NORTHWESTERN UNIVERSITY

Machine Learning for Materials Discovery and Design

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

For the degree

DOCTOR OF PHILOSOPHY

Field of Materials Science and Engineering

By

LOGAN TIMOTHY WARD

EVANSTON, ILLINOIS

March 2017

Abstract

The impacts of many important technologies are limited by the availability of better-performing materials. One factor limiting the ability of engineers to develop better materials is the speed at which they can search through possible formulations and processing schemes. Recently, machine learning algorithms have emerged as a possible route to reusing existing materials data to guide the design process. In this thesis, we discuss work towards addressing three major challenges in the use of machine learning in materials engineering. First, we implemented an automated toolkit for solving crystal structures and use that to improve the quality of an existing materials database. Second, we developed general-purpose methods for creating machine learning models from materials data, which will simplify and accelerate the development of new models. Third, we created open-source software for making these machine learning techniques more readily-accessible to the materials community. Along with addressing these challenges, we also demonstrate how machine learning can be applied to optimize existing and discover new Bulk Metallic Glass alloys. It is our vision that the methods developed in this work will help enable the application of machine learning to a wider variety of problems and, potentially, be used to improve materials employed in many different technologies.

Acknowledgements

I have many people to thank for their role in making this PhD possible. Here is my best attempt at succinctly thanking (in loosely chronological order)…

my family, Amanda, Tim, and Jennifer, for always encouraging my curiosity.

my wife, Meagan, for being an excellent sounding board and partner in adventure.

many great teachers and professors throughout my education, especially Bill Jervey for excelling at teaching math and also encouraging me to learn advanced math on my own.

several excellent mentors (Jim Mueller, Rajiv Berry, David Mollenhauer, Dan Miracle) and colleagues at AFRL for introducing me to computational materials science.

my Master's advisers, Wolfgang Windl and Kathy Flores, for teaching me to be a scientist.

the Wolverton research group and other graduate students at Northwestern for being such great friends and colleagues, and for being willing to share their varied expertise.

my supervisors and colleagues at Los Alamos National Laboratory, especially Bob Hackenberg, for sharing their insight and helpful career advice.

my PhD adviser, Chris Wolverton, for always being able to ask great questions and knowing how to distill and communicate a great story.

Additionally, I would like to acknowledge support from the Center for Hierarchical Materials Design (CHiMaD), as well as the National Defense Science and Engineering Graduate (NDSEG) Fellowship, Weertman Fellowship, and Ryan Fellowship.

Table of Contents

List of Figures

List of Tables

1 Introduction

The development of entirely new classes of materials can have revolutionary effects on society, but even the slight improvement of existing materials can have profound impacts. Considering only energy applications, there are plenty of examples of technologies limited by available materials. New cathode materials for batteries could make electric cars with comparable ranges and costs to gasoline-powered vehicles. Additions to the composition or changes in the processing of nickel alloys could make jet engines and gas turbine power plants dramatically more efficient. The identification of oxides made of inexpensive constituents and the right tolerance of oxygen deficiency could make it feasible to create fuel from only water and sunlight. What limits the ability of engineers to find materials that suit these applications is their ability to quickly search through the countless design space of new materials.

The many possible and many interrelated effects leading to the observed properties of materials makes them complicated to engineer. Many properties (e.g., fracture toughness) can be driven by the effects of structures from the atomic scale up to centimeters, and be controlled by processing conditions from the initial selection of constituents all the way through how it is joined to other components in a device. The practical consequences of this complexity are that understanding how a material behaves or optimizing its properties can require a large number of tests and design iterations. Consequently, the time for a new material to go from initial discovery in the laboratory to actual use is often on the scale of decades.[1,2] One route for accelerating the materials design process and, thereby, the development of new technologies, is to supplement or replace many of the required tests with computational tools.

Owing to the wide variety of processes that affect materials behavior, a wide variety of computational methods are available for predicting material properties. The tools range from atomic-scale modelling techniques, such as Density Functional Theory (DFT), to finite-element models for predicting the mechanical behavior of entire components. Combined together, these tools make it possible to optimize many aspects of a material concurrently and with minimal amount of input from experiments. While these conventional tools are powerful and have been employed to design many new materials, they have limitations. Some methods are very computationally expensive (e.g., DFT), others require extensive experimental measurements of a certain system before being useful (e.g., computational thermodynamics), and there are some properties that currently lack computational tools that can predict them. Recently, machine-learning-based computational tools have emerged as a route to addressing some of these limitations.

While the use of machine learning is becoming pervasive in many other aspects of society, their application to materials – often called “Materials Informatics” or “Materials Data Analytics” – is still in its initial stages.[3,4] Machine learning algorithms, which automatically learn predictive rules from data, offer the ability to create materials models that can be much faster than conventional, physics-based computational tools and can create models for properties where physical models do not yet exist. However, their widespread use in materials engineering is limited by several factors. For one, there is a lack of the basic ingredient of machine learning models for many materials applications: high-quality, digital data. Furthermore, the process for creating these models is complicated by the lack of general-

purpose techniques for turning that raw data into a useful model. Finally, there is no software that will simplify integrating these models into materials design tools. If solved, these problems could enable machine learning as another commonly-used tool in the computational design of new materials.

In this thesis, I will describe progress towards addressing the availability of high-quality materials data and the development of general-purpose materials data analytics methods. The next chapter will describe the theory behind the methods employed in this work, and cover what progress has already been made in these fields. Chapters 3 and 4 will describe how we used automated crystal-structure solution tools to fill in missing data in materials databases. Chapters 5 and 6 will introduce methods we created for constructing machine learning models based on the composition and crystal structure of a material, respectively, that can be applied to model a wide variety of material properties. Chapter 7 is a case study describing how we applied these machine learning methods to develop new Bulk Metallic Glass alloys. Finally, Chapter 8 will discuss an open-source tool developed to make materials informatics methods more widely available to the materials engineering community at large. Together, we envision that these advancements will help make machine learning tools more commonly used in the engineering of new materials.

2 Background

The following sections cover the theory and historical developments in the areas of science associated with this dissertation. The first section includes a description of atomic scale computational modeling, which is used in most of the subsequent chapters. The second section is a description of how crystal structures are determined from experimental data. Finally, this background section is concluded with a discussion of machine learning methods and their application to materials engineering, with a particular focus on the intersection between machine learning and atomic-scale modeling.

2.1 Atomic-Scale Modeling

At their core, atomic-scale modeling methods are based solely on the calculation of the energy of a collection of atoms or electrons. As atomic-scale processes are the most fundamental root of nearly all materials phenomena, these calculation techniques enable the prediction of the behavior of a material with few assumptions about what effects actually lead to the observed properties. Owing to the versatility inherently connected to their predictive ability, atomic-scale simulation tools are widely used in materials engineering. For example, these tools are commonly employed to predict the rate of diffusion of elements within a material,[5,6] to study the behavior of materials in conditions where experiments are difficult (e.g., under shock loading[7]), and predicting the atomic scale structure of materials.[8]

The core requirement of atomic scale calculations is a method to evaluate the energy or electronic structure of the material as a function of atomic positions. Once one can express the energy of a system as function of atomic positions, it becomes possible to also express many

macro-scale, observable properties as a function of atomic positions. For example, one can determine elastic constants by computing the curvature of energy as a function of strain applied to the system. Elastic constants are a simple property to compute compared to what is common practice today. One can use Newton's Laws of Motion and the forces acting on each atom (accessible by derivatives of energy with respect to atomic displacement) to model the trajectories of atoms over time and determine properties such as the melting temperature.[9] Additionally, one can use the electronic density of states and vibrational spectra of materials to assess the electrical conductivity and stability as a function of temperature.[10] While the ways to measure macro-scale properties can be quite complex, the majority can be condensed to expressions of the energy of the system as a function of changes in atomic or electronic structure.

The methods to create these energy functions are quite varied, but can be generally grouped into classical and quantum-mechanical. The "classical" energy functions, as referred to as empirical or interatomic potentials, the energy is expressed directly as a function of the atomic positions. These interatomic potentials could be a simple as a function where energy is determined by the distances between all pairs of atoms in a structure. For example, the Lennard-Jones potential can be written as

$$E = \sum_{i \neq j} 4\epsilon \left(\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right) \quad (1)$$

where the sum is over all pairs of atoms in the calculation and r_{ij} is the distance between each pair of atoms, i and j. The coefficients in these interatomic potentials (ex: σ and ϵ in

Equation 1), are determined by adjusting them so that the properties computed using the potential match either experimental data or quantum-mechanical calculations. The classical functions have the advantage of being fast to compute (compared to quantum-mechanical approaches), which allows them to model processes involving trillions of atom or time scales up to milliseconds.[11,12] However, they require being fit to some other data before being useful and do not allow one to determine the electronic structure of the material.

2.1.1 Density Functional Theory

In contrast, quantum-mechanics-based calculations compute the interactions between atoms by solving the electronic structure of the material – a process which requires zero prior knowledge about the material. While it is possible to analytically solve or numerically approximately solutions to the Schrödinger Equation for systems with small numbers of electrons, it becomes computationally intractable for systems with even a few dozen electrons (i.e., most technologically-relevant materials). A variety of approaches exist to overcome this computational intractability, including Quantum Monte Carlo and Hartree-Fock methods.[13,14] The method that is most widely used for inorganic materials is Density Functional Theory (DFT).

DFT is based on the theory that the many-body interactions between electrons can be expressed as functionals of the density of all electrons in the material.[15] In doing so, one can solve the wavefunctions of each electron independently, which is actually computationally tractable.[16] In order to make this approximation, one must be able to express two distinct effects as functionals of electron density: *exchange,* an effect of the Pauli exclusion principle;

and *correlation*, the repulsion between individual electrons. As the exact form of these functionals is not known,[17] one must rely on approximations that are often based on the magnitudes and spatial derivatives of the electron density. While there will likely always be room to improve these approximations (often at greater computational cost), advances in DFT methods and computing have made it possible to accurately compute many properties with DFT. For example, the accuracy formation enthalpy of crystalline compounds computed from DFT are comparable to the levels of errors between different experimental methods.[18] Consequently, DFT has found broad usage across many domains of materials engineering.[19]

2.1.2 High-Throughput Density Functional Theory

Starting around the early 2000s, computing power and the quality of DFT codes advanced far enough that it became possible to automatically perform 1000s of DFT calculations. This automated approach to DFT – often called High-Throughput Density Functional Theory – can be used to evaluate the properties of many materials within a single study.[20] One of the earliest examples of this work assessed the formation enthalpy of tens of thousands of ordered face-centered-cubic compounds, and found several that might prove interesting as strengthening precipitates in superalloys.[21,22] As computer power has advanced even further, it is now possible to evaluate hundreds of thousands of materials in a single study and predict even more complex properties of materials in an automated manner.[23–25] An example of such an approach that involves assessing the energy density and ease of Li-ion extraction of battery materials is shown in Figure 2.1. Such studies have enabled the discovery of new materials for

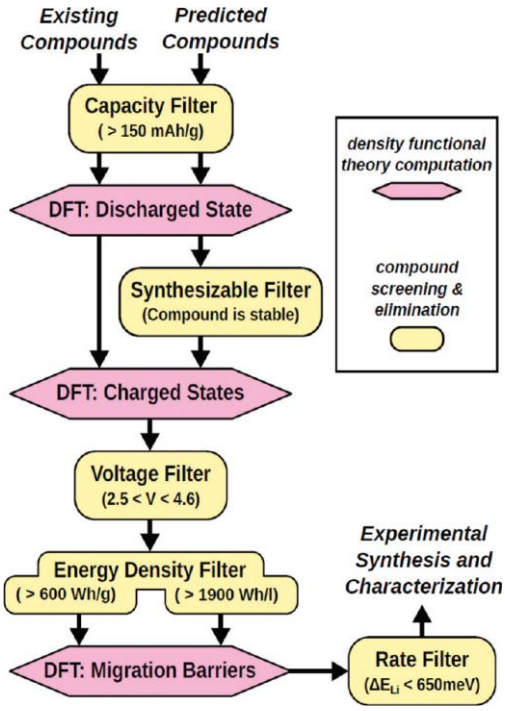


Figure 2.1. [From Ref. [27]] Schematic of an automated scheme used to assess the suitability of various materials for Li-ion battery cathodes. The voltage, storage capacity, and Li-migration energy of known and hypothetical materials are computed with DFT, and then used to identify interesting candidates for use as Li-ion battery cathodes.

applications such as thermochemical water splitting,[26] Li-ion batteries,[27–31] and many other applications.[20,29,32–35]

The results of high-throughput DFT studies are often stored in a database, which makes it easier to reuse the results of the calculations for different problems. Many research groups have gone so far as to make the data openly available on the internet.[18,36–41] Consequently, these databases are actually used by groups unassociated with their creators. For example, data form the OQMD – a database created by the Wolverton group at Northwestern – has been used in the study of the effects of Li on the elastic properties of Mg alloys,[42] and data from the Materials Project was used in the design of Fe-Ag-based amorphous alloys.[43] These examples demonstrate the utility of sharing data from high-throughput DFT calculations and, additionally, are also a great success story of open data in materials engineering. As these databases grow in size and start to include even more material properties (e.g., elastic constants [25]), the use of high-throughput DFT data may become pervasive in materials engineering.

2.1.3 Limitations of Atomic-Scale Modeling

The main limitation of atomic scale modeling is the high computational cost. Computing the energy of a material with DFT requires on the order of CPU minutes to hours on a modern computer, depending on the number of electrons in the structure and choice of exchange-correlation functional. Considering that modern supercomputers have hundreds of thousands of processors, a few CPU hours is trivial. Where atomic-scale modelling becomes expensive is when assessing properties that require large numbers of energy evaluations. For example, computing the melting temperature of a material can require tens of thousands of CPU hours for a single material because it involves many different calculations that describe the trajectory of atoms over time.[9] While there are certainly plenty of applications within the reach of current computing power for both classical and quantum-mechanical computational tools, expanding the reach of atomic scale calculations by developing new and more efficient methods is still a very active area of research.[44–48]

2.2 Crystal Structure Solution

The only input needed from experiment to assess the properties of a crystalline material with DFT is its atomic-scale structure, which can be difficult to determine. One of the most common means for assessing the crystal structure of a material is X-ray diffraction. X-ray diffraction techniques are based on the fact that x-rays diffract off the periodic lattice of a crystal most strongly at specific angles, which are defined by shape of the unit cell and the symmetry of the crystal. For some cases (such as NaCl and Cu), only one possible arrangement of atoms within the unit cell is possible given the unit cell shape, symmetry, and the measured

density of the material. In such cases, the determination of the structure is simple. However, such simplicity is not always the case. For many diffraction patterns, many possible arrangement of atoms are possible, and determining which is the actual structure requires finding which structure matches the intensities of the diffraction peaks. As these intensities also depend on many other factors, such as differences between the thermal vibrations of each atom, solving the structure of the material can be quite difficult. However, over century since the discovery of x-ray diffraction, crystallographers have developed many sophisticated methods.

The process of solving the structure of a material from diffraction data, as shown in Figure 2.2, can be broken down into at least two major steps: peak indexing and structure solution.[49,50] During the peak indexing step, one first determines the diffraction angles of each diffraction peak and then uses the relative angles of the peaks and Bragg's law to associate each reflection with a crystallographic plane. Through knowledge about how the symmetry of the crystal leads to some reflections overlapping or being invisible, one can then

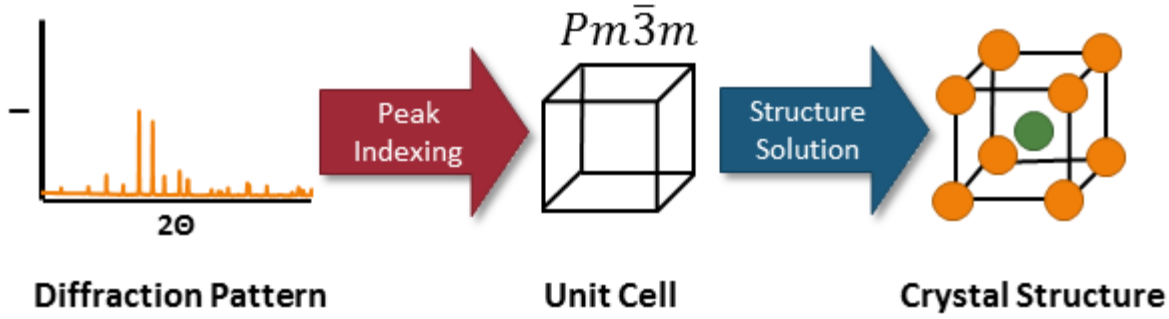


Figure 2.2. Schematic of the crystal structure solution process. After a diffraction pattern is measured, the locations of the diffraction peaks are used to determine the shape and symmetry of the unit cell. Next, structure solution methods are used to determine the arrangements of atoms within the unit cell that best explain the observed diffraction pattern.

determine both the shape and symmetry of the unit cell. The indexing problem has been studied extensively, and there are several commonly-used computer programs that perform this task automatically.[50–52] In contrast, crystal structure solution is often more difficult.

2.2.1 Structure Solution Methods

Once the unit cell and symmetry of a crystal are determined, the next step is to determine the positions atoms within the unit cell – a process known as "structure solution." Many different approaches to this problem exist, and they are often broken into at least two distinct classes: "direct" and "direct-space." The goal of direct methods is to reconstruct the electron charge density that leads to the observed diffraction pattern, which can be related to the observed diffraction peaks with the relationship

$$\rho(\vec{r}) = \sum_{\vec{h}} \left|F_{\vec{h}}\right|^2 \cos[2\pi(\vec{h}\cdot\vec{r}) + \alpha_{\vec{h}}] \quad (2)$$

where $F_{\vec{h}}$ is the diffracted intensity for a reflection associated with a plane in the crystal, $\vec{h}$, and $\alpha_{\vec{h}}$ is the phase angle associated with the diffracted wave.[51] Direct methods seek to retrieve these phase angles, which cannot be measured. Once the charge density of the system is determined, the local peaks in density are attributed to the location of atoms. A variety of techniques are available for the direct solution of these phase angles, including a method by Hauptman and Karle in 1954 and the charge-flipping method from the 2000s.[53,54] Direct methods work best when one can create a large single crystal sample of a material. Using direct methods with data collected from powder diffraction data, in comparison, is complicated by many different diffraction peaks overlapping, but techniques for estimating the contributions

from each independent reflection (i.e., to access the values of $F_{\vec{h}}$) do exist.[55–57] As the result of many advancements in the theory and practice of these methods, solutions of structures containing hundreds of atoms with data from single crystal experiments is routine and can also be automated.[58–61]

The second class of structure solution methods, known as “direct space” methods, function by finding an optimal arrangement of atoms whose calculated diffraction pattern matches the observed pattern. In contrast to direct methods, no attempt is made to retrieve phase angles or reconstruct the electron density. When the shape of the unit cell and its contents (i.e., type and number of atoms) are known, the determination of an optimal arrangement of atoms becomes a global optimization problem where the objective is to minimize the difference between the calculated and measured X-ray diffraction pattern. A wide collection of methods is available to perform this task, which mostly vary in the choice of global optimization algorithm and cost function.[55] For example, some methods use genetic algorithms to search through possible structures and others use simulated annealing.[49,62] Furthermore, some methods only use the agreement the diffraction pattern and others include additional information about the structure (e.g., potential energy [49,63–65]) to guide the optimization. As with the peak indexing and direct methods, many of these direct-space methods are available in commercial or open-source software.[55]

2.2.2 First-Principles-Assisted Structure Solution (FPASS) Method

One direct-space crystal structure solution method that is especially important to the work described in this thesis is the First-Principles-Assisted Structure Solution (FPASS) method.[63]

The FPASS method has two major features that make it particularly useful in the solution of crystal structures from powder diffraction data: a symmetry-restricted optimization algorithm and the use of *ab initio* DFT calculations to compute the energy of candidate structures. First of all, the ability to include all of the information from peak indexing (i.e., both unit cell shape and symmetry) into defining constraints for the search space drastically simplifies the search for the optimal structure. FPASS integrates these constraints with a specially-designed genetic algorithm (GA) that ensures all the candidates tested by the algorithm have the observed symmetry. Additionally, the use of DFT makes it possible to assume that the structure that best matches the diffraction pattern is also the lowest in energy within the unit cell and symmetry constraints (given the reliability of DFT in identifying the ground state structures of materials [18]). The specially-designed GA and integration of DFT makes it possible to solve structures that were difficult to determine using conventional structure solution techniques.[63]

2.2.3 Crystallographic Databases

Other important developments in crystallography include the creation of standards for sharing crystallographic data and the aggregation of solved crystal structures into databases. The most widely-accepted format for crystallographic data is the Crystallographic Information File (cif), which can be used to store the solved crystal structure and over 500 other pieces of metadata to describe other aspects of the structure and how they were gathered.[66] Having access to an accepted standard for sharing data helped to enable the rise of common repositories of crystallographic data, such as the Crystallography Open Database (COD) and

Inorganic Crystal Structure Database (ICSD).[67,68] These databases have made it possible to easily integrate information about known crystal structures into techniques including the high-throughput determination of phase diagrams from X-ray data and the assessment of the properties of thousands of known materials with DFT.[18,69] Additionally, these databases highlight the significant advancements and room for future progress in crystal structure solution. These databases contain hundreds of thousands of crystal structures for a wide variety of materials, but also many incompletely-determined structures that have yet to be solved.

2.3 Machine Learning and Materials Engineering

Machine learning algorithms automatically generate computer programs from training data, and can improve them without being explicitly programmed by humans. The ability to automatically generate the rules necessary to create complex software makes it possible to construct programs that would be impossible to create manually. For example, machine learning algorithms are used by Netflix to recommend videos to their customers because "machines have a much better ability to learn from vastly bigger data pools than expert humans."[70] For reasons similar to those cited by engineers at Netflix, machine learning has become an established part of many fields of science, such as chemistry and biology.[71,72] However, machine learning has yet to become widely adopted in materials engineering.

2.3.1 Machine Learning

Machine learning algorithms encompass a wide variety of tools, which are often characterized by the type of problem they are created to solve or how the models are created. For example, "unsupervised" learning algorithms are designed to recognize whether a dataset

is composed of distinct clusters of data with similar characteristics. There are also "reinforcement learning" algorithms that are designed to create programs by iteratively producing programs and then adjusting to optimize performance (via a "reward" system), which have recently been used to create software for playing the board game Go.[73] Of particular relevance to the work described in this dissertation are "supervised" learning algorithms.

Supervised machine learning algorithms are designed to create models that map the inputs of some unknown process to its outputs. As a simple example, linear regression algorithms map the inputs of a process to its output by fitting slopes related to each input variable. Of course, significantly more advanced methods such as neural networks, decision trees, and ensemble approaches exist and are often better choices than linear regression for many problems.[74–78] However, in all cases, the model is created by providing the algorithm a set of training examples consisting of the inputs and outputs to the process. However similar, each particular algorithm method does offer unique advantages and no algorithm works best for every problem. For example, nearest neighbor algorithms train quickly and Gaussian process regression models offer uncertainties for each prediction at little additional computational cost. It is the task of the user to determine which algorithm works best for a particular problem.

Key concepts needed to understand how to select the optimal supervised learning algorithm are “generalization error” and overfitting. It is trivial to create a machine learning model that perfectly describes the input data, but this model may be limited in its ability to make accurate predictions – that is to be able to generalize to unseen data. To illustrate this concept, Figure 2.3 shows the effect of increasing the degree of a polynomial fit to some data generated based on a cubic polynomial. As the number of terms is increased past 3, the model starts to perform worse on other data generated with the same function even though it fits data perfectly. These high-order polynomial models do not predict data from outside the training examples well, and are referred to as being “overfit” to the training data. Often, testing whether models are overfit is accomplished via cross-validation, where a section of available data is withheld from training a model and used to evaluate the performance of the model. By adjusting the parameters of the learning algorithm (e.g., the number of terms in a polynomial),

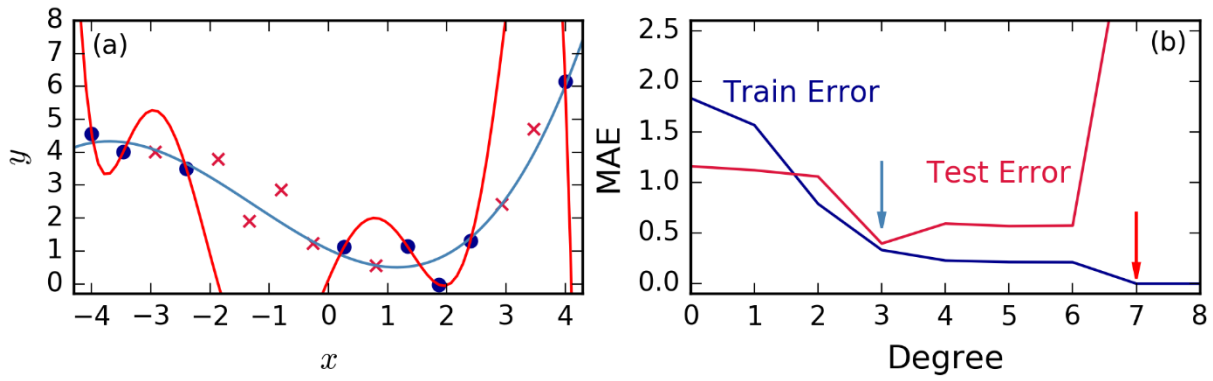


Figure 2.3. Illustration of overfitting a polynominal model to data generated from a cubic polynomial with Gaussian noise. (a) A plot of 3rd (light blue) and 7th (red) degree polynomials fit to a series of training points (blue circles). A series of points generated using same cubic function but not used to fit the model are shown by red crosses, which were used to test the performance of the model. (b) The mean absolute error (MAE) between the value predicted by a polynomial and the training points (blue) and withheld points (red). The 7th degree polynomial is overfit because error the training points lower than the 3rd degree model but much higher for those in the test set.

one can maximize the performance of the model on this validation set and ensure the algorithm does not overfit to the training data.

2.3.2 Representations and Machine Learning

Another key concept in understanding the application of machine learning to materials is the idea of representations. While many machine learning problems are based on non-numerical data like text or the compositions of semiconductors, machine learning algorithms generally expect data to be expressed as a list of numerical values with a fixed length. Converting non-numerical data into a numerical form is known as creating a representation. For example, text can be converted into a string of numbers by counting how many times certain words appear – the "bag of words" approach.[76] To give a practical example, a machine learning algorithm could develop a model that employs the number of times "horrible" appears in a movie review to determine whether the review was positive or not. Devising these kind of informative variables for materials is still an outstanding problem.

There has at least been general agreement on what characteristics these representations of materials should have, which are often broken into 4 distinct traits:[45,79–81]

1. **Complete:** Representations should distinguish different materials from each other
2. **Descriptive:** Representations should contain attributes that relate to the physical factors leading to the observed properties
3. **Simple**: Representations should be faster to compute that the method used to generate the data
4. **Unique:** Each material should have exactly one representation

Depending on the author, there is some variation how these rules are described. For example, Jain *et al.* assert that representations must not just be unique but also be reversible, so that it is possible to easily identify a material that has certain combination of attributes.[81] Additionally, these rules should not be viewed as mandatory. Montavon *et al*. have demonstrated how it is possible to create models with a non-unique representation.[82] Regardless of the particularities of each view of these rules, these desirable characteristics do give at least broad outlines of how to approach designing representations and understanding of how to best select a suitable representation for a certain problem.

Of these requirements, one rule that is particularly complicated to satisfy for materials is "completeness." Owing to the large differences in scales of structure and variety of processing conditions that affect materials properties, the minimum amount of representation to differentiate materials can be drastically different. For example, if one is modeling the solution energies of different elements in Zirconia, then it is sufficient to differentiate each training entry based on properties of each element.[83] However, when trying to predict the fatigue strength of steel it is necessary to include the fraction of each element in the material and the processing history into the representation.[84] However, it is not necessary to include *all* available information about a material when building a machine learning model. For instance, Meredig *et al*. only used information based on the composition of a material even though the crystal structure was available because their objective was to identify compositions where it was likely to be able to form a new compound.[85] Consequently, when selecting a

representation for materials, it is important to first consider both what information is necessary to build a model but also how the model will be used.

2.3.3 Applications in Materials Engineering

Machine learning algorithms have been used across many different domains in materials science, and for several different purposes. In general, the applications of machine learning are reflective of its advantages compared to conventional computational tools: (1) creating fast enough models to search over large numbers of possible materials, and (2) identifying the important factors leading to observed material properties from many possible options. This background section is only meant to cover the major advancements in this area to give context to the developments described in this work. Consequently, I would suggest for the reader to consult several recent reviews on this subject for a more comprehensive picture.[4,45,81,86-90]

A demonstrative example of the use of machine learning to be able to evaluate an exceptionally large search space is work by Faber *et al* in 2016 to find new crystalline materials based on the Elpasolite crystal structure.[91] Considering only main group elements, there are approximately 2 million possible materials with the Elpasolite structure – far too many to evaluate the stability of all using DFT. So, in order to make it possible to search the entire space, the authors computed the formation enthalpies of 10^4 of the possibilities using DFT, trained a machine learning model on that set, and then employed that model to evaluate the remaining $\sim 2 \times 10^6$ entries. In doing so, the authors identified 90 new, stable Elpasolite compounds, including one ($NaFAl_2Ca_6$) that features Al in a surprising oxidation state. This study

demonstrates the advantages of machine learning – the ability to perform searches too large to approach with conventional tools that may yield unanticipated solutions. This sort of surrogate model optimization is not an uncommon method in materials informatics and has been used to discover a number of other crystalline materials, among other applications.[85,92–95]

Another common application for machine learning in materials is to identify which factors out of a large number of possibilities are most predictive of materials behavior. A good example of this sort of approach is described in a paper by Isayev *et al*., where the authors attempted to find structural traits of high-temperature superconductors.[96] As a representation, the authors collected many different structural fragments of known superconductors using the Simplex Representation of Molecular Structure (SiRMS) and identified that the factor that is the strongest predictor of a material being a high temperature superconductor is the presence of copper atoms near two oxygen atoms. While this result is not particularly surprising in light of the prevalence of high-temperature cupric oxide superconductors, it is promising that the machine learning algorithm detected this automatically. Machine learning has also been used to understand behavior including the friction constants in materials and factors influencing solubility of different elements in zirconia.[41,79,83,86,97–101]

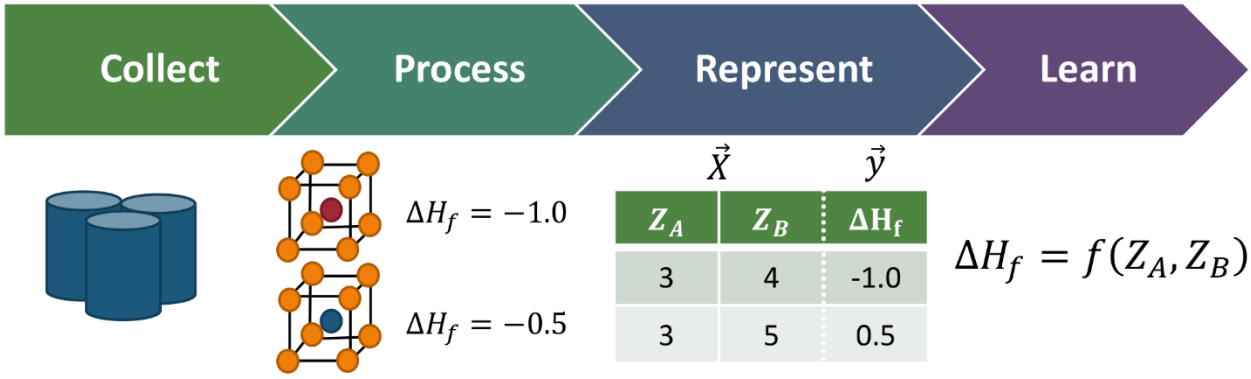


Figure 2.4. Schematic of the workflow for materials informatics, which outlines the process from data collection through selecting a training set and representation to finally training the machine learning algorithm.

2.3.4 Workflow for Materials Informatics

The materials informatics methods described in the previous section can be generalized into a four step process, as shown in Figure 2.4. The first step in any method is the collection of data, which often includes the extraction of information from papers or handbooks. Once sufficient information about a particular material is collected, the next step is to organize that data into a training dataset for the model. At this point, the main goal is to define the desired inputs and outputs to the model, and to ensure the dataset does not contain any duplicates or erroneous data. Once the inputs are defined, one must decide which attributes of the material should be used as inputs to the machine learning model (i.e., the representation), which must meet the requirements specified in Section 2.3.2 (e.g., adequately differentiate materials in the training set). Finally, one decides on the appropriate machine learning model and trains it. Depending on the target application, it may be necessary to select algorithms for reasons besides predictive accuracy, such as training speed, differentiability, or human-interpretability. While the specific strategies employed in each step and the eventual use of the model vary

between each study, this general process was used to create all of the models described in this work and, I would assert, all other materials informatics models in the literature.

2.3.5 Challenges in Materials Data Analytics

There are many outstanding challenges slowing the adoption of machine learning techniques into materials engineering. Foremost is the limited availability of the raw ingredients of the machine learning models: well-structured materials data. Additionally, the lack of re-usable techniques and general-purpose software for materials data analytics means both that researchers are unnecessarily recreating infrastructure and that these techniques are difficult to employ for the non-expert. Furthermore, there is no established procedure for determining whether materials a machine learning models is reliable enough for a given application. While the challenges are significant, there is also an extensive and growing research community working towards addressing these problems. For example, there are now tools for simplifying the storage of materials data,[102–104] there are researchers releasing materials informatics software,[105] and efforts devoted to understanding the uncertainties of machine learning models created from materials data.[106] With time, data analytics tools may yet become prevalent in materials engineering.

3 Completing Incompletely-Solved Structures with FPASS

3.1 Abstract

The structures of three Na-Pb compounds, γ, δ and δ', have remained incompletely solved for nearly 60 years. The space group, lattice parameters, and positions of Pb atoms of these three structures have been determined, but the positions of the Na atoms are still unknown. In this work, we used the First-Principles Assisted Structure Solution (FPASS) method (Meredig & Wolverton, *Nature Materials,* 2013) to complete the description of these three structures using only experimental information available from the literature as input. We also discuss the relative advantages of FPASS in comparison to conventional crystal structure prediction methods in reference to their abilities to complete the solution of other unsolved structures.

3.2 Introduction

While modern tools for determining crystal structures are quite advanced, it is not uncommon that the structure of a compound cannot be determined with the available experimental data. In fact, thousands of entries in the Powder Diffraction File (PDF) and Inorganic Crystal Structure Database (ICSD) are not associated with a crystal structure. Each of these incomplete entries represents a gap in the scientific knowledge and a material whose properties cannot be better understood by assessing their atomic-scale structures. In many of these cases, it was possible to determine at least some information about the crystal, such as its composition and symmetry group. As an example, Weston and Shoemaker attempted to solve the structures of three Na-Pb compounds in 1957 and failed.[107] They were able to determine the lattice parameters, space group, and even the positions of the Pb atoms, but

were unable to solve the positions for the Na atoms and, to this day, the structures have yet to be solved.

In the case of these unsolved Na_xPb_y compounds, the diffraction data is no longer available. Rather than repeating the diffraction experiments required to use conventional crystal structure solution techniques, we propose that these structures can be solved using crystal structure prediction algorithms. Crystal structure prediction (CSP) algorithms are designed to determine the lowest-energy crystal structure when provided with at least the composition of the structure in question.[8] These CSP algorithms have the advantage of requiring no experimental input to determine the ground state structure of a compound, which makes them ideal for solving the Na_xPb_y structures considered in this work and, possibly, useful tools in addressing the large number of other unsolved structures.

As the number of possible crystal structures for a given composition is too large to be exhaustively evaluated,[108] CSP algorithms rely on evaluating a subset of these possibilities that is likely to contain the true ground state, and are quite varied in their approaches to determining that subset. For instance, there are a wide variety of CSP methods that employ a specially-designed global optimization algorithms to search efficiently through the space.[108–114] Alternatively, one could assume that the ground state structure is similar to a structure that has already been observed experimentally and evaluate a list of already-known crystal structure types as potential solutions.[95,100,115,116] Both of these classes of CSP techniques have been used extensively in the literature to determine the structure of compounds when it was impractical to do so experimentally.[8,95,117,118]

A third, computationally more efficient approach is to use information that is already known about a crystal structure to constrain the search for the correct solution and employ a CSP algorithm to locate the lowest energy structure within those constraints.[49,63,64] The concepts behind such constrained methods are that (1) employing these constraints speeds the calculation by reducing the number of possible candidates and (2) evaluating candidates based on both energetic feasibility and consistency with experimental measurements (e.g., lattice parameters, diffraction patterns) will eliminate spurious, low-energy solutions that are inconsistent with experimental observation.[49] One such combined method, the First-Principles-Assisted Structure Solution (FPASS) method, uses a genetic algorithm to search for materials that both match a powder diffraction pattern and have minimum energy according to *ab initio* Density Functional Theory (DFT) calculations. This method has been used previously to solve structures that proved difficult with conventional crystal structure prediction and solution techniques,[63] and has the ability to constrain based on symmetry. Given that the space group, lattice parameters, and Pb atomic positions are known for the compounds studied in this work, FPASS is a suitable tool for solving their structures.

In this chapter, we present the solutions to three long-unsolved Na-Pb crystal compounds, the γ, δ, and δ’ phases.[119] Additionally, we investigate the effect of supplying FPASS with different amounts of experimental information and discuss the relative advantages of constrained methods compared to crystal structure prediction strategies with reference to their ability to be used to solve incompletely-determined structures.

Table 3.1. Structures for the γ, δ, and δ' phases, as determined using FPASS. The composition, lattice parameters, space group, and Pb positions were originally determined by Weston and Shoemaker.[107]

Phase	δ	δ'	γ
Composition	Na_5Pb_2	Na_9Pb_4	$Na_{13}Pb_5$
Space group	$R\bar{3}m$ (166)	$P6_3/mmc$ (194)	$P6_3/mmc$ (194)
Lattice Parameters	a = b = 5.54 Å c = 23.15 Å	a = b = 5.47 Å c = 30.41 Å	a = b = 5.51 Å c = 40.39 Å
Atom Positions	Na (0 0 0) Na (0 0 0.785) Na (0 0 0.357) Pb (0 0 0.070)	Na (0 0 0.183) Na (1/3 2/3 3/4) Na (1/3 2/3 0.635) Na (1/3 2/3 0.091) Na (1/3 2/3 0.519) Pb (0 0 0.050) Pb (1/3 2/3 0.020)	Na (0 0 0) Na (1/3 2/3 0.211) Na (0 0 0.083) Na (1/3 2/3 0.617) Na (0 0 0.167) Na (1/3 2/3 0.710) Na (1/3 2/3 0.531) Pb (0 0 1/4) Pb (1/3 2/3 0.05) Pb (1/3 2/3 0.13)

3.3 Methods

3.3.1 Experimental Data

The three structures of interest in this work are Na-rich binary Na_xPb_y compounds originally discovered by Weston and Shoemaker (W&S) in 1957.[107] W&S were able to determine some information about the structures and published them as an abstract for a presentation at the 4th IUCr Congress, which is the only source of data used for solving these structures. Since then,

these structures have remained incompletely solved. For clarity, we describe them using the notation used in the phase diagram reported by Hultgren:[119]

γ - $Na_{13}Pb_5$: This phase was originally reported to have a stoichiometric composition of Na_5Pb_2. W&S were able to determine the space group, lattice parameters, and positions of Pb atoms (shown in Table 3.1) using a combination of powder and single-crystal X-ray diffraction techniques.[107] The authors were unable to determine the positions of Na atoms. This phase is known to have a composition of approximately 71.4 at% Na – Na_5Pb_2 – at its melting temperature.[107] W&S were also able to determine that the sodium positions are likely to be partially occupied and the composition of γ with all sites fully-occupied is 72.2 at% Na ($Na_{13}Pb_5$). In this work, we used this information to simplify the solution process by assuming all sites are fully occupied and the composition of γ is $Na_{13}Pb_5$.

δ - Na_5Pb_2: This is a high temperature phase with a composition near Na_9Pb_4. This phase is known to have a space group of $R\bar{3}m$ with lattice parameters and Pb positions shown in Table 3.1. As with the γ phase, the positions of Na atoms are yet unknown. As with the γ structure, W&S hypothesized that the Na positions are partially occupied and proposed that the structures had a composition of 71.4 at% Na (Na_5Pb_2). As with the γ phase, we assume all sites are fully-occupied and the composition is Na_5Pb_2 when solving the structure of δ.

δ'- Na_9Pb_4: This low-temperature, hexagonal phase ($P6_3/mmc$) has a composition of Na_9Pb_4. As with δ and γ, the lattice parameters, Pb positions (but not those of the Na atoms), and space group were also determined using X-Ray Diffraction techniques, and are shown in Table 3.1.

3.3.2 Structure Solution Method

We employed the recently-developed First-Principles-Assisted Structure Solution (FPASS) method to solve each structure.[63] FPASS works by using a genetic algorithm to locate the lowest-energy crystal structure out of all structures that match any existing, known structural information, which can include lattice parameters and space group. In cases where diffraction data is available, this search is further guided by preferentially evaluating candidate structures that are better matches to the powder diffraction pattern of the compound. The inclusion of both diffraction pattern matching and constraining searches to a certain symmetry group has been shown to allow FPASS to resolve the correct structure when both conventional crystal structure solution and crystal structure prediction methods are unable to determine the correct crystal structure with certainty.[63]

We used the results from a study by Weston and Shoemaker as a starting point for our solution process,[107] as described in the previous section. For all three cases (γ, δ, and δ'), the space group, lattice parameters, and the positions of the Pb atoms were known. Unless otherwise mentioned, all of this information was employed to define the space of possible crystal structures evaluated using FPASS. While both single crystal and powder X-ray diffraction were used to characterize each compound in the original study from 1957, the diffraction data was not reported in the original papers, and therefore is not available to help solve the structures.

We used a population size of 10 structures for the genetic algorithm and the optimization was halted once the energy of the optimal structure failed to change by more than 5 meV/atom

after 5 generations. At each generation, the best performing structure from the previous generation was kept in the population. Mutation and crossover operations were slightly different than those used in the original FPASS paper, and are described in the following chapter. Mutation probabilities of 50% were used for both the Wyckoff site combinations and atom positions. Wyckoff-site biasing, as described in Ref. [63], was not found to be necessary for these structures. FPASS was run 10 times with different random number seeds for each compound, and the structure with the lowest energy out of all runs was selected to be the candidate solution. The software used to perform FPASS is available under an open source license from: http://github.com/materials/mint.

