure. Reproduced with permission.[102] Copyright 2018 American Chemical Society.

network model, yielding an error of 0.12 eV. This low error is within the screening energy range of potential catalysts. Similar features have been used in the prediction of CO and OH adsorption on bimetallic surfaces to identify several transition metal alloys and local environments with theoretical performance better than Pt in direct methanol fuel cells.[275] Similarly, Gasper et al.[276] have studied the CO adsorption on the Pt nanoparticles using a ML approach. One of the obstacles of studying the Pt nanoparticles is that the low symmetry nanoclusters tend to be more energetically stable. The authors adopted a genetic algorithm using a bond-order potential to automatically find such low energy structures, which were then confirmed by DFT calculations. A GBR ML model was used for predicting the CO adsorption energy on Pt nanoclusters using three types of descriptors, including the electronic structure descriptors calculated from DFT (averaged d-band center), structural information (average nearest-neighbor bond length, generalized coordination number, cluster radius of gyration, etc.), and the CO-frozen adsorption energy. The authors showed that with adding all features, the prediction error could be reduced to 0.12 eV compared to 0.23 eV when only the d-band center was used. All presented work have shown that site descriptors with ML models predict the adsorption energy better than the d-band theory alone. However, these accurate models require descriptors calculated from DFT, slowing down the screening process substantially. To further accelerate the catalyst discovery and to circumvent DFT calculations, Noh et al.[277] have used d-band width of the muffin-tin orbital theory and the electronegativity as alternative descriptors combined with neural networks and KRR models to predict CO adsorption on alloys. These models achieved a very low MAE of 0.05 eV after a training process involving active learning. Using this model, the authors found Cu_3Y@Cu* as an effective CO_2 reduction catalyst lowering the overpotential by 1 eV compared to the precious metal Au.

ML models can also provide useful catalyst design insights. Wexler et al.[102] have predicted the hydrogen evolution reaction (HER) activity on nonmetal doped Ni_3P_2 termination of $Ni_2P(0001)$ surfaces. The H binding energy on the pristine surfaces was too strong for catalysis and the authors found that by doping the surface with nonmetal elements (As, B, C, N, O, S, Se, Si, and Te) the binding energy of H can be tuned. A list of descriptors from DFT-relaxed structures was compiled including Ni–Ni bond lengths, Ni–Ni–Ni bond angles, Löwdin charges, elemental data, and their summary statistics, and other geometric parameters. These descriptors were used as features in a regularized random forests (RRF) model for predicting the H adsorption free energy ΔG_H. The model RMSE was only 0.09 eV using threefold CV on 55 observations and 29 descriptors (**Figure 10**a). The model-derived feature importance pointed out that the top two descriptors were a particular Ni–Ni bond length and the average Ni–Ni bond length (Figure 10b,c). In addition, seven out of the top ten were related to the geometry of the adsorption Ni_3-hollow site. These results suggested that the chemical pressure plays a critical rule in altering the catalytic activity. On this basis, a Ni–Ni bond length of 2.97 to 3.07 Å should produce thermoneutral H adsorption and optimal intrinsic activity for HER in an electrocatalyst with Ni_3 motif (Figure 10d). Such findings have provided a useful descriptor for high-throughput screening of Ni-nonmetal catalysts for HER. O'Connor et al.[278] has also used LASSO regression to understand the factors governing the interactions between single metal atoms and the oxide supports in single-atom catalysts. The authors discovered that in addition to the known relationships between binding energy and the oxide formation enthalpy of the metal adatoms ($\Delta H_{f,\,ox}$), the binding energy was also correlated with the oxygen vacancy formation energy (ΔE_{vac}) of the oxide support. The authors devised a series of descriptors that included the atomic properties of the adatom and the support. All those features were further mathematically transformed, producing a feature space with 333 932 descriptors. After the LASSO feature selection in a repeated random shuffle CV with 10% test data, the top five 1D descriptor always contained $\Delta H_{f,\,ox}$ and ΔE_{vac}. In addition, the ratio of them, i.e., $|\Delta H_{f,\,ox}/\Delta E_{vac}|$, always appeared in the 1D descriptor multiplied by a second term. The descriptor and model can be potentially used for screening metal/support combinations as catalysts.

An interesting application of active learning aims not to replace DFT calculations, but rather to guide DFT-based searches for ideal intermetallic catalysts for CO_2 reduction and H_2 evolution.[119] Tran and Ulissi[119] obtained 1499 intermetallic

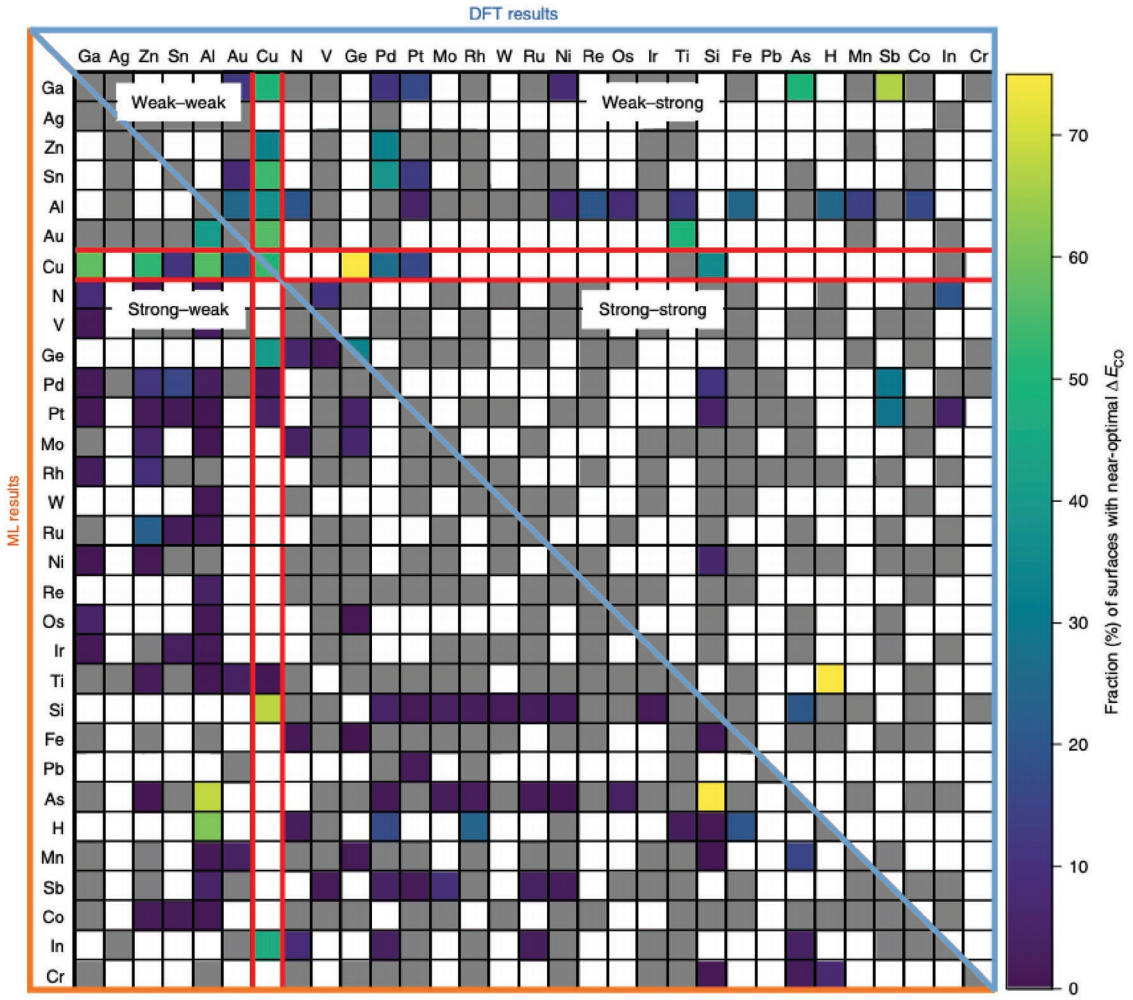

Figure 11. Predicted CO_2 reduction activity map for two-component intermetallics from ML and DFT. The elements are sorted according to the binding energy of CO on elemental surfaces, where the adsorption energy on monometallic Cu is nearest to the optimal value of –0.67 eV. The elements above Cu show relatively weak binding and the ones below it show strong binding. The near optimal ΔE_{CO} is defined as the ±0.1 eV range of the optimal value, i.e., (–0.77, –0.57) eV. Reproduced with permission.[119] Copyright 2018 Nature Publishing Group.

crystals from the Materials Project[10] covering 31 elements (50% d-block and 33% p-block). From a total of 1 684 908 unique adsorption sites as the candidate pool, an iterative active learning approach and the ML models only worked as guides for finding the next candidates to perform DFT calculations of the adsorption energy of CO and H. In each iteration, an ML model was fitted using previous data and the model predicted candidates with the strongest binding energies close to the optimal value (–0.67 eV for CO and –0.27 eV for H) for DFT calculations. **Figure 11** shows the resulting CO_2 reduction activity map of bimetallic surfaces from ML and DFT. While the ML models were able to provide results consistent with DFT, their large errors (MAE of 0.29 eV for CO_2 reduction and 0.24 eV for H_2 evolution) render them unsuitable for direct prediction. The main reason lies in the fact that the absorption energy strongly depends on atomic positions, which is not known prior to performing a structural relaxation.

Structural relaxations and their effect on absorption energies and other properties pose a particular challenge in catalysis problems. The model structures in catalysis typically comprise slabs or nanoparticles comprising many more atoms and lower symmetry than bulk crystals, which renders DFT relaxations much more expensive. This has prompted several innovative approaches to circumvent the need to perform structural relaxations in building ML models. For example, Jinnouchi et al.[140,279] have used local environment similarity as features to prediction absorption energies of N, O, and NO on $Rh_{1-x}Au_x$ nanoparticles, with the assumption that similar local environments lead to similar binding energies. The authors used the SOAP local environment descriptor on unrelaxed structures and combined with Bayesian linear regression to predict the binding energy and formation energy (**Figure 12**a). The formation energy ML model was used to predict the stable metal element distribution in the nanoparticles via Monte Carlo simulations. The predicted MAEs for the binding energies of species and formation energies of alloys were about 100 meV and 20 meV per atom, respectively. The reaction turnover frequency (TOF) from predicted energies results showed a volcano-like shape with Au fraction x in $Rh_{1-x}Au_x$ Figure 12b, and with decreasing nanoparticle size,

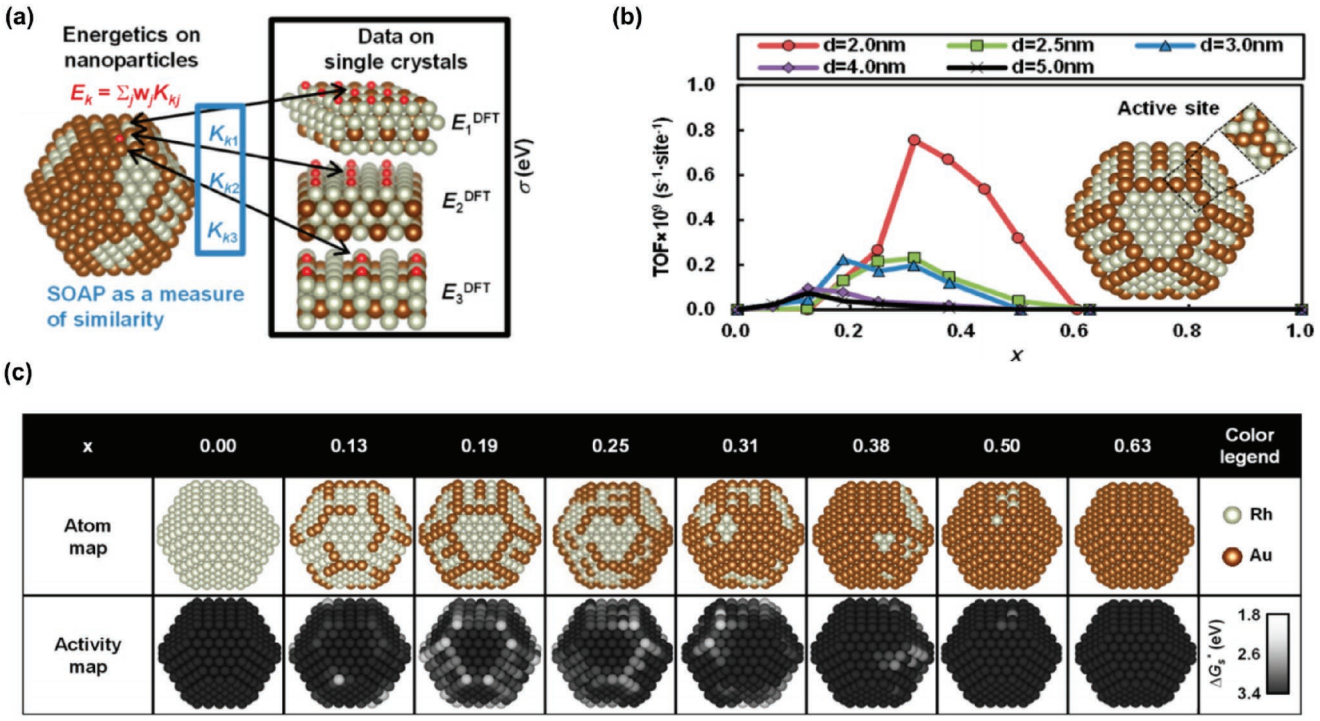

Figure 12. a) Bayesian linear regression scheme with kernel computed between unrelaxed structures and model targets being binding energies from DFT-relaxed structures. b) TOF as a function of x in $Rh_{1-x}Au_x$ and nanoparticle size. c) Activity map on atoms with different Au concentration. Adapted with permission.[140] Copyright 2017 American Chemical Society.

