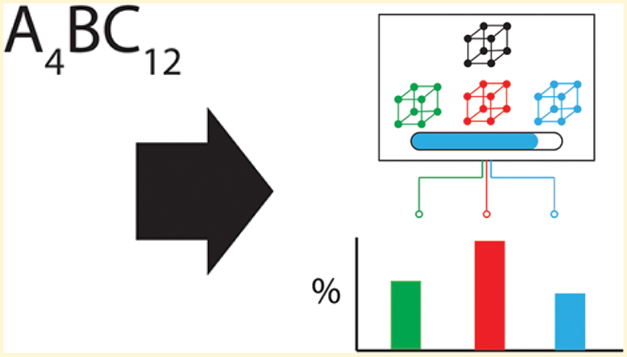
CHEMISTRY OF MATERIALS
Review
Cite This: Chem. Mater. 2018, 30, 3601−3612
pubs.acs.org/cm

Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons

Jake Graser,[†] Steven K. Kauwe,[†] and Taylor D. Sparks*,[†]

[†]Department of Material Science and Engineering, University of Utah, Salt Lake City, Utah 84112, United States

Supporting Information



ABSTRACT: Predicting crystal structure has always been a challenging problem for physical sciences. Recently, computational methods have been built to predict crystal structure with success but have been limited in scope and computational time. In this paper, we review computational methods such as density functional theory and machine learning methods used to predict crystal structure. We also explored the breadth versus accuracy of building a model to predict across any crystal structure using machine learning. We extracted 24 913 unique chemical formulas existing between 290 and 310 K from the Pearson Crystal Database. Of these 24 913 formulas, there exists 10 711 unique crystal structures referred to as entry prototypes. Common entries might have hundreds of chemical compositions, while the vast majority of entry prototypes is represented by fewer than ten unique compositions. To include all data in our predictions, entry prototypes that lacked a minimum number of representatives were relabeled as “Other”. By selecting the minimum numbers to be 150, 100, 70, 40, 20, and 10, we explored how limiting class sizes affected performance. Using each minimum number to reorganize the data, we looked at the classification performance metrics: accuracy, precision, and recall. Accuracy ranged from 97 ± 2 to 85 ± 2%; average precision ranged from 86 ± 2 to 79 ± 2%, while average recall ranged from 73 ± 2 to 54 ± 2% for minimum-class representatives from 150 to 10, respectively.

INTRODUCTION

Scientific exploration into chemical whitespace has always been a challenging process due to the high risk, high reward nature of research into untested territory. Materials discovery and characterization is a very time intensive process. Synthesis of untested materials requires a large amount of trial and error to determine optimum synthesis conditions with some chemical reactions taking days to weeks to perform. Many of these untested materials use exotic elements or compounds which can be expensive. In addition to the cost of reagents, samples must then be characterized for crystal structure and microstructure. Techniques such as diffraction, spectroscopy, and electron microscopy can be very time intensive.

Once a material is finally synthesized and characterized, its properties can be evaluated in the engineering design process. However, most applications require an optimization of multiple properties which may be interrelated. If we look at the field of thermoelectrics, for example, materials are compared to one another using a figure of merit, $zT = \sigma S^2\kappa^{-1}T$, where S is the Seebeck coefficient, σ is the electrical conductivity, κ is the thermal conductivity, and T is temperature. The material properties σ, κ, and S are all interrelated. For example, electrical conductivity requires high carrier concentration, whereas Seebeck coefficient requires low carrier concentration to increase zT. In addition, thermal conductivity also increases with carrier concentration which in turn decreases zT. Therefore, optimization of thermoelectric materials requires a compromise between these properties. Some of the most significant advances in this field have come from identifying new compounds which exhibit a better intrinsic balance in these properties.

The need to discover new materials is not unique to the field of thermoelectrics. Similar challenges are seen across many material science fields such as superconductivity,[1,2] lithium-ion batteries,[3,4] solid oxide fuel cells,[5,6] catalysts,[7] high strength materials,[8,9] and others. In these fields, a relatively small number of materials is being actively investigated compared to the tens of thousands of known potential compounds in databases such as the Inorganic Crystal Structure Database (ICSD). In many instances, some of the most exciting and promising new materials have been discovered via fortuity and luck. Critical engineering materials such as vulcanized rubber,[10] Teflon,[11] and synthetic plastics[12] to everyday luxuries such as artificial dyes,[13] super glue,[14] and synthetic sweeteners[15] were all discovered though chance.

This current approach to materials discovery and deployment is far too slow and expensive to meet the demands that we face in the 21st century. Instead, we need a rational and structured method to explore chemical whitespace. This new method

Received: December 22, 2017
Revised: May 18, 2018
Published: May 19, 2018

ACS Publications

DOI: 10.1021/acs.chemmater.7b05304

needs to be not only economical but also quick, precise, and accurate.

Consider The National Academy of Engineering's Grand Challenges. These include such challenges as making solar energy economical, providing access to clean water, or developing carbon capture and sequestration methods, among others.[16] Solutions to these challenges will undoubtedly require radically improved materials to be developed as quickly as possible. To this end, the President of the United States implemented The Materials Genome Initiative (MGI) in 2011 to deploy new materials "twice as fast at a fraction of the cost."[17] The MGI proposes to achieve this goal by enhancing collaboration between experimental and computational materials scientists. Computational resources can screen and reduce the total number of experiments necessary rather than experimentally testing every composition. Although the MGI has been in existence for a only short time, we are already seeing key successes from techniques rooted in MGI principles. For example, the Ford Motor Company has employed an MGI-based approach known as Integrated Computational Materials Engineering (ICME) to reduce the time for deploying engine aluminum casting, saving them a hundred million USD.[18] Other examples include QuesTek Innovations using ICME to develop new aviation components such as high strength steel for landing struts or helicopter rotors,[19] GE Aviation developing gas turbine components without rhenium to reduce cost,[20] Ford and General Motors researching materials to improve powertrain castings,[21] and investigation into distortions caused by welding in ship building.[22]

Critical to most MGI-based techniques is knowing the crystal structure of a candidate material a priori and then using this structure to calculate performance for a given property. Indeed, understanding the specific relationships between crystal structure, processing, and materials properties such as electrical, optical, and mechanical performance is at the heart of the materials science discipline. However, predicting crystal structure itself for any given composition has been a surprisingly vexing challenge for materials scientists, chemists, and physicists for over a century.

While some general rules have been identified which offer insight, such as Pauling's five rules for crystal structures,[23] there are numerous exceptions to these rules, and predicting the structure of some simple and most complex compounds still challenges scientists today.[24] Pauling's rules are as follows: (1) the radii ratio and radius sum rule for polyhedra formation; (2) the electrostatic valence principle for electroneutrality related to the coordination number of the cation; (3) the stability of the crystal related to polyhedra sharing of corners, edges, and faces; (4) the lack of sharing of polyhedra when multiple cations with large valence and small coordination number are present in the crystal; and (5) the multiplicity of constituents in the crystal will be small. Later, in the early 1980s, Pierre Villars built multiple three-dimensional stability diagrams by the determination of three specific atomic properties to help separate binary and ternary alloys. By using the difference of Zunger's pseudopotential radii sums, a difference in Martynov–Bastsanov electronegativity, and the sum of the valence electrons, Villars was able to build a predictive model to predict thousands of binary and ternary compounds.[25−28] Modern day prediction techniques now rely heavily on computational materials science and take many forms such as simulated annealing, genetic algorithms, and density functional theory (DFT).[29,30]

In this paper, we performed a literature review of multiple algorithms with an overview of the basics of each algorithm, a brief focus on the history, key breakthroughs, and modern examples of crystal structure prediction. We then discuss new approaches for crystal structure predictions based on machine learning. This paper gives an overview of the promise and challenges in using machine learning to predict structures.

1.1. Energy-Based Algorithms for Predicting Crystal Structure.

DFT, simulated annealing, and genetic algorithms all require a crystal structure suggestion or a randomly generated atomic configuration as a starting point to begin calculations from first principles.[29] The algorithms then search for the lowest energy structure using energetic potentials unique to each algorithm.[29] The lowest energy states are assumed to be the ground energy states and thus a compound's most likely thermodynamically stable crystal structure. It is important to note that due to these algorithms' focus on energy minimization, only ground state structures can be calculated. These algorithms cannot be used to determine metastable states or structures that require external temperature or pressure to remain stable.