3.3.3 Energy Calculations

We used Density Functional Theory [15,16], as implemented in the Vienna *Ab Initio* Software Package (VASP)[120,121], to evaluate the energy of each candidate crystal structure. In particular, we employed the projector augmented-wave method [122] with the Perdew-Burke-Ernzerhof generalized-gradient approximation for the exchange-correlation energies.[123] We used pseudopotentials for Na and Pb that treat the $3s^1$ and $6s^26p^2$ electrons as valence, respectively, with a cutoff energy of 102 eV and a gamma-centered mesh of 1000 k-points per reciprocal atom in all calculations.[121,122]

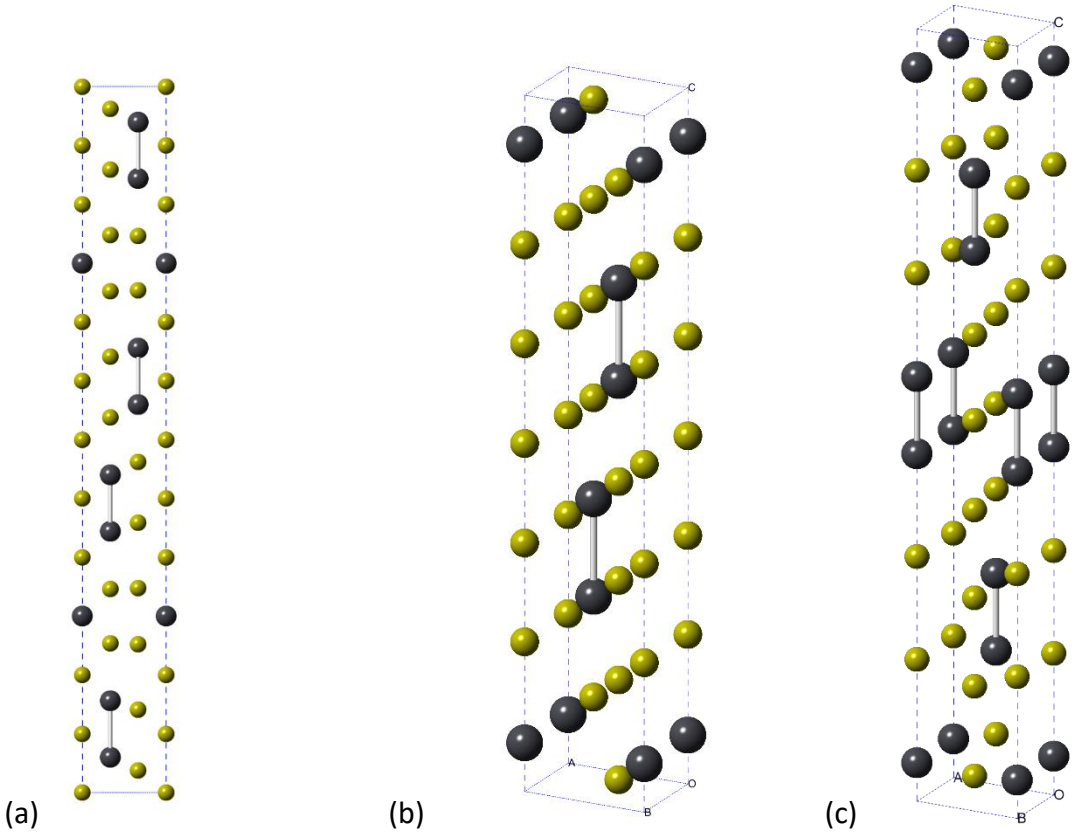
When comparing the energy of our proposed solutions against those of other Na_xPb_y compounds, we used the same DFT settings of the Open Quantum Materials Database (OQMD).[32] These more accurate parameters include a higher cutoff energy of 520 eV and a k-point mesh of 8000 points per reciprocal atom. Additionally, performing energy calculations

with these settings made it possible to use energies available in the OQMD directly in the analysis of our results without repeating any calculations. It is worth noting that while these DFT settings are more accurate, they are too computationally expensive to be practical for use with FPASS.

3.4 Results and Discussion

3.4.1 δ Phase

The δ phase is a high temperature phase with a rhombohedral structure that reversibly transforms into hexagonal δ' below 190°C.[107,119] (Below, we consider the δ' phase.) W&S



(a) (b) (c)

Figure 3.1. Proposed solution for (a) γ, (b) δ, and (c) δ', as determined using FPASS. As hypothesized by Weston and Shoemaker[107], atoms lie along the $\langle 0,0,z\rangle$, $\langle\frac{1}{3},\frac{2}{3},z\rangle$, and $\langle\frac{2}{3},\frac{1}{3},z\rangle$ lines in all three cases.

found that the stoichiometry of this structure is Na_5Pb_2 with all Na sites fully occupied, the space group of the structure is $R\bar{3}m$ (166), and the locations of all Pb atoms in the structure. The only missing piece of information about the structure is the positions of the Na atoms. To find these positions, we used FPASS to locate the structure with the lowest energy that satisfies all of the known information about the structure (i.e., lattice parameters, Pb positions, and space group). Each FPASS solution requires evaluating between 120 and 200 candidate structures, which required only 3 hours for all 10 runs of FPASS on two 8-core, 2.6 GHz processors. All ten runs returned the same structure: a crystal that is iso-structural with Li_5Tl_2 and Li_5Sn_2.[124,125] The structural parameters of our proposed solution are listed in listed in Table 3.1.

As with the γ structure, we were able verify that our structure matches other quantitative characteristics determined by the original investigators. According to Weston and Shoemaker, the Na_5Pb_2 structure should have 6 atoms on the lines $[0,0,z]$, $\left[\frac{1}{3},\frac{2}{3},z\right]$, and $\left[\frac{2}{3},\frac{1}{3},z\right]$.[107] Our final structure satisfies this geometric constraint. However, we should note that every structure that matches the number of atoms in the unit cell, Pb positions, and space group from W&S automatically fits this requirement. W&S also proposed that this structure is a supercell of b.c.c., which we were able to confirm using the newly-solved Na positions. We found that the (1 1 0) plane of this structure is parallel to the (1 1 0) plane of the underlying b.c.c. lattice, which has a lattice parameter of approximately $a \approx 3.9$ Å. While each Pb atom in the structure features exactly 1 Pb nearest neighbor (as originally suggested by W&S), the structure features Na atoms with between 0 and 4 Pb nearest neighbors.

As further test of our solution, we performed a test where we provided FPASS a space group of lower symmetry than what was determined experimentally ($P\bar{3}m1$) and a second test where no symmetry information was provided. By easing the symmetry requirements, we allow the algorithm to test a larger number of possible configurations to see whether any are lower energy solutions that do not fit all of the provided constraints. Even though it was possible for Na atoms to be located off of the lines predicted by W&S in these tests, the FPASS result in both cases was identical to the result found when FPASS was provided full symmetry information. Finding the correct structure in these cases both supports the space group determination of W&S and demonstrates how FPASS can be used with incomplete symmetry information. We also found that FPASS predicts the same structure when the Pb positions from Weston and Shoemaker were not used and the space group was assumed to be $R\bar{3}m$. As with the tests with reduced symmetry information, finding the same structure as the fully-constrained test supports our conclusion that we have found the correct structure for the δ phase.

The fact that FPASS returns the same structure in each test shows that the algorithm is capable of finding the solution even with limited initial data. However, the real advantage of being able to employ already-known information about a crystal in FPASS is reduced computational time. When using only the Pb positions and lattice parameters, a single FPASS calculation to solve this structure requires approximately 10.5 hours of computing time. By incorporating only lattice parameters and symmetry (i.e., no Pb positions), the required time decreases to 1.3 hours. If we provided FPASS with all of the known information about the structure, we can increase the speed of the solution to only 22 minutes per calculation – an

acceleration of over 30x the test without symmetry information and 3.5x faster than without Pb positions.

3.4.2 γ Phase

The crystal structure of the γ phase was determined by W&S to have the symmetry group of $P6_3/mmc$ (194) and a stoichiometry of $Na_{13}Pb_5$ with 36 atoms in the unit cell when all sites are fully occupied. Additionally, they were able to determine the positions of all 10 Pb atoms. In this work, we completed the description of this structure by solving for the lowest-energy positions of the Na atoms in structures that fit these constraints using FPASS (see Figure 3.1a and Table 3.1). Three out of ten FPASS calculations found this structure, which had the lowest DFT energy of all candidate structures for this phase. Each solution required, on average, 6 hours on two eight-core, 2.6 GHz processors.

We were able to validate our solution using a few characteristics about the structure that were determined by Weston and Shoemaker: (1) 12 atoms exist along the $[0,0,z]$, $\left[\frac{1}{3},\frac{2}{3},z\right]$, and $\left[\frac{2}{3},\frac{1}{3},z\right]$ lines through the unit cell, and (2) four-fifths of the Pb atoms have exactly one Pb nearest neighbor. Our structure meets both criteria. The second criterion is satisfied by the Pb positions provided as input to FPASS, and our proposed solution trivially meets this requirement as a result. In contrast, the fact that our structure satisfies first criterion (which was not predetermined by the input parameters) shows that our solution matches the experimental data first determined by W&S and provides validation of the structure's accuracy. We found that this structure, like the δ phase, is based on a distorted b.c.c. superstructure, as

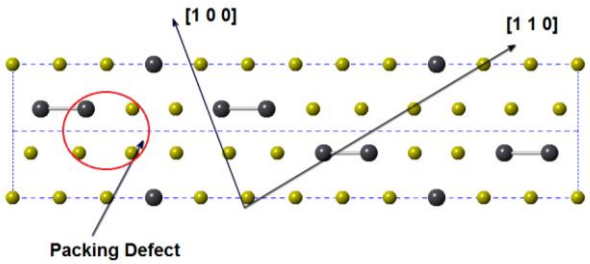


Figure 3.2. (1 1 0) plane of the *γ* structure. The black lines indicate the approximate [1 0 0] and [1 1 0] directions in the underlying Na b.c.c. lattice, which is heavily distorted by the presence of the Pb atoms. The region containing a defect from this b.c.c. lattice is indicated with a red ellipse.

was originally hypothesized by W&S. In this case, the lattice is not only distorted but also contains a defect from the ideal lattice, as shown in Figure 3.2.

There is no other structure in the ICSD that has the same stoichiometry ($A_{13}B_5$), number of atoms in the unit cell (36), and space group ($P6_3/mmc$) as the one found we found. The only other structure with $A_{13}B_5$ stoichiometry that matches the second criterion (80% of the Pb atoms present in pairs) given by Weston and Shoemaker is that of $Li_{13}Sn_5$, which was found to be nearly degenerate (~2 meV/atom lower in energy) with the γ structure we determined (with Na/Pb replacing Li/Sn). However, the space group of $Li_{13}Sn_5$ and the positions of the Pb atoms are different than what was found by Weston and Shoemaker. Additionally, while the $Li_{13}Sn_5$ structure is also a superstructure of b.c.c. [126], it lacks the deviation from perfect packing found in our solution. Assuming that the original space group determination was correct, the solution of the γ phase structure shows the unreliability of simply relying on energy and searching only known prototypes when solving a crystal structure. Had we relied only evaluated the energy of known structures, we would have incorrectly concluded the $Li_{13}Sn_5$ structure was the solution for the gamma phase structure.

3.4.3 δ' Phase

The δ’ phase is stable at low temperatures and has a composition of Na_9Pb_4. The space group of its structure ($P6_3/mmc$), lattice parameters, and Pb positions were determined by Weston and Shoemaker and we found the Na positions using FPASS. Our solution for the structure of δ’, shown in Table 3.1c and Table 3.1, fits the descriptions supplied by Weston and Shoemaker: 8 atoms along the $[0,0,z]$ line, 9 atoms along the $\left[\frac{1}{3},\frac{2}{3},z\right]$ and $\left[\frac{2}{3},\frac{1}{3},z\right]$ lines, and all Pb atoms have exactly one Pb nearest neighbor. Each solution required, on average, 1.6 hours on two eight-core, 2.6 GHz processors. Four out ten FPASS solutions found the same ground state structure. Our proposed solution for δ’ is not isostructural to any other phase in the ICSD, so simply searching a database of known crystal structure prototypes would have failed to correctly solve this compound. The other A_4B_9 hexagonal crystals in the ICSD do not match the criteria given by W&S and, according to DFT calculations, are higher in energy than our solution by at least 100 meV/atom.

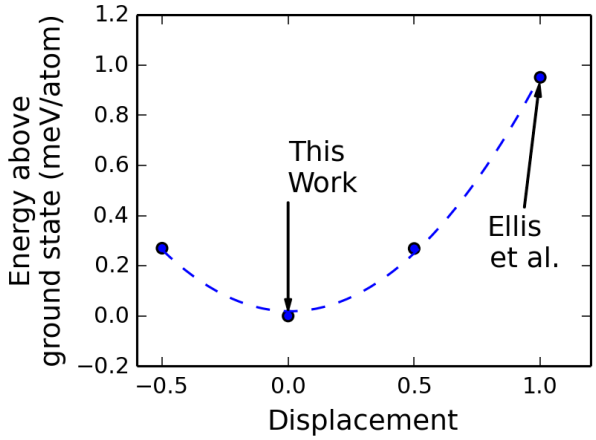


Figure 3.3. DFT Energy of structures that interpolate between the structure for δ’ determined in this work and a distorted version proposed by Ellis *et al*.[127] A displacement of 0.0 corresponds to our hexagonal solution, and 1.0 to the orthorhombic structure of Ellis *et al*. Energy is shown to increase with displacement, which demonstrates that the structure proposed by Ellis *et al* is dynamically unstable.

Recently, Ellis *et al* proposed that δ’ has the orthorhombic, $Na_{13}Sn_5$ structure (which has a fully-occupied stoichiometry of A_4B_9) – a distorted version of our solution.[127] In order to determine whether this distortion is

real or just an artifact of an incomplete structure refinement from X-ray data, we first relaxed the atomic positions and lattice parameters of the Ellis *et al* structure to their minimum DFT energy values. Then, we adjusted the atomic positions in the structure so that they matched the positions of our higher symmetry solution. Next, we calculated the energy of several structures whose atomic positions interpolated between those of the higher-symmetry structure and Ellis's solution. As shown in Figure 3.3, we found that our structure is more stable and that the energy of the structure increases with larger displacements. Consequently, we conclude that the structure proposed by Ellis *et al*. is dynamically unstable and that our solution – an undistorted version of the $Na_{13}Sn_5$ structure - is a better representation of the δ' phase.

3.4.4 T=0K Na-Pb Ground State Phase Diagram

As an additional step of validation, we compared the energy of each structure at 0K (computed using DFT) to that of every other known compound in the Na-Pb binary system. The energies of the other compounds (Na, $Na_{15}Pb_4$, NaPb, $NaPb_3$, and Pb) were taken directly from the Open Quantum Materials Database (OQMD) or computed using its associated toolkit.[18,32] The DFT-calculated formation enthalpies are shown in Figure 3.4 along with the convex hull (solid black line), which represents the energy of the lowest-energy combination of phases at a certain composition.

Two of the three proposed structures (δ and δ') are stable at 0 K with respect to any combination of all other known phases. Since the δ' phase is known to be stable at low temperatures, the fact that we found its structure to be the 0 K ground state supports that we have found the correct structure. The δ phase is only known to be stable at elevated temperatures and is observed to exist at an off-stoichiometric composition near $Na_{0.69}Pb_{0.31}$. We did find this structure to be stable at 0 K at its stoichiometric composition of Na_5Pb_2, which suggests that it is energetically feasible to form at high temperature. The fully-occupied structure of the third phase, γ, was found to be unstable 0 K, which is consistent with it only being known to be stable at high temperatures and is observed to have a composition of Na_5Pb_2. Even so, the γ phase structure is only unstable by 4 meV/atom, which is not unfeasibly large. Given that *kT* at room temperature is around 25 meV/atom, it is possible for γ to be stabilized by entropic contributions to the free-energy at modest temperatures. As a result, we conclude that our solution for its structure is also energetically reasonable.

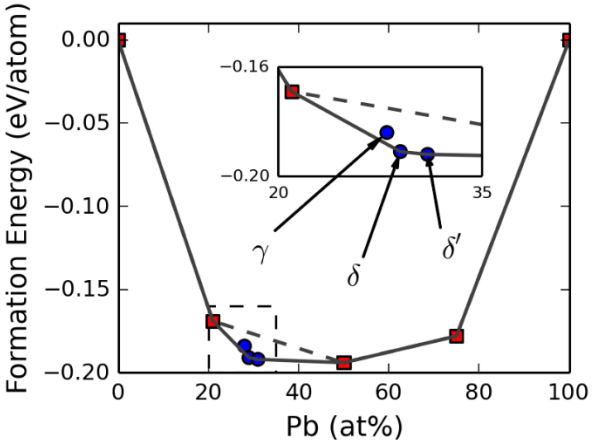


Figure 3.4. Phase diagram of Na-Pb calculated using DFT showing the formation energies of compounds with already-known structures (red squares) and those solved in this work (blue circles). Solid line indicates convex hull for this system. Dashed line represents the convex hull before introducing the compounds solved in this work. The region highlighted by the inset is shown in black dashed lines. All three of the structures that were solved in this work were found to be low in energy and either stable (i.e. on the solid black line) or close to it, which suggests they are energetically feasible.

Finding additional stable compounds in the Na-Pb system impacts the calculated

phase diagram and the corresponding chemical potentials in that composition region, which affects both the accuracy and feasibility of atomistic-simulation-based studies. For example, these chemical potentials are of great importance when determining defect energies with DFT – calculations which have been used to guide the doping of Na into PbTe thermoelectric materials.[128] Additionally, Pb has recently been studied as a possible anode material for non-aqueous sodium-ion batteries.[127] A more-complete database of Na-Pb structures now makes it possible to study electrochemical reactions in this battery system with atom-scale modeling.

3.4.5 Advantages of Constraining Structure Search

The solution of the Na-Pb compounds in this work demonstrates that FPASS is a suitable tool for solving crystal structures when limited information about the structure is already available and, in general, highlights the advantages of constraining a CSP algorithm using that information. While originally designed to solve structures given diffraction data, we have shown that FPASS is robust enough to solve structures lacking this information and, in some cases, even lacking complete symmetry information about the structure. Furthermore, we found that it is possible to solve the structure of phases that are only stable at high temperatures using FPASS, as demonstrated by the solution of the structure of the γ phase. By restricting the search to only structures that match experimental measurements, spurious solutions that happen to be lower in energy at 0 K are avoided – which could be recurring problem during the solution high-temperature phases (as suggested by the solution of the structure of the γ phase).

The solution of the Na-Pb compounds in this work also highlights the deficiencies of using crystal structure prediction techniques that do not enforce consistency with experimental

observations. In the solution of the structure of the γ phase, a prediction method that only considers energy would have found the $Li_{13}Sn_5$ structure because it is lower in energy even though this structure does not match the experimentally-determined space group. Additionally, by constraining based on space group and lattice parameter, the next lowest energy solution is at least 50 meV/atom higher in energy than the best solution in each case, which is sufficiently large to confidently select that solution as the true ground state. In contrast, finding several, nearly-degenerate ground states (e.g., at least $Li_{13}Sn_5$ and our solution when solving *γ*) would complicate selecting the true solution.

Furthermore, techniques that do not consider already-available information about a crystal structure could be drastically slower. As an example, the solution of the δ’ structure required 30x more time when the experimental symmetry group was not used – and this figure would only increase if the Pb positions and lattice parameters were also ignored. This results suggests that constraining a structure search using symmetry and known positions could have performance benefits in other methods.

3.5 Conclusions

In this chapter, we demonstrated how the First Principles Assisted Structure Solution (FPASS) can be used to solve incompletely-determined crystal structures. In particular, we used FPASS to solve the structures of three Na-Pb compounds (γ, δ, and δ’) that had remained unsolved since in 1957.[107] Through these solutions, we show that FPASS is able to solve structures that are unstable at 0 K and can be used to determine the correct structure even with incomplete symmetry information and without a diffraction pattern.

4 Automated Structure Solution from Powder Diffraction Data

4.1 Abstract

High-throughput *ab initio* computational methods offer the ability to automatically predict the properties of materials, provided their crystal structures are known. However, there are many compounds for which the structure is unknown and, consequently, many potentially-useful materials that are unable to be assessed by high-throughput searches. Here, we demonstrate an automated tool to solve the structures of materials from powder diffraction patterns based on the First-Principles-Assisted Structure Solution (FPASS) method. We first validated this tool by using it to solve approximately 90 known crystal structures, and then applied it to the solution of two dozen unsolved crystal structures. Of these candidates, we were able to successfully solve the structures of 10 materials and found, using high-throughput DFT, several are interesting candidates for semiconductors.

4.2 Introduction

Determining the crystal structure of a material is often the first step in being able to understand or predict its properties. In fact, crystal structure is the only requirement to predict the properties of a material with Density Functional Theory (DFT). As demonstrated by the recent advancements in high-throughput DFT, DFT can be used to automatically evaluate whether a material is an interesting candidate for many potential applications.[19,20,23,27,31,32,38] However, despite significant advances in techniques to solve crystal structures from powder diffraction data, there are many materials for which the structure is currently unknown. For example, there are thousands of diffraction patterns that

are not associated with a crystal structure in the Powder Diffraction File (PDF) and tens of thousands of entries in the Inorganic Crystal Structure Database (ICSD) that are incomplete.[68,129] Beyond simply filling in gaps in scientific knowledge, solving these structures would significantly expand the databases of material properties computed using high-throughput DFT.[18,32,37,38]

Reconstructing 3-dimensional crystal structures from 1-dimensional powder diffraction data is a nontrivial problem even in an ideal case.[51] Conventionally, solving a crystal structure involves first determining the shape, symmetry, and content of the unit cell, and then performing the "structure solution" step to determine atomic positions.[49] A variety of techniques exist for performing structure solution, which all require varying degrees of expert judgement to employ.[55] Given the large number of unsolved structures, the amount of time required by experts must be kept as small as possible. Specifically, what would enable the solution of these crystals is a set of automated tools for crystal structure solution.

One potential tool for automated crystal structure solution is the First-Principles-Assisted Structure Solution (FPASS) method.[63] The FPASS method functions by using energies calculated from DFT to guide the search for structures that have both minimal potential energy and an optimal match to experimental data. As FPASS is designed to solve crystal structures given composition, lattice parameters, and symmetry, it fills exactly the need for the many unsolved crystal structures in the PDF that already have this information. Furthermore, as FPASS uses DFT to compute the energy of candidate crystal structures, no problem-specific selection of an empirical potential is required – making it possible to run large numbers of

FPASS calculations without any need to first validate an empirical potential.[49,64,65] However, to date, FPASS has only been tested in a few case studies[63,130,131] and no automated implementation of this method exists.

In this work, we present a new implementation of the FPASS method and demonstrate that it can be used to automatically solve crystal structures. Our method includes a modified version of the original genetic algorithm and software capable of matching candidate solutions to raw diffraction patterns. To evaluate the performance of this method, we first validate the ability of the algorithm to solve nearly 100 common crystal structures given the known diffraction pattern, unit cell, and symmetry. During these validation tests, we determined that crystals structures with large numbers of possible combinations of Wyckoff sites are difficult to solve with FPASS and show how the algorithm that be tuned to perform better on such cases. With this knowledge, we applied FPASS to dozens of unsolved crystal structures from the PDF and were able to solve a significant fraction of them automatically. We then added these structures to the Open Quantum Materials Database (OQMD)[18,32] and used high-throughput DFT to predict their properties, and found several were promising for semiconductor applications. As our method requires minimal human interaction, we plan to continuously apply it to unsolved structures and then automatically predict the properties of these materials using high-throughput DFT.

4.3 Methods

A general outline of how crystal structures are solved from powder diffraction data using FPASS is shown in Figure 4.1. FPASS is used to determine atomic positions in a crystal after the

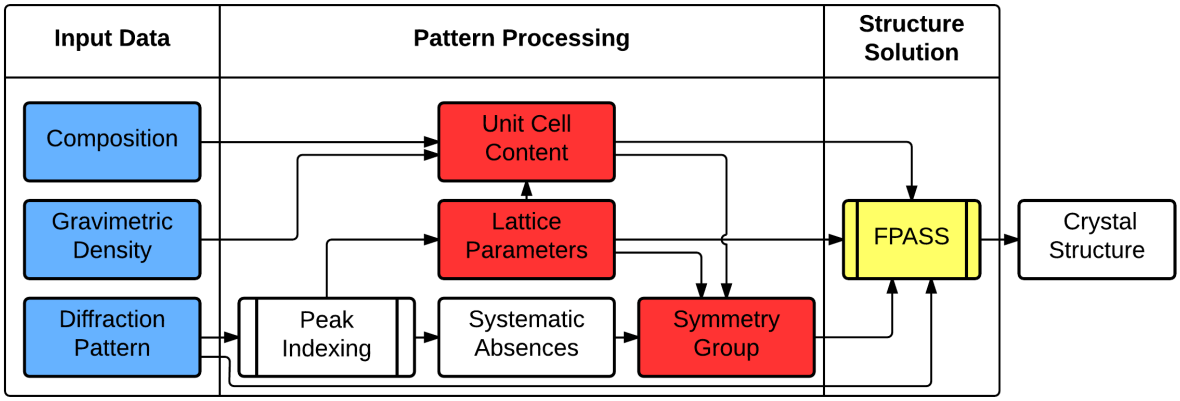


Figure 4.1. Flow chart describing the process of solving a crystal structure from powder diffraction data using FPASS. The typical solution process starts by measuring the composition, diffraction pattern, and gravimetric density of a compound, which are then used to determine the unit cell contents, lattice parameters, and symmetry group of a crystal. Once this information is determined, FPASS is used to find the lowest energy structure within these constraints.

lattice parameters, unit cell contents, and space group have been determined with peak indexing techniques. In this section, we will describe the theory and techniques behind FPASS. In particular, we will discuss the method used to compute the energy of a candidate structure, how our software determines how well a structure matches a powder diffraction pattern, and the details of the genetic algorithm used to efficiently solve the structure.

4.3.1 FPASS Software

The FPASS algorithm is implemented as part of the Materials Interface (Mint) software, which performs all steps in FPASS automatically. Mint itself was designed to perform many common tasks in the atomic-scale simulation of materials, including symmetry determination and generating input files for various simulation packages. Consequently, many of the analyses required by FPASS (e.g., symmetry determination) were already present before we implemented FPASS. Mint is written in C++ and available freely under the LGPL license.[132]

4.3.2 Computing Energy of Candidate Structures

We use Density Functional Theory (DFT) to compute the energy of candidate structures, which enables the reliable prediction of energies without the need to fit or select appropriate empirical potentials. DFT only requires the crystal structure in order to compute energy and is known to be able to reliably predict the ground state structure and formation energy of many inorganic systems.[18,32] As a result, DFT makes an excellent choice for a high-throughput solution tool – one can be confident in the accuracy of the calculated energies for a broad variety of crystal structures without needing to validate the energy calculation method before each solution.

For this work, we performed all DFT calculations using the Vienna *Ab Initio* Simulation Package (VASP).[120,133–135] Unless otherwise stated, all calculations were performed with projector augmented-wave basis sets,[121,122] the GGA exchange-correlation functionals of Perdew, Burke, and Ernzerhof,[123] a cutoff energy of 1.3 times the maximum cutoff energy of all of the provided pseudopotentials, and 1000 K-points per reciprocal atom. We employed the DFT settings used by the Open Quantum Materials Database (OQMD), a collection of the structures and DFT-predicted formation energy of hundreds of thousands of crystalline materials, when comparing the stability of a candidate crystal structure against others.[18]

4.3.3 Matching Candidate Structures against Powder Diffraction Patterns

Another component of FPASS is methods to determine of how well a proposed structure agrees with the experimental diffraction pattern. Comparing a structure to an XRD pattern requires being able to compute XRD patterns for hypothetical structures, processing

experimentally-measured patterns, and adjusting a structure to better match the diffraction pattern. Each of these techniques are detailed in the following subsections.

4.3.3.1 Diffraction Pattern Calculation

Techniques to calculate the diffraction pattern of a crystal structure are well-established in the crystallography community.[136,137] At a high level, the powder diffraction pattern is calculated by first finding all reflections that will occur within a certain range of diffraction angles. Next, reflections that would cancel each other out in a powder diffraction pattern due to the symmetry of the crystal are removed and symmetrically-identical peaks are grouped together for computational efficiency. These two steps generate a list of diffraction peaks that should be observed in a powder diffraction pattern, and are automatically performed by Mint based on the unit cell and symmetry of the structure. Once the list of peaks is generated, the intensity of each individual peak is computed based on the atomic positions, number of overlapping peaks, thermal factors describing the thermal oscillations of each atom (we use isotropic thermal factors), and the Lorentz and polarization factors following standard methods.[51]

4.3.3.2 Raw Diffraction Pattern Processing

Comparing a calculated diffraction pattern to an experimental pattern is easiest when comparing the integrated intensities of each peak. As the experimental data available for use may be the raw diffraction signal (intensity as a function of angle), we needed to implement an automated scheme for detecting the positions and integrating the intensities of each peak. Our

implementation uses a version of the processing algorithm described by Pecharsky and Zavalij:[51]

1. **Noise filter**: The raw X-ray pattern is first passed through a noise filter that smooths the data by averaging the diffracted intensities of points with similar diffraction angles.
2. **Background removal:** First, the background signal determined by calculating a running average where every point in a 2 degree window is assigned a weight inversely proportional to the 4th power of the intensity at that point. This background signal is then subtracted from the smoothed data from step 1.
3. **Peak detection**: Once the background has been removed, the locations of diffraction peaks are found by identifying local minima in the second derivative of intensity with respect to diffraction angle. By identifying peaks based on the second derivative, we can easily separate peaks that are slightly overlapped.[51]
4. **Intensity extraction**: A pseudo-Voight function, which is known to describe the peaks in X-ray diffraction patterns well,[51] is fit to the intensity values for each peak. We then integrate the area under each function to determine the intensity of each diffraction peak. The fitting functions corresponding to overlapping peaks are fitted concurrently in order to accurately determine the contribution of each individual peak.

4.3.3.3 Matching Structures to Diffraction Pattern

Before measuring how well a structure matches an experimental diffraction pattern, we adjust the candidate structures so that its computed pattern better matches the reference pattern – a technique known as structure refinement. Refinement is accomplished by minimizing the function

$$\mathrm{R} = \frac{\sum(I_{calc} - s \times I_{obs})^2}{\sum s \times I_{obs}} \quad (3)$$

where I_{calc} is the calculated integrated intensity of each peak, I_{obs} is the integrated intensity of the same peak in the experimental pattern, and s is a scaling factor. Each peak in the observed pattern is assigned to the closest peak of any peak in the reference pattern that is within 0.15 degrees. If multiple observed peaks match a single peak in the reference pattern, their

intensities values are added together. Unassigned peaks from both the observed and reference pattern are treated as being matched to a hypothetical peak of zero intensity. The computed intensity, I_{calc}, is a function of several different factors (e.g., atomic positions, thermal factors, texturing), which we optimize using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, as implemented in Dlib.[138] Following the procedure described by Pecharsky and Zavalij,[51] we sequentially add more fitting parameters (starting with the scale factor) during the optimization. After refinement, we use the optimized R to describe the match to the diffraction pattern.

4.3.3.4 Rietveld Refinement

When matching the computed XRD pattern against the raw intensity measurements, we employ a method known as Rietveld refinement.[139] In contrast to the previous section where only the integrated intensities of each peak are considered, the computed X-ray pattern of a structure is compared to the entire experimental X-ray diffraction pattern in Rietveld refinement. Here, the match between two patterns is defined by

$$R_p = \sum|I(obs.)_i - s \times I(calc.)_i| \,/\, \sum|s \times I(obs)_i| \tag{4}$$

where I_i is the observed or calculated diffracted intensity above the background signal at angle $2\theta_i$.[140] In order to compute this quantity, one must subtract the background signal from the pattern and evaluate the sum only over regions where $I(obs.)$ is positive. To do so, it is necessary to determine the background signal in the diffraction pattern, which we perform by fitting the background signal to Chebyshev polynomials. Additionally, we use pseudo-Voight functions to describe the shape of each diffraction peak. The parameters for the background

signal and peak shapes are fit in addition to the parameters described in the previous section. As this optimization problem is significantly costlier than optimizing using only integrated peak intensities, we only employ this technique when automated pattern processing has failed and for when reporting the match to diffraction data when validating a proposed structure (i.e., not during the structure solution process).

4.3.4 Efficiently Locating the Optimal Crystal Structure Solution

The FPASS method is based on a genetic algorithm designed to efficiently search through candidate crystal structures. FPASS requires the unit cell parameters and content (i.e., number of atoms of each type) as input, and can use the measured diffraction pattern, space group, and the known position of any atoms in the structure to guide and constrain the search. In particular, including the symmetry as a constraint can dramatically accelerate the solution process.[63,130] With this information, the FPASS algorithm can be used to determine the positions of atoms that minimize the energy. Genetic algorithms, in general, work by mimicking natural selection: better-performing solutions are mixed to create new candidates that are similar to them. As shown in Figure 4.2, this process is repeated for several generations until the algorithm converges on an optimal solution. In the following sections, we will describe the two parts of the GA that are unique for FPASS: how the initial population is generated and how new generations are created.

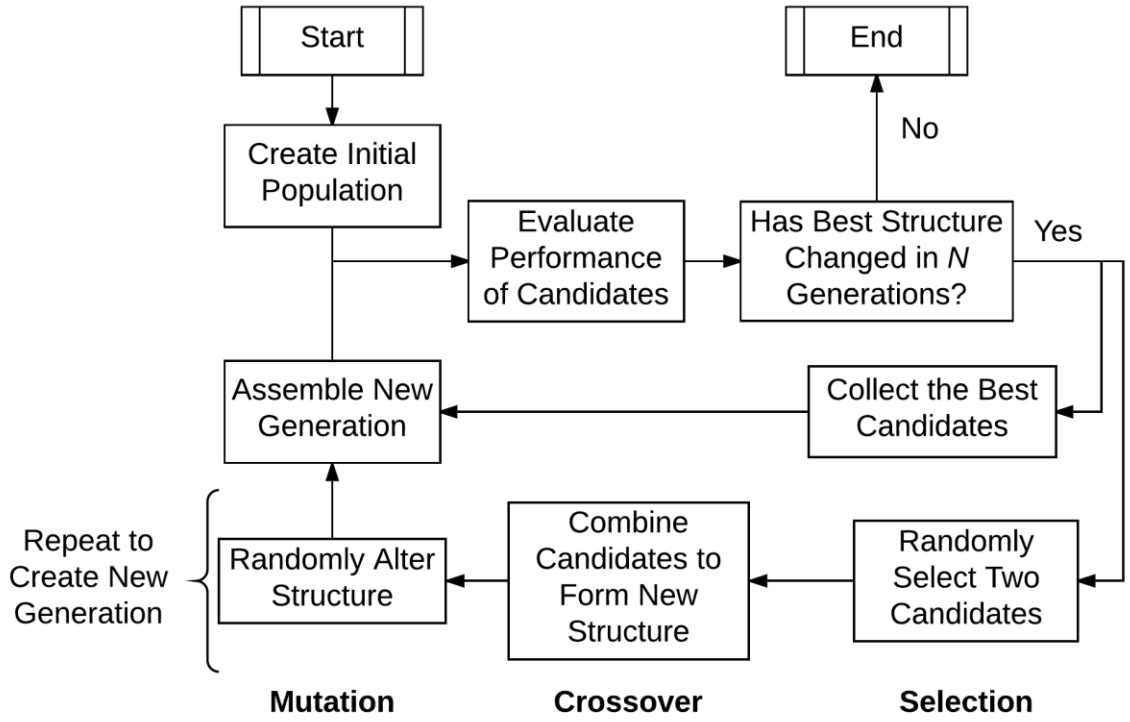


Figure 4.2. Flowchart for genetic algorithm used by FPASS. The algorithm starts by generating a random population of candidate crystal structures, and then evaluating their properties. After the initial population, new generations are created by a mixture of the best-performing compounds from the previous generation and compounds created using genetic operators. These new structures are then evaluated, and the process is repeated until the best structure does not change after N generations.

4.3.4.1 Generating Initial Population

The genetic algorithm used in our implementation of FPASS starts with an initial population of randomly-generated structures. Each initial structure is created by first selecting a combination of Wyckoff sites from the selected space group that lead to a unit cell with the correct numbers of atoms. Then, we assign random positions to each site. This process is repeated until we have the desired number of structures.

In order to more efficiently search the structure space, we bias the selection of the Wyckoff sites to be statistically similar to information mined from a large number of known crystal structures. By analyzing the Inorganic Crystal Structure Database,[68] we found that 85% of crystal structures have the minimum possible number of Wyckoff sites consistent with space group and number of atoms in the structure, as shown in Figure 4.3. To bias our initial population to have more structures with small numbers of Wyckoff sites, we first generate a list of all possible combinations of Wyckoff sites that will lead to the correct number of atoms in the unit cell, provided the symmetry group given as input to FPASS. Each of those possibilities is assigned a weight related to the fraction of structures in the ICSD with the same ratio between the number of Wyckoff sites in that crystal to the fewest-possible number of sites. These weights are used to influence the selections of combinations of sites when creating random crystal structures, and can be tuned using one of the input parameters for the genetic algorithm.

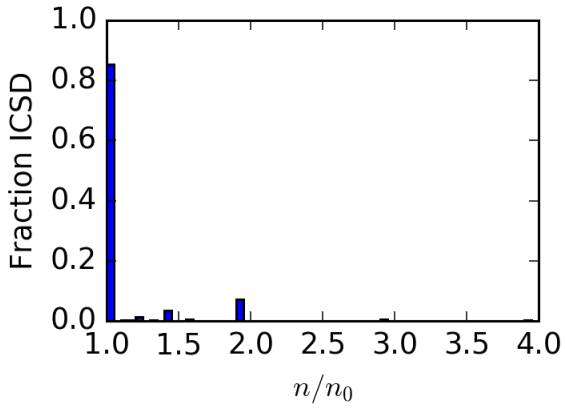


Figure 4.3. Distribution of structures in the ICSD based on the ratio between number of unique sites (n) in the crystal structure and the minimum possible number of sites (n_0). The minimum possible number of unique sites can be determined using number of atoms in unit cell and symmetry group. For the majority of structures (approximately 85%), the actual and minimum number of unique atoms are the same.

4.3.4.2 Creating New Generations

Each new generation in FPASS is generated by a combination the best-performing structures from the previous generation, and structures created using genetic operations. Before creating the new

generation, we first rank each structure in the previous generation based on the sum of its rank in both energy and match to diffraction pattern (ex: the structure with the lowest energy and second best match to the pattern would have a score of 3). The “elite” are included to ensure the best-performing structure is always considered when making new guesses.

To create the rest of the generation, we generate structures using genetic operations. For each new structure, we first select two parent structures out of the list of better-performing structures from the previous generation. To bias our selection to only the better-performing candidates, we randomly select parents from only the top $0.6\sqrt{n}$ structures, where n is the number of entries in the population. Each time we generate a new structure, we first select two parents from a list containing these top candidates, then generate a new structure with crossover, and, finally, randomly perturb that structure.

Crossover is performed by combining groups of symmetrically-equivalent atoms from either parent. For example, suppose one parent structure has a total of 8 Na atoms on the 4c Wyckoff position and the second parent has 4 Na atoms on the 4c position, 2 atoms on the 2b, and 2 on the 2a position. Our crossover method could produce a child structure that includes one of the two groups of Na atoms on the 4c position from the first parent and the atoms on the 4c position from the second parent. Or, it could generate structure that contains the second group of Na atoms on the 4c position in the first parent and the atoms on the 2a and 2b positions in the second. In total, there are 6 possible ways of combining the Na atoms from the two parents that will have the same total number of Na atoms. This procedure is repeated for

each type of atom and will create a new structure with the same symmetry and composition as the parent structures.

After crossover, for randomly-selected children, we perform one of two possible mutations: (1) perturbing the atomic positions of a group of symmetrically-equivalent atoms, or (2) selecting a different combination of Wyckoff sites. Both of these operations are designed to preserve the original symmetry of the structure. The probability of performing either type of mutation is adjustable, and the random magnitude of each perturbation allows the mutations to range from small alterations to completely-random structures.

4.4 Testing Validity and Improving Efficiency of FPASS

Before automating FPASS, we first determined whether FPASS can reliably determine crystal structures for a broad variety of structures and adjusted the settings of the algorithm in order to minimize the computational cost. The results from our validation and tuning efforts are described below.

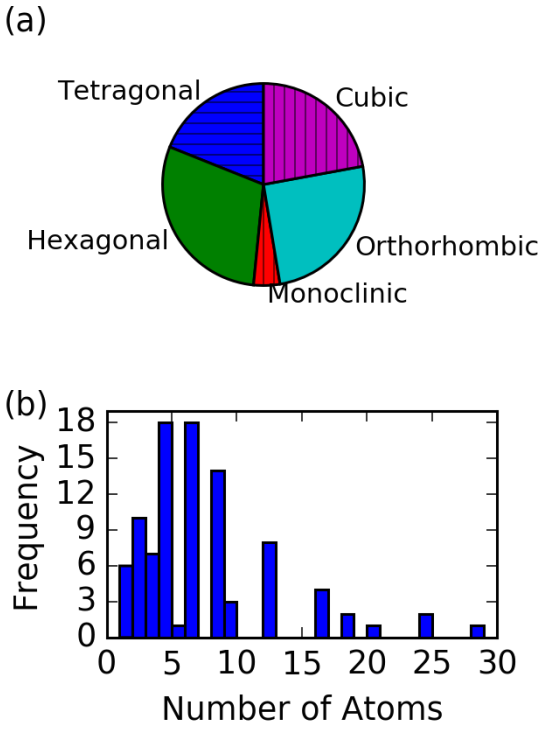


Figure 4.4. (a) Distribution of crystal families and (b) number of atoms in the primitive cell for all 95 compounds used to validate FPASS. Candidates were intentionally chosen to sample a wide variety of cell sizes, sizes, and chemistries.