the reactivity increased. The mechanism was probed by mapping the reactivity on the nanoparticle surfaces. It was found that Au segregated to the corners and edges at low Au fraction in the nanoparticles which roughly corresponded to the reaction activity map (Figure 12c). With decreasing particle size, the atom fractions at the corners and edges increased. All these calculations and analyses were enabled with fast ML surrogate models. Similarly, Ulissi and co-workers[280–282] have used a modified CGCNN model[93] whereby the crystal graph is constructed using Voronoi tessellation to determine the edges, obviating the need for specific bond lengths. The model showed an error of only 0.15 eV in the prediction of CO and H_2 adsorption energies after training on 12 000 data for each molecule. Back et al.[280] further studied the oxygen evolution reaction (OER) on the IrO_2 and IrO_3 surfaces using the modified model. The authors discovered from DFT that the less stable low-index surfaces such as (100), two unique terminations of (111), and all active sites of (121) were more active than the most stable rutile (110) surfaces in IrO_2. The calculated data were subsequently fed into training GCNN models[281] using unrelaxed structures as inputs and DFT adsorption energies as the targets. Test MAEs of 0.07 eV (300 training data points) and 0.18 eV (500 training data points) were achieved for coverage and OER calculations, respectively. The same methodology has also been used in predicting intermetallic surface energies.[281]

Another way of circumventing DFT calculations is to relax the structures using ML-IAPs. For example, an NNP has been applied in the study of oxygen coverage on Pd (111) surface.[283] Ulissi et al.[284] have studied CO_2 reduction on nickel gallium bimetallic facets. The authors fitted an NNP to relax structures to a local minimum as well as predict the adsorption energy. They showed that with only 10% of the total required DFT calculations in a conventional study, the NNP can predict the binding energy of CO on Ni–Ga surfaces with RMSE approaching DFT accuracy of 0.2 eV. Recently, Chen et al.[285] have adopted a similar ML-IAP approach to accelerate the discovery of active sites for CO_2 reduction on Au nanoparticles and dealloyed Au_3Fe core–shell surfaces. The CO adsorption energy and HOCO transition state formation energy trained using 1100 data points reach an error level of 0.05 and 0.06 eV, respectively, on the test set. The trained models were subsequently applied to 11 537 surface sites, which represents a factor of ten in computational time reduction.

6.3.2. Reaction Pathways

Besides the binding energies, the energetics of the reaction pathway fundamentally determine the rate of the catalyzed reaction. An understanding of the reaction pathway is key to identifying the critical intermediates and the mechanism, leading to insights for the development of new catalysts. Ulissi et al.[286] have applied GPR to the study of syngas reaction on Rh(111). Although the reaction has more than two thousand possible pathways, a GPR model was able to start from a few DFT calculations and iteratively predict all intermediates. The most probable reaction network was identified with acetaldehyde and CO_2 as the reaction products for H_2 and CO_2 reactants.

6.3.3. Morphology

The surface/nanoparticle morphology can have a significant impact on the efficiency of a catalyst. In traditional approaches, a brute force computational search is typically used to determine the shape with the lowest energy as well as the orientation of nonatomic reactants with respect to the substrate. Genetic algorithms have been the common choice to assist predicting metal nanocluster configurations using computational methods.[287–289] Experimentally, atomic local environments can be inferred from X-ray adsorption spectroscopy (XAS). Several works have tried to predict the local environments by combining ML models with high-throughput computational or experimental XAS data.[172,290–294] Timoshenko et al.[290] have used neural networks to predict the Pt nanoparticle structure from the L-edge X-ray absorption near-edge spectra. The authors first constructed an experimentally verified computational database of L-edge XANES. Then the mapping from the L-edge XANES to the coordination number up to the fourth coordination shell was learned using neural networks. The NN model was successfully applied to experimental XANES. This example shows the power of ML beyond the capability of traditional methods. Conventionally, the study of coordination beyond the first shell is carried out using extended X-ray absorption fine structure (EXAFS), which has a relatively weak signal and is challenging in the probe of the local environment of systems at high temperatures, dilute samples, and in complex environments. The combination of ML with L-edge XANES solves this problem with high accuracy and can enable on-the-fly data and local environment acquisition. Predicting the local environments from X-ray absorption data will likely facilitate the mechanistic understanding of catalysis on an atomistic level.

In a study of C_{60} adsorption on a TiO_2 (101) surface, Todorović et al.[295] have devised a Bayesian optimization structure search (BOSS) scheme for addressing this interface problem by mapping the adsorption energy surface using only several DFT computations. The BOSS scheme was able to find energy minima using 12 data points in 1D and 45 data points in 2D, more than twice the efficiency of grid search. In higher dimensions, the grid search becomes intractable, yet the BOSS scheme was still able to find the minima with 700 data points in 5D.

Finally, surface coverage and energies, and consequently, morphology, can change with the chemical environment of the catalyst. Ulissi et al.[296] have used a GPR model to rapidly predict the free energies of different surface coverage configurations and thus construct the surface diagram. The computational cost was reduced by three times in constructing the Pourbaix diagrams of IrO_2 and MoS_2 using the ML-based approaches.

6.3.4. Summary

ML approaches have partially succeeded in solving the problems associated with large surface configurational space in catalysis. Nevertheless, the model errors are currently still too large to be a reliable surrogate to DFT calculations. For adsorption energies on chemically diverse surfaces, the prediction errors are above 0.2 eV.[119] Such large errors may misrepresent an unstable surface site as a stable one and change the reaction mechanisms. The high errors are an intrinsic difficulty for surface adsorption predictions, which require the relaxation of atoms to their equilibrium positions, while such relaxations are only possible with expensive ab initio methods or accurate ML transferable force fields. To reach errors below 0.2 eV, geometric descriptors based on local coordination counting are unlikely to yield good results and those that involve certain levels of advanced yet cheap calculations seem to be a better choice, as seen by the examples where adding d-band centers leads to more accurate models.

6.4. Thermoelectrics

Thermoelectrics, which are materials that can convert heat to electricity and vice versa, has long been seen as a potential approach to greatly enhance the efficiency of industrial processes and transportation, as well as a means of more efficient heating and cooling. However, the development of thermoelectric devices has long been constrained by materials performance. The performance of thermoelectric material is given by its figure of merit zT, defined as

$$zT = \frac{\sigma S^2 T}{\kappa} = \frac{\sigma S^2 T}{\kappa_{\text{electron}} + \kappa_{\text{phonon}}} \tag{18}$$

where S is the Seebeck coefficient, κ, κ_{electron}, and κ_{phonon} are the total, electron, and phonon thermal conductivities, respectively, σ is the electrical conductivity, and T is the temperature. Another commonly used metric is the power factor (PF), defined as σS^2. As σ and κ_{electron} are positively correlated, it can be seen that the ideal thermoelectric material would have a high electrical conductivity, high Seebeck coefficient and low phonon thermal conductivity, i.e., "phonon-glass electron-crystal" structures.[297] Despite intense research efforts, the best zT achieved to date remains ≈2.6 for SnSe.[298]

Here, we will review ML works aimed at specifically identifying thermoelectrics through property predictions. There are several works that aim to identify new intermetallic structures (e.g., Heuslers being a major class) and their stability, but make no attempt to predict other critical thermoelectric properties.[117,299,300] These works will not be reviewed here.

The Seebeck coefficient and electrical conductivity of a material are usually calculated within the framework of Boltzmann transport theory, using, for example, the BoltzTrap code.[301] Making use of the Gaultois's database,[302] which contains ≈1000 experimentally characterized thermoelectrics, Furmanchuk et al.[184] have developed a RF model to predict Seebeck coefficients. There were initially 452 features describing each structure in both crystal level and atomic level, but the best performing model only needed 187. The RMSE of the best model was 84 μV K^{-1}, which indicates an uncertainty of 11% in the Seebeck coefficient ranging from −400 to 400 μV K^{-1}. The predictions on an external test set of 20 materials from Gaultois's database were found to be accurate with a $R^2 \geq 0.88$, which is evidence of the generalizability of the model. It was also found (using feature importance) that the thermal conductivity of constituent elements in their ground-state crystal structures is important in determining the overall Seebeck coefficient. Choudhary et al.[303] have also attempted to develop ML models

for the classifications of Seebeck coefficients as well as the PF. For data at 600 K and 10^{20} cm^{-3} doping concentration, thresholds of −100 μV K^{-1}, 100 μV K^{-1}, and 1000 μW $(mK)^{-2}$ were set for n-type Seebeck coefficients, p-type Seebeck coefficients, and PF, respectively. Using CFID features with GBDT models, the classification AUCs are generally above 0.8.

The thermal conductivity is usually obtained by computing the phonon dispersion curve and applying the phonon Boltzmann transport equation (BTE). Obtaining an accurate phonon dispersion curve and force constants is highly computationally expensive. Seko et al.[142] have developed a Bayesian optimization method to locate the compound with the lowest lattice thermal conductivity. They computed lattice thermal conductivity for 101 compounds and used them as observations for a kriging search using GPR. The descriptors were volume (V) and density (ρ) of the crystal, along with a 34-digit one-hot-encoding of the elements. As a demonstration, the model successfully identified PbSe and LiI as the compounds with the lowest thermal conductivity using only 11 and 19 observed data, respectively. As a comparison, random search required 55 and 65 data points. The authors further screened 54 779 compounds from Materials Project,[10] and 221 were expected to have lower thermal conductivity than rocksalt PbSe (0.9 W m^{-1} K^{-1}). First-principles verification confirmed that all of the top five predicted compounds have thermal conductivity less than 0.2 W m^{-1} K^{-1}. While Seko et al.[142] used computed thermal conductivity as the learning target, Chen et al.[304] have developed a model to learn from experimentally measured thermal conductivity. The data set contains 100 experimentally measured thermal conductivities from a data set with diverse compositions and space groups. From a total of 63 features including elemental properties, DFT relaxed structural information, number of valence electrons for each subshell, DFT computed bulk modulus, and space group number, the 29 features with the highest importance were identified via recursive feature elimination. The as-trained GPR model was found to have RMSE of 0.18 for training set and 0.28 in test set for $\log\kappa_L$. Finally, the overall thermal resistance of a thermoelectric device is the sum of the bulk thermal resistance and the thermal boundary resistance (TBR) at the interfaces of the materials. Zhan et al.[305] have developed four ML models to predict experimentally measured TBR. The four models deployed different algorithms, namely, generalized linear regression (GLR), LASSO-GLR, GPR, and SVR. The data set contained total of 876 TBRs measured for 368 interfaces of 45 different materials, of which the predicted TBRs by acoustic mismatch model (AMM) and diffuse mismatch model (DMM) only showed weak correlation with experimental values and the coefficients were R^2 = 0.6 and R^2 = 0.62, respectively. The GPR and SVR models are the best performing, with R^2 of 0.92 and RMSE of 13.2 and 13.9 10^{-9} m^2 K W^{-1}, respectively. Furthermore, by analyzing the feature importance, the film thickness was found to be an important descriptor, which is in accordance with intuition.