1.1.1. Density Functional Theory.

Density functional theory is the most well-known predictive algorithm currently used by material scientists and researchers. Density functional calculations investigate the electronic structure of many-body systems at the ground state. Rather than simulating the interaction between every subatomic particle, it uses approximations. These approximations are nucleonic potentials for atoms and use an electron density rather than calculating each individual electron interaction. DFT achieves relatively high accuracy though the quantum mechanical modeling of the spatially dependent electron density in a system. Modern DFT is based on noninteracting electrons moving in a system-wide electronic potential. The potential is constructed using the structure and elemental composition of the system with their interelectronic interactions. The potential is evaluated to determine the energy cost for each state, or configuration, of the system. The correct ground state electron density is determined when the energy of the system has been minimized. DFT requires knowing a candidate crystal structure before making a calculation. Therefore, researchers will create a list of possible structures and then use DFT quantum calculations to determine each structure's energy at zero kelvin.[31] The lowest energy is then determined to be the most stable, and thus most likely structure.[31] Therefore, this is a zero-kelvin approximation and does not work for high temperatures which can include room temperature.[32] Researchers have merged different techniques, such as molecular dynamics,[33] to overcome this issue.

The pioneer of DFT was Douglas Rayner Hartree when he created a self-consistent field for electrons to solve for the wave function using the field around the nucleus.[30] This was expanded by two of his graduate students, Fock and Slater, in 1930 by replacing the equation with a determinant.[30] This became the Hartree–Fock equation and Slater determinants. Earlier, in 1927, Thomas and Fermi developed a model to calculate atomic properties.[30] This used a local density (LD) approximation for kinetic energy.[30] This set the foundation for solving the wave equation for atoms using density functionals. In 1964, Hohenberg and Kohn introduced a variational principle for energy, which showed a relationship between the ground state density of electrons and the wave function.[30] This was a huge accomplishment and the start of modern DFT. Formalism and refinement of the densities helped refine the algorithm. Yet, acceptance of this algorithm did not occur until around 1990 due to wariness within the field of chemistry.[30] The creation of simple to use software packages greatly improved DFT acceptance and use.[30] Between 1990 and 2015 there has been over 160 000 publications in the field of chemistry alone.[30]

DOI: 10.1021/acs.chemmater.7b05304
Chem. Mater. 2018, 30, 3601−3612

Yet, DFT is not without its shortcomings. Computational costs remain a large issue for each DFT calculation, while numerous tests must be run to determine the proper energy functional as different approximations will affect the outcome.[30] DFT also struggles with highly correlated electron systems, large scale systems, modeling weak or van der Waals forces, time-dependent dynamics, and properties that are not observable at the ground state such as excited states or room temperature bandgap energy.[32] Critically, DFT also cannot calculate disordered structures necessary for many unique material properties (partial cation occupancy, oxygen vacancies, etc.). Recently, better programmed algorithms and approximations coupled with the introduction of machine learning have helped offset the computational cost and time requirement limitations.[32,34]

Density functional theory has had many successes in material property predictions ranging from batteries,[35,36] capacitors,[37] thermoelectrics,[38−40] superconductors,[41,42] photovoltaics,[37,43] and catalysts.[44] DFT calculations continue to improve accuracy. and upward of 15 000 density functional theory papers are published every year.[45]

1.1.2. Force Field Models. Another method to calculate the energy of atomic configurations are empirical force field models. These simulate interatomic interactions in unique ways depending on the model.[46,47] These models are used when the system grows in complexity where ab initio calculations become too computationally demanding.[48] These models are considered critical for nanoparticle structure predictions to reduce computational time due to the large size of each predicted system.[46,47]

However, the accuracy of the calculation is heavily dependent on which model is used as well as the accuracy of the model's approximations.[48] Swamy et al.[48] used two separate force field models to predict all known polymorphs of TiO_2 with varying success depending on the specific polymorph. It was shown that one model could predict low pressure polymorphs accurately but struggled with high pressure polymorphs, while the other model could predict high pressure polymorphs accurately but had difficulty with low pressure polymorphs.

1.1.3. Global Energy Optimization Algorithms. Simulated annealing is a computational approach that is based on the process of physical annealing wherein a material is heated to modify its crystal or microstructure. Modifying the crystal and microstructure requires atomic mobility by overcoming energy barriers. This is made possible by increasing the temperature. These atoms then settle into their lowest energy state in the crystal at the given temperature. This allows the atoms to move and adjust the crystal structure to reach the minimization of the Gibbs function assuming constant temperature and volume. In simulated annealing, the same concept is applied. We allow these atoms to settle into their lowest energy state as their simulated energy, commonly labeled temperature, is slowly decreased in the model. It is important to note that this temperature is an arbitrary energy unit and not related to actual temperatures.[49] Minimum energy is found by randomizing the motion of simulated atoms using either Monte Carlo statistics, molecular dynamics, or other techniques.[29] The initial temperature is selected so that the kinetic energy of each atom is high enough to allow the system to overcome local energy barriers so global minima can be found.[29] Simulated annealing has had success in predicting inorganic structures and can make predictions of partially disordered materials, something DFT cannot do.[24,49,50]

First introduced by Kirkpatrick in 1983, a relation of physical annealing and statistical mechanic relations lead him to introduce the algorithm.[51] By using the algorithm, he showed predictions for the classic traveling salesman problem to the physical determination of wiring and cooling within a computer.[51] The algorithm was quickly adopted due to its robust nature and ease of use.[29,52] The algorithm was expanded upon with new techniques to increase the accuracy of the algorithm such as automated assembly of secondary building units and a hybrid method using Monte Carlo basin hopping.[29]

However, simulated annealing is computationally intensive, which led to many researchers developing workarounds such as parallelization processing techniques.[52] The most concerning problems with simulated annealing are slow energy landscape exploration, the inability to focus on specific problems of interest due to the randomization of atomic placement, and computational limitations.[29,53] In theory, simulated annealing can explore the entire energy landscape and find the energy minimum regardless of starting position, but the time and computational resources required for this complete exploration can be excessive.[54] The algorithm must start in a specific point, and it could miss low energy regions as the algorithm reduces temperature. To offset this limitation, many simulated annealing runs are done sequentially, and different starting spots are selected to better explore the energy landscape.[29]

Genetic algorithms are another energy-based technique used by researchers to predict crystal structures. Created by John Holland in 1975, genetic algorithms are a subset of evolutionary algorithms.[55] These algorithms are based on the concept of evolution where the strong can procreate offspring while the weaker will not. An initial population of structures is generated with constrained but randomized atomic placement or prearranged atomic configurations.[56] The algorithm then selects parents from the population to exchange crystallographic information. This is done using a random selection code such as tournaments or a roulette wheel, where the percentage chance of selection is dependent on how well it meets certain criteria specified by the algorithm, referred to as a fitness function.[29] The parent structures share information in either a random pattern or by mixing together, which results in an offspring structure.[56] This process repeats until the desired number of offspring is achieved. Mutations can be introduced to add diversity to the population by changing random properties of the crystal structures. Mutations and offspring then have their individual energies minimized and are added to a new data set called a generation. This new data set also includes the samples of the data set that are lower in energy than the new offspring and mutants. The old generation is replaced, and the process is continued until the energy converges to some final criteria or by reducing the training set at each generation until only a single crystallographic structure remains.[29,56] The practice of applying genetic algorithms to material science has been growing rapidly in the past few years.[56−59] After their initial introduction, multiple variations of genetic algorithms have been introduced such as adaptive genetic algorithm.[59] Specific examples for crystal structure predictions exists such as global spacegroup optimization known as GSGO[58] and the genetic algorithm for structure and phase prediction also known as GASP.[57]

A popular technique for the creation of offspring structures is the cut and splice method.[60] This method creates a new generation by splitting a chosen structure at an arbitrary plane. Members of the generation can be sliced and combined to form a new randomized structure, or offspring. The cut and splice method can also be used to generate mutations by rotating a section of the given structure at an arbitrary angle. These new structures have their energy minimized and are added to the new generation.[60]

DOI: 10.1021/acs.chemmater.7b05304
Chem. Mater. 2018, 30, 3601−3612

Repeating this process allows us to optimize the crystal structure with each new generation eventually reaching a minimum.

Like all optimization techniques, the selection of algorithm parameters will affect performance. For example, convergence to a global minimum can be discovered if the initial population size is large enough, the creation of the offspring population is set at an appropriate rate to explore space but not to over saturate a local minimum, and a mutation rate is significant enough to shift the algorithm out of local minimums. However, this leads to large computational times. Thus, a promising region should be determined with a small population size to reduce computational time. If selected incorrectly, this can lead to finding only a local minimum.[61]

There are many simulation software packages available for these algorithms mentioned above. The most common software package used for DFT in crystal structure predictions is the Vienna Ab Initio Simulation Package, or VASP for short. Historically, it has had success in crystal structure predictions.[31,34,57,59,62,63] As mentioned above, DFT needs a starting structure to calculate energy. To generate this starting structure for DFT, there are multiple methods and algorithms available. USPEX (Universal Structure Predictor: Evolutionary Crystallography) has been used to determine high pressure phases of $CaCO_3$,[64] while CALYPSO (Crystal Structure Analysis by Particle Swarm Optimization) is another software package that recently has been used to determine structures of noble gases at high pressure.[63] Finally, the AIRSS (Ab Initio Random Structure Searching) method, which has been used for determining the crystal structure for lithium-based crystal structures.[65,66]

1.2. Machine Learning Algorithms for Predicting Crystal Structure. All of the previous techniques involve calculating and comparing energies for crystal structures to make predictions about which crystal structures are most thermodynamically stable. This stability is with respect to certain given conditions, defined by either the algorithm or user, such as constant entropy or constant volume. A fundamentally different approach exists which relies on machine learning. Machine learning is a heavily data-centric technique where large amounts of data are collected and analyzed. A predictive model is then trained on this large collection of data. This model can then be fed inputs similar to the collected data to predict probabilistic results. These inputs can be categorical as well as numerical. Thus, all supervised learning algorithms can be characterized as classification or regression problems. This ability to segregate crystal structure data of all types is key for allowing us to predict crystal structure.

