TOPICAL REVIEW

Materials discovery through machine learning formation energy

Gordon G C Peterson\(^1\)\(\dagger\) and Jakoah Brgoch\(^1,2\)\(\dagger\)

\(^1\) Department of Chemistry, University of Houston, Houston, TX 77204, United States of America
\(^2\) Texas Center for Superconductivity, University of Houston, Houston, TX 77204, United States of America

E-mail: jbrgoch@uh.edu

Keywords: machine learning, intermetallics, neural network, support vector machine, random forest, phase diagram, materials informatics

Abstract
The budding field of materials informatics has coincided with a shift towards artificial intelligence to discover new solid-state compounds. The steady expansion of repositories for crystallographic and computational data has set the stage for developing data-driven models capable of predicting a bevy of physical properties. Machine learning methods, in particular, have already shown the ability to identify materials with near ideal properties for energy-related applications by screening crystal structure databases. However, examples of the data-guided discovery of entirely new, never-before-reported compounds remain limited. The critical step for determining if an unknown compound is synthetically accessible is obtaining the formation energy and constructing the associated convex hull. Fortunately, this information has become widely available through density functional theory (DFT) data repositories to the point that they can be used to develop machine learning models. In this Review, we discuss the specific design choices for developing a machine learning model capable of predicting formation energy, including the thermodynamic quantities governing material stability. We investigate several models presented in the literature that cover various possible architectures and feature sets and find that they have succeeded in uncovering new DFT-stable compounds and directing materials synthesis. To expand access to machine learning models for synthetic solid-state chemists, we additionally present MatLearn. This web-based application is intended to guide the exploration of a composition diagram towards regions likely to contain thermodynamically accessible inorganic compounds. Finally, we discuss the future of machine-learned formation energy and highlight the opportunities for improved predictive power toward the synthetic realization of new energy-related materials.

1. Introduction

The quest for superior solid-state materials spans from Bronze Age forgers searching for harder, more durable tools to modern-day researchers optimizing metal catalysts and hunting for the first room-temperature superconductor [1–4]. Scientific research methods have transitioned over this period from early empirical approaches to complex theoretical models and computational modeling. Materials chemistry has recently come to the cusp of a fourth paradigm of compound discovery, shifting toward data-driven materials informatics [5]. The study of data science has long existed, but the exploration of solid-state compounds through artificial intelligence has only recently been realized. Machine learning is a specific subset of data science, generally referring to the process by which a computer algorithm builds a mathematical model for predicting an unknown target through exposure to available training data. Thousands of individual training data with a multitude of numerical and/or categorical features are generally required to produce a well-trained model. Fortunately, the information in materials related data repositories like the Materials Project (MP) [6], Inorganic Crystal Structure Database (ICSD) [7], Pearson's Crystal Data (PCD) [8], Open Quantum Materials Database (OQMD) [9], Automatic-FLOW for Materials Discovery (AFLOW) [10], and NOvel MAterials Discovery (NOMAD) [11] has exploded over the past decades. For instance, the MP has grown from less than 10 000 entries in 2013 to contain more than 130 000 inorganic...
understand a compound’s Gibbs Free Energy, being synthetically accessible. Therefore, out of all physical properties, it is perhaps most important to understand one of the largest contributing factors to a compound’s Gibbs Free Energy, \( G \), which directly measures thermodynamic stability.

\[
G = H - TS.
\]  
(1)

In equation (1), \( G \) is classically defined with \( H \), \( S \), and \( T \) representing the enthalpy, entropy, and temperature of the system, respectively. Historically, entropic effects are often considered to have a small effect on overall stability, especially in the solid-solid transformations that dominate inorganic materials chemistry. Thus, this contribution is safely omitted. The change in \( G \) for any reaction may, therefore, be approximated as \( \Delta H \) according to equation (2), where \( H_{\text{prod}} \) and \( H_{\text{react}} \) represent the enthalpy of the products and reactants of a given reaction.

\[
\Delta G \approx \Delta H = H_{\text{prod}} - H_{\text{react}}.
\]  
(2)

By assuming a negligible change in volume for a solid-state reaction, \( \Delta H \) can be further simplified to the change in internal energy, \( \Delta U \), following equation (3).

\[
\Delta H = \Delta U - p \Delta V \approx \Delta U.
\]  
(3)

This simplification essentially sets \( \Delta G \) and \( \Delta H \) equal to \( \Delta U \), and most importantly, allows the total internal energy to approximate thermodynamic stability at 0 K.