6.4.1. Recommendation Engines

In addition to using models to predict single property, Gaultois et al.[306] have developed an ML-based recommendation engine for the discovery of thermoelectric materials, which took into account of four properties, i.e., the Seebeck coefficient, electrical resistivity, thermal conductivity, and bandgap (**Figure 13**). The data used in this work was fairly comprehensive, including

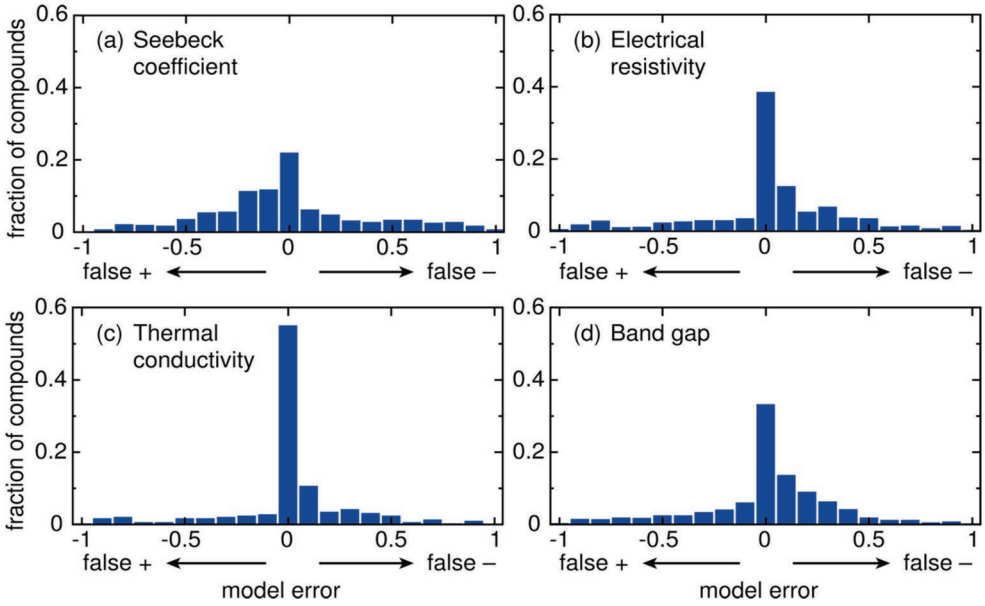

Figure 13. Distribution of leave-one-out cross validation errors of the recommendation engine on four key properties: a) Seebeck coefficient; b) electrical resistivity; c) thermal conductivity; d) bandgap. For the given material and property, the engine outputs the confidence score between 0 and 1 that the property falls within predefined ideal windows. The errors approaching 1 represent false negatives, meaning the property is poor when the engine predicts it to be promising. The errors approaching −1 represents false positive where the property is ideal when the engine predicts it to be poor. The errors close to 0 represent the model's predictions are in accordance with the ground truth. Therefore the peaks around 0 for four properties indicate the high reliability of the engine. Reproduced according to the terms of the CC-BY license.[306] Copyright 2016, The Authors.

both experimental and first-principles data, from almost all main-stream materials databases.[10,302] The descriptors of choice were elemental properties, with heavier weights assigned to the ones that displayed periodic table principles. The engine makes decision with the help of RF models on the input materials by analyzing whether the properties fall within certain criteria. To demonstrate the engine's ability in guiding experimental design, the authors chose two materials ($RE_{12}Co_5Bi$, RE = Er, Gd) out of high-ranking candidates for experimental synthesis. The model suggested their high probability of achieving high electrical and low thermal conductivity, and low probability of having a large Seebeck coefficient. Full characterization of the obtained materials showed that both materials have noncompetitive Seebeck coefficients, relatively high thermal conductivity, but promising electron conductivity. The overall zT is around 0.03 W m^{-1} K^{-1} at 400 K, which is higher than 30% structures in Gaultois's database.[302] While the zT of the two materials are not state-of-the-art, it should be noted that common thermoelectric materials contain heavy p-block metalloids like Sn, Sb, and Te, and less portion of d- and f-block metals. Therefore, the two $RE_{12}Co_5Bi$ compounds are counterintuitive and thus showcases the capability of ML guidance in expanding the chemical space of the field. Other more successful experimental verification included the Heusler $TiRu_2Ga$, whose preliminary results indicated that it had low thermal conductivity, and transition-metal germanides $Mn(Ru_{0.4}Ge_{0.6})$, whose thermal conductivity could be as low as 2 W m^{-1} K^{-1}.[117] In addition, two quaternary rare-earth germanides RE-M-M′-Ge, i.e., $Nd_4Mn_2InGe_4$ and $Nd_4Mn_2AgCe_4$, have also been synthesized[307] and confirmed to have low thermal conductivities, in agreement with the engine prediction.

6.4.2. Adaptive Design

One important and often used technique in discovering optimum energy materials is chemical composition alternation, meaning to tune the ratio of certain constituent elements in order to optimize property. It has been reported experimentally that $Al_2Fe_3Si_3$ compound can exhibit low lattice thermal conductivity and high Seebeck coefficient at different temperatures. However, various measurements of the overall PF of this compound are always less than 1 mW m^{-1} K^{-2}. Recent evidence shows that the PF can be improved by controlling the conduction type of $Al_2Fe_3Si_3$, which is further dictated by the ratio of Al/Si. Hou et al.[308] have employed a GPR model to predict PF for unknown compositions from existing experimental data. The input features were composition x in $Al_{23.5+x}Fe_{36.5}Si_{40-x}$ and temperature. Initially, temperature-dependent PFs (T = 300–840 K) were obtained for five compositions of $Al_{23.5+x}Fe_{36.5}Si_{40-x}$ (x = 0.0, 1.5, 1.8, 2.0, 2.2). The model was trained using the initial data, and based on the prediction, the new composition with the highest PFs in the temperature range of 450–650 K would be synthesized and the measured data were added to the data set for further training of the model. The iteration was stopped with convergence at x = 0.9, which demonstrated higher PF than its neighbor ratios. The PF at ≈510 K with x = 0.9 was ≈670 μW m^{-1} K^{-2}, which is 40% higher than x = 0, the original composition. The authors explained the improvements from two aspects. First, increasing x (Al content) results in a higher Seebeck coefficient because Al has fewer valence electrons than Si, so introducing more Al reduces carrier concentration. Second, the presence of a secondary metallic phase of $\tau 8$-$Al_2Fe_3Si_4$ might be the reason why electrical conductivity of x = 0.9 is higher than x = 0.7. This work made a contribution in presenting a new phase of $Al_2Fe_3Si_3$ that shows superior performance. The adaptive Bayesian analysis is an interesting approach to reduce the number of trials.

6.4.3. Text Mining

Encouraged by the advancement in the natural language processing field, attempts have been made to extract data and insights from literature using text mining. In thermoelectrics, the relevant literature consists of 3.3 million published papers between 1922 and 2018 in more than 1000 journals.[52] Tshitoyan et al.[52] has shown that, by feeding these abstracts to a "machine," one can obtain embeddings of words that incorporate the chemical relationships. The dot-product of the embeddings of two words can be interpreted as the likelihood of the words co-occurring in scientific abstracts. The authors therefore calculated the dot-products of 7663 material embeddings with the word "thermoelectric". The average maximum PF (calculated by DFT) of 40.8 μW cm^{-1} K^{-2} for these top ten predictions was 3.6 times larger than the average of candidate materials (11.5 μW cm^{-1} K^{-2}) and 2.4 times larger than the average of known thermoelectrics (17.0 μW cm^{-1} K^{-2}). Also, these embeddings are highly readable. For example, $CsAgGa_2Se_4$ appeared close to words like "chalcogenide," "bandgap," "optoelectronic," and "photovoltaic applications" in the coordinates of the embedding (**Figure 14**c). It can explain why it is close to "thermoelectric:" $CsAgGa_2Se_4$ belongs to chalcogenides which are often good thermoelectric materials. In addition, the bandgap is often an examined property for thermoelectric applications, and thermoelectric materials often have overlap with optoelectronics and PV applications. The trained embeddings can be used as a powerful tool for discovering new candidates, and more importantly, the methodology described in the paper could be easily transplanted for other tasks in materials design and benefit the field enormously.

6.4.4. Summary

As with other applications, the data quantity and quality are critical for training ML models. For thermoelectrics, as pointed out by Furmanchuk et al.,[184] the Seebeck coefficient is sensitive to doping, meaning that missing or incorrect report of doping information could lead to unexpected modeling results. This imposes a strong requirement on data curation if experimental properties are the targets. In DFT calculations, usually constant relaxation time approximation is used which cannot evaluate the carrier scattering. This leads to overestimation of bipolar conduction and underestimation of Seebeck coefficients. Such errors can be case-dependent, leading to unexpected results in ML modeling. Therefore, more efforts shall be devoted on improving data quality.

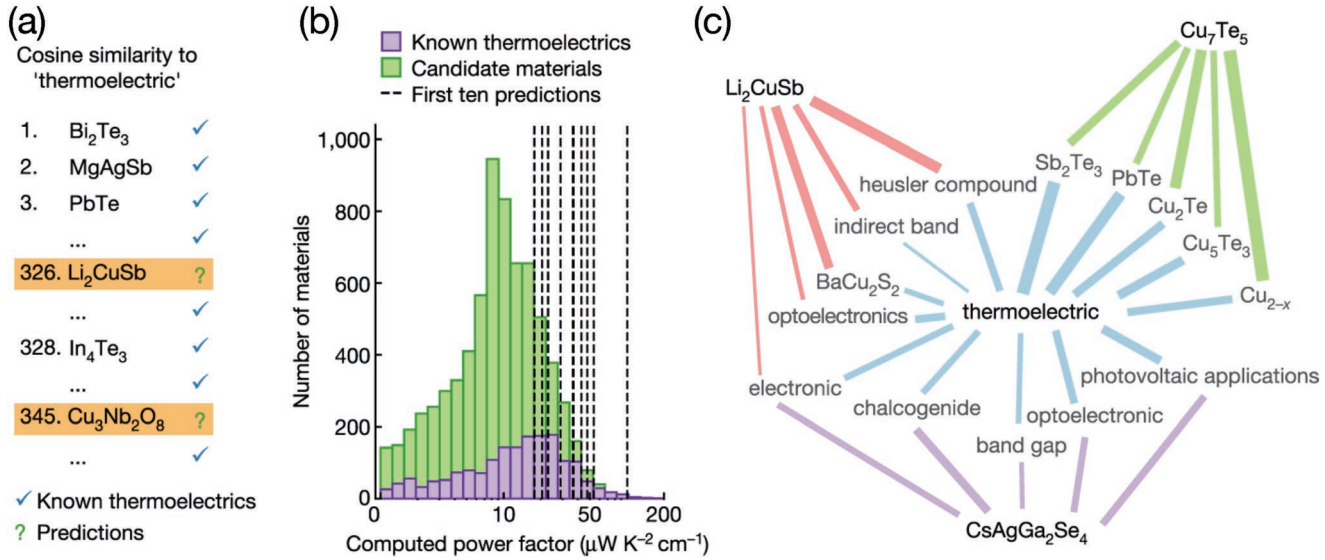

Figure 14. Prediction of new thermoelectric materials using text mining. a) Ranking of the cosine similarities of thermoelectric materials with the embedding of "thermoelectric." b) Distributions of the DFT-calculated power factors for 1820 known thermoelectrics (purple) and 7663 unreported candidates (green). The black dashed lines indicate the first ten predictions that have not been studied as thermoelectric. c) Illustration of how the context words of the materials connect to the word "thermoelectric." The width of the edges in the figure is proportional to the cosine similarity between the two connected words. Examination of the context words indicates that the algorithm makes decision based on structure type, co-mention of other thermoelectric materials, association of other related applications, and description of the materials' properties. Reproduced with permission.[52] Copyright 2019 Nature Publishing Group.

6.5. Piezoelectrics

Converting mechanical pressure into electricity (and vice versa) using piezoelectric materials is of significant technological interest for numerous applications in actuators, sensors, transducers, etc. The direct piezoelectric effect was firstly found in single crystal quartz, and scientists have further expanded the categories to perovskite ceramics (barium titanate, lead zirconate titanate, or PZT), ferroelectric ilmenite (lithium niobate, lithium tantalite, etc.), and polymer (polyvinylidene fluoride). Despite the variety of piezoelectric materials, intensive studies are conducted primarily on PZT-based compounds because of their large piezoelectric coefficients[309] and vertical morphotropic phase boundary (MPB).[310]

Balachandran et al.[311] have attempted to develop multivariate models for predicting Curie temperature (T_c) at MPB of perovskite $BiMeO_3$–$PbTiO_3$ solid solutions (Me is single cation with charge 3+ or a combination of two different cation with an average charge 3+) and discover high-temperature perovskite piezoelectric materials. The authors reduced 30 descriptors associated with crystal geometry, bonding, thermodynamics, and electronic structure to 6 key attributes with the PCA approach. These attributes were used in a linear multivariate model fitting on 15 data using PLS. According to the formalism, the authors predicted the high-temperature $BiTmO_3$ and $BiLuO_3$ with T_c of 730 and 705 °C, respectively (see **Figure 15**a). Then a tree-based classification model was used to probe the importance of these attributes to determine structural stability, showing that the A–O bond length in ABO_3 system is the most significant attribute (see Figure 15b). Similarly, Nelson and Sanvito[312] have predicted the experimental T_c for ≈2500 known magnets with RF algorithm and features only encoding the information of chemical composition. They used composition-weighted quantities, mode, and absolute deviation of elemental properties and predicted the target with an MAE of 57 K among all chemical compositions included in the test set without systematic bias toward any particular chemical compositions. When they tried to incorporate the structural information in the feature space, the prediction performance was poorer than that of using chemical composition only, which is likely due to the sparse training data given a relatively high-dimensionality of the feature space.

While T_c represents one critical property in piezoelectric materials, the "verticality" of the MPB, which determines the temperature insensitivity property of the materials, is another. For years scientists have made intensive efforts in searching for alternatives to PZT-based compounds because of the environmental concerns. However, these substitutes suffer from inferior temperature reliability compared to PZT,[313] which originates from a tilted/curved phase boundary in the composition-temperature phase diagram. Xue et al.[314] have discovered a monotonically decreasing relationship between the slope of the MPB with an increasing unit cell volume ratio (r_V) of the tetragonal and rhombohedral ends of composition-temperature phase diagram in $BaTiO_3$-based systems and $PbTiO_3$-based systems (see Figure 15c). They then defined the ratio of ionic displacements of the tetragonal and rhombohedral ends as a polarization-related descriptor (r_P) and revealed the positive linear correlation between the slope of MPB and r_P (see Figure 15d). While these studies demonstrated effective descriptors in searching piezoelectric materials, the scope of their application was restricted to certain structures/chemical systems and by a limited amount of data.