Companies such as Amazon and Netflix already collect an enormous amount of data related to consumer interest, browsing habits, viewing history, etc. and are already incorporating machine learning into their websites as highly efficient ways to recommend products or entertainment options to consumers. These algorithms do not need to know exactly why the consumer is interested, it only needs to predict the probabilistic likelihood of a consumer being interested. The mechanistic details of the relationships, which are essential to a technique such as DFT, are not even necessarily known in machine learning algorithms. Instead, only the probabilistic determination of a given outcome from input data matters. Yet, these algorithms should be viewed as tools to assist rather then to replace experimental and computational materials scientists. Ultimately, the algorithm will only make predictions, and some of these could be correspond to completely unstable or even physically impossible compounds. Therefore, chemical intuition will still need to be utilized to determine what is valuable and what to ignore.

Machine learning algorithms rely on building predictive models from empirical data or calculated data.[31,67] The data used for supervised machine learning is organized in tables referred to as training data sets. Each row describes a single entity or observation, and each column represents a commonly shared feature or attribute. We label these columns as features. These commonly shared features include a column which contains the key property you are attempting to predict, such as crystal structure. The column features can be numerical or categorical. A sample of features used to predict crystal structure could include compositional thresholds, bond character, or average number of valence electrons, among many others. For data to be useful in machine learning, each row needs a value for the features that will be included in the model. When features are missing too many values, they are generally discarded, although there are methods to estimate the missing values. For example, imputation is the procedure of replacing empty values in a data set.[68] Imputation is typically handled by filling empty cells with the mean for continuous numerical data, the mode for categorical number data, or the most common string for written categorical data. There are also more sophisticated procedures which involve building nested predictive models to fill in the missing attributes.[68]

Like the energy-based algorithms mentioned above, multiple machine learning algorithms exist such as random forest algorithms, support vector machines, and artificial neural networks. They all share the ability to use a collection of data to build a predictive algorithm. Each of these algorithms build prediction models in different ways, and the data requirements, such as size and formatting, differ per algorithm.

A popular machine learning technique used for predictive model building is the random forest algorithm.[69] The random forest algorithm utilizes many independent decision trees trained from collected data. Training is the process of using the input data to create a criteria-based prediction model that has predictive power. A decision tree is trained by using a subset of features from the data. The training process compares feature values for all inputs and attempts to segregate the input data. The features separate the input data at different feature values creating successively more homogeneous groups.[69]

As discussed above, there are two types of commonly used decision trees: classification and regression.[70] Classification decision trees create predictions that attempt to classify categorical data, an example being crystal structure such as fluorite, spinel, etc. Regression decision trees predict continuous numerical outputs, such as thermal conductivity. In the random forest algorithm, all the trees in the "forest" have different structure because they sample different data and random features.[70] The trees are composed of unique nodes and branches. The nodes are a way to represent splitting points in the data. The initial node is referred to the root of the tree. Feature values from the data are used to separate an input data group. The groups that result from separation are called branches. Each subsequent node receives an input group from the branch above it. That separation is output to nodes below it until all the groups are homogeneous. These final nodes are referred to as leaves. The tree structures that are built to separate the experimental data can then be used as a model for separating future data;[69] an example is shown in Figure 1.

Random forest has already been used as a high-throughput material screening process for thermal conductivity or energetically favorable compositions.[71,72] For example, Oliynyk et al.[73] built a model that predicted whether 21 ternary compositions were either full-Heusler, inverse Heusler, or not Heusler. Of these 21, 19 were confirmed though experimental results.

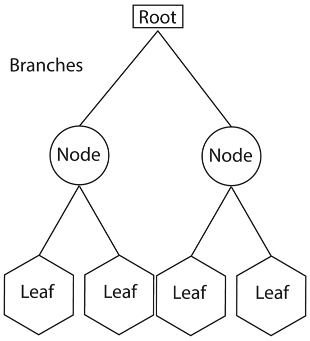


Figure 1. Graphical description of a decision tree with the root, node, and leaf sections labeled. The data is passed from one section to another along branches.

The challenge with differentiating these classes is that they all look nearly indistinguishable via powder diffraction, and single crystals are difficult to grow and therefore rarely used to characterize structure. Even with these difficulties, the algorithm had an accuracy of 94%. Balachandran et al. also used decision trees as well as support vector machines, which we will discuss below, to predict wide band gap binary structures as well as transition-metal intermetallic compounds. These algorithms had accuracies ranging from 86.7 to 96.7% for the decision trees and 86.7 to 93.3% for the support vector machines (SVM).[74]

An SVM is another machine learning algorithm based on the segregation of data. SVMs can segregate regression or classification data like the random forest model discussed above. SVMs accomplish this task by plotting the data into an *n*-dimensional space. The algorithm then attempts to create a hyperplane to segregate the data. This hyperplane is determined by maximizing the vector normal to the hyperplane, usually labeled *W*, and the closest data point to create the largest gap possible.[75] By graphing the data with respect to different physical properties, the algorithm can compare the hyperplane separation with respect to physical properties. The hyperplane with the largest split is the defining feature relative to the physical properties and thus the most important feature to segregate the data. To help with this process, an error function can be introduced to allow the algorithm to ignore a few data points to plot the hyperplane. The distance of this separation can be used as a method to optimize the algorithm.[76] A graphical example is shown in Figure 2.

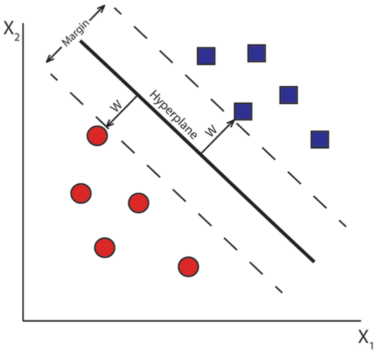


Figure 2. Graphical example of support vector machine. The data are being separated by the hyperplane that maximizes the vector *W*.

The strength of SVM algorithms is the ability to optimize itself by adjusting its kernel function and thus adjusting the dimensionality of the problem to help with segregation.[77,78] The kernel function is a symmetric and continuous function where, if the restrictions of Mercer's theorem is met, can define the dot product in a specific space.[79] This allows the program to increase dimensionality without calculating the dot product continuously, allowing the algorithm to expand its dimensionality without impacting computational time. Yet, this leads to the major disadvantage of SVMs. The choice of the kernel function is a critical and challenging process, and unlike certain other machine learning techniques, SVMs can still be computationally demanding depending on the dimensionality of the problem with respect to other machine learning algorithms.[75]

SVMs have been used in the prediction of protein sequences,[80] residue-position importance,[81] domain boundaries in protein structures,[82] microstructure imaging,[83] and atmospheric corrosion of materials.[84] SVMs have also been used in crystal predictions of binary and intermetallic compounds with success.[74]

With regards to predicting crystal structure specifically, Oliynyk et al. built a support vector machine to predict the crystal structure of binary as well as ternary compounds. The binary algorithm achieved an accuracy of 93.2% with a training set of 706 compounds. The authors provided further experimental validation by synthesizing one compound, rhodium cadmium, which was predicted to have the cesium chloride structure, and confirming via X-ray diffraction that the predicted structure was correct.[77] The ternary algorithm achieved an accuracy of 96.9% with 1556 unique compounds.[85]

Artificial neural networks (ANNs) are another machine learning method used in material informatics. ANNs are capable of modeling essentially any complex relationship given enough data.[86] ANNs tend to perform well for large amounts of data, experiencing performance saturation later than other machine learning models. They are particularly capable of dealing with data that must have spatial and temporal relations, such as images and text processing.

Artificial neural networks are based on a collection of connected units called neurons. Neural networks rely on layering of neurons to allow for processing of complex patterns. Most ANN models consist of an input layer, hidden neuron layers, and an output layer. ANN models work by processing input values though massive connected networks called hidden layers. Each connection in the network is called a synapse, and it transmits information to the neurons downstream. Information is passed starting from the input layer until the output neurons are reached. The neurons derive value from the synapse connections while also converting the data into nonlinear space if needed. ANN models are trained by adjusting the weights of each synapse until the output of the network is close to the output of the training data.[87] A graphical example is shown in Figure 3.