The calculation of \( \Delta U \), and therefore \( \Delta H \), can be achieved using DFT to determine the relative stability of a compound with respect to its constituent elements using equation (4).

\[
\Delta H_f^{\varepsilon_1 \varepsilon_2 \cdots \varepsilon_n} = H_f^{\varepsilon_1 \varepsilon_2 \cdots \varepsilon_n} - \sum_{i=1}^{n} (x_i \times H_f^{\varepsilon_i}).
\]  
(4)

In this equation, \( \Delta H \) for a compound containing \( n \) elements \( \varepsilon_1, \varepsilon_2, \cdots \varepsilon_n \) is calculated on a per atom basis as the difference between the total enthalpy of the product \( H_f^{\varepsilon_1 \varepsilon_2 \cdots \varepsilon_n} \), and the sum of enthalpies for each constituent element, \( H_f^{\varepsilon_i} \), scaled by that element’s stoichiometric fraction in the product. This specific construction is notated as \( \Delta H_f \), commonly called the formation energy, and is generally used to measure a compound’s thermodynamic stability starting from elemental precursors. Thus, with relatively straightforward DFT calculations, \( H \) can be obtained for any compound and its elements, allowing a direct approximation of \( \Delta H_f \).

A reaction is favorable when the total energy for the products is more negative than the sum of the energies of the reactants. In this way, an overall negative value for \( \Delta H_f \) corresponds to a potentially synthetically accessible compound with respect to the elements. Conversely, a positive value for formation energy indicates thermodynamic instability and a preference towards decomposition. However, the starting materials are not the only competition for phase stability; it is also necessary to consider all neighboring compounds’ relative energetic stabilities within a given chemical system. This problem can be resolved by constructing a thermodynamic convex hull, drawn as the largest convex object connecting plotted \( \Delta H_f \) values against a compositional axis [30]. This concept is illustrated in figure 1. Here, the hypothetical compound \( \text{AB}_3 \) lies lower in formation energy than other compounds with nearby stoichiometries and forms part of the thermodynamically stable convex hull. \( \text{AB}_3 \) should, therefore, be synthetically accessible.
and form as a pure product phase when one part A and three parts B are allowed to reach thermodynamic equilibrium. A₄B, on the other hand, lies well above the convex hull. Although it has a lower formation energy than the constituent elements, A₄B must also compete with the nearby A₃B phase, which lies at a substantially more negative formation energy. Overall, a 4:1 A–B system will find thermodynamic stability as a heterogeneous mixture of the two nearest phases that lie on the convex hull, in this case, A₃B and elemental A. Thus, A₄B is likely inaccessible using standard synthetic approaches.

In practice, the line between what can be synthesized and what cannot be is a blurry divide. Sun et al found that 80% of experimentally verified inorganic solids listed in the ICSD are not predicted to be on the DFT-calculated hull and instead were found up to 36 meV atom⁻¹ above the calculated hull [31]. This effect is mostly attributed to DFT calculations being performed at 0 K, whereas solid-state synthesis is carried out at very elevated temperatures. Room temperature alone can account for an additional 26 meV atom⁻¹ of available thermal energy [12]. Additionally, the average difference between DFT and experimental measurements for the internal energy of elemental solids has been estimated around ∼24 meV atom⁻¹, although individual DFT error can vary greatly between compounds [32]. As a result, assuming that compounds within a buffer of about 50 meV atom⁻¹ above the predicted convex hull may, in fact, be thermodynamically stable is generally wise. For instance, in figure 1, the compound AB lies within this range just above the hull, and so despite its DFT-calculated instability with respect to the convex hull, it may be synthetically accessible.

Calculated DFT energies are available in high-throughput computational databases and are generally reliable; unfortunately, this information is difficult to employ to explore entirely unknown compositional spaces because of the massive compute hours necessary for DFT calculations and a prerequisite knowledge of the crystal structure. However, using DFT calculations as training data for supervised machine learning models of formation energy that make predictions based only on compositional information is extremely promising. This review focuses on the prediction of formation energy using machine learning and its role in guiding the synthesis of new inorganic materials for energy-relevant applications. This discussion begins by investigating the needs, benchmarks, design choices, and goals for a generalized machine model of formation energy. It continues with a look at the various choices of the algorithms used in model construction. In each case, published machine learning methods that have shown notable success are highlighted. The shortcomings of current models’ ability to accurately predict experimental phase diagrams are also discussed. In addition to reviewing the state-of-the-art models available, we present the newly developed MatLearn program, a web-based machine learning model and visualization package for the quick and generalizable
construction of predicted binary and ternary phase diagrams. This tool can help synthetic solid-state chemists by highlighting the most statistically favorable regions of a phase diagram for novel energy materials.

2. Data collection and features

A machine learning model’s predictive power is mostly dependent on the size, diversity, and quality of its training dataset. Fortunately, the dramatic expansion of materials data collected in online repositories addresses this need. Databases like the ICSD and PCD focus on collecting experimental crystal structure data [7, 8]. These are valuable resources and research tools, especially for synthetic chemists searching for experimentally verified compounds. Other repositories such as The MP [6], the OQMD [9], and AFLOW [10] focus on aggregating DFT-calculated quantities like formation energy. Additionally, the Materials Platform for Data Science [33], NIMS Material Database [34], and NOMAD [11] contain experimental data that has been extracted from the literature. These databases contain optimized structures and a plethora of computationally and experimentally derived property data, including thermoelectric power and conductivity coefficients for promising energy materials. They also have mechanical property data, including hardness and complex electronic information, magnetic descriptors for superconductivity, and magnetocaloric materials. This information is usually easily extracted using sleek application programming interfaces (APIs) that interface with common computing languages like Python. Employing APIs allows researchers to quickly pull troves of data that can be used for subsequent machine learning.