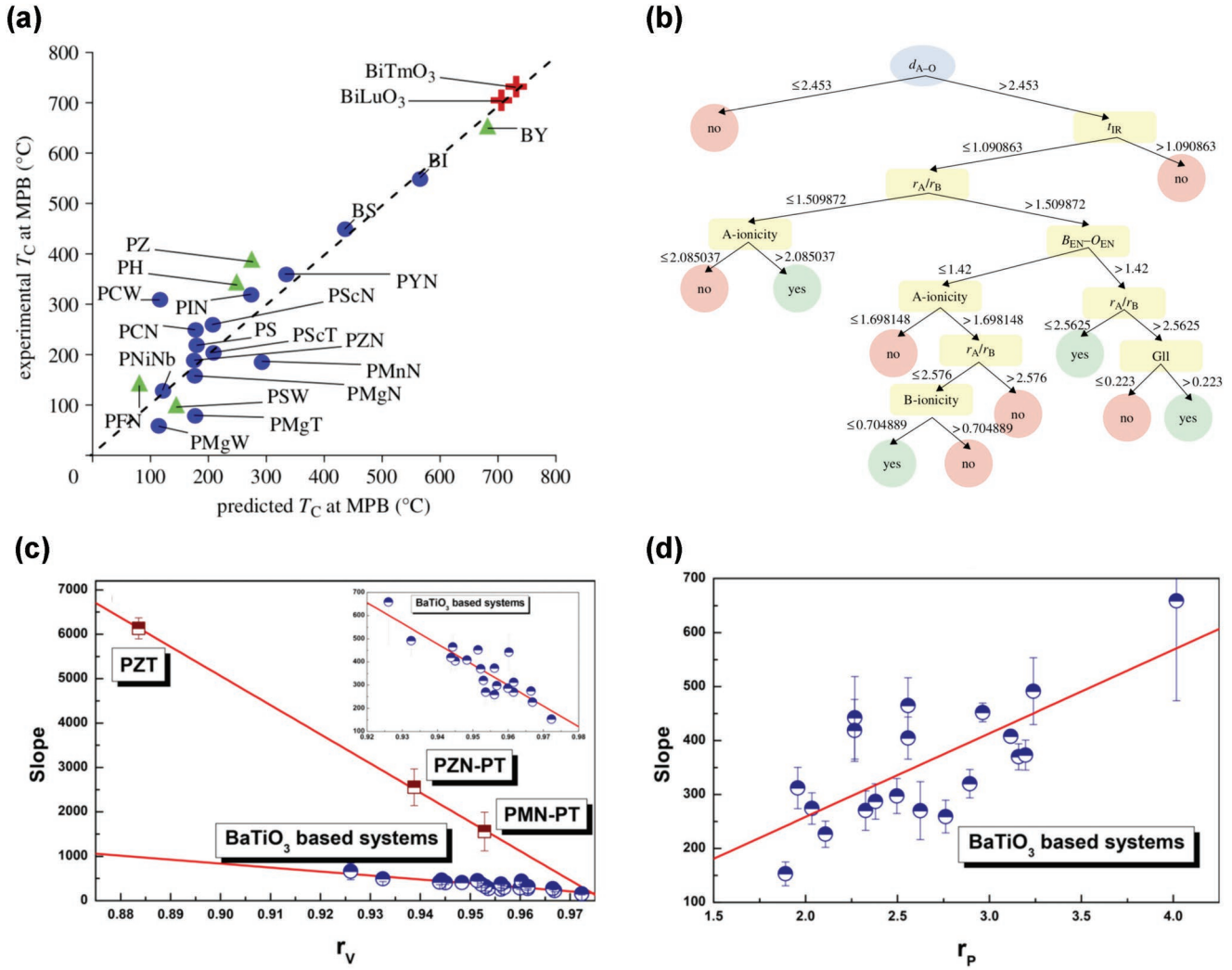

Figure 15. a) The PLS linear model trained on 15 systems and tested on 5 systems. b) The dendrogram classification model used Shannon entropy as a selection criterion to discriminate different descriptors importance. Reproduced according to the terms of the CC-BY license.[311] Copyright 2011, The Authors. c) The relationship between the slope of MPB and descriptor r_V (the ratio of the unit cell volume at two ends). d) The relationship between the slope of MPB and descriptor r_P (the ratio of ionic displacements at two ends).[314] Reproduced with permission.[314] Copyright 2017, American Institute of Physics.

Subsequently, Xue et al.[315] have applied Bayesian learning framework with uncertainties estimation and took linear regression as surrogate model to guide the discovery of new $BaTiO_3$-based materials with desirable temperature reliability (see **Figure 16**). The authors used the atomic, crystal, and electronic structure properties of the tetragonal and rhombohedral ends of composition-temperature phase diagram as features and the change of composition *dx* along each MPB when temperature decreases by 100 K from room temperature (298 K) as the target. The Landau functional provides a quadratic relationship between temperature τ_{MPB} and composition *x*. The "best" candidate with the smallest *dx* was chosen for experimental validation and the model was updated with augmentation of the newly measured *dx*. This strategy has helped the authors find the $(Ba_{0.5}Ca_{0.5})TiO_3$–$Ba(Ti_{0.7}Zr_{0.3})O_3$ system with better MPB verticality but at the expense of poorer piezoelectric response d_{33}. Similarly, Yuan et al.[316] have used Bayesian learning to lead the synthesis of the piezoelectric $(Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$ compound with largest electrostrain of 0.23% in the $BaTiO_3$-based family and $(Ba_{0.85}Ca_{0.15})(Ti_{0.91}Zr_{0.09})O_3$ compound with "optimal" d_{33} of 362 pC N^{-1}.[317] While these studies paved a way in accelerating the discovery of targeted piezoelectric materials, one objective optimization may sometimes lead to the expense of other properties not in the optimized targets[315] or failure in reproducing the best available results.[317]

To simultaneously explore multiple (potentially competing) properties, Gopakumar et al.[318] have provided a way to identify the points on characteristic property boundary where one property cannot be improved without the expense of degrading the other property, i.e., the Pareto front. The authors considered the Pareto front between bandgap and piezoelectric modulus of 704 piezoelectric materials from Materials Project[10,319] to be "unknown" and used adaptive learning to identify this "unknown" Pareto front from initial known data with as few

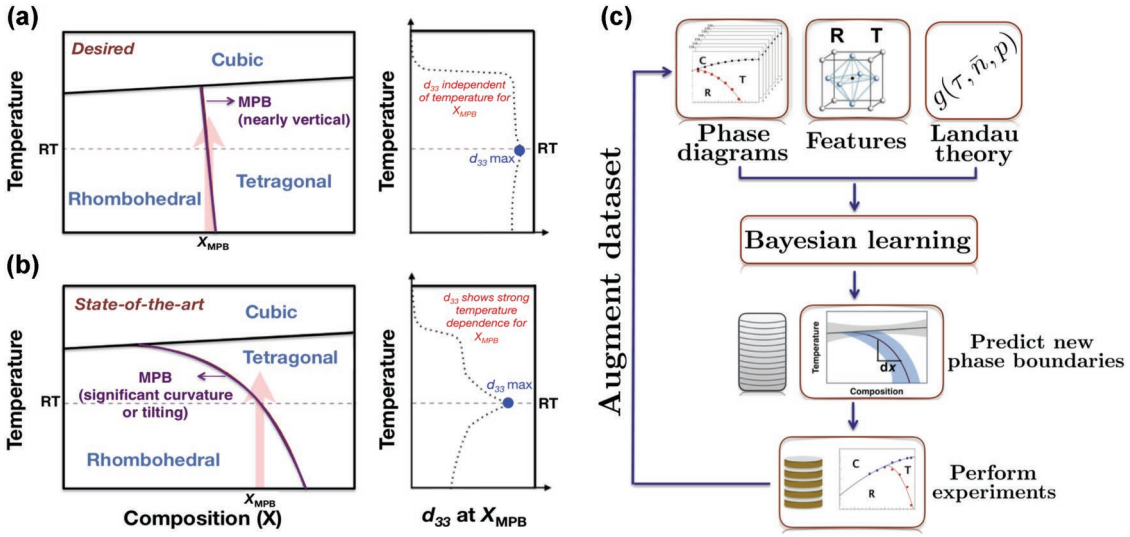

Figure 16. a) A desirable vertical MPB provides temperature-independent piezoelectric response d_{33} for the MPB composition (X_{MPB}). b) An undesirable tilted MPB brings up to highly temperature-sensitive piezoelectric response d_{33}. c) Bayesian learning framework for materials design. The Bayesian linear regression model relates phase boundaries to materials' features while the Landau functional serves as prior knowledge to constraint the model space. The loop is repeated until the material with the desired response is discovered. Reproduced with permission.[315] Copyright 2016, National Academy of Sciences.

measurements as possible. They have applied GPR and SVR with Gaussian radial basis function (RBF) kernel as surrogate models and expected improvement coupling the trade-off between "exploitation" and "exploration"[320] as learning criteria to optimally choose the next data point and update the model with new measured data until learning criteria is reached. Through the multiobjective optimization scheme, more than half of the Pareto-frontal points were found within the first 50 measurements. In contrast to previous works, the work by Balachandran et al.[320] explored a larger chemical space beyond titanates. Recently, large-scale screening for piezoelectrics has been attempted by combining high throughput screening and ML predictions.[321] Regression models targeting highest infrared frequency and maximum Born-effective charges, and classification models for maximum piezoelectric and average dielectric tensors have been built using CFID features and GBDT models to accelerate the materials discovery process. The model MAEs for highest infrared frequency and maximum Born-effective charges are 68.7 cm^{-1} and 0.6, respectively. Using thresholds of 1 C cm^{-2} for piezoelectric and 10 for dielectric, the classification model AUCs are 0.86 and 0.92, respectively.

6.5.1. Summary

ML in piezoelectrics is relatively new and initial efforts have been mainly on improving and understanding the existing titanate-based materials. Unlike other fields, the calculations of piezoelectrics are expensive. It is until recently, thousands of materials have been investigated computationally[319,321] and a large portion of known materials remains to be explored. Furthermore, the piezoelectric constants are tensorial properties, which obey crystal symmetry and bandgap constraints, i.e., only structures lack inversion symmetry and have electronic bandgaps can exhibit piezoelectric behavior. Such information is not considered in the existing models, and hence the models are not likely to fully capture the intrinsic piezoelectric behavior. This also calls for the fusion of physics and domain knowledge to ML applications in material science. Learning a tensorial property is also challenging and few attempts have been made to solve this issue.[322] The development of models that can predict tensors with symmetry constraints will likely see great interest.

6.6. Superconductors

Superconductivity describes electrical charge flow inside a material without resistance. Finding high-temperature, ideally, room-temperature, superconductors is of profound importance to the society on many aspects including reducing the electrical energy loss during transmission as well as saving the electricity waste in everyday appliances.[323] Improvements in first-principles calculations of electron–phonon spectra have made it possible to predict the critical temperature T_{cr} in silico. The error for predicting T_{cr} has been shown to be within 1 K in some elemental systems, ranging from weak-coupling (Mo,Al,Ta) to strong-coupling (Nb,Pb) superconductors.[324] This has led to the successful prediction and confirmation of new superconductors, including high pressure Si[325,326] and Li,[327–329] FeB_4,[330,331] H_2S (or H_3S),[332,333] $SnBi_2Se_4$[334] and $PbBi_2Te_4$,[335] and recently NaH_{10}.[336–339] These successes, coupled with the development of databases of superconductors,[44,340,341] have made ML an interesting new tool in the search of novel superconductors, particularly since the mechanisms for superconducting behavior differ from system to system.[342] For example, simple SVR models using only the lattice parameters as inputs has worked reasonably well on Fe-based superconductors[343] and doped MgB_2,[344] although the studies were limited in scope. Other than predicting the critical temperatures, ML