As of the writing of this work, neural networks are not used in crystal structure determination. There have been a few preliminary classifications related to structure such as protein secondary structure investigations during folding.[88,89] When it comes to inorganic crystal structures, neural networks mainly focus on data interpretation and categorizing. Recently, Timoshenko et al. built an artificial neural network to decipher metallic nanoparticle structures from experimental data.[90] Due to the large requirement of information required for artificial neural networks, they created a data set of simulations that were verified against experimental data. The accuracy of artificial neural networks allowed them to predict an average coordination number up to the fourth coordination shell and thus the size and shape of the nanoparticle.

Regardless of the type of machine learning algorithm utilized, success is measured by the ability to forecast and predict accurately. There are many different and unique ways to test the accuracy of a machine learning algorithm. One of the most

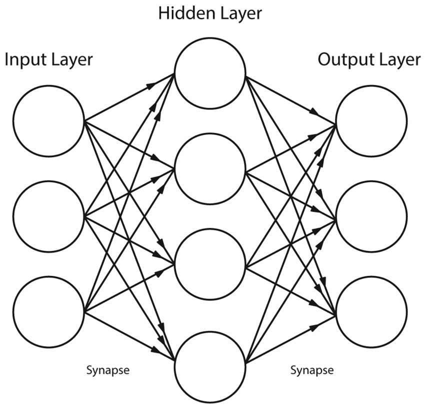


Figure 3. Graphical definitions of artificial neural networks with the input layer, the hidden layer, and the output layer. Each layer connects to each other layer though multiple paths called synapses.

common and simple methods is the *k*-fold cross-validation. The idea behind *k*-fold validation is the separation of the training data into *k* equal data sets. The algorithm is trained on $k - 1$ data sets and is used to predict the data set that was not used in the training. The actual values are compared to the predicted values. For a real valued property, root mean squared error is the commonly reported error metric.

$$\text{root mean square error} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

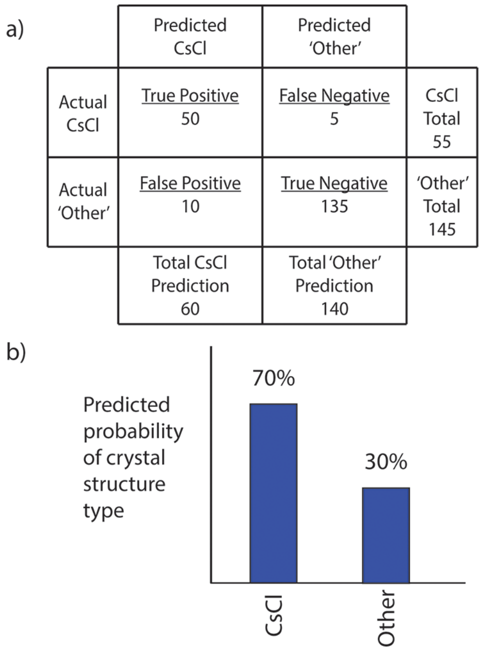
where $\hat{y}_i$ is the observed value, y_i is the predicted value, and n is the total amount of predictions performed by the algorithm. A quick way to examine your algorithm is by comparing the results to a random guessing algorithm. If the probability of the algorithm is the same or worse than a random selection of each class, then rebuilding the algorithms, focusing on features or shifting to a different type of algorithm, is required.

When model predictions are categorical as opposed to real values, a more useful accuracy evaluation tool is the confusion matrix. The confusion matrix compares the known values from the training set and the predicted values from the algorithm. From this you can compare how often the algorithm classifies the data correctly to determine its accuracy, in-class precision, recall, and false positives/negatives. The overall model accuracy is the ratio of the total number of correct predictions versus the total predictions. In class precision is the accuracy of a specific prediction in the model. In class recall is the number of times a class was predicted correct over the total number of instances of that class. A false positive or negative is when the algorithm is wrong in its prediction. We can define these equations of accuracy, in-class recall, and in-class precision as seen below. For clarity we define true positive as TP, true negative as TN, false positive as FP, and false negative as FN.

$$\text{in-class precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{in-class recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$



	Predicted CsCl	Predicted 'Other'	
Actual CsCl	True Positive 50	False Negative 5	CsCl Total 55
Actual 'Other'	False Positive 10	True Negative 135	'Other' Total 145
	Total CsCl Prediction 60	Total 'Other' Prediction 140	

Figure 4. (a) Example of a confusion matrix with CsCl defined as positive outcome and Other as a negative outcome. (b) Example of predicted probability of a specific data point from the algorithm. Due to CsCl-type structure having the larger percentage, the algorithm would categorize this data point as CsCl-type structure.

To help illustrate the idea of a confusion matrix, let us consider an algorithm that predicts if the crystal structure will be cesium chloride-type structure (CsCl) or another category we shall label "Other" (see Figure 4). For false positives/negatives, we will define that the CsCl-type structure outcome is positive, and the Other outcome is negative. Out of 200 compounds, let us have 55 CsCl-type structures compounds and 145 Other compounds in our training set. Let us assume the algorithm predicts 60 CsCl-type structure compounds and 140 Other compounds. The accuracy of this algorithm would be the sum of the number of times it guessed correctly over the total training set. That would be 50 correct CsCl-type structure compounds plus 135 correct Other compounds divided by the total training set of 200 to give us 92.5% accuracy. The in-class precision for CsCl-type structure would be 50 divided by 60 or 83.33%. In-class recall for CsCl-type structure would be 50 divided by 55 or 90.91%. CsCl-type structure would have a false positive of 10, while Other would have a false negative of 5.

1.3. Synergy vs Competition in Energy-Based vs Machine Learning Approaches. Researchers have started using machine learning techniques to explore chemical white-space focused on crystal structure with success.[71,73,91,92] The databases of information online, such as the International Crystal Structure Database or the Pearson's Crystal Database, give large amounts of physical parameters that can be used to build training data sets. These can then be used to build a prediction algorithm. This is very attractive to researchers for high throughput material exploration. Yet, problems still exist within physical sciences with machine learning algorithms because large and diverse training sets are required as well as knowledge of coding and algorithm deployment. The building of training sets requires a large amount of time and effort, while energy-based algorithms still struggle with calculation time and cost. It is not surprising then that most researchers do a combination of each technique to offset the weakness of each type of algorithm. Using machine learning, researchers can search chemical whitespace quickly and single

DOI: 10.1021/acs.chemmater.7b05304

out interesting or promising materials. These are then fed into energy-based functionals or calculations to create a more refined prediction of the material properties and characteristics.[31,34,93−95] Others use these energy-based algorithms such as DFT to generate data sets for unknown or ill-defined chemical compounds upon which to train their machine learning data set.[96,97]

As described earlier with Heusler and basic binary structure predictions, machine learning has been used for a very select few specific crystal structure predictions. However, a general, universal structure type prediction algorithm has never been deployed using machine learning. Therefore, in this paper, we extend previous efforts to determine the extent to which machine learning could predict any crystal structure type. We accomplish this by training off all crystal structure data available in Pearson's Crystal Database to predict the structure for any composition.

2. METHODS

In this work, we use a machine learning algorithm from the open source program H2O FLOW. A database of 24 215 unique formulas and associated entry prototype (EP), Pearson symbol, space group number, phase prototype, etc. was assembled from the original 24 913 entries obtained from the Pearson Crystal Database. This was the result of removing formula with exotic elements such as polonium, astatine, protactinium, etc. These exotic elements lacked sufficient elemental properties for our machine learning method. These were organized into identification numbers in order of decreasing size. A graphical representation of the specific entry to the number of representatives per entry is shown in Figure 5. This was then screened for materials near room temperature (290−310 K) with duplicates removed. The chemical formula for each entry was then separated into composition-weighted elemental and atomic properties to allow the model to explore any chemical composition as all features were elementally-based. These formulas-based features were then uploaded into H2O FLOW and were then used as a training set in the random forest machine learning model. This model was selected due to its ease of use and scalability for the size of the data. Error metrics were calculated in accordance to the *k*-fold validation methods discussed above.

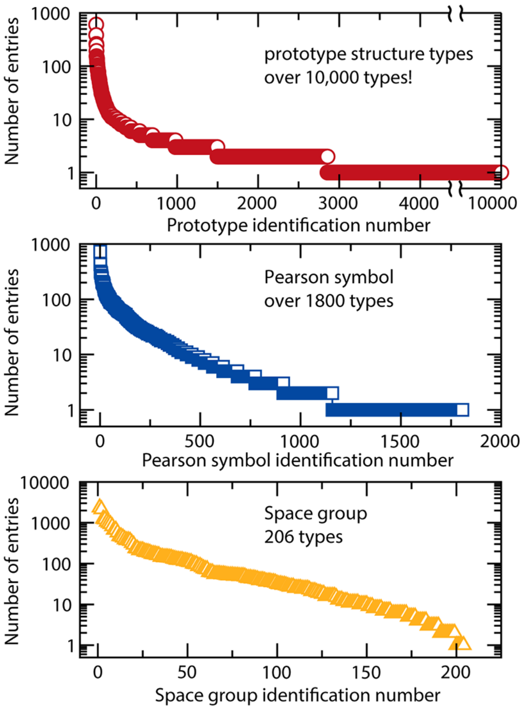


Figure 5. Histogram of entry size versus entry prototype, Pearson's symbol, and space group. The size of each category drops quickly with the majority of each category having only a few entries.