Data sets extracted from these repositories are generally reliable and complete, but it is always necessary to thoroughly ‘clean’ data before using it to train a machine learning model. For example, duplicate data are a common issue that can mislead or bias the model if multiple target values are assigned to a single compound. In the case of formation energy, duplicates are often the result of multiple DFT calculations run on the same or similar structural arrangement for a particular composition. As a result, it is usually considered best practice to keep only the most negative value among those available for constructing the convex hull. Missing values are another possible error when data is extracted in bulk and may also bias training. If data set size is not a concern, offending compounds can be removed; alternatively, they can be completed with an average feature value if necessary to minimize bias. It may additionally be useful to cross-reference the dataset against a repository of experimentally verified phases such as the ICSD or PCD to remove hypothetical compounds predicted by DFT calculation. The process of data cleaning is laborious and often the most time-intensive aspect of developing a machine learning model. Nevertheless, this is crucial for obtaining an unbiased and robust result.

Once a dataset has been collected and cleaned, the next step to consider is what features will represent the compounds. There are many features to consider when training a formation energy machine learning model, but the simplest way to describe an arbitrary compound is often by its composition. Because models for inorganic compounds generally perform best when the features are given as numerical rather than categorical values, compositions can be written as features through an approach called ‘one-hot encoding’ [35]. This process, illustrated in figure 2, transforms chemical formulas into vectors, with each value corresponding to a particular element’s fractional composition. The result effectively captures chemical formulas in a machine-readable format.

Beyond one-hot encoded fractional compositions, additional compositional information about each compound’s constituent elemental and physical properties can be included. Table 1 provides a list of the properties used in various models, covering periodic table information like atomic number and group, measurements of atomic radii, electronic values like electronegativity and electron counts, and physical properties like density and heat of fusion [12, 36–38]. Compounds can then be described through numerical operations of these values like the average, difference, largest value, smallest value of the compositional information. The inclusion of these property descriptors gives the model important information for each compound, and a well-chosen algorithm will not risk overfitting the data. Importantly, these features can
Table 1. Selected compositional features for solid state compounds derived from the elements.

| Elemental Compositional Descriptors | Atomic Size                        |
|-------------------------------------|------------------------------------|
| **Periodic Table**                  | **Atomic Size**                    |
| Atomic Number                       | Atomic Radius                      |
| Atomic Weight                       | Covalent Radius                    |
| Period                               | Zunger Radii Sum                   |
| Family                               | Ionic Radius                       |
| Group                                | s Orbital Pseudopotential Radius   |
| Mendeleev Number                    | p Orbital Pseudopotential Radius   |
| **Electronic**                      | van der Waals Radius               |
| Pauling Electronegativity (EN)      | Crystal Radius                     |
| Martynov-Batsanov EN                |                                    |
| Allen EN                             |                                    |
| Gordy EN                             |                                    |
| Mulliken EN                          |                                    |
| Allred-Rockow EN                     |                                    |
| Metallic Valence                    |                                    |
| Valence Electron Count              |                                    |
| Valence Electron Gilmour No.        |                                    |
| Valence s Electron Count            |                                    |
| Valence p Electron Count            |                                    |
| Valence d Electron Count            |                                    |
| Valence f Electron Count            |                                    |
| Outer Shell Electron Count          |                                    |
| Unfilled s States                   |                                    |
| Unfilled p States                   |                                    |
| Unfilled d States                   |                                    |
| Unfilled f States                   |                                    |
| Total Unfilled States               |                                    |
| Fraction of s Valence Electrons     |                                    |
| Fraction of p Valence Electrons     |                                    |
| Fraction of d Valence Electrons     |                                    |
| Fraction of f Valence Electrons     |                                    |

also be determined with zero knowledge of the target compound other than its composition. Sticking to composition-only based descriptors is vital to a formation energy machine learning model because non-compositional (especially structural) information is generally not available when searching for novel compounds across unknown regions of a phase diagram. Models that include structural descriptions have improved accuracy, but their scope is generally limited to investigating a single structure type, and researchers are still working on ways to encode this information for generic materials [39–41].

3. Model architecture

Perhaps the most critical decision in designing a machine learning model is the choice of learning algorithm, also called the model architecture. Algorithmic complexity ranges from simple linear models and kernel-based regression schemes to non-linear methods such as decision trees and neural networks. Each of these approaches comes in various flavors, often with the same basic learning unit assembled differently. There is no universally ideal algorithm, and deciding which to use depends on the problem at hand. In this section, some of the most common machine learning algorithms used for predicting formation energy are
discussed broadly. Moreover, several available models and learning architectures for predicting formation energy are highlighted.