4.4.1 Validating the Algorithm

While the FPASS method has been shown to be able to accurately determine the

crystal structures in a few test cases,[63,130] we further validated our new implementation before using it in an automated manner. To further validate our method, we measured how often FPASS determines the correct structure for 95 compounds with a wide variety of symmetry groups and cell sizes. As shown in Figure 4.4, we selected compounds from all crystal families with the exception of triclinic and with sizes ranging from between 1 and 30 atoms in the primitive cell. This wide variety of structures also enabled us to study how symmetry and unit cell size affect the performance of FPASS.

For each test case, we ran at least 10 individual FPASS calculations to determine how often the algorithm finds the correct structure. In each test, FPASS was supplied with the lattice parameters, symmetry group, powder diffraction pattern from the PDF, and the number of atoms of each type in the unit cell. In all test cases, we found that the lowest energy candidate structure for that compound agreed with the known structure. As shown in Figure 4.5a, we found that FPASS returns the correct structure at least 90% of the time in 67 (71%) of the tests. In several of these test cases, the high success rate of FPASS is not surprising because many of the solutions had less than 5 possible structures that match the known space group and number of atoms. Even so, FPASS was still able to determine the correct structure of the face-centered-orthorhombic, 72-atom unit cell of GeS_2 8 out of 10 times.

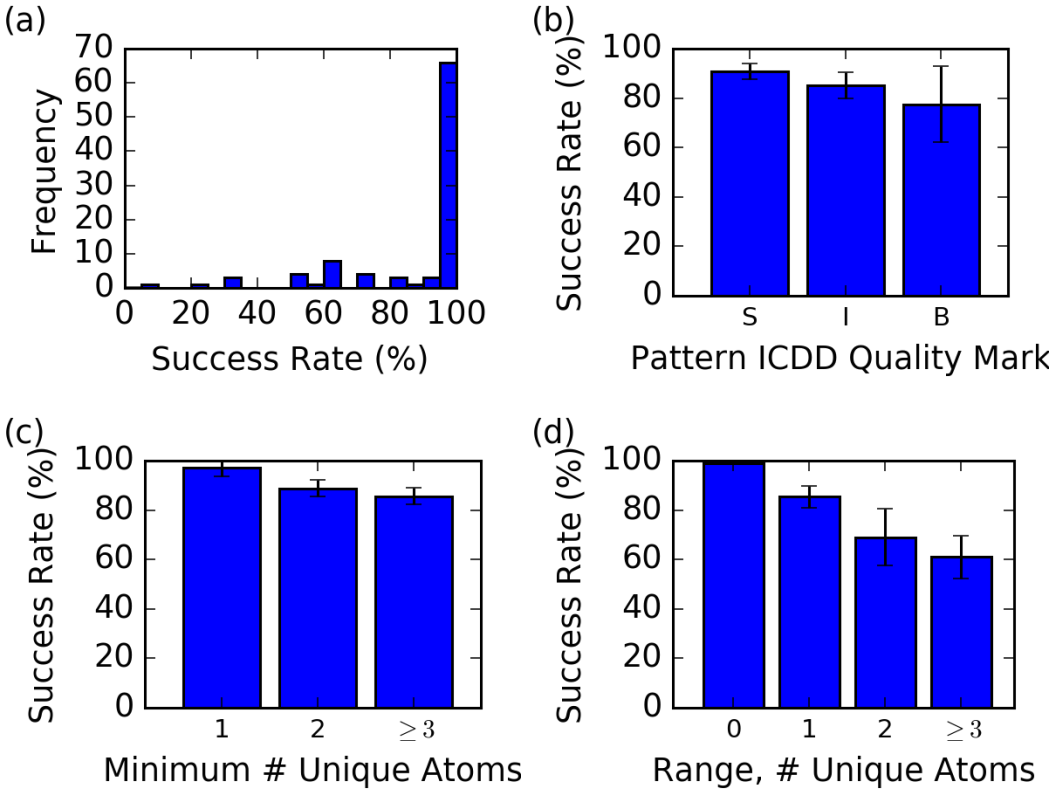


Figure 4.5. (a) Histogram of in how often FPASS determines the correct structure for all 95 test cases. (b-d) Variation in how often FPASS determines the correct structure as a function of (b) quality of diffraction pattern, (c) minimum possible number of unique atoms, and (d) difference between maximum and minimum number of unique atoms. Generally, the success rate for FPASS is the worst when poor-quality x-ray data is provided and for crystals with large unit cells.

We also studied the characteristics of structures that correspond to especially low and high success rates. As shown in Figure 4.5, the success rate decreases with worse pattern quality, larger minimum possible number of symmetrically-unique atoms, and larger difference in the maximum and minimum number of Wyckoff sites. In fact, the compound that FPASS solved correctly the least often, $MgNi_2$, has a large number of minimum number of unique atoms (4), a large difference between the maximum and minimum number of unique atoms (4), and a poor quality X-ray pattern (a "B" rating from the ICDD). Many of the other structures with low FPASS success rates have similar characteristics. In general, we found that FPASS performs best when

solving structures with high symmetry and small number of atoms in the unit cell and is provided with a high quality X-ray pattern.

4.4.2 Tuning the Algorithm

In order to improve the reliability of FPASS for difficult-to-solve crystal structures, we adjusted parameters of the genetic algorithm and evaluated changes in success rate for 10 of the more difficult test cases. For this process, we selected a set of problems we determined to be more difficult based on the metrics studied in the previous section: α-Mn, β-Quartz, Mg_2Ni, CdI_2 GeS_2, PdS, CuS, NiS, α-Np, and SiU_3. FPASS had high rates of determining the correct structure for four of these structures (α-Mn, GeS_2, PdS, NiS), even though they were expected to be difficult based on the metrics established in the previous section. For the other 6 test cases, FPASS had low success rates. By selecting cases with a variety of success rates for FPASS, we can ensure that changes in the parameters that improve the performance on difficult cases do not negatively affect other cases.

We first tuned factors that do not directly affect the computation time: the Wyckoff-site biasing factor and mutation probabilities. As described in the Methods section, these parameters correspond to how much we bias our initial population and how new generations are created. Originally, we used a Wyckoff site and atomic position mutation probabilities of 0.5 and a biasing parameter of 0.5. In order to tune these parameters, we first adjusted the mutation probabilities and biasing parameter and found that the biasing parameter had the strongest impact on success rate. When then held all mutation parameters fixed, and found that the performance of FPASS was the best with the biasing turned off (i.e., a parameter of 0).

Table 4.1. Adjustable parameters for the FPASS algorithm, and their recommended values.

Name	Description	Recommendation
gaoptPopSize	Size of population	10
gaoptConverge	Number of generations after which if no better structure is found, the optimization is terminated	7
WyckoffBias	The biasing factor used when selecting new combinations in new Wyckoff site combinations. Larger values of this parameter bias selection towards fewer Wyckoff sites	0
gaoptWyckMutProb	Probability that a new structure will be mutated by selecting new Wyckoff sites	50%
gaoptPosMutProb	Probability that a new structure will be mutated by perturbing atomic positions	50%
gaoptNumToKeep	Number of top entries to retain in new generation	1

We then held the bias factor fixed at 0, and repeated the tuning process in order to adjust the mutation probabilities and found that the original selections of 0.5 for each type of mutation were optimal. By adjusting mutation probabilities and biasing, we were only able to increase the average success rate of FPASS slightly from 64% of these difficult cases to 68%.

Once we finished tuning the mutation probabilities and biasing factor, we iteratively increased the population size. As the population size directly controls the calculation time, we also considered computational efficiency when adjusting this parameter. We found that by increasing the population size from 10 to 20 we could increase the success rate to 82%. Achieving a one part in a million chance of not finding the correct structure in at least one calculation would require 12 calculations for a success rate of 68% and only 8 for an 82% success rate. However, this increase in population size increases the total time for each calculation at a faster rate. Considering that the average FPASS calculation for a population size of 10 evaluated only 119 structures before converging and the average for a generation size of

20 was 219, the population size of 20 would require evaluating more structures and, thereby, more resources to achieve a certain likelihood of finding the correct structure. For that reason, we recommend a population size of 10 and running large numbers of FPASS calculations for more-difficult solutions. A summary of our recommended values for each parameter are shown in Table 4.1.

4.5 Automated Solution of Crystal Structures

After validating and tuning FPASS, we employed it to solve the structures of several entries that lacked crystal structures in both the Powder Diffraction File and OQMD. To minimize the amount of human effort required to perform each solution, we created a software package designed to automate starting FPASS calculations, checking output for errors, and performing several validation checks. This automation software, named “fpassmgr,” is available in under an open-source license. Using fpassmgr, we run FPASS at least 10 times and until at least 5 FPASS results are identical, and then compute the stability and equilibrium volume for each candidate solution using qmpy.[18] Once the calculation is complete, the code generates a webpage summary that includes structure files for the proposed solution as well as all of the validation results. For most cases, the only human interaction required is starting the calculation and reviewing the validation summary in order to decide whether FPASS has found a correct solution.

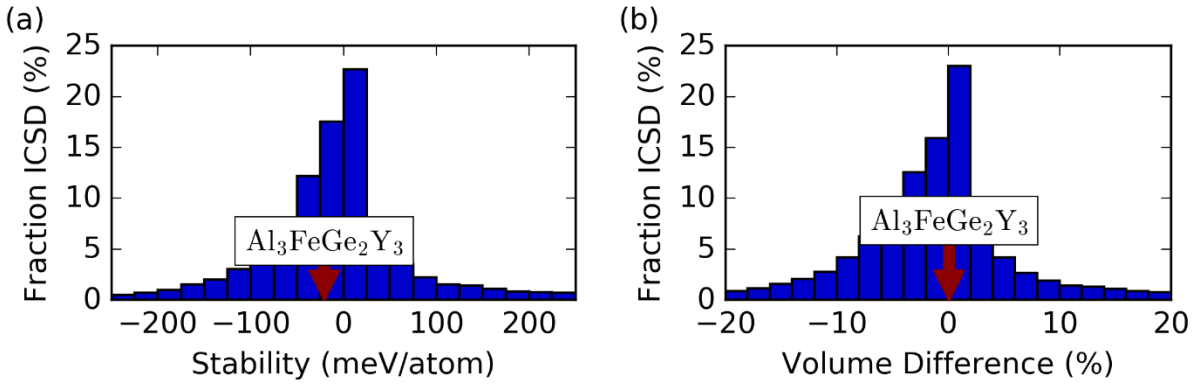


Figure 4.6. Distributions of (a) DFT-computed stability and (b) fractional difference between the measured and DFT-computed volume of all compounds from the Inorganic Crystal Structure Database (ICSD) in the OQMD. Stability was measured as the difference between the computed formation enthalpy of a compound and the minimum-energy combination of all other phases in the OQMD at the same composition. The red arrows indicated the measured stability and fractional difference in volume for our proposed solution for the structure of $Al_3FeGe_2Y_3$, which lies well within the observed range of these two characteristics.

4.5.1 Validation Strategy

Our validation strategy is based on several different tests: (1) agreement between several independent FPASS calculations, (2) match to experimental diffraction pattern, (3) energetic stability, and (4) difference between the experimentally-measured and DFT-predicted volume. First, we run FPASS at least 10 times and conclude a solution has been found when at least 4 other calculations agree with the lowest-energy structure. Once we reach this level of agreement, we select the lowest energy solution as the putative structure, refine the structure to best match the experimental pattern, and measure the R factor. As our Rietveld refinement code is relatively simple, we assume a R of approximately 0.3 is a satisfactory match (typically, this is value can be refined to below 0.05).

We also validate the structure by assessing the DFT-computed formation enthalpy and equilibrium volume of the candidate structure. To assess energetic feasibility, we compute the

Table 4.2. Compositions, symmetry groups, and DFT-predicted properties of structures solved using our automated implementation of FPASS. The stability is the difference between the with respect to the convex hull of the formation enthalpies in the OQMD.[18,32] Negative stability indicates that a compound is stable against decomposition into other structures.

Composition	**Space Group**	**Volume (Å^3/atom)**			ΔH_f **(eV/atom)**	**Stability (meV/atom)**	E_g **(eV)**
		Experiment	DFT	δ (%)			
Al_3CePt	$I4mm$	19.84	20.44	3.0	-0.840	-32	0
Pb_2ZnTeO_6	$Fm\bar{3}m$	12.78	12.85	0.5	-1.375	70	1.3
Sr_2TaZnO_6	$Fm\bar{3}m$	13.36	12.79	-4.2	-2.721	58	0
CaCoSO	$P6_3mc$	17.02	16.99	-0.2	-1.872	46	1.1
$Al_3FeGe_2Y_3$	$P\bar{6}2c$	19.46	19.47	0.1	-0.686	-20	0
Ba_2CdTeO_6	$Fm\bar{3}m$	14.61	14.73	0.8	-2.159	-100	1.2
KFe_2Se_2	$I4/mmm$	21.44	20.80	-3.0	-0.504	31	0
Mo_2NaTmO_8	$I\bar{4}$	12.55	12.24	-2.5	-2.663	-55	3.1
$LiSbO_3$	$C2/m$	11.55	11.47	-0.7	-1.954	1	2.9
$Tb_2O_2CN_2$	$P\bar{3}m1$	14.25	14.03	-1.5	-2.175	-206	4.1

stability with respect to decomposition to all other phases in the OQMD.[141] We then compare this stability to that of all structures from the Inorganic Crystal Structure Database in the OQMD, and determine whether it is within similar ranges – as shown in Figure 4.6a. While the low stability value is no guarantee that our solution is the ground state, it does verify that it is energetically feasible. Once we have computed the formation enthalpy, we also compute the change in volume during relaxation and compare the fractional change to the fractional changes of all compounds in the ICSD (see Figure 4.6b).

4.5.2 Solved Structures

The following subsections are descriptions of the 10 structures we were able to solve using FPASS. The complete structures of some of this compounds had been solved previously but were not present in the OQMD, which made them ideal candidates for further testing our solution and structure validation strategy. For other compounds, like $NaTmMo_2O_8$ and $LiSbO_3$, the unit cell and symmetry of the structure have been determined, but the atomic positions were not known. In all cases, the solution of the structures with FPASS enabled adding these materials to the OQMD and making it more complete. As shown in the summary table, Table 4.2, and Figure 4.7, these sample a broad variety of types of compounds: two of these materials are poly-anionic compounds (CaCoSO and $Tb_2O_2CN_2$), one is a lithium-containing oxide ($LiSbO_3$), and two are intermetallics ($CeAl_3Pt$ and $Al_3FeGe_2Y_3$). Once we solved the structures and added these compounds to the OQMD, we found that several of the materials have band gap energies in the desired range for photovoltaics or thermoelectrics.

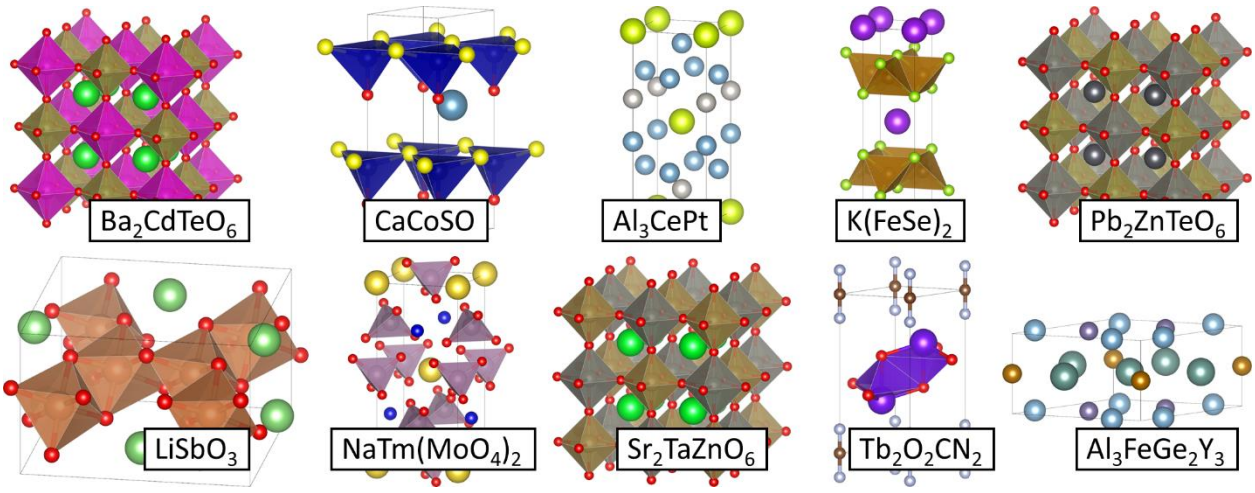


Figure 4.7. Crystal structures determined in this work using an automated implementation of the FPASS algorithm.

4.5.2.1 $Al_3FeGe_2Y_3$

We were unable to find a crystal structure for $Al_3FeGe_2Y_3$ or even reports of its synthesis in the literature. As a starting point for our solution, we used the diffraction pattern from the PDF, along with the already-known composition, space group, and lattice parameters. Using FPASS, we found this compound to be isostructural to the chemically-similar compound, $Al_3NiGe_2Y_3$.[142] We repeated the FPASS solution 10 times and found that each repetition returned the same structure. As shown in Figure 4.6, the stability (measured with respect to all other competing phases) is negative, which indicates that it is thermodynamically stable at T = 0 K. The fractional difference between the experimentally-determined and DFT-predicted volume of this structure is also well within the distribution of other structures from the ICSD, which also indicates the structure is reasonable. Finally, as shown in Figure 4.8, the agreement between the computed and measured powder diffraction pattern is qualitatively excellent. Overall, these validation tests suggest our structure is likely the correct solution for $Al_3FeGe_2Y_3$.

4.5.2.2 Al_3CePt

Al_3CePt was originally synthesized in 1994, is known to have the $BaNiSn_3$-type crystal structure,[143,144] and our FPASS calculation also finds this structure. While this structure has been solved before, the solution was not present in the Powder Diffraction File when we performed FPASS

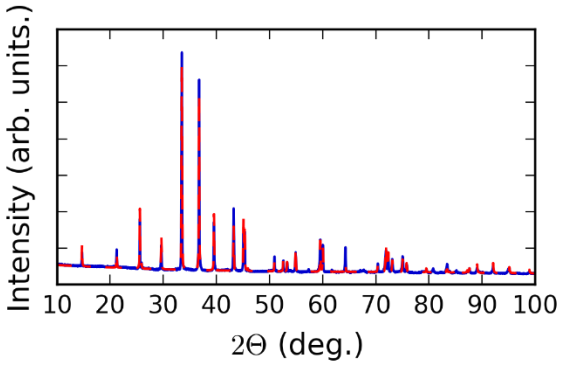


Figure 4.8. Calculated (red, dashed line) and measured (blue, solid line) powder diffraction patterns of our proposed solution for the structure $Al_3NiGe_2Y_3$, as calculated using the Materials Interface (Mint).

and is currently listed only under the $ThCr_2Si_2$ structure type in the ICSD. Both $ThCr_2Si_2$ and $BaNiSn_3$ are based on the $BaAl_4$ structure (with Ce in the body center position), but differ in the fact that $ThCr_2Si_2$ is centrosymmetric and $BaNiSn_3$ is not.[144] As in the solution for $Al_3FeGe_2Y_3$, we found the same structure in all 10 FPASS calculations, the structure is stable in DFT, and that the DFT and experimental volumes of the structure are in agreement. Also considering the acceptable match to the X-ray diffraction pattern, we conclude our solution is correct. While our work is not the first solution for this structure, our solution does match the literature and we were still successful in adding a missing crystal structure to the OQMD.

4.5.2.3 Pb_2ZnTeO_6, Ba_2CdTeO_6, and Sr_2TaZnO_6

We found Pb_2ZnTeO_6, Ba_2CdTeO_6, and Sr_2TaZnO_6 to all have the double perovskite structure. Of these, Pb_2ZnTeO_6 and Ba_2CdTeO_6 are both known to be double perovskites,[145,146] but were not present in the ICSD and OQMD. As far as we could tell, Sr_2TaZnO_6 has not yet been reported in the literature and is not present in the ICSD or Crystallography Open Database (COD).[67,68] For all three cases, FPASS found the same structure 10 out of 10 calculations, and all of our other validation checks indicate these structures are reasonable. Each structure was either stable or slightly metastable (<75 meV/atom) and the volume difference between the experimental were all within bounds (<5%) – as shown in Table 4.2. According to our FPASS calculations, there are only two likely structures given the $A_2B'B''O_3$ stoichiometry, the known symmetry, and the number of atoms in the unit cell: a structure where each A atom is coordinated with 12 O atoms, and one where A is coordinated with six. As the only difference between these two structures is the positions of

O atoms, these two structures have similar diffraction patterns when O scatters X-rays weakly compared to other atoms in the structure. However, the energy difference between the structures can be quite large (> 1 eV/atom for Pb_2ZnTeO_6), which makes the solution unambiguous with FPASS.

4.5.2.4 CaCoSO

According to its entry in the Powder Diffraction File, CaCoSO is a hexagonal structure with a space group of $P6_3mc$. Using this information and the unit cell parameters available in the PDF, we found the structure to be similar to that of CaClOH, with Co occupying the tetrahedrally-coordinated site.[147] This finding is consistent with the structure solution of Salter *et al*. in 2016.[148] We found this structure in 9 out of 10 FPASS calculations, and were able to confirm that it is only slightly metastable (46 meV/atom). Therefore, we agree with the structure proposed by Salter *et al*.

4.5.2.5 KFe_2Se_2

KFe_2Se_2 is the nominal composition of a superconducting compound discovered in 2010, and is known to have the $ThCr_2Si_2$ structure.[149] At the time we performed our FPASS calculation, this structure was not available in the OQMD but already established in the literature.[149] We did confirm the $ThCr_2Si_2$ structure with our FPASS calculation, and found it to pass all validation checks. As in the solution of Al_3CePt, we did not solve this structure for the first time but were able to improve the OQMD by adding this structure to our database.

4.5.2.6 $NaTmMo_2O_8$

We found the first report of Mo_2NaTmO_8 to be in a 1964 paper by Ayala *et al*.[150] The authors reported the unit cell parameters of the structure and that Mo_2NaTmO_8 was based on the scheelite ($CaWO_4$) structure, but the not the atomic positions.[151] Using FPASS, we confirmed the structure of Mo_2NaTmO_8 is based on the scheelite structure and equivalent to that of Li_2CaHfF_8.[150] FPASS found this structure in 10 out of 10 calculations. Our validation DFT calculations found a similar volume to that observed in experiment and that this structure is stable. Considering that the diffraction pattern match is also acceptable, we conclude we have determined the correct structure.

4.5.2.7 $LiSbO_3$

The monoclinic phase for $LiSbO_3$ was discovered by Nalbandyan *et al*. in 2006.[152] While the authors were unable to determine its crystal structure, they hypothesized $LiSbO_3$ is a distorted rock salt structure.[152] Using the known unit cell parameters and symmetry group ($C2/m$) proposed by Nalbandyan *et al*. as input to FPASS, we found a layered structure similar to Li_2MnO_3 – a distorted rocksalt structure consistent with the hypothesis of Nalbandyan *et al*.[153] The Sb atoms form a 2D network of face sharing octahedra, with the Li atoms occupying tetrahedral sites in the space between these layers (as opposed to octahedral sites in Li_2MnO_3). This structure is nearly degenerate with the known, orthorhombic phase of $LiSbO_3$,[154] being only 1 meV/atom higher in energy than the orthorhombic structure according to our DFT calculations. We ran FPASS thirty times and found the layered structure in 8 of the calculations. The other, higher energy solutions, include a version of this structure

where the Li is in octahedral sites (as in Li_2MnO_3), which is slightly higher in energy (42 meV/atom). Also considering the acceptable match with the experimental diffraction pattern and small difference between experimental and DFT-predicted volumes, we conclude our solution to $LiSbO_3$ – a layered structure with Li on the tetrahedral sites – is correct.

4.5.2.8 $Tb_2CN_2O_2$

According to the Powder Diffraction File, $Tb_2CN_2O_2$ has a hexagonal unit cell with $P\bar{3}m1$ symmetry. Starting with this information and the lattice parameters listed in the PDF, we found that this compound shares the same crystal structure as 11 other lanthanoid dioxymonocyanamides.[155,156] All 10 FPASS calculations we performed found this structure, and we also found it to be stable in the OQMD. The volume change on relaxation and match to diffraction pattern are also reasonable, which lead us to conclude that this is the correct structure for $Tb_2CN_2O_2$.

4.6 Current Limitations of FPASS

Ten of the 20 solutions we attempted with FPASS failed at least one of the validation tests, which could each be a result of several factors. First of all, our technique is based on the assumption that the hypothesized lattice parameters and space group are correct. If any of these are inaccurate, our algorithm may converge to an incorrect solution. Also, FPASS may have failed to find the correct solution within the search space – though this is unlikely if the algorithm returned the same structure from multiple runs. Furthermore, our implementation of FPASS currently only supports perfectly-ordered materials. If the true solution is disordered (e.g., mixing between two types of elements on a single site), FPASS will fail to find the correct

structure. Overall, these failures highlight potentials avenues for improving FPASS (e.g., accounting for disorder) and the necessity of automated validation tests.

4.7 Conclusion

In this work, we described an implementation of the FPASS algorithm capable of being used to automatically solve incompletely-determined crystal structures. We validated this algorithm by determining the structures of over 90 known crystal structures, and found that FPASS identified the correct structure in each case. Once validated, we tuned the algorithm to increase its reliability for crystal structures that are difficult for the algorithm to solve and then applied it to solve of dozens of yet-undetermined structures from the Powder Diffraction File. To date, we have solved the structures of 10 compounds and added these structures to the OQMD; thereby increasing the completeness of this database.

5 General-Purpose, Composition-Based Representations of Materials

5.1 Abstract

A very active area of materials research is to devise methods that use machine learning to automatically extract predictive models from existing materials data. While prior examples have demonstrated successful models for some applications, many more applications exist where machine learning can make a strong impact. To enable faster development of machine-learning-based models for such applications, we have created a framework capable of being applied to a broad range of materials data. Our method works by using a chemically diverse list of attributes, which we demonstrate are suitable for describing a wide variety of properties, and a novel method for partitioning the data set into groups of similar materials in order to boost the predictive accuracy. In this chapter, we demonstrate how this new method can be used to predict diverse properties of crystalline and amorphous materials, such as band gap energy and glass-forming ability.

5.2 Introduction

Rational design of materials is the ultimate goal of modern materials science and engineering. As part of achieving that goal, there has been a large effort in the materials science community to compile extensive datasets of materials properties in order to provide scientists and engineers with ready access to the properties of known materials. Today, there are databases of crystal structures[68], superconducting critical temperatures,[157] physical properties of crystalline compounds,[18,32,37,38] and many other repositories containing useful materials data. Recently, it has been shown that these databases can also serve as

resources for creating predictive models and design rules – the key tools of rational materials design.[3,4,86,88,106,158] These databases have grown large enough that the discovery of such design rules and models is impractical to accomplish by relying simply on human intuition and knowledge about material behavior. Rather than relying directly on intuition, machine learning offers the promise of being able to create accurate models quickly and automatically.

To date, materials scientists have used machine learning to build predictive models for a handful of applications.[79,80,85,94,100,159–168] For example, there are now models to predict the melting temperatures of binary inorganic compounds,[163] the formation enthalpy crystalline compounds,[41,79,85] which crystal structure is likely to form at a certain composition,[37,100,115,116,169] band gap energies of certain classes of crystals,[170,171] and the mechanical properties of metal alloys.[94,166] While these models demonstrate the promise of machine learning, they only cover a small fraction of the properties used in materials design and the datasets available for creating such models. For instance, no broadly-applicable, machine-learning-based models exist for the band gap energy or glass forming ability even though large-scale databases of these properties have existed for years.[18,172]

Provided the large differences between the approaches used in the literature, a systematic path forward to creating accurate machine learning models across a variety of new applications is not clear. While techniques in data analytics have advanced significantly, the development of routine methods for transforming raw materials data into the quantitative descriptions required for employing these algorithms has yet to emerge. In contrast, the chemoinformatics community benefits from a rich library of methods for describing molecular structures, which

allow for standard approaches for deciding inputs into the models and, thereby, faster model development.[173–175] What is missing are similar flexible frameworks for building predictive models of material properties.

In this work, we present a general-purpose machine-learning-based framework for predicting the properties of materials based on their composition. In particular, we focus on the development of a set of attributes – which serve as input to the machine learning model – that could be reused for a broad variety of materials problems. Provided a flexible set of inputs, creating a new material property model can be reduced to finding a machine learning algorithm that achieves optimal performance – a well-studied problem in data science. Additionally, we employ a novel partitioning scheme to enhance the accuracy of our predictions by first partitioning data into similar groups of materials and training separate models for each group. We show that this method can be used regardless of whether the materials are amorphous or crystalline, the data is from computational or experimental studies, or the property takes continuous or discrete values. In particular, we demonstrate the versatility of our technique by using it for two distinct applications: predicting novel solar cell materials using a database of DFT-predicted properties of crystalline compounds and using experimental measurements of glass-forming ability to suggest new metallic glass alloys. Our vision is that this framework could be used as a basis for quickly creating models based on the data available in the materials databases and, thereby, initiate a major step forward in rational materials design.

5.3 General Purpose Method to Create Materials Property Models

Machine learning (ML) models for materials properties are constructed from three parts: training data, a set of attributes that describe each material, and a machine learning algorithm to map attributes to properties. For the purposes of creating a general purpose method, we focused entirely on the attributes set because the method needs to be agnostic to the type of training data and because it is possible to utilize already-developed machine learning algorithms. Specifically, our objective is to develop a general set of attributes based on the composition that can be reused for a broad variety of problems.

The goal in designing a set of attributes is to create a quantitative representation that both uniquely defines each material in a dataset and relates to the essential physics and chemistry that influence the property of interest.[79,80] As an example, the volume of a crystalline compound is expected to relate to the volume of the constituent elements. By including the mean volume of the constituent elements as an attribute, a machine learning algorithm could recognize the correlation between this value and the compound volume, and use it to create a predictive model. However, the mean volume of the constituent elements neither uniquely defines a composition nor perfectly describes the volumes of crystalline materials.[176] Consequently, one must include additional attributes to create a suitable set for this problem. Potentially, one could include factors derived from the electronegativity of the compound to reflect the idea that bond distances are shorter in ionic compounds, or the variance in atomic radius to capture the effects of polydisperse packing. The power of machine learning is that it is

not necessary to know which factors actually relate to the property and how before creating a model – those relationships are discovered automatically.

The materials informatics literature is full of successful examples of attribute sets for a variety of properties.[79,85,100,159,163,170,177] We observed that the majority of attribute sets were primarily based on statistics of the properties of constituent elements. As an example, Meredig, Agrawal *et al.* described a material based on the fraction of each element present and various intuitive factors, such as the maximum difference in electronegativity, when building models for the formation energy of ternary compounds.[85] Ghiringhelli *et al*. used combinations of elemental properties such as atomic number and ionization potential to study the differences in energy between zinc-blende and rocksalt phases.[79] We also noticed that the important attributes varied significantly depending on material property. The best attribute for describing the difference in energy between zinc-blende and rocksalt phases was found to be related to the pseudopotential radii, ionization potential, and electron affinity of the constituent elements.[79] In contrast, melting temperature was found to be related to atomic number, atomic mass, and differences between atomic radii.[163] From this we conclude that a general-purpose attribute set should contain the statistics of a wide variety of elemental properties in order to be adaptable.

Building on existing strategies, we created an expansive set of attributes that can be used for materials with any number of constituent elements. As we will demonstrate, this set is broad enough to capture a sufficiently-diverse range of physical/chemical properties in order to be used to create accurate models for many materials problems. In total, we use a set of 145

attributes, which are compared against other attribute sets in the Appendix, that fall into four distinct categories:

1. **Stoichiometric** attributes that depend only on the fractions of elements present and not what those elements actually are. These include the number of elements present in the compound and several L^p norms of the fractions.
2. **Elemental Property Statistics**, which are defined as the mean, mean absolute deviation, range, minimum, maximum, and mode of 22 different elemental properties. This category includes attributes such as the maximum row on periodic table, average atomic number, and the range of atomic radii between all elements present in the material.
3. **Electronic Structure** attributes, which are the average fraction of electrons from the *s*, *p*, *d*, and *f* valence shells between all present elements. These are identical to the attributes used by Meredig, Agrawal *et al*.[85]
4. **Ionic Compound** attributes that include whether it is possible to form an ionic compound assuming all elements are present in a single oxidation state, and two adaptations of the fractional "ionic character" of a compound based on an electronegativity-based measure.[178]

For the third ingredient, the machine learning algorithm, we evaluate many possible methods for each individual problem. Previous studies have used machine learning algorithms including partial least-squares regression,[115,159] Least Absolute Shrinkage and Selection Operator (LASSO),[79,171,179] decision trees,[85,100] kernel ridge regression,[80,160,161,180] Gaussian process regression, [161–163,181] and neural networks.[94,164,165] Each method offers different advantages, such as speed or interpretability, which must be weighed carefully for a new application. We generally approach this problem by evaluating the performance of several algorithms to find one that has both reasonable computational requirements (i.e., can be run on available hardware in a few hours) and has low error rates in cross-validation – a

process that is simplified by the availability of well-documented libraries of machine learning algorithms.[74,138] We often find that ensembles of decision trees (e.g., rotation forests[78]) perform best with our attribute set. These algorithms also have the advantage of being quick to train, but are not easily interpretable by humans. Consequently, they are less suited for understanding the underlying mechanism behind a material property but, owing to their high predictive accuracy, excellent choices for the design of new materials.

We also utilize a partitioning strategy that enables a significant increase in predictive accuracy for our ML models. By grouping the dataset into chemically-similar segments and training a separate model on each subset, we boost the accuracy of our predictions by reducing the breadth of physical effects that each machine learning algorithm needs to capture. For example, the physical effects underlying the stability intermetallic compounds are likely to be different than those for ceramics. In this case, one could partition the data into compounds that contain only metallic elements and another including those that do not. As we demonstrate in the examples below, partitioning the dataset can significantly increase the accuracy of predicted properties. Beyond using our knowledge about the physics behind a certain problem to select a partitioning strategy, we have also explored using an automated, unsupervised-learning-based strategy for determining distinct clusters of materials.[83] Currently, we simply determine the partitioning strategy for each property model by searching through a large number of possible strategies and selecting the one that minimizes the error rate in cross-validation tests.

5.3.1 Justification for Large Attribute Set

The main goal of our technique is to accelerate the creation of machine learning models by reducing or eliminating the need to develop a set of attributes for a particular problem. Our approach was to create a large attribute set, with the idea that it would contain a diverse enough library of descriptive factors such it is likely to contain several that are well-suited for a new problem. To justify this approach, we evaluated changes in the performance of attributes for different properties and types of materials using data from the Open Quantum Materials Database (OQMD). As described in greater detail in the next section, the OQMD contains the DFT-predicted formation energy, band gap energy, and volume of hundreds of thousands of crystalline compounds. The diversity and scale of the data in the OQMD make it ideal for studying changes in attribute performance using a single, uniform dataset.

We found that the attributes which model a material property best can vary significantly depending on the property and type of materials in the dataset. To quantify the predictive ability of each attribute, we fit a quadratic polynomial using the attribute and measured the root mean squared error of the model. We found the attributes that best describe the formation energy of crystalline compounds are based on the electronegativity of the constituent elements (e.g., maximum and mode electronegativity). In contrast, the best-performing attributes for band gap energy are the fraction of electrons in the *p* shell and the mean row in the periodic table of the constituent elements. Additionally, the attributes that best describe the formation energy vary depending on the type of compounds. The formation energy of intermetallic compounds is best described by the variance in the melting temperature

and number of *d* electrons between constituent elements, whereas compounds that contain at least one nonmetal are best modelled by the mean ionic character (a quantity based on electronegativity difference between constituent elements). Taken together, the changes in which attributes are the most important between these examples further supports the necessity of having a large variety of attributes available in a general-purpose attribute set.

It is worth noting that the 145 attributes described in this paper should not be considered the complete set for inorganic materials. The chemical informatics community has developed thousands of attributes for predicting the properties of molecules.[173] What we present here is a step towards creating such a rich library of attributes for inorganic materials. While we do show in the examples considered in this work that this set of attributes is sufficient to create accurate models for two distinct properties, we expect that further research in materials informatics will add to the set presented here and be used to create models with even greater accuracy. Furthermore, many materials cannot be described simply by average composition. In these cases, we propose that the attribute set presented here can be extended with representations designed to capture additional features such as structure (ex: Coulomb Matrix[80] for atomic-scale structure) or processing history. We envision that it will be possible to construct a library of general-purpose representations designed to capture structure and other characteristics of a material, which would drastically simplify the development of new machine learning models.

5.4 Example Applications

In the following sections, we detail two distinct applications for our novel material property prediction technique in order to demonstrate its versatility: predicting three physically-distinct properties of crystalline compounds and identifying potential metallic glass alloys. In both cases, we use the same general framework, i.e., the same attributes and partitioning-based approach. In each case, we only needed to identify the most-accurate machine learning algorithm and find an appropriate partitioning strategy. Through these examples, we discuss all aspects of creating machine-learning based models to design a new material: assembling a training set to train the models, selecting a suitable algorithm, evaluating model accuracy, and employing the model to predict new materials.

5.4.1 Accurate Models for Properties of Crystalline Compounds

Density Functional Theory (DFT) is a ubiquitous tool for predicting the properties of crystalline compounds, but is fundamentally limited by the amount of computational time that DFT calculations require. In the past decade, DFT has been used to generate several databases containing the T = 0 K energies and electronic properties of ~10^5 crystalline compounds,[18,32,37,38,182] which each required millions of hours of CPU time to construct. While these databases are indisputably-useful tools, as evidenced by the many materials they have been used to design,[20,27,29,32–35] machine-learning-based methods offer the promise of predictions at several orders of magnitude faster rates. In this example, we explore the use of data from the DFT calculation databases as training data for machine learning models that

Table 5.1. Comparison of the ability of several machine learning algorithms to predict properties of materials from the OQMD. Data represents the mean absolute error in a 10-fold cross-validation test of a single model trained on the properties predicted using DFT of 228,676 crystalline compounds.

		Machine Learning Algorithm			
		Linear Regression	Reduced-Error Pruning Tree (REPTree)	Rotation Forest[78] + REPTree	Random Subspace[183] + REPTree
Property	Volume (Å^3/atom)	1.22	0.816	0.593	0.563
	Formation Energy (eV/atom)	0.259	0.126	0.0973	0.0882
	Band gap Energy (eV)	0.202	0.0701	0.0643	0.0645

can be used rapidly assess many more materials than what would be feasible to evaluate using DFT.

Training Data: We used data from the Open Quantum Materials Database (OQMD), which contains the properties of around 300,000 crystalline compounds as calculated using DFT.[18,32] We selected a subset of 228,676 compounds from OQMD that represent the lowest-energy compound at each unique composition to use as a training set. As a demonstration of the utility of our method, we developed models to predict the three physically-distinct properties currently available through the OQMD: band gap energy, specific volume, and formation energy.

Method: To select an appropriate machine learning algorithm for this example, we evaluated the predictive ability of several algorithms using 10-fold cross-validation. This technique randomly splits the dataset into 10 parts, and then trains a model on 9 partitions and attempts to predict the properties of the remaining set. This process is repeated using each of

the 10 partitions as the test set, and the predictive ability of the model is assessed as the average performance of the model across all repetitions. As shown in Table 5.1, we found that creating an ensemble of reduced-error pruning decision trees using the random subspace technique had the lowest mean absolute error in cross-validation for these properties among the 10 ML algorithms we tested (of which, only 4 are listed for clarity).[183] Models produced using this machine learning algorithm had the lowest mean absolute error in cross validation, and had excellent correlation coefficients of above 0.91 between the measured and predicted values for all three properties.

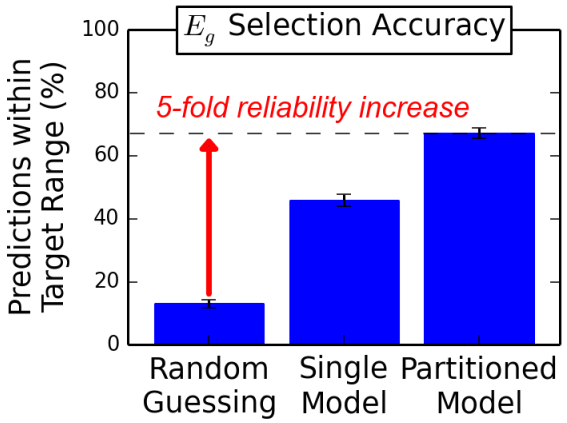


Figure 5.1. Performance of three different strategies to locate compounds with a band gap energy within a desired range: randomly-selecting nonmetal-containing compounds, and two strategies using the machine-learning-based method presented in this work. The first machine learning strategy used a single model trained on the computed band gap energies of 22667 compounds from the ICSD. The second method a model created by first partitioning the data into groups of similar materials, and training a separate model on each subset. The number of materials that were actually found to have a band gap within the desired range after 30 guesses was over 5 times larger when using our machine learning approach than when randomly selecting compounds. Error bars represent the 95% confidence interval.

As a simple test for how well our band gap model can be used for discovering new materials, we simulated a search for compounds with a band gap within a desired range. To evaluate our the ability of our method to locate compounds that have band gap energies within the target range, we devised a test where a model was trained on 90% of the dataset and then

was tasked with selecting which 30 compounds in the remaining 10% were most likely to have a band gap energy in the desired range for solar cells: 0.9 – 1.7 eV.[184] For this test, we selected a subset of the OQMD that only includes compounds that have been reported to be possible to be made experimentally in the ICSD (a total of 25085 entries) so that only band gap energy, and not stability, needed to be considered.

