NORTHWESTERN UNIVERSITY

High Throughput Computational Materials Discovery

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

Sean Darius Griesemer

EVANSTON, ILLINOIS

June 2023

ABSTRACT

The field of materials discovery is undergoing an unprecedented transition from laboratory to computer. Behind this transition is the new ability to accurately compute material properties, especially energetic stability, from first principles with density functional theory (DFT). However, DFT remains computationally expensive, and DFT-based materials discovery is intractable, especially in high throughput, when the search space comprises a combinatorically explosive number of possible compositions and structures with many degrees of freedom. In this thesis, we develop ways to accelerate two legs of computational materials discovery: crystal structure solution and the search for new stable compounds, and employ our methods to conduct materials discovery in high throughput. For the first leg, crystal structure solution, we develop a novel method of rapidly solving crystal structures from experimental diffraction data by searching for candidate prototypes from materials databases and evaluating their DFT stabilities and diffraction pattern matches, and then deploy this method in high throughput to solve 521 structures of compounds with existing diffraction data. For the second leg, the search for stable compounds, we compare and improve the workflows of previously developed search methods based on data mining and machine-learned formation energy prediction, and then deploy the methods in high throughput to discover thousands of new compounds that DFT predicts to lie on the convex hull of stability. Finally, we provide a comprehensive literature review of recent efforts to develop artificial intelligence for the accelerated discovery of new materials that are stable at zero and finite temperature.

ACKNOWLEDGEMENTS

I would like to give my sincere thanks to those who assisted and supported me during my Ph.D.

First, I would like to thank my advisor, Prof. Chris Wolverton, for nurturing my research interests, asking thought-provoking questions, critiquing my work in detail, shaping up my scientific writing and presentation, and giving valuable advice.

Next, I would like to thank my committee members: Prof. Vinayak Dravid, Prof. Mercouri Kanatzidis, and Prof. James Rondinelli, for providing helpful feedback during my Ph.D.

I would like to thank my advisors of my undergraduate research: Dr. Binhua Lin and Prof. Stuart Rice, for training me and sparking my interest in materials science.

I would also like to thank the students of the Wolverton group, with whom I formed collaborative relationships and lifelong friendships. I give special thanks to post-doctoral researchers: Dr. Eric Isaacs, Dr. Koushik Pal, Prof. Yi Xia, and Prof. Yizhou Zhu for being an invaluable part of my training; former graduate students: Dr. Logan Ward and Dr. Vinay Hegde for laying the foundation of my thesis work; and Ruijie Zhu, as my experience guiding his Master's thesis work greatly influenced my own development.

I would like to thank my family, without whom none of this would be possible. I thank my mom, dad, and brother for nurturing my growth, sharing our love of science and technology, and always encouraging me to do my best; as well as all of my aunts, uncles, cousins, and grandparents who shaped me into who I am.

Finally, I would like to acknowledge financial support from the Center for Hierarchical Materials Design (CHiMaD). I also acknowledge computational resources: Quest at Northwestern, Extreme Science and Engineering Discovery Environment (XSEDE; now known as ACCESS), and National Energy Research Scientific Computing Center (NERSC).

PUBLICATIONS

[1] S. D. Griesemer, L. Ward, and C. Wolverton, "High-throughput crystal structure solution using prototypes," *Phys. Rev. Mater.*, vol. 5, p. 105 003, 10 Oct. 2021.

[2] J. Shen *et al.*, "Reflections on one million compounds in the open quantum materials database (OQMD)," *Journal of Physics: Materials*, vol. 5, no. 3, p. 031 001, Jul. 2022.

TABLE OF CONTENTS

ABBREVIATIONS

AFLOW	Automatic Flow database
AIRSS	Ab-Initio Random Structure Searching
ANN	Artificial neural network
CALPHAD	Calculation of Phase Diagrams
CALYPSO	Crystal Structure AnalYsis by Particle Swarm Optimization
CCA	Canonical-correlation analysis
CEF	Compound energy formalism
CGCNN	Crystal Graph Convolutional Neural Network
CHD	Convex hull distance
CM	Coulomb matrix
CPU	Central processing unit (of a computer)
DFT	Density functional theory
DMSP	Data Mining Structure Predictor
EBEF	Effective bond energy formalism
EFA	Entropy forming ability
EHH	Effective harmonic Hamiltonian

ESP	Element Structure Predictor
FPASS	First-Principles-Assisted Structure Solution
GA	Genetic algorithm
GAP	Gaussian approximation potential
GGA	Generalized gradient approximation
GPR	Gaussian process regression
GRDF	Generalized radial distribution function
HEA	High entropy alloy
HT-DFT	High throughput density functional theory
iCGCNN	Improved Crystal Graph Convolutional Neural Network
ICSD	Inorganic Crystal Structure Database
IFC	Interatomic force constant
IM	Intermetallic compound
ISP	Ion Structure Predictor
JARVIS	Joint Automated Repository for Various Integrated Simulations
KNN	K-nearest neighbor
KRR	Kernel ridge regression

LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Local density approximation
MAE	Mean absolute error
MBTR	Many-body tensor representation
ME	Mean error
MEGNet	MatErials Graph Network
ML	Machine learning
MLIP	Machine learned interatomic potential
MODNet	Material Optimal Descriptor Network
MP	Materials Project database
MPEA	Multi-principle element alloy
MTP	Moment tensor potential
NNP	Neural network potential
NOMAD	Novel Materials Discovery Repository
OPTIMADE	Open Databases Integration for Materials Design
OQMD	Open Quantum Materials Database
PDF	Powder Diffraction File

PRDF	Partial radial distribution function
QHA	Quasiharmonic approximation
RF	Random Forest
SISSO	Sure Independence Screening and Sparsifying Operator
SNAP	Spectral neighbor analysis potential
SOAP	Smooth Overlap of Atomic Positions descriptors
SQS	Special Quasirandom Structure
SS	Solid solution phase
SVM	Support vector machine
TI	Thermodynamic integration
TU-TILD	Two-Stage Upsampled Thermodynamic Integration Using Langevin Dynamics
USPEX	Universal Structure Predictor: Evolutionary Xtallography
VASP	Vienna Ab-Initio Simulation Package
XC	Exchange correlation functional
XRD	X-ray diffraction

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

Fundamental to the field of materials science and engineering is our collective knowledge of crystalline materials. For centuries, tens of thousands of unique crystals have been dug up from the Earth's crust and their properties studied and exploited, and more recently novel materials are being conjured up in laboraties and meticulously engineered to advance exciting new technologies. Among countless examples of this are the engineering of semiconducting inorganic compounds with record high figure of merit for high-performance and inexpensive thermoelectric devices, [1, 2] the discovery of electrode and solid-state electrolyte materials with targeted properties for zero-emissions electrochemical energy storage, [3–5] and the development of design guidelines for heteroanionic compounds that exhibit unique phenomena owing to their complex bonding chemistry. [6] Central to these exciting developments is our public knowledge of materials that can be made, and their crystal structures which are the key to unlocking their properties.

As of today, tens of thousands of unique crystals have been described in the literature, and thanks to recent advances in computational power, the complete set of materials can be obtained online and in publicly available, searchable materials databases including the Inorganic Crystal Structure Database (ICSD) and the Powder Diffraction File (PDF). As our many collective efforts to find materials have become well integrated, we are now able to study chemical trends and analyze properties over the entire set of known materials. For instance, databases of highly accurate first-principles density functional theory calculations, such as the Open Quantum Materials

Database (OQMD) and Materials Project (MP), have enabled an automatic calculation of phase diagrams and various electronic properties with just crystal structures as input. [7–11]

A crucial question remains unanswered: how many materials are still out there for us to find, and what are their crystal structures? According to the OQMD, there are 45366 unique structures in the ICSD portion of the database (downloaded in 2011), and the rate at which new structures are introduced has been steadily increasing in recent years. [12] By no means is this the complete number, or is it anywhere close to the complete number of inorganic materials that could hypothetically be made and used. New materials, for the most part, are still found the conventional way: dug up from the Earth, or synthesized in a laboratory, both very slow and laborious undertakings. There are, however, endeavors to automate the process of materials discovery. What's stopping us from quickly finding the other materials via automated methods? Two obstacles are holding us back: (1) crystal structure solution from experimental diffraction patterns remains a challenge, frustrating our ability to explore the properties of materials via first-principles study of their crystal structures; and (2) predicting new materials that can be made in a lab is overwhelming to say the least, given the enormity of the search space and the computational expense of simulations.

In this document, we describe novel automated methods to accelerate the process of crystal structure solution and materials prediction, and execute the automated methods in a high-throughput manner to resolve thousands of outstanding structure solution problems and discover new materials. In the following two sections (Section 1.1.1 and 1.1.2), we give a more thorough overview of the barriers to accelerated materials discovery and how our novel approaches seek to overcome those barriers. Next, In Section 1.2, we present DFT and OQMD stability evaluation, which are the backbone of our computational materials discovery methods. In Chapter 2, we present our novel method of solving crystal structures by searching for and evaluating prototypes in the OQMD database, and present the results of using the method to solve the structures of 521

compounds from the PDF for which x-ray diffraction data exists but structure is missing. In Chapter 3, we test the performance of already-developed materials discovery methods, based on data mining and machine learning, and construct workflows to greatly improve their performance in finding new stable compounds. In Chapter 4, we provide an overview of the tens of thousands of newly predicted stable compounds in OQMD, a large fraction of which were obtained using the materials discovery methods discussed in Chapter 3. In Chapter 5, we provide a comprehensive review of how artificial intelligence is being developed by various research teams to accelerate the prediction of stable materials, both at zero temperature and finite temperature. Finally, in Chapter 6, ...

1.1.1 Solving Structures

To date, there are numerous materials that have been discovered and synthesized, but their crystal structures remain unknown. Consequently, such materials are often not included in materials databases, especially crystal structure and DFT databases, and are therefore not considered by researchers who routinely use these data sources. For example, the ICSD is a very popular source in the materials community, but by definition does not include materials whose crystal structures are unknown. Furthermore, knowledge of crystal structure is required in order to understand the material's properties, and is a required input for first principles calculations of materials properties. So, why are we often missing crystal structure, which is such a crucial piece of information?

The answer lies in the difficulties of the experiments and interpretation of the output data. In attempt to obtain crystal structure, the common experiment is x-ray or neutron diffraction, where an x-ray/neutron beam is fired at a crystalline sample, and due to the high translational symmetry of the crystal, the beam diffracts at specific angles. Each lattice plane with atoms lying on it will diffract the beam at an angle specified by Bragg's law. Thus, a diffractometer will detect intensity

peaks at the relevant angles away from the incident beam; the data set is called a "diffraction pattern." This diffraction pattern must then be inverted (from reciprocal space to real space) in order to obtain the crystal structure. At this step, there can be complications that make the diffraction pattern hard to interpret. For one, the sample may contain impurities such that the diffraction pattern cannot be mapped to a perfect crystal. In many cases, the sample may have multiple types of crystals that must somehow be separated from the data. Another complication is "texturing," or preferred grain orientation of the crystal. Though not always achievable, the way to have a standardized, one-to-one mapping between diffraction pattern and crystal structure is to either have a single crystal or a uniformly powdered crystal. If neither of these conditions is met, then the diffraction pattern will look different from the standardized one (specifically, the relative peak heights will be different and some peaks may appear/disappear where they shouldn't). Another problem arises if the crystal contains very light elements that diffract very weakly, like hydrogen, or the crystal contains heavy elements that dwarf the signal of lighter elements. Furthermore, even for a perfect sample, the crystal structure can still be very difficult to determine. The process of inverting diffraction pattern to crystal structure removes the phase of diffracted waves, and so there is not necessarily a one-to-one mapping that can be straightforwardly computed. What crystallographers do firstly is to determine the unit cell parameters and space group, by investigating which peaks from the data are allowable based on symmetry and the distance between peaks. This step alone can be difficult, if an expert (or even a computer) cannot identify a specific symmetry group from the data. From here, one then determines the atomic coordinates that correspond to the peak heights. Here, crystal structure solution algorithms are used to find a set of atomic coordinates that matches the diffraction pattern. If the search space is too large, then the algorithm fails to find the right solution. There can be multiple sets of atomic coordinates that correspond to the same diffraction pattern. This also requires knowledge of the crystal's composition, which is not always

known in some experiments. If the diffraction pattern is recognizable from previous experiments, one can use databases to find an isostructural material (*i.e.*has the same "prototype") with the same diffraction pattern. But even when this is possible, it may be difficult to refine the exact atomic coordinates, particularly when the sample/data quality is too poor.

There are numerous crystal structure solution algorithms, and recent algorithms have utilized DFT in order to obtain structures that are physically plausible. [13–16]. The First-Principles-Assisted Structure Solution (FPASS) method, developed previously by our group, combines DFT with R-factor (*i.e.*the match to diffraction pattern) in order to obtain solutions that are both physically plausible and consistent with experimental diffraction data. [17–20] While these methods have demonstrated success at solving many problems, they are very computationally expensive and limited to solving just a few structures at a time.

We are interested in employing a method to more rapidly solve the numerous unsolved crystal structures so that we can expand crystal structure databases and our DFT database, OQMD. In Chapter 2, we describe our novel method that combines DFT, XRD analysis techniques, and DFT databases and employ the method to quickly and cheaply solve 521 previously unsolved crystal structures from the PDF diffraction database. In this method, we search for unique structure prototypes in the DFT database that satisfy the constraints known from XRD analysis: stoichiometry, space group, and/or number of atoms per unit cell. We then construct structures from the candidate prototypes, and evaluate their DFT stability against competing phases in the DFT database as well as R-factor. If the best candidate is below a threshold DFT stability and R-factor, we deem the structure to be solved.

1.1.2 Searching for New Materials

Numerous inorganic exist naturally in the earth, and over the years these have been (and still are being) extracted, isolated, characterized (such as by diffraction methods), and catalogued. Meanwhile, methods of material synthesis have been used for thousands of years, such as mixing and heating precursors, and over the last century, greater scientific understanding and industrialization have enabled the development of more sophisticated synthesis methods, such as ball milling and solution synthesis. All of these methods are routinely used to discover new materials, but they are laborious and time-consuming, and if the recipe and products haven't been catalogued, it is generally not known what outcome will be (although chemical intuition can be helpful). As of today, there are a great number of materials that could hypothetically be synthesized, but we don't know what they are or how to make them.