**Decision Trees.** The most common algorithm class for creating formation energy models thus far are based on decision trees. A decision tree begins by splitting initial data into two groups based on a logical operation (node). Subsequent nodes split the resulting subgroups by additional operations into two branches at a time until the final leaves of the tree contain only suitably identical items. This concept is visualized in figure 3(a), where the items at each node are separated by logical choices like color, shape, and size until only groups of identical objects remain. The presence or absence of key elements or the magnitude of specific physical properties could represent analogous logical operations for a tree trained to sort solid-state materials. Decision trees offer the inherent advantage that their process is transparent, and visualizing a tree’s decision process allows comparison of the trained model to the chemist’s intuition. Decision trees are often applied towards classification problems, such as determining if a compound is a conductor, a semiconductor, or an insulator [42]. Given sufficient data, decision trees can also be used for regression problems to predict a continuous numerical value like formation energy.

Individual decision tree learners can be quite prone to overfitting, modeling noise rather than the training data signal. This can be addressed by adjusting the hyperparameters governing the tree, such as tree depth and features considered at each node. Often the best solution to this problem is through the use of decision tree ensembles. Random forests [43, 44], extremely randomized trees [45], extreme gradient boosting [46], and rotation forests [47] are tree ensemble methods that work by creating many individual decision trees and inducing some aspect of randomness to the training of each tree, ultimately averaging their results. A random forest, for example, uses a randomly split subsample of the data to train each tree, lowering the weight of outlier data points relative to the general trend [43, 44]. Gradient boosting methods, notably the XGBoost algorithm [46, 48], instead arrange decision trees in sequence, with each trained on the aggregated knowledge of those before it.

Many of the leading formation energy models present in the literature are built on a decision tree ensemble architecture. Ward et al constructed Matminer, a random forest model built on top of a machine learning data pipeline to predict formation energy among other material properties [38]. Figure 4(a) shows a ‘parity plot’ for this model trained on an accompanying data set of entries from the OQMD [9]. In this diagram, predicted formation energy values for compounds withheld from the model during training are plotted against the known DFT-calculated values. The result gives a visual estimate of the model’s accuracy, with perfectly accurate predictions approaching the identity (1:1) line. Meredig et al offered an alternative approach to these decision-tree-based predictions by developing a rotation forest ensemble model specifically created to predict formation energy [12]. This architecture divides initial data into random subsets, with principal component analysis subsequently applied to each subset [47, 49]. Another rotation forest-based formation energy model also developed by Ward et al expands on Meredig’s composition-derived feature sets by presenting the Materials Agnostic Platform for Informatics and Exploration (Magpie) library for the computation of materials attributes [37].

All of these decision tree-based models have shown promising results for the anticipation of new synthetically accessible compounds. Meredig’s work, for example, evaluated a list of 1.6 million physically reasonable ternary candidate compounds and identified 4500 new phases predicted to be thermodynamically stable, eight of which are confirmed to be DFT stable when tested in various candidate structures. Ward applied their rotation forest model towards the anticipation of metallic glasses, undertaking a broad sweep.
through 24 million candidate ternary alloys and identifying eight likely new glass-forming phases. These results emphasize the accelerated rate at which machine learning models can explore phase spaces.

**Support Vector Machines (SVMs).** One of the oldest machine learning methods, the SVM concept, was introduced in 1995 by Cortes and Vapnik to classify complex data [50]. The defining quality of a SVM is that the input features, which may not be linearly separable, are transformed into higher dimensional spaces where they can be divided by a hyperplane as illustrated in figure 3(b) [51]. Only those data points closest to this hyperplane, dubbed the support vectors, control where the decision boundary is drawn [52]. This distinguishes support vector models from most other techniques, where the entire data set is used to determine the best fit.

While support vector techniques are often best-suited to classification models, Drucker et al first demonstrated the effectiveness of support vector regression (SVR) algorithms to predict continuous numerical targets with high-dimensionality data [53]. Lotfi and Zhang et al developed an SVR formation energy model designed to guide synthesis in underexplored compositional spaces [36]. This model was trained using a combined OQMD and MP dataset of 279,943 unique compounds encoded into features as a mean, minimum, maximum, and difference of 35 compositional attributes. Using this method, the authors reported a mean absolute error of 0.085 eV atom$^{-1}$, as highlighted by the parity plot in figure 4(b). To test its merit, predicted formation energy values were determined for a grid containing 253 compositions on a $Y$–Ag–$X$ ($X$ = B, Al, Ga, In) ternary intermetallic diagram. The convex hulls generated from these predictions guided the researchers toward the synthesis of $YAg_{0.65}In_{1.35}$, a new phase in the $Y$–Ag–In system. Lotfi and Zhang’s synthetic validation of their predictions, as opposed to using DFT, is somewhat unique and emphasizes the method’s applicability. Their result is encouraging for the future success of machine learning derived formation energy predictions and provides a pathway for discovering synthetically accessible compounds.