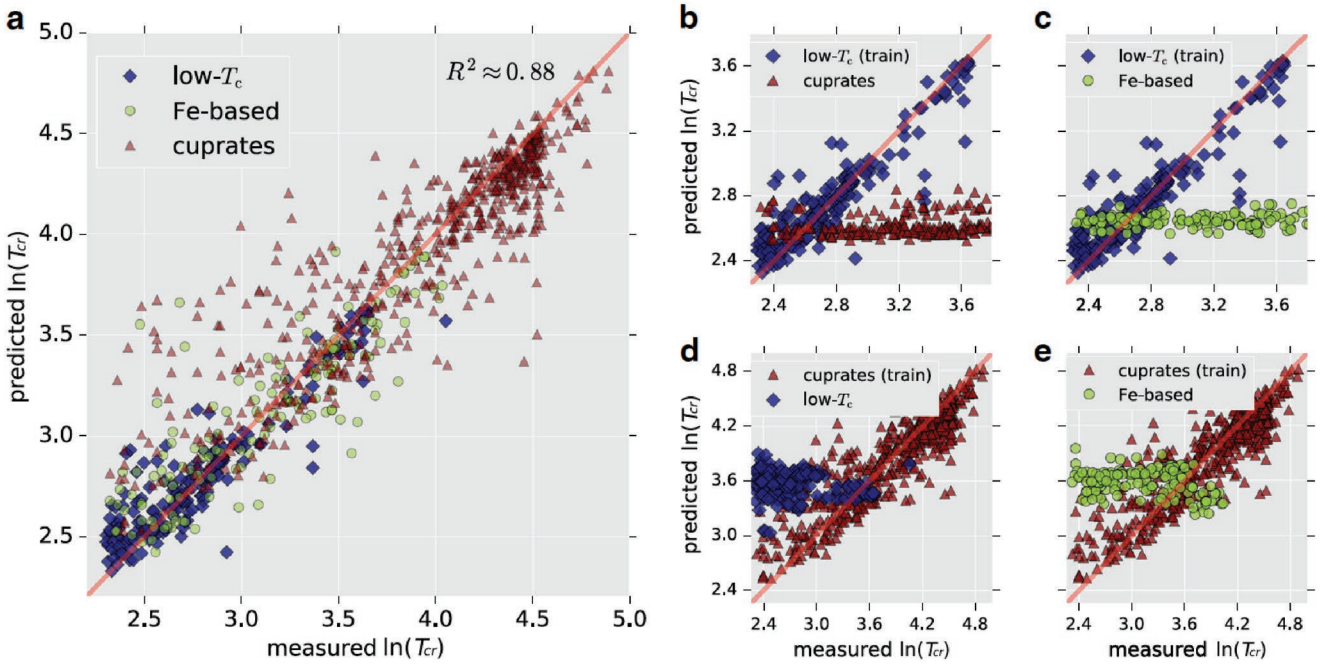

Figure 17. a) Parity plot between ML predicted and measured ln T_{cr} using all data groups with $T_{cr} > 10$ K. The performance of model trained on b) low-T_{cr} on cuprates and c) Fe-based materials. The performance of model trained on cuprates on d) low-T_{cr} materials and e) Fe-based materials. Adapted according to the terms of the CC-BY license.[30] Copyright 2018, The Authors.

models have also been used in predicting the electron–phonon coupling constants of elemental systems,[345] and elucidating the electronic ordering from experimental electronic structure images in superconducting copper oxide.[346]

Using electronic band structure derived features, Isayev et al.[29] have visualized relationships between different materials in "cartograms." The authors found $Ba_2Cu_3XO_7$ (where X is a lanthanide) lies at the center of the data and materials with highest T_{cr}, e.g., $Ba_2Ca_2Cu_3HgO_8$ (T_{cr} = 133 K) and $Ba_2CaCu_2HgO_6$ (T_{cr} = 125 K), were close in the feature space. The authors then used a modified Simplex (SiMRS) method for extracting structural features and used them together with electronic features for regression to predict T_{cr} and classification to find materials with $T_{cr} > 20$ K. The regression model errors were high, with a cross-validation determination coefficient of only 0.66. Nevertheless, due to particular data distribution that has a clear boundary at $T_{cr} = 20$ K, a classifier was able to achieve a cumulative accuracy of 0.94. Stanev et al.[30] have utilized the SuperCon database, which currently hosts data for more than 12 000 superconductors,[341] and constructed a series of models and workflow pipelines for finding potential superconductors in the existing structural database. The authors noted that if using only the SuperCon database, the model will be inevitably biased toward predicting superconductors and cannot tell the ones that do not exhibit such behavior (negative samples). These samples are particularly important from an ML perspective. The authors then included a pool of 300 non-superconducting materials to their model pool. The SuperCon database, however, does not provide structural information for the material. Thus compositional features using Magpie descriptors[72] were used with RF models for classifying high temperature superconductor by setting a threshold. The results showed that with a threshold of 10 K, the prediction accuracy was about 92%, and by keeping only the five most informative descriptors (from a full set of 145), the accuracy maintained at about 90%. RF regression models showed a surprisingly high accuracy in predicting T_{cr} with a R^2 of 0.88 between the predicted and measured T_{cr}, even though the features were only compositional (**Figure 17**a). However, if only one group of superconductors was used as training data, the models failed to predict other groups, as shown in Figure 17b–e. This finding is consistent with a later study by Meredig et al.,[180] where the authors noted that the superconductivity data formed distinct groups and the error of the models on a given group would increase sharply if no training data from this group was included in the model. The conclusion seems to suggest that different mechanisms are at play for different superconductor groups and in fact is known in the community. In addition, the authors tried to include AFLOW features to selected compounds with known structures and the inclusion was found to increase the model recall.

A distinct approach was taken by Konno et al.[347] to learn the T_{cr} of compounds from the composition using deep learning. The authors treated the periodic table as a rectangular image and filled the pixels of the image using the corresponding atomic fractions in the compound formula. The image was further separated into four channels, corresponding to s, p, d, and f blocks. Thus each compound could be encoded into a same-sized image with four channels. These image inputs were fed to CNN models for predicting T_{cr} and the predicted results showed an R^2 value of 0.92. However, like the previous work using the SuperCon database, the crystal structure information was not included, making it difficult to draw conclusions even the model error was low. In addition, similar approaches to

projecting the periodic table to an image have been adopted by Zheng et al.[348] to find novel X_2YZ compounds. The prediction of T_{cr} has also been attempted by Matsumoto and Horide[349] on ternary systems and below.

6.6.1. Summary

Predicting the superconducting critical temperature T_{cr}, particularly for nonelements, remains a major challenge. The foremost reason is that there is a lack of a universal theory and physical model of superconductivity. Furthermore, current data compilations of T_{cr} seem to lack even basic structural information, which is a critical gap that needs to be addressed for the development of reliable ML models.

7. Perspectives

In the preceding sections, we have attempted to give a critical review of the application of ML in the study and design of energy materials. While neither ML nor energy materials science can be considered "new" fields, a confluence of factors, namely, the substantial advances in ML, particularly, in deep learning and the parallel advancements in high-throughput first-principles computations and large federated computed materials property databases, have led to an intensification of research activity in this area in the past 4–5 years. From Section 6, it can be seen that there have already been substantial successes demonstrated, with ML models learning novel energy materials and insights. Here, we will provide our perspectives on the key challenges and opportunities in this nascent field. It should be noted that many aspects of the review (e.g., ML techniques and applications) as well as the following perspectives are readily generalizable to other application domains, and indeed, many energy materials have nonenergy related applications as well.

7.1. Data

As is evident from Section 6, the problem of limited data plague many application domains. ML models trained on fewer than 50 data points are common, especially when the target property is highly specific and difficult to measure/compute (e.g., alkali ionic conductivities or migration barriers, binding energies on surfaces, etc.). The outlook is better for more general properties. For instance, energetic data, such as the formation energy and other derived energy quantities such as E_{hull}, as well as electronic structure (bandgaps, density of states, band structure, etc.) are widely available for broad chemistries and crystal structures. Indeed, state-of-the-art graph-based deep learning models such as CGCNN,[93] SchNet,[259] and MEGNet[94] are already able to achieve MAE <0.04 eV per atom on the formation energy and MAE <0.35 eV on bandgaps, across data sets of ≈60 000–100 000 crystals, and this should serve as a benchmark for future general or specialized ML models. Sizable data sets also exist for elastic constants and moduli, diffraction spectra, X-ray absorption spectra, superconducting T_{cr}, etc. For the properties that are more difficult/expensive to measure/compute, there are several approaches worth exploring.

7.1.1. Data Fusion/Multifidelity Approaches

If it is possible to obtain low-fidelity data easily/cheaply, low fidelity data can provide extra information to high-fidelity ML models. The classic example is the bandgap of materials, for which a spectrum of techniques exist at different tradeoffs between cost and accuracy.[246,350] Yet another example is energies from different levels of theory, e.g., DFT versus CCSD(T).[351]

7.1.2. Transfer Learning

Somewhat related to the data fusion concept is transfer learning, whereby the knowledge gained from one model is used in a different but related problem. From Figure 2, there remains a substantial difference in data quantity between different properties, for example, there is an order of magnitude fewer elastic constants than energies and band structures in the Materials Project. Chen et al.[94] recently showed the elemental embeddings from a MEGNet model trained on formation energies can be transferred to accelerate the training and improve the accuracy of ML models for the bandgaps and elastic moduli. We note that the term "transfer learning" has been used somewhat inaccurately in the materials science domain to describe multifidelity approaches.

7.1.3. Crowd-Sourcing/Text Mining

Crowd-sourcing can be a means to improve the quantity of data available. For example, the MPContribs from Materials Project,[352] the NIST materials data repository,[353] citrination platform,[354] and the materials data facility (MDF)[355] are providing interfaces and tools for users to share their materials data. A second approach to substantially increase the amount of available data is via text-mining/natural language processing approaches. While text mining has been used to some success in identifying broad application/synthesis trends,[48,50,52,53] there have been no works attempting to extract materials properties.

Data scarcity is further compounded by "prejudice" bias. For example, the majority of computed energies and electronic structure properties are for experimentally known materials. Furthermore, there is much larger quantity of data on perovskites for a variety of applications (solar cells, fuel cells, etc.) compared to most other crystal types owing to the intense experimental focus on these structures and the fact that they are relatively easy to compute (the cubic perovskite unit cell only has five atoms). Similar biases in data exist in nearly all application domains. While it is impossible to compute all possible structural and compositional variations, care has to be taken to ensure that the chemical space outside of the immediate application domain is sufficiently sampled to avoid nongeneralizable models, obvious or spurious trends, or trivial findings, e.g., finding minor modifications of known materials with modest

improvements. First-principles computations can help address this issue to a certain extent, since such computations are not limited to already synthesized/synthesizable materials.

7.2. Models

As mentioned in Section 5.1, simpler, interpretable models are often more desirable in scientific disciplines. Many ML models are treated as "black boxes," with no relational constraints (well-established physical/chemical relationships) within their construction.

Relational constraints can be imposed within the model architecture itself as well as in feature selection. For instance, graph-based models implicitly encode bonding relationships between atoms in both crystals and molecules, while local-environment-based ML-IAPs impose locality of interactions between atoms. Indeed, once the model architecture encodes such relations between atoms, the selection of features becomes much more straightforward. For instance, it has been shown that graph-based models can achieve among the lowest MAEs on broad categories of crystal/molecule properties[94] with just the atomic number as atom feature and the bond distance as the bond feature. In the drive to achieve lower MAEs, it is not uncommon to see materials ML works where a large number of features (numbering in the hundreds) are fed into a model trained with similarly sized data sets. In more egregious cases, many of the features are highly correlated with each other, affecting model generalizability. One frequently used approach is to identify the most important features subsequently, but such techniques may end up arbitrarily choosing one correlated feature over another within the training data set. Nevertheless, there are approaches that can aid in model interpretability to some degree. For example, the symbolic regression method[356,357] is able to construct a mathematical formula for predicting target values, and such simple formulas may shed light on the physical processes. In deep learning for images, there have been attempts trying to unlock the "black box" of CNN models using class activation maps[358] or activation atlas maps.[359] Similar approaches that map certain molecular properties to the molecule geometries have been applied using GCNN models[360] and are likely to gain more popularity.

To date, uncertainty in the models is often overlooked except in a few cases where uncertainty is built in the model, for example, in GPR. However, in a materials discovery process, the desired material with superior property is more likely an "outlier" from existing data and thus in the model prediction on new materials, the credibility of the prediction should be assessed. This is especially true for black-box ML models. In addition to measuring the credibility of the prediction, the analysis of uncertainty also provides valuable information regarding the data selection for the model. Peterson et al.[361] devised a bootstrap approach to randomly sample the data with replacement for model construction. In each sampling out of a total number of 50, the same number of data points as the full data was sampled to construct a ML-IAP and those models were then used to predict energy and forces, forming an ensemble. The authors found that a large half ensemble spread on the sample (defined by the half of the spread between 5% and 95% of the predictions) was correlated to high uncertainty of model prediction on the sample and thus creative suggestions can be made by including the sample to the training data. However, creating an ensemble of models can be costly. In GPR models, such uncertainty analysis is carried out natively. For example, Bayesian optimization usually takes an iterative approach where, depending on the exploration and exploitation settings, high uncertainty samples are generally favored to be included in a data acquisition step. In neural networks, the analysis of uncertainty can be analyzed simply via keeping dropout in the prediction.[362] We believe that uncertainty quantification will play a major role in providing model insights, guiding the model construction and data selection in the near future.

8. Conclusion

To conclude, ML has already had a major impact on the study and discovery of energy materials in recent years. Nevertheless, there remains major scope and opportunities for improvements in both predictive performance as well as model interpretability. We are cautiously optimistic that with continued data improvements (e.g., better theoretical methods,[239–241,363] collection, curation, uncertainty quantification, etc.) and ML architecture advances (e.g., incorporating relational biases, known physics and chemistry, etc.), these challenges are surmountable and ML will become an integral complementary tool to existing experimental and computational techniques for materials science.

Acknowledgements

The authors acknowledge the support from the Materials Project, funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231: Materials Project program KC23MP.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

energy materials, machine learning, materials design

Received: October 4, 2019
Revised: December 13, 2019
Published online:

[1] A. L. Tarca, V. J. Carey, X. Chen, R. Romero, S. Drăghici, *PLoS Comput. Biol.* **2007**, *3*, e116.
[2] J. Carrasquilla, R. G. Melko, *Nat. Phys.* **2017**, *13*, 431.
[3] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, *arXiv:1903.10563* **2019**.
[4] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291.