Computational modelling of materials is maturing quickly, and is now highly useful for predicting hypothetical materials that can be synthesized. With DFT, the formation energy of any hypothetical (fully-ordered) crystal structure can be computed with high enough accuracy that stability analysis, considering any competing phases, can be done. With the recent development of comprehensive DFT databases such as OQMD, only one DFT calculation of a hypothetical compound is needed and the stability analysis using already-computed competing phases can be done straightforwardly. With this infrastructure in place, we can proceed to evaluate the stabilities of a large number of hypothetical compounds in high throughput and quickly determine which ones are stable. However, the difficulty of high-throughput DFT (HT-DFT) is that there is a combinatorial explosion of the number of possible compounds one can make, considering all possible crystal structures and all elements in the periodic table. In Chapter 3, we explore methods recently developed to intelligently sample this enormous search space for the hypothetical compounds that are most likely to be stable. As there are several methods, we test them side-by-side on a complete

data set of DFT-calculated stabilities of all possible Heusler compounds that can be formed from elements on the periodic table, and develop protocols to greatly improve the performance of the state-of-the-art methods. Then, in Chapter 4, we use these methods to conduct a full-scale HT-DFT search of hypothetical materials across the whole range of structures and prototypes that are known to date, and discover tens of thousands of new stable compounds.

1.2 Density Functional Theory

Over the years, many physical models have been used to describe materials. Models that analytical functions, such as Lennard-Jones potential, Coulomb potential, and embedded atom model, can sometimes provide a fair accuracy in the description of certain classes of materials. However, such simple models cannot capture all of the physics of a material that is required to understand its properties (although "machine-learned" interatomic potentials have demonstrated promise in the last decade [21]). One such property, formation energy, is of special interest in this work to determine which candidate crystal structure is lowest in energy for a given composition. Crystal structure candidates can differ in energy by very small amounts, *e.g.*tens of meV/atom, [12] which simple models cannot resolve accurately.

The exact, or "first-principles" description of materials at the atomic level, based on the interactions of its electrons and nuclei, is given by the many-body Schrödinger equation. While this partial differential equation has an analytical solution for the simplest of cases, such as a single hydrogen atom, it generally can only be solved numerically, with approximations, in real many-body systems.

To accomplish this, materials researchers today rely on density functional theory (DFT). DFT was borne out of the Hohenberg-Kohn (H-K) theorems, which state that in a system of N electron kinetic energies and Coulomb interaction, the external potential is a unique functional of the elec-

tron density. [22] This leads to a reduction from $3N$ degrees of freedom to a much more tractable 3 spatial coordinates. Due to variational principle, the electron density that minimizes the total energy is the ground state of the system. In addition, Kohn and Sham showed that the system can be equivalently represented with non-interacting electron kinetic and electrostatic energy functionals of electron density as well as an unknown exchange correlation (XC) functional. [23] The XC functionals used today are often obtained by the local density approximation (LDA), density gradient (GGA), [24] and in some cases other approximations like meta-GGA, [25] PBEsol, [26] and hybrid functionals. [27]

All of the DFT calculations in this work were performed with the Vienna Ab-Initio Simulation Package (VASP), [28, 29] with pseudopotentials generated using the projector augmented-wave method [30] and PBE (GGA) exchange correlation functionals. The specific settings for the VASP calculations were all pre-determined during the initial development of the OQMD database, [7] so that the DFT-computed properties of $>$1 million materials in the database today [31] can be directly compared. After a VASP calculation is completed, we obtain a total energy E for the compound. From this, we compute a formation energy for the compound by subtracting the chemical potentials of its constituent elements:

$$\Delta H = E - \sum_{i}^{\text{elements}} n_i \mu_i \tag{1.1}$$

where ΔH is formation energy, n_i is the fraction of the constituent element in the compound, and μ_i is the chemical potential. The elemental chemical potentials were pre-determined to be the total energies of the elemental ground state structures, with an adjustment based on fitting to experimental formation energy data and to STP reference states. Using these chemical potentials, the DFT-calculated formation energies of known compounds were found to match experimental formation energies with a mean absolute error (MAE) of just 91 meV/atom. [12] A negative formation energy indicates that the compound is more energetically stable than the linear combination of its

elemental reference states. However, the most interesting quantity for this work is the compound's distance to the convex hull of formation energies with respect to composition. [32] To illustrate how this works, we show an example phase diagram of the Hf-Bi convex hull in Figure 1.1. The convex hull marked by a dashed black line indicates the phases from the Inorganic Crystal Structure Database (ICSD) that are in stable equilibrium: Hf, Hf_2Bi, $HfBi_2$, and Bi. However, when we solved the crystal structure of Hf_8Bi_9 not in the ICSD (see chapter 3), this compound turned out to also be stable, and so a new convex hull (marked by solid green line) is constructed. The convex hull construction is particularly important for this work, since it will tell us whether hypothetical (not experimentally known) compounds, that are not yet in materials databases, are energetically stable.

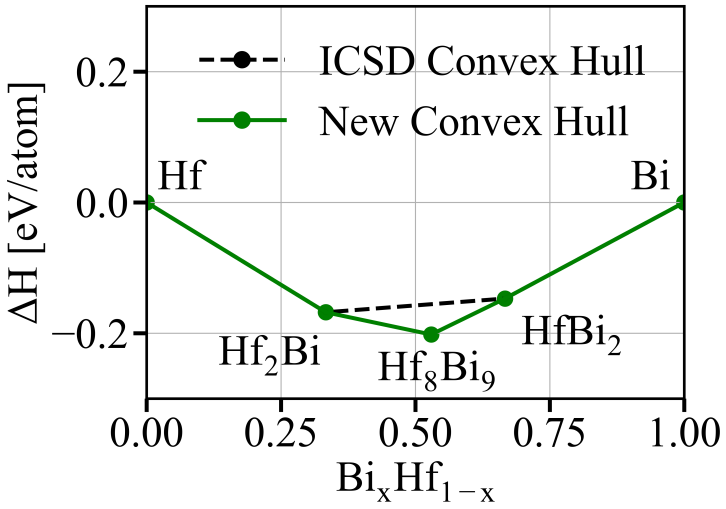

Figure 1.1: Hf-Bi phase diagram of formation energies with respect to composition. The convex hull is traced by lines: dashed black for when only ICSD compounds are used, and solid green for when a compound Hf_8Bi_9, not in the ICSD, is included.

CHAPTER 2
HIGH THROUGHPUT CRYSTAL STRUCTURE SOLUTION USING PROTOTYPES

Most of this chapter is taken directly from our published paper. [33]

2.1 Background

Crystal structure is a fundamental descriptor of inorganic compounds and is necessary input for first principles calculations. Indeed, the composition and crystal structure of a compound, comprising of unit cell vectors and atomic coordinates, are the only inputs required for a DFT calculation of the compound's energetic, electronic, and magnetic properties. Thanks to knowledge of crystal structures obtained by experiment, databases of high-throughput DFT calculations, such as the Open Quantum Materials Database (OQMD), Materials Project, and Automatic Flow (AFLOW), have enabled the calculation of phase diagrams, screening of materials for future applications, and prediction of novel materials. [7–9, 34–39] However, due to challenges in extracting crystal structure from experimental diffraction data, there are many known compounds with unknown crystal structures. For example, there are thousands of diffraction patterns in the Powder Diffraction File (PDF) without an associated crystal structure, meaning that compounds have been synthesized, diffraction patterns measured, and yet there is still no solved structure. [40] Identifying the structure of these materials would enable DFT calculations of their properties and open the door to a full exploration of their potential.

There are several reasons why a complete crystal structure is not always obtained in a diffraction experiment. For instance, while the unit cell parameters, space group, formula units per unit cell, and elemental composition can often be determined from high quality data by indexing the

diffraction peaks, the determination of atomic coordinates, known as structure solution, is especially challenging because the process of measuring a diffraction intensity does not capture the phase of diffracted waves, complicating the inversion from reciprocal space to real space. [41] Another common reason is that the compound either is part of a multi-phase material, is impure or contains elements that only weakly scatter x-rays, hindering the ability to capture relevant information in the diffraction pattern. When attempting to solve structures, crystallographers routinely use structure optimization algorithms, in which atomic coordinates are optimized to match the diffraction pattern, *i.e.*minimize the R-factor. When only R-factor is used for the objective function, structure optimization algorithms are sometimes challenged by the existence of multiple solutions with similar R-factors. A promising workaround is to supplement R-factor with DFT calculations in order to rule out candidates that are unphysically high in energy. [42] For example, the First-Principles-Assisted Structure Solution (FPASS) method, which uses a genetic algorithm with R-factor and DFT energy as a combined objective function, has been used to effectively resolve several long-standing problems. [17–20] Other DFT-based structure optimization algorithms that can be constrained using experimental input include USPEX, [13] CALYPSO, [14] XtalOpt, [15] and AIRSS. [16] DFT provides a highly accurate estimate of the energetic stability of candidate structures; however, DFT is computationally expensive to use in structure optimization algorithms like FPASS, USPEX and others, where as many as hundreds or thousands of structures are considered over the course of optimization.

On the other hand, a simpler and cheaper way to solve structures is to search existing databases. The OQMD contains DFT calculations of over 800,000 compounds, including experimentally observed compounds from a 2011 version of the ICSD as well as many hypothetical compounds. As we will show, structures from the ICSD portion of the OQMD can be grouped into 10,203 prototypes, distinguished by space group, stoichiometry, and Wyckoff site occupancies. This grouping

allows us to identify a relatively small number of symmetrically distinct prototypes as candidates for a given unsolved structure, according to the experimentally determined stoichiometry, number of formula units and space group. Furthermore, we find that 83% of distinct compounds in the ICSD share a common prototype with at least one other ICSD compound, giving us confidence that we can solve many (but not all) new structures using a "prototype searching" method. In this prototype searching method, we search for candidate prototypes in the OQMD, select a representative structure for each prototype, decorate the structures with the experimental composition, and evaluate them by computing R-factors and DFT energies. A related prototype searching method has been used to predict new compounds for hydrogen storage applications; [43–46] however, without experimental input to constrain the search, the prototype searching method is still computationally expensive. On the other hand, when used for structure solution with experimental input, our prototype searching method usually requires evaluating up to 3 prototypes, far fewer than what is needed for other structure solution methods, allowing us to solve structures at low cost. We note while structure optimization algorithms like FPASS, USPEX and others can leverage experimental input to constrain the search, as optimization algorithms they still typically require DFT calculations of many structures over the search space, including highly unphysical structures whose atomic coordinates are consistent with the experimental space group, stoichiometry, and Wyckoff site occupancies. On the other hand, our prototype searching method gets us straight to the answer with just a few DFT calculations for prototypes that are known to exist in the ICSD. Since the prototype searching method is inexpensive, it can be used to quickly solve numerous unsolved compounds and expand crystal databases with a limited computational budget.

In this work, we leverage the low computational cost of the prototype searching method to solve the structures of 521 compounds from experimental diffraction patterns in the PDF. All 521 compounds were missing from the ICSD and OQMD, and thus are newly solved, and constitute

a 1.4% expansion of all experimentally known compounds in the OQMD. Confident that we have identified structures that both match experimental input and are energetically stable, we open the door to analyzing the properties of these materials and considering their use in a wide range of future applications.

2.2 The Prototype Searching Method

In this section, we detail the prototype searching method to solve the structure of a compound given experimental data.

2.2.1 Searching for Candidate Structures

A completely solved structure is one where we know all descriptive details; in particular, the unit cell parameters and the coordinates of all atoms in the unit cell. For the compounds we address in this paper, we have from experimental data the unit cell parameters, elemental composition, space group, and number of formula units per unit cell, but we do not have the atomic coordinates, suggesting that the diffraction peaks were successfully indexed but the structure solution step was not completed. Our approach to solve for the atomic coordinates of the structure is to take the stoichiometry, space group, and the number of formula units per unit cell, and search the OQMD for prototypes with the same attributes. We define the prototype of a crystal structure as the set of the following attributes:

- Stoichiometry, *e.g.*ABC_3
- Space group
- Set of Wyckoff site occupancies in the unit cell

For example, the calcite prototype ($CaCO_3$) has ABC_3 stoichiometry, $R\bar{3}c$ space group, six atoms on the 6a $(1/4, 0, 0)$ Wyckoff site, six atoms on the 6b $(0, 0, 0)$ site, and eighteen atoms on the 18e $(x, 0, 1/4)$ site. Leveraging this definition allows us to classify 32 compounds within the OQMD with these attributes as having the prototype of calcite, irrespective of the elements comprising $\{$A,B,C$\}$, value of x, or unit cell parameters. This classification allows us to treat this group as one, symmetrically unique candidate solution to an experimental structure.

Having identified a relatively small number of prototypes as possible solutions to the experimental structure, we proceed to generate candidate structures by populating the prototypes with the experimentally determined unit cell parameters and elements from the composition. We consider all possible arrangements of elements in the structure, *e.g.*the two distinct ways to swap Ca and C onto the Wyckoff sites of the calcite prototype. We must also account for the fact that a single prototype can produce multiple structures that, while symmetrically identical, have different local geometries. For example, the structures C23, C25, C29, and C37 ($PbCl_2$, $HgCl_2$, SrH_2, and Co_2Si respectively) have the same stoichiometry, space group, and Wyckoff site occupancies (AB_2, Pnma, $\{$4c,4c,4c$\}$), but are distinct structures. In order to decide which of these structures to select as a candidate, we compute the R-factor of all compounds with this prototype in the OQMD but with the target composition and experimental lattice parameters substituted in. Since the calculation of R-factor is very fast, we can quickly select the structure with the lowest R-factor as the candidate. By the end of our procedure, we have generated a set of candidate structures, usually no more than seven structures across three prototypes, as possible solutions for the experimental structure.

We note that we initially assume the experimental structure does not have any partially occupied sites. In some cases, this assumption will be inevitably incorrect. We can justify the assumption if we obtain a structure that has a satisfyingly low energy and R-factor; otherwise, we say that none of our candidate structures are valid solutions.

2.2.2 Computing Match to Diffraction Pattern

We can determine which candidate structures from OQMD are valid solutions to unsolved PDF diffraction patterns by computing how well their simulated patterns match the experimental pattern. This computed pattern match is known as an R-factor and is routinely used by the crystallography community. Our method to compute R-factor is implemented in the Mint (Materials Interface) code; [47] here, we detail the method. [19]

Based on a candidate structure's easily-determined space group and lattice parameters, we make a list of all allowable diffraction peaks (each representing the distance between parallel lattice planes). We then use an equation from Pecharsky and Zavalij (eq. 8.41) to compute the intensity I of a peak located at (hkl): [41]

$$I_{hkl} = K \times m_{hkl} \times \text{LP}(\theta) \times T_{hkl} \times |F_{hkl}|^2 \tag{2.1}$$

where K is a fitted scale factor, m_{hkl} is the number of lattice planes for that peak, LP(θ) is the Lorentz-polarization factor corresponding to the angle of the peak, T_{hkl} is the March-Dollase function [48] to describe texturing (preferred grain orientation), and F_{hkl} is the structure factor. The R-factor is then computed as a sum of differences between the candidate structure (calc) and experimentally measured (obs) peak intensities:

$$R = \frac{\sum_{\text{peaks}} \left(I_{\text{calc}} - I_{\text{obs}}\right)^2}{\sum_{\text{peaks}} I_{\text{obs}^2}} \tag{2.2}$$

An R-factor of zero indicates a perfect match between the two patterns. The experimental peaks were taken from the PDF4+ software, which obtained the peak intensities by integrating over the angle interval of the peak profile. The Mint code matches peaks between experimental and calcu-

lated pattern by summing together all intensities within a 0.15° interval and pairing the intervals between the two patterns. Finally, the R-factor is locally relaxed by optimizing the free parameters of equation 2.2 (including atomic coordinates in the structure factor) using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, implemented in Dlib. [49]

We show an example pattern matching in Figure 2.1 for the experimental pattern of Hf_8Bi_9 and its solution. It is clear that the peaks align well with the two patterns, and the resulting R-factor of 0.21 is low.