Considering the success of both tree-based and SVM methods, there is no single superior machine learning approach. This inspired Dunn et al to develop the Automatminer algorithm, which addresses the ambiguity between models by automating the selection process using a 28-fold regression model space [54]. The accompanying Matbench suite contains preprocessed and cleaned training data useful for predicting 13 different physical properties, including a general formation energy dataset of 132,752 compounds extracted from the MP. The prediction accuracy of Automatminer for formation energy yields a mean absolute error of 0.12 eV atom$^{-1}$ [54]. These errors cannot be directly compared between models using different training data sets, as is apparent from examining learning curves showing the change in prediction error with data set size [55, 56]. Nonetheless, Automatminer appears to perform similarly well to the other decision tree- and support vector-based models seen in the literature. Automatminer’s flexibility and streamlined process also provide access to novel researchers in the field who want to have access to robust machine learning models.

**Neural Networks.** Neural networks are yet another class of machine learning models that have become increasingly popular due to their extremely accurate predictions when trained on large datasets [57, 58]. A simple neural network consists of an interconnected series of nodes arranged in layers, where data is passed from the initial input layer, through a hidden layer or layers, and finally to an output node. The process by which data is passed between layers and nodes is learned through training, and optimization continues until a requisite condition is met. The strength of the network approach lies in its ability to manipulate
high-dimensional input vectors into accurate outputs, but networks can often lose most or all interpretability as data is passed through hidden layers. Additionally, unlike hyperparameter tuning in other approaches, there is no straightforward method to optimize layer construction, so searching for the ideal architecture is a laborious process. Still, the resulting models boast some of the highest prediction accuracies, and formation energy prediction models are no exception.

*ElemNet* is one such formation energy neural network presented by Jha et al [53]. This model is a 17-layer deep neural network trained on 275,778 unique compositional data points from the OQMD [9]. When benchmarked against a battery of other conventional machine learning models trained on this dataset, the authors found that *ElemNet* generates predictions with a mean error of only 0.055 eV atom\(^{-1}\). It was also used to produce true-to-experiment phase diagrams for the Ti–O and Na–Fe–Mn–O systems in holdout tests. *ElemNet* was further used to scan 450 million hypothetical quaternary compounds, and several phases predicted to be thermodynamically stable were confirmed using DFT. Another model, *Roost*, is a neural network for predicting formation energy trained with an OQMD dataset [59]. Goodall and Lee demonstrated a weighted graph representation for compositional descriptors and find that this feature construction allows *Roost* to achieve a staggeringly accurate mean error of 0.0241 eV atom\(^{-1}\). Both of these neural network models credit their high degree of accuracy with large training data sets and should see continued improvement as repositories of DFT calculated formation energies continue to grow.

### 3.1. *MatLearn*: a tool for the synthetic exploration of novel materials

The models discussed above show a variety of optimized approaches to predict formation energy through machine learning. Assuredly, improvements beyond the current capabilities of these highly complex algorithms like *Roost*, *ElemNet*, and *Automatminer* will only be possible through the growth of DFT data repositories or a significant innovation in feature or architecture design. A common thread among these methods is that they are designed for and by computational chemists and data scientists. Each has focused on maximizing the model’s quality using DFT to confirm predictions. Indeed, it is only through this lens that such achievements are possible. However, there remains a great need to develop a machine learning-guided synthetic workflow created for researchers with little to no knowledge of computational methods or model design. Designed to meet this need, *MatLearn* is a web-based machine learning model and visualization suite that guides the synthetic exploration of composition space through the rapid creation of predicted phase diagrams for an arbitrary chemical system.

*MatLearn* is modeled after the initial work by Lotfi and Zhang [36] and uses a training dataset that consists of compositional and thermodynamic data from the MP. The objective of *MatLearn* is to provide users with no prior understanding of the machine learning process and no knowledge of coding languages like Python access to this formation energy prediction tool. *MatLearn* is available for use online. From the home page, a user can select any combination of available elements and choose ‘Create Predictions’ to generate the machine learning output. *MatLearn* includes a visualization process for predicted binary and ternary systems.

To train the *MatLearn* model, the formation energy data for all available compounds in the MP was collected, and duplicate phases were manually removed, keeping only the item at each composition with the lowest DFT-calculated formation energy. Compositions including noble gas elements, Te, or an element with \(Z > 83\) (except for Th and U) were also removed, as they present significant challenges to synthesis. Each of the 87,614 phases in the finalized training dataset was next encoded into a vector of 225 composition-based features. The first 85 items of each vector are the one-hot encoded composition for the compound, as applied to the 85 elements considered. The rest of the included features correspond to the mean, maximum, minimum, and difference of 35 physical properties for the elements in each phase (full list of elemental physical property values used available at *MatLearn*).