[5] R. Gómez-Bombarelli, A. Aspuru-Guzik, in *Handbook of Materials Modeling* (Eds: W. Andreoni, S. Yip), Springer International Publishing, Berlin **2018**, pp. 1–24.
[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, *Nature* **2017**, *550*, 354.
[7] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5.
[8] G. Ho Gu, J. Noh, I. Kim, Y. Jung, *J. Mater. Chem. A* **2019**, *7*, 17096.
[9] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, *58*, 218.
[10] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[11] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*, 1501.
[12] M. Hellenbrandt, *Crystallogr. Rev.* **2004**, *10*, 17.
[13] J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722.
[14] P. Villars, N. Onodera, S. Iwata, *J. Alloys Compd.* **1998**, *279*, 1.
[15] P. S. White, J. R. Rodgers, Y. Le Page, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2002**, *58*, 343.
[16] F. H. Allen, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2002**, *58*, 380.
[17] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864.
[18] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133.
[19] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191.
[20] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2011**, *50*, 2295.
[21] G. Hautier, A. Jain, S. P. Ong, *J. Mater. Sci.* **2012**, *47*, 7317.
[22] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31*, 977.
[23] S. P. Ong, L. Wang, B. Kang, G. Ceder, *Chem. Mater.* **2008**, *20*, 1798.
[24] S. P. Ong, V. L. Chevrier, G. Hautier, A. Jain, C. Moore, S. Kim, X. Ma, G. Ceder, *Energy Environ. Sci.* **2011**, *4*, 3680.
[25] S. P. Ong, Y. Mo, W. D. Richards, L. Miara, H. S. Lee, G. Ceder, *Energy Environ. Sci.* **2013**, *6*, 148.
[26] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, G. Ceder, *Sci. Adv.* **2016**, *2*, e1600225.
[27] W. Ye, C. Chen, Z. Wang, I. H. Chu, S. P. Ong, *Nat. Commun.* **2018**, *9*, 3800.
[28] L. Wang, T. Maxisch, G. Ceder, *Phys. Rev. B* **2006**, *73*, 195107.
[29] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **2015**, *27*, 735.
[30] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, *npj Comput. Mater.* **2018**, *4*, 29.
[31] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, *Sci. Data* **2018**, *5*, 180053.
[32] G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, *J. Chem. Inf. Model.* **1983**, *23*, 66.
[33] P. Villars, *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds*, ASM International, Materials Park, OH **2007**.
[34] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, *J. Appl. Cryst.* **2009**, *42*, 726.
[35] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[36] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177.
[37] T. Fink, H. Bruggesser, J. L. Reymond, *Angew. Chem., Int. Ed.* **2005**, *44*, 1504.
[38] T. Fink, J. L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342.
[39] L. C. Blum, J. L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732.
[40] L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864.
[41] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202.
[42] L. Glasser, *J. Chem. Educ.* **2016**, *93*, 542.
[43] H. Landolt, R. Börnstein, *Physikalisch-Chemische Tabellen*, J. Springer, Berlin **1883**.
[44] D. R. Lide, *CRC Handbook of Chemistry and Physics*, Vol. 85, CRC Press, Boca Raton, FL **2004**.
[45] W. Martienssen, H. Warlimont, *Springer Handbook of Condensed Matter and Materials Data*, Springer Science & Business Media, Berlin **2006**.
[46] S. Kasap, P. Capper, *Springer Handbook of Electronic and Photonic Materials*, Springer Science & Business Media, Berlin **2017**.
[47] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist, J. Schrier, *Nature* **2019**, *573*, 251.
[48] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, A. Valencia, *Chem. Rev.* **2017**, *117*, 7673.
[49] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **2016**, *56*, 1894.
[50] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436.
[51] Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS Cent. Sci.* **2019**, *5*, 892.
[52] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.
[53] A. Torayev, P. C. M. M. Magusin, C. P. Grey, C. Merlet, A. A. Franco, *J. Phys. Mater.* **2019**, *2*, 044004.
[54] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227.
[55] D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Nørskov, K. W. Jacobsen, *Comput. Sci. Eng.* **2012**, *14*, 51.
[56] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 15010.
[57] NOMAD Repository, https://nomad-repository.eu/ (accessed: July 2019).
[58] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, *Comput. Mater. Sci.* **2016**, *111*, 218.
[59] K. Choudhary, Q. Zhang, A. C. E. Reid, S. Chowdhury, N. Van Nguyen, Z. Trautt, M. W. Newrock, F. Y. Congo, F. Tavazza, *Sci. Data* **2018**, *5*, 180082.
[60] J. S. Hummelshøj, F. Abild-Pedersen, F. Studt, T. Bligaard, J. K. Nørskov, *Angew. Chem., Int. Ed.* **2012**, *51*, 272.
[61] R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. Buongiorno Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **2014**, *93*, 178.
[62] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K. A. Persson, *Comput. Mater. Sci.* **2015**, *97*, 209.
[63] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241.
[64] NREL MatDB, https://materials.nrel.gov/ (accessed: June 2019).
[65] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, *MRS Bull.* **2016**, *41*, 399.
[66] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.

[67] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
[68] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G. M. Rignanese, G. Hautier, D. Gunter, K. A. Persson, *Concurr. Comput. Pract. Exp.* **2015**, *27*, 5037.
[69] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. H. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson, A. Jain, *Comput. Mater. Sci.* **2017**, *139*, 140.
[70] B. Hammer, J. K. Nørskov, *Advances in Catalysis*, Impact of Surface Science on Catalysis, Vol. 45, Academic Press, San Diego, CA **2000**, pp. 71–129.
[71] Y.-L. Lee, J. Kleis, J. Rossmeisl, Y. Shao-Horn, D. Morgan, *Energy Environ. Sci.* **2011**, *4*, 3966.
[72] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, *2*, 16028.
[73] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **2016**, *117*, 135502.
[74] D. Jha, L. Ward, A. Paul, W. Liao, A. Choudhary, C. Wolverton, A. Agrawal, *Sci. Rep.* **2018**, *8*, 17593.
[75] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *89*, 094104.
[76] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[77] O. A. vonLilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, *Int. J. Quantum Chem.* **2015**, *115*, 1084.
[78] E. Kauderer-Abrams, *arXiv:1801.01450* **2017**.
[79] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[80] B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2016**, *145*, 161102.
[81] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2017**, *13*, 5255.
[82] D. Weininger, *J. Chem. Inf. Model.* **1988**, *28*, 31.
[83] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742.
[84] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K. R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326.
[85] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[86] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[87] A. V. Shapeev, *Multiscale Model. Simul.* **2016**, *14*, 1153.
[88] K. Choudhary, B. DeCost, F. Tavazza, *Phys. Rev. Mater.* **2018**, *2*, 083801.
[89] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, *Sci. Adv.* **2017**, *3*, e1701816.
[90] B. Hammer, J. K. Nørskov, *Nature* **1995**, *376*, 238.
[91] D. Bonchev, *Chemical Graph Theory: Introduction and Fundamentals*, Vol. 1, CRC Press, Boca Raton, FL **1991**.
[92] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K. R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
[93] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[94] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 3564.
[95] M. R. Filip, F. Giustino, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 5397.
[96] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, *Sci. Adv.* **2019**, *5*, eaav0693.
[97] L. Pauling, *J. Am. Chem. Soc.* **1929**, *51*, 1010.
[98] V. M. Goldschmidt, *Naturwissenschaften* **1926**, *14*, 477.
[99] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*, 105503.
[100] C. Kim, G. Pilania, R. Ramprasad, *Chem. Mater.* **2016**, *28*, 1304.
[101] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, *Phys. Rev. Mater.* **2018**, *2*, 083802.
[102] R. B. Wexler, J. M. P. Martirez, A. M. Rappe, *J. Am. Chem. Soc.* **2018**, *140*, 4678.
[103] J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, H. Chang, *npj Comput. Mater.* **2019**, *5*, 37.
[104] J. Yang, V. Honavar, in *Feature Extraction, Construction and Selection: A Data Mining Perspective* (Eds: H. Liu, H. Motoda), The Springer International Series in Engineering and Computer Science, Springer US, Boston, MA **1998**, pp. 117–136.
[105] N. Wagner, J. M. Rondinelli, *Front. Mater.* **2016**, *3*, 28.
[106] F. Song, Z. Guo, D. Mei, in *2010 Int. Conf. on System Science, Engineering Design and Manufacturing Informatization*, Vol. 1, IEEE, Piscataway, NJ **2010**, pp. 27–30.
[107] C. Chen, Z. Deng, R. Tran, H. Tang, I. H. Chu, S. P. Ong, *Phys. Rev. Mater.* **2017**, *1*, 043603.
[108] B. C. Yeo, D. Kim, C. Kim, S. S. Han, *Sci. Rep.* **2019**, *9*, 1.
[109] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579.
[110] H. Park, R. Mall, F. H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail, F. El-Mellouhi, *Phys. Chem. Chem. Phys.* **2019**, *21*, 1078.
[111] M. Nuñez, *Comput. Mater. Sci.* **2019**, *158*, 117.
[112] G. Landrum, *RDKit: Open-Source Cheminformatics* **2006**.
[113] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2019**, 106949.
[114] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[115] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Materiomics* **2017**, *3*, 159.
[116] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, *J. Phys. Mater.* **2019**, *2*, 032001.
[117] A. O. Oliynyk, A. Mar, *Acc. Chem. Res.* **2018**, *51*, 59.
[118] Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duke, J. Brgoch, *Nat. Commun.* **2018**, *9*, 4377.
[119] K. Tran, Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696.
[120] K. Choudhary, M. Bercx, J. Jiang, R. Pachter, D. Lamoen, F. Tavazza, *Chem. Mater.* **2019**, *31*, 5900.
[121] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.
[122] R. Jalem, M. Nakayama, T. Kasuga, *J. Mater. Chem. A* **2014**, *2*, 720.
[123] R. Jalem, M. Kimura, M. Nakayama, T. Kasuga, *J. Chem. Inf. Model.* **2015**, *55*, 1158.
[124] A. D. Sendek, Q. Yang, E. D. Cubuk, K. A. N. Duerloo, Y. Cui, E. J. Reed, *Energy Environ. Sci.* **2017**, *10*, 306.
[125] A. D. Sendek, G. Cheon, M. Pasta, E. J. Reed, *arXiv:1904.08996* **2019**.
[126] C. Chen, Z. M. Baiyee, F. Ciucci, *Phys. Chem. Chem. Phys.* **2015**, *17*, 24011.
[127] C. Chen, D. Chen, F. Ciucci, *Phys. Chem. Chem. Phys.* **2015**, *17*, 7831.
[128] C. Chen, Z. Lu, F. Ciucci, *Sci. Rep.* **2017**, *7*, 40769.
[129] N. Kireeva, V. S. Pervov, *Phys. Chem. Chem. Phys.* **2017**, *19*, 20904.
[130] E. V. Podryabinkin, A. V. Shapeev, *Comput. Mater. Sci.* **2017**, *140*, 171.
[131] Z. Deng, C. Chen, X.-G. Li, S. P. Ong, *npj Comput. Mater.* **2019**, *5*, 1.
[132] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **2016**, *7*, 11241.
[133] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York **2006**.
[134] T. Mueller, A. G. Kusne, R. Ramprasad, *Reviews in Computational Chemistry* (Eds: A. L. Parrill, K. B. Lipkowitz), John Wiley & Sons, Inc., Hoboken, NJ **2016**, pp. 186–273.
[135] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 1.
[136] Y. Zhang, C. Ling, *npj Comput. Mater.* **2018**, *4*, 25.