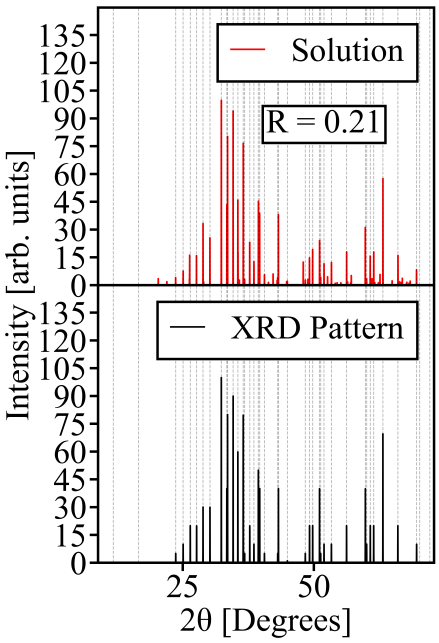

Figure 2.1: Experimental XRD pattern and solved structure pattern of Hf_8Bi_9, an entry in PDF with missing atomic coordinates. The two patterns agree very well and the resulting R-factor of 0.21 is low.

2.2.3 Choosing a Structure As the Solution

After computing the energy and R-factor of each candidate structure, we select the best performing structure as the final solution. To do so, we take all candidate structures with an R-factor within 0.2 above the lowest R-factor found, and then select the lowest-energy structure among these. We then decide whether the final solution is valid, based on the values of energy and R-factor; we provided a detailed description of the validation procedure in the Section 2.3.4.

A schematic diagram of our prototype searching method is given in Figure 2.2 for an example PDF entry, VI_3 (PDF# 00-023-0719), that contained a diffraction pattern, space group, and unit cell parameters, but no atomic coordinates. We obtain three candidate prototypes (FeF_3-type, PdF_3-type, BiF_3-type) from the OQMD, generate one structure of VI_3 for each prototype, and evaluate their DFT formation energies and R-factors. The BiI_3-type structure has both the lowest formation energy and the lowest R-factor and is thus the best-performing prototype of the three. The BiI_3-type structure also has sufficiently low energy and R-factor according to our validation criteria (see Section 2.3.4), and so we declare it to be the solution of the VI_3 measurement.

2.3 Results

2.3.1 Prevalence of Prototypes Among Known Inorganic Compounds

The OQMD contains DFT calculations of experimentally observed inorganic compounds from a 2011 version of the ICSD, excluding those with partial occupancy or very large unit cells. Using the definition of a prototype outlined in Section 2.2.1, we build a database of all prototypes that exist among 36807 nonduplicate, stoichiometric, and inorganic compounds in the ICSD portion of the OQMD. An exhaustive database like this one can be compared to existing prototype databases such as the one built from AFLOW. [50–52] The AFLOW prototype database distinguishes prototypes

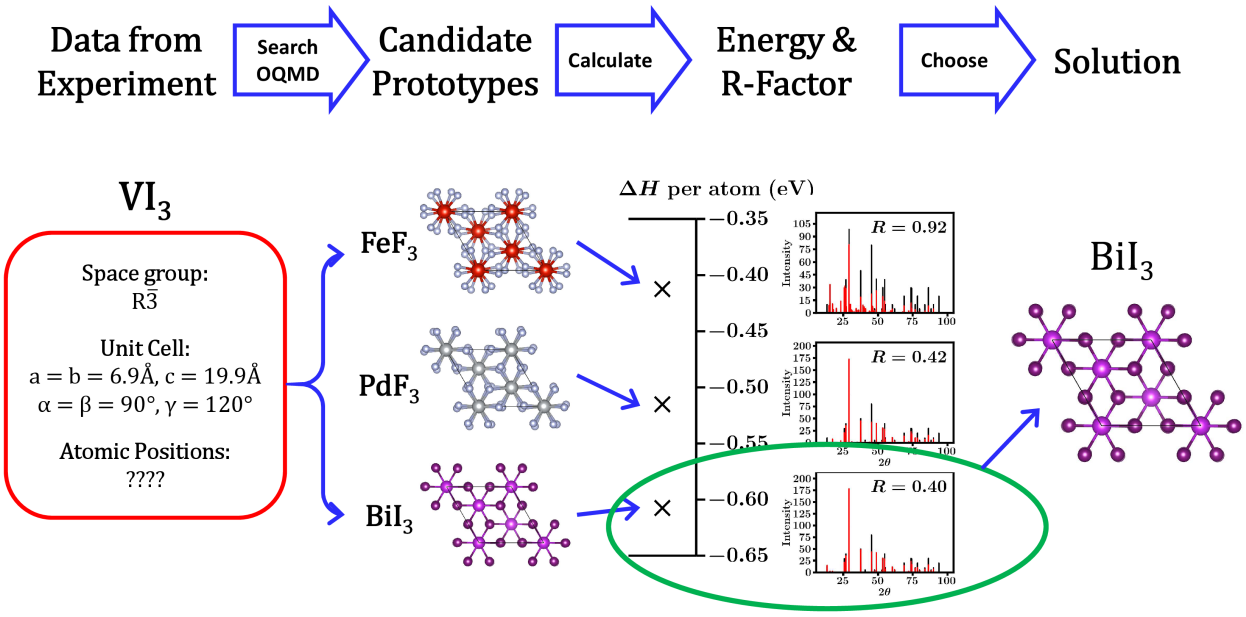

Figure 2.2: Flow chart of the prototype searching method to solve structures. The compound VI_3 (PDF #: 00-023-0719) is presented here as an example. Using experimentally determined structure attributes absent atomic coordinates, we search the OQMD for all prototypes with the space group ($R\bar{3}$), stoichiometry (AB_3), and formula units per cell (6). We then evaluate each of the three prototypes found (FeF_3-type, PdF_3-type, BiI_3-type) using DFT and R-factor. We find that the BiI_3 prototype is the most plausible solution because it has the lowest formation energy and R-factor.

in a similar manner, *i.e.*by space group, stoichiometry, and Wyckoff sites, but also distinguishes prototypes with different local geometries, *e.g.*C23, C25, C29, and C37. A key distinction of our prototype database is that it is exhaustive and contains many more prototypes than the 1100 prototypes in AFLOW. From 36807 nonduplicate, stoichiometric compounds, of which 7852 are binary, 18482 are ternary, and 8076 are quaternary, we identify a total of 10203 prototypes, of which 1617 are binary, 4120 are ternary, and 3062 are quaternary. Although this implies that there is an average of 3.6 compounds per prototype, some prototypes are shared by hundreds of compounds. In Figure 2.3, we plot the sorted number of compounds per prototype. There are 77 prototypes with fifty or more compounds, accounting for 27% of the total number of compounds in the ICSD set. A table of these prototypes is given in our paper on this work. [33] Such prototype-

sharing reflects that compounds with similar chemistries tend to arrange in the same or similar geometries. For example, binary compounds containing a cation and an anion most commonly have NaCl, $PbCl_2$, CaF_2, and CdI_2 prototypes, while half-Heusler and related prototypes are often observed for compounds with metals and metalloids where the sum of valence electrons equals 8 or 18. [53, 54]

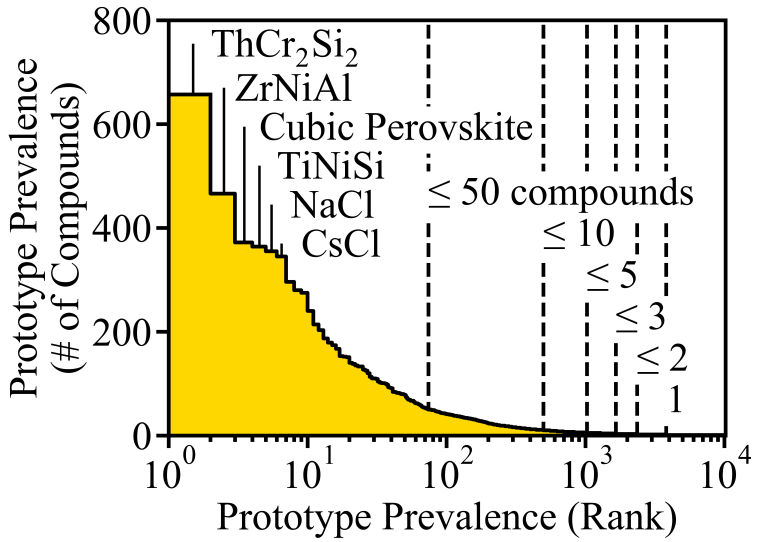

Figure 2.3: Sorted numbers of compounds associated with prototypes in the 2011 version of the ICSD, present in the OQMD. The total number of compounds in this ICSD set is 36807. The most prevalent (rank 1) prototype is $ThCr_2Si_2$, with 657 compounds in the ICSD; and the second-most prevalent (rank 2) prototype is ZrNiAl, with 466 compounds. Beginning at rank 10 or so, the trend in prototype prevalence smoothly decays with a wide tail.

While most compounds share common prototypes, there are also 6394 prototypes, of which 981 are binary, 2428 are ternary, and 1928 are quaternary, that are associated with just a single compound in the ICSD. These "one-hit wonders" highlight a shortcoming of the prototype searching method for solving crystal structures: some structures, $\sim$17% of the ICSD, are unique and cannot be solved by searching for already-known prototypes. We can acquire an insight about the one-hit wonders by investigating their statistics. For instance, we note that hydrogen is disproportionately

represented among the one-hit wonders: 37% of compounds containing H are one-hit wonders, much higher than the average of 17%. Other elements that are disproportionately represented are N (34%), F (31%), and Xe (56%). Nonmetals and alkali metals in general are disproportionately represented ($\gtrsim$20%), while lanthanide and actinide elements rarely occur by themselves ($\leq$10%), except for La (15%) and U (18%). Previous studies have identified which element pairs commonly appear together in compounds of the same prototype. [55, 56] In addition, many of the one-hit wonders have unique stoichiometries, such as $Fe_{107}O_{125}$. We find that 941 compounds do not share the same stoichiometry with any other compound in the ICSD. Many-component compounds tend to be unique as well: 955 of 2156 compounds (44%) with five or more components are one-hit wonders. One-hit wonders tend to also have larger unit cells: 29% of compounds with forty or more atoms are one-hit wonders, compared to 16% of compounds between twenty and forty atoms and just 6% of compounds with fewer than twenty atoms. Furthermore, most space groups are rarely observed. We find that 159 of 230 space groups have an above average proportion of one-hit wonders (>17%), and 11 space groups are not observed at all. On the other hand, a select few space groups account for a much larger proportion of ICSD compounds. One such space group is $Fm\bar{3}m$, which is found in 1464 compounds, of which only 33 (2%) are one-hit wonders.

2.3.2 Description of Target Compounds from the Powder Diffraction File

As the prototype searching method is cheap, often costing only a few DFT calculations, we leverage this approach to perform "high-throughput" structure solution for numerous entries from the International Centre for Diffraction Data (ICDD) database within the PDF for which the atomic coordinates are missing but other structure details are known. We start with 80624 entries missing atomic coordinates in the 2018 version of the PDF4+ software. We screen for entries that satisfy the following criteria:

- Entry is "primary" status as identified by the PDF4+ software, *i.e.*is not an alternative to another similar entry.
- Diffraction experiment was performed under ambient conditions. We note that the enthalpy of high-pressure compounds can be accounted for in DFT by supplying external pressure to the stress tensor. Furthermore, the enthalpies of high-pressure compounds can be compared to those of other compounds in the OQMD. [57]
- Compound is binary, ternary, or quaternary.
- Compound is inorganic and does not contain noble gases, actinides, or radioactive elements.
- Diffraction data quality is listed as "star," "good," or "indexed," indicating that the diffraction pattern represents a single-phase crystal with minimal impurities. While structures with poorer quality diffraction patterns can still be solved, their R-Factors may be less useful.
- Space group and number of formula units per unit cell are already known.
- Reduced cell volumes are less than 3000$Å^3$ and unit cells contain few enough atoms to be cheaply assessed by high-throughput DFT:
  - Cubic, hexagonal, trigonal, and tetragonal cells contain 80 or fewer atoms.
  - Orthorhombic cells contain 40 or fewer atoms.
  - Monoclinic and triclinic cells contain 20 or fewer atoms.
- The structure does not evidently contain partially occupied sites, *i.e.*, the listed composition contains only natural numbers and it is possible to generate a structure with a set of fully occupied Wyckoff sites given the listed space group and number of formula units per unit cell. We note that a PDF entry satisfying these conditions may not necessarily represent a

fully occupied structure. We can justify the validity of our prototype structures based on DFT energy and R-factor.

- There is no existing OQMD nor ICSD compound with the same composition and space group.
- There is at least one prototype in the OQMD matching the known stoichiometry, space group, and number of formula units per unit cell.

We find 603 PDF entries that satisfy the above constraints. We additionally find hundreds of entries that satisfy all the above constraints except for the last one, *i.e.*, there is no prototype in the OQMD that matches the provided stoichiometry, space group, and number of atoms per unit cell. However, it is possible that the listed space group is incorrect, and hence we attempt to solve these by searching within the crystal system, *e.g.*, tetragonal space groups, instead of the listed space group.

2.3.3 Summary of Structures Obtained by Prototype Searching

For 603 PDF entries with diffraction data but no structure, we find at least one prototype in the OQMD that matches the provided space group, stoichiometry, and number of formula units per unit cell. In Figure 2.4a, we plot a distribution of the number of prototypes found per PDF entry. The highest number of prototypes is only ten. In most cases (386, or 64%), only one prototype is found. Although the number of candidate prototypes is almost always very few, each prototype can produce multiple structures representing the possible ways to arrange elements onto the prototype's Wyckoff sites. Despite this, there are rarely more than a dozen structures to evaluate (see Figure 2.4b).