To make a given structure prediction, there must be multiple example compositions or instances having that structure type to correlate composition to a structure type. In our data set, there were 10 711 unique entry prototypes and 97.5% of the entry prototypes had fewer than 10 instances. However, a mere 2.5% of the entry prototypes, those most common structures such as perovskite or spinel, encompassed 28.5% of the data. This led us to question at what point, in terms of number of prototype entries, can we build accurate models. Moreover, because machine learning model accuracy generally increases with number of instances per class type, up to a point, we can study the expected trade-off between model breadth and model accuracy. Specifically, to handle the uneven distribution of entry prototypes, a minimum number of instances was set at an arbitrary cutoff. This cutoff was then varied for different models. Entry prototypes with fewer than the required number of instances were categorized into a single class named "Other". When the minimum cutoff value was varied between 150 and 10, the Other class encompassed between 92.5 and 64.1% of the input data, respectively. The database was prepared multiple times with separate cutoff values with minimum bin size of 150, 100, 70, 40, 20, and 10.

Although 97.5% the entry prototypes exist below the cutoff limit of 10, we still find the classification of Other to be useful information. With a cutoff limit of 10 entry prototypes, a prediction of Other leads to a rare crystal structure with less than 10 known similar compounds listed within the Pearson's Crystal Database. If a researcher is looking for a very rare crystal structure for a specific property, that crystal structure most likely will exist within the Other category. Moreover, the model is able to make accurate predictions with moderate recall for the most common crystal structure types.

The prepared data sets were all analyzed with a distributed random forest algorithm. All the algorithms had a limitation of 150 trees with a maximum depth of 40. Each prepared data set resulted in a unique model. Error metrics were calculated using fivefold cross-validation in accordance to the *k*-fold validation methods discussed above. Predictions were plotted as a confusion matrix.

For error analysis, each cutoff model was built with five different random seeds. The error metric we compared is the mean per class error. Mean class error is defined as one minus recall, as seen in Figure 6. The difference between the largest and smallest metric is calculated to determine error range.

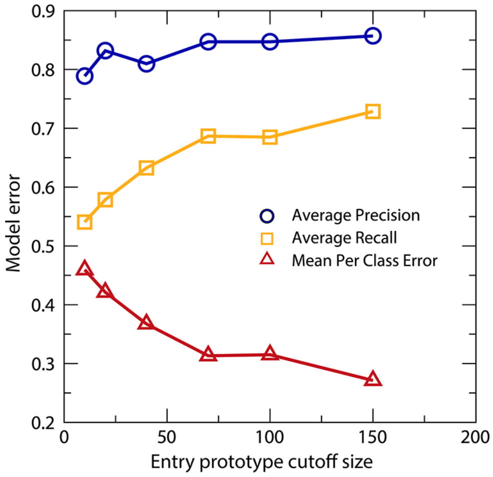


Figure 6. Model error with respect to cutoff size. Each point is a specific cutoff with guidelines inserted between points. As cutoff size increases, the model's overall accuracy increases, as expected. Error bars (2%) are smaller than marker size.

3. RESULTS AND DISCUSSION

Before describing model accuracy, we first remark on the model speed. As described in the Introduction, one of the key advantages of machine learning is the speed of prediction. In this model, we trained our algorithm on 25 000 different entries with 90 columns of metadata each. Therefore, our overall data set exceeded 2.2 million data values. Nevertheless, training the model on 25 000 entry prototypes only took up to 2 h depending on the minimum bin size. Once the model was trained, 15 000 entry prototypes were predicted in under 10 s on a laptop (Intel-I7, 2.6 GHz processor 16GB RAM, 64-bit Windows 10). These composition-based predictions included the assigned entry prototype (structure with highest probability) and a breakdown of the probabilities for each other entry prototype. The most probable class is selected as the final prediction. For example, the model output for $Cr_{0.12}Ru_{0.88}SrO_3$, which has the entry prototype $GdFeO_3$-type structure, would have a distribution across all possible entry prototypes. $GdFeO_3$-type structure has the highest probability, so it would be selected as the entry prototype. The graphical representation of these data can be seen in Figure 7.

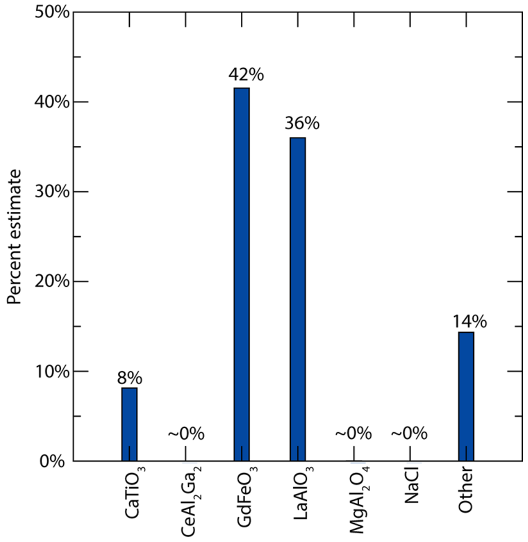


Figure 7. Graphical representation of the algorithms probabilities for entry prototype. With $GdFeO_3$ having the largest probability, it will be selected as the algorithm's entry prototype prediction.

For compounds with multiple crystal structures possible, the model predicts the most common structure at room temperature due to the model and training set having been built at room temperature.

To determine the error range of the model, each cutoff model was built with a different random seed five separate times. The range was determined by the difference in the largest and smallest error. This difference ranged from 0.5% for accuracy, 2% for recall, and 1.8% for precision. To be conservative with the error range, the largest error was adopted for all our percentage errors discussed below.

As expected, as the minimum cutoff size was reduced for each entry prototype, from 150 to 10, fewer data were available, and the mean in-class error increased slowly. This mean class error ranged from 27 ± 2 to 46 ± 2% for a minimum-bin size ranging from 150 to 10, respectively. In comparison, random guessing mean-class error ranged from 99.8 to 83.3%. If the algorithm only selected the Other category, its mean class error ranged from 99.9 to 86.7%. We can see that regardless of cutoff, all showed an overall error far lower than random guessing or only selecting Other. Although mean in-class accuracy describes the overall performance of the model, the reliability of a prediction is better understood by evaluating the accuracy for predicting individual entry prototypes. For the 150-cutoff section, the largest class error of the entry prototypes was $CaTiO_3$-type structure at 52%, while others are much more accurate such as $CeAl_2Ga_2$-type structure and $MgAl_2O_4$-type structure with classification errors of only 14 and 19%, respectively. To clarify, when we discuss classification errors, we are describing the percent of the time the algorithm categorized a prediction in the wrong category. For example, if the algorithm predicts a $CaTiO_3$-type structure prediction as a $GdFeO_3$-type structure, that would be a classification error. An alternative metric used is precision, or one minus the classification error. Overall, the in-class error is surprisingly low, even when we only include as few as 10 entry prototypes in training data with classes such as $BaNiO_3$-type structure with only six entries, which has 100% precision. However, some specific classes with only one or two entry prototypes predictions have zero precision. In other words, we see evidence that when the model thinks a composition belongs to a given class, it will predict it with very good precision, but in a significant number of cases where only one or two data points exist, it will just call it Other.

Some entry prototype predictions are more consistently correct than others. When these high-accuracy prototypes are predicted, we can have high confidence that the prediction is correct. If we consider the entry prototype cutoff of 10, we can see examples of these high-accuracy entry prototypes, including $CeAl_2Ga_2$-type structure, $MgCuAl_2$-type structure, and $CeNiSi_2$-type structure, which all perform at a precision above 90%, which is 20% above the average model precision. Similarly, we can express doubt for predictions involving entry prototypes that are frequently predicted incorrectly in the model. Low-accuracy entry prototypes are rarely valuable as predictions; some examples include TiNiSi-type structure, Th_3P_4-type structure, and BiF_3-type structure, which scored 20% below the model's average precision. Some classes with very few entries have precisions of 20% or lower, further confirming the benefit of larger amounts of representatives in data sets.

Although the average precision was stable, the average recall dropped off steadily with smaller cutoff sizes. The average recall ranged from 73 ± 2 to 54 ± 2%, showing that as the number of classes increased, the algorithms ability to classify the known training data diminished. This is due to certain classes having only one or two entries after the removal of exotic elements. These are usually categorized as Other, again showing the necessity of classes with many entries. An outline of the errors is shown in Figure 6.