[137] L. J. Nelson, G. L. W. Hart, F. Zhou, V. Ozoliņš, *Phys. Rev. B* **2013**, *87*, 035125.
[138] A. Lunghi, S. Sanvito, *Sci. Adv.* **2019**, *5*, eaaw2210.
[139] C. K. Williams, C. E. Rasmussen, *Gaussian Processes for Machine Learning*, Vol. 2, MIT Press, Cambridge, MA **2006**.
[140] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*, 4279.
[141] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. J. Fisher, H. Moriwake, I. Tanaka, *Adv. Energy Mater.* **2013**, *3*, 980.
[142] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, *Phys. Rev. Lett.* **2015**, *115*, 205901.
[143] S. M. Halawani, *Int. J. Tech. Res. Appl.* **2014**, *2*, 127.
[144] B. Medasani, A. Gamst, H. Ding, W. Chen, K. A. Persson, M. Asta, A. Canning, M. Haranczyk, *npj Comput. Mater.* **2016**, *2*, 1.
[145] V. Botu, A. B. Mhadeshwar, S. L. Suib, R. Ramprasad, *Information Science for Materials Discovery and Design* (Eds: T. Lookman, F. J. Alexander, K. Rajan), Springer Series in Materials Science, Springer International Publishing, Berlin **2016**, pp. 157–171.
[146] A. Furmanchuk, A. Agrawal, A. Choudhary, *RSC Adv.* **2016**, *6*, 95246.
[147] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, *Phys. Rev. X* **2014**, *4*, 011019.
[148] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947.
[149] B. Meredig, C. Wolverton, *Chem. Mater.* **2014**, *26*, 1985.
[150] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[151] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*, 2545.
[152] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, S. P. Ong, *arXiv:1906.08888* **2019**.
[153] C. Zheng, C. Chen, Y. Chen, S. P. Ong, *arXiv:1911.01358* **2019**.
[154] G. Marcus, *arXiv:1801.00631* **2018**.
[155] Y. Bengio, A. Courville, P. Vincent, *arXiv:1206.5538* **2012**.
[156] K. Hornik, M. Stinchcombe, H. White, *Neural Networks* **1989**, *2*, 359.
[157] H. Li, Z. Zhang, Z. Liu, *Catalysts* **2017**, *7*, 306.
[158] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.
[159] N. Artrith, A. Urban, G. Ceder, *Phys. Rev. B* **2017**, *96*, 014112.
[160] L. Zhang, J. Han, H. Wang, R. Car, W. E, *Phys. Rev. Lett.* **2018**, *120*, 143001.
[161] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **1998**, *86*, 2278.
[162] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Advances in Neural Information Processing Systems 25* (Eds: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger), Curran Associates, Inc., Red Hook, NY **2012**, pp. 1097–1105.
[163] S. Kajita, N. Ohba, R. Jinnouchi, R. Asahi, *Sci. Rep.* **2017**, *7*, 16991.
[164] J. Hoffmann, L. Maestrati, Y. Sawada, J. Tang, J. M. Sellier, Y. Bengio, *arXiv:1909.00949* **2019**.
[165] 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, *ACS Cent. Sci.* **2018**, *4*, 268.
[166] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, *preprint arXiv:1511.05493* **2015**.
[167] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595.
[168] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in *Proc. of the 34th Int. Conf. on Machine Learning* (Eds: D. Precup, Y. W. Teh), Vol. 70, JMLR.org, **2017**, pp. 1263–1272.
[169] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.
[170] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, R. Pascanu, *arXiv:1806.01261* **2018**.
[171] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513.
[172] C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson, S. P. Ong, *npj Comput. Mater.* **2018**, *4*, 12.
[173] A. R. Oganov, A. O. Lyakhov, M. Valle, *Acc. Chem. Res.* **2011**, *44*, 227.
[174] C. W. Glass, A. R. Oganov, N. Hansen, *Comput. Phys. Commun.* **2006**, *175*, 713.
[175] D. C. Lonie, E. Zurek, *Comput. Phys. Commun.* **2011**, *182*, 372.
[176] Y. Wang, J. Lv, L. Zhu, Y. Ma, *Comput. Phys. Commun.* **2012**, *183*, 2063.
[177] P. Jaccard, *New Phytol.* **1912**, *11*, 37.
[178] Y. Xu, J. Ma, A. Liaw, R. P. Sheridan, V. Svetnik, *J. Chem. Inf. Model.* **2017**, *57*, 2490.
[179] D. H. Wolpert, *Neural Comput.* **1996**, *8*, 1341.
[180] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, *Mol. Syst. Des. Eng.* **2018**, *3*, 819.
[181] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA **2012**.
[182] H. Sahu, W. Rao, A. Troisi, H. Ma, *Adv. Energy Mater.* **2018**, *8*, 1801032.
[183] I. Takigawa, K. Shimizu, K. Tsuda, S. Takakusagi, *RSC Adv.* **2016**, *6*, 52587.
[184] A. Furmanchuk, J. E. Saal, J. W. Doak, G. B. Olson, A. Choudhary, A. Agrawal, *J. Comput. Chem.* **2018**, *39*, 191.
[185] T. Xie, J. C. Grossman, *J. Chem. Phys.* **2018**, *149*, 174111.
[186] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[187] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, in *12th USENIX Symp. on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association, Savannah, GA **2016**, pp. 265–283.
[188] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, *Neural Information Processing Systems*, NeurIPS-Workshop, Long Beach, CA **2017**.
[189] An Automatic Engine for Predicting Materials Properties: Hacking-materials/Automatminer, Hacking Materials Research Group.
[190] B. Kolb, L. C. Lentz, A. M. Kolpak, *Sci. Rep.* **2017**, *7*, 1192.
[191] A. Khorshidi, A. A. Peterson, *Comput. Phys. Commun.* **2016**, *207*, 310.
[192] N. Artrith, A. Urban, *Comput. Mater. Sci.* **2016**, *114*, 135.
[193] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2019**, *15*, 448.
[194] Z. Li, J. R. Kermode, A. De Vita, *Phys. Rev. Lett.* **2015**, *114*, 096405.
[195] S. V. Kalinin, B. G. Sumpter, R. K. Archibald, *Nat. Mater.* **2015**, *14*, 973.
[196] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K. S. Sohn, *IUCrJ* **2017**, *4*, 486.
[197] J. Timoshenko, A. I. Frenkel, *ACS Catal.* **2019**, *9*, 10192.
[198] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, L. Emsley, *Nat Commun* **2018**, *9*, 1.
[199] K. Mizushima, P. C. Jones, P. J. Wiseman, J. B. Goodenough, *Mater. Res. Bull.* **1980**, *15*, 783.
[200] A. K. Padhi, *J. Electrochem. Soc.* **1997**, *144*, 1188.
[201] M. S. Whittingham, R. F. Savinell, T. Zawodzinski, *Chem. Rev.* **2004**, *104*, 4243.
[202] M. Armand, J. M. Tarascon, *Nature* **2008**, *451*, 652.
[203] M. S. Whittingham, *Chem. Rev.* **2014**, *114*, 11414.
[204] Z. Deng, Y. Mo, S. P. Ong, *NPG Asia Mater.* **2016**, *8*, e254.

[205] Y. Mo, S. P. Ong, G. Ceder, *Chem. Mater.* **2012**, *24*, 15.
[206] R. Jalem, T. Aoyama, M. Nakayama, M. Nogami, *Chem. Mater.* **2012**, *24*, 1357.
[207] I. G. Chong, C. H. Jun, *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103.
[208] K. Baba, I. Enbutu, M. Yoda, in *1990 IJCNN Int. Joint Conf. on Neural Networks*, Vol. 3, IEEE, Piscataway, NJ **1990**, pp. 155–160.
[209] M. Nakayama, K. Kanamori, K. Nakano, R. Jalem, I. Takeuchi, H. Yamasaki, *Chem. Rec.* **2019**, *19*, 771.
[210] H. Chen, L. L. Wong, S. Adams, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2019**, *75*, 18.
[211] R. Jalem, K. Kanamori, I. Takeuchi, M. Nakayama, H. Yamasaki, T. Saito, *Sci. Rep.* **2018**, *8*, 5845.
[212] Y. Zhang, X. He, Z. Chen, Q. Bai, A. M. Nolan, C. A. Roberts, D. Banerjee, T. Matsunaga, Y. Mo, C. Ling, *Nat. Commun.* **2019**, *10*, 1.
[213] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, M. Asta, *Sci. Data* **2015**, *2*, 150009.
[214] Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, *ACS Cent. Sci.* **2018**, *4*, 996.
[215] Z. Ahmad, V. Viswanathan, *Phys. Rev. Mater.* **2017**, *1*, 055403.
[216] W. Li, Y. Ando, E. Minamitani, S. Watanabe, *J. Chem. Phys.* **2017**, *147*, 214106.
[217] N. Artrith, A. Urban, G. Ceder, *J. Chem. Phys.* **2018**, *148*, 241711.
[218] A. Thompson, L. Swiler, C. Trott, S. Foiles, G. Tucker, *J. Comput. Phys.* **2015**, *285*, 316.
[219] X. Li, C. Hu, C. Chen, Z. Deng, J. Luo, S. P. Ong, *Phys. Rev. B* **2018**, *98*, 094104.
[220] M. A. Wood, A. P. Thompson, *J. Chem. Phys.* **2018**, *148*, 241721.
[221] S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, G. Csányi, *J. Chem. Phys.* **2018**, *148*, 241714.
[222] V. L. Deringer, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203.
[223] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J. E. Peralta, *ACS Appl. Mater. Interfaces* **2019**, *11*, 18494.
[224] M. A. Green, A. Ho-Baillie, H. J. Snaith, *Nat. Photonics* **2014**, *8*, 506.
[225] A. Kojima, K. Teshima, Y. Shirai, T. Miyasaka, *J. Am. Chem. Soc.* **2009**, *131*, 6050.
[226] W. S. Yang, B.-W. Park, E. H. Jung, N. J. Jeon, Y. C. Kim, D. U. Lee, S. S. Shin, J. Seo, E. K. Kim, J. H. Noh, S. I. Seok, *Science* **2017**, *356*, 1376.
[227] B. W. Park, S. I. Seok, *Adv. Mater.* **2019**, *31*, 1805337.
[228] C. Li, K. C. K. Soh, P. Wu, *J. Alloys Compd.* **2004**, *372*, 40.
[229] L. M. Feng, L. Q. Jiang, M. Zhu, H. B. Liu, X. Zhou, C. H. Li, *J. Phys. Chem. Solids* **2008**, *69*, 967.
[230] Q. Sun, W.-J. Yin, *J. Am. Chem. Soc.* **2017**, *139*, 14905.
[231] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.-J. Yin, *Adv. Funct. Mater.* **2019**, *29*, 1807280.
[232] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, *Front. Mater.* **2016**, *3*, 19.
[233] W. Shockley, H. J. Queisser, *J. Appl. Phys.* **1961**, *32*, 510.
[234] S. Rühle, *Sol. Energy* **2016**, *130*, 139.
[235] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[236] J. M. Crowley, J. Tahir-Kheli, W. A. Goddard, *J. Phys. Chem. Lett.* **2016**, *7*, 1198.
[237] J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, A. Görling, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2801.
[238] O. Allam, C. Holmes, Z. Greenberg, K. C. Kim, S. S. Jang, *ChemPhysChem* **2018**, *19*, 2559.
[239] J. Heyd, G. E. Scuseria, M. Ernzerhof, *J. Chem. Phys.* **2003**, *118*, 8207.
[240] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, J. G. Ángyán, *J. Chem. Phys.* **2006**, *124*, 154709.
[241] F. Aryasetiawan, O. Gunnarsson, *Rep. Prog. Phys.* **1998**, *61*, 237.
[242] M. L. Agiorgousis, Y.-Y. Sun, D.-H. Choe, D. West, S. Zhang, *Adv. Theory Simul.* **2019**, *2*, 1800173.
[243] O. Gritsenko, R. van Leeuwen, E. van Lenthe, E. J. Baerends, *Phys. Rev. A* **1995**, *51*, 1944.
[244] M. Kuisma, J. Ojanen, J. Enkovaara, T. T. Rantala, *Phys. Rev. B* **2010**, *82*, 115106.
[245] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
[246] G. Pilania, J. E. Gubernatis, T. Lookman, *Comput. Mater. Sci.* **2017**, *129*, 156.
[247] L. Yu, A. Zunger, *Phys. Rev. Lett.* **2012**, *108*, 068701.
[248] F. Tran, P. Blaha, *Phys. Rev. Lett.* **2009**, *102*, 226401.
[249] L. Meng, Y. Zhang, X. Wan, C. Li, X. Zhang, Y. Wang, X. Ke, Z. Xiao, L. Ding, R. Xia, H.-L. Yip, Y. Cao, Y. Chen, *Science* **2018**, *361*, 1094.
[250] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, *Adv. Mater.* **2006**, *18*, 789.
[251] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik, *Sci. Data* **2016**, *3*, 160086.
[252] E. O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, *Adv. Funct. Mater.* **2015**, *25*, 6495.
[253] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K. R. Müller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.
[254] D. Padula, J. D. Simpson, A. Troisi, *Mater. Horiz.* **2019**, *6*, 343.
[255] H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang, H. Ma, *J. Mater. Chem. A* **2019**, *7*, 17480.
[256] E. O. Pyzer-Knapp, G. N. Simm, A. A. Guzik, *Mater. Horiz.* **2016**, *3*, 226.
[257] S. A. Lopez, B. Sanchez-Lengeling, J. de GoesSoares, A. Aspuru-Guzik, *Joule* **2017**, *1*, 857.
[258] A. Paul, D. Jha, R. Al-Bahrani, W.-k. Liao, A. Choudhary, A. Agrawal, *arXiv:1903.03178* **2019**.
[259] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, K.-R. Müller, *Advances in Neural Information Processing Systems 30* (Eds: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., Red Hook, NY **2017**, pp. 991–1001.
[260] H. S. Stein, D. Guevarra, P. M. Newhouse, E. Soedarmadji, J. M. Gregoire, *Chem. Sci.* **2019**, *10*, 47.
[261] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, T. Buonassisi, *Joule* **2018**, *2*, 1410.
[262] Z. Shi, E. Tsymbalov, M. Dao, S. Suresh, A. Shapeev, J. Li, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 4117.
[263] G. Rothenberg, *Catalysis: Concepts and Green Applications*, John Wiley & Sons, New York **2017**.
[264] S. Kito, T. Hattori, Y. Murakami, *Appl. Catal., A* **1994**, *114*, L173.
[265] M. Sasaki, H. Hamada, Y. Kintaichi, T. Ito, *Appl. Catal., A* **1995**, *132*, 261.
[266] T. Hattori, S. Kito, *Catal. Today* **1995**, *23*, 347.
[267] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, *QSAR Comb. Sci.* **2004**, *23*, 767.
[268] E.-J. Ras, G. Rothenberg, *RSC Adv.* **2014**, *4*, 5963.
[269] J. R. Kitchin, *Nat. Catal.* **2018**, *1*, 230.
[270] B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel, C. Sutton, *AIChE J.* **2018**, *64*, 2311.
[271] P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, *ChemCatChem* **2019**, *11*, 3581.