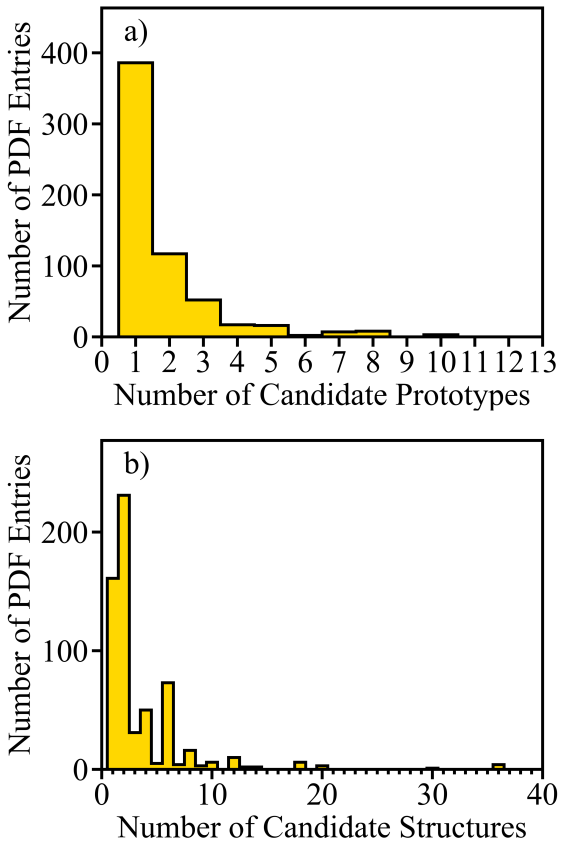

Figure 2.4: Distribution of the number of a) OQMD prototypes and b) candidate structures matching stoichiometry, space group, and number of formula units per unit cell of 603 PDF entries with missing atomic coordinates.

After computing the DFT energy and R-factor of all candidate structures, we select the structure with the lowest DFT energy among all candidates within 0.2 of the lowest R-factor. We are thus left with 603 structure candidates, each one outperforming other candidates for every attempted PDF entry. For 10 of the 603 entries, we find a candidate with a different space group within the same crystal system that outperforms all candidates with the reported space group. In these 10 cases, the structure with the same space group fails our validation checks of energy and R-factor (described in Section 2.3.4), while the structure with a different space group passes these

checks; we thus opt to present the 10 structures with a different space group. In addition, we find that for 21 of the PDF entries, while there is no prototype in the OQMD that matches the reported space group, stoichiometry, and number of formula units per cell, there is a candidate with a different space group within the same crystal system that passes our validation checks. In total, we present 624 structures (603 + 21) in the following analysis. Of these, 521 pass our validation checks of energy and R-factor, and we thus declare them to be solved.

2.3.4 Analysis of Structures Obtained by Prototype Searching

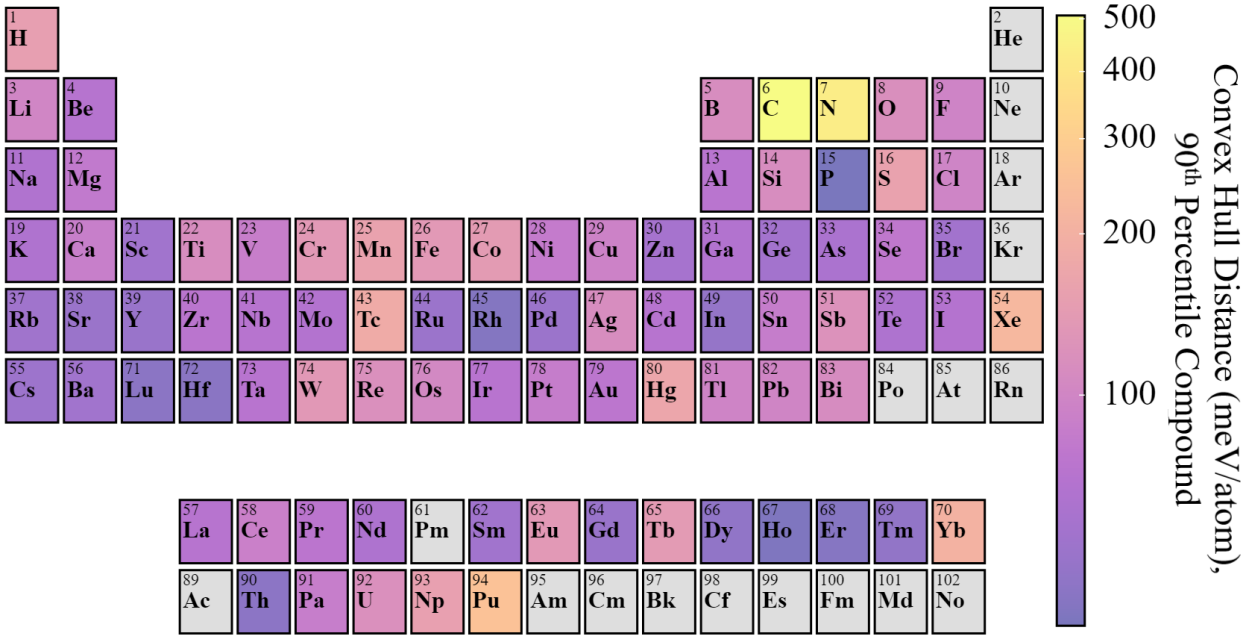

Figure 2.5: DFT-computed 90th percentile convex hull distance (meV/atom) of ICSD compounds containing each element on the periodic table. The metastability of materials is chemistry dependent, with carbides and nitrides standing out as being particularly high in energy. Gray shaded elements are excluded from this analysis.

After selecting the 624 best-performing candidate structures, one for each PDF entry, we proceed to assess their validity by examining the values of energy and R-factor. As for energy, we

are interested in the difference between the structure's formation energy and the OQMD convex hull at the relevant composition with this structure included. If this value is 0 meV/atom, then our candidate structure is stable and thus highly plausible. However, metastable compounds with nonzero convex hull distances are also common in nature. Although not all hypothetical structures with nonzero convex hull distances can be synthesized, they should be considered as potentially valid solutions in our structure search. Analyses of experimentally known metastable compounds calculated by DFT have revealed that, while most metastable compounds are within 100 meV/atom of the convex hull, the values of convex hull distance are highly dependent on chemistry. [58, 59] In Figure 2.5, for each element, we plot the 90th percentile convex hull distance for ICSD compounds containing that element. There is a stark contrast in the convex hull distances as a function of element; carbides and nitrides are especially metastable. [60–62] We thus opt to use these values of convex hull distance as cutoff values in determining whether the structures we obtain from the prototype searching method are valid based on energy. Specifically, for a compound of interest, *e.g.*Ba_2CeSnO_6 (PDF #: 00-056-0332) solved in this work, we use the highest of the four 90th percentile convex hull distance values as the cutoff: 116 meV/atom for oxygen. Since our best-performing candidate structure for Ba_2CeSnO_6 is 102 meV/atom above the convex hull, we deem this structure valid based on energy.

In Figure 2.6a, we plot the convex hull distances of 624 compounds with structures obtained by prototype searching in this work, along with those of ICSD compounds. If all 624 of these compounds were correctly solved, then we would expect that their convex hull distances would line up well with the ICSD distribution. Although the proportion of our compounds that lie on the convex hull is high (277 compounds within 5 meV/atom of the hull), this proportion is shy of the ICSD, where 61% of compounds are within 5 meV/atom of the convex hull. We find that 543 compounds (87%) pass our validation criterion for energy, compared to 93% of the ICSD.

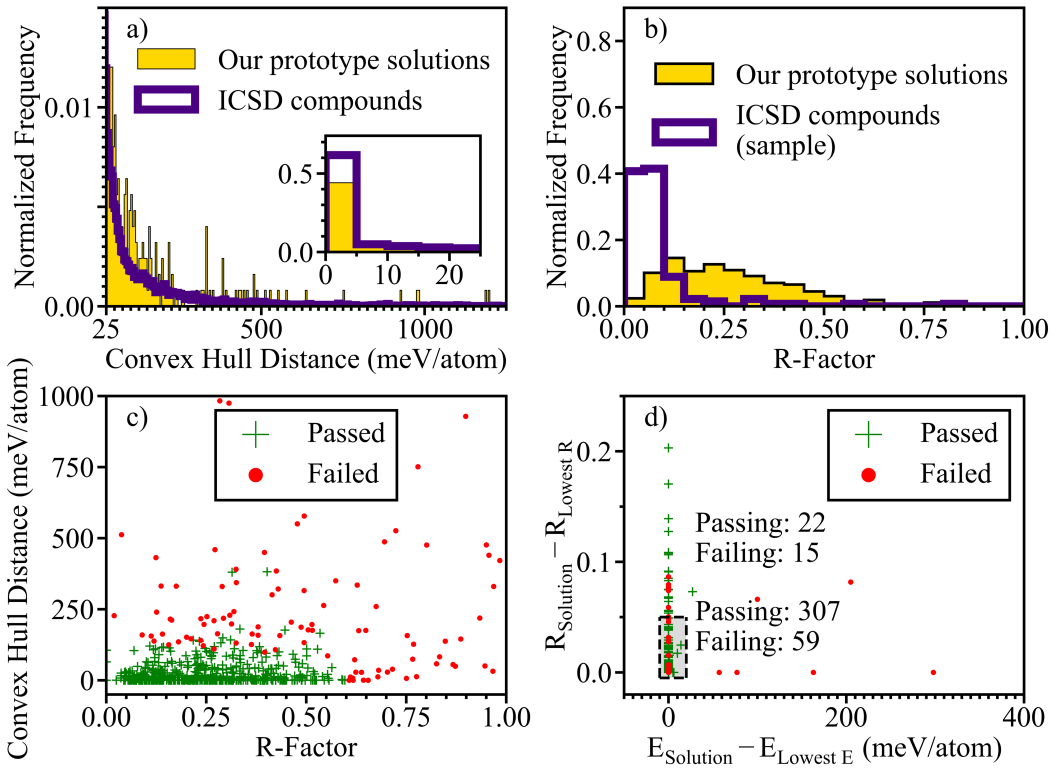

Figure 2.6: a) Distribution of convex hull distances of 624 compounds with structures obtained by prototype searching in this work, compared to 23247 ICSD compounds that have been calculated in the OQMD. Inset is the same distribution between 0 and 25 meV/atom; almost half of the 624 compounds lie within 5 meV/atom of the convex hull, somewhat shy of the ICSD. b) Distribution of R-factors of 624 compounds with structures obtained by prototype searching in this work, compared to 136 randomly chosen solved compounds from the PDF. c) Convex hull distances and R-factors of 624 compounds with structures obtained by prototype searching. Green pluses and red dots are compounds that passed and failed validation checks, respectively. d) Discrepancies in the best-performing candidate structure energies and R-factors with the lowest-energy and lowest-R candidate structures. Only cases with multiple candidate structures are shown (403 of 624 PDF compounds). The vast majority (91%) of cases lie within the shaded region; in other words, the best-performing candidate structure usually is close to the lowest energy and lowest R-factor of all candidate structures. Cases with low discrepancy are also more likely to pass our validation checks (84%) than cases with high discrepancy (59%).

However, we also expect that the prototype searching method will fail to solve compounds that are "one-hit wonders," *i.e.*compounds that do not share a prototype with any other in the ICSD

(described in Section 2.3.1). Since as many as 17% of known compounds are one-hit wonders, this inevitable shortcoming of our approach might explain why our convex hull distances are higher than those of ICSD compounds, on average.

R-factors of all 624 compounds are plotted in Figure 2.6b. To give context to our R-factor values, we overlay a distribution of R-factors for 136 randomly selected already-solved ICSD compounds with diffraction patterns stored in the PDF. With a median value of 0.25, our structures have higher R-factors overall than the already-solved structures (median of 0.06). We argue that this discrepancy does not suggest an issue with our prototype solutions, because many of our solutions with R-factor greater than 0.05 are clearly the right answers by inspection. For example, 44 of our compounds evidently have the elpasolite (K_2NaAlF_6) prototype, since they have A_2BCD_6 stoichiometry, space group of $Fm\bar{3}m$, and four formula units per unit cell. The only other possible prototype is typically very high in energy. Indeed, we find that 17 of the elpasolite compounds lie on the convex hull, despite R-factors ranging from 0.05 to 0.52. The high R-factors are not due to any issue with our refinement code either; despite elpasolite having only one degree of freedom to refine (the x coordinate of the 24e site), we still obtain high R-factors. We argue that the high R-factors highlight an issue with the diffraction patterns, not with our prototype searching approach. Because we cannot impose a strict R-factor validation criterion, we look to the relationship with energy values to decide on a cutoff R-factor value. Stable compounds tend to have low R-factors: 51% of compounds with R-factor below 0.1 lie on the convex hull; 54% with R-factor between 0.1 and 0.2 lie on the convex hull; 50% between 0.2 and 0.3. Following these intervals, we have 41%, 36%, 36% between 0.5 and 0.6, 24% 0%, 0%, and 0% between 0.9 and 1.0. As the proportion of stable compounds begins dropping off at 0.6, we opt to use an R-factor of 0.6 as the cutoff value for validation. This works out to be a generous cutoff value: 580 of our compounds (93%) have R-factor less than 0.6.

Combining our validation checks, we declare that 520 of 624 (83%) of our compounds are "solved" based on low convex hull distance and R-factor less than 0.6. The convex hull distances and R-factors of all 624 compounds are plotted in Figure 2.6c. Although most of our compounds simultaneously pass both energy and R-factor validation criteria, there are cases that pass only one of the criteria. Compounds with high energy and low R-factor might have structures that happen to exhibit a close match to diffraction data while being theoretically unphysical. On the other hand, compounds with low energy and high R-factor could be polymorphs of the "true" structure observed in experiment. It is also possible that compounds with low energy and high R-factor are, in fact, correctly solved; indeed, we are using an atypically high cutoff for R-factor. Despite the high R-factors, we argue that the R-factors are helpful in distinguishing structures that best match experimental data. In Figure 2.6d, we demonstrate that even though many of our structures have high R-factor, they are most often both the lowest-energy and lowest-R-factor candidate out of all possible candidates. Considering 403 cases where more than one possible candidate structure exists, we find that 366 (91%) of our best-performing candidates lie within the shaded region, *i.e.*are within 20 meV/atom of the lowest-energy candidate and 0.05 of the lowest-R-factor candidate. Compounds that pass our validation criteria are even more likely to lie within the shaded region (93%) than failing compounds (80%). This result demonstrates that even when all candidate structures have high R-factor, we can still use R-factor to distinguish the best structure from other candidates; however, DFT energy is often helpful in determining which candidates are physical.