For this test, we compared three selection strategies for finding compounds with desirable band gap energies: randomly selecting nonmetal-containing compounds (i.e., without machine learning), using a single model trained on the entire training set to guide selection, and a model created using the partitioning approach introduced in this manuscript. As shown in Figure 5.1, randomly selecting a nonmetal-containing compound would result in just over 12% of the 30 selected compounds to be within the desired range of band gap energies. Using a single model trained on the entire dataset, this figure dramatically improves to approximately 46% of selected compounds having the desired property. We found the predictive ability of our model can be increased to around 67% of predictions actually having the desired band gap energy by partitioning the dataset into groups of similar compounds before training. Out of the 20 partitioning strategies we tested, we found the best composite model works by first partitioning the dataset using a separate model trained to predict the expected range, but not the actual value, of the band gap energy (e.g., compounds predicted to have a band gap between 0 and 1.5 eV are grouped together), and then on whether a compound contains a halogen, chalcogen, or pnictogen (as shown in Figure 5.2). By partitioning the data into smaller subsets, each of the individual machine learning models only evaluates compounds with similar

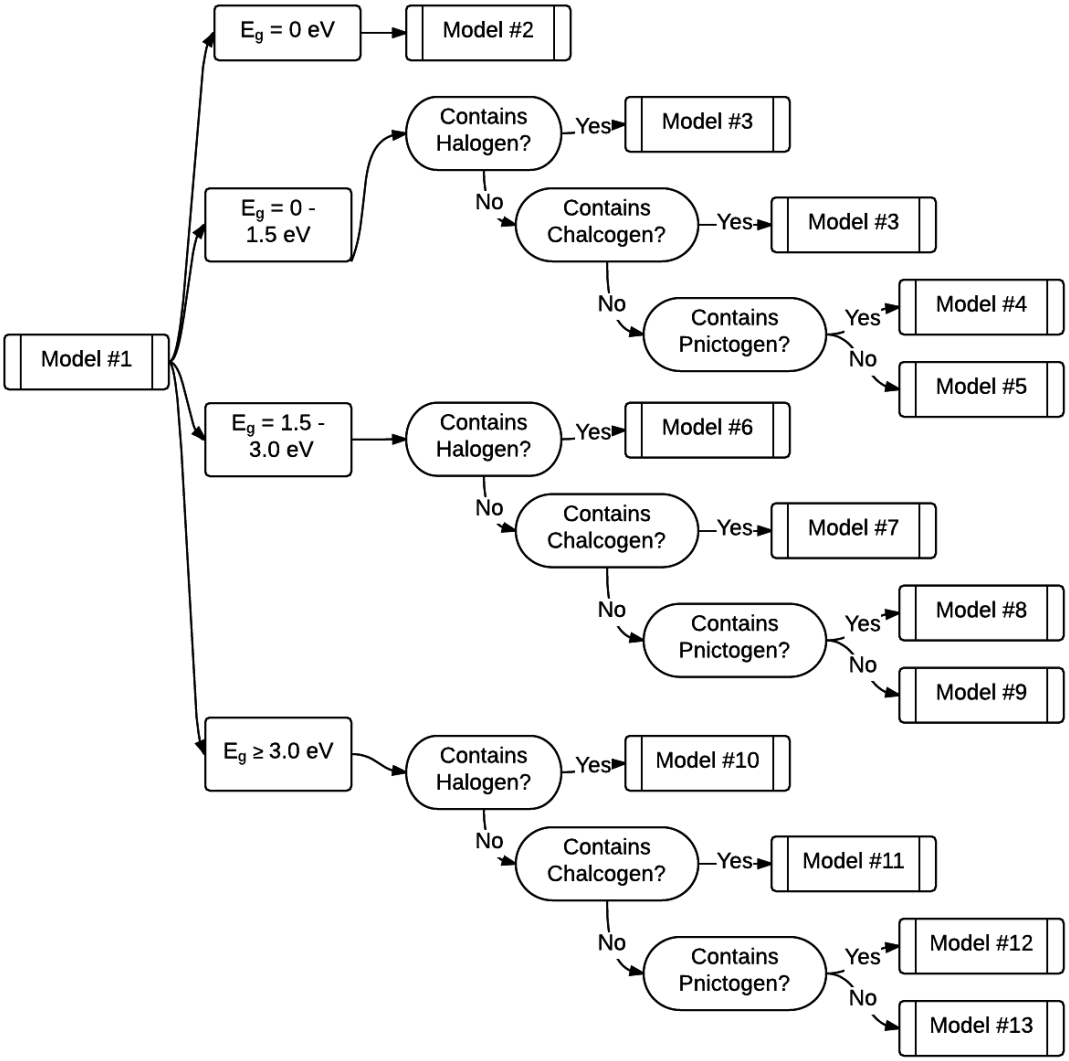


Figure 5.2. Hierarchical model used to predict band gap energies of crystalline compounds. Each rectangle with rounded corners represents a machine learning model. The model on the far left (Model #1) is trained to predict the mostly-likely range for the band gap of a compound. The models on the right are trained to predict actual value of the band gap energy. Depending on results of Model #1 and the composition of an entry, a different machine model would be used. For example, Model #3 will be used for all halogen-containing compounds predicted to have a band gap energy between 0 and 1.5 eV by Model #1.

chemistries (e.g. halogen-containing compounds with a band gap expected to be between 0

and 1.5 eV), which we found enhances the overall accuracy of our model.

Table 5.2. Compositions and predicted band gap energies of materials predicted using machine learning to be candidates for solar cell applications. Compositions represent the nominal compositions of novel ternary compounds predicted by using methods developed in Ref. [85]. Band gap energies were predicted using a machine learning model trained on DFT band gap energies from the OQMD[18] using methods described in this work.

Composition	E_g (eV)
$ScHg_4Cl_7$	1.26
$V_2Hg_3Cl_7$	1.16
Mn_6CCl_8	1.28
$Hf_4S_{11}Cl_2$	1.11
VCu_5Cl_9	1.19

Once we established the reliability of our model, we used it to search for new compounds (i.e., those not yet in the OQMD) with a band gap energy within the desired range for solar cells: 0.9 – 1.7 eV. To gain the greatest predictive accuracy, we trained our band gap model on the entire OQMD dataset. Then, we used this model to predict the band gap energy of compositions that were predicted by Meredig, Agrawal *et al.*[85] to be as-yet-undiscovered ternary compounds. Out of this list of 4500 predicted compounds, we found that 223 are likely to have favorable band gap energies. A subset with the best stability criterion (as reported in Ref. [85]) and band gap energy closest to 1.3 eV are shown in Table 5.2. As demonstrated in this example and by recent work from Sparks *et al.*,[93] having access to several machine learning models for different properties can make it possible to rapidly screen materials based on many design criteria. Provided the wide range of applicability of the machine learning technique demonstrated in this work and the growing availability of material property data, it may soon be possible to screen for materials based on even more properties than those considered here using models constructed based on several different datasets.

5.4.2 Locating Novel Metallic Glass Alloys

Metallic glasses possess a wide range of unique properties, such as high wear resistance and soft magnetic behavior, but are only possible to create at special compositions that are difficult to determine *a priori*.[185] The metallic glass community commonly relies on empirical rules (e.g., systems that contain many elements of different sizes are more likely to form glasses[186]) and extensive experimentation in order to locate these special compositions.[187] While searches based on empirical rules have certainly been successful (as evidenced by the large variety of known alloys[188]), this conventional method is known to be slow and resource-intensive.[187] Here, we show how machine learning could be used to accelerate the discovery of new alloys by using known experimental datasets to construct predictive models of glass forming ability.

Data: We used experimental measurements taken from "Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys," a volume of the Landolt-Börnstein collection.[172] This dataset contains measurements of whether it is possible to form a glass using a variety of experimental techniques at thousands of compositions from hundreds of ternary phase diagrams. For our purposes, we selected 5369 unique compositions where the ability to form an amorphous ribbon was assessed using melt spinning. In the event that multiple measurements for glass forming ability were taken at a single composition, we assume that it is possible to form a metallic glass if at least one measurement found it was possible to form a completely-amorphous sample. After the described screening steps, 70.8% of the entries in the training dataset correspond to metallic glasses.

Method: We used the same set of 145 attributes as in the band gap example and ensembles of Random Forest classifiers[77] created using the random subspace technique as the machine learning algorithm, which we found to be the most accurate algorithm for this problem. This model classifies the data into two categories (i.e., can and cannot form a metallic glass) and computes the relative likelihood that a new entry would be part of each category. For the purposes of validating the model, we assume any composition predicted to have a greater than 50% probability of glass formation to be a positive prediction of glass forming ability. Using a single model trained on the entire dataset, we were able to create a model with 90% accuracy in 10-fold cross-validation.

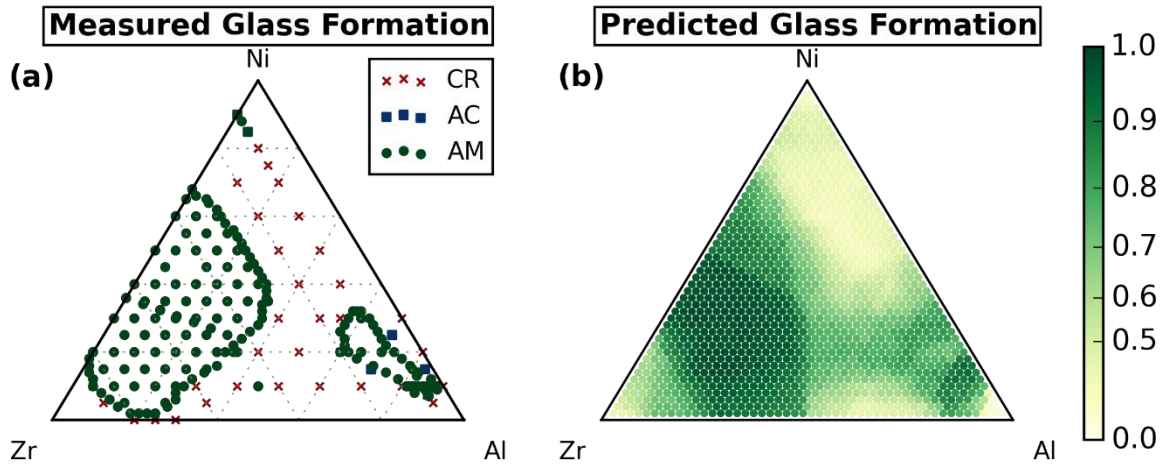


Figure 5.3. (a) Experimental measurements of metallic glass forming ability in the Al-Ni-Zr ternary, as reported in Ref. [172]. Green circles (AM) mark compositions at which it is possible to create a fully-amorphous ribbon via melt spinning, blue squares (AC) mark compositions at which only a partially-amorphous ribbon can be formed, and red squares (CR) mark compositions where it is not possible to form any appreciable amount of amorphous phase. (b) Predicted glass forming ability from our machine learning model. Points are colored based on relative likelihood of glass formation, where 1 is the mostly likely and 0 is the least. The model used to make these predictions was developed using the methods outlined in this work, and was not trained on any measurements from the Al-Ni-Zr ternary or any of its constituent binaries.

As a test of the ability of our method to predict new alloys, we removed all entries that contained exclusively Al, Ni, and Zr (i.e., all Al-Ni-Zr ternary compounds, and any binary formed by any two of those elements) from our training dataset and then predicted the probability of an alloy being able to be formed into the amorphous state for the Al-Ni-Zr ternary system. As shown in Figure 5.3a, it is possible to form amorphous ribbons with melt spinning in one region along the Ni-Zr binary and in a second, Al-rich ternary region. Our model is able to accurately predict both the existence of these regions and their relative locations (see Figure 5.3b), which shows that models created using our method could serve to accurately locate favorable compositions in yet-unassessed alloy systems.

We further validated the ability of our models to extrapolate to alloy systems not included in the training set by iteratively using each binary system as a test set. This procedure works by excluding all alloys that contain both of the elements in the binary, training a model on the remaining entries, and then predicting the glass-forming ability of the alloys that were removed. For example, if the Al-Ni binary were being used as a test set, then $Al_{50}Ni_{50}$ and $Al_{50}Ni_{25}Fe_{25}$ would be removed but $Al_{50}Fe_{50}$ and $Al_{50}Fe_{25}Zr_{25}$ would not. This process is then repeated for all 380 unique binaries in the dataset. We measured that our model has an 80.2% classification accuracy over 15318 test entries where 71% of entries were measured to be glasses – in contrast to the 90.1% measured in 10-fold cross-validation with a similar fraction of glasses in the test set. We also found that by training separate models for alloys that contain only metallic elements and those that contain a nonmetal/metalloid it is possible to slightly increase the prediction accuracy to 80.7% - a much smaller gain than that observed in the band

gap example (23%). Overall, this exclusion test strongly establishes that our model is able to predict the glass forming ability in alloy systems that are completely unassessed.

In order to search for new candidate metallic glasses, we used our model to predict the probability of glass formation for all possible ternary alloys created at 2 at% spacing by any combination of elements found in the training set. Considering that the dataset included 51 elements, this space includes approximately 24 million candidate alloys, which required approximately 6 hours to evaluate on 8, 2.2 GHz processors. In order to remove known alloys from our prediction results, we first removed all entries where the L_1 distance between the composition vector (i.e., $\langle x_H, x_{He}, x_{Li}, \ldots \rangle$) of the alloy and any amorphous alloy in the training set was less than 30 at%. We then found the alloys with the highest predicted probability of glass formation in each binary and ternary. Eight alloys with the highest probability of glass formation are shown in Table 5.3. One top candidate, $Zr_{0.38}Co_{0.24}Cu_{0.38}$, is particularly promising considering the existence of Zr-lean Zr-Co and Zr-Cu binary alloys and Zr-Al-Co-Cu bulk metallic glasses.[189] To make the ability to find new metallic glasses openly available to the materials science community, we have included all of the software and data necessary to use this model in the Supplementary Information and created an interactive, web-based tool.[190]

Table 5.3. Compositions of candidate metallic glass alloys predicted using a machine learning model trained on experimental measurements of glass forming ability. These alloys were predicted to have the highest probability being able to be formed into an amorphous ribbon via melting spinning out of 24 million candidates.

Alloy Composition	
$Zr_{0.38}Co_{0.24}Cu_{0.38}$	$Hf_{0.7}Si_{0.16}Ni_{0.14}$
$V_{0.16}Ni_{0.64}B_{0.2}$	$Hf_{0.48}Zr_{0.16}Ni_{0.36}$
$Zr_{0.46}Cr_{0.36}Ni_{0.18}$	$Zr_{0.48}Fe_{0.46}Ni_{0.06}$
$Zr_{0.5}Fe_{0.38}W_{0.12}$	$Sm_{0.22}Fe_{0.54}B_{0.24}$

5.5 Conclusions

In this work, we introduced a general-purpose machine learning framework for predicting the properties of a wide variety of materials and demonstrated its broad applicability via illustration of two distinct materials problems: discovering new potential crystalline compounds for photovoltaic applications and identifying candidate metallic glass alloys. Our method works by using machine learning to generate models that predict the properties of a material as a function of a wide variety of attributes designed to approximate chemical effects. The accuracy of our models is further enhanced by partitioning the dataset into groups of similar materials. In this manuscript, we show that this technique is capable of creating accurate models for properties as different as the electronic properties of crystalline compounds and glass formability of metallic alloys. Creating new models with our strategy requires only finding which machine learning algorithm maximizes accuracy and testing different partitioning strategies, which are processes that could be eventually automated.[191] We envision that the versatility of this method will make it useful for a large range of problems, and help enable the quicker deployment and wider-scale use machine learning in the design of new materials.

6 Voronoi-Tessellation-Based Representations for Crystal Structures

6.1 Abstract

While high-throughput Density Functional Theory (DFT) has become a prevalent tool for materials discovery, it is limited by the large computational cost. In this paper, we explore using DFT data from high-throughput calculations to create faster, approximate models with machine learning (ML) that can be used to guide new searches. Our method works by using decision tree models to map length-scale-independent attributes derived from the Voronoi tessellation of a crystal structure to its DFT-predicted formation enthalpy. We found that models created using this method have half the error rate cross-validation and similar computation times to models created with the Coulomb matrix and Pair Radial Distribution Function (PRDF) methods. For a dataset of 435k entries taken from the OQMD, our model achieves a mean absolute error (MAE) of 80 meV/atom in cross-validation, which is lower than the approximate error between DFT and experiment formation enthalpies and below 15% of the mean absolute deviation of the training set. We also demonstrate our method can accurately estimate the formation energy of materials outside of the training set before determining their equilibrium crystal structure (MAE: 136 ± 64 meV/atom). We propose that our models can be used to accelerate the discovery of new materials by identifying the most promising materials to study with DFT at little additional computational cost.

6.2 Introduction

Especially in the past decade, high-throughput atomistic calculation methods have proven to be powerful tools for discovering new materials.[18,20,32,38,192,193] These methods

generally work by employing an accurate computational tool, often Density Functional Theory (DFT), to predict the properties of large numbers of experimentally-observed and hypothetical inorganic compounds created by substituting different elements into known compounds. Given the advances in computing technology, the largest of these databases contain the predicted properties of on the order of 10^6 distinct crystalline materials. The results of these predictions are often stored in publically-accessible databases,[32,37,39,40,194] which makes it possible to many researchers to quickly search for materials that warrant further investigation (e.g., via experimental synthesis). This strategy of combinatorial replacement and high-throughput calculations has already enabled the discovery of new materials for a host of applications, including Li-ion batteries, thermoelectrics, and structural alloys.[20,26,27,33,34,195–198]

While combinatorial searches are evidently useful, they are intrinsically limited by available computational power. Evaluating only the zero temperature, ground state properties of a material using DFT can require hours of processor time per compound. Consequently, the space of possible combinations is too large to evaluate every candidate for some types of compounds. For example, the combinations of every element in a quaternary crystal structure results in at least 2 million materials (more if there are inequivalent sites in the crystal), which far outstrips the capability of today's computational resources. For more complex properties (e.g., elastic constants), evaluating 2 million compounds is certainly impractical. At some point, it is necessary to selectively evaluate only the parts of the search space that are likely to contain promising candidates.

Past work has shown that it is possible to use optimization algorithms, such as genetic algorithms, to efficiently search for stable structures.[21,22,199] For example, Jóhannesson *et al*. used a genetic algorithm to search for order structures on an FCC lattice with low formation enthalpies.[21,22] While successful, these techniques are generally restricted to searching over a single, defined structure for a material and are designed as if no information relevant to the optimization problem has be gathered previously. One must start from the beginning each for each new search. Considering the vast resources of previous examples of DFT calculations,[18,37,38,41] it could be much more effective to integrate knowledge from these databases into new searches.

Machine learning (ML) offers a route for creating fast surrogate models from databases, and has proven to be a viable route for estimating the results of DFT calculations.[45] For example, previous work from our group has demonstrated how to predict the formation enthalpies from DFT calculations given composition [85,200], and have successfully used those models to predict the compositions of undiscovered compounds.[85] However, because the methods used in these papers are based on the composition of a material, they require expensive crystal-structure prediction algorithms to determine the structure of the material in order to validate the predictions. There has also been work showing how to predict some computationally expensive properties given the results of single DFT calculations, including elastic constants,[101,201,202] thermal conductivity,[24,203] and melting temperature.[163] Additionally, several studies have predicted new materials with a desired crystal structure by training a model on data including materials with the same stoichiometry or same crystal

structure, and using that to identify promising materials in a much larger set.[91,92,204] To fully leverage the amount of information available in high-throughput databases to discover new materials, one needs a reliable and fast method for predicting properties given any crystal structure – and such a model remains elusive.

Several different strategies for building ML models based on the crystal structure of a material have already been proposed. These methods are composed of two main components: (1) a numerical representation that describes each crystal structure, and (2) a choice of machine learning algorithm. These methods include work by Faber *et al.*[80] that uses a modified Coulomb matrix representation and kernel ridge regression (KRR) to predict formation energies, and work by Schütt *et al.*[160] that uses Partial Radial Distribution Functions (PRDF) to represent each material and KRR to predict the density of states at the Fermi level. However, the best reported machine learning model to date for using crystal structure to predict formation energy has a mean absolute error (MAE) of 370 meV/atom in [80], which is too large to be used in place of DFT calculations because the median formation enthalpy of compounds in the ICSD is only slightly larger, at 800 meV/atom (using data from the OQMD [18,32]). Furthermore, in order to use these models to actually predict new materials, they cannot be reliant on having already used DFT to compute the equilibrium structure of the material. The ability of these strategies to perform well with only an estimate of the fully-equilibrated structure has not been established. Additionally, as we will demonstrate in this manuscript, these methods are impractical to use with the datasets as large as those currently available.

Overall, while promising, there is a need for improvements in methods that can link crystal structure and properties with machine learning.

In this work, we demonstrate an approach for predicting properties of crystalline compounds using attributes derived from the Voronoi tessellation of its structure that is both twice as accurate as existing methods and can scale to large training set sizes. Additionally, we designed our representations for the crystal structure to be length-scale invariant, which makes it possible to predict the properties of the crystal without needed to first compute the equilibrium structure as input into the model. In this manuscript, we will benchmark this new method against existing techniques in the literature (the Coulomb matrix and PRDF methods) using cross-validation with data from the OQMD. Then, in order to understand limitations of our approach, we employ cross-validation to assess whether it is actually learning the effect of structural characteristics of materials and to determine which types of compounds yield the highest error rates. Finally, we validate the ability of our model to make predictions of the formation enthalpy of materials outside our currently-available training data and to identify materials with strongly-negative formation enthalpies given only an estimate for the equilibrium atomic positions. We envision that this model can be used to screen potential materials based on stability before more expensive calculation techniques are used and, thereby, enable faster high-throughput searches for new materials.

6.3 Methods

Our approach is composed of two distinct steps, (1) representing a crystal structure as a set of quantitative attributes, and (2) using machine learning to extract patterns that relate those

attributes to the property of interest. We describe both steps in this section, along with the resource used to provide training data for these models.

6.3.1 Training Data

All training data for the machine learning models created in this work was extracted from the Open Quantum Materials Database (OQMD).[18,32] At the time the data used here was extracted, the OQMD contains the results DFT calculations for 435k structurally-unique crystals all performed with the Vienna *Ab Initio* Simulation Package (VASP).[120,121] Detailed settings used in these calculations are described in Ref. [18]. The OQMD contains over 30k entries corresponding to entries from the Inorganic Crystal Structure Database (ICSD)[68] and the remainder are predominantly hypothetical structures created by replacing elements in known crystal structures with different elements. As described in later subsections, we use several unique subsets of this database, which include using only the entries from the ICSD.

6.3.2 Representing Crystal Structures

The representation of a material is designed to transform raw materials data (e.g., a crystal structure) into a list of quantitative attributes and, functionally, are what serve as input into a machine learning model. Following previous discussions of the desired features of representations for materials,[47,91,160,205,206] we also assert that representations for crystalline compounds should be quick to compute and capture all relevant features of a structure in a compact list of attributes. Additionally, we suggest several other desirable features specific to building representations for crystal structures. First, these attributes should also be insensitive to the choice of a unit cell (i.e., primitive cells, conventional cells, and

supercells of the same structure should all have the same representation). Additionally, as our goal in using these models is to estimate the stability of a crystal structure before employing DFT, we also assert that it should fill two other requirements to be predictive without first computing the equilibrium atomic positions and lattice vectors with DFT. For one, the representation should at least be invariant to changes due to simple dilation or contraction of the lattice in order to minimize the effect of changes in atomic structure (assuming that relaxations do not significantly affect coordination environments). Also, the representation should be designed such that small changes in the structure (e.g., perturbations in atomic position) do not result in unphysical, discontinuous changes in attributes.

Considering all of these constraints, we created a representation for crystalline compounds based on the Voronoi tessellation of the structure.[207] The Voronoi tessellation of a crystal partitions space into the so-called Wigner-Seitz cells of each atom, which encompass the region closer to that atom than any other atom.[208] This tessellation is uniquely defined for a crystal structure and is insensitive to the choice of unit cell. The faces of a Voronoi polyhedron correspond to the nearest neighbors of an atom, which provides an unambiguous way of describing its local environment. To create attributes, we compute many characteristics of the local environment of each atom (described below) and then measure statistics about the distribution of these characteristics across all atoms in the unit cell. These attributes are designed in such a way that they are unaffected by unit cell selection or by changing the volume of the unit cell. Furthermore, we also weigh the contribution of each neighboring atom to each attribute according to the area of its corresponding face on the Voronoi cell. In this way, the

attributes are stable against discontinuities caused by addition or removal of faces, which can be caused by small deformations in the structure.

We use the Voronoi tessellation and composition of the structure to create several different categories of attributes:

1. **Effective Coordination Number Attributes** based on the mean, maximum, minimum, and mean absolute deviation in the effective coordination number of each atom, which is computed using the equation

$$CN_{eff} = \frac{1}{\sum f_i^2} \quad (1)$$

where f_i is the fraction of surface area corresponding to face i. This formula reverts to the number of faces on the cell for cells with equally-sized faces (e.g., 12 for FCC) and leads to smaller coordination numbers for structures with unequal faces (ex: 11.96 rather than 14 for BCC).

2. **Structural Heterogeneity Attributes** that measure the variation in local environments around each atom. Includes statistics regarding the mean bond length about each atom, the variation in bond length between each neighbor of an atom, and variation the volume between each Voronoi cell

3. **Chemical Ordering Attributes** that are computed using the Warren-Cowley ordering parameters[209] of the first, second, and third neighbor shells, weighted according to face sizes of each neighboring atom. To make the number of attributes the same regardless of the number of elements in the crystal, we measure the mean

absolute value of ordering parameters for each atom in the lattice for each type in the crystal. Consequently, crystals with ordered arrangements (e.g., rocksalt) will have values of this attribute closer to 1, and more random arrangements will be closer to zero.

4. **Maximum Packing Efficiency**, which can be computed by finding the largest sphere that fits inside each Voronoi cell. For example, the maximum packing efficiency for FCC is 0.74 by this definition.

5. **Local Environment Attributes** that are computed by comparing the elemental properties of the element of each atom to those of its nearest neighbors using the relationship

$$\widehat{p_i} = \frac{\sum A_n * |p_n - p_i|}{\sum A_n} \quad (5)$$

where A_n is the area of face n of the Voronoi cell of atom i, and p_n and p_i are the properties of the atom corresponding to face n and atom i, respectively. For this study, we compute the mean, mean absolute deviation, maximum, minimum, and range of this property for all atoms in a structure for 22 different elemental properties (e.g., atomic number), which are listed in Table 10.1. For example, each atom for NaCl in the rocksalt structure is surrounded by only atoms of the opposite type. So, difference between the electronegativity of each atom and its neighbors is 2.23 (the difference between Na and Cl) and the mean across the entire structure is also 2.23.

6. **Composition-Based Attributes** based on the fractions of each element present in the structure. These attributes are described in recent work by Ward *et al*. [200]:

a. **Stoichiometric attributes** that depend on the fractions of each element and not what those elements actually are
b. **Elemental-property-based attributes** that are based on statistics of the elemental properties of all atoms in the crystal.
c. **Electronic structure attributes,** which depend on the fraction of electrons in the *s, p*, *d*, and *f* shells of the constituent elements. These are based on work by Meredig *et al*.[85]
d. **Ionicity attributes** derived from differences in electronegativity between constituent elements and whether the material can form a charge-balanced ionic compound if all elements have on common oxidation states

Further details about the attributes are described in Appendix. In total, our method describes each material with 271 attributes. Each of these attributes can be computed using the Materials-Agnostic Platform for Informatics and Exploration (Magpie) and the Versatile Atomic-Scale Structure Analysis Library (Vassal), which are both freely available under open source licenses.

6.3.3 Machine Learning Technique

For the machine learning algorithm, we chose to use the Random Forests (RF) algorithm proposed by Breiman due to its superior performance and robustness against overfitting.[77] The RF algorithm works by aggregating the results of several decision trees, each built from a random subset of training examples and attributes. Each decision tree is composed of a series of decision rules (e.g., Packing Efficiency > 0.5) learned by recursively partitioning data into

pieces and assigning to each piece a value that minimizes the error to the training set. The decision rules used to create partitions are determined finding a rule that minimizes intra-subset variation of class values, which, in this case, are formation enthalpies. This decision tree generation process is repeated several times with a different subset of the training set, and the predictions made from all decisions trees are averaged in order to predict the class value of new data.

In modeling our problem, we used an ensemble of 50 decision trees for all machine learning models created based on the ICSD dataset and 100 decisions trees for machine learning models created with the full OQMD dataset. We also investigated increasing the number of trees as the training data increases but no notable improvement was observed. Models were constructed using the Scikit-Learn library in Python [75] and the Weka machine learning library in Java.[74]

6.3.4 Coulomb Matrix and Pair Radial Distribution Function Methods

In this work, we compared our new method against the Coulomb Matrix[80] (CM) and Partial Radial Distribution Function[160] Matrix (PRDF) approaches. Both of these methods utilize Kernel Ridge Regression (KRR) as the base machine learning algorithm, which performs linear regression where the input into the linear model are differences between a new observation and each entrt in the training set. This difference metric is often designed specifically for a particular problem, and the CM and PRDF methods primarily vary in the choice of this difference metric used to compare two crystal structures.

The PRDF method expresses the difference between two structures based on a matrix defined by the Partial Radial Distribution Functions between each pair of atoms in the structure.[160] Each row of this matrix corresponds to the radial distribution function between a different pair of elements. For instance, one row is the Li-Cl RDF, which describes the frequency of Li and Cl atoms a certain distance apart in the structure. To compute the difference between two structures, one generates this matrix for both structures and computes the Frobenius norm of the difference between the two matrices.

The Coulomb Matrix method is based on a representation that was originally developed for molecules.[180] In this representation, one computes a matrix that is related to the Coulomb repulsion between the atomic nuclei in the material

$$C_{ij} = \begin{cases} 0{,}5Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{r_{ij}} & \text{if } i \neq j \end{cases} \quad (6)$$

where Z_i is the atomic number of atom i and r_{ij} is the distance between atoms i and j. In order to compare two structures, one first computes the eigenvalues of the Coulomb matrix for both structures and then subtracts the two lists of eigenvalues (padding with zeros to make them the same length). More recently, Faber *et al.* proposed several modifications to the Coulomb matrix to account for periodic boundary conditions.[80] Of their proposed modifications, we use the Sine Matrix approximation, which they found to lead to the lowest cross-validation error when predicting formation enthalpy.

For both methods, we optimized the metaparameters for the KRR learning algorithm and, for the PRDF matrix, the cutoff radius and bin size used for the RDF. In both cases, we used a

grid search technique. All parameters were varied in order to maximize the performance of each model at a training set size of 3000 entries. With this technique, we were able to reproduce the observed cross-validation error of the Coulomb matrix reported in Ref. [80]. Our implementation for both of these methods is available as part of Magpie.[210]

6.4 Characterizing Model Performance

In this section we characterize several different aspects of our new machine learning technique. First of all, we compare our technique to existing methods by comparing their cross-validation accuracy. Then, we analyze the predictions where our model performs least accurately in order to determine where this model can be best applied. Finally, we study the effect of structural information in our representation in order to determine whether the model is actually learning the effect of structural traits on formation energy.

6.4.1 Comparison to Existing Techniques with Cross-Validation

We first used cross-validation to study the ability of our technique to model the formation energy of inorganic compounds and compare its performance to existing methods. As a training set, we used the equilibrium structures and formation energies of all compounds in the Inorganic Crystal Structure Database (ICSD)[211] available in the Open Quantum Materials Database (OQMD).[18,32] Our dataset includes 32111 compounds and, with minor exception, represents an unbiased sampling of all known compounds with a primitive cell size smaller than 40 atoms. In order to assess the effect of increasing training set size, we constructed models using randomly-selected training sets with between 1 to 30,000 entries and evaluated the performance of the model on a distinct set of 1,000 entries. This test strategy was selected in

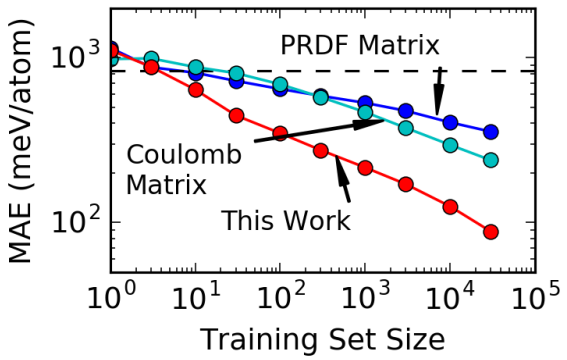


Figure 6.1. Mean absolute error (MAE) measured using cross-validation of models created using the PRDF, [35] Coulomb Matrix (CM), [34] and the method presented in this work. Each model was trained on the DFT formation energies of a set of randomly-selected compounds from the ICSD and used to evaluate 1000 distinct compounds that were also selected at random. The black, dashed line indicates the expected error from guessing the mean formation energy of the training set for all structures.

order to assess the effect of training set size on model performance. Each cross-validation test was repeated 20 times for each training set size, and the performance of the model was taken to be the average over all tests.

We found that the models created using our approach were more accurate than those based on the CM and PRDF methods for all training sets larger than 3 entries. As shown in Figure 6.1, models based on our method were found to have an MAE of 170 meV/atom at a training set size of 3000 entries. In contrast, the CM and PRDFs models were found to have 2.2x and 2.8x larger errors, respectively. At a training set size of 30000 entries, the MAE of our model (88 meV/atom) is still significantly lower than those from the other two methods. Since the error of our models decreases with increasing training set size at a similar rate to those of the CM method and faster than those from the PRDF method, we expect our models to be more accurate even when trained with the largest available DFT formation energy datasets of between 10^5 – 10^6 compounds.[32,37]

In order to determine whether the increased accuracy is a result of the new representation or the use of the RF algorithm, we repeated the comparison between the Coulomb Matrix and

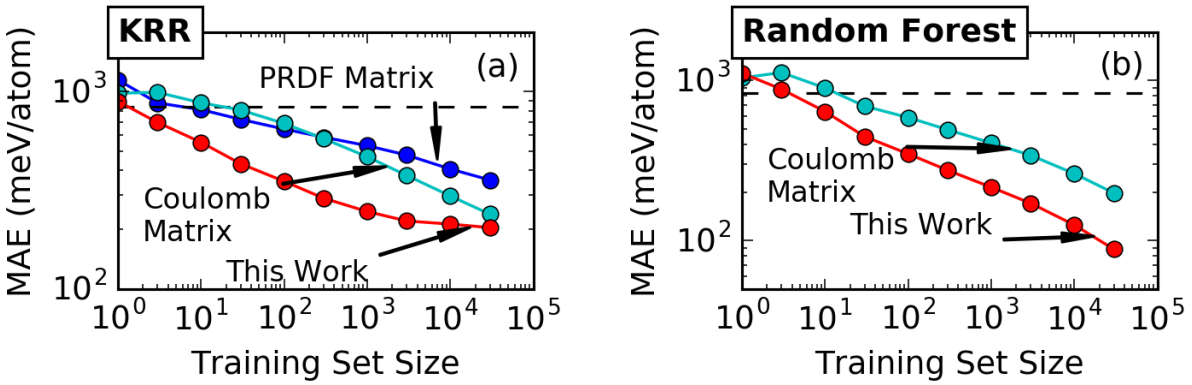


Figure 6.2. Performance of machine learning models for formation enthalpy created with the same machine learning algorithm but different representations. Each graph shows Mean Absolute Error (MAE) for (a) Kernel Ridge Regression (KRR) model and (b) Random Forest algorithm in a cross-validation test where the model was trained on progressively larger training sets and validated against a separate test set of 1000 entries. For each algorithm, we compare the performance using the Voronoi-tessellation based representation proposed in this work is compared against Coulomb Matrix [80] and the Partial Radial Distribution Function (PRDF) Matrix representations.

our new representation using both Kernel Ridge Regression (KRR) and RF as the learning algorithm for both representations. For the KRR test, the MAE for the model using our new representation was significantly higher than when we used the RF algorithm, but still lower than the CM and KRR model (see Figure 6.2). In contrast, the error rate of models created using our representation was lower than those using the CM by a factor of 2 when we employed RF as the learning algorithm. From this, we conclude the improved accuracy of our models is a result of the new representation, and not only the choice of machine learning algorithm.

Additionally, we found that the training time of our method scales better with increasing data size and has similar evaluation speed than the PRDF and CM methods. As shown in Figure 6.3, as the size of the training set reaches 10000 and more, the time taken to train and run models created using our method is comparable to the PRDF and CM methods. The training and run time of our model is dominated by the time required to compute the Voronoi tessellation used to generate the attributes, which requires approximately 0.1 s per compound on our test system and accounts for ~98% of the model training time and >99% of the run time. While we observe a $O(N)$ scaling for training time due to the large calculation time for the representation, the training time formally scales with $O(N \log N)$ scaling for the Random Forest algorithm. For small dataset sizes, the time to compute this attributes makes it slower to train and run than both of the competing methods. However, this is not true for large datasets and we observe parity between the two methods for training set sizes around 10^4. Considering that the training time for the CM and PRDF is scale at the faster rate of $O(N^3)$ for KRR, our approach will remain more feasible to train for even larger datasets. For datasets with only 30000 entries in the training set, our method is faster to train by approximately a factor of 10 and is only slightly slower to run than the Coulomb Matrix model – although we found (see Figure 6.3) differences in run

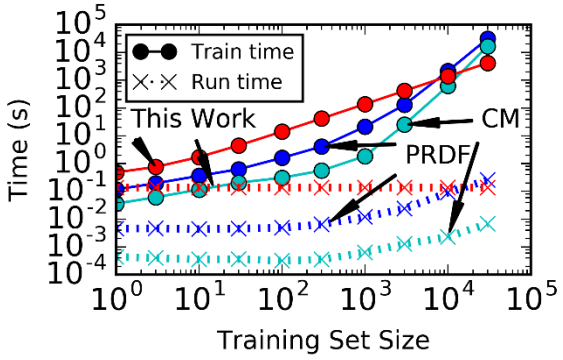


Figure 6.3. Comparison of model training and running time of three different techniques to predict the formation energy of inorganic compounds. Training time is the sum of attribute generation and model construction with given data. Run time is the average time taken to compute the requisite attributes and evaluate the machine learning model for a single compound.

speed are also likely to close with increasing training set size.

6.4.2 Testing for Systematic Errors in Our Models

In order to understand where our machine learning model can be used the most effectively, we ran a cross-validation test and studied the compounds where the model had the highest error rates. In this cross-validation test, we withheld a random selection of 25% of the ICSD dataset used in the previous section for a test set and trained the model on the remaining 75% of the data. We then repeated this test 100 times, and measured the MAE for each compound over all times it appeared in the test set. Then, we selected the 643 compounds with highest 2% of MAE values (above 446 meV/atom) in order to determine which compounds our model are persistently the worst at predicting accurately. First, we found that many of these outliers are compounds with extraordinarily small or positive formation enthalpies (see Figure 6.4a). The fact our model performs poorly for these compounds is unsurprising, since these compounds are evidently outliers compared to the rest of the ICSD training set. However, we cannot use the DFT formation enthalpy of a compound to predict whether the model will be unreliable *a priori,* which makes it an ineffective descriptor for determining where this model can be employed.

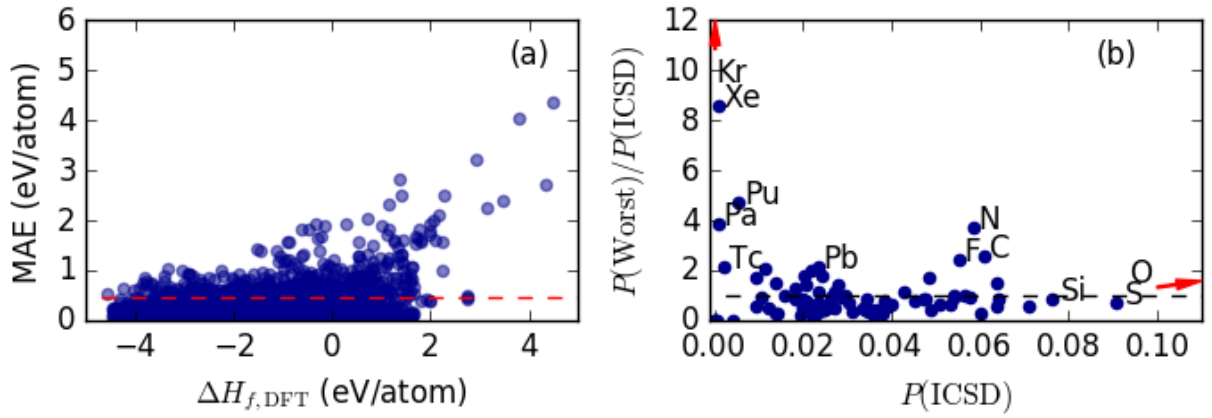


Figure 6.4. (a) The DFT-computed formation enthalpy of a compound compared to the mean absolute error (MAE) between the DFT and machine-learning-predicted formation enthalpy of that compound during a cross-validation test. The red, dashed line indicates the 98th percentile of the mean absolute error. (b) Comparison of the fraction of compounds that contain a certain element in our ICSD training set P(ICSD) to the ratio between the fraction of compounds in the 98th percentile of MAE and the fraction in the training set.

We also found that compounds containing elements that appear least frequently in our training set are overrepresented in the compounds with the worst MAEs. Figure 6.4b shows the probability of finding a compound containing a certain element in our entire dataset ($P(\mathrm{ICSD})$) and the ratio between the probability of finding that element in the entries with the highest MAE ($P(\mathrm{Worst})$) and the probability of finding it in the entire dataset. Of all elements present in the training set, Kr, Xe, and Pa have the highest overrepresentation (a ratio of 14 for Kr) and are among the least frequently appearing elements in the original dataset. Several other infrequently appearing elements (He, Ne, Ar, Pm, Ac) violate this trend because they appear infrequency appear in both the training set and the list of worst predictions. In the case of the noble gas elements in this “surprisingly good” category, they only appear as elemental compounds in training set and our model correctly identifies those compounds as having near-zero formation enthalpies. From these results, we conclude our model performance is expected

to be worst for compounds containing noble gases and compounds with infrequently occurring elements (e.g., Tc, actinides).

The elements that are both frequently occurring and overrepresented in our worst-performing materials are C, N, and F. Out of the 643 worst compounds, there are 43 that contain either C or N. This set mostly includes compounds that contain rarely-observed elements (e.g., Th, Pu), and instances where with each of these elements and some combination of C, O, or F. As the C-, O-, and F- compounds may have a tendency to exhibit covalent bonding, this suggests that our model could be improved by including attributes that capture characteristics such as bond angles or using electron counting rules to characterize the types of bonds present in the structure. Beyond identifying regions to improve this model, our analysis of its failures also identifies where it can be applied with the greatest likelihood of success: compounds with commonly-occurring elements without significant covalent bonding.