A confusion matrix was generated for each model. The confusion matrix for the algorithm with a cutoff of 100 is shown in Figure 8. The error matrix for each class was trained with entry prototype imbalances in mind. This was done by normalizing the predicted value by the total number of predictions in the class. In this paper, we focus on the confusion matrix with a bin size of 100 due to the large confusion matrix generated for smaller bin sizes. The algorithm with a cutoff of 100 showed a mean precision of 85 ± 2% with a mean recall of 68 ± 2%. In other words, the average ability for the algorithm to correctly predict a certain structure was 85 ± 2%, while the average ability for the algorithm to predict a true positive rate was 68 ± 2%. To clarify further the idea behind recall and precision, let us look at $CaTiO_3$-type structure in the entry prototype data set with a cutoff of 150. Out of 162 guesses, the algorithm classified $CaTiO_3$-type structures correctly 123 times. This gives the precision of 123/162 or 76%. The recall is the

	$Ca(Ca_{0.5}Nd_{0.5})_2NbO_6$	$Ca_2Nb_2O_7$	$CaTiO_3$	$CeAl_2Ga_2$	Cu	CuZrSiAs	FeAs	$GdFeO_3$	K_2NiF_4	$LaAlO_3$	$MgAl_2O_4$	$MgCu_2$	NaCl	$NaFeO_2$	TiNiSi	Other	Recall
$Ca(Ca_{0.5}Nd_{0.5})_2NbO_6$	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	0.686
$Ca_2Nb_2O_7$	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0.655
$CaTiO_3$	0	0	133	0	0	0	0	12	0	5	0	0	0	0	1	105	0.522
$CeAl_2Ga_2$	0	0	0	161	0	0	0	0	0	0	0	0	0	0	0	28	0.847
Cu	0	0	0	0	56	0	0	0	0	0	0	0	0	0	0	70	0.444
CuZrSiAs	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	21	0.816
FeAs	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	15	0.854
$GdFeO_3$	0	0	9	0	0	0	0	454	0	19	0	0	1	0	0	120	0.753
K_2NiF_4	0	0	0	0	0	0	0	3	81	2	0	0	0	0	0	56	0.570
$LaAlO_3$	0	0	2	0	0	0	0	33	1	92	0	0	0	0	0	27	0.594
$MgAl_2O_4$	0	0	0	0	0	0	0	0	0	0	315	0	0	0	0	69	0.820
$MgCu_2$	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	53	0.523
NaCl	0	0	0	0	0	0	0	1	0	0	1	0	140	1	0	81	0.625
$NaFeO_2$	0	0	0	0	0	0	0	0	0	0	0	0	0	105	0	34	0.755
TiNiSi	0	0	0	1	0	0	3	0	0	0	0	0	0	0	45	65	0.395
Other	0	5	29	5	37	2	18	59	21	16	31	14	21	12	8	20984	0.986
Total	105	100	173	167	93	95	109	562	103	134	103	72	162	118	54	21821	0.950
Precision	0.895	0.950	0.769	0.964	0.602	0.979	0.807	0.808	0.786	0.687	0.908	0.806	0.864	0.890	0.833	0.962	

Figure 8. Confusion matrix of algorithm with a cutoff size of 100. A perfect confusion matrix would have all nondiagonal sections zero. Precision and recall were rounded to three decimal places.

amount of times actual $CaTiO_3$-type structures was classified correctly. For example, out of 255 known $CaTiO_3$ entry prototypes, only 123 were correctly classified giving a recall of 123/255 or 48%.

To further understand the misclassification issues, we compared $CaTiO_3$, $LaAlO_3$, and $GdFeO_3$, which are shown in Figure 9. All lattices show remarkable structural similarities. While they are all variations of the standard cubic structure of perovskites, $CaTiO_3$ and $GdFeO_3$ are distorted orthorhombic structures, while $LaAlO_3$ is a trigonal structure. The essential structures are similar in terms of polyhedra, bond distances, and polyhedral connectivity but vary in terms of polyhedral tilting or rotation. $CaTiO_3$ and $GdFeO_3$ experience this tilting of the octahedra due to calcium and gadolinium being too small to form the cubic structure. $GdFeO_3$ also shows more distortion along the *c* axis for gadolinium then $CaTiO_3$ does with calcium. $LaAlO_3$ deviates from the ideal cubic structure by experiencing a rotation of the octahedra due to the length of the aluminum oxygen bond.[98]

Future work would be to take individual structures that are quite similar and group them by "family", which would increase the size of representative entries per family while reducing Other category percentage. This would help spread the data across multiple classes and create a more balanced training set. We believe this would help increase recall in the algorithm. Other possible future work would be to extend this approach to identify what synthetic routes would result in different structure types. Finding specific information on synthesis methods can be a difficult proposal due to the vast amount of research papers using different methods to achieve the same goal. However, Kim et al. recently published a machine learning paper to collect and organize this data as an interesting approach to overcoming this challenge.[99] However, their work covered only select oxide materials and required 640 000 papers to build a learning model. Some materials of interest may not have sufficiently established literature publications.

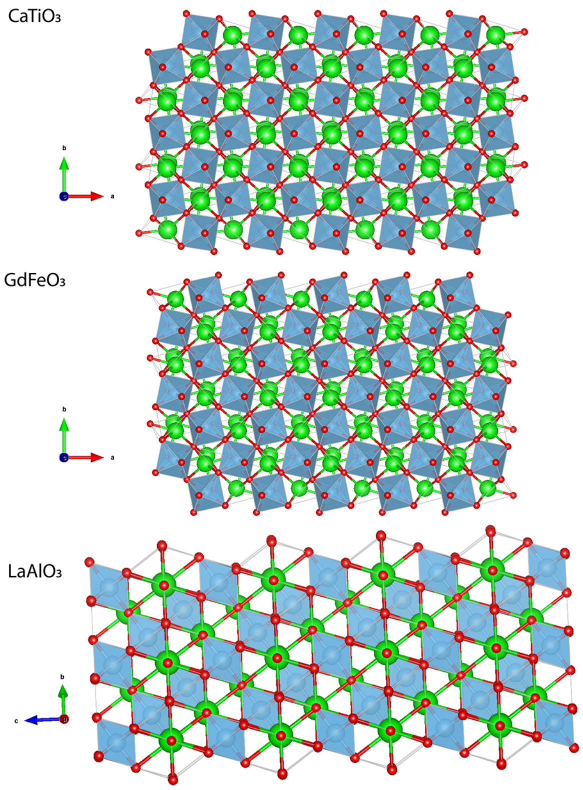


Figure 9. Comparison of misclassification of $GdFeO_3$ with $CaTiO_3$ and $LaAlO_3$. Calcium, gadolinium, and lanthanum are represented in green. Titanium, iron, and aluminum are represented as blue. Oxygen is represented as red.

4. CONCLUSION

The ability to predict crystal structures remains a challenging problem. The capability to engineer specific materials with certain properties requires the ability to predict crystal structures. Currently, characterization techniques such as diffraction or spectroscopy are the standard for assessing a compound's crystal structure, but these require a premade physical sample to measure. First principle calculations to predict crystal structure, on the other hand, could be used to screen materials prior to synthesis. These approaches have grown in recent years but are hindered by long computational times, limited scope, and cost. Machine learning offers a fundamentally new approach that can operate in concert with the experimental and first principle approaches mentioned earlier by rapidly offering probabilistic predictions of crystal structure rather than calculations.

Previous publications have introduced the possibility of machine learning-based crystal structure predictions but have been very limited in scope. For example, previous publications dealt only with a range from 3 to 208 specific crystal structures. These were limited to binary structures, ternary structures, or Heusler/inverse Heusler compounds.[73,74,77,85] Moreover, previous work built machine learning models which incorporated only training and validation sets limited from 55 to 1948 entries.[73,74,77,85] In contrast, consider that large inorganic material databases such as PCD or ICSD feature around a quarter of a million entries dispersed over more than 10 000 unique crystal structures at room temperature alone. Therefore, in this contribution, we built machine learning models that not only extend far beyond previous work but also begin to address what are the limitations and trade-offs in predicting any arbitrary crystal structure. To do so, we incorporated 24 215 of the 24 913 structure entries we obtained from the Pearson Crystal Database. The 24 215 entries were the result of simplifying feature development aspects of the machine learning process and included over 10 000 unique entry prototypes. With these models, we explored the implication of massively imbalanced entry prototype distributions and quantify the model performance associated with compromising model breadth for accuracy. The most notable trade-off is recall, which dropped quickly with a range from 73 ± 2 to 54 ± 2% for minimum-class sizes ranging from 150 to 10, respectively. These values drastically outperform simple metrics such as random guessing, which has a mean-class error ranged from 83.3 to 99.8% and fixing the prediction to the dominant class Other, which results in a mean class error from 86.7 to 99.9%. Reducing the scope of the model had little effect on average precision or accuracy, which was consistent across all the algorithms with a range of 86 ± 2 to 79 ± 2% and from 97 ± 2 to 85 ± 2%, respectively. Although the model struggled to exhaustively capture all members of a crystal structure, particularly with decreasing class size, the consistently high prediction accuracy is notable.