Upon inspecting our prototypes selected for the PDF compounds, we noticed that they are quite often chemically similar to other ICSD compounds with the same prototype. For example, the solution to Ag_7SbS_6 (PDF #: 00-021-1333) is the prototype of Ag_7AsS_6, found in ICSD. We can quantify "chemical similarity" by taking advantage of a data mined Pettifor chemical scale developed by Glawe and co-workers. [56] They computed a chemical similarity metric P_{AB} for all

pairs of elements A and B on the periodic table. To compute the chemical similarity between two compounds, *e.g.*Ag_7SbS_6 and Ag_7AsS_6, we take the product

$$P = \prod P_{AB} \tag{2.3}$$

of chemical similarities of the closest-matching element pairs in the two compounds, setting $P_{AB} = 1$ when the elements are identical and 0 if the element pairs rarely or never occur in the ICSD. For all of our chosen prototypes, we searched for the ICSD compound of the same prototype with the highest chemical similarity; the results are plotted in Figure 2.7. The trends in the plots demonstrate that compounds that pass our validation criteria are more likely to be chemically similar to ICSD compounds than compounds that fail. The chemical similarities we find here give us an extra layer of confidence in our solutions.

All 521 compounds solved in this work are provided in the Supplemental Material of Ref. 33 along with a complete tabular summary of all 624 attempts. In addition, all compounds can be found in the OQMD, which can be accessed via the web at oqmd.org or directly downloaded. As there are 36807 unique ICSD compounds already in the OQMD, we have expanded the set of all experimentally observed compounds in the OQMD by 1.4%. The simplicity and efficiency of the prototype searching method presented in this paper has thus enabled us to significantly expand the set of experimentally observed compounds accessible to DFT. It will be of interest to further study the properties of these materials. For example, as shown in Figure 2.8, 284 of our solved compounds have nonzero bandgaps within 4 eV, making them potential candidates for semiconductor applications. In addition to the 521 newly solved compounds, we find 33 PDF "unsolved" compounds where there is either no matching prototype in the OQMD or no prototype matching the reported space group that passes our validation checks, but there is solution with a

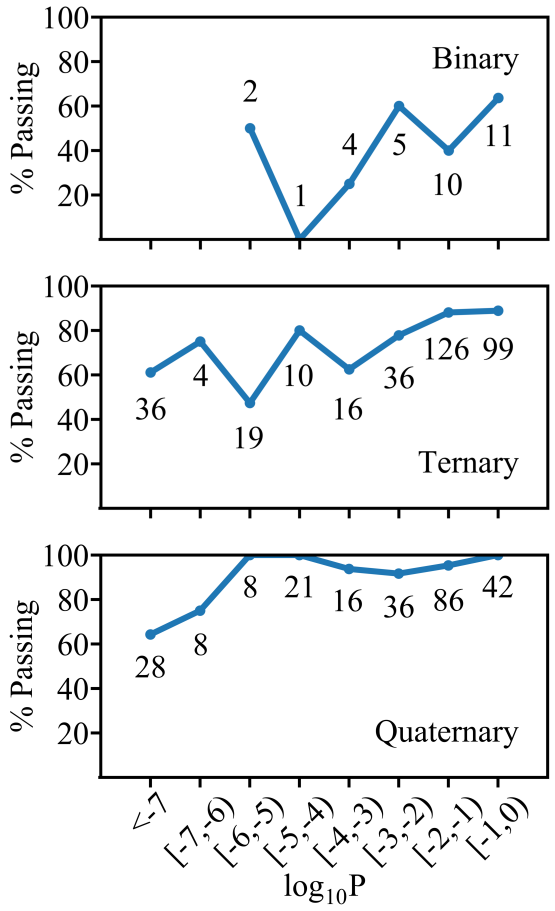

Figure 2.7: The percentage of the prototypes chosen for each of 624 PDF compounds that pass our validation criteria, plotted against the chemical similarities P of these compounds to ICSD compounds of the same prototype. The chemical similarities P are binned by decades on a log scale; P is defined in equation 2.3. The top, middle, and bottom plots focus on binary, ternary, and quaternary compounds, respectively. The numbers of compounds that fall within each range of chemical similarities are shown beside the data points. The trends demonstrate that compounds that pass our validation criteria are more likely to be chemically similar to ICSD compounds than compounds that fail.

different space group within the same crystal system that not only passes our validation checks but also already exists in the ICSD. These 33 solutions are provided in a separate table in the Supplemental Material of Ref. 33.

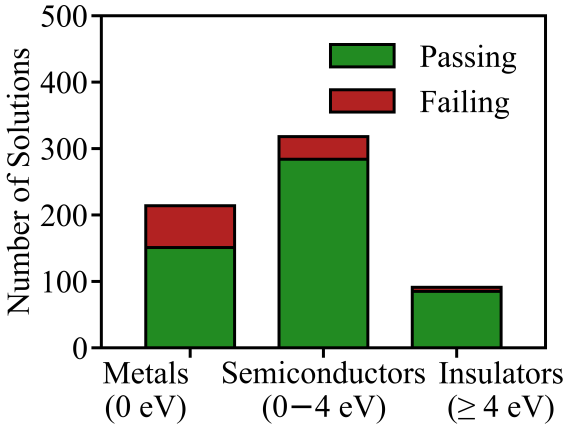

Figure 2.8: 151 of our solutions that pass validation criteria are metals (0 eV), 284 are semiconductors (0-4 eV), and 85 are insulators ($\geq$4 eV); band gap was not determined for 1 solution.

2.3.5 Detailed Description of Selected Solutions

In this section, to demonstrate the prototype searching method at work, we discuss nine PDF compounds that we solved. An illustration of the solved compounds is shown in Figure 2.9. All of the nine compounds in this section pass our validation criteria of energy and R-factor and are chemically similar to other compounds in the ICSD with the same prototype. For some of these compounds, the paper describing the diffraction experiment stated the name of the prototype that matches our solution but did not present atomic coordinates. Although the prototypes of these compounds were already known, our prototype searching method enabled us to obtain atomic coordinates for all structures and expand the OQMD.

2.3.5 Hf_8Bi_9

In the PDF entry for Hf_8Bi_9 (#: 00-051-0679), a diffraction pattern is supplied along with a space group (P4/nmm), unit cell, and formula units (Z = 2), but atomic coordinates are missing. [63] Because the atomic coordinates are missing, this compound did not previously exist in the ICSD

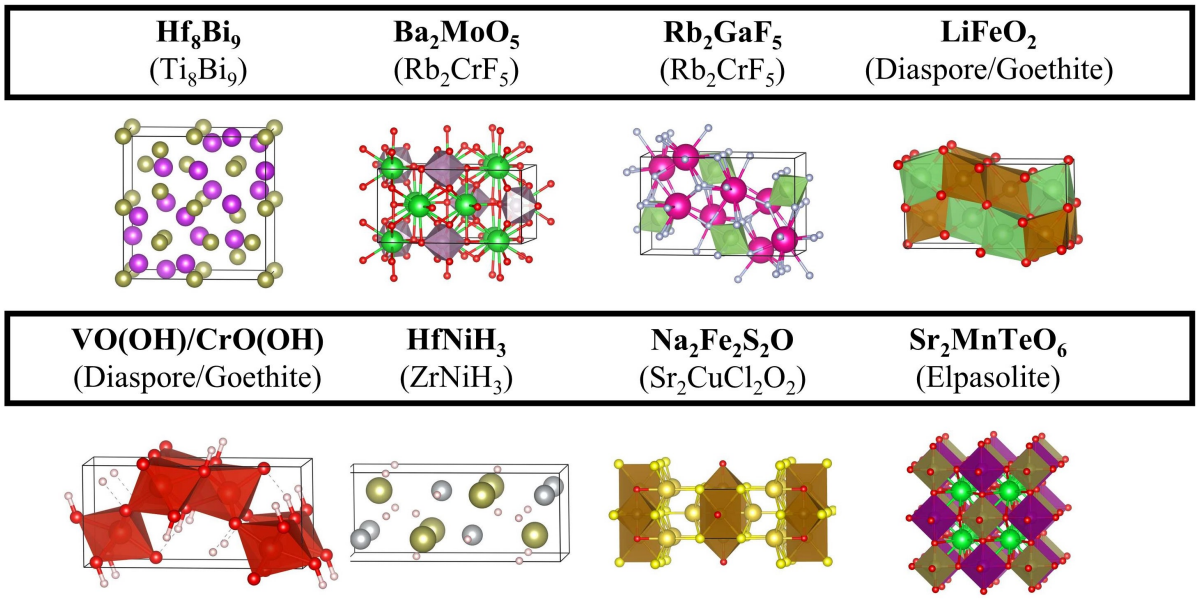

Figure 2.9: Crystal structures of 9 of the 521 materials solved using prototypes. The compositions of the solved materials are in bold, and the prototypes are in parentheses. Note that some of the solutions presented here have the same prototype, specifically Ba_2MoO_5 and Rb_2GaF_5 as well as $LiFeO_2$, VO(OH), and CrO(OH).

nor OQMD and has thus been excluded from DFT studies. However, in the reference for this entry, the authors presented the then-new prototype Ti_8Bi_9, complete with atomic coordinates, and stated that Hf_8Bi_9 has the same prototype as Ti_8Bi_9. As Ti_8Bi_9 is indeed already in the OQMD, we use the prototype searching method to complete the structure of Hf_8Bi_9. Specifically, our crystal structure for Hf_8Bi_9 consists of the unit cell parameters provided by the PDF entry for Hf_8Bi_9, and the DFT-relaxed atomic coordinates of Bi plus the atomic coordinates of Hf substituted for Ti in the already-solved compound Ti_8Bi_9. We find that this structure matches the diffraction pattern well (R-factor = 0.21, see Figure 2.1) and is on the convex hull (Figure 1.1).

2.3.5 *Ba_2MoO_5 and Rb_2GaF_5*

The reference for Ba_2MoO_5 provided by the PDF (#: 00-025-0011) for this diffraction pattern describes the structure as isostructural with $K_2VO_2F_3$ with Pnma symmetry and 4 formula units per unit cell but does not provide atomic coordinates. [64] The OQMD prototype Rb_2CrF_5 is indeed isostructural with $K_2VO_2F_3$ in that it has the same space group and Wyckoff site occupancies (though with different stoichiometry), and, with the elements Ba, Mo and O substituted in, lies on the Ba-Mo-O convex hull and has an R-factor of 0.16, indicating it is a highly plausible solution. An existing ICSD compound, Ba_2WO_5, has the same prototype and is highly chemically similar to Ba_2MoO_5 (P = 0.25). We also considered three other candidate prototypes: $BaSi_2O_5$ (convex hull distance = $+$19 meV/atom, R-factor = 0.28), KPd_2F_5 (hull distance = $+$102 meV/atom, R-factor = 0.42), and CsN_2H_5 (DFT failed to converge, R-factor = 0.19). Since the Rb_2CrF_5 prototype has both the lowest energy and lowest R-factor out of all candidates and passes our validation criteria, we deem it to be the solution to Ba_2MoO_5.

The compound Rb_2GaF_5 from the PDF (#: 00-032-0914) has the same story. [65] The prototype Rb_2CrF_5 is the solution because it is on the convex hull and has R-factor of 0.43, lower than other candidates. The ICSD compound Rb_2FeF_5, with the same prototype, is the most chemically similar to Rb_2GaF_5 (P = 0.04).

2.3.5 *$LiFeO_2$ polymorph, VO(OH), and CrO(OH)*

While several polymorphs of $LiFeO_2$ are known, to our knowledge there are no reports of the atomic coordinates of the Pnma polymorph of $LiFeO_2$ listed in the PDF (#: 00-052-0698), and consequently its structure was not previously in the OQMD. The reference listed in the PDF reported that the goethite polymorph of $LiFeO_2$ is rechargeable in lithium cells. [66] We find that goethite, or FeO(OH), is the correct prototype for this polymorph of $LiFeO_2$ when Li atoms are

substituted for H, since the convex hull distance is only $+36$ meV/atom and the R-factor is 0.13. We reject another candidate, YPd_2Si (convex hull distance = $+272$ meV/atom, R-factor = 0.29), because it is highly unstable.

Single crystals of VO(OH) (PDF #: 00-011-0152), found in montroseite, were examined by x-ray crystallography in 1953 were found to be isostructural with diaspore, or AlO(OH). [67] Diaspore and goethite are the same prototype (as is chalcostibite). An incomplete structure for VO(OH) having only V and O positions can be found in the ICSD; [68] H positions are missing, presumably since they cannot be detected in the x-ray pattern, and as a result, the properties of VO(OH) have not been studied with DFT. We obtain a complete structure for VO(OH), including H positions, by substituting V, O, and H into the sites of the diaspore structure and find it to be nearly stable (convex hull distance = $+8$ meV/atom, R-factor = 0.48). We similarly apply our prototype searching method to fill in the H coordinates of the CrO(OH) structure (PDF #: 00-025-1497), which was previously found to resemble diaspore. [69] Our structure for CrO(OH) is close to the convex hull ($+8$ meV/atom), but has poor match to diffraction pattern (R-factor = 0.62).

2.3.5 $HfNiH_3$

We report several stable hydrides in this work, including four lanthanide hydrogen chalcogenides. It is tricky to solve the hydrogen positions from x-ray diffraction data since hydrogen scattering is too weak to detect in an x-ray diffraction pattern. In the case of $HfNiH_3$ (PDF #: 00-047-1412), the peak indices could be matched to those of space group Cmcm. The authors inferred that the H atoms situate within the HfNi structure (space group Cmcm, 8 atoms per unit cell). [70] Separate DFT studies of $HfNiH_3$ utilized the assumption that H atoms occupy octahedral and tetrahedral interstices between Hf and Ni atoms in order to estimate the positions of H. [71, 72] We find ten unique prototypes having Cmcm space group and 20 atoms per cell, but the best

performing prototype is that of $ZrNiH_3$ (convex hull distance = 0, R-factor = 0.45). This is indeed a superstructure of HfNi, in which nine Hf-H bonds constitute edge-sharing polyhedra. Notably, the other nine prototypes with much higher energy are not hydrides. The $ZrNiH_3$ structure in the OQMD, complete with H positions, was obtained using neutron diffraction; [73] we utilize the solution from this past neutron diffraction study to complete the structure of $HfNiH_3$.

2.3.5 $Na_2Fe_2S_2O$

The diffraction pattern for the mixed anion compound $Na_2Fe_2S_2O$ was obtained through an ICDD Grant-In-Aid (PDF #: 00-065-0329). The atomic positions are missing from the entry, but the space group and number of formula units were reported to be I4/mmm and Z = 2, respectively. We conclude that the $Sr_2CuCl_2O_2$ prototype is a convincing solution. Since there are $3! = 6$ unique ways to arrange the elements Na, Fe and S onto the 4c, 4e and 4e Wyckoff sites of the $Sr_2CuCl_2O_2$ prototype, we check each one individually and find that the best arrangement is on the convex hull and has R-factor of 0.21. Interestingly, this arrangement has cation Na^{1+} occupying the anion Cl^{1-} site of $Sr_2CuCl_2O_2$, and likewise has anion O^{2-} occupying the cation Cu^{2+} site. Such an arrangement could be a direct consequence of the balancing of oxidation states in $Na_2Fe_2S_2O$. Other arrangements are significantly higher in energy, so they are ruled out.