6.4.3 Assessing Whether Algorithm Has Learned Structural Effects

As many of the attributes employed in our representation are not dependent on structure, it is important to determine whether the structure dependent terms actually have an effect on the accuracy of our machine learning models. If these structural attributes have a negligible effect, it is possible that the model is only learning structurally-invariant (i.e., composition-based) attributes. To test the effect of including structure-dependent attributes, we replicated the cross-validation described in the previous section and trained a Random Forest algorithm with three sets of attributes: (i) only the composition-based (i.e., structure-independent) attributes, (ii) only the structure-dependent attributes, and (iii) all 271. As a reference, we also

include the results of a Random Forest using the Coulomb Matrix model. As shown in Figure 6.5a, there is little difference between the error rate of a model trained using all the attributes and the structure-independent ones. The structure-dependent attributes do lead to a machine learning model with superior performance to the Coulomb Matrix representation. Consequently, we do conclude that the Voronoi-based attributes do carry useful information about a material, but we are unable to determine whether they leave to an *improved* model compared to a purely composition-dependent model.

One possible explanation for the similar performance between model trained on composition-only and composition-and-structure representations is that the ICSD dataset contains too few examples of multiple structures at the same composition. Consequently, there could be insufficient training data to build a model that benefits from the additional structural information. To test this hypothesis, we repeated the cross-validation test using a dataset

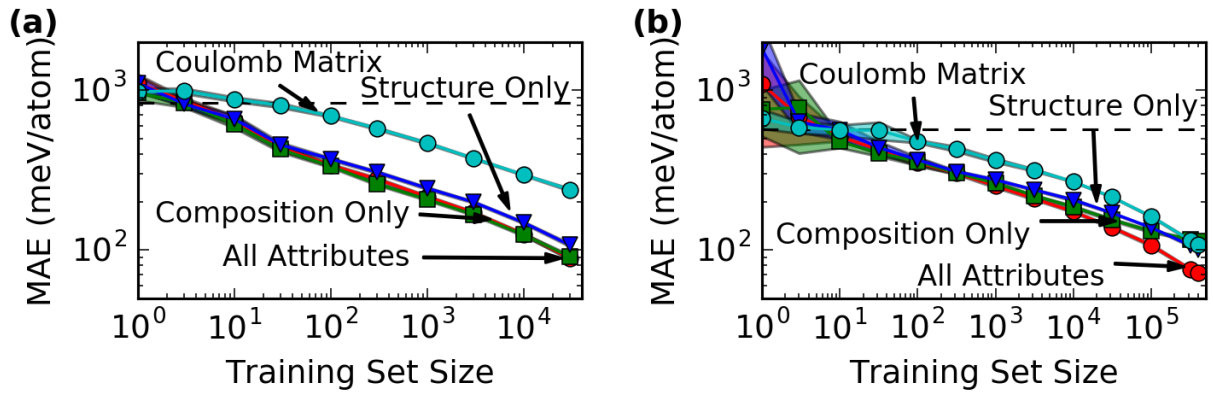


Figure 6.5. Performance of machine learning models trained on various kinds of representations in cross-validation tests using data from the (a) ICSD subset of the OQMD and (b) the entire OQMD. These include models trained using all of the attributes in our proposed representation and, separately, models created using only the composition-dependent terms and only the structure-dependent terms. The results of a model created using the Coulomb Matrix and Random Forest is shown for comparison. Shaded regions represent the 90% confidence intervals.

comprised of all non-duplicate entries from the entire OQMD (435792 entries), which contains dramatically more examples of multiple structures at a single composition. Only 51% of the training entries in this dataset lack another structure at the same composition, which is lower than the 70% of entries without another example structure in the ICSD dataset used previously. With the larger dataset, we observed a significant improvement when using both the structure- and composition-based attributes rather than either subset of attributes alone (as shown in Figure 6.5b). At a training set size of 400000 entries, the model using structural and composition-based attributes has an error rate around 35% lower than the composition only models.

The increased accuracy of the "all attributes" model on the OQMD dataset is not merely an effect of training set size. At a training set size of 10^4, the composition-only model trained on the OQMD dataset (with fewer compositions with only 1 structure) has a 7% larger MAE than the "all attributes" model (185±3.5 vs 174±2.0) meV/atom. For the same training set size and the ICSD training set, the composition-only and "all attributes" model have approximately the same MAE (125±2.2 vs 126±1.8 meV/atom, respectively). The difference between the composition-only and "all attributes" model in our full OQMD test only becomes larger with increasing sample size. This lower error suggests there is indeed an advantage to introducing structure-based attributes to our machine learning models but this effect is only significant in datasets where there are sufficient training examples. Additionally, this result also demonstrates that our model has learned the effect of structural features in the crystal structure and not just from the composition.

6.5 Applying Method to Predicting New Materials

In this section, we explore using this model to assess the performance of our machine learning models in two applications: (1) predicting the formation enthalpy of experimentally-observed compounds yet to be included in the OQMD, and (2) identifying which materials are most likely to be stable out of a list of compounds studied via a high-throughput search. In both cases, we also seek to determine whether our models can perform well when provided with only the unrelaxed structures that serve as input into DFT calculations. In contrast, we used the fully-relaxed structures generated as output from a DFT calculation as input into our machine learning model in the cross-validation tests in the previous section.

6.5.1 Validation with Yet-Unevaluated Materials

One unresolved question from our cross-validation test is whether our models can predict the formation enthalpy of a material without knowledge of the equilibrium structure. To answer this question, we used our model to predict the formation enthalpies of compounds from the ICSD that have yet to be included in the OQMD. These compounds generally have large unit cell sizes, which leads to high computational costs to evaluate with DFT and makes the ability to predict their energies with machine learning particularly useful. To make our model as accurate as possible, we trained a machine learning model on the full OQMD dataset. We then used this model to evaluate the 12667 entries from the ICSD that had not yet been added to the OQMD, which required less than 2 hours on a 2.2 GHz CPU. We then selected a total of 30 entries from this list to validate with DFT using three different strategies: (1) randomly-selecting entries, (2) selecting entries predicted to have the most negative ΔH_f, and

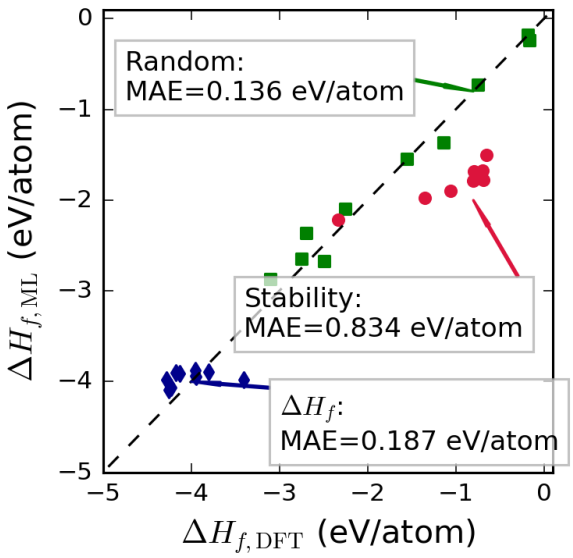


Figure 6.6. Comparison of formation enthalpies (ΔH_f) predicted using machine learning (ML) and computed using Density Functional Theory. The machine learning model was trained on the formation enthalpies of all 435792 non-duplicate entries from the OQMD. Each material was selected from a list of 12667 entries from the ICSD that have yet to be included in the OQMD using three different strategies: (green squares) random selection, (blue diamonds) predictions with the lowest ΔH_f, and (red circles) with the largest, negative difference between the predicted ΔH_f and the OQMD convex hull.

(3) selecting those predicted have the largest stability (farthest below the energy of OQMD convex hull at that composition [141]). By studying these three different strategies separately, we can also assess how best to use our machine learning model in practice.

As shown in Figure 6.6 and Table 6.1, we observed the best performance of the model in the entries that were randomly selected from the dataset – a MAE of 136 ± 64 meV/atom. This is excellent accuracy when considering that these predictions were made before determining the equilibrium structure of the material. The change in the predicted formation enthalpy between the model given the input structure and fully-relaxed structure was below 25 meV/atom for 9 out of 10 materials – far below MAE of 80 meV/atom observed in cross-validation. These results show that our machine learning model can predict the formation enthalpy of unstudied compounds with an accuracy on the order of 100 meV/atom and the predictions of our model are relatively insensitive to structural relaxations.

The MAE for materials selected by finding those with minimal ΔH_f was generally higher, 187 ± 99 meV/atom, but our model was successful in locating materials with especially-large formation enthalpies. The worst-performing entry in this dataset, CeF_4, is likely is an outlier because the DFT calculations in the OQMD treat Ce with only 3 valence electrons.[18] Consequently, the Ce^{4+} is not modeled correctly and formation enthalpy for CeF_4 will be more positive than what might be expected based on Ce in other oxidation states and the behavior of other metal-fluoride salts. There are four examples of Ce^{4+} in the worst 2% of predictions in the previous section (CeO_2, $BaCeN_2$, Li_2CeN_2, and Ce_2SeN_2) and the ML predicts a more negative formation enthalpy in all cases, just as observed for CeF_4. Provided enough training examples, our model should be able to learn the abnormal behavior of Ce^{4+} but it apparently lacks the ability with the current training set. If we exclude this compound from the analysis, the error rate in our test is only 144 ± 65 meV/atom. Regardless, the accuracy levels observed in these test is sufficiently high to successfully identify materials with exceptionally low ΔH_f. All of the compounds selected based on minimal formation enthalpy are within the 97th percentile (99.4th without CeF_4) and, on average, above the 99th percentile of all compounds in the ICSD. While the numerical accuracy of predictions is worse when preferentially selecting large formation enthalpy materials, we do find it sufficient to identify which materials are most likely to have a large, negative formation enthalpy out of a large dataset.

Table 6.1. Performance of machine learning algorithm in predicting the formation ethalpy (ΔH_f) of 30 materials outside of the training set that were three different strategies. The DFT computed value is compared to the ML prediction using the input structure to DFT (Before Relaxation) and the relaxed, output structure.

			Before Relaxation		After Relaxation		
	Composition	ΔH_f, DFT (eV/atom)	ΔH_f, ML (eV/atom)	Error (eV/atom)	ΔH_f, ML (eV/atom)	Change (eV/atom)	Error (eV/atom)
Largest ΔH_f	$SrMgF_4$	-3.952	-3.876	0.077	-3.862	0.014	0.091
	CeF_4	-3.400	-3.982	0.583	-3.887	0.095	0.488
	Sr_2ScF_7	-4.175	-3.902	0.273	-3.924	-0.022	0.251
	$RbLu_3F_{10}$	-4.275	-3.978	0.297	-4.001	-0.023	0.274
	$BaAlF_5$	-3.956	-3.936	0.020	-3.949	-0.013	0.007
	$ThZrF_8$	-4.223	-4.066	0.157	-4.039	0.027	0.183
	KU_2F_9	-3.800	-3.891	0.091	-3.869	0.022	0.069
	$RbTh_2F_9$	-4.252	-4.091	0.161	-4.104	-0.013	0.148
	Ba_2ZrF_8	-4.125	-3.912	0.213	-3.914	-0.002	0.212
	$Ba_7Cl_2F_{12}$	-3.939	-3.943	0.004	-3.943	0.000	0.004
Random	$CrHg_3Pb_2O_8$	-1.139	-1.368	0.229	-1.370	-0.002	0.231
	Y8Co56B4	-0.181	-0.176	0.005	-0.174	0.002	0.006
	$YH_3C_3S_2O_{12}F_9$	-1.555	-1.541	0.014	-1.535	0.006	0.021
	$CuH_{12}C_5S_4N$	-0.168	-0.239	0.071	-0.216	0.023	0.048
	$Rb_2Tc_3Se_6$	-0.750	-0.730	0.020	-0.727	0.003	0.022
	Li_6CaCeO_6	-2.257	-2.092	0.165	-1.373	0.719	0.885
	$Na_5Ti_2VSi_2O_{13}$	-2.747	-2.643	0.104	-2.656	-0.013	0.091
	ErP_5O_{14}	-2.692	-2.359	0.334	-2.368	-0.009	0.325
	$Cs_2USi_6O_{15}$	-3.100	-2.868	0.232	-2.854	0.014	0.246
	$RbVP_2O_8$	-2.490	-2.674	0.184	-2.672	0.002	0.182
Largest Stability	$CeTl_5Fe_2N_{12}O_{24}$	-0.804	-1.782	0.978	-1.754	0.028	0.951
	$YTl_5Cu_2N_{12}O_{24}$	-0.697	-1.673	0.976	-1.652	0.021	0.955
	$Rb_2BiCl_5O_{20}$	-0.655	-1.502	0.847	-1.502	0.000	0.847
	$YTl_5Co_2N_{12}O_{24}$	-0.752	-1.757	1.005	-1.727	0.030	0.974
	$TmAu_2F_9$	-2.331	-2.212	0.119	-2.211	0.002	0.120
	VXe_2F_{34}	-1.348	-1.978	0.629	-2.072	-0.095	0.724
	$CeTl_5Ni_2N_{12}O_{34}$	-0.777	-1.754	0.977	-1.780	-0.026	1.003
	$CsXe_3O_3F_{36}$	-0.687	-1.776	1.088	-1.782	-0.006	1.094
	$ScH_3Cl_2O_{10}$	-1.058	-1.895	0.837	-1.929	-0.034	0.872
	$CeAg_6(NO_3)_9$	-0.794	-1.679	0.885	-1.667	0.012	0.873

Of our three selection strategies, the accuracy of our predictions was worst when selecting materials predicted to be the most stable relative to other compounds. In this test case, our

error rates were approximately 850 meV/atom, which is approximately the error expected when guessing the mean formation enthalpy of the OQMD training set for all predictions. This poor performance could be a result of the biasing effect described by Faber *et al*.[91] In their paper, Faber *et al*. observed a low success rate when selecting Elpasolite materials based on the predicted stability with reference to other compounds. They attributed this low success rate to this strategy of selecting materials "systematically favor[ing] those [predictions] with negative ML formation energy errors."[91] Consistent with their observation, nearly all of our predictions made with this strategy have negative formation enthalpy errors and are well within the 99th percentile of magnitude of errors observed in our cross-validation test. This poor performance suggests that identifying materials based on the difference between ML-predicted formation energy and the energies of competing phases is problematic. Consequently, we recommend either searching for new stable materials by selecting those with minimal formation energies or directly predicting the stability in reference to other phases with machine learning.

Overall, this validation test was particularly successful. We were able to observe formation energy errors of approximately 130 meV/atom for randomly-selected materials, and were able to successfully locate materials with exceptionally low formation enthalpies. In these cases, making the ML predictions required only a tiny fraction of the thousands of CPU hours of DFT calculations necessary to validate them for these limited test cases. It is also worth emphasizing that we were able to observe these high accuracies without knowledge of the equilibrium DFT structure. Across all 30 predictions, the median difference between the prediction of our model with the initial guess provided to DFT and with the fully-relaxed structure was only

13 meV/atom – far below the error expected in the prediction from the cross-validation experiment at 80 meV/atom. This result demonstrates that our models can be used effectively when only an approximate model of the relaxed geometry is known – a very important feature when searching for new crystalline materials using machine learning.

6.5.2 Application to Combinatorial Searches

In order to test how our models could be applied to the high-throughput materials discovery process, we simulated the results of searching for new compounds based on several common crystal structures. First, we trained each model using data from all 32111 compounds in the OQMD that are based on entries from the ICSD. Then, we used this model to evaluate the formation enthalpies of all entries in the OQMD with the B2, $L1_0$, and ilmenite prototype structures. In order to simulate how this model would be used in practice, we evaluated the formation energy of the compound using the same input geometry provided to the DFT calculation. These three structural prototypes were chosen as separate test cases in order to sample structures that have a variety of local environments and that are known to be stable for compounds with both metallic and ionic bonding. Furthermore, the B2 and $L1_0$ datasets were created by generating all possible combinations of elements into the structure, which is useful for testing the ability of the model to evaluate a broad range of chemistries. In contrast, the ilmenite dataset is limited to only ABO_3 metal oxides and predominately includes materials with negative formation enthalpies, which will allow us to evaluate the performance over a more-restricted space.

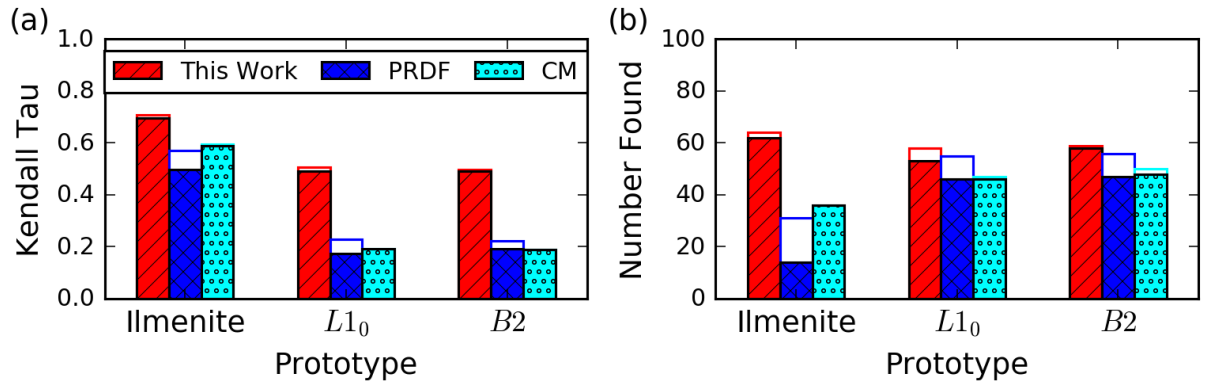


Figure 6.7. Comparison of the ability of different machine learning methods to rank different types of compounds based on DFT formation energy, measured using two different metrics. (a) The Kendall Tau ranking correlation coefficient, which is based on how well the model ranks the entire dataset. A correlation value of 1.0 corresponds to perfect ranking. (b) How many of the 100 compounds with the lowest DFT formation energy were predicted by the model to be within the lowest 100 compounds. Each model was trained on the DFT-predicted formation energy of 32111 inorganic compounds from the ICSD. The solid bar indicates the ranking performance using the input structure provided to DFT. The line above each bar indicates the ranking accuracy when provided with the fully-relaxed output from DFT.

To evaluate the ability of each machine learning algorithm to rank compounds from most to least stable, we measured the Kendall Tau ranking correlation coefficient between the predicted and actual formation enthalpies for each prototype structure. The Kendall Tau, which is defined as the difference in the fraction of pairs in a list that are correctly and incorrectly ordered,[212] allows us to understand how well the algorithm could be employed to prioritize compounds that are likely to be stable. As shown in Figure 6.7a, the model created using our new method has the highest ranking correlation coefficient for all three considered test cases. For the $L1_0$ structure, our model performs twice as well as the CM model and almost three times better than the PRDF model, and the differences are similarly large in the B2 test case.

The performance for all three machine learning methods was best for the ilmenite test case, where the dataset was restricted to metal oxides with mostly (99.4%) negative formation

enthalpies. In that example, our model had very strong ranking performance – a nearly 85% success rate. This exceptional ranking performance is likely a result of the test set containing mostly materials that have negative formation enthalpies. If we repeat the ranking test for B1 and $L1_0$ using only compounds with negative formation enthalpies in the test set, we observe improved performance for all three machine learning techniques. The improved performance on a dataset containing only low-enthalpy materials is consistent with previous finding that the model performs worst materials with positive formation enthalpies (see Figure 6.4a). Consequently, we purpose that the selection performance of each model could be included by first screening the space based on heuristic chemical rules (e.g., are the elements in reasonable oxidation states?). This could eliminate compounds that are more likely to be extremely unstable at the risk of potentially missing exciting materials that violate established rules (as in Ref.[91]).

One factor leading to improved performance of our method is the insensitivity of our representation to changes in volume. In the case of B2, the only degree of freedom in the crystal structure is the volume. Consequently, our predictions are not dependent on the quality of the initial guess for the equilibrium volume and, incidentally, the accuracy of the CM method is also only negligibly affected. In contrast, the predictive accuracy of the PRDF method increases significantly when we use the fully-relaxed geometry as input to the model. In the case of ilmenite and $L1_0$, our representation and, therefore, predicted enthalpies depend upon relaxation because there are other degrees of freedom in the structure. Even so, the mean change between the initial and relaxed structures in the predicted ΔH_f is less than

55 meV/atom and the correlation coefficient between the two predictions is approximately 99% for both structure types. Correspondingly, the ranking performance only changes slightly. Furthermore, the fact our model performs best when provided with the relaxed structure shows that we can expect our model to have the best accuracy even when provided a perfect estimate of the relaxed structure. For that reason, we conclude our model is the best choice for this ranking task.

In practice, these machine learning models might only be used to select the entries with the lowest predicted formation enthalpy. To measure the ability of each model to identify entries with the largest formation enthalpies, we measured the number of entries predicted by our machine learning model to have the 100 largest formation enthalpies that actually were within the top 100 of the test set. As shown in Figure 6.7b, the model created using our method performs the best according to this metric for all three cases and over half of the predictions made with our model are actually within the top 100. What this high predictive accuracy suggests is that it is possible to use a machine learning model trained on data with dissimilar crystal structures (e.g., the entire OQMD) to predict stable compounds with a target crystal structure without having to first create a new, problem-specific training set – as is common practice in previous accelerated searches for stable compounds.[21,91,92,204] By using existing data and our machine learning technique, we can quickly make predictions of which materials are most likely to be stable and use that knowledge to accelerate high-throughput DFT searches for new materials.

6.6 Conclusions

In this work, we present a strategy for predicting the formation energy of crystalline, inorganic compounds using characteristics derived from the Voronoi tessellation of its structure and machine learning. We demonstrate that these models are more accurate in cross-validation and better at ranking unseen compounds from most to least stable than those produced using the Coulomb Matrix[80] and Partial Radial Distribution Function[160] methods, and equivalently as fast. Furthermore, we show that our model is actually learning the effect of structure on formation enthalpy and can accurately predict the formation enthalpy of materials without knowledge of the exact ground-state crystal structure. Provided the high predictive accuracy of this method and the ability to utilize large training datasets, we envision it will be possible to employ this method to identify new, stable materials at a low computational cost.

7 Engineering Bulk Metallic Glass Alloys with Machine Learning

7.1 Abstract

Bulk metallic glasses (BMGs) are a unique class of material that have proven to be excellent choices for many applications, but the wide-scale use of these materials is limited by the lack of tools for engineering their properties. To address this issue, we developed a framework for designing metallic glasses using machine learning (ML) models that predict three key properties of metals using only the composition as input: the ability to form in the amorphous state, critical casting diameter (D_{max}), and supercooled liquid range (ΔT_x). These models were created from a database assembled from several dozen papers and handbooks, and contains the compositions and properties of several thousand metallic glasses. We employ these models to optimize the properties of known alloys, and to identify new metallic glasses in yet-unstudied alloy systems. We validated our predictions using commercial injection molding equipment and found several of our ML-predicted compositions can form glasses and exceed existing alloys in one of our two design variables, ΔT_x. We envision that these machine learning models will enable quickly tailoring bulk metallic glass alloys for new applications.

7.2 Introduction

While the amorphous structure of Bulk Metallic Glasses (BMGs) gives them combinations of properties that are impossible in conventional polycrystalline metals, it also makes them difficult to engineer. The lack of atomic-scale order leads to high mechanical strength, high corrosion resistance, low magnetic hysteresis, and the ability to be formed using net-shape casting techniques – among many other useful characteristics. [186] This unique set of

properties make them promising materials for applications such as surgical tools, pressure sensors, flight control surfaces, automotive components, and more.[213] However, also owing to the energetic metastability of the amorphous structure, few alloys have the right characteristics to make bypassing crystallization easy enough to form bulk amorphous castings. Furthermore, alloys capable of forming into bulk amorphous castings, much less ones with desirable properties, are difficult to locate *a priori*. What would enable the wider-scale technological use of metallic glasses is the ability to quickly locate new alloy compositions with optimal properties.

Conventional alloy design for metallic glasses relies on empirical rules and extensive experimentation. For example, it is known that BMGs tend to form in systems with a large diversity in atomic sizes[186] or near eutectic compositions.[214] These empirical rules can also be informed using physics-based models (e.g., computational thermodynamics to predict the driving force for crystallization),[215–223] which often require both large computational resources and experimental measurements from each alloy system being assessed. However, even with the availability of these rules, extensive experimentation is often required to locate alloys with optimal properties.[187] Furthermore, there are some important properties (e.g., supercooled liquid range), for which neither empirical rules nor physical models linking the composition and properties of the material exist. Fortunately, the fact that there are known empirical rules for metallic glass formation and large resources of experimental data available make a new tool for alloy design feasible: machine learning.

Machine learning models are created by employing algorithms that automatically discover how characteristics of a material relate to its properties, provided sufficient training data. These models offer the advantage of being able to link the structure and properties of a material when empirical intuition or physical models do not exist, and can create models fast enough to enable the rapid evaluation of millions of candidate materials.[85,91,93,200] Machine learning has been employed to design and discover a wide variety of materials, including crystalline compounds,[85,91] thermoelectrics,[93] shape memory alloys,[224] and many others.[24,94,95,170,203,204,225]

The application of machine learning techniques to the development of metallic glasses has yet to become prevalent. To date, Ward *et al*. have developed a model for predicting whether a composition can be formed into a thin amorphous ribbon,[200] Tripathi *et al*. have devised a scheme for linking the composition of an alloy to its critical temperatures (e.g., glass transition temperature),[226] and there have been two studies that employed regression algorithms to determine a link between critical casting diameter and other experimentally-measured properties of an alloy.[227,228] However a machine learning tool for predicting the ability to form a bulk metallic glass and the other key properties of an alloy remains elusive. Furthermore, there has yet to be any experimental validation of predictions from machine learning models for metallic glasses.

In this paper, we describe and experimentally validate a machine learning framework for accelerating the design and discovery of bulk metallic glasses. To do so, we used machine learning to create models for three key properties of BMGs: the ability to form a metallic glass,

the critical casting diameter, and the supercooled liquid range. These models were created with data from the literature, and were designed to incorporate empirical rules about metallic glass formation. We then employed these models to suggest compositions of alloys that could have both large critical casting diameters and supercooled liquid ranges. These tools will enable the development of new, cheaper metallic glasses and improve the ability of engineers to tailor the properties of BMGs for specific applications. The ability to create optimized alloys could then help accelerate the incorporation of BMGs into a wider variety of technologies.

7.3 Methods

Machine learning models are composed of three separate components: a resource of training data, a representation to describe key characteristics of each entry, and a machine learning algorithm. In this section, we will describe these three components separately.

7.3.1 Training Data

Our training set was assembled from publically available data collected from many different articles and handbooks. The data collection process was significantly aided by previous efforts in collecting large resources of metallic glass data, such as the Landolt-Börnstein Handbook on “Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys”[172] and papers by Long *et al.*[229] and Tripathi *et al.*[226] The data we collected was partitioned into three distinct training sets: a dataset of (i) whether it was possible to create a bulk metallic glass, (ii) the critical casting diameters, and (iii) the supercooled liquid ranges.

7.3.1.1 Glass-Forming Ability Classification Dataset

Our glass-forming ability (GFA) training set was composed of 6315 measurements of whether it was reported to be possible to form a BMG, an amorphous ribbon via melt spinning, or not at all. This dataset was assembled by combining measurements of whether it was possible to form a glass with melt spinning, taken from the Landolt-Börnstein Handbook, with the compositions of known BMGs collected from 41 different papers.[189,229–268] Where conflicting reports exist about the glass-forming ability of a certain composition, we select the most optimistic report (i.e., BMG over Ribbon over None) for inclusion within our dataset.

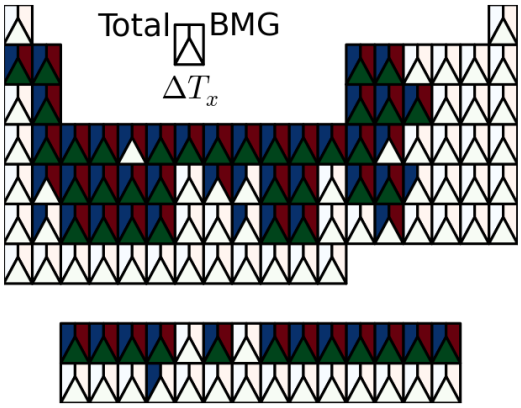


Figure 7.1. Elements present in each training dataset. The blue fill in the top left segment indicates the element is present in any dataset. The red fill in the top right segment indicates the element is present in BMG alloys included in our training set. The green fill on the bottom indicates the element is present in the supercooled liquid range dataset.

The entries in our dataset span a broad range of chemistries. As shown in Figure 7.1, the dataset contains 55 different elements, including transition metals and metalloids, and nearly all of these elements are present in examples of BMGs. Furthermore, the dataset includes examples of interactions between 513 pairs of elements, spanning metal/metal and metal/metalloid glasses.

There are two major biases in our data that need to be considered when training GFA models. First of all, the only negative examples of glass-formation are from Ref. [172], which contains only alloys with 3 or fewer elements. Consequently, the number of elements in an

alloy should not be used as a factor in our classification models. Additionally, as the ability to form a bulk glass was not tested during the ribbon forming experiments, a positive measurement of ribbon forming ability does not prove a material cannot be formed using bulk techniques. This issue limits the ability of our models to distinguish between bulk and ribbon-forming glasses, but not the ability to distinguish between glass-forming and non-glass-forming compositions.

7.3.1.2 Critical Casting Diameter Dataset

The critical casting diameter (D_{max}) defines the largest diameter at which it is possible to cast a fully-amorphous rod of metallic glass. Our training dataset was created using the same literature source as the classification dataset, and has 5916 entries. For alloys extracted from the BMG literature, we used the value of critical casting diameter reported in the paper, if available. For alloys that were classified as ribbon-forming in the Landolt-Börnstein handbook,[172] we assumed a single, small value of 0.2 mm for the critical casting diameter of each alloy. Any alloy classified as "not glass-forming" was assigned a diameter of 0. When multiple values of the critical diameter were available, we used the average of all available measurements.

7.3.1.3 Supercooled Liquid Range Dataset

The supercooled liquid range (ΔT_x) is defined as the difference between the glass transition temperature (T_g) and the onset temperature at which an alloy begins crystallizing rapidly on heating (T_x), and is correlated with how easily a metallic glass can be shaped after solidification. At 621 entries, this training dataset is the smallest of the three, and includes 45 different

elements and 307 different pairs (see Figure 7.1). Because ΔT_x measurements can vary significantly for single alloys (ex: reports of $Fe_{77}B_{22}Y_6$ vary by 40 K[242,243]), we use the average of all measurements for a particular composition.

7.3.2 Representation: Attributes to Describe Materials

The representation of a material is a set of quantitative attributes that describe the material and are what serve as input to a machine learning model. The goal in selecting a representation for materials is to construct a set of attributes that both differentiates materials and captures factors that could be related to the properties of interest. By including attributes that are known to be correlated with the property of interest as input into the model, it becomes possible for a machine learning model to automatically recognize these correlations and, thereby, create a more predictive model.

As our objective is to find new alloy compositions for BMGs, we chose to differentiate materials based on composition. To create factors derived from the composition that reflect empirical knowledge about glass formation, we started with the set of composition-based attributes developed by Ward *et al.*[200] These attributes are primarily based on statistics of elemental properties (e.g., range in atomic radius of constituent elements), and include terms that reflect known empirical rules, such as the polydispersity of atomic radii and average number of valence electrons.[186,269] Additionally, we developed new attributes to reflect other empirical rules developed by the BMG community:

Cluster Packing Efficiency Attributes, which are based on the hypothesis that bulk metallic glasses are composed of special arrangements of atoms that are both energetically-

stable and symmetrically-incommensurate with long-range order.[270,271] These special clusters occur where the ratio between the radius of the central atom and the average radius of the first neighbor shell is close to the ideal ratio for a cluster with the same number of atoms in the neighbor shell (similar to Pauling's rule for ionic crystals).[221,272] To adapt this concept into attributes, we first compute the compositions of these special clusters based on a simple, geometric approximation.[221] We then compute the distance between the composition of each alloy and the first 1, 3, and 5 of these special clusters to use as attributes. Furthermore, we also estimate the packing efficiency assuming that the first neighbor shell of each atom has the same composition as the alloy, which has also been related to the formation of bulk metallic glasses.[247]

Proximity to Crystalline Compound Attributes, which reflect the idea that the driving force for crystallization is correlated to distance between an alloy composition and nearby crystalline compounds.[217,273] We represented this effect using data from the Open Quantum Materials Database (OQMD),[18,32] which contains the DFT-predicted T=0 K energies of hundreds of thousands of experimentally-observed and hypothetical structures. We compute the formation enthalpy and several measures of the distance between an alloy and the nearest phases to use as attributes.

Probability of Forming a Glass, which were computed using our glass-forming ability model and employed as input into the models for the critical casting thickness and ΔT_x. Specifically, we use the predicted label from the GFA model (either BMG, ribbon-forming, or none) and the probabilities for each individual label as attributes.

In order to account for the sampling bias resulting in the only negative examples of glass-formation having less than 4 elements (see Training Data section), we excluded attributes that are based on the number of elements for the glass-forming ability classification model. These attributes include the "stoichiometric" attributes introduced by Ward *et al.*, and the number of crystals in equilibrium at a particular composition (i.e., due to Gibbs' Phase Rule).

7.3.3 Machine Learning Algorithms

The final ingredient of our models is a machine learning algorithm to create a function that maps the attributes in the representation to the property of interest. In order to select the optimal algorithm for each problem, we evaluated the performance in a 10-fold cross-validation test for over 10 different machine learning algorithms and selected the one with the lowest Mean Absolute Error (MAE) or highest classification accuracy.

7.4 Results and Discussion

7.4.1 Validating Machine Learning Models

We found that decision tree algorithms, specifically the Random Forest algorithm,[77] resulted in the most quantitatively accurate models in cross-validation for all three datasets. In the case of the D_{max} and ΔT_x models, we were able to further boost the accuracy in cross-validation using the additive regression technique.[274]

We first validated the predictive performance of our algorithms using 10-fold cross-validation. We found our classification model to have an accuracy of 89% (fraction of entries correctly labelled), which corresponds to a False Positive Rate (FPR) for incorrectly predicting a non-glass-forming composition as glass-forming of only 7% and True Positive Rate of

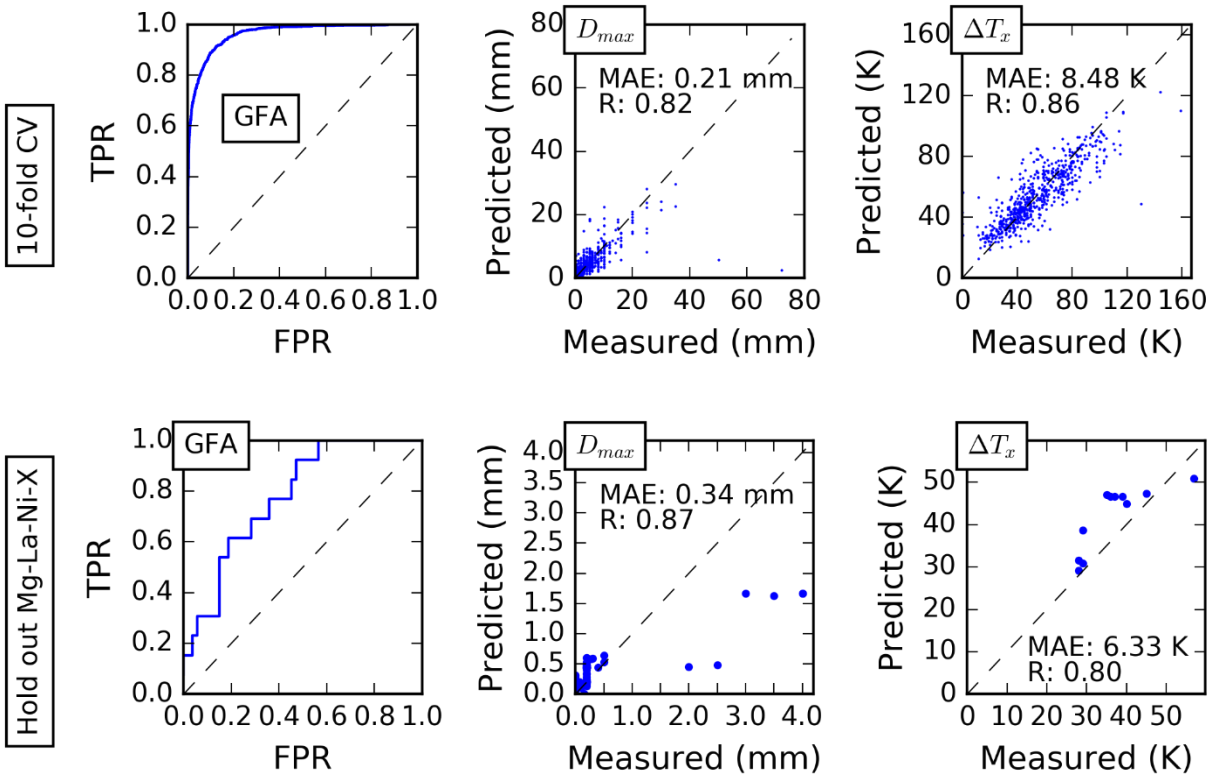


Figure 7.2. Performance of machine learning models designed to predict the glass-forming ability (GFA), critical casting diameter (D_{max}), and supercooled liquid range (ΔT_x) evaluated using two different cross-validation tests. The upper panel shows the algorithm performance in 10-fold cross validation. The bottom panels are the results from test where all data from the Mg-La-Ni ternary or any derivative quaternary was held out as a test set. The GFA classification model was characterized using a Receiver Operating Characteristic (ROC) curve, which shows the True and False Positive Rates of labelling metallic glasses as a function of the threshold at which an entry is labeled "glass-forming." The D_{max} and ΔT_x charts show the values of the experimentally-measured properties and machine-learning-predicted values for each entry in the dataset.

approximately 90%. We also found good predictive accuracy of our model for D_{max} and ΔT_x regression models with correlation coefficients greater than 80%, as shown in Figure 7.2.

To study the ability of our model to predict properties in alloy systems not included in the training set, we withheld all data from the Mg-Ni-La ternary system and any derivative quaternaries to use as a test set. We chose this system in particular because the D_{max} and ΔT_x

for all BMGs were assessed by the same research group using the same casting technique with the same quality materials.[237,238] In this test, our regression models produced similar levels of accuracy to those reported in the cross-validation test (see Figure 7.2). Additionally, out of the 66 alloys in the test set, the 20 where the classification model was most confident that the alloy would be glass-forming were all indeed metallic glasses. From this test, we conclude that our model is able to predict the properties of BMGs and glass-forming ability in yet-unstudied systems.

7.4.2 Optimizing Existing and Discovering New BMG Alloys

We applied our models to two separate design tasks: optimizing the properties of existing alloys, and identifying BMGs in yet-unstudied alloy systems.

7.4.2.1 Tuning Established Zr-based Alloys

The first application we considered was using our models to optimize the performance of two established alloys that are known to be viable with our target processing method: LM105 ($Zr_{52.5}Ti_5Cu_{17.9}Ni_{14.6}Al_{10}$) and LM601 ($Zr_{51}Cu_{36}Ni_4Al_9$). Here, our goal is to find small adjustments to the alloy compositions that lead to alloys with improved casting diameters or supercooled liquid range. The search space we defined for these two alloys is shown in Table 7.1.

Table 7.1. Composition ranges considered for each element varied for the optimization of LM105 and LM601. Between both tests, only the ranges of acceptable compositions changed and not the step sizes. All values are in at. %.

	LM105		**LM106**		
Element	*Minimum*	*Maximum*	*Minimum*	*Maximum*	*Step Size*
Zr	47	60	45	65	0.5
Ti	2	8	0	20	0.5
Cu	12	24	0	30	0.5
Ni	0	30	0	30	0.5
Al	5	15	0	20	0.5

From the several million possible candidates, we identified the alloys that were predicted to have the optimal balance of D_{max} and ΔT_x using a Pareto analysis.[275] This search required only a few CPU-days of computer time, but would require decades to perform experimentally. As shown in Figure 7.3, this technique allowed us to identify alloys where no other alloy is predicted to have better performance by both metrics. In total, we identified 6 Pareto-optimal alloys for further testing, which are listed in Table 7.2.