Successful performance in predicting crystal structure validates this machine learning approach for the exploration of chemical whitespace. We created a tool that rapidly and efficiently predicts one of the critical factors for physical phenomenon in a material. The output of our machine learning-based model is useful to influence or validate other crystal structure approaches. We see particular value when used synergistically with other machine learning algorithms based around physical property prediction.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemmater.7b05304.

Elemental properties used to generate features and importance rankings of features for each cutoff size (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sparks@eng.utah.edu.

ORCID

Taylor D. Sparks: 0000-0001-8020-7711

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the NSF CAREER Award DMR 1651668. Special thanks to Anton O. Oliynyk for his assistance in feature creation and counselling as well as Bryce Meredig and Christopher H. Borg from Citrine Informatics for helpful discussions and insight.

■ REFERENCES

(1) Gurevich, A. Challenges and opportunities for applications of unconventional superconductors. *Annu. Rev. Condens. Matter Phys.* **2014**, *5*, 35−56.

(2) Foltyn, S.; Civale, L.; MacManus-Driscoll, J.; Jia, Q.; Maiorov, B.; Wang, H.; Maley, M. Materials science challenges for high-temperature superconducting wire. *Nat. Mater.* **2007**, *6*, 631−642.

(3) Tarascon, J.-M.; Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **2001**, *414*, 359−367.

(4) Tarascon, J.-M. Key challenges in future Li-battery research. *Philos. Trans. R. Soc., A* **2010**, *368*, 3227−3241.

(5) Mori, D.; Hirose, K. Recent challenges of hydrogen storage technologies for fuel cell vehicles. *Int. J. Hydrogen Energy* **2009**, *34*, 4569−4574.

(6) Shao, Y.; Yin, G.; Wang, Z.; Gao, Y. Proton exchange membrane fuel cell from low temperature to high temperature: material challenges. *J. Power Sources* **2007**, *167*, 235−242.

(7) Kärkäs, M. D.; Åkermark, B. Water oxidation using earth-abundant transition metal catalysts: opportunities and challenges. *Dalton Trans.* **2016**, *45*, 14421−14461.

(8) Kaner, R. B.; Gilman, J. J.; Tolbert, S. H. Designing superhard materials. *Science* **2005**, *308*, 1268−1269.

(9) Zhao, Z.; Xu, B.; Tian, Y. Recent advances in superhard materials. *Annu. Rev. Mater. Res.* **2016**, *46*, 383−406.

(10) Babcock, E. WHAT IS VULCANIZATION? *Ind. Eng. Chem.* **1939**, *31*, 1196−1199.

(11) Garrett, A. B. *The flash of genius*; Van Nostrand: 1963, pp 14−15.

(12) Baekeland, L. The Bakelizer: National Museum of American History, Smithsonian Institution: a National Historic Chemical Landmark, November 9, 1993. *American Chemical Society* **1993**, 1−4.

(13) Ball, P. Chemistry: Perkin, the mauve maker. *Nature* **2006**, *440*, 429−429.

(14) Champagne, C. Serendipity, Super Glue and Surgery: Cyanoacrylates as Hemostatic Aids in the Vietnam War. In *The Proceedings of the 18th Annual History of Medicine Days Conference 2009*; Peterman, L., Sun, K., Stahnisch, F. W., Eds.; Cambridge Scholars Publishing: Alberta, Canada, 2009; pp 155−176.

(15) Munn, O. D.; Beach, A. E. The Inventor of Saccharine. In Munn & Co. *Sci. Am.* **1886**, *LV*, 36.

(16) National Academy of Engineering. *Grand Challenges for Engineering*; National Academies Press: Washington, D.C., 2008.

(17) Patel, P. Materials Genome Initiative and energy. *MRS Bull.* **2011**, *36*, 964−966.

(18) Seshadri, R.; Sparks, T. D. Perspective: Interactive material property databases through aggregation of literature data. *APL Mater.* **2016**, *4*, 053206.

(19) Sangid, M. D.; Matlik, J. F. ANALYSIS A better way to engineer aerospace components. *Astronaut. Aeronaut.* **2016**, *54*, 40−43.

(20) Fink, P. J.; Miller, J. L.; Konitzer, D. G. Rhenium reduction—alloy design using an economically strategic element. *JOM* **2010**, *62*, 55−57.

(21) Joost, W. J. Reducing vehicle weight and improving US energy efficiency using integrated computational materials engineering. *JOM* **2012**, *64*, 1032−1038.

(22) Rieger, T.; Gazdag, S.; Prahl, U.; Mokrov, O.; Rossiter, E.; Reisgen, U. Simulation of welding and distortion in ship building. *Adv. Eng. Mater.* **2010**, *12*, 153−157.

(23) Pauling, L. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **1929**, *51*, 1010−1026.

(24) Schön, J. C.; Jansen, M. First Step Towards Planning of Syntheses in Solid-State Chemistry: Determination of Promising Structure Candidates by Global Optimization. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1286−1304.

(25) Villars, P. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *J. Less-Common Met.* **1983**, *92*, 215−238.

(26) Villars, P. A three-dimensional structural stability diagram for 1011 binary AB2 intermetallic compounds: II. *J. Less-Common Met.* **1984**, *99*, 33−43.

(27) Villars, P. A semiempirical approach to the prediction of compound formation for 3486 binary alloy systems. *J. Less-Common Met.* **1985**, *109*, 93−115.

(28) Villars, P. A semiempirical approach to the prediction of compound formation for 96446 ternary alloy systems: II. *J. Less-Common Met.* **1986**, *119*, 175−188.

(29) Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937−946.

(30) Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **2015**, *87*, 897−923.

(31) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **2006**, *5*, 641−646.

(32) Jain, A.; Shin, Y.; Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **2016**, *1*, 15004.

(33) Ong, S. P.; Andreussi, O.; Wu, Y.; Marzari, N.; Ceder, G. Electrochemical windows of room-temperature ionic liquids from molecular dynamics and density functional theory calculations. *Chem. Mater.* **2011**, *23*, 2979−2986.

(34) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762−3767.

(35) Kang, K.; Meng, Y. S.; Bréger, J.; Grey, C. P.; Ceder, G. Electrodes with high power and high capacity for rechargeable lithium batteries. *Science* **2006**, *311*, 977−980.

(36) Anasori, B.; Xie, Y.; Beidaghi, M.; Lu, J.; Hosler, B. C.; Hultman, L.; Kent, P. R.; Gogotsi, Y.; Barsoum, M. W. Two-dimensional, ordered, double transition metals carbides (MXenes). *ACS Nano* **2015**, *9*, 9507−9516.

(37) Sharma, V.; Wang, C.; Lorenzini, R. G.; Ma, R.; Zhu, Q.; Sinkovits, D. W.; Pilania, G.; Oganov, A. R.; Kumar, S.; Sotzing, G. A. Rational design of all organic polymer dielectrics. *Nat. Commun.* **2014**, *5*, 4845.

(38) Yan, J.; Gorai, P.; Ortiz, B.; Miller, S.; Barnett, S. A.; Mason, T.; Stevanović, V.; Toberer, E. S. Material descriptors for predicting thermoelectric performance. *Energy Environ. Sci.* **2015**, *8*, 983−994.

(39) Zhu, H.; Hautier, G.; Aydemir, U.; Gibbs, Z. M.; Li, G.; Bajaj, S.; Pöhls, J.-H.; Broberg, D.; Chen, W.; Jain, A. Computational and experimental investigation of TmAgTe 2 and XYZ 2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *J. Mater. Chem. C* **2015**, *3*, 10554−10565.

(40) Madsen, G. K. Automated search for new thermoelectric materials: the case of LiZnSb. *J. Am. Chem. Soc.* **2006**, *128*, 12140−12146.

(41) Kolmogorov, A.; Shah, S.; Margine, E.; Bialon, A.; Hammerschmidt, T.; Drautz, R. New superconducting and semiconducting Fe-B compounds predicted with an ab initio evolutionary search. *Phys. Rev. Lett.* **2010**, *105*, 217003.

(42) Li, Y.; Hao, J.; Liu, H.; Li, Y.; Ma, Y. The metallization and superconductivity of dense hydrogen sulfide. *J. Chem. Phys.* **2014**, *140*, 174712.

(43) Yan, F.; Zhang, X.; Yonggang, G. Y.; Yu, L.; Nagaraja, A.; Mason, T. O.; Zunger, A. Design and discovery of a novel half-Heusler transparent hole conductor made of all-metallic heavy elements. *Nat. Commun.* **2015**, *6*, 1−8.

(44) Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **2006**, *5*, 909−913.

(45) Lejaeghere, K.; Bihlmayer, G.; Björkman, T.; Blaha, P.; Blügel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A. Reproducibility in density functional theory calculations of solids. *Science* **2016**, *351* (6280), aad3000.