*MatLearn* uses a random forest algorithm implemented in Python 3.7.6 using scikit-learn [60]. Hyperparameter values were optimized, with the final model using 50 estimators in the random forest ensemble and a cutoff of 48 max features. A parity plot for *MatLearn* comparing predicted and DFT-calculated formation energies for a holdout set is shown in figure 4(c), yielding a mean error of 0.0799 eV atom\(^{-1}\).

An example of a predicted binary phase diagram for the Y–In system is shown in figure 5(a). Here, for the binary system Y–In, the predicted formation energies are plotted against percent Y for compositions varying from elemental In (0% Y) to elemental Y (100% Y) in 5% steps. Predicted values are determined by averaging the estimates of \(n\) random forest regressors trained independently on subsets of the training data, where the user controls the value of \(n\). Shown as blue shading in the plot, the estimated standard deviation measures the model’s precision for any given prediction. Finally, the convex hull is drawn as a dashed orange line, passing through composition points (red circles) where thermodynamic stability is expected.
Figure 5. Predicted phase diagrams generated using MatLearn for (a) the Y–In binary system and (b) the Ca–Cu–Al ternary system. The circle size in the ternary plot indicates estimated uncertainty of the points on or near the convex hull.

MatLearn is also capable of creating predicted ternary phase diagrams using the Plotly graphing library [61]. Figure 5(b) shows an example for the Ca–Cu–Al phase diagram, where we see the triangular ternary phase diagram with elemental phases at each vertex, binary phases along outer edges, and ternary compositions in the central triangle. The first layer of this diagram shows the predicted values of formation energy interpolated into a heatmap, where blue corresponds to regions of high energy (less thermodynamically favorable) and red shows compositions expected to have lower, usually more negative values of formation energy (more thermodynamically favorable). The predicted convex hull for the system is drawn on top of this heat map as red dots are connected by phase area tie lines. Faded gray dots show compositions that lie within 0.050 eV atom\(^{-1}\) of the convex hull, which are considered thermodynamically feasible and, therefore, synthetically accessible, at least at elevated temperatures. The data points are drawn such that their radii are proportional to that prediction’s standard deviation: larger points show less precise measurements, while smaller points show predictions that are consistent between random forest models. Connecting points on the convex hull are tie lines that show the three phases into which compositions within a triangle are predicted to decompose. Finally, white ‘X’ marks are included on the plot to show experimental compounds found in PCD, indicating compositions that have already been investigated and demonstrated to be experimentally accessible.

MatLearn is not the most accurate method for predicting the formation energy using machine learning currently available. However, it is useful as a tool for the synthetic chemist to speed up the examination of underexplored phase diagrams. Researchers could identify the location of compounds suggested on (or near) the convex hull and use this information to guide their synthetic efforts. Critical in its use is also the understanding that predictions from MatLearn are useful for guiding synthesis but do not imply physical reality. For example, regions often show physically unrealistic features such as dense clusters of compositions expected to be on the hull. This prediction is still useful for directing synthetic efforts towards (or away from) areas of the phase diagram.

In the Ca–Cu–Al phase diagram, shown in figure 5(b), this bears out. Using only compositional information, MatLearn predicts a dense cluster of compositions on the convex hull in the phase diagram’s Al-rich region. Therefore, MatLearn would recommend a synthetic investigation avoiding Ca- and Cu-rich compositions. Indeed, recent results in the literature have identified the stable phase, Ca\(_3\)Cu\(_{7.8}\)Al\(_{26.2}\), in the Al-rich region of this system, as predicted [62]. Of the three ternary Ca–Cu–Al phases contained in the MatLearn training data (Ca\(_3\)Al\(_2\)Cu\(_2\), Ca(Al\(_2\)Cu\(_4\)), and Ca\(_{10}\)Al\(_2\)Cu\(_7\)), only Ca\(_3\)Al\(_2\)Cu\(_2\) is in the Al-rich region of the phase diagram and exists at a significantly different composition than the discovered Ca\(_3\)Cu\(_{7.8}\)Al\(_{26.2}\). Thus, the formation energy prediction of Ca\(_3\)Cu\(_{7.8}\)Al\(_{26.2}\) demonstrates MatLearn’s ability to identify a synthetically realizable intermetallic phase without prior knowledge of its stability. In this way, MatLearn can accelerate the realization of energy-relevant materials by guiding synthetic efforts towards regions of the phase diagram likely to contain novel compounds.
4. Perspective and outlook

In addition to the formation energy models described above, machine learning has shown many promising experimentally-confirmed results for predicting crystal structure type and materials property predictions. For example, Oliynyk et al. have tested a model for the structural categorization of equiatomic ternaries through synthesis and characterization of TiFeP, finding that the model correctly identifies it is structural polymorphism [63]. Oliynyk and Gzyl additionally investigated the atomic ordering and stability of Heusler and half-Heusler materials, identifying two new arrangements and six novel ternary compounds flagged by the machine learning model [64–66]. Other studies have used machine learning models to produce a plethora of experimentally verified materials with ferroelectric [67], hydrogen storage [68], and superhard properties [69], or those with applications to batteries, lighting, and other devices [70–72]. Yet, the field of materials informatics is still in the early stages of development.