[272] J. K. Nørskov, F. Abild-Pedersen, F. Studt, T. Bligaard, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 937.
[273] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069.
[274] X. Ma, Z. Li, L. E. K. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2015**, *6*, 3528.
[275] Z. Li, S. Wang, W. Shan Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5*, 24131.
[276] R. Gasper, H. Shi, A. Ramasubramaniam, *J. Phys. Chem. C* **2017**, *121*, 5612.
[277] J. Noh, S. Back, J. Kim, Y. Jung, *Chem. Sci.* **2018**, *9*, 5152.
[278] N. J. O'Connor, A. S. M. Jonayat, M. J. Janik, T. P. Senftle, *Nat. Catal.* **2018**, *1*, 531.
[279] R. Jinnouchi, H. Hirata, R. Asahi, *J. Phys. Chem. C* **2017**, *121*, 26397.
[280] S. Back, K. Tran, Z. W. Ulissi, *ACS Catal.* **2019**, *9*, 7651.
[281] S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran, Z. W. Ulissi, *J. Phys. Chem. Lett.* **2019**, *10*, 4401.
[282] A. Palizhati, W. Zhong, K. Tran, Z. Ulissi, *ChemRxiv* **2019**, *10.26434/chemrxiv.8709566.v1*.
[283] J. R. Boes, J. R. Kitchin, *Mol. Simul.* **2017**, *43*, 346.
[284] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*, 6600.
[285] Y. Chen, Y. Huang, T. Cheng, W. A. Goddard, *J. Am. Chem. Soc.* **2019**, *141*, 11651.
[286] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*, 14621.
[287] D. Tian, H. Zhang, J. Zhao, *Solid State Commun.* **2007**, *144*, 174.
[288] S. Lysgaard, D. D. Landis, T. Bligaard, T. Vegge, *Top. Catal.* **2014**, *57*, 33.
[289] X. Huang, Y. Su, L. Sai, J. Zhao, V. Kumar, *J. Cluster Sci.* **2015**, *26*, 389.
[290] J. Timoshenko, D. Lu, Y. Lin, A. I. Frenkel, *J. Phys. Chem. Lett.* **2017**, *8*, 5091.
[291] K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong, K. A. Persson, *Sci. Data* **2018**, *5*, 180151.
[292] S. Kiyohara, T. Miyata, K. Tsuda, T. Mizoguchi, *Sci. Rep.* **2018**, *8*, 13548.
[293] M. R. Carbone, S. Yoo, M. Topsakal, D. Lu, *Phys. Rev. Mater.* **2019**, *3*, 033604.
[294] Y. Suzuki, H. Hino, M. Kotsugi, K. Ono, *npj Comput. Mater.* **2019**, *5*, 39.
[295] M. Todorović, M. U. Gutmann, J. Corander, P. Rinke, *npj Comput. Mater.* **2019**, *5*, 35.
[296] Z. W. Ulissi, A. R. Singh, C. Tsai, J. K. Nørskov, *J. Phys. Chem. Lett.* **2016**, *7*, 3931.
[297] D. M. Rowe, *CRC Handbook of Thermoelectrics*, CRC Press, Boca Raton, FL **1995**.
[298] L.-D. Zhao, S.-H. Lo, Y. Zhang, H. Sun, G. Tan, C. Uher, C. Wolverton, V. P. Dravid, M. G. Kanatzidis, *Nature* **2014**, *508*, 373.
[299] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, A. Mar, *Chem. Mater.* **2016**, *28*, 7324.
[300] F. Legrain, J. Carrete, A. van Roekeghem, G. K. Madsen, N. Mingo, *J. Phys. Chem. B* **2018**, *122*, 625.
[301] G. K. H. Madsen, J. Carrete, M. J. Verstraete, *Comput. Phys. Commun.* **2018**, *231*, 140.
[302] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, D. R. Clarke, *Chem. Mater.* **2013**, *25*, 2911.
[303] K. Choudhary, K. Garrity, F. Tavazza, *arXiv:1906.06024* **2019**.
[304] L. Chen, H. Tran, R. Batra, C. Kim, R. Ramprasad, *arXiv:1906.06378* **2019**.
[305] T. Zhan, L. Fang, Y. Xu, *Sci. Rep.* **2017**, *7*, 7109.
[306] M. W. Gaultois, A. O. Oliynyk, A. Mar, T. D. Sparks, G. J. Mulholland, B. Meredig, *APL Mater.* **2016**, *4*, 053213.
[307] D. Zhang, A. O. Oliynyk, G. M. Duarte, A. K. Iyer, L. Ghadbeigi, S. K. Kauwe, T. D. Sparks, A. Mar, *Inorg. Chem.* **2018**, *57*, 14249.
[308] Z. Hou, Y. Takagiwa, Y. Shinohara, Y. Xu, K. Tsuda, *ACS Appl. Mater. Interfaces* **2019**, *11*, 11545.
[309] R. Guo, L. E. Cross, S.-E. Park, B. Noheda, D. E. Cox, G. Shirane, *Phys. Rev. Lett.* **2000**, *84*, 5423.
[310] A. Bouzid, E. Bourim, M. Gabbay, G. Fantozzi, *J. Eur. Ceram. Soc.* **2005**, *25*, 3213.
[311] P. V. Balachandran, S. R. Broderick, K. Rajan, *Proc. R. Soc. A* **2011**, *467*, 2271.
[312] J. Nelson, S. Sanvito, *arXiv:1906.08534* **2019**.
[313] W. Liu, X. Ren, *Phys. Rev. Lett.* **2009**, *103*, 257602.
[314] D. Xue, P. V. Balachandran, H. Wu, R. Yuan, Y. Zhou, X. Ding, J. Sun, T. Lookman, *Appl. Phys. Lett.* **2017**, *111*, 032907.
[315] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13301.
[316] R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, *Adv. Mater.* **2018**, *30*, 1702884.
[317] R. Yuan, D. Xue, D. Xue, Y. Zhou, X. Ding, J. Sun, T. Lookman, *IEEE Trans. Ultrason., Ferroelect., Freq. Control* **2019**, *66*, 394.
[318] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2018**, *8*, 3738.
[319] M. de Jong, W. Chen, H. Geerlings, M. Asta, K. A. Persson, *Sci. Data* **2015**, *2*, 150053.
[320] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, *Sci. Rep.* **2016**, *6*, 19660.
[321] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. H. Walker, F. Tavazza, *arXiv:1910.01183* **2019**.
[322] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, M. Ceriotti, *ACS Cent. Sci.* **2019**, *5*, 57.
[323] J. A. Flores-Livas, L. Boeri, A. Sanna, G. Profeta, R. Arita, M. Eremets, *arXiv:1905.06693* **2019**.
[324] M. A. L. Marques, M. Lüders, N. N. Lathiotakis, G. Profeta, A. Floris, L. Fast, A. Continenza, E. K. U. Gross, S. Massidda, *Phys. Rev. B* **2005**, *72*, 024546.
[325] K. J. Chang, M. L. Cohen, *Phys. Rev. B* **1984**, *30*, 5376.
[326] K. J. Chang, M. M. Dacorogna, M. L. Cohen, J. M. Mignot, G. Chouteau, G. Martinez, *Phys. Rev. Lett.* **1985**, *54*, 2375.
[327] A. Y. Liu, M. L. Cohen, *Phys. Rev. B* **1991**, *44*, 9678.
[328] J. B. Neaton, N. W. Ashcroft, *Nature* **1999**, *400*, 141.
[329] K. Shimizu, H. Ishikawa, D. Takao, T. Yagi, K. Amaya, *Nature* **2002**, *419*, 597.
[330] A. N. Kolmogorov, S. Shah, E. R. Margine, A. F. Bialon, T. Hammerschmidt, R. Drautz, *Phys. Rev. Lett.* **2010**, *105*, 217003.
[331] H. Gou, N. Dubrovinskaia, E. Bykova, A. A. Tsirlin, D. Kasinathan, W. Schnelle, A. Richter, M. Merlini, M. Hanfland, A. M. Abakumov, D. Batuk, G. Van Tendeloo, Y. Nakajima, A. N. Kolmogorov, L. Dubrovinsky, *Phys. Rev. Lett.* **2013**, *111*, 157002.
[332] Y. Li, J. Hao, H. Liu, Y. Li, Y. Ma, *J. Chem. Phys.* **2014**, *140*, 174712.
[333] A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov, S. I. Shylin, *Nature* **2015**, *525*, 73.
[334] R. Matsumoto, Z. Hou, H. Hara, S. Adachi, H. Takeya, T. Irifune, K. Terakura, Y. Takano, *Appl. Phys. Express* **2018**, *11*, 093101.
[335] R. Matsumoto, Z. Hou, M. Nagao, S. Adachi, H. Hara, H. Tanaka, K. Nakamura, R. Murakami, S. Yamamoto, H. Takeya, T. Irifune, K. Terakura, Y. Takano, *Sci. Technol. Adv. Mater.* **2018**, *19*, 909.
[336] F. Peng, Y. Sun, C. J. Pickard, R. J. Needs, Q. Wu, Y. Ma, *Phys. Rev. Lett.* **2017**, *119*, 107001.
[337] H. Liu, I. I. Naumov, Z. M. Geballe, M. Somayazulu, J. S. Tse, R. J. Hemley, *Phys. Rev. B* **2018**, *98*, 100102.
[338] A. P. Drozdov, P. P. Kong, V. S. Minkov, S. P. Besedin, M. A. Kuzovnikov, S. Mozaffari, L. Balicas, F. F. Balakirev,

D. E. Graf, V. B. Prakapenka, E. Greenberg, D. A. Knyazev, M. Tkacz, M. I. Eremets, *Nature* **2019**, *569*, 528.
[339] M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, R. J. Hemley, *Phys. Rev. Lett.* **2019**, *122*, 027001.
[340] C. K. Poole, H. A. Farach, R. J. Creswick, *Handbook of Superconductivity*, Elsevier, Amsterdam **1999**.
[341] Superconducting Material Database (SuperCon), https://supercon.nims.go.jp/index_en.html (accessed: October 2019).
[342] I. I. Mazin, *Nature* **2010**, *464*, 183.
[343] T. Owolabi, K. Akande, S. Olatunji, *Adv. Phys. Theor. Appl.* **2014**, *35*, 12.
[344] T. O. Owolabi, K. O. Akande, S. O. Olatunji, *J. Supercond. Nov. Magn.* **2015**, *28*, 75.
[345] Z. Alizadeh, M. R. Mohammadizadeh, *Phys. C* **2019**, *558*, 7.
[346] Y. Zhang, A. Mesaros, K. Fujita, S. D. Edkins, M. H. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. C. S. Davis, E. Khatami, E. A. Kim, *Nature* **2019**, *570*, 484.
[347] T. Konno, H. Kurokawa, F. Nabeshima, R. Ogawa, M. Iwazume, I. Hosako, A. Maeda, *arXiv:1812.01995* **2018**.
[348] X. Zheng, P. Zheng, R.-Z. Zhang, *Chem. Sci.* **2018**, *9*, 8426.
[349] K. Matsumoto, T. Horide, *Appl. Phys. Express* **2019**, *12*, 073003.
[350] M. L. Hutchinson, E. Antono, B. M. Gibbons, S. Paradiso, J. Ling, B. Meredig, *arXiv:1711.05099* **2017**.
[351] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087.
[352] P. Huck, A. Jain, D. Gunter, D. Winston, K. Persson, in *2015 IEEE 11th Int. Conf. on E-Science*, IEEE, Piscataway, NJ **2015**, pp. 535–541.
[353] Materials Data Repository Home, https://materialsdata.nist.gov/ (accessed: July 2019).
[354] Citrination, https://citrination.com/ (accessed: July 2019).
[355] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, *JOM* **2016**, *68*, 2045.
[356] S. Sun, R. Ouyang, B. Zhang, T. Zhang, *MRS Bull.* **2019**, *44*, 559.
[357] Y. Wang, N. Wagner, J. M. Rondinelli, *MRS Commun.* **2019**, *9*, 793.
[358] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE, Piscataway, NJ **2016**, pp. 2921–2929.
[359] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, C. Olah, *Distill* **2019**, *4*, e15.
[360] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in Neural Information Processing Systems 28* (Eds: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett), Curran Associates, Inc., Red Hook, NY **2015**, pp. 2224–2232.
[361] A. A. Peterson, R. Christensen, A. Khorshidi, *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978.
[362] Y. Gal, Z. Ghahramani, *arXiv:1506.02142* **2015**.
[363] J. Sun, A. Ruzsinszky, J. P. Perdew, *Phys. Rev. Lett.* **2015**, *115*, 036402.