2.3.5 Double Perovskites

Many materials presented in this work share the same prototypes with one another. Forty-four of the materials in this work have the elpasolite structure, or K_2NaAlF_6, which is an ordered double perovskite. Elpasolite is one of two prototypes that are possible given the experimentally known $Fm\bar{3}m$ space group, ABC_2D_6 stoichiometry, and 40 atoms per unit cell. The other possibility is the same as elpasolite but with the D_6 atoms occupying the 24d Wyckoff site rather than the 24e

Wyckoff site; this prototype is rare in the ICSD and is typically higher in energy by 1000-2000 meV/atom. Elpasolite is the most common quaternary prototype in nature, with 179 examples from the ICSD subset of the OQMD. All of the elpasolite-type compounds we present here are within $+114$ meV/atom of the OQMD convex hull (22 are on the hull), and have R-factors below 0.52 (28 had R-factors below 0.20), indicating that they were all stable or metastable and had reasonable pattern matches. For the metastable cases, the ground state is often a distortion of double perovskite; in the case of Sr_2MnTeO_6 (PDF #: 00-029-0897), the ground state is monoclinic (P21/c) double perovskite, which is 24 meV/atom lower in energy than the elpasolite decoration. Recently there has been interest in identifying more elpasolite compounds. It is difficult to perform high-throughput DFT calculations of elpasolite structures using elemental substitution, since there are millions of permutations. Faber *et al*.developed a machine learning model to predict the energies of elpasolite compounds, and found 90 structures on the convex hull, after considerable model training and DFT calculations of 2133 candidates. [74] We note that one of our 44 elpasolites is in their set of 90: Cs_2KGaF_6 (PDF #: 00-021-0849).

2.4 Discussion

Structure solution is a challenging roadblock to materials discovery. Thankfully, crystal structures are rarely unique, and a successful structure solution can often be obtained by searching among a relatively small number of prototypes as valid candidates. We apply this simple and inexpensive strategy to solve 521 structures taken from the PDF. Utilizing the OQMD as an exhaustive database of prototypes as well to validate the energetic stability of candidates along with R-factor, we have identified potential solutions to these materials, and we have a high degree of confidence in our solutions.

The prototype strategy employed in this work can be improved upon in many ways. One way

is to tweak the definition of a prototype to distinguish different structures more effectively. In our approach, we define the prototype of a structure as the combination of its stoichiometry, space group, and Wyckoff site occupancies. All structures from the OQMD sharing these characteristics are grouped into one prototype. However, within these constraints, there can be many degrees of freedom in atomic coordinates and lattice parameters, and it is possible for two structures with the same prototype, as defined in this paper, to in fact have very different local geometries, a problem described at length by Trimarchi *et al*.[75] Our workaround is to choose the OQMD compound whose structure, with its elements replaced by the target elements, gives the lowest R-factor, since the calculation of R-factor is nearly instantaneous compared to DFT. A more reliable workaround would be to devise a stricter prototype definition capable of properly distinguishing structures with different local geometries. For instance, some definitions apply additional restrictions on unit cell axial ratios and angles. [76] One could also quantify the difference between structures using a distance metric, such as one devised from radial distribution functions [77] or atomic/molecular matching algorithms [78–80] Moreover, if a given prototype has many internal degrees of freedom, one could conceivably develop an algorithm to optimize DFT and R-factor within the search space of that prototype.

Another way to improve the performance of the prototype searching method is to recommend the most plausible prototypes first, prior to evaluating them with DFT. There was no need to do so for this work, since constraining the search to the PDF-provided space group, composition, and number of atoms per unit cell of all solved materials reduced the number of candidate prototypes fewer than three in most cases. If, on the other hand, we could not constrain the search as much, there would have been too many candidates to evaluate. Existing techniques for recommending prototypes as candidates for an unsolved compound involve machine learning [81] as well as data-mined ion substitution. [55]

Furthermore, we suggest incorporating prototypes as initial guesses to structural optimization algorithms as a way to improve their performance. If an existing prototype is indeed the correct answer, as is the case for most compounds in nature, then optimization algorithms would converge immediately without wasting computational resources.

2.5 Conclusion

In this work, we outline a novel prototype searching method and use it to solve the structures of 521 PDF diffraction patterns. For each diffraction pattern, we obtain all prototypes in the OQMD satisfying the known stoichiometry, space group, and number of atoms per unit cell that are provided by the PDF, and select a structure based on DFT energy and R-factor. We then validate each structure by assessing its energetic stability with respect to competing phases in the OQMD as well as the R-factor. The 521 solved compounds, along with a table of descriptive details, can be found in the Supplemental Material of Ref. 33, and the compounds are also available in the latest release of OQMD. Identifying structures for these experimentally observed materials enables us to explore their properties from first-principles and unveil their potential for a wide variety of future applications.

CHAPTER 3
HOW TO DISCOVER STABLE INORGANIC COMPOUNDS MORE EFFICIENTLY

Most of this chapter is taken directly from our unpublished manuscript.

3.1 Background

The design space of inorganic compounds is unfathomably large, and only a very small fraction of compounds in the design space are energetically stable. For most of history, the gold-standard way to discover compounds has been to either find them in the ground or to synthesize them manually. However, these approaches are slow, and even our best efforts will fail to sufficiently sample the entire design space. It is now much faster to use computational methods, especially DFT, to compute the stabilities of candidate hypothetical compounds. DFT has had an impressive track record of accurately computing the formation energies of a wide range of experimentally known compounds. [12] One can then construct a convex hull of the formation energies of all competing phases [32] to determine which of the compounds are stable; compounds that lie on the convex hull are stable, and compounds that lie above but close to the convex hull are likely to be metastable, *i.e.*stable under certain environmental conditions. [58, 59] The DFT-calculated convex hull of inorganic compounds can now be computed with high precision due to the recent development of large DFT databases such as the OQMD, [7] Materials Project, [8] AFLOW, [9] Joint Automated Repository for Various Integrated Simulations (JARVIS), [10] Novel Materials Discovery (NOMAD), [82, 83] and Open Databases Integration for Materials Design (OPTIMADE). [11] For example, OQMD now contains DFT calculations of over one million compounds, both experimentally known compounds from the Inorganic Crystal Structure Database (ICSD) and hypothetical compounds. [31]

These developments have enabled the rapid discovery of thousands of new compounds, often with superior properties for applications in renewable energy and other technologies. [4, 39, 62, 74, 84–96] While DFT has greatly accelerated the discovery of stable compounds, its computational expense prevents us from studying all hypothetical compounds in design space. For this reason, we still have vast, largely untapped regions of design space where stable compounds have not yet been discovered. As a thought experiment, if we assume the remaining undiscovered stable compounds share prototype with an experimentally known compound (83% of experimental compounds share prototype with another compound), then by decorating 76 technologically relevant elements from the periodic table into the 10203 experimental prototypes, [33] we have 5055990 possible binary compounds, 425033800 ternary, 9463223600 quaternary, and 94683555000 quinary.

To accelerate the discovery of the remaining stable compounds, we turn to recommendation engines, or methods to identify likely-stable candidate compounds prior to DFT confirmation. Over the last few decades, many recommendation engines have been developed. Among the earliest examples are phenomenological structure maps, in which the structure prototypes of compounds are clustered according to elemental properties; [97–99] these structure maps were limited to common prototypes of binary alloys. Fischer *et al*.developed a recommendation engine, which they called the data mining structure predictor (DMSP), that is not limited to any particular set of chemistries or structures. [81] Their structure mapping is based on correlations between elemental compositions and the corresponding prototypes that appear in known phase diagrams. For example, supposing we didn't know the structure of Ni_3Pt, we could successfully predict that its prototype is that of Cu_3Au given that NiPt and $NiPt_3$ have the same prototypes as CuAu and $CuAu_3$, respectively. Hautier *et al*.developed a recommendation engine, which we will refer to as the ion substitution predictor (ISP), to exploit the apparent substitutability of certain ions in compounds with the same prototype in order to discover new ionic compounds. [55] For example, supposing we didn't know

the structure of SnSe, we could successfully predict that its prototype is that of SnS because S^{2-} and Se^{2-} play similar roles in chemical environments. Glawe *et al*.devised a related substitutability concept applicable to any type of inorganic chemistry, not just ionic compounds; [56] we will refer to this method as the element substitution predictor (ESP). Additionally, over the last decade we have seen the development of numerous machine learning (ML) models to predict formation energy. [100–118] In our work, we consider the version of iCGCNN (improved crystal graph convolutional neural network) presented in Ref. 103 due to its demonstrated low mean absolute error (MAE) of 46.5 meV/atom in predicting formation energies of 230000 diverse compounds from OQMD.

While many recommendation engines have been developed, little work has been done to compare their performances in predicting stable compounds in a side-by-side manner. Bartel *et al*.[119] compared seven formation energy ML models on the same Materials Project data sets, finding that models with lower MAE in predictions of diverse 85014 compounds exhibited better performance in recovering the stable compounds (*i.e*.convex hull distance of zero). However, when applied to a sparse chemical space of 267 Li-Mn-TM-O compounds (TM $\in$ {Ti, V, Cr, Fe, Co, Ni, Cu}), all models except for CGCNN (crystal graph convolutional neural network) [101] failed to correctly predict the 9 stable compounds.

In our work, we compare the performances of recommendation engines in recovering stable compounds from several chemical spaces for which OQMD has extensive DFT calculations. Specifically, the recommendation engines we examine are DMSP, ISP, ESP, and iCGCNN, and the chemical spaces we use are full and half Heuslers, binary AB_3 prototypes, and two distortions of ABO_3 perovskite; these are detailed in Table 3.1. As part of our systematic comparisons, we explore strategies to improve the performance of the recommendation engines. For example, we find that the ISP and ESP methods perform significantly better when we carry out an iterative feedback

Prototype	Source	No. expt. comps	No. expt. stable comps (i.e. ≤ 5 meV/atom of convex hull)	No. hypo. comps	No. hypo. stable comps
$L2_1$ Full Heusler	Ref 12	280	143	130106	1324
$C1_b$ Half Heusler		169	85	113186	258
$D0_3$ BiF_3		44	19	5656	40
$D0_{19}$ Ni_3Sn		38	28	5670	122
$D0_{22}$ Al_3Ti		23	14	5677	40
$L1_2$ Cu_3Au		296	178	5457	64
P4mm ABO_3 Perovskite	Ref 87	3	1	3207	11
Pnma ABO_3 Perovskite		153	64	3207	60

Table 3.1: Statistics of OQMD compounds with the prototypes of interest for this study.

loop where "newly predicted" hypothetical stable compounds are added to the set of known (initially experimental) compounds after each iteration. We also find that iCGCNN performs better with appropriate design of the training set, particularly consisting of experimental compound DFT energies and a representative sample of hypothetical compound DFT energies. Ultimately, with these strategies in effect, we find that iCGCNN is the best-performing recommendation engine, while ISP and ESP are very strong alternatives for the wide variety of chemical spaces under examination. Although we anticipate DMSP could become a strong choice when carried out in a feedback loop, we could not do so due to computational expense. Finally, we examine the current state of materials discovery in OQMD, which has been largely driven by the use of the above recommendation engines. Analysis using ESP suggests that, while many of the most likely-stable hypothetical compounds have been calculated in OQMD, there remain numerous uncalculated likely-stable compounds across a wide range of prototypes, some more than others.

3.2 Methods

We frame the goal of the recommendation engines, described in the following sections, as to compute a likelihood quantity P_{HC} of a hypothetical compound being stable. With likelihood quantities for every hypothetical compound, we can proceed to sort the compounds by likelihood and run DFT on the compounds with highest likelihood of stability. This process of computational materials discovery is illustrated in Figure 3.1. Since the exact values of P_{HC} can have a different interpretation depending on the recommendation engine, we are instead only concerned with the *sorting* of P_{HC} for the purpose of comparing recommendation engines side-by-side.

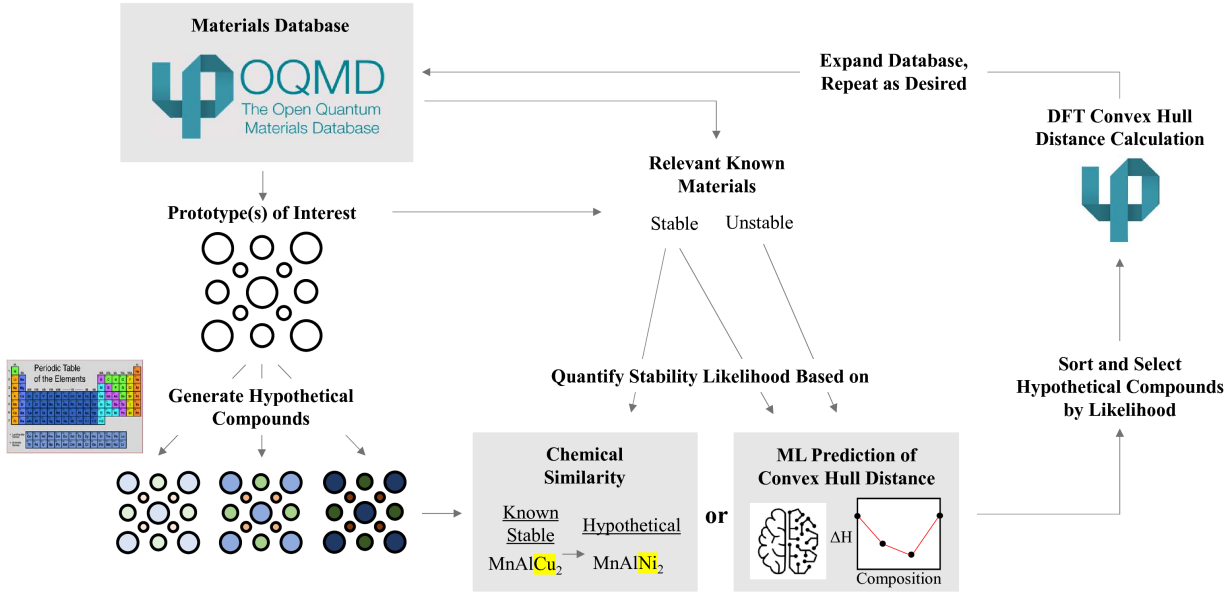

Figure 3.1: Schematic illustrating efficient computational materials discovery workflow employed in this section.