One possible mechanism for increasing a machine learning model’s predictive power is to increase the information from which the model can learn by expanding the feature space. The various models discussed above rely only on compositional features, which is ideal for predicting systems where only chemical composition is known. However, models that incorporate additional non-compositional descriptors, especially structure descriptors, are substantially more accurate than compositional methods alone. For example, Ward et al. report that the addition of a Voronoi tessellation descriptor to an otherwise composition-based random forest model results in a mean absolute prediction error of 0.08 eV atom$^{-1}$ [41]. Such models promise to speed up the search for materials, yet their applications to phase diagram exploration where the crystal structure is unknown remain a challenge. One idea to address this need is by combining structure prediction via methods like CALYPSO [73], USPEX [74], IM$^2$ODE [75], and others [76] with machine learning prediction methods and then using the resulting structure to generate features for the model. While encouraging, this approach requires errors in formation energy predictions with much lower error than currently available machine learning models, limiting its success thus far.

Another essential consideration for predicting formation energy is the subtle difference between predictions of formation energy and the true thermodynamic stability. The latter is quantified by the distance between a compound’s energy and the surface of the convex hull. As pointed out by Bartel et al. [30], computing this distance by subtracting DFT-calculated formation energy values benefits from systematic cancellation of DFT error, resulting in highly accurate predictions of thermodynamic stability. However, because errors in machine-learned formation energies are random and not systematic, when they are used to create a convex hull there is no analogous beneficial cancellation. The result is that the machine learning-based predictions of distance from the convex hull, and therefore stability, are significantly worse than their DFT counterparts. This effect can be seen in Ward’s Matminer (figure 6(a)), Lotfi & Zhang’s SVM model (figure 6(b)), and MatLearn (figure 6(c)). Here, each model is trained on an OQMD dataset to predict DFT-calculated distance from the hull directly. Though the mean errors of these distance-from-the-hull predictions are roughly the same as those of formation energy, the quality of prediction suffers significantly due to the values of the former value being much smaller than those of latter. In addition, distance above the hull is not only determined by the energy of an individual compound, but also the energies of compositionally similar phases, making the prediction less accurate. These results suggest that predicting thermodynamic stability with machine learning will be less reliable even though machine learning can accurately reproduce DFT formation energy.

An additional limitation of applying machine learning formation energy predictions to guide experimental research is the high prevalence of false-positives. As discussed for MatLearn and the Ca–Cu–Al system, the machine learning algorithm dramatically overestimates the number of compositions lying on the convex hull. This is represented by the many closely related compositions predicted to be thermodynamically stable (shown by the clusters of red circles). While this is not physically impossible, in practice, one of these phases often dominates the rest. Usually, these batches of false positives occur in areas where the model predicts something should form, but the model is not sure exactly where the energy surface is most favorable. If the training data includes a specific composition in this region of the phase diagram, the model is more likely to recognize this known phase as stable. However, in areas with little or no data, the model may lack the ability to sharpen the potential energy surface features and identify distinct minima. Still, this result can serve to inspire and guide investigations of underexplored phase diagrams.

These thermodynamic predictions are all made without considering entropy. As a result, the machine learning predictions are restricted to 0 K. Because phase diagrams are not invariant across temperatures, these 0 K phase diagrams may mislead synthetic efforts. Ideally, new machine learning models could be trained to predict $\Delta G$ directly. However, this requires training entropic- and work-related components that were previously assumed negligible. Unfortunately, the necessary thermodynamic training data are still sparse, although the compounds with calculated phonon values in repositories such as the MP continues to
grow. Some models like the MatErials Graph Network have shown positive results for predicting $\Delta G$ for small organic molecules with graph networks [77], but a model that can predict thermodynamic stability as the temperature is varied is still needed for solid-state materials.

5. Conclusions

The achievements of computational chemistry have been a boon for innovation in energy-related materials; however, the demand for ever-higher functioning materials continues to grow. Machine learning methods represent a solution to this problem: a shift to a new data-driven paradigm enabled by the expansion of materials chemistry data repositories. In particular, accurate predictions of formation energy can help inspire the synthesis of new functional materials. Multiple published machine learning models have shown the ability to make high-quality predictions of formation energy-based solely on compositional information despite differences in feature selection process and model architecture. To supplement these works, MatLearn was developed to be an openly available web-based application for directing synthetic exploration of solid-state phase diagrams through machine learning. MatLearn uses a random forest model trained on MP data and includes a visualization method for predicted binary and ternary phase diagrams. These models have all shown the ability to accelerate synthetic efforts by anticipating compositions and compositional regions likely to contain new materials. In the rapidly developing field of materials informatics, machine learning models are a key advancement for discovering the next generation of energy materials.