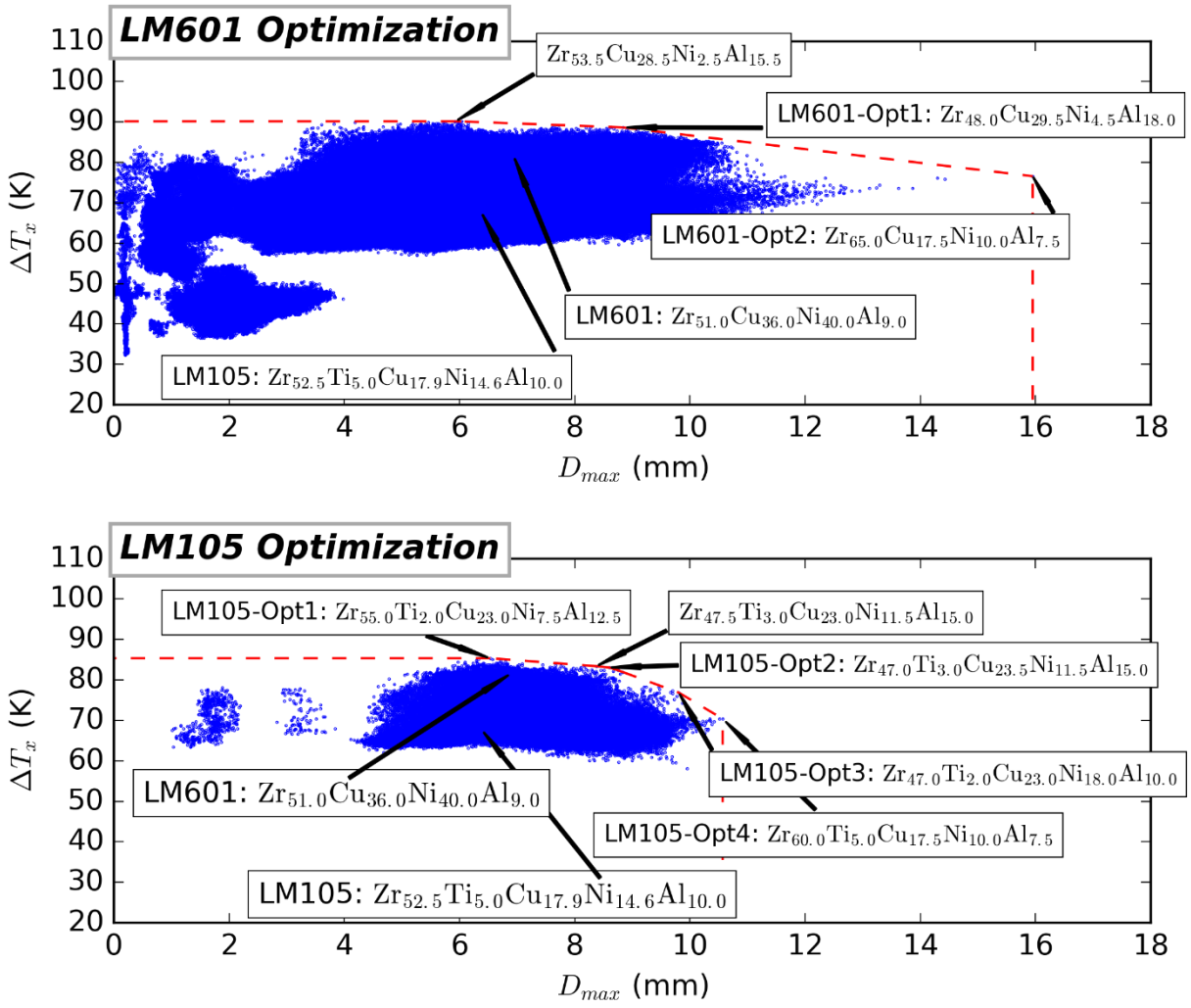


Figure 7.3. Machine-learning-predicted properties of alloys evaluated during the optimization of two established BMG alloys: LM601 (top) and LM105 (bottom). The red, dashed line in each plot represents the Pareto surface of the predicted alloys, which was used to identify alloys with optimal levels of critical casting diameter (D_{max}) and supercooled liquid range (ΔT_x). The properties, names, and compositions of alloys tested in this work are labeled with arrows.

We were able to successfully cast all alloys using arc melting and an Engel e-motion 110 Liquidmetal® injection molding machine. For each alloy, we determined the Strength Limited Casting Thickness (SLCT). The SLCT relates to the rod diameter at which the yield strength of the alloy begins to degrade from the fully-amorphous value, which need not be related to the casting diameter at which crystallization starts to occur. As shown in Figure 7.4, we found little

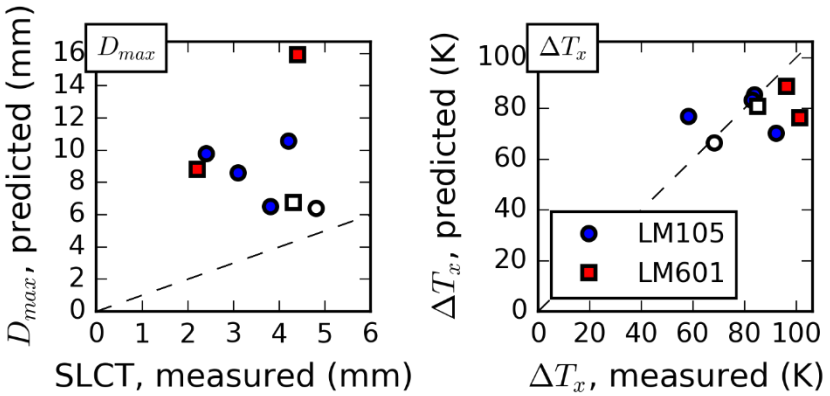


Figure 7.4. Comparison of measured and predicted values of the critical casting diameter and supercooled liquid ranges of alloys tested in this work. Filled points represent materials predicted by our machine learning model. Hollow points represent the base alloys. Materials with improved properties are located on the right side of each chart.

to no correlation between the predicted D_{max} of an alloy and the measured SLCT. Even so, all of the alloys achieved an SLCT of above 1 mm and several had values comparable to the original alloys.

Table 7.2. Measured and predicted properties of alloys evaluated in this work. Alloys designated with "-Op" are optimized compositions predicted in this work.

Name	Composition (at. %)					ΔT_x (K)		SLCT (mm)	D_{max} (mm)
						Measured	*Predicted*	*Measured*	*Predicted*
LM601-Op1	Zr	Cu	Ni	Al		96.1	88.8	2.2	8.8
	48.00%	29.50%	4.50%	18.00%					
LM601-Op2	Zr	Cu	Ni	Al		101.3	76.6	4.4	16.0
	65.00%	17.50%	10.00%	7.50%					
LM601	Zr	Cu	Ni	Al		85.0	81.1	4.3	6.8
	50.75%	36.23%	4.03%	9.00%					
LM105-Op1	Zr	Ti	Cu	Ni	Al	83.8	85.5	3.8	6.5
	55.00%	2.00%	23.00%	7.50%	12.50%				
LM105-Op2	Zr	Ti	Cu	Ni	Al	82.8	83.2	3.1	8.6
	47.00%	3.00%	23.50%	11.50%	15.00%				
LM105-Op3	Zr	Ti	Cu	Ni	Al	58.8	76.9	2.4	9.8
	47.00%	2.00%	23.00%	18.00%	10.00%				
LM105-Op4	Zr	Ti	Cu	Ni	Al	92.3	70.2	4.2	10.6
	60.00%	5.00%	17.50%	10.00%	7.50%				
LM105	Zr	Ti	Cu	Ni	Al	68.3	66.5	4.8	6.4
	52.50%	5.00%	17.90%	s14.60%	10.00%				

The low correlation between the predicted D_{max} and measured SLCT of our alloys could have several explanations. First, the differences could be a result of statistical errors of the model. Second, while both SLCT and D_{max} describe the maximum casting thickness of an alloy, it has not been established whether they are correlated. Third, the casting technique used to test our prediction was different than what was used in the literature. Further experimental work would be required to determine which of these effects, if any, are the most important in understanding the limitations of our model. Consequently, while we are unable to validate the ability of our model to predict D_{max}, we can at least conclude that it is not a strong predictor of SLCT. However, it was sufficiently accurate enough that our predicted alloys each had acceptable critical casting diameters.

We also measured the ΔT_x of each alloy using Differential Scanning Calorimetry (DSC) and found several alloys that outperformed our base materials. As shown in Figure 7.5, both of our new LM601 variants exceeded the base alloy in this property, and 3 of the 4 new LM105 variant alloys had larger supercooled liquid ranges than the base material. We also observed a weak (R = 0.33) correlation between the measured and predicted values for this property. Considering the large variation in this property between different reports of the same alloys in the literature[242,243] and small sample size in our validation test, a lower-than-expected correlation coefficient is not surprising. The mean absolute error of our predictions, 10.1 K, is also only slightly worse to what was measured in the cross-validation tests. From these results, we conclude that our model is suitable for optimizing the ΔT_x of our alloys.

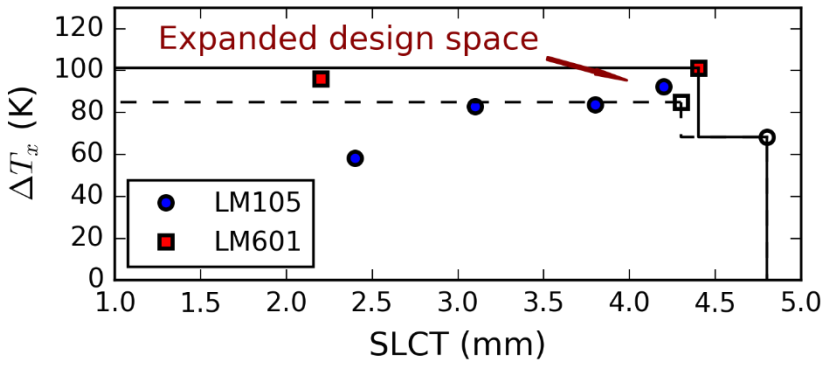


Figure 7.5. Comparison of the measured supercooled liquid range (ΔT_x) and strength limiting casting thickness (SLCT) of alloys predicted in this work. The unfilled shapes represent the base alloys and the filled shapes the alloys predicted in this work. The dotted line indicates the Pareto surface of the original alloys. Three of our new alloys exceed this boundary, and create a larger region (solid line) where it is possible to create an alloy that exceeds certain minimum values for both properties.

Overall, our new alloys expand the design space for Zr-based metallic glasses. While no new material exceeds both of the known alloys in both properties, we were able to locate materials with improved ΔT_x and similar critical casting diameters (see Figure 7.5). In doing so, we expanded the space of critical diameters and supercooled liquid ranges where there is a known alloy that exceeds both design properties. This demonstrates the power of machine learning methods to interpolate between known materials and locate new materials with properties tailored for different requirements.

7.4.2.2 Locating Novel Bulk Metallic Glasses

The second application we explored was using our models to find alloys that are different from known BMGs, and have large critical casting diameters and supercooled liquid ranges. This application is more difficult than the previous example because we no longer constrain the search space to a region known to be compatible with injection molding. For this test, we decided to evaluate alloys created using all combinations of elements out of a list of 27 suitable

for commercial production (see Table 7.3). From this list, we tested all combinations of two or three elements at 2 at% spacing, which yields a total of 3.5 million candidate materials.

Table 7.3. List of elements included in search for new alloys

ELEMENTS CONSIDERED IN ALLOY SEARCH						
Mo	W	Ta	Hf	Ti	Zr	P
Ca	Li	Mg	Nb	Zn	Si	B
Co	Fe	Mn	Cr	V	Pb	C
Sn	In	Ag	Cu	Al	Ni	

Search Strategy: Our procedure for finding new alloys was split into four separate screens. First, we removed any alloys where the L_1 distance between the composition of that alloy and any known ribbon or bulk-forming glass was less than 30 at%. We computed the L_1 distance by summing the difference between the fractions of each element in the two alloys (e.g.: the distance between $Cu_{50}Zr_{50}$ and $Cu_{50}Zr_{46}Al_4$ is 8 at% by this measure). Next, we used the classification model to eliminate alloys that are predicted to have greater than a 5% chance of not being able to form a glass. Then, we used the regression models to locate alloys with a ΔT_x above the 80th percentile of the training data (66 K) and a D_{max} greater than 1 mm.

Search Results: The search space included 3.5 million alloy candidates, which required about 2 days on 8 2.2 GHz processors to fully evaluate. Only 38361 (approximately 1%) alloys passed all of the filters. We then sorted each ternary system based on the number of alloys that passed all filters, as a proxy for the size of the glass-forming region. Out of these top alloys, we selected the Cu-Hf-Mg and Cu-Hf-Ti systems. Finally, we enumerated alloys in these two systems at a much finer spacing (0.5 at%) and identified alloy compositions within these systems with the highest predicted D_{max} and ΔT_x (see Figure 7.6).

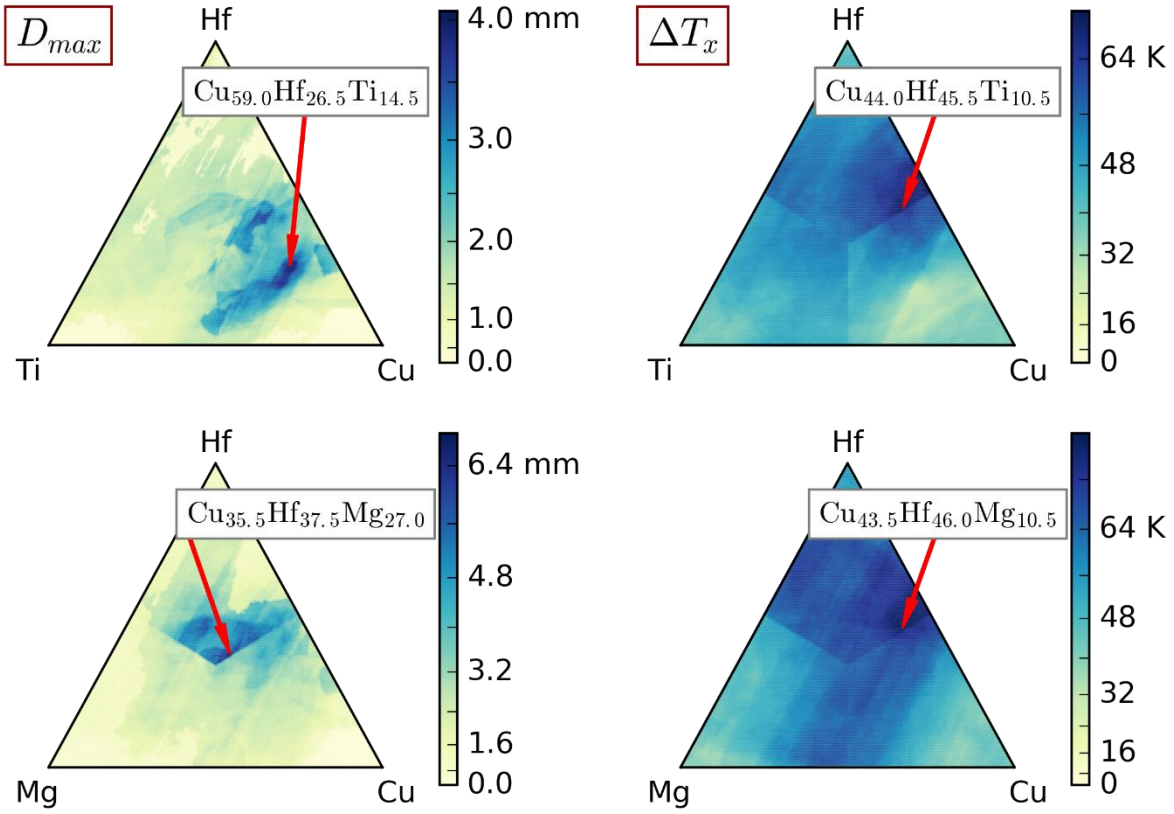


Figure 7.6. Critical casting thickness (D_{max}) and supercooled liquid ranges (ΔT_x) of Cu-Hf-Ti and Cu-Hf-Mg alloys, as predicted using a machine learning model. Arrows indicate the compositions with the maximum value of a property in each of the ternary diagrams.

As in the previous application, we attempted to fabricate these materials using arc melting and injection molding. Of the four alloys, we were only able to successfully melt and cast the Cu-Hf-Ti alloys, and were unable to successfully alloy the Mg into the Cu-Hf-Mg alloys. In the case of the Cu-Hf-Mg alloys, the Mg evaporated before becoming incorporated into the melt. For the alloys that were able to be cast (those in the Cu-Hf-Ti system), we were unable to form fully-amorphous samples in the injection molding machine.

In this application, we found processing requirements to be an important factor when identifying new alloys and have several recommendations for how to incorporate that

information. First of all, we recommend limiting the search space based on the target fabrication process. In contrast to the screening steps employed earlier in this work, these filters would be based on human knowledge about which materials work best for each casting step. For example, if we removed high vapor pressure elements (e.g., Mg) from the search space, we could have selected alloys better suited for arc melting.

Secondly, future searches for new BMGs should consider processing conditions as input into the machine learning models. While we found $Cu_{59}Hf_{26.5}Ti_{14.5}$ could not be formed with injection molding, previous experiments have reported that a very similar alloy, $Cu_{60}Hf_{25}Ti_{15}$, could be made using copper mold casting.[276] This variation in processability demonstrates the strong impact of fabrication method on observed glass-forming ability. The first step in incorporating processing conditions into the machine learning model would be to gather examples of metallic glasses created with several different methods. Next, processing conditions need to be included as input into the machine learning model. The simplest method for introducing processing conditions as inputs would be adding a categorical variable that indicates the method used when attempting to fabricate a BMG. Provided enough examples of materials that were able or were not able to be created using each processing method, a machine learning model could predict both whether it is possible to create a glass at a certain composition and which processing technique would be the most effective.

7.5 Conclusion

In this chapter, we presented a set of machine-learning-based design tools for Bulk Metallic Glasses (BMGs) and initial work on validating these predictions using commercially-viable

fabrication methods. To create these tools, we developed new attributes based on empirical rules from the metallic glass community and used them to create machine learning models that link the composition of an alloy to its glass forming ability, critical casting diameter, and supercooled liquid range. With these new tools, we optimized existing BMGs to create new alloys with superior supercooled liquid ranges and comparable critical casting diameters to the original alloys. Our attempts to discover BMGs in new alloy systems were unsuccessful, which we suggest could be remedied by incorporating processing method into the machine learning model. We envision that this framework, with modification, could lead to the discovery of new BMGs and provide a path for engineering these materials for new applications.

8 Magpie: A Materials-Agnostic Platform for Informatics and Exploration

8.1 Abstract

Machine learning methods have started to show significant promise as a route for designing new materials. However useful, the adoption and widespread use of these techniques is limited by the lack of publically-available software. In this work, we present an open-source software package, Magpie, designed to simplify the use of machine learning for materials engineering by providing a user-friendly interface, mechanisms for sharing models, and the ability to be easily extended to include new methods. In this chapter, we describe the key features of Magpie, its user interface, and give a worked example of using this code to model metallic glass formation. It is our goal for this tool can be employed by the community at large in order to make machine learning models a readily-accessible tool for materials design.

8.2 Introduction

Free and open-source software are widely-used in many areas of computational materials engineering. Considering atomic-scale modelling, there are the popular LAMMPS and Quantum-ESPRESSO codes that allow users to perform state-of-the-art molecular dynamics and DFT calculations.[277,278] There is also an open-source code for performing computational thermodynamics and a variety of software for performing phase field simulations.[279,280] Having access to open source codes not only lowers the barrier to accessing these powerful methods, but also improves the reproducibility of studies performed using these techniques.[281] However, there are few studies that employ machine learning to study or

design materials – a field often called “Materials Informatics” [3] – that release the software necessary to recreate or reuse their methods.

As noted in a study by Bhadeshia *et al*, the dissemination of neural networks models used for materials science applications is infrequent.[94] By analyzing 100 papers that published a neural network model, they found that well over half did not publish the requirements necessary to recreate models: the training data and coefficients of the neural network. Considering the wide breadth of machine learning methods besides neural networks in the literature, we also assert that the software that performs the machine learning algorithm needs to be released as well. Fortunately, there are several positive examples of papers that are at least partially open. Several papers include links to online tools where one can use the model for new tasks.[93,116,202,282] There are also a few open source codes that have been released recently,[105,283] and papers that release their training data along with the study.[24,79,163] However, openness is certainly not the norm.

In this chapter, we describe the development of an open-source software tool, Magpie, designed to simplify the creation of machine learning models from materials data. The primary motivation when developing Magpie was to create a library of materials informatics techniques that can be used by the non-expert. Additionally, Magpie is also designed to simplify the sharing of models either through operating-system-independent files or custom web applications. In the following sections, we describe the key features of Magpie as well as demonstrate how it can be applied to building a model for metallic glass formation. It is our view that these features

will make creating, employing, and sharing machine learning models from materials data easier for both experts and new users of materials informatics.

8.3 Technical Details

Magpie, which is an abbreviation for "Materials Agnostic Platform for Informatics and Exploration," is a Java Library with a custom command-line interface and ability to be integrated to other codes using a flexible API. Magpie is written to be compatible with Java 8, which is available for most operating systems and computer architectures. One advantage of using Java is that Magpie can be compiled once and run on any other architecture that supports Java. The source code for Magpie is available on BitBucket under an MIT License, and is also distributed with over 400 unit tests to ensure that Magpie is working as originally designed.

Most of the numerical, statistical, and machine-learning algorithms employed by Magpie are provided by linking to other libraries. The Apache Commons Math library is used to perform many of the analyses of model performance and a few of the simpler machine learning algorithms (e.g., linear regression). Magpie also links to the Weka library and has the capability to train and run models created with scikit-learn, which allows access to dozens of modern machine learning algorithms.[74,75]

Magpie also contains the ability to perform some computations in parallel. The Java programming language natively supports shared-memory parallelization, which we employ to allow generating representations, running models, and performing high-throughput screening in parallel. As emphasized in a later section, many calculations performed when using a machine learning model are embarrassingly parallelizable and we can therefore achieve high

parallel efficiency. Work is currently underway to support distributed memory parallelization via Apache Spark, which will allow Magpie to be deployed on high-performance clusters.

Besides interacting with Magpie via its text interface or the Java API, one can also create tools in other programming languages via a REST API. The central component of the REST API for Magpie is a server that holds all information necessary to run a collection of materials informatics models. Other programs can then submit simple commands (e.g., “run formation energy model for X, Y, Z materials”) via HTTP web requests to this server, which will automatically perform them and send back data in a JSON format. As many programming languages support generating HTTP requests and parsing JSON files, it is possible to create software that utilize models running in Magpie in many different languages. For example, the websites described in Section 8.5.3 are built using the REST API.

8.4 User Interface

The main user interface for Magpie is a text-based input file loosely modeled after the Bash programming language. Users can create variables that represent components of a machine learning model, such as a dataset to store training and test data, and interact with them via simple commands. An example of reading in data from a text file, training a model with Weka, and saving the model in an operating-system-independent format is shown in Figure 8.1. Most typical machine learning tasks, such as performing cross-validation and running the model on a search space, can be performed from this interface with only a few commands. A worked example of a script to build a machine learning model through this interface is available in a later section.

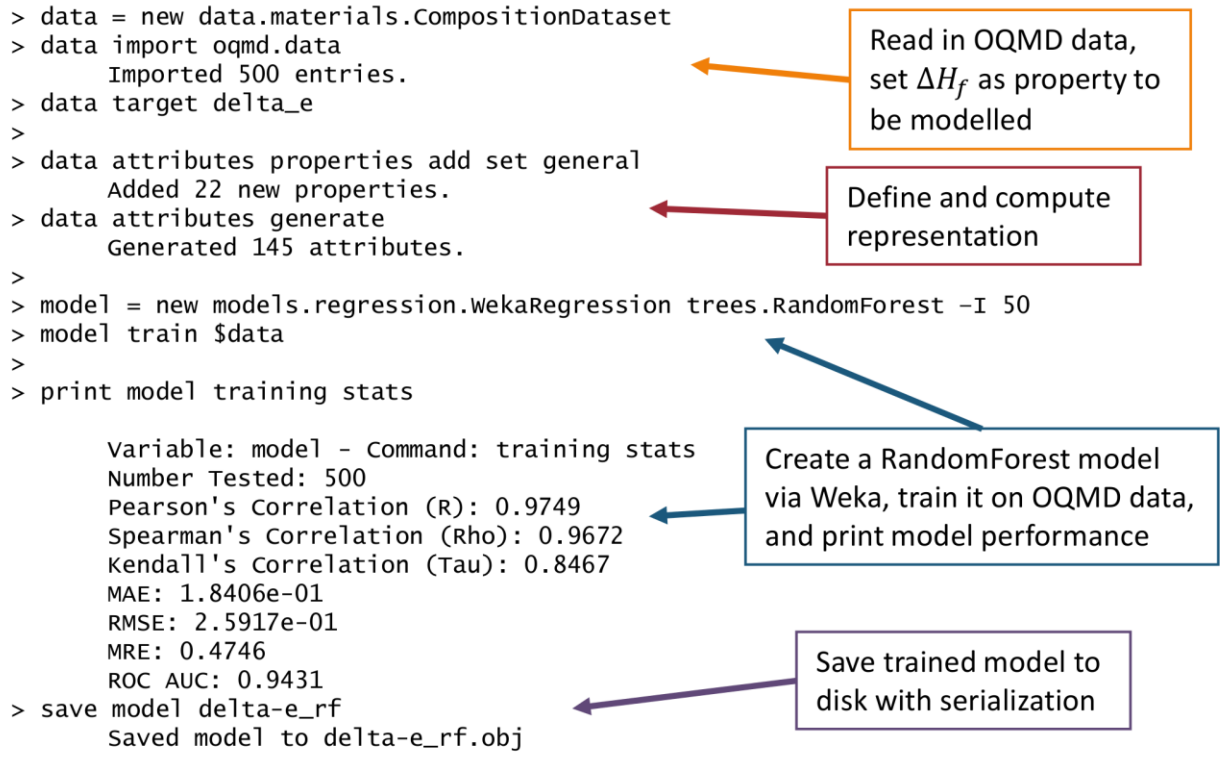


Figure 8.1. Example input and output from Magpie for a script that creates a model to predict the formation enthalpy of materials given data from the OQMD. Users can create variables that store the dataset and model object, and manipulate them with simple text commands.

While the command line interface has the advantage of running without installing any additional software, it is limited in the kind of operations that can be performed. For example, it does not support loops for performing parameter sweeps often used to optimize a model. However, as this software is based in Java, you have several other options to create more complicated analysis software. For example, one can link to Magpie from other Java code or create scripts using the Scala programming language. Doing so requires either compiling new software or having Scala installed on a particular system, but they do offer more advanced functionality to expert users.

8.5 Key Features of Magpie

The following subsections contain information about the features of Magpie which are likely to be used by the largest audience.

8.5.1 Library of Material Representations

The most important feature of Magpie is a library of methods to create representations of materials. The purpose of computing these representations is to transform raw materials data (e.g., the fraction of each element in the material) into a form that reflects the physical effects leading to observed material properties, which makes it possible to create simpler and more accurate models.[79,200] Representations are composed of a set of quantitative attributes, and the development of these attributes has been a subject of much research over the past decade.[41,79,80,160,200,284] One of our goals when creating Magpie is to implement these methods all in a single library so that it will be easier to use them for new problems.

As of the time of writing this manuscript, Magpie contains a diverse set of methods for representing materials based on their composition and crystal structure. For example, the attribute sets used by Meredig *et al*. to predict the formation enthalpy of crystalline materials,[85] those employed by Ward *et al*. to predict the glass-forming ability of metal alloys,[200] and many of the attributes developed by Deml e*t al*. to predict the total energy of ionic compounds are all included in Magpie.[41] Furthermore, Magpie contains the Atomic-Property-Weighted Pair Radial Distribution Function (AP-RDF) method and the Coulomb Sine Matrix representations for crystal structures.[80,285] Our future vision is to add in existing

attribute sets (e.g., microstructure representations [286,287]) into Magpie and continually upgrade it to contain new techniques as they are developed.

8.5.2 Linkages to Machine Learning Libraries

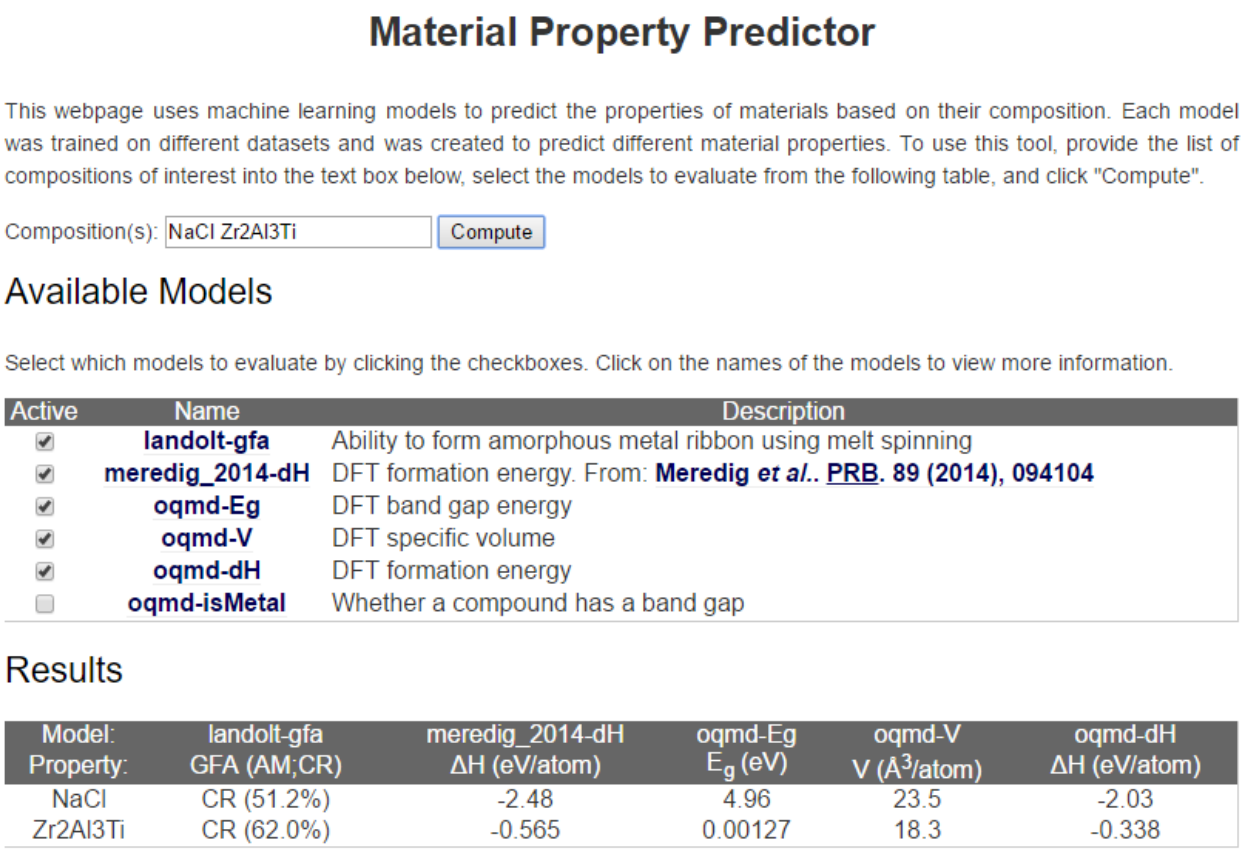
Besides raw materials data and a numerical representation for that data, the other key ingredient of a materials informatics method is a machine learning algorithm. Our approach for allowing users access to modern machine learning algorithms is to directly link to two extensive libraries: Weka and Scikit-Learn.[74,75] By providing a link to other libraries, it will be possible for Magpie users to take advantages of continual improvements and additions to these libraries and, therefore, be able to utilize state-of-the-art machine learning methods as they become available. Magpie does contain custom code for some machine learning techniques, including the LASSO-based feature selection technique used by Ghiringhelli *et al*. and the cumulant expansion method employed by Fischer *et al*.[79,116] However, in general, our guiding philosophy is to use existing code where possible.

8.5.3 Easy Sharing of Models

Magpie provides users three different methods for sharing their models, each to target a different audience. First of all, it is possible to share the input script and training data files used to generate and test the model. Secondly, one can save the model and dataset objects, which contain the machine learning model and code to generate the representation for new data, into an operating-system independent file format with the serialization mechanism of the Java programming language. These two techniques require a user to install and be able to run the

Magpie software and, therefore, require some familiarity with the principles of machine learning to be useful.

To share models with the widest-possible user base, Magpie also provides software to simplify the creation of custom web applications. Figure 8.2 shows an example webpage where users can run composition-based machine learning models through a simple form. This web tool also links to another webpage (shown in Figure 8.3) that describes the provenance of the model, which includes when the model was created, a description of the training data and validation statistics, and a list of recommended citations if this model is used in an article.



Material Property Predictor

This webpage uses machine learning models to predict the properties of materials based on their composition. Each model was trained on different datasets and was created to predict different material properties. To use this tool, provide the list of compositions of interest into the text box below, select the models to evaluate from the following table, and click "Compute".

Composition(s): NaCl Zr2Al3Ti Compute

Available Models

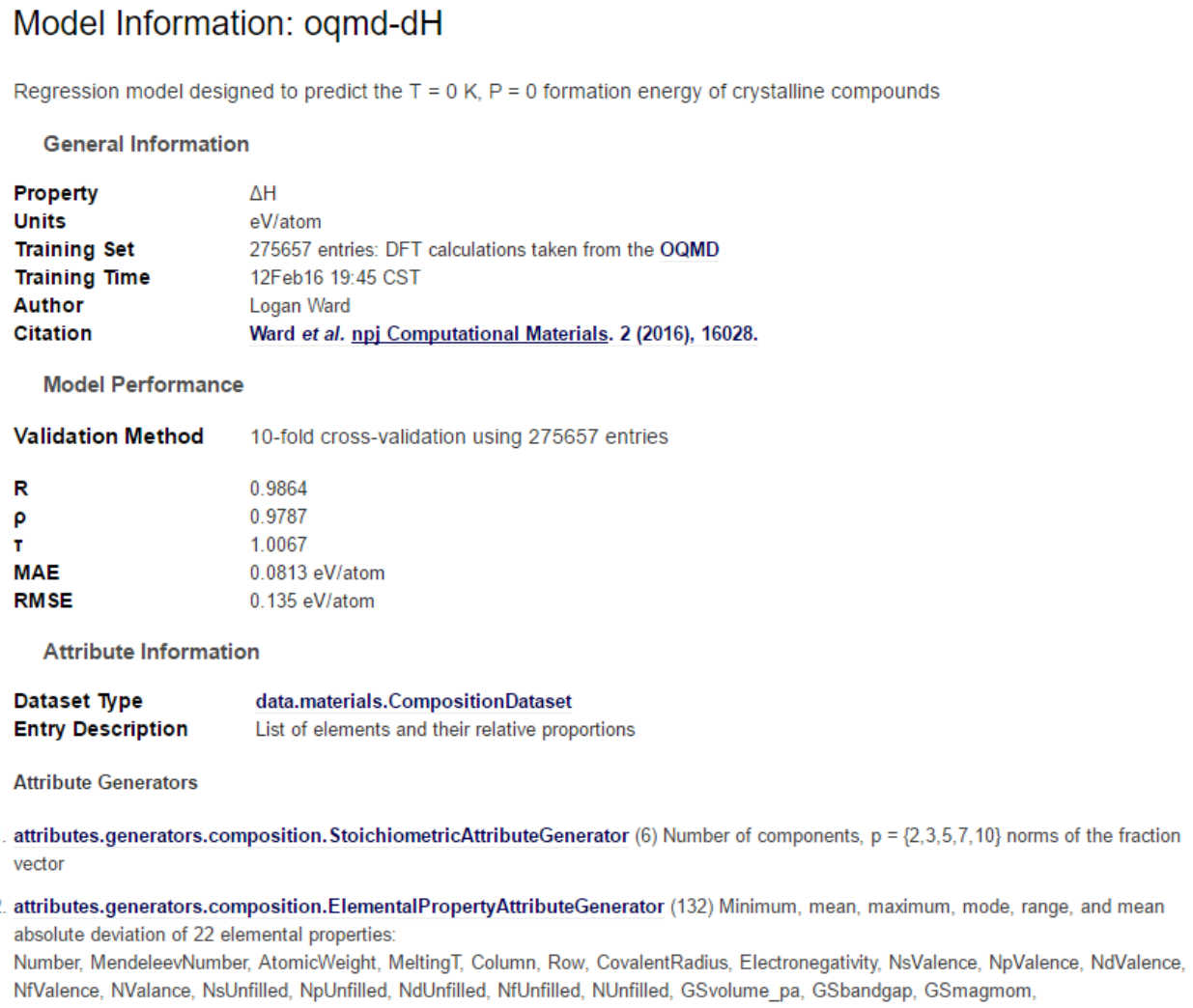
Select which models to evaluate by clicking the checkboxes. Click on the names of the models to view more information.

Active	Name	Description
☑	**landolt-gfa**	Ability to form amorphous metal ribbon using melt spinning
☑	**meredig_2014-dH**	DFT formation energy. From: **Meredig *et al.*. PRB. 89 (2014), 094104**
☑	**oqmd-Eg**	DFT band gap energy
☑	**oqmd-V**	DFT specific volume
☑	**oqmd-dH**	DFT formation energy
☐	**oqmd-isMetal**	Whether a compound has a band gap

Results

Model: Property:	landolt-gfa GFA (AM;CR)	meredig_2014-dH ΔH (eV/atom)	oqmd-Eg E_g (eV)	oqmd-V V ($Å^3$/atom)	oqmd-dH ΔH (eV/atom)
NaCl	CR (51.2%)	-2.48	4.96	23.5	-2.03
Zr2Al3Ti	CR (62.0%)	-0.565	0.00127	18.3	-0.338

Figure 8.2. Sample webpage for interacting with materials informatics models through a simple interface. Users input compositions into a text box, select which models they want to run from a list of checkboxes, and then click compute. All necessary steps to evalaute the models are performed automatically and the results are displayed in a clear, tabular format.



Model Information: oqmd-dH

Regression model designed to predict the T = 0 K, P = 0 formation energy of crystalline compounds

General Information

Property	ΔH
Units	eV/atom
Training Set	275657 entries: DFT calculations taken from the OQMD
Training Time	12Feb16 19:45 CST
Author	Logan Ward
Citation	Ward *et al.* npj Computational Materials. 2 (2016), 16028.

Model Performance

Validation Method	10-fold cross-validation using 275657 entries
R	0.9864
ρ	0.9787
τ	1.0067
MAE	0.0813 eV/atom
RMSE	0.135 eV/atom

Attribute Information

Dataset Type	data.materials.CompositionDataset
Entry Description	List of elements and their relative proportions

Attribute Generators

1. attributes.generators.composition.StoichiometricAttributeGenerator (6) Number of components, p = {2,3,5,7,10} norms of the fraction vector
2. attributes.generators.composition.ElementalPropertyAttributeGenerator (132) Minimum, mean, maximum, mode, range, and mean absolute deviation of 22 elemental properties:
   Number, MendeleevNumber, AtomicWeight, MeltingT, Column, Row, CovalentRadius, Electronegativity, NsValence, NpValence, NdValence, NfValence, NValance, NsUnfilled, NpUnfilled, NdUnfilled, NfUnfilled, NUnfilled, GSvolume_pa, GSbandgap, GSmagmom,

Figure 8.3. Example webpage that displays proveance information about a machine learning model. This automatically-generated page contains information about when the model was created, who to cite when using it, validation information, and descriptions of how to reproduce the model.

Additionally, this webpage contains information about how to recreate this model, including an automatically-generated description (see Figure 8.4) of the attributes and machine learning algorithm used to create the model. Magpie is distributed with example web pages so that creating these tools for new models is easier. Our vision that by creating these model sharing tools, there will be little need to recreate the same infrastructure for each new model.

8.5.4 Proper Attribution

It is our view that researchers should be recognized for their contributions to the methods employed in Magpie regardless of whether they implemented the code used in Magpie or not. In order to make it easier for users to recognize who developed the methods employed by a model, we have designed the ability to add a "get citations" command to any object in Magpie. If implemented, Magpie can automatically prepare a list of recommended citations. Upon calling the "citations" command through the user interface, information about these citations for each variable used in the current script is printed to screen (see Figure 8.4). With this functionality, we can make sure any users of Magpie are properly informed about the methods they are using and have the information necessary to acknowledge contributions from others when, for example, citing papers in a journal article.

```
> citations
Suggested citations for selector:

        Reason: Introduced the LASSO attribute selector
        Component: magpie.attributes.selectors.LassoAttributeSelector
        Type: Article
        Authors: L. Ghiringhelli, et al.
        Title: Big Data of Materials Science: Critical Role of the Descriptor
        URL: http://link.aps.org/doi/10.1103/PhysRevLett.114.105503

> print selector description

Variable: selector
Uses compressed sensing (LASSO) and linear regression to identify a subset of
attributes that make the best linear model. First, a 10 attribute model is created
with LASSO and these attributes are used as a starting set. The final subset of 1 to
4 attributes is selected based on which, out of all possible subsets, generates a
linear regression model with the lowest mean squared error over 100 iterations of a
hold-10.0%-out cross-validation test.
```

Figure 8.4. Example of the functionality in Magpie for automatic generating recommended citations and human-readable descriptions. This example output from Magpie shows recommended citations and a description for a tool using the attribute selection technique of Ghiringhelli *et al*.[79] Color added for clarity.

8.5.5 Efficient High-Throughput Screening

One of the common uses of machine learning tools is to evaluate exceptionally large numbers of candidate materials, and Magpie is designed to simplify that process. For one, Magpie includes the methods to generate large search spaces. For example, it contains code to generate compositions given a list of possible constituent elements or generating new crystal structures based on a user-defined prototype.

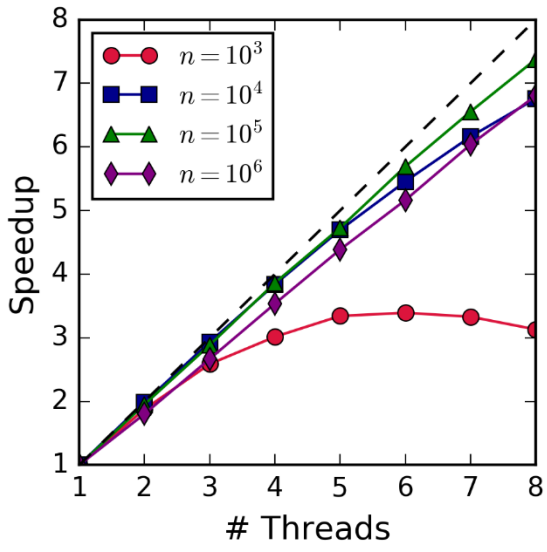


Figure 8.5. Parallel performance of Magpie in performing a combinatorial search for new metallic glass alloys. Speedup factor (ratio between runtime and serial calculation time) shown as a function of number of threads for different batch sizes (n).

Magpie also contains methods for evaluating large search spaces efficiently. As each entry can be evaluated individually, computing the entire search space can be trivially parallelized. Magpie can generate the representation for each material and run the models in parallel using shared-memory parallelization. We also support processing the data in batches because storing the representation of the entire search space in memory could be impractical, due to memory constraints, and is unnecessary. For example, the search in Ref. [200] would require over 32 GB of memory to run if the representation for all entries were computed at once ($26 \text{ million entries} \times 145 \frac{\text{attributes}}{\text{entry}} \times 8 \frac{\text{Bytes}}{\text{attribute}} \approx 32 \text{ GB}$). By selecting an appropriate batch size (trading off between efficiencies gained by processing larger batches against more-equitable division of work between threads), the parallel efficiency can be above 90% - as

shown in Figure 8.5. Using both the parallel computing and batch processing methods built in to Magpie, one can quickly assess search spaces with more than 10^7 entries.

8.5.6 Extensibility

Magpie is also designed to be easily extensible in the future. Each major type of variable (e.g., machine learning model, dataset filter) is based on a supertype that contains all of the reusable parts of the code. For example, someone creating a new filter need only implement the method to label whether each entry in a dataset passes the filter. The other, generic code for actually removing the non-passing entries and performing labelling in parallel can be automatically reused. The documentation describing which operations need to be implemented for each variable type and what those operations should perform are provided in the online documentation. Following this object-oriented software design, we hope Magpie can easily be updated by a large community.