3.2.1 Data Mining Structure Predictor

The Data Mining Structure Predictor (DMSP) was developed in 2006 [81] as an attempt to automate the construction of Pettifor maps [97–99] and predict the phases of all possible compositions in all binary phase diagrams. We use the Magpie implementation of the DMSP method. [111] The output of this method is a likelihood quantity, $P(x_{ck}|\mathbf{X})$ that a phase diagram $\mathbf{X}$ with already-known phases also contains an unknown phase x_{ck}. As an example, we look at the experimental Ni-Pt phase diagram excluding Ni_3Pt. The phase diagram is represented as:

$$\mathbf{X} = (x_{E1}, ..., x_{EC}, x_{c1}, ..., x_{cn}) \tag{3.1}$$

where there are $C = 2$ elements (x_{E1} = Ni, x_{E2} = Pt) and n known phases binned by composition (x_0 = fcc, x_1 = fcc, $x_{0.5}$ = CuAu-type, $x_{0.25}$ = Cu_3Au-type). The phase we wish to predict is Ni_3Pt of type $x_{ck} = x_{0.25}$ = Cu_3Au-type. The probability of this phase is expressed as:

$$P(x_{ck}|\mathbf{X}) = \frac{P\left(\mathbf{X} + x_{ck}\right)}{P\left(\mathbf{X}\right)} \tag{3.2}$$

Likelihood quantities P are computed based on other phase diagrams $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_N}$ in the experimental database (ICSD compounds within 5 meV/atom of the convex hull). The likelihood quantities are expressed as a 2-order approximation to the generalized cumulant expansion

$$P\left(\mathbf{X}\right) = \frac{1}{Z} \prod_i P\left(x_i\right) \prod_{j<k} g^{(2)}\left(x_j, x_k\right) \tag{3.3}$$

where Z is a partition function to ensure normalization; $P\left(x_i\right)$ is the likelihood associated with one variable x_i, either element identity or prototype name; and $g^{(2)}\left(x_j, x_k\right)$ is the pair correlation

between two variables x_i and x_j. The subscripts i, j, k range across all possible element identities and prototypes in the experimental database, and we are looking for how frequently variables and variable pairs occur. The terms in equation 3.3 are given by

$$P\left(x_i\right) = \frac{N_{x_i} + 1/I}{N_{\mathrm{PD}} + 1} \tag{3.4}$$

$$P\left(x_i, x_j\right) = \frac{N_{x_i,x_j} + 1/\left(JK\right)}{N_{\mathrm{PD}} + 1} \tag{3.5}$$

$$g^{(2)}\left(x_j, x_k\right) = \frac{P\left(x_i, x_j\right)}{P\left(x_i\right) P\left(x_j\right)} \tag{3.6}$$

where N_{PD} is the total number of phase diagrams in the experimental database; N_{x_i} is the number of phase diagrams that contain x_i; N_{x_j,x_k} is the number of phase diagrams simultaneously containing x_j and x_k; and I, J, K are the numbers of possible values of x_i, x_j, and x_k, respectively. As a pratical requirement, prototypes are binned by stoichiometry; for example, only one prototype with stoichiometry equal or close to 0.25 can represent $x_{ck} = x_{0.25}$.

3.2.2 Ion Substitution Predictor

To explain the ISP approach, [55] we use CaF_2 and BaF_2 as an exmaple of two ionic compounds that share the same prototype, fluorite. Suppose for a moment that BaF_2 is not experimentally known but CaF_2 is. We predict a likelihood quantity for BaF_2 in the fluorite prototype based on the (quite common) substitution of Ba^{2+} into the Ca^{2+} sites of the CaF_2 structure. Given any Ca^{2+}-containing compound, regardless of its prototype, the likelihood that Ba^{2+} can be substituted into Ca^{2+} is given by

$$P\left(b|a\right) = \frac{\exp\left(\lambda_{a,b}\right)}{\sum_i \exp\left(\lambda_{a,i}\right)} \tag{3.7}$$

where a, b are the two ions (e.g. Ca^{2+} and Ba^{2+} respectively), i runs over all possible ions, and $\lambda_{a,b}$ is a weight related to the frequency that two ions appear on the same Wyckoff sites of known isostructural compounds. Each of the possible ion pairs is associated with a corresponding weight, and the weights are symmetric, i.e., $\lambda_{a,b} = \lambda_{b,a}$. Considering all ions between the two compounds, we express the likelihood as

$$P\left(x'|x\right) = P\left(a, b, ...|e, f, ...\right) = P\left(a|e\right) \cdot P\left(b|f\right) \cdot ... \tag{3.8}$$

where x' is the hypothetical compound, x is the known compound, and $a, b, ...$ and $e, f, ...$ are their respective ions. When a host ion stays the same, i.e. $P(i|i)$, we set its likelihood to 1. We use the set of weights Λ provided by the *pymatgen* module. [120] These weights were determined [55] so as to maximize the log-likelihood L of the experimental database D (ICSD), represented as a set of m ion pairs that are found in prototype-sharing compounds:

$$D = \left\{\left(a, b\right)^1, ..., \left(a, b\right)^m\right\} \tag{3.9}$$

$$L\left(D|\Lambda\right) = \sum_t^m \log P\left(a, b|\Lambda\right)^t \tag{3.10}$$

There are dozens of fluorite-type compounds in the ICSD, but of those, CaF_2 yields the highest likelihood value for BaF_2 because of their obvious chemical similarity. In general, to obtain a likelihood measure for any hypothetical compound x' like BaF_2, we compute

$$P_{\mathrm{HC}} = \max_x P\left(x'|x\right) \tag{3.11}$$

using all experimental fluorite compounds x in the database.

3.2.3 Element Substitution Predictor

The ESP [56] is conceptually similar to the ISP in that it concerns the substitution of chemically similar species between prototype-sharing compounds, but with a major difference that the ESP considers just element identity *without oxidation states* as possible species. In other words, whether Co is being substituted by Fe^{+2} versus Fe^{+3} does not affect likelihood; only 'Fe' matters. This allows for the treatment of metallic compounds, which don't obey valence charge balancing, in addition to ionic compounds. The likelihood of a pair of elements a, b substituting for one another is given by

$$P\left(a,b\right)=\sqrt{\frac{S_{a,b}^{2}}{\left(\sum_{i}S_{a,i}\right)\left(\sum_{j}S_{j,b}\right)}} \tag{3.12}$$

where $S_{a,b}$ is the number of occurrences in the ICSD where prototype-sharing compounds contain the two elements on the same Wyckoff sites, and the indices i, j run over elements from the periodic table. The likelihood of substituting a compound with multiple elements is given by Equation 3.8, as with the ISP method. We then use Equation 3.11 to compute the likelihood P_{HC} of a hypothetical compound given all experimental compounds of the same prototype. In this work, we use the $P\left(a,b\right)$ values provided by Ref 56, which were trained on ICSD data.

3.2.4 Crystal Graph Convolutional Neural Network

The iCGCNN [102] builds upon the original CGCNN [101]. The general framework is to model the crystal as a graph, with node embeddings containing atomic information and edge embeddings containing bond information. In the iCGCNN, the nodes are connected by their Voronoi neighbors, and the bond information is encoded in the attributes of the Voronoi polyhedra such as solid angle, area, and volume. To account for periodicity, multiple edges can exist between nodes. The crystal graph is then fed into a graph convolution neural network, where node and edge embeddings are

iteratively optimized via convolution functions, which are designed to capture both two- and three-body correlations between atoms. After the convolution steps, all node and edge embeddings are combined via a pooling layer and subsequently passed through a hidden layer neural network to predict the target property, in our case formation energy. We use the 'scale-invariant' version of iCGCNN [103] designed to predict formation energy after volumetric relaxation of the input crystal structure (which is always unrelaxed if the compound is not yet DFT-calculated). Note that scale-invariant version handles isotropic relaxation of the unit cell, i.e. preserving axial ratios and angles, and does not handle relaxation of individual atoms. In this version, the crystal graph is associated with an additional scale factor representing the smallest interatomic distance in the crystal, and this scale factor is simultaneously optimized and predicted along with formation energy. Initially, the unit cell is rescaled such that the scale factor becomes 1. Then, during the convolution steps, the scale factor is iteratively updated as a function of the node embeddings, and edge embeddings are rescaled according to the scale factor (specifically, facet areas and polyhedral volumes). The scale-invariant iCGCNN was found to exhibit lower MAE than CGCNN and the original iCGCNN on a diverse set of 230000 OQMD relaxed-volume formation energies using unrelaxed structures as input.

To predict new stable compounds, we use iCGCNN to predict their formation energies, then compute P_{HC} as the difference between their formation energies and the current OQMD convex hull energy at the same compositions. In contrast with DMSP, ISP, and ESP, here the most stable compounds should have lowest P_{HC} rather than highest. We note that to some degree, our performance assessment of iCGCNN, relying on these predicted convex hull differences, depends on the current state of the OQMD convex hull, which is constantly being refined over time as the database grows.

3.3 Results

3.3.1 Improving the Performance of ESP and ISP by Iterative Feedback Loop

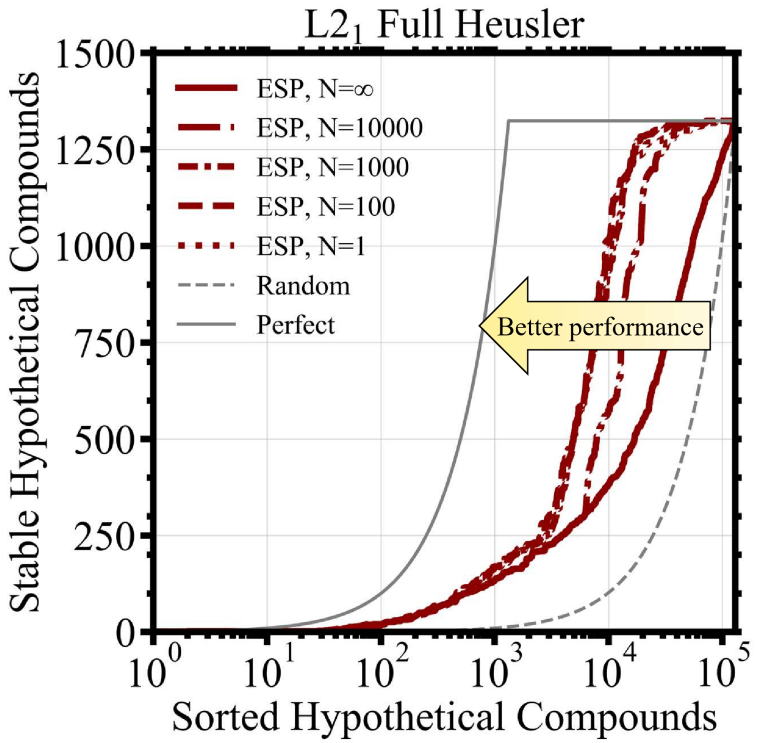

Figure 3.2: Performance of ESP-based recommendation engines in recovering stable $L2_1$ full Heusler-type compounds in OQMD. The x-axis is the index of hypothetical compounds sorted according to the recommendation engine, and the y-axis is the cumulative number of stable hypothetical compounds found upon calculating all compounds up to the sorting index. Shown here are ESP-based engines with varying sizes of N in the iterative feedback loop. Also shown are curves representing a random sorting of hypothetical compounds as well as perfect sorting (i.e. stable compounds are ranked highest in the sorting).

We choose to assess performance of recommendation engines with plots like shown in Figure 3.2. The x-axis of these plots are the indices of hypothetical compounds that are DFT-calculated in the order that the recommendation engine would suggest. For example, in the case of the full Heusler prototype (Figure 3.2a), there are a total of 130106 hypothetical compounds to calculate.

The y-axis is the cumulative number of hypothetical compounds that are stable up to the sorting index of the x-axis; e.g., there are a total of 1324 stable hypothetical full Heusler compounds. Also shown in the plots are the curves that would be obtained if the hypothetical compounds were randomly sorted or perfectly sorted (i.e. all stable compounds lie at the top of the list). Generally, a good-but-not-perfect recommendation engine would produce a curve that lies between the random-sorting and perfect-sorting curves.

Here, we demonstrate that executing the ESP in an iterative feedback loop improves its performance in discovering stable compounds. Note that the same concept of an iterative feedback loop can apply to the ISP method, but here we demonstrate it with ESP. In this strategy, illustrated in Figure 3.3, we first sort the hypothetical compounds by P_{HC} computed by equation 3.11, where initially $x \in X$ consists of stable experimental compounds of the relevant prototype. Then, we perform DFT calculations of the top N hypothetical compounds in the sorted-by-P_{HC} list. Next, we take the stable hypothetical compounds found among the N calculated compounds and add them to X. We then re-compute P_{HC} for the remaining uncalculated hypothetical compounds, run N more DFT calculations, and repeat until all hypothetical compounds have been calculated. The hypothesis behind this strategy is that stable compounds are clustered in the space of elemental composition, and while initial set X of experimental compounds may not contain all clusters, adding the other clusters to X during a feedback loop iteration will speed up discovery of the remaining stable compounds. The drawback of this strategy is that it de-parallelizes DFT calculations, which are normally performed on a high-performance computing cluster. Results for full Heuslers are plotted in Figure 3.2. We find that for the ternary full Heuslers, $N = 1000$ yields superior performance; not only does it remain reasonably parallel, but it discovers stable compounds at several times higher rate than $N = \infty$ (no feedback loop), up to the first 10000 or so DFT calculations.

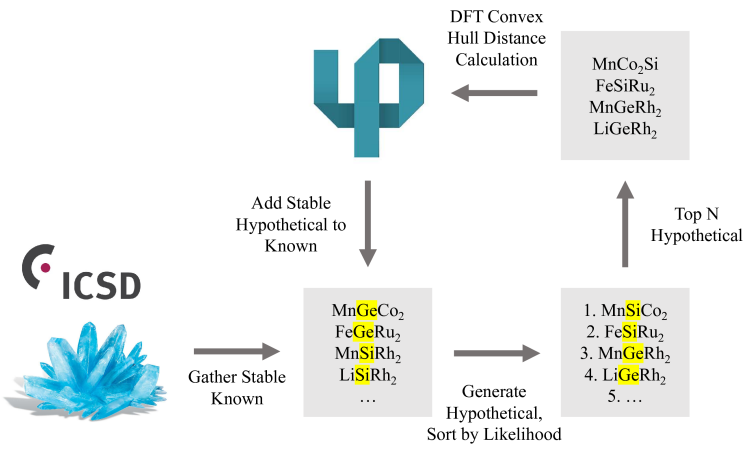

Figure 3.3: Schematic of the iterative feed loop to improve performance of ESP and ISP methods.