Acknowledgments

The authors gratefully acknowledge the generous financial support provided by the University of Houston Division of Research through a High Priority Area Research Seed Grant. The authors also thank the National Science Foundation (DMR-1847701), the Welch Foundation (E-1981), and the Texas Center for Superconductivity at the University of Houston (TcSUH). The authors would also like to thank the team at LaPraim Digital Agency for collaborating on the development of MatLearn.org. The authors declare no competing financial interest.

ORCID iDs

Gordon G C Peterson https://orcid.org/0000-0003-3638-8652
Jakoah Brgoch https://orcid.org/0000-0002-1406-1352

References

[1] Noe M and Steurer M 2007 High-temperature superconductor fault current limiters: concepts, applications, and development status Supercond. Sci. Technol. 20 R13
[2] Hull J R and Murakami M 2004 Applications of bulk high-temperature superconductors Proc. IEEE 92 1705–18
[3] Kalsi S S 2011 Applications of High Temperature Superconductors to Electric Power Equipment (New York, NY: Wiley)
[4] Snider E, Daeonbrylic-Gammnon N, McBride R, Debsaei M, Vindana H, Vencatasamy K, Lawler K V, Salamat A and Dias R P 2020 Room-temperature superconductivity in a carbonaceous sulfur hydride Nature 586 373–7
Agrawal A and Choudhary A 2016 Perspective: materials informatics and big data: realization of the 'fourth paradigm' of science in materials science Appl. Mater. 4 053208

Iain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D and Ceder G 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation Appl. Mater. 1 011002

Bergnerhoff G, Brown I and Allen F 1987 Crystallographic databases vol 360 (Chester: International Union of Crystallography) pp 77–95

Villars P 2007 Pearson’s Crystal Data®: Crystal Structure Database for Inorganic Compounds (Materials Park, OH: ASM International)

Kirklin S, Saal J E, Meredith B, Thompson A, Doak J W, Aylol M, Ruhl S and Wolverton C 2015 The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies npj Comput. Mater. 1 1–15

Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor R H, Nelson I J, Hart G L, Sanvito S and Buongiorno-Nardelli M 2012 AFLWLIB. ORG: a distributed materials properties repository from high-throughput ab initio calculations Comput. Mater. Sci. 58 227–35

Dralç I and Scheffler M 2019 The NOMAD laboratory: from data sharing to artificial intelligence J. Phys. Mater. 2 036001

Meredig B, Agrawal A, Kirklin S, Doak J, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 Combinatorial screening for new materials in an unconstrained composition space with machine learning Phys. Rev. B 89 094104

Draxl C and Scheffler M 2019 The NOMAD laboratory: from data sharing to artificial intelligence J. Phys. Mater. 2 036001

Isayev O, Fourches D, Muratov E N, Oses C, Rasch K, Tropsha A and Curtarolo S 2015 Materials cartography: representing and visualizing large materials databases J. Phys. Mater. 77–95

AFLOWLIB. ORG: a distributed materials properties repository from high-throughput ab initio calculations Comput. Mater. Sci. 58 227–35

Materials Genome Initiative: a materials genome approach to accelerating materials innovation Energy AI 1 1–11

Juneja R, Yumnam G, Satsangi S and Singh A K 2019 Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity Chem. Mater. 31 5145–51

Ando H, Yonemitsu Y and Ueno T 2018 Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm npj Comput. Mater. 5 1–11

Carrete J, Li W, Mingo N, Wang S and Curtarolo S 2014 Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling Phys. Rev. X 4 011019

Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D and Ceder G 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation Appl. Mater. 1 011002

Norris A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D and Ceder G 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation Appl. Mater. 1 011002

Jain A and Bligaard T 2018 Atomic-position independent descriptor for machine learning of material properties Phys. Rev. B 98 214112

Juneja R, Yumnam G, Satsangi S and Singh A K 2019 Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity Chem. Mater. 31 5145–51

Lejaeghere K, Van Speybroeck V, Van Oost G and Cottenier S 2014 Error estimates for solid-state density-functional theory calculations J. Chem. Phys. 140 124107

Lotfi S, Zhang Z, Visswanathan G, Fortenberry K, Tehrani A and Brgoch J 2020 Targeting productive composition space through machine-learning-directed inorganic synthesis Matter 3 261–72

Ward L, Agrawal A, Choudhary A and Wolverton C 2016 A general-purpose machine learning framework for predicting properties of inorganic materials npj Comput. Mater. 2 1–7

Draxl C and Scheffler M 2019 The NOMAD laboratory: from data sharing to artificial intelligence J. Phys. Mater. 2 036001

Meredig B, Agrawal A, Kirklin S, Saal J E, Doak J, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 Combinatorial screening for new materials in an unconstrained composition space with machine learning Phys. Rev. B 89 094104

Wu S, Kondo Y, Kakimoto M and Ong S P 2020 A critical review of machine learning of energy materials npj Comput. Mater. 6 1–11

Jain A and Bligaard T 2018 Atomic-position independent descriptor for machine learning of material properties Phys. Rev. B 98 214112