8.6 Worked Example: Predicting New Metallic Glasses

To give a demonstration of an example use case of Magpie, we will describe how to replicate the machine learning model for predicting metallic glasses described in Ref. [200]. The starting point for this study is a simple data file, where the composition of each material in the dataset is listed along with whether it was measured to be able to form a completely-amorphous sample (labelled AM), a mix between amorphous and crystalline phases (AC), and only crystalline phases (CR). The first few lines of this file are

```
comp gfa{AM,AC,CR}
Ag20Al25La55 AM
Ag15Al10Mg75 AM
Ag25Al10Mg65 AM
```

where the first line defines measured property (gfa), followed by its possible values (AM, AC, and CR). To store this data, one uses the CompositionDataset object, which stores data where entries can be differentiated based on their composition. Creating the object to store the data and reading in the dataset are accomplished by two lines

```
data = new data.materials.CompositionDataset
data import landolt.data
```

where the first line creates the object stores it to the variable named "data" and the second reads in the text file. The first word in each line for the text interface to Magpie is either a command word (e.g., "save") or the name of a variable followed by a series of words that define the operation and some of its options. For this example, "import" is the command for reading in data from the filesystem and "landolt.data" is the only option – the path to the data. All other commands in Magpie follow this structure.

The next step in creating the model is cleaning the data. The first step is to define "gfa" as the property to be modeled with the command

```
data target gfa
```

The second step is to reduce the number of possible class from three to only two: was a material fully-amorphous or not. This modification is accomplished by invoking the "ClassEliminationModifier" class to replace all examples entries labeled as "AC" to be labelled as "CR." To do so, one employs the "modify" command of the dataset object

```
data modify ClassEliminationModifier AC CR
```

where the options to the command are the name of the modifier class to be run and any options for that modifier. Magpie contains many such "modifier" classes to perform other data cleaning and processing algorithms. The third step is to remove duplicate compositions. If duplicates for a single composition have different labels (e.g., some are labelled crystalline and others glassy), the authors assigned a label of AM (amorphous) to the entry. This is accomplished by calling the "duplicates" command of the dataset and telling the duplicate resolution code to select the entry with the minimum value of the "gfa" property (corresponding to the first label in the data file, AM). This duplicate resolution strategy can be enforced with the command

```
data duplicates RankingDuplicateResolver minimize &
  PropertyRanker gfa SimpleEntryRanker
```

where the "&" breaks the command on to two different lines for readability. After invoking these two commands, the dataset is reduced by 1467 entries and the machine learning problem is simplified into a binary classification problem.

Once the data is cleaned, the next step is to generate the representation that serves as input into the machine learning model. In this work, the authors employed the default representation for composition data in Magpie – the attribute set proposed by Ward *et al*.[200] For this, one simply needs to specify the list of elemental properties (e.g., electronegativity) that are used to generate the representation with the command

```
data attributes properties add set general
```

where the general set is the 22 properties employed by Ward *et al*.[200] Then, to compute all 145 attributes, the command is

```
data attributes generate
```

Computing the attributes takes only a few seconds and, after it is complete, it is possible to start building machine learning models.

The final step is to train and validate a machine learning model. The first step in this process is to create a new variable to store the model object. To create a RandomForest model via Weka, this is accomplished by the command

```
model = new models.classification.WekaClassifier &
trees.RandomForest -I 100
```

that creates a model with 100 individual trees. The "train" command for this object invokes the training operation with

```
model train $data
```

which takes a dataset variable as an input option. The fact that the word "data" should correspond to a variable in Magpie is indicated with the $ in front. The next step is to save the model and a copy of the dataset, which contains the information needed to compute the representation, using the serialization feature of Java

```
save model gfa-model_model
save data gfa-model_data template
```

The "save" command in Magpie takes up to three arguments: the name of the variable being saved, the file name, and the format. These commands will produce two files, gfa-model_model.obj and gfa-model_data.obj, that can be read in with the load command

```
model = load gfa-model_model.obj
```

on any other operating system. These model files can be included in scripts that perform high-throughput screening, loaded in to create interactive webpages, or simply shared as Supplementary Information with a journal article so that anyone can use the exact same model created in the original work.

8.7 Summary

In this chapter, we introduced the Materials Agnostic Platform for Informatics and Exploration (Magpie) – a tool designed to simplify the creation of machine learning models from materials data. We described the user interface of Magpie, highlighted its main features, and then provided an example of how Magpie can be used to create a classification model for identifying glass-forming metallic alloys. In the future, we plan to continually update Magpie with newer materials informatics approaches and envision integrating this software into other computational materials science tools. In this way, we hope that Magpie will help enable the wider-scale use of materials informatics methods and lead to the accelerated development of many new materials.

9 Summary and Outlook

In this thesis, we presented several advancements that will improve the ability of researchers to design new materials with machine learning methods. As described in the, my work focused on two main areas: (i) the use of automated crystal structure solution algorithms to fill in missing information in material property databases, and (ii) developing tools to simplify the development of machine learning models from materials data.

9.1 Automated Crystal Structure Solution

We have shown how to solve incompletely-determined crystal structures with the First-Principles-Assisted Structure Solution (FPASS) method (Chapters 3 and 4). By solving these structures, we added several new compounds to the Open Quantum Materials Database (OQMD), which are now available as candidate materials for any new searches made using this resource. Additionally, any future machine learning models made using the OQMD will be based on more accurate training data. As we have automated the methods necessary to perform these solutions and made the automation tools publically available, it will be easier to continue to populate the OQMD with more information about known crystalline materials.

The most promising route for future work in this area is to continue using the automated solution tools to solve more incomplete structures. There are plenty of examples of compounds in both the Powder Diffraction File and Inorganic Crystal Structure Database whose structures have yet to be determined. Additionally, there are plenty of methodological advancements that would make FPASS more widely applicable. The area I would recommend the most strongly would be devising ways to reduce the number of DFT calculations required to attain a solution.

Given the large numbers of DFT calculations performed in the solution of a structure, one could fit an empirical potential to the previous calculations in order to either speed up the structural relaxation (as in Ref. [288]) or to be used when suggesting candidate crystal structures. Additionally, it may be beneficial to improve some aspects of the optimization algorithm with methods employed by the crystal structure prediction community (e.g., minima hopping [289]). As a long term goal, FPASS could be combined with other crystallographic tools, such as *ab initio* peak indexing software, to make a tool that can be used to automatically solve structures from raw diffraction data with limited human input.

9.2 Machine Learning

In terms of machine learning, we have created new methods for transforming materials data into a form compatible with machine learning (Chapters 5 and 6), demonstrated how to apply those methods to design metallic glasses (Chapter 7), and created open source software to make these methods available to the materials community (Chapter 8). Our vision is that these advancements will simplify applying machine learning methods to new materials problems. For example, we found methods we created for generating representations based on the composition of materials can be applied to many different problems, which reduces the need to create a new representation for each application. We also show how these machine learning models can be used to create new commercial Bulk Metallic Glass alloys. Finally, the open source software we released will reduce the learning curve for using machine learning and amount of infrastructure that needs to be created when applying machine learning to new applications.

In terms of machine learning, a good next step would be to start finding new applications for the techniques developed in this work. There are plenty of wonderful, curated resources of materials data (e.g., created by NIMS, NIST, and ASM) that could be used to create useful machine learning models. On a technical level, the main question I have about employing machine learning in materials design is how to decide when a machine learning model can be used and how much error to expect in each prediction. Both of these questions could be answered with reliable error estimates for individual predictions from a machine learning model, and there has been some research towards this question in materials science. [47,106,163,290] Another impactful task would be to implement more materials informatics methods into Magpie and using Magpie to recreate more machine learning models from the literature. Doing so would make it easier for others to utilize machine learning in their own work, which would only benefit the materials community as a whole.

Another promising area of machine learning research could be the development of representations for other types of materials data. In this thesis, we have shown methods for using the composition or crystal structure of a material to build machine learning models. Other groups have developed methods to construct machine learning models from microstructure data.[286,287] However, these are far from the only types of materials data. One may want to build models that incorporate the thermomechanical processing history. It would be very beneficial to have a collection of techniques to create useful representations from this variety of data types. If such a library of techniques were created, the process to create a machine learning model from *any* materials data would be drastically simplified.

References

[1] OSTP, National Science and Technology Council, Materials Genome Initiative for Global Competitiveness, 2011.

[2] D. Apelian, A. Alleyne, C.A. Handwerker, D. Hopkins, J.A. Isaacs, G.B. Olson, et al., Accelerating Techology Transition: Bridge the Valley of Death for Materials and Processes in Defense Systems, 2004.

[3] K. Rajan, Materials informatics, Mater. Today. 8 (2005) 38–45. doi:10.1016/S1369-7021(05)71123-8.

[4] S.R. Kalidindi, M. De Graef, Materials Data Science: Current Status and Future Outlook, Annu. Rev. Mater. Res. 45 (2015) 171–193. doi:10.1146/annurev-matsci-070214-020844.

[5] S.S. Naghavi, V.I. Hegde, A. Saboo, C. Wolverton, Energetics of cobalt alloys and compounds and solute–vacancy binding in fcc cobalt: A first-principles database, Acta Mater. 124 (2017) 1–8. doi:10.1016/j.actamat.2016.10.065.

[6] T. Angsten, T. Mayeshiba, H. Wu, D. Morgan, Elemental vacancy diffusion database from high-throughput first-principles calculations for fcc and hcp structures, New J. Phys. 16 (2014). doi:10.1088/1367-2630/16/1/015018.

[7] H. Zong, X. Ding, T. Lookman, J. Sun, Twin boundary activated α → ω phase transformation in titanium under shock compression, Acta Mater. 115 (2016) 1–9. doi:10.1016/j.actamat.2016.05.037.

[8] S.M. Woodley, R. Catlow, Crystal structure prediction from first principles, Nat. Mater. 7 (2008) 937–946. doi:10.1038/nmat2321.

[9] Q.-J. Hong, A. van de Walle, Solid-liquid coexistence in small systems: A statistical method to calculate melting temperatures., J. Chem. Phys. 139 (2013) 94114. doi:10.1063/1.4819792.

[10] S. Bhattacharya, G.K.H. Madsen, High-throughput exploration of alloying as design strategy for thermoelectrics, Phys. Rev. B. 92 (2015) 85205. doi:10.1103/PhysRevB.92.085205.

[11] P.L. Freddolino, F. Liu, M. Gruebele, K. Schulten, Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain, Biophys. J. 94 (2008) L75–L77. doi:10.1529/biophysj.108.131565.

[12] T.C. Germann, K. Kadau, Trillion-atom Molecular Dynamics becomes a Reality, Int. J. Mod. Phys. C. 19 (2008) 1315–1319. doi:10.1142/S0129183108012911.

[13] K.P. Esler, J. Kim, D.M. Ceperley, W. Purwanto, E.J. Walter, H. Krakauer, et al., Quantum Monte Carlo algorithms for electronic structure at the petascale; the Endstation project, J. Phys. Conf. Ser. 125 (2008) 12057. doi:10.1088/1742-6596/125/1/012057.

[14] D. Sholl, J. Steckel, Density Functional Theory: A Practical Introduction, Wiley, 2009.

[15] P. Hohenberg, W. Kohn, Inhomogeneous electron gas, Phys. Rev. 136 (1964) B864–B871. doi:10.1103/PhysRev.136.B864.

[16] W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (1965) A1133. doi:10.1103/PhysRev.140.A1133.

[17] J.P. Perdew, Climbing the ladder of density functional approximations, MRS Bull. 38 (2013) 743–750. doi:10.1557/mrs.2013.178.

[18] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, et al., The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, Npj Comput. Mater. 1 (2015) 15010. doi:10.1038/npjcompumats.2015.10.
[19] J. Hafner, C. Wolverton, G. Ceder, Toward Computational Materials Design: The Impact of Density Functional Theory on Materials Research, MRS Bull. 31 (2006) 659–668. doi:10.1557/mrs2006.174.
[20] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (2013) 191–201. doi:10.1038/nmat3568.
[21] G. Jóhannesson, T. Bligaard, A. Ruban, H. Skriver, K.W. Jacobsen, J. Nørskov, Combined Electronic Structure and Evolutionary Search Approach to Materials Design, Phys. Rev. Lett. 88 (2002) 255506. doi:10.1103/PhysRevLett.88.255506.
[22] T. Bligaard, M.P. Andersson, K.W. Jacobsen, H.L. Skriver, C.H. Christensen, J.K. Nørskov, Electronic-Structure-Based Design of Ordered Alloys, MRS Bull. 31 (2006) 986–990. doi:10.1557/mrs2006.225.
[23] S. Kirklin, J.E. Saal, V.I. Hegde, C. Wolverton, High-throughput computational search for strengthening precipitates in alloys, Acta Mater. 102 (2016) 125–135. doi:10.1016/j.actamat.2015.09.016.
[24] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, Finding Unprecedentedly Low-Thermal-Conductivity Half-Heusler Semiconductors via High-Throughput Materials Modeling, Phys. Rev. X. 4 (2014) 11019. doi:10.1103/PhysRevX.4.011019.
[25] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, et al., Charting the complete elastic properties of inorganic crystalline compounds, Sci. Data. (2015) 1–13. doi:10.1038/sdata.2015.9.
[26] A.A. Emery, J.E. Saal, S. Kirklin, V.I. Hegde, C. Wolverton, High-Throughput Computational Screening of Perovskites for Thermochemical Water Splitting Applications, Chem. Mater. 28 (2016) 5621–5634. doi:10.1021/acs.chemmater.6b01182.
[27] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, et al., Carbonophosphates: A New Family of Cathode Materials for Li-Ion Batteries Identified Computationally, Chem. Mater. 24 (2012) 2009–2016. doi:10.1021/cm203243x.
[28] M. Aykol, S. Kirklin, C. Wolverton, Thermodynamic Aspects of Cathode Coatings for Lithium-Ion Batteries, Adv. Energy Mater. 4 (2014) 1400690. doi:10.1002/aenm.201400690.
[29] M. Liu, Z. Rong, R. Malik, P. Canepa, A. Jain, G. Ceder, et al., Spinel compounds as multivalent battery cathodes: a systematic evaluation based on ab initio calculations, Energy Environ. Sci. 8 (2014) 964–974. doi:10.1039/C4EE03389B.
[30] S. Kirklin, M.K.Y. Chan, L. Trahey, M.M. Thackeray, C. Wolverton, High-throughput screening of high-capacity electrodes for hybrid Li-ion–Li–O 2 cells, Phys. Chem. Chem. Phys. 16 (2014) 22073–22082. doi:10.1039/C4CP03597F.
[31] D.H. Snydacker, M. Aykol, S. Kirklin, C. Wolverton, Lithium-Ion Cathode/Coating Pairs for Transition Metal Containment, J. Electrochem. Soc. 163 (2016) A2054–A2064. doi:10.1149/2.1101609jes.
[32] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials Design and Discovery

with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), Jom. 65 (2013) 1501–1509. doi:10.1007/s11837-013-0755-4.
[33] S. Kirklin, B. Meredig, C. Wolverton, High-Throughput Computational Screening of New Li-Ion Battery Anode Materials, Adv. Energy Mater. 3 (2013) 252–262. doi:10.1002/aenm.201200593.
[34] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T.O.L. Sunde, et al., Prediction and accelerated laboratory discovery of previously unknown 18-electron ABX compounds, Nat. Chem. 7 (2015) 308–316. doi:10.1038/nchem.2207.
[35] K. Yang, W. Setyawan, S. Wang, M. Buongiorno Nardelli, S. Curtarolo, A search model for topological insulators with high-throughput robustness descriptors., Nat. Mater. 11 (2012) 614–9. doi:10.1038/nmat3332.
[36] L. Lin, Materials Databases Infrastructure Constructed by First Principles Calculations: A Review, Mater. Perform. Charact. 4 (2015) 148–169. doi:10.1520/MPC20150014.
[37] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, et al., AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58 (2012) 227–235. doi:10.1016/j.commatsci.2012.02.002.
[38] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, et al., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 11002. doi:10.1063/1.4812323.
[39] http://nomad-repository.eu/cms/
[40] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, AiiDA: automated interactive infrastructure and database for computational science, Comput. Mater. Sci. 111 (2016) 218–230. doi:10.1016/j.commatsci.2015.09.013.
[41] A.M. Deml, R.O. Hayre, C. Wolverton, V. Stevanovic, Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression, Phys. Rev. B. 93 (2016) 85142. doi:10.1103/PhysRevB.93.085142.
[42] O. Pavlic, W. Ibarra-Hernandez, I. Valencia-Jaime, S. Singh, G. Avendaño-Franco, D. Raabe, et al., Design of Mg alloys: The effects of Li concentration on the structure and elastic properties in the Mg-Li binary system by first principles calculations, J. Alloys Compd. 691 (2017) 15–25. doi:10.1016/j.jallcom.2016.08.217.
[43] T. Nagase, M. Suzuki, T. Tanaka, Amorphous phase formation in Fe-Ag-based immiscible alloys, J. Alloys Compd. 619 (2015) 311–318. doi:10.1016/j.jallcom.2014.08.212.
[44] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (2016) 170901. doi:10.1063/1.4966192.
[45] L. Ward, C. Wolverton, Atomistic calculations and materials informatics: A review, Curr. Opin. Solid State Mater. Sci. (2016). doi:10.1016/j.cossms.2016.07.002.
[46] C. Jiang, B.P. Uberuaga, Efficient Ab initio Modeling of Random Multicomponent Alloys, Phys. Rev. Lett. 116 (2016) 105501. doi:10.1103/PhysRevLett.116.105501.
[47] V. Botu, R. Ramprasad, Adaptive machine learning framework to accelerate ab initio molecular dynamics, Int. J. Quantum Chem. 115 (2015) 1074–1083. doi:10.1002/qua.24836.
[48] J.C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, Finding Density Functionals with Machine Learning, Phys. Rev. Lett. 108 (2012) 253002.

doi:10.1103/PhysRevLett.108.253002.
[49] H. Putz, J.C. Schön, M. Jansen, Combined method for ab initio structure solution from powder diffraction data, J. Appl. Crystallogr. 32 (1999) 864–870. doi:10.1107/S0021889899006615.
[50] K.D.M. Harris, M. Tremayne, Crystal Structure Determination from Powder Diffraction Data, Chem. Mater. 8 (1996) 2554–2570. doi:10.1021/cm960218d.
[51] V. Pecharsky, P. Zavalij, Fundamentals of Powder Diffraction and Structural Characterization of Materials, Springer US, Boston, MA, 2009. doi:10.1007/978-0-387-09579-0.
[52] A. Altomare, M. Camalli, C. Cuocci, C. Giacovazzo, A.G.G. Moliterni, R. Rizzi, Advances in space-group determination from powder diffraction data, J. Appl. Crystallogr. 40 (2007) 743–748. doi:10.1107/S0021889807027501.
[53] H. Hauptman, J. Karle, Solution of the phase problem for space group P \overline{1}, Acta Crystallogr. 7 (1954) 369–374. doi:10.1107/S0365110X54001053.
[54] G. Oszlányi, A. Süto, Ab initio structure solution by charge flipping., Acta Crystallogr. A. 60 (2004) 134–41. doi:10.1107/S0108767303027569.
[55] R. Černý, V. Favre-Nicolin, Direct space methods of structure determination from powder diffraction: principles, guidelines and perspectives, Zeitschrift Für Krist. 222 (2007) 105–113. doi:10.1524/zkri.2007.222.3-4.105.
[56] A. Le Bail, H. Duroy, J.L. Fourquet, Ab-initio structure determination of LiSbWO6 by X-ray powder diffraction, Mater. Res. Bull. 23 (1988) 447–452. doi:10.1016/0025-5408(88)90019-0.
[57] A. Altomare, R. Caliandro, C. Cuocci, C. Giacovazzo, A.G.G. Moliterni, R. Rizzi, et al., Direct methods and simulated annealing: a hybrid approach for powder diffraction data, J. Appl. Crystallogr. 41 (2008) 56–61. doi:10.1107/S0021889807054192.
[58] C. Giacovazzo, Direct Methods and Powder Data: State of the Art and Perspectives, Acta Crystallogr. Sect. A Found. Crystallogr. 52 (1996) 331–339. doi:10.1107/S0108767395013651.
[59] http://www.iucr.org/resources/commissions/crystallographic-computing/software-museum
[60] P.D. Adams, P. V Afonine, G. Bunkóczi, V.B. Chen, N. Echols, J.J. Headd, et al., The Phenix software for automated determination of macromolecular structures., Methods. 55 (2011) 94–106. doi:10.1016/j.ymeth.2011.07.005.
[61] S.R. Ness, R. a G. de Graaff, J.P. Abrahams, N.S. Pannu, CRANK: new methods for automated macromolecular crystal structure solution., Structure. 12 (2004) 1753–61. doi:10.1016/j.str.2004.07.018.
[62] B.M. Kariuki, H. Serrano-González, R.L. Johnston, K.D.M. Harris, The application of a genetic algorithm for solving crystal structures from powder diffraction data, Chem. Phys. Lett. 280 (1997) 189–195. doi:10.1016/S0009-2614(97)01156-1.
[63] B. Meredig, C. Wolverton, A hybrid computational-experimental approach for automated crystal structure solution., Nat. Mater. 12 (2013) 123–7. doi:10.1038/nmat3490.
[64] O.J. Lanning, S. Habershon, K.D.M. Harris, R.L. Johnston, B.M. Kariuki, E. Tedesco, et al., Definition of a “guiding function” in global optimization : a hybrid approach combining energy and R -factor in structure solution from powder diffraction data, Chem. Phys. Lett.

(2000) 296–303.
[65] V. Brodski, R. Peschar, H. Schenk, A Monte Carlo approach to crystal structure determination from powder diffraction data, J. Appl. Crystallogr. 36 (2003) 239–243. doi:10.1107/S0021889802023208.
[66] S.R. Hall, F.H. Allen, I.D. Brown, The crystallographic information file (CIF): a new standard archive file for crystallography, Acta Crystallogr. Sect. A Found. Crystallogr. 47 (1991) 655–685. doi:10.1107/S010876739101067X.
[67] S. Gražulis, D. Chateigner, R.T. Downs, A.F.T. Yokochi, M. Quirós, L. Lutterotti, et al., Crystallography Open Database – an open-access collection of crystal structures, J. Appl. Crystallogr. 42 (2009) 726–729. doi:10.1107/S0021889809016690.
[68] A. Belsky, M. Hellenbrandt, V.L. Karen, P. Luksch, New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design, Acta Crystallogr. Sect. B Struct. Sci. 58 (2002) 364–369. doi:10.1107/S0108768102006948.
[69] A.G. Kusne, D. Keller, A. Anderson, A. Zaban, I. Takeuchi, High-throughput determination of structural phase diagram and constituent phases using GRENDEL, Nanotechnology. 26 (2015) 444002. doi:10.1088/0957-4484/26/44/444002.
[70] C.A. Gomez-Uribe, N. Hunt, The Netflix Recommender System, ACM Trans. Manag. Inf. Syst. 6 (2015) 13. doi:10.1145/2843948.
[71] A.R. Leach, V.J. Gillet, An Introduction to Chemical Informatics, Springer, 2003. doi:10.1002/3527603743.ch.
[72] J.J. Ramsden, Bioinformatics: An introduction, Springer London, 2009. doi:10.1111/j.1468-5922.2010.01872_2.x.
[73] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, et al., Mastering the game of Go with deep neural networks and tree search, Nature. 529 (2016) 484–489. doi:10.1038/nature16961.
[74] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software, ACM SIGKDD Explor. Newsl. 11 (2009) 10. doi:10.1145/1656274.1656278.
[75] F. Pedregosa, G. Varoquaux, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. doi:10.1007/s13398-014-0173-7.2.
[76] I.H. Witten, E. Frank, M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, 3rd ed., Morgan Kaufmann, 2011.
[77] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.
[78] J.J. Rodríguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: A new classifier ensemble method., IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1619–30. doi:10.1109/TPAMI.2006.211.
[79] L.M. Ghiringhelli, J. Vybiral, S. V Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. 114 (2015) 105503. doi:10.1103/PhysRevLett.114.105503.
[80] F. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, Crystal structure representations for machine learning models of formation energies, Int. J. Quantum Chem. 115 (2015)

1094–1101. doi:10.1002/qua.24917.
[81] A. Jain, G. Hautier, S.P. Ong, K. Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, J. Mater. Res. 31 (2016) 977–994. doi:10.1557/jmr.2016.80.
[82] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, et al., Machine learning of molecular electronic properties in chemical compound space, New J. Phys. 15 (2013) 95003. doi:10.1088/1367-2630/15/9/095003.
[83] B. Meredig, C. Wolverton, Dissolving the Periodic Table in Cubic Zirconia: Data Mining to Discover Chemical Trends, Chem. Mater. 26 (2014) 1985–1991. doi:10.1021/cm403727z.
[84] A. Agrawal, P.D. Deshpande, A. Cecen, G.P. Basavarsu, A.N. Choudhary, S.R. Kalidindi, Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters, Integr. Mater. Manuf. Innov. 3 (2014) 8. doi:10.1186/2193-9772-3-8.
[85] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, et al., Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B. 89 (2014) 94104. doi:10.1103/PhysRevB.89.094104.
[86] K. Rajan, Materials Informatics: The Materials “Gene” and Big Data, Annu. Rev. Mater. Res. 45 (2015) 153–169. doi:10.1146/annurev-matsci-070214-021132.
[87] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science, APL Mater. 4 (2016) 53208. doi:10.1063/1.4946894.
[88] S. V. Kalinin, B.G. Sumpter, R.K. Archibald, Big–deep–smart data in imaging for guiding materials design, Nat. Mater. 14 (2015) 973–980. doi:10.1038/nmat4395.
[89] P.M. Voyles, Informatics and data science in materials microscopy, Curr. Opin. Solid State Mater. Sci. (2016). doi:10.1016/j.cossms.2016.10.001.
[90] K. Rajan, ed., Informatics for Materials Science and Engineering, Elsevier, 2013. doi:10.1016/B978-0-12-394399-6.00021-7.
[91] F.A. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, Machine Learning Energies of 2 Million Elpasolite ABC2D6 Crystals, Phys. Rev. Lett. 117 (2016) 135502. doi:10.1103/PhysRevLett.117.135502.
[92] A.O. Oliynyk, E. Antono, T.D. Sparks, L. Ghadbeigi, M.W. Gaultois, B. Meredig, et al., High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds, Chem. Mater. 28 (2016) 7324–7331. doi:10.1021/acs.chemmater.6b02724.
[93] T.D. Sparks, M.W. Gaultois, A. Oliynyk, J. Brgoch, B. Meredig, Data mining our way to the next generation of thermoelectrics, Scr. Mater. 111 (2015) 10–15. doi:10.1016/j.scriptamat.2015.04.026.
[94] H.K.D.H. Bhadeshia, R.C. Dimitriu, S. Forsik, J.H. Pak, J.H. Ryu, Performance of neural networks in materials science, Mater. Sci. Technol. 25 (2009) 504–510. doi:10.1179/174328408X311053.
[95] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Finding Nature’s Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory, Chem. Mater. 22 (2010) 3762–3767. doi:10.1021/cm100795d.
[96] O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, et al., Materials

Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints, Chem. Mater. 27 (2015) 735–743. doi:10.1021/cm503507h.
[97] E.W. Bucholz, C.S. Kong, K.R. Marchman, W.G. Sawyer, S.R. Phillpot, S.B. Sinnott, et al., Data-Driven Model for Estimation of Friction Coefficient Via Informatics Methods, Tribol. Lett. 47 (2012) 211–221. doi:10.1007/s11249-012-9975-y.
[98] A.M. Deml, A.M. Holder, R.P. O'Hayre, C.B. Musgrave, V. Stevanović, Intrinsic Material Properties Dictating Oxygen Vacancy Formation Energetics in Metal Oxides, J. Phys. Chem. Lett. 6 (2015) 1948–1953. doi:10.1021/acs.jpclett.5b00710.
[99] V. Botu, A.B. Mhadeshwar, S.L. Suib, R. Ramprasad, Optimal Dopant Selection for Water Splitting with Cerium Oxides: Mining and Screening First Principles Data, in: Inf. Sci. Mater. Discov. Des., 2016: pp. 157–171. doi:10.1007/978-3-319-23871-5.
[100] C.S. Kong, W. Luo, S. Arapan, P. Villars, S. Iwata, R. Ahuja, et al., Information-theoretic approach for the discovery of design rules for crystal chemistry., J. Chem. Inf. Model. 52 (2012) 1812–20. doi:10.1021/ci200628z.
[101] C.S. Kong, S.R. Broderick, T.E. Jones, C. Loyola, M.E. Eberhart, K. Rajan, Mining for elastic constants of intermetallics from the charge density landscape, Phys. B Condens. Matter. 458 (2015) 1–7. doi:10.1016/j.physb.2014.11.002.
[102] A. Dima, S. Bhaskarla, C. Becker, M. Brady, C. Campbell, P. Dessauw, et al., Informatics Infrastructure for the Materials Genome Initiative, JOM. 68 (2016) 2053–2064. doi:10.1007/s11837-016-2000-4.
[103] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, The Materials Data Facility: Data Services to Advance Materials Science Research, JOM. 68 (2016) 2045–2052. doi:10.1007/s11837-016-2001-3.
[104] J. O'Mara, B. Meredig, K. Michel, Materials Data Infrastructure: A Case Study of the Citrination Platform to Examine Data Import, Storage, and Access, JOM. 68 (2016) 2031–2034. doi:10.1007/s11837-016-1984-0.
[105] W. Daniel, B. David, F. Tony, K. Surya, R. Andrew, PyMKS: Materials Knowledge System in Python, (2014). doi:10.6084/m9.figshare.1015761.
[106] T. Lookman, F.J. Alexander, A.R. Bishop, Perspective: Codesign for materials science: An optimal learning approach, APL Mater. 4 (2016) 53501. doi:10.1063/1.4944627.
[107] M.E. Weston, D.P. Shoemaker, The crystal structures of three phases in the Na-Pb system, Acta Crystallogr. 10 (1957) 775.
[108] A.R. Oganov, C.W. Glass, Crystal structure prediction using ab initio evolutionary techniques: principles and applications., J. Chem. Phys. 124 (2006) 244704. doi:10.1063/1.2210932.
[109] E. Majzoub, V. Ozoliņš, Prototype electrostatic ground state approach to predicting crystal structures of ionic compounds: Application to hydrogen storage materials, Phys. Rev. B. 77 (2008) 104115. doi:10.1103/PhysRevB.77.104115.
[110] A. Togo, I. Tanaka, Evolution of crystal structures in metallic elements, Phys. Rev. B. 87 (2013) 184104. doi:10.1103/PhysRevB.87.184104.
[111] Y. Wang, J. Lv, L. Zhu, Y. Ma, Crystal structure prediction via particle-swarm optimization, Phys. Rev. B. 82 (2010) 94116. doi:10.1103/PhysRevB.82.094116.
[112] D.C. Lonie, E. Zurek, XtalOpt: An open-source evolutionary algorithm for crystal structure

prediction, Comput. Phys. Commun. 182 (2011) 372–387. doi:10.1016/j.cpc.2010.07.048.
[113] K.J. Michel, C. Wolverton, Symmetry building Monte Carlo-based crystal structure prediction, Comput. Phys. Commun. 185 (2014) 1389–1393. doi:10.1016/j.cpc.2014.01.015.
[114] C.J. Pickard, R.J. Needs, Ab initio random structure searching., J. Phys. Condens. Matter. 23 (2011) 53201. doi:10.1088/0953-8984/23/5/053201.
[115] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, Predicting Crystal Structures with Data Mining of Quantum Calculations, Phys. Rev. Lett. 91 (2003) 135503. doi:10.1103/PhysRevLett.91.135503.
[116] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics., Nat. Mater. 5 (2006) 641–6. doi:10.1038/nmat1691.
[117] A.R. Oganov, S. Ono, Theoretical and experimental evidence for a post-perovskite phase of MgSiO3 in Earth's D" layer., Nature. 430 (2004) 445–8. doi:10.1038/nature02701.
[118] C.J. Pickard, R.J. Needs, When is H_2O not water?, J. Chem. Phys. 127 (2007) 244503. doi:10.1063/1.2812268.
[119] R. Hultgren, Selected Values of the Thermodynamic Properties of Binary Alloys, American Society for Metals, Metals Park, OH, 1973.
[120] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B. 47 (1993) 558–561. doi:10.1103/PhysRevB.47.558.
[121] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B. 59 (1999) 1758. doi:10.1103/PhysRevB.59.1758.
[122] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B. 50 (1994) 17953. doi:10.1103/PhysRevB.50.17953.
[123] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77 (1996) 3865. doi:10.1103/PhysRevLett.77.3865.
[124] J. Stöhr, H. Schafer, Die Kristallstrukturen von Li3In2, Li5Tl2 und Li3Tl, Zeitschrift Für Krist. 34 (1979) 653–656.
[125] U. Frank, W. Muller, H. Schafer, Die Struktur der Phase Li5Sn2, Zeitschrift Für Krist. 30 (1975) 1–5.
[126] U. Frank, W. Müller, Darstellung und Struktur der Phase Li13Sn5 und die strukturelle Verwandtschaft der Phasen in den Systemen Li-Sn und Li-Pb, Zeitschrift Für Naturforsch. B. 30 (1975) 316–322. doi:10.1515/znb-1975-5-605.
[127] L.D. Ellis, B.N. Wilkes, T.D. Hatchard, M.N. Obrovac, In Situ XRD Study of Silicon, Lead and Bismuth Negative Electrodes in Nonaqueous Sodium Cells, J. Electrochem. Soc. 161 (2014) A416–A421. doi:10.1149/2.080403jes.
[128] J. He, I.D. Blum, H.-Q. Wang, S.N. Girard, J. Doak, L.-D. Zhao, et al., Morphology Control of Nanostructures: Na-Doped PbTe–PbS System, Nano Lett. 12 (2012) 5979–84. doi:10.1021/nl303449x.
[129] http://www.icdd.com/products/pdf4.htm.
[130] L. Ward, K. Michel, C. Wolverton, Three new crystal structures in the Na–Pb system: solving structures without additional experimental input, Acta Crystallogr. Sect. A Found. Adv. 71 (2015) 542–548. doi:10.1107/S2053273315012516.

[131] G.L.W. Hart, L.J. Nelson, R.R. Vanfleet, B.J. Campbell, M.H.F. Sluiter, J.H. Neethling, et al., Revisiting the revised Ag-Pt phase diagram, Acta Mater. 124 (2017) 325–332. doi:10.1016/j.actamat.2016.10.053.
[132] https://github.com/materials/mint.
[133] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B. 54 (1996) 11169. doi:10.1103/PhysRevB.54.11169.
[134] G. Kresse, J. Hafner, Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, Phys. Rev. B. 49 (1994) 14251. doi:10.1103/PhysRevB.49.14251.
[135] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50. doi:10.1016/0927-0256(96)00008-0.
[136] M. Ladd, R. Palmer, Structure Determination by X-ray Crystallography, 4th ed., Kluwer Academic / Plenum Publishers, New York, N.Y., 2003.
[137] C. Giacovazzo, H.L. Monaco, G. Artioli, D. Viterbo, G. Ferraris, G. Gilli, et al., Fundamentals of Crystallography, 2nd ed., Oxford University Press, New York, N.Y., 2002.
[138] D. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[139] H. Rietveld, A profile refinement method for nuclear and magnetic structures, J. Appl. Crystallogr. 2 (1969) 65.
[140] E. Jansen, W. Schäfer, G. Will, R values in analysis of powder diffraction data using Rietveld refinement, J. Appl. Crystallogr. 27 (1994) 492–496. doi:10.1107/S0021889893012348.
[141] A.R. Akbarzadeh, V. Ozoliņš, C. Wolverton, First-Principles Determination of Multicomponent Hydride Phase Diagrams: Application to the Li-Mg-N-H System, Adv. Mater. 19 (2007) 3233–3239. doi:10.1002/adma.200700843.
[142] J.T. Zhao, E. Parthé, Y3NiAl3Ge2, a quaternary substitution variant of the hexagonal Fe2P type, Acta Crystallogr. Sect. C Cryst. Struct. Commun. 46 (1990) 2273–2276. doi:10.1107/S0108270190005194.
[143] C. Schank, F. Jährling, L. Luo, A. Grauel, C. Wassilew, R. Borth, et al., 4f-conduction electron hybridization in ternary Ce_3TMAl compounds, J. Alloys Compd. 207–208 (1994) 329–332. doi:10.1016/0925-8388(94)90234-8.
[144] C. Franz, A. Senyshyn, A. Regnat, C. Duvinage, R. Schönmann, A. Bauer, et al., Single crystal growth of CeTAl3 (T = Cu, Ag, Au, Pd and Pt), J. Alloys Compd. 688 (2016) 978–986. doi:10.1016/j.jallcom.2016.07.071.
[145] S. Vasala, M. Karppinen, A2B'B''O6 perovskites: A review, Prog. Solid State Chem. 43 (2015) 1–36. doi:10.1016/j.progsolidstchem.2014.08.001.
[146] E.D. Politova, Y.N. Venevtsev, No Title, Dokl. Akad. Nauk SSSR. 209 (1973) 838.
[147] S. Westman, P.-E. Werner, T. Schuler, W. Raldow, P.H. Nielsen, X-Ray Investigations of Ammines of Alkaline Earth Metal Halides. I. The Structures of CaCl2(NH3)8, CaCl2(NH3)2 and the Decomposition Product CaClOH., Acta Chem. Scand. 35a (1981) 467–472. doi:10.3891/acta.chem.scand.35a-0467.
[148] E.J.T. Salter, J.N. Blandy, S.J. Clarke, Crystal and Magnetic Structures of the Oxide Sulfides

CaCoSO and BaCoSO, Inorg. Chem. 55 (2016) 1697–1701. doi:10.1021/acs.inorgchem.5b02615.
[149] J. Guo, S. Jin, G. Wang, S. Wang, K. Zhu, T. Zhou, et al., Superconductivity in the iron selenide KxFe2Se2 (0<x<1), Phys. Rev. B. 82 (2010) 180520. doi:10.1103/PhysRevB.82.180520.
[150] A.P. Ayala, C.W.. Paschoal, J.-Y. Gesland, J. Ellena, E.E. Castellano, R.L. Moreira, Single-crystal structure determination and infrared reflectivity study of the Li 2 CaHfF 8 scheelite, J. Phys. Condens. Matter. 14 (2002) 5485–5495.
[151] M.M. Schieber, Growth of Rare Earth Scheelites by the Flux Method, Inorg. Chem. 4 (1965) 762–763. doi:10.1021/ic50027a040.
[152] V.B. Nalbandyan, M. Avdeev, A.A. Pospelov, Ion exchange reactions of NaSbO3 and morphotropic series MSbO3, Solid State Sci. 8 (2006) 1430–1437. doi:10.1016/j.solidstatesciences.2006.05.017.
[153] A. Boulineau, L. Croguennec, C. Delmas, F. Weill, Reinvestigation of Li 2 MnO 3 Structure: Electron Diffraction and High Resolution TEM, Chem. Mater. 21 (2009) 4216–4222. doi:10.1021/cm900998n.
[154] M. Edstrand, N. Ingri, The Crystal Structure of the Double Lithium Antimony(V)oxide LiSbO3., Acta Chem. Scand. 8 (1954) 1021–1031. doi:10.3891/acta.chem.scand.08-1021.
[155] Y. Hashimoto, M. Takahashi, S. Kikkawa, F. Kanamaru, Syntheses and Crystal Structures of Trigonal Rare-Earth Dioxymonocyanamides,Ln2O2CN2(Ln= Ce, Pr, Nd, Sm, Eu, Gd), J. Solid State Chem. 125 (1996) 37–42. doi:10.1006/jssc.1996.0261.
[156] M. Li, W. Yuan, J. Wang, C. Gu, H. Zhao, Syntheses and crystal structures of trigonal rare-earth dioxymonocyanamides, Ln2O2CN2 (Ln=Dy, Ho, Er, Tm, Yb), Powder Diffr. 22 (2007) 59–63. doi:10.1154/1.2424475.
[157] http://supercon.nims.go.jp/
[158] G.J. Mulholland, S.P. Paradiso, Perspective: Materials informatics across the product lifecycle: Selection, manufacturing, and certification, APL Mater. 4 (2016) 53207. doi:10.1063/1.4945422.
[159] S. Srinivasan, K. Rajan, “Property Phase Diagrams” for Compound Semiconductors through Data Mining, Materials (Basel). 6 (2013) 279–290. doi:10.3390/ma6010279.
[160] K.T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.R. Müller, E.K.U. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Phys. Rev. B. 89 (2014) 205118. doi:10.1103/PhysRevB.89.205118.
[161] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning., Sci. Rep. 3 (2013) 2810. doi:10.1038/srep02810.
[162] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. 104 (2010) 136403. doi:10.1103/PhysRevLett.104.136403.
[163] A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids, Phys. Rev. B. 89 (2014) 54303. doi:10.1103/PhysRevB.89.054303.
[164] Z.-Y. Hou, Q. Dai, X.-Q. Wu, G.-T. Chen, Artificial neural network aided design of catalyst

for propane ammoxidation, Appl. Catal. A Gen. 161 (1997) 183–190. doi:10.1016/S0926-860X(97)00063-X.
[165] B.G. Sumpter, D.W. Noid, On the Design, Analysis, and Characterization of Materials Using Computational Neural Networks, Annu. Rev. Mater. Sci. 26 (1996) 223–77. doi:10.1146/annurev.ms.26.080196.001255.
[166] S. Chatterjee, M. Murugananth, H.K.D.H. Bhadeshia, <I>δ</I> TRIP steel, Mater. Sci. Technol. 23 (2007) 819–827. doi:10.1179/174328407X179746.
[167] G. Hautier, Data Mining Approaches to High-Throughput Crystal Structure and Compound Prediction, in: Top. Curr. Chem., 2013: pp. 139–179. doi:10.1007/128_2013_486.
[168] L. Yang, G. Ceder, Data-mined similarity function between material compositions, Phys. Rev. B. 88 (2013) 224107. doi:10.1103/PhysRevB.88.224107.
[169] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Data mined ionic substitutions for the discovery of new compounds., Inorg. Chem. 50 (2011) 656–63. doi:10.1021/ic102031h.
[170] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, et al., Informatics-aided bandgap engineering for solar materials, Comput. Mater. Sci. 83 (2014) 185–195. doi:10.1016/j.commatsci.2013.10.016.
[171] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 19375. doi:10.1038/srep19375.
[172] Y. Kawazoe, J.Z. Yu, A.-P. Tsai, T. Masumoto, eds., Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys, Springer-Verlag, Berlin/Heidelberg, 1997. doi:10.1007/b58222.
[173] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH Verlag GmbH, Weinheim, Germany, Germany, 2000. doi:10.1002/9783527613106.
[174] Y.B. Ruiz-Blanco, W. Paz, J. Green, Y. Marrero-Ponce, ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins., BMC Bioinformatics. 16 (2015) 162. doi:10.1186/s12859-015-0586-0.
[175] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: An easy approach to molecular descriptor calculations, Match Commun. Math. Comput. Chem. 56 (2006) 237–248.
[176] A.R. Denton, N.W. Ashcroft, Vegards law, Phys. Rev. A. 43 (1991) 3161–3164. doi:10.1103/PhysRevA.43.3161.
[177] P. Villars, K. Cenzual, J. Daams, Y. Chen, S. Iwata, Data-driven atomic environment prediction for binaries using the Mendeleev number, J. Alloys Compd. 367 (2004) 167–175. doi:10.1016/j.jallcom.2003.08.060.
[178] W.D. Callister, Materials Science and Engineering: An Introduction, 7th ed., Wiley, 2007.
[179] A. Seko, A. Takahashi, I. Tanaka, Sparse representation for a potential energy surface, Phys. Rev. B. 90 (2014) 24101. doi:10.1103/PhysRevB.90.024101.
[180] M. Rupp, A. Tkatchenko, K.-R. Müller, V. Lilienfeld, O. Anatole, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. 108 (2012) 58301. doi:10.1103/PhysRevLett.108.058301.
[181] E.O. Pyzer-Knapp, G.N. Simm, A. Aspuru Guzik, A Bayesian approach to calibrating high-

throughput virtual screening results and application to organic photovoltaic materials, Mater. Horiz. 3 (2016) 226–233. doi:10.1039/C5MH00282F.
[182] A. Jain, G. Hautier, C.J. Moore, S. Ping Ong, C.C. Fischer, T. Mueller, et al., A high-throughput infrastructure for density functional theory calculations, Comput. Mater. Sci. 50 (2011) 2295–2310. doi:10.1016/j.commatsci.2011.02.023.
[183] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844. doi:10.1109/34.709601.
[184] W. Shockley, H.J. Queisser, Detailed Balance Limit of Efficiency of p-n Junction Solar Cells, J. Appl. Phys. 32 (1961) 510. doi:10.1063/1.1736034.
[185] W.H. Wang, C. Dong, C.H. Shek, Bulk metallic glasses, Mater. Sci. Eng. R Reports. 44 (2004) 45–89. doi:10.1016/j.mser.2004.03.001.
[186] A. Inoue, Stabilization of metallic supercooled liquid and bulk amorphous alloys, Acta Mater. 48 (2000) 279–306. doi:10.1016/S1359-6454(99)00300-6.
[187] S. Ding, Y. Liu, Y. Li, Z. Liu, S. Sohn, F.J. Walker, et al., Combinatorial development of bulk metallic glasses, Nat. Mater. 13 (2014) 494–500. doi:10.1038/nmat3939.
[188] J.F. Löffler, Formation of Bulk Metallic Glasses and Their Composites, MRS Bull. 32 (2007) 624–628.
[189] T. Wada, T. Zhang, A. Inoue, Formation and High Mechanical Strength of Bulk Glassy Alloys in Zr-Al-Co-Cu System, Mater. Trans. 44 (2003) 1839–1844. doi:10.2320/matertrans.44.1839.
[190] http://oqmd.org/static/analytics/glass_search.html
[191] C. Thornton, F. Hutter, H.H. Hoos, K. Leyton-Brown, Auto-WEKA, in: Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '13, ACM Press, New York, New York, USA, New York, USA, 2013: p. 847. doi:10.1145/2487575.2487629.
[192] G. Ceder, K. Persson, The Stuff of Dreams, Sci. Am. 309 (2013) 36–40. doi:10.1038/scientificamerican1213-36.
[193] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sánchez-Carrera, A. Gold-Parker, et al., The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid, J. Phys. Chem. Lett. 2 (2011) 2241–2251. doi:10.1021/jz200866s.
[194] X. Qu, A. Jain, N.N. Rajput, L. Cheng, Y. Zhang, S.P. Ong, et al., The Electrolyte Genome project: A big data approach in battery materials discovery, Comput. Mater. Sci. 103 (2015) 56–67. doi:10.1016/j.commatsci.2015.02.050.
[195] G.L.W. Hart, S. Curtarolo, T.B. Massalski, O. Levy, Comprehensive Search for New Phases and Compounds in Binary Alloy Systems Based on Platinum-Group Metals, Using a Computational First-Principles Approach, Phys. Rev. X. 3 (2013) 41035. doi:10.1103/PhysRevX.3.041035.
[196] I.E. Castelli, T. Olsen, S. Datta, D.D. Landis, S. Dahl, K.S. Thygesen, et al., Computational screening of perovskite metal oxides for optimal solar light capture, Energy Environ. Sci. 5 (2012) 5814. doi:10.1039/c1ee02717d.
[197] W. Chen, J.-H. Pöhls, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, et al., Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment, J. Mater. Chem. C. (2016). doi:10.1039/C5TC04339E.

[198] A. Bhatia, G. Hautier, T. Nilgianskul, A. Miglio, J. Sun, H.J. Kim, et al., High-Mobility Bismuth-based Transparent p -Type Oxide from High-Throughput Material Screening, Chem. Mater. 28 (2016) 30–34. doi:10.1021/acs.chemmater.5b03794.
[199] A. Franceschetti, A. Zunger, The inverse band-structure problem of finding an atomic configuration with given electronic properties, Nature. 402 (1999) 60–63. doi:10.1038/46995.
[200] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials, Npj Comput. Mater. 2 (2016) 16028. doi:10.1038/npjcompumats.2016.28.
[201] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, et al., A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds, Sci. Rep. 6 (2016) 34256. doi:10.1038/srep34256.
[202] A. Furmanchuk, A. Agrawal, A. Choudhary, Predictive analytics for crystalline materials: Bulk modulus, RSC Adv. (2016). doi:10.1039/C6RA19284J.
[203] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization, Phys. Rev. Lett. 115 (2015) 205901. doi:10.1103/PhysRevLett.115.205901.
[204] G. Pilania, P. V Balachandran, C. Kim, T. Lookman, Finding New Perovskite Halides via Machine Learning, Front. Mater. 3 (2016). doi:10.3389/fmats.2016.00019.
[205] L. Yang, S. Dacek, G. Ceder, Proposed definition of crystal substructure and substructural similarity, Phys. Rev. B. 90 (2014) 54102. doi:10.1103/PhysRevB.90.054102.
[206] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B. 87 (2013) 184115. doi:10.1103/PhysRevB.87.184115.
[207] G. Voronoi, Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs., J. Für Die Reine Und Angew. Math. 134 (1908) 198–287.
[208] E. Wigner, F. Seitz, On the constitution of metallic sodium, Phys. Rev. 43 (1933) 804–810. doi:10.1103/PhysRev.43.804.
[209] J. Cowley, An Approximate Theory of Order in Alloys, Phys. Rev. 77 (1950) 669–675. doi:10.1103/PhysRev.77.669.
[210] https://bitbucket.org/wolverton/magpie/.
[211] Y. Xu, M. Yamazaki, P. Villars, Inorganic Materials Database for Exploring the Nature of Material, Jpn. J. Appl. Phys. 50 (2011) 11RH02. doi:10.1143/JJAP.50.11RH02.
[212] G.S. Shieh, A weighted Kendall's tau statistic, Stat. Probab. Lett. 39 (1998) 17–24. doi:10.1016/S0167-7152(98)00006-6.
[213] D.C. Hofmann, Bulk Metallic Glasses and Their Composites : A Brief History of Diverging Fields, 2013 (2013).
[214] C. Suryanarayana, A. Inoue, Bulk Metallic Glasses, CRC, Boca Raton, FL, 2011.
[215] X. Yan, A thermodynamic approach for predicting the tendency of multicomponent metallic alloys for glass formation, Intermetallics. 9 (2001) 535–538. doi:10.1016/S0966-9795(01)00036-X.
[216] D. Kim, B.-J. Lee, N.J. Kim, Prediction of composition dependency of glass forming ability

of Mg–Cu–Y alloys by thermodynamic approach, Scr. Mater. 52 (2005) 969–972. doi:10.1016/j.scriptamat.2005.01.038.
[217] S. Gorsse, G. Orveillon, O.N. Senkov, D.B. Miracle, Thermodynamic analysis of glass-forming ability in a Ca-Mg-Zn ternary alloy system, Phys. Rev. B. 73 (2006) 224202. doi:10.1103/PhysRevB.73.224202.
[218] L. Ge, X. Hui, E. Wang, G. Chen, R. Arroyave, Z. Liu, Prediction of the glass forming ability in Cu–Zr binary and Cu–Zr–Ti ternary alloys, Intermetallics. 16 (2008) 27–33. doi:10.1016/j.intermet.2007.07.008.
[219] S.-W. Kao, C.-C. Hwang, T.-S. Chin, Simulation of reduced glass transition temperature of Cu–Zr alloys by molecular dynamics, J. Appl. Phys. 105 (2009) 64913. doi:10.1063/1.3086623.
[220] M.F. de Oliveira, F.S. Pereira, C. Bolfarini, C.S. Kiminami, W.J. Botta, Topological instability, average electronegativity difference and glass forming ability of amorphous alloys, Intermetallics. 17 (2009) 183–185. doi:10.1016/j.intermet.2008.09.013.
[221] K.J. Laws, K.F. Shamlaye, K. Wong, B. Gun, M. Ferry, Prediction of Glass-Forming Compositions in Metallic Systems: Copper-Based Bulk Metallic Glasses in the Cu-Mg-Ca System, Metall. Mater. Trans. A. 41 (2010) 1699–1705. doi:10.1007/s11661-010-0274-7.
[222] L. Yang, G. Guo, L.Y. Chen, C.L. Huang, T. Ge, D. Chen, et al., Atomic-Scale Mechanisms of the Glass-Forming Ability in Metallic Glasses, Phys. Rev. Lett. 109 (2012) 105502. doi:10.1103/PhysRevLett.109.105502.
[223] C. Tang, P. Harrowell, Anomalously slow crystal growth of the glass-forming alloy CuZr, Nat. Mater. 12 (2013) 1–5. doi:10.1038/nmat3631.
[224] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. 7 (2016) 11241. doi:10.1038/ncomms11241.
[225] P. Raccuglia, K.C. Elbert, P. Adler, C. Falk, M. Wenny, A. Mollo, et al., Machine-learning-assisted materials discovery using failed experiments, Nature. 533 (2016) 73–76. doi:10.1038/nature17439.
[226] M.K. Tripathi, P.P. Chattopadhyay, S. Ganguly, Multivariate analysis and classification of bulk metallic glasses using principal component analysis, Comput. Mater. Sci. 107 (2015) 79–87. doi:10.1016/j.commatsci.2015.05.010.
[227] M.K. Tripathi, S. Ganguly, P. Dey, P.P. Chattopadhyay, Evolution of glass forming ability indicator by genetic programming, Comput. Mater. Sci. 118 (2016) 56–65. doi:10.1016/j.commatsci.2016.02.037.
[228] L.-M. Wang, Y. Tian, R. Liu, W. Wang, A “universal” criterion for metallic glass formation, Appl. Phys. Lett. 100 (2012) 261913. doi:10.1063/1.4731881.
[229] Z. Long, H. Wei, Y. Ding, P. Zhang, G. Xie, A. Inoue, A new criterion for predicting the glass-forming ability of bulk metallic glasses, J. Alloys Compd. 475 (2009) 207–219. doi:10.1016/j.jallcom.2008.07.087.
[230] W.K. An, A.H. Cai, J.H. Li, Y. Luo, T.L. Li, X. Xiong, et al., Glass formation and non-isothermal crystallization of Zr62.5Al12.1Cu7.95Ni17.45 bulk metallic glass, J. Non. Cryst. Solids. 355 (2009) 1703–1706. doi:10.1016/j.jnoncrysol.2009.06.040.
[231] W.K. An, X. Xiong, Y. Liu, J.H. Li, A.H. Cai, Y. Luo, et al., Investigation of glass forming

ability and crystallization kinetics of Zr63.5Al10.7Cu10.7Ni15.1 bulk metallic glass, J. Alloys Compd. 486 (2009) 288–292. doi:10.1016/j.jallcom.2009.06.134.
[232] A.H. Cai, W.K. An, Y. Luo, T.L. Li, X.S. Li, X. Xiong, et al., Glass forming ability, non-isothermal crystallization kinetics, and mechanical property of Zr61.5Al10.7Cu13.65Ni14.15 metallic glass, J. Alloys Compd. 490 (2010) 642–646. doi:10.1016/j.jallcom.2009.10.125.
[233] A. Cai, X. Xiong, Y. Liu, W. An, J. Tan, Y. Pan, Design of new Zr–Al–Ni–Cu bulk metallic glasses, J. Alloys Compd. 468 (2009) 432–437. doi:10.1016/j.jallcom.2008.01.016.
[234] H.W. Chang, Y.C. Huang, C.W. Chang, C.C. Hsieh, W.C. Chang, Soft magnetic properties and glass formability of Y–Fe–B–M bulk metals (M=Al, Hf, Nb, Ta, and Ti), J. Alloys Compd. 472 (2009) 166–170. doi:10.1016/j.jallcom.2008.05.014.
[235] Z.Y. Chang, X.M. Huang, L.Y. Chen, M.Y. Ge, Q.K. Jiang, X.P. Nie, et al., Catching Fe-based bulk metallic glass with combination of high glass forming ability, ultrahigh strength and good plasticity in Fe–Co–Nb–B system, Mater. Sci. Eng. A. 517 (2009) 246–248. doi:10.1016/j.msea.2009.03.082.
[236] B. Dong, S. Zhou, D. Li, C. Lu, F. Guo, X. Ni, et al., A new criterion for predicting glass forming ability of bulk metallic glasses and some critical discussions, Prog. Nat. Sci. Mater. Int. 21 (2011) 164–172. doi:10.1016/S1002-0071(12)60051-3.
[237] S. González, I.A. Figueroa, H. Zhao, H.A. Davies, I. Todd, P. Adeva, Effect of mischmetal substitution on the glass-forming ability of Mg–Ni–La bulk metallic glasses, Intermetallics. 17 (2009) 968–971. doi:10.1016/j.intermet.2009.04.012.
[238] S. González, I.A. Figueroa, I. Todd, Influence of minor alloying additions on the glass-forming ability of Mg–Ni–La bulk metallic glasses, J. Alloys Compd. 484 (2009) 612–618. doi:10.1016/j.jallcom.2009.05.002.
[239] S. Guo, C.T. Liu, Phase stability in high entropy alloys: Formation of solid-solution phase or amorphous phase, Prog. Nat. Sci. Mater. Int. 21 (2011) 433–446. doi:10.1016/S1002-0071(12)60080-X.
[240] S. Guo, Y. Shen, Design of high strength Fe-(P, C)-based bulk metallic glasses with Nb addition, Trans. Nonferrous Met. Soc. China. 21 (2011) 2433–2437. doi:10.1016/S1003-6326(11)61032-7.
[241] N. Hua, R. Li, H. Wang, J. Wang, Y. Li, T. Zhang, Formation and mechanical properties of Ni-free Zr-based bulk metallic glasses, J. Alloys Compd. 509 (2011) S175–S178. doi:10.1016/j.jallcom.2011.01.078.
[242] X.M. Huang, C.T. Chang, Z.Y. Chang, a. Inoue, J.Z. Jiang, Glass forming ability, mechanical and magnetic properties in Fe-W-Y-B alloys, Mater. Sci. Eng. A. 527 (2010) 1952–1956. doi:10.1016/j.msea.2009.11.042.
[243] X.M.M. Huang, C.T.T. Chang, Z.Y.Y. Chang, X.D.D. Wang, Q.P.P. Cao, B.L.L. Shen, et al., Formation of bulk metallic glasses in the Fe–M–Y–B (M=transition metal) system, J. Alloys Compd. 460 (2008) 708–713. doi:10.1016/j.jallcom.2007.09.063.
[244] W. Jiao, D.Q. Zhao, D.W. Ding, H. Bai, W.H. Wang, Effect of free electron concentration on glass-forming ability of Ca – Mg – Cu system, J. Non. Cryst. Solids. 358 (2012) 711–714. doi:10.1016/j.jnoncrysol.2011.10.033.
[245] H. Kato, H.S. Chen, A. Inoue, Relationship between thermal expansion coefficient and

glass transition temperature in metallic glasses, Scr. Mater. 58 (2008) 1106–1109. doi:10.1016/j.scriptamat.2008.02.006.
[246] I. Kucuk, M. Aykol, O. Uzun, M. Yildirim, M. Kabaer, N. Duman, et al., Effect of (Mo, W) substitution for Nb on glass forming ability and magnetic properties of Fe–Co-based bulk amorphous alloys fabricated by centrifugal casting, J. Alloys Compd. 509 (2011) 2334–2337. doi:10.1016/j.jallcom.2010.11.011.
[247] K.J. Laws, D.B. Miracle, M. Ferry, A predictive structural model for bulk metallic glasses, Nat. Commun. 6 (2015) 8123. doi:10.1038/ncomms9123.
[248] K.J. Laws, K.F. Shamlaye, M. Ferry, Synthesis of Ag-based bulk metallic glass in the Ag-Mg-Ca-[Cu] alloy system, J. Alloys Compd. 513 (2012) 10–13. doi:10.1016/j.jallcom.2011.10.097.
[249] D.M. Lee, J.H. Sun, D.H. Kang, S.Y. Shin, G. Welsch, C.H. Lee, A deep eutectic point in quaternary Zr–Ti–Ni–Cu system and bulk metallic glass formation near the eutectic point, Intermetallics. 21 (2012) 67–74. doi:10.1016/j.intermet.2011.09.006.
[250] G.H. Li, W.M. Wang, X.F. Bian, J.T. Zhang, R. Li, J.Y. Qin, Correlation between thermal expansion coefficient and glass formability in amorphous alloys, Mater. Chem. Phys. 116 (2009) 72–75. doi:10.1016/j.matchemphys.2009.02.041.
[251] H.X. Li, Z.B. Jiao, J.E. Gao, Z.P. Lu, Synthesis of bulk glassy Fe–C–Si–B–P–Ga alloys with high glass-forming ability and good soft-magnetic properties, Intermetallics. 18 (2010) 1821–1825. doi:10.1016/j.intermet.2010.01.021.
[252] J.W. Li, A.N. He, B.L. Shen, Effect of Tb addition on the thermal stability, glass-forming ability and magnetic properties of Fe–B–Si–Nb bulk metallic glass, J. Alloys Compd. 586 (2014) S46–S49. doi:10.1016/j.jallcom.2012.09.087.
[253] Y.H. Li, W. Zhang, C. Dong, J.B. Qiang, K. Yubuta, A. Makino, et al., Unusual compressive plasticity of a centimeter-diameter Zr-based bulk metallic glass with high Zr content, J. Alloys Compd. (2010) 2–5. doi:10.1016/j.jallcom.2010.02.069.
[254] Z. Liu, K.C.C. Chan, L. Liu, Enhanced glass forming ability and plasticity of a Ni-free Zr-based bulk metallic glass, J. Alloys Compd. 487 (2009) 152–156. doi:10.1016/j.jallcom.2009.08.030.
[255] Z. Long, Y. Shao, F. Xu, H. Wei, Z. Zhang, P. Zhang, et al., Y effects on magnetic and mechanical properties of Fe-based Fe–Nb–Hf–Y–B bulk glassy alloys with high glass-forming ability, Mater. Sci. Eng. B. 164 (2009) 1–5. doi:10.1016/j.mseb.2009.04.010.
[256] M. Malekan, S.G. Shabestari, W. Zhang, S.H. Seyedein, R. Gholamipour, A. Makino, et al., Effect of Si addition on glass-forming ability and mechanical properties of Cu-Zr-Al bulk metallic glass, Mater. Sci. Eng. A. 527 (2010) 7192–7196. doi:10.1016/j.msea.2010.07.067.
[257] H. Peng, S.S. Li, Y.P. Qi, T.Y. Huang, Mg–Ni–Gd–Ag bulk metallic glass with improved glass-forming ability and mechanical properties, Intermetallics. 19 (2011) 829–832. doi:10.1016/j.intermet.2010.11.029.
[258] O.N. Senkov, Effect of the atomic size distribution on glass forming ability of amorphous metallic alloys, Mater. Res. Bull. 36 (2001) 2183–2198. doi:10.1016/S0025-5408(01)00715-2.
[259] K.K. Song, P. Gargarella, S. Pauly, G.Z. Ma, U. Kühn, J. Eckert, Correlation between glass-

forming ability, thermal stability, and crystallization kinetics of Cu-Zr-Ag metallic glasses, J. Appl. Phys. 112 (2012) 63503. doi:10.1063/1.4752263.
[260] S. Tao, Z. Ahmad, H. Jian, T. Ma, M. Yan, Synthesis, thermal stability and properties of [(Fe1-xCo x)72Mo4B24]94Dy 6 bulk metallic glasses, J. Alloys Compd. 509 (2011) 3843–3846. doi:10.1016/j.jallcom.2010.12.109.
[261] S. Tao, T. Ma, H. Jian, Z. Ahmad, H. Tong, M. Yan, Glass forming ability, magnetic and mechanical properties of (Fe72Mo4B24)100−xDyx (x=4–7) bulk metallic glasses, Mater. Sci. Eng. A. 528 (2010) 161–164. doi:10.1016/j.msea.2010.08.092.
[262] W.H. Wang, The elastic properties, elastic models and elastic perspectives of metallic glasses, Prog. Mater. Sci. 57 (2012) 487–656. doi:10.1016/j.pmatsci.2011.07.001.
[263] L. Wu, S. Li, J. Fang, Q. Chen, K. Peng, Enhancement of the glass forming ability of La–Al–Cu glassy alloys by partial substitution of Al by Mg, J. Alloys Compd. 504 (2010) S38–S40. doi:10.1016/j.jallcom.2010.03.102.
[264] K.F. Xie, K.F. Yao, T.Y. Huang, Preparation of (Ti0.45Cu0.378Zr0.10Ni0.072)100−xSnx bulk metallic glasses, J. Alloys Compd. 504 (2010) S22–S26. doi:10.1016/j.jallcom.2010.02.199.
[265] K.-F. Xie, K.-F. Yao, T.-Y. Huang, A Ti-based bulk glassy alloy with high strength and good glass forming ability, Intermetallics. 18 (2010) 1837–1841. doi:10.1016/j.intermet.2010.02.036.
[266] F. Xu, H.B. Lou, X.D. Wang, S.Q. Ding, Q.P. Cao, J.Z. Jiang, Glass forming ability and crystallization of Zr–Cu–Ag–Al–Be bulk metallic glasses, J. Alloys Compd. 509 (2011) 9034–9037. doi:10.1016/j.jallcom.2011.02.107.
[267] G.-H. Zhang, K.-C. Chou, A criterion for evaluating glass-forming ability of alloys, J. Appl. Phys. 106 (2009) 94902. doi:10.1063/1.3255952.
[268] Q.S. Zhang, W. Zhang, D.V. Louzguine-Luzgin, a. Inoue, Effect of substituting elements on glass-forming ability of the new Zr48Cu36Al8Ag8 bulk metallic glass-forming alloy, J. Alloys Compd. 504 (2010) S18–S21. doi:10.1016/j.jallcom.2010.02.052.
[269] Q. Jiang, B.Q. Chi, J.C. Li, A valence electron concentration criterion for glass-formation ability of metallic liquids, Appl. Phys. Lett. 82 (2003) 2984. doi:10.1063/1.1571984.
[270] D.B. Miracle, A structural model for metallic glasses., Nat. Mater. 3 (2004) 697–702. doi:10.1038/nmat1219.
[271] H.W. Sheng, W.K. Luo, F.M. Alamgir, J.M. Bai, E. Ma, Atomic packing and short-to-medium-range order in metallic glasses., Nature. 439 (2006) 419–25. doi:10.1038/nature04421.
[272] D.B. Miracle, E.A. Lord, S. Ranganathan, Candidate Atomic Cluster Configurations in Metallic Glass Structures, Mater. Trans. 47 (2006) 1737–1742. doi:10.2320/matertrans.47.1737.
[273] M. Falcão de Oliveira, A simple criterion to predict the glass forming ability of metallic alloys, J. Appl. Phys. 111 (2012) 23509. doi:10.1063/1.3676196.
[274] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378. doi:10.1016/S0167-9473(01)00065-2.
[275] K. Lejaeghere, S. Cottenier, V. Van Speybroeck, Ranking the Stars: A Refined Pareto Approach to Computational Materials Design, Phys. Rev. Lett. 111 (2013) 75501.

doi:10.1103/PhysRevLett.111.075501.
[276] A. Inoue, W. Zhang, T. Zhang, K. Kurosaka, High-strength Cu-based bulk glassy alloys in Cu-Zr-Ti and Cu-Hf-Ti ternary systems, Acta Mater. 49 (2001) 2645–2652. doi:10.1016/S1359-6454(01)00181-1.
[277] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, J. Comput. Phys. 117 (1995) 1–19. doi:10.1006/jcph.1995.1039.
[278] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys. Condens. Matter. 21 (2009) 395502. doi:10.1088/0953-8984/21/39/395502.
[279] B. Sundman, U.R. Kattner, M. Palumbo, S.G. Fries, OpenCalphad - a free thermodynamic software, Integr. Mater. Manuf. Innov. 4 (2015) 1. doi:10.1186/s40192-014-0029-1.
[280] A.M. Jokisaari, P.W. Voorhees, J.E. Guyer, J. Warren, O.G. Heinonen, Benchmark problems for numerical implementations of phase field models, Comput. Mater. Sci. 126 (2017) 139–151. doi:10.1016/j.commatsci.2016.09.022.
[281] D.C. Ince, L. Hatton, J. Graham-Cumming, The case for open computer programs, Nature. 482 (2012) 485–488. doi:10.1038/nature10836.
[282] R. Liu, L. Ward, C. Wolverton, A. Agrawal, W.-K. Liao, A. Choudhary, Deep Learning for Chemical Compound Stability Prediction, Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2016).
[283] https://github.com/libAtoms/QUIP.
[284] W.T. Hong, R.E. Welsch, Y. Shao-Horn, Descriptors of Oxygen-Evolution Activity for Oxides: A Statistical Evaluation, J. Phys. Chem. C. (2015) acs.jpcc.5b10071. doi:10.1021/acs.jpcc.5b10071.
[285] M. Fernandez, N.R. Trefiak, T.K. Woo, Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity, J. Phys. Chem. C. 117 (2013) 14095–14105. doi:10.1021/jp404287t.
[286] T. Fast, S.R. Kalidindi, Formulation and calibration of higher-order elastic localization relationships using the MKS approach, Acta Mater. 59 (2011) 4595–4605. doi:10.1016/j.actamat.2011.04.005.
[287] B.L. DeCost, E. a. Holm, A computer vision approach for automated analysis and classification of microstructural image data, Comput. Mater. Sci. 110 (2015) 126–133. doi:10.1016/j.commatsci.2015.08.011.
[288] S.Q. Wu, M. Ji, C.Z. Wang, M.C. Nguyen, X. Zhao, K. Umemoto, et al., An adaptive genetic algorithm for crystal structure prediction, J. Phys. Condens. Matter. 26 (2014) 35402–6. doi:10.1088/0953-8984/26/3/035402.
[289] M. Amsler, S. Goedecker, Crystal structure prediction using the minima hopping method, J. Chem. Phys. 133 (2010) 224104. doi:10.1063/1.3512900.
[290] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, Adaptive Strategies for Materials Design using Uncertainties, Sci. Rep. 6 (2016) 19660. doi:10.1038/srep19660.
[291] http://reference.wolfram.com/language/note/ElementDataSourceInformation.html.

10 Appendix: Formulae for Attributes

The first step in creating a machine learning model of a material property is to compute attributes that reflect physical effects potentially are influence that property. These attributes are designed to enable a machine learning algorithm to construct general rules that can possibly "learn" chemistry and reflect some kind of chemical intuition. In this thesis, we have developed many categories of attributes, which are described in detail in the following subsections.

10.1 Stoichiometric Attributes (6 in total)

These attributes capture the fraction of the elements present and are not affected by what those elements are. All are based on L^p norms (i.e. $\|x\|_p = (\sum_{i=0}^{n}|x_i|^p)^{1/p}$) of a vector representing the atomic fraction of the material corresponding to each element. In this work, we use the p=0 norm (which is equivalent to the number of components) and the p=2, 3, 5, 7, and 10 norms. Such a broad range was selected to create attributes that respond to changes in fractions with varied strengths. As an example, the p=7 norm for Fe_2O_3 is:

$$\|x\|_7 = \left(\left(\frac{2}{5}\right)^7 + \left(\frac{3}{5}\right)^7\right)^{1/7} \cong 0.605$$

10.2 Elemental-Property-Based Attributes (115 in total)

Most of the attributes created using our method are based on statistics of the elemental properties listed in Table 10.1. For each property, the minimum, maximum, and range of the values of the properties of each element present in the material is computed along with the

fraction-weighted mean, average deviation, and mode (i.e. the property of the most prevalent element). The mean and average deviation are calculated using the following formulae:

$$\bar{f} = \sum x_i f_i \tag{S7}$$

$$\hat{f} = \sum x_i \left| f_i - \bar{f} \right| \tag{S8}$$

where f_i is the property of element *i*, x_i is the atomic fraction, $\bar{f}$ is the mean, and $\hat{f}$ is the average deviation. As an example, the mean of and average deviation in the atomic number of Fe_2O_3 are:

$$\bar{f} = \frac{2}{5}(26) + \frac{3}{5}(8) = 15.2$$

$$\hat{f} = \frac{2}{5}|26 - 15.2| + \frac{3}{5}|8 - 15.2| = 8.16$$

Table 10.1. Elemental properties used to compute elemental-property-based attributes. Elemental property is taken from that dataset available with the Wolfram programming language,[291] unless otherwise specified.

Atomic Number	Mendeleev Number[177]	Atomic Weight	Melting Temperature	Column
Row	Covalent Radius	Electronegativity *	# s Valence Electrons	# p Valence Electrons
# d Valence Electrons	# f Valence Electrons	Total # Valance Electrons	# Unfilled s States†	# Unfilled p States†
# Unfilled d States†	# Unfilled f States†	Total # Unfilled States†	Specific Volume of 0 K Ground State‡	Band Gap Energy of 0 K Ground State‡
Magnetic Moment (per atom) of 0 K ground state‡	Space Group Number of 0 K Ground State‡			

*Electronegativities for Eu, Yb, Tb, Pm taken to be the average of that of the element with one greater and one less atomic number (e.g. the average of Sm and Gd is used for Eu)
†Computed as the number of electrons in a partially-occupied orbital subtracted from the total number of electrons allowed in that orbital. Unoccupied orbitals always count as 0. Example: an element with an electronic configuration of $[Ar]3d^3 4s^2$ has 0 unfilled s orbitals, 7 filled d orbitals, and 0 unfilled p and f orbitals by the measure defined here.
‡Data taken from OQMD.org

10.3 Valance Orbital Occupation Attributes (4 in total)

These attributes are the fraction-weighted average of the number of valance electrons in each orbital divided by the fraction-weighted average of the total number of valance electrons. This attribute is exactly equivalent to the one employed by Meredig, Agrawal *et al*.[85] As an example, the fraction of p electrons for Fe_2O_3 is computed by

$$F_p = \frac{\frac{2}{5}(0) + \frac{3}{5}(4)}{\frac{2}{5}(8) + \frac{3}{5}(6)} = \frac{6}{17} \cong 0.352$$

10.4 Ionic Compound Attributes (3 in total)

These attributes are designed to determine whether a material is ionically bonded. The first measure is a Boolean denoting whether it possible to form a neutral, ionic compound assuming each element takes exactly one of its common charge states. The other two are based on the "ionic character" of a binary compound, which is computed from the electronegativity difference between its two constituent elements using the relation

$$I(X_A, X_B) = 1 - \exp(-0.25(X_A - X_B)^2) \quad \text{(S9)}$$

where I is the fraction of ionic character, X_A is the electronegativity of element A, and X_B is the electronegativity of element B.[178] The first attribute we used is the maximum ionic character between any two elements in the material. The second is the mean ionic character, which is computed using

$$\bar{I} = \sum x_i x_j * I(X_i, X_j) \quad \text{(S10)}$$

10.5 Effective Coordination Number Attributes

We define the effective coordination number of an atom as a function of the sizes of faces on its Voronoi cell:

$$CN_{eff} = \frac{S^2}{\sum A_i^2} \quad (11)$$

where A_i is the area of face i and S is the total surface area of the cell. For a shape with equally-sized faces, the effective coordination number is exactly equal to the number of faces. Additionally, the introduction of a small face will only have a small influence on both the total

surface area and the sum of squared surface areas and, therefore, a small change in the effective coordination number.

We compute the maximum, minimum, mean, and mean absolute deviation in coordination number as attributes. The mean absolute deviation of a quantity is computed as:

$$\hat{f} = \frac{1}{N}\sum_i |f_i - \bar{f}| \quad (12)$$

where $\hat{f}$ is the mean absolute deviation, f_i is the value of sample i, N is the number of samples, and $\bar{f}$ is the mean. The mean absolute deviation was selected to measure variance in the properties of atoms in a structure (e.g., coordination number) because it is insensitive to unit cell selection:

1) All symmetrically-distinct images of an atom in a lattice will have the same property
2) All unit cell choices have the same proportion of each type of symmetrically-distinct atoms
3) Consequently, the mean property for any choice of unit cell will have the same mean
4) Therefore, the deviation between the mean property in a unit cell and the property for each atom is unchanged by unit cell choice

If the deviation in the property properties do not change with unit cell choice, the mean deviation will also be insensitive to unit cell choice.

10.6 Structural Heterogeneity Attributes

These attributes are designed to reflect variation in the shape of local bonding environments. The first set of attributes are based on maximum, minimum, and mean absolute deviation (see Eq. 12) in average bond length of each atom. Here, we define bond length as the Voronoi-face-area-weighted average of the distance between an atom and each neighbor:

$$\overline{l_i} = \frac{\sum A_n * \|\overrightarrow{r_n} - \overrightarrow{r_i}\|_2}{\sum A_n} \quad (13)$$

where $\overline{l_i}$ is the mean bond length of an atom *i*, $\overrightarrow{r_i}$ is the position of atom *i*, and A_n and $\overrightarrow{r_n}$ is the area and of the n^{th} neighbor of atom *i*. To make these attributes insensitive to scaling the volume of a unit cell, they are all normalized by the average $\overline{l_i}$ of all atoms.

Additionally, we create attributes based on the mean, maximum, minimum, and mean absolute deviation in the bond length variance of each atom. The bond length variance captures the distribution in bond lengths between each neighbor of an atom, and is computed by:

$$\widehat{l_i} = \frac{\sum |A_n * \|\overrightarrow{r_n} - \overrightarrow{r_i}\|_2 - \overline{l_i}|}{\overline{l_i} * \sum A_n} \quad (14)$$

where $\widehat{l_i}$ is the bond length variance and other terms are the same as in Eq. 13. As the variance is normalized by the mean bond length, this term is also insensitive to scaling the volume of the unit cell.

The mean absolute deviation of the volume of the Voronoi cell about each atom is also used as an attribute. This attribute is normalized by the mean volume of all cells in order make it insensitive to changes in the cell volume.

10.7 Chemical Ordering Attributes

These attributes are based on Warren-Cowley ordering parameters, which measure how the distribution of atoms differs from purely-random.[209] We first compute all *N*-length, non-backtracking paths originating from each atom in the crystal. For each step in these paths, we

assign the step a fractional weight corresponding to the size of its face compared to all faces corresponding to other possible (i.e., non-backing steps):

$$w_i = \frac{f_i}{\sum_a f_a - \sum_b f_b} \tag{15}$$

where f_i is the area of face i, and the two sums in the denominator are over the faces corresponding to all allowed and back-tracking steps, respectively. The total weight of a path is determined by multiplying the weight of these steps, which results in the sum over the weights of all paths being equal to 1. Consequently, the weight for each path can be envisioned as the probability a walker will take a certain path if its probability of making each step is proportional to the area of the face being traversed.

After determining the paths and their effective weights, sum the total weight of all paths ending on each type of atom. If the arrangement of atoms on the lattice is purely random, the likelihood of a type of atoms being at the end of any path is equal to the fraction of atoms of that type in the material. Consequently, the Warren-Cowley ordering parameter can be expressed as

$$\alpha_i(t, s) = 1 - \frac{\sum_p w_p \delta(t - t_p)}{x_t * n_s} \tag{16}$$

where $\alpha_i(t, s)$ is the weighted ordering parameter for type t in the s^{th} shell about atom i, , x_t is the atomic fraction of type t in the crystal, w_p is weight of path p, t_p is the type of atom at the end of path p, and δ is the delta function.

In order to generate attributes that describe the entire cell, the mean of the absolute values of the ordering parameter over all types and each atom is computed for the 1st, 2nd, and 3rd nearest-neighbor shells.

10.8 Maximum Packing Efficiency

The largest sphere centered on the position of an atom that can fit inside its Voronoi cell has a radius equal to the distance between the center of an atom and center of the closest face of the cell. In order to compute the maximum packing efficiency, the sum of the largest possible spheres for each atom is divided by the cell volume.

10.9 Local Environment Attributes

These attributes constitute the majority of the structure-based attributes used in our method, and are based on the difference in elemental properties between an atom and each neighbor. The local property difference for each atom is defined as the face-area-weighted mean of the absolute difference in elemental properties between an atom and each of its neighbors

$$\widehat{p_i} = \frac{\sum A_n * |p_n - p_i|}{\sum A_n} \tag{17}$$

where $\widehat{p_i}$ is the local property difference (for a hypothetical property *p*) of atom *i*, p_i is the elemental property of atom *i*, and A_n is the area of the face corresponding to neighbor *n*.

To create attributes, we compute the mean, mean absolute deviation (Eq. 12), maximum, and minimum in the local property difference for each atom considering 22 different elemental properties (listed in Table 10.1).

Publication List

Before PhD

1. **L. Ward,** D. Miracle, W. Windl, O.N. Senkov, K. Flores. "Structural evolution and kinetics in Cu-Zr metallic liquids from molecular dynamics simulations." Physical Review B. 88 (2013), 134205.

2. D. Mollenhauer, **L. Ward**, E. Iarve, S. Putthanarat, K. Hoos, S. Hallett, X. Li. "Simulation of discrete damage in composite overheight compact tension specimens." Composites Part A: Applied Science and Manufacturing. 41 (2012), 1667-1679.

3. R.C. Kramb, **L.T. Ward**, K.E. Jensen, R.A. Vaia, D.B. Miracle. "The structure of Cu-Zr glasses using a colloidal proxy system." Acta Materialia. 61 (2013), 2025-2032.

4. R.C. Kramb, **L.T. Ward**, K.E. Jensen, R.A. Vaia, D.B. Miracle. "Structural property comparison of Ca-Mg-Zn glasses to a colloidal proxy system." Acta Materialia. 61 (2013), 6911-6917.

5. A. Agrawal, R. Mishra, **L. Ward**, K.M. Flores, W. Windl. "An embedded atom method potential of beryllium." Modelling and Simulation in Materials Science and Engineering. 21 (2013), 085001.

During PhD

1. **L. Ward**, K. Michel, C. Wolverton. "Three New Crystal Structures in the Na-Pb System: Solving Structures without Additional Experimental Input." Acta Crystallographica Section A. 71 (2015), 542-548.

2. **L. Ward**, A. Agrawal, A. Choudhary, C. Wolverton. "A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials." npj Computational Materials. 2 (2016), 16028.

3. **L. Ward**, C. Wolverton. "Atomistic Calculations and Materials Informatics: A Review." Current Opinion in Solid State and Materials Science, *in press*

4. **L. Ward**, R. Liu, A. Krishna, V. Hegde, A. Agrawal, A. Choudhary, C. Wolverton. "Models for the Formation Energy of Crystalline Compounds using Voronoi Tessellations and Machine Learning.", *in preparation*

5. **L. Ward**, K. Michel, C. Wolverton. "High-Throughput Structure Solution from Powder Diffraction Data with the First-Principles-Assisted Structure Solution Method.", *in preparation*

6. **L. Ward**, S.C. O'Keeffe, J. Stevick, G.R. Jelbert, M.Aykol, C. Wolverton. "A Machine Learning Approach for Discovering and Engineering Bulk Metallic Glass Alloys.", *in preparation*

7. **L. Ward,** R.E. Hackenberg**. "**A Digital Workflow for Constructing Aging Models: Microstructure-Aware Hardness Prediction for U-Nb Alloys," *in preparation*

8. R. Liu, **L. Ward**, C. Wolverton, A. Agrawal, W.-K. Liao, A. Choudhary. "Deep Learning for Chemical Compound Stability Prediction," In the Workshop on Large-Scale Deep Learning for Data Mining, held in conjunction with the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016. *In press*

9. S. Tepavcevic, H. Zheng, R. Klee, **L. Ward**, C. Stoumpos, D.G. Hinks, C. Wolverton, N. Markovic and J.F. Mitchell. "Topotactic Chemical and Electrochemical Synthesis and Structure of Single Crystal Na3Ir2O6 and NaIrO3 Honeycomb Iridates", *in preparation*

10. S. Tepavcevic, H. Zheng, Z. Lu, **L. Ward**, B. Key, C. Wolverton, C. Stoumpos, J.F. Mitchell and N. Markovic.. "Sodium Single Crystal Battery," *in preparation*

11. Z. Yao, S. Kim, M. Aykol, Q. Li, J. Wu, J. He, **L. Ward**, V. Dravid, M.M. Thackery, C. Wolverton. "Revealing the Conversion Mechanism of Co_3O_4 during Lithiation from First Principles Calculations," *in preparation*