3.3.2 Improving the Performance of iCGCNN by Training Set Design

Here we explore how appropriate design of the training set can improve the performance of iCGCNN in finding stable compounds. In particular, we consider the inclusion of ICSD, i.e. a diverse set of 37525 experimentally known compounds calculated in OQMD, as well as a varying number of randomly sampled hypothetical compounds into the training set (see Figure 3.4 for an illustration of the training set). In Figure 3.5, we show the errors (MAE: mean absolute error and ME: mean error) of the iCGCNN models trained on the various training sets. When the training set consists of only ICSD compounds, both the MAE and ME are very high (respectively: 296 and +294 meV/atom; a positive ME indicates underestimation). The reason for these strikingly underestimated energies is that the ICSD training set contains only synthesizable low-energy compounds whereas the testing set contains mostly high-energy hypothetical Heusler compounds. We fix this problem by adding randomly selected hypothetical Heusler compounds to the training set. Upon adding just 0.1% (or 156) of hypothetical full Heusler compounds, the ME drops drastically to +58 meV/atom, while the MAE drops less drastically to 110 meV/atom.

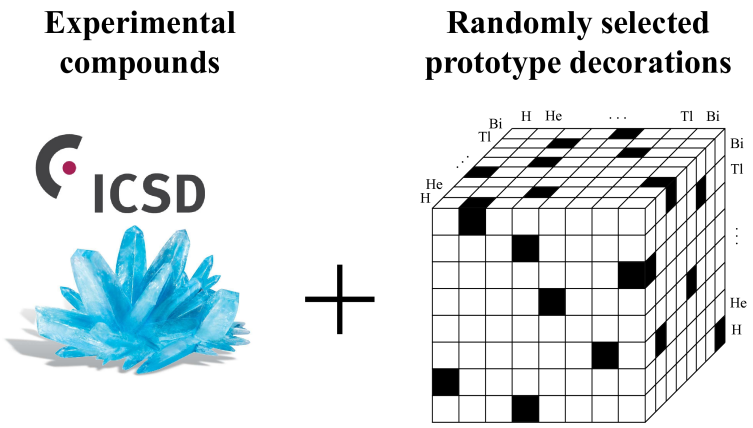

Figure 3.4: Training set design for good performance of iCGCNN model in predicting new stable compounds. The training sets in this work consist of experimentally known compounds taken from the ICSD in addition to randomly selected compositions within the search space.

Of course, adding more compounds to the training set comes with significant computational expense. To assess the cost-benefit of this strategy, we plot the performance of such models in finding stable compounds in Figure 3.6. For models with $C\%$ of hypothetical compounds added to the training set, the first $C\%$ of compounds along the x-axis are sorted randomly, and the number of stable compounds is approximately C% of the total; then, the performance rapidly increases as the remaining $1-C\%$ of compounds are sorted by P_{HC} according to the improved models. Obviously, the model with $C = 0.1\%$ performs better than $C = 1\%$ for at least the first 1% of compounds, but the 1% model overtakes 0.1% in performance within 10^4 sorted Heusler compounds.

When applying this strategy for a real materials search, it would not be clear which value of C to use. Our intuition is that $C = 1\%$ is a solid choice when it's affordable; otherwise, especially in ternary and quarternary search spaces, using just 0.1% can be sufficient to obtain greatly improved performance.

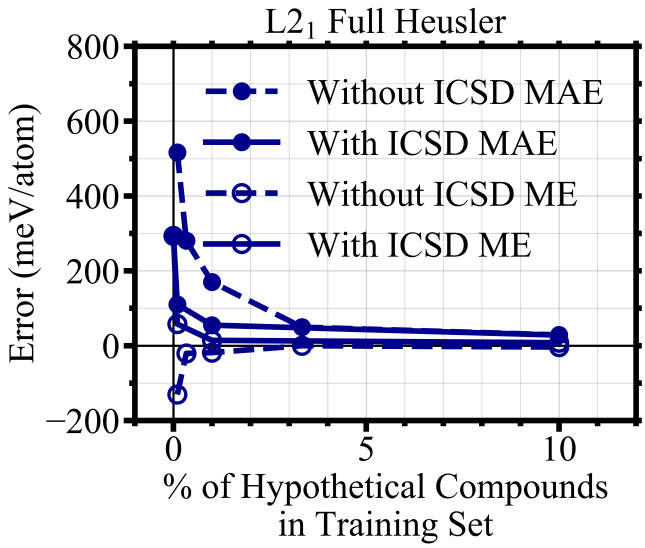

Figure 3.5: Performance of iCGCNN-based recommendation engines in predicting the formation energies of hypothetical $L2_1$ full Heusler-type compounds. The recommendation engines vary in terms of whether they include diverse ICSD compounds and percentages of randomly-chosen hypothetical compounds.

3.3.3 Comparing Performance of DMSP, ESP, and iCGCNN on Metallic Compounds

In Figure 3.7, we compare the performances of the ESP, iCGCNN, and DMSP in recovering hypothetical stable full Heusler compounds. Based on our conclusions from the previous two sections, we opted to use $N = 1000$ for ESP and $C = 1\%$ as well as ICSD compounds for iCGCNN training set. We see that the ESP method produces dozens of stable compounds through the first $\sim$1000 or so DFT-evaluated compounds, at which the iCGCNN method overtakes ESP in performance. Thus, we recommend ESP for the most immediate results when computational expense prohibits thousands of DFT calculations, and iCGCNN when thousands or more compounds can be calculated. On the other hand, the DMSP method did not perform best in any situation we studied. This could be because we did not explore ways to improve the performance of DMSP, as we found DMSP to be quite expensive to execute.

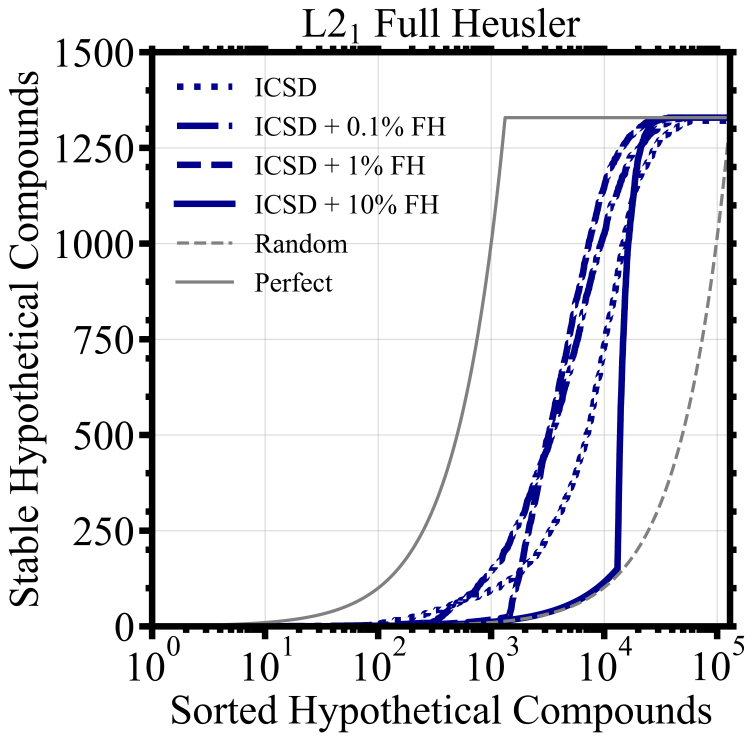

Figure 3.6: Performance of iCGCNN-based recommendation engines in recovering stable $L2_1$ full Heusler-type compounds. The x-axis is the index of hypothetical compounds sorted according to the recommendation engine, and the y-axis is the cumulative number of stable hypothetical compounds found upon calculating all compounds up to the sorting index. Also shown are curves representing a random sorting of hypothetical compounds as well as perfect sorting (i.e. stable compounds are ranked highest in the sorting).

3.3.4 Comparing Performance of ISP and ESP on Ionic Compounds

Up until this point, we excluded ISP from the performance comparisons because it would not be sensible to use ISP to discover the mostly metallic compounds that form in the full Heusler prototype. We now assess the performance of ISP using an ionic chemical system for which OQMD has extensive DFT data: ABO_3 perovskites. Since ISP and ESP methods work similarly, with the primary exception that ISP distinguishes oxidation states of chemical species, we seek to determine whether ISP is advantageous over ESP for the discovery of ionic compounds. The extensive

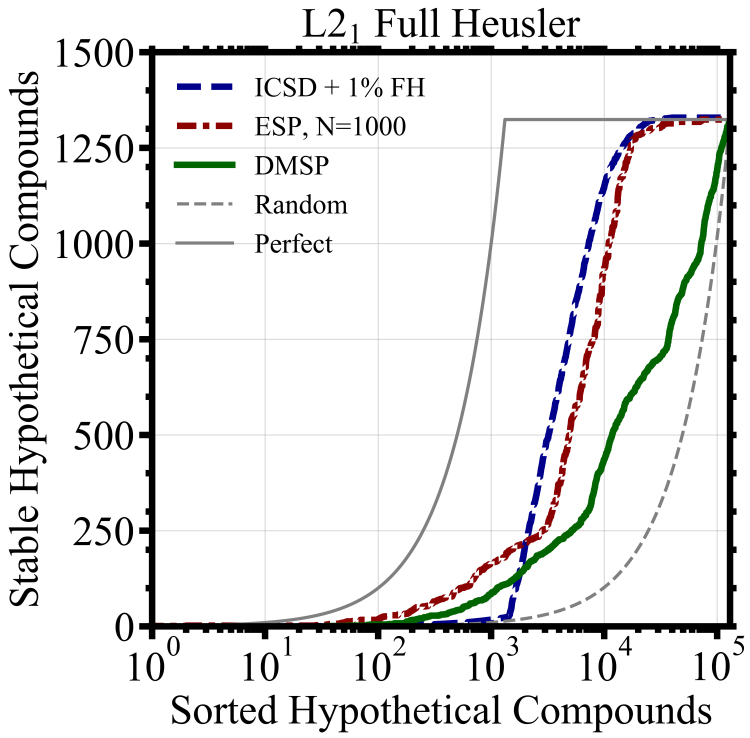

Figure 3.7: Performance of iCGCNN-, ESP-, and DMSP-based recommendation engines in recovering stable full Heusler-type compounds. The x-axis is the index of hypothetical compounds sorted according to the recommendation engine, and the y-axis is the cumulative number of stable hypothetical compounds found upon calculating all compounds up to the sorting index. Also shown are curves representing a random sorting of hypothetical compounds as well as perfect sorting (i.e. stable compounds are ranked highest in the sorting).

OQMD calculations of ABO_3 compounds are introduced in Ref 86. In short, the authors first calculated all cubic perovskite ($Pm\bar{3}m$) compounds ABO_3 such that A,B are 73 metals from the periodic table (although we exclude the 6 actinides from our study). They then took all compositions for which the cubic perovskite was within 500 meV/atom of the convex hull and calculated these compositions at 3 perovskite distortions: $R\bar{3}c$, P4mm, and Pnma, asserting that the remaining compositions are unlikely to be stable at any distortion. We follow this assertion and treat the remaining compositions as unstable for our study. In Figure 3.8, we plot the performance of ISP

and ESP methods with $N = 100$ iterative feedback loop on the Pnma perovskite distortion. The general finding is that both methods perform very well, with no clear advantage of one method over the other. We do note a possible advantage of ISP in certain situations where a stable compound contains an uncommon ionic species. For example, while sulfur is most often $2-$ oxidation state, it can in rare occasions be $6+$, as in SF_6. However, in our perovskite study, we could not find examples of 'rare-ion-containing' compounds that were more easily predicted by ISP than by ESP.

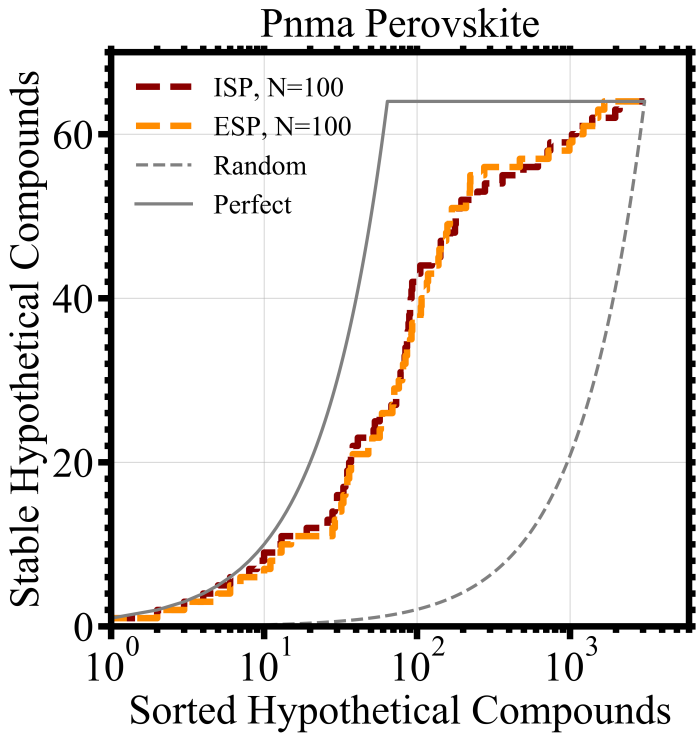

Figure 3.8: Performance of ISP- and ESP-based recommendation engines in recovering stable Pnma perovskite compounds. The x-axis is the index of hypothetical compounds sorted according to the recommendation engine, and the y-axis is the cumulative number of stable hypothetical compounds found upon calculating all compounds up to the sorting index. Also shown are curves representing a random sorting of hypothetical compounds as well as perfect sorting (i.e. stable compounds are ranked highest in the sorting).

3.4 Status of Materials Discovery in OQMD

Over the last decade, the OQMD has seen a rapid expansion to tens of thousands of hypothetical stable compounds, which today outnumber experimental stable compounds 2-to-1. [31] A large majority of the new stable compounds come from the execution of recommendation engines explored in this work. For instance, more than 10000 stable compounds come from the use of ESP by other research teams [84, 121]. We have also extensively used ESP, ISP, DMSP, and iCGCNN extensively in a number of published and unpublished search projects. With such progress made, we can stop to ask where we currently stand in the overall quest to identify all possible stable compounds. Here, we examine a few examples of prototypes for which many hypothetical stable compounds (but not all) have been added to OQMD: ZrNiAl, plotted in Figure 3.9. In this plot, the x-axis is the sorting of hypothetical compounds according to ESP method with feedback loop ($N = 1000$); in this case we include all possible compositions, not just the ones that have been calculated in OQMD as we did in the previous figures of this section. The left-hand side of the y-axis is the number of stable hypothetical compounds up to the corresponding sorting index of the x-axis, and the right-hand side is the number of calculated compounds up to the same sorting index.

In the case of ZrNiAl, nearly all of the first 1000 sorted compounds have been calculated in OQMD, indicating that recommendation engine-based searching of the ZrNiAl space has already been attempted, and over 600 stable compounds have been found. After 1000 compounds, the number of calculated compounds (blue curve) begins to deviate from the diagonal, representing all compounds (dashed gray line), indicating that there is a significant number of likely-stable hypothetical compounds that have not been calculated in OQMD yet.