(46) Pittaway, F.; Paz-Borbón, L. O.; Johnston, R. L.; Arslan, H.; Ferrando, R.; Mottet, C.; Barcaro, G.; Fortunelli, A. Theoretical studies of palladium− gold nanoclusters: Pd− Au clusters with up to 50 atoms. *J. Phys. Chem. C* **2009**, *113*, 9141−9152.

(47) Shao, G.-F.; Tu, N.-N.; Liu, T.-D.; Xu, L.-Y.; Wen, Y.-H. Structural studies of Au−Pd bimetallic nanoparticles by a genetic algorithm method. *Phys. E* **2015**, *70*, 11−20.

(48) Swamy, V.; Gale, J. D.; Dubrovinsky, L. Atomistic simulation of the crystal structures and bulk moduli of TiO2 polymorphs. *J. Phys. Chem. Solids* **2001**, *62*, 887−895.

(49) Harris, K. J.; Foster, J. M.; Tessaro, M. Z.; Jiang, M.; Yang, X.; Wu, Y.; Protas, B.; Goward, G. R. Structure Solution of Metal-Oxide Li Battery Cathodes from Simulated Annealing and Lithium NMR Spectroscopy. *Chem. Mater.* **2017**, *29* (13), 5550−5557.

(50) Naserifar, S.; Zybin, S.; Ye, C.-C.; Goddard, W. A., III Prediction of structures and properties of 2, 4, 6-triamino-1, 3, 5-triazine-1, 3, 5-trioxide (MTO) and 2, 4, 6-trinitro-1, 3, 5-triazine-1, 3, 5-trioxide (MTO3N) green energetic materials from DFT and ReaxFF molecular modeling. *J. Mater. Chem. A* **2016**, *4*, 1264−1276.

(51) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.

(52) Azencott, R. *Simulated Annealing: Parallelization Techniques*; Wiley-Interscience: 1992; Vol. *27*, pp 1−33.

(53) Ingber, L. Simulated annealing: Practice versus theory. *Mathematical and computer modelling* **1993**, *18*, 29−57.

(54) Bertsimas, D.; Tsitsiklis, J. Simulated annealing. *Statistical science* **1993**, *8*, 10−15.

(55) Kumar, M.; Husian, M.; Upreti, N.; Gupta, D. Genetic algorithm: Review and application. *International Journal of Information Technology and Knowledge Management* **2010**, *2*, 451−454.

(56) Lloyd, L. D.; Johnston, R. L.; Salhi, S. Strategies for increasing the efficiency of a genetic algorithm for the structural optimization of nanoalloy clusters. *J. Comput. Chem.* **2005**, *26*, 1069−1078.

(57) Singh, A. K.; Revard, B. C.; Ramanathan, R.; Ashton, M.; Tavazza, F.; Hennig, R. G. Genetic algorithm prediction of two-dimensional group-IV dioxides for dielectrics. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 155426.

(58) Trimarchi, G.; Freeman, A. J.; Zunger, A. Predicting stable stoichiometries of compounds via evolutionary global space-group optimization. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *80*, 092101.

(59) Wu, S.; Ji, M.; Wang, C.-Z.; Nguyen, M. C.; Zhao, X.; Umemoto, K.; Wentzcovitch, R.; Ho, K.-M. An adaptive genetic algorithm for crystal structure prediction. *J. Phys.: Condens. Matter* **2014**, *26*, 035402.

(60) Froltsov, V. A.; Reuter, K. Robustness of 'cut and splice'genetic algorithms in the structural optimization of atomic clusters. *Chem. Phys. Lett.* **2009**, *473*, 363−366.

(61) Pham, D.; Karaboga, D. *Intelligent Optimization Techniques: Genetic Algorithms, Tabu Search, Simulated Annealing and Neural Networks*; Springer Science & Business Media: 2012, pp 1−50.

(62) Brgoch, J.; Hermus, M. Pressure-Stabilized Ir3−in a Superconducting Potassium Iridide. *J. Phys. Chem. C* **2016**, *120*, 20033−20039.

DOI: 10.1021/acs.chemmater.7b05304

(63) Miao, M.-s.; Wang, X.-l.; Brgoch, J.; Spera, F.; Jackson, M. G.; Kresse, G.; Lin, H.-q. Anionic chemistry of noble gases: formation of Mg−NG (NG= Xe, Kr, Ar) compounds under pressure. *J. Am. Chem. Soc.* **2015**, *137*, 14122−14128.

(64) Oganov, A. R.; Glass, C. W.; Ono, S. High-pressure phases of CaCO3: crystal structure prediction and experiment. *Earth Planet. Sci. Lett.* **2006**, *241*, 95−103.

(65) Morris, A. J.; Grey, C.; Pickard, C. J. Thermodynamically stable lithium silicides and germanides from density functional theory calculations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 054111.

(66) See, K. A.; Leskes, M.; Griffin, J. M.; Britto, S.; Matthews, P. D.; Emly, A.; Van der Ven, A.; Wright, D. S.; Morris, A. J.; Grey, C. P. Ab Initio Structure Search and in Situ 7Li NMR Studies of Discharge Products in the Li−S Battery System. *J. Am. Chem. Soc.* **2014**, *136*, 16368−16377.

(67) Curtarolo, S.; Morgan, D.; Persson, K.; Rodgers, J.; Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **2003**, *91*, 135503.

(68) Batista, G. E.; Monard, M. C. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* **2003**, *17*, 519−533.

(69) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, And Prediction*, 2nd ed.; Springer: New York, NY, 2009, pp 587−604.

(70) Roberts, N. A.; Walker, D. G. Phonon Transport in Asymmetric Sawtooth Nanowires. In ASME-JSME Thermal Engineering Joint Conference; Honolulu, HI, March 13−17, 2010.

(71) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 094104.

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

(73) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324−7331.

(74) Balachandran, P. V.; Theiler, J.; Rondinelli, J. M.; Lookman, T. Materials prediction via classification learning. *Sci. Rep.* **2015**, *5*, 1−16.

(75) Clarke, B.; Fokoue, E.; Zhang, H. H. *Principles and Theory for Data Mining and Machine Learning*; Springer Science & Business Media: 2009, pp 73−113.

(76) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; ACM: 1992; pp 144−152.

(77) Oliynyk, A. O.; Adutwum, L. A.; Harynuk, J. J.; Mar, A. Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis. *Chem. Mater.* **2016**, *28*, 6672−6681.

(78) Chen, N. *Support Vector Machine in Chemistry*; World Scientific: 2004; pp 53−59.

(79) Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Academic Press: London, 1999, pp 701−763.

(80) Wang, L.; Brown, S. J. Prediction of RNA-Binding Residues in Protein Sequences Using Support Vector Machines. *Engineering in Medicine and Biology Society, EMBS'06, 28th Annual International Conference of the IEEE*; IEEE: 2006; pp 5830−5833.

(81) Janda, J.-O.; Busch, M.; Kück, F.; Porfenenko, M.; Merkl, R. CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinf.* **2012**, *13* (55), 1−11.

(82) Redfern, O. C.; Harrison, A.; Dallman, T.; Pearl, F. M.; Orengo, C. A. CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.* **2007**, *3* (11), e232.

(83) Sundararaghavan, V.; Zabaras, N. Classification and reconstruction of three-dimensional microstructures using support vector machines. *Comput. Mater. Sci.* **2005**, *32*, 223−239.

(84) Fang, S.; Wang, M.; Qi, W.; Zheng, F. Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials. *Comput. Mater. Sci.* **2008**, *44*, 647−655.

(85) Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Lotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.; Brgoch, J. Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC. *J. Am. Chem. Soc.* **2017**, *139* (49), 17870−17881.

(86) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative structure−property relationship modeling of diverse materials properties. *Chem. Rev.* **2012**, *112*, 2889−2919.

(87) Basheer, I.; Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *J. Microbiol. Methods* **2000**, *43*, 3−31.

(88) Patel, M. S.; Mazumdar, H. S. Knowledge base and neural network approach for protein secondary structure prediction. *J. Theor. Biol.* **2014**, *361*, 182−189.

(89) Holley, L. H.; Karplus, M. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 152−156.

(90) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised Machine Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8* (20), 5091−5098.

(91) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.

(92) Oliynyk, A. O.; Sparks, T. D.; Gaultois, M. W.; Ghadbeigi, L.; Mar, A. Gd12Co5. 3 Bi and Gd12Co5 Bi, Crystalline Doppelgänger with Low Thermal Conductivities. *Inorg. Chem.* **2016**, *55*, 6625−6633.

(93) Seko, A.; Maekawa, T.; Tsuda, K.; Tanaka, I. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 054303.

(94) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **2017**, *29* (12), 5090−5103.

(95) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *93*, 115104.

(96) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **2016**, *6*, 20952.

(97) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 1−6.

(98) Tilley, R. J. *Perovskites: Structure−Property Relationships*; John Wiley & Sons: 2016, pp 1−40.

(99) Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **2017**, *4*, 170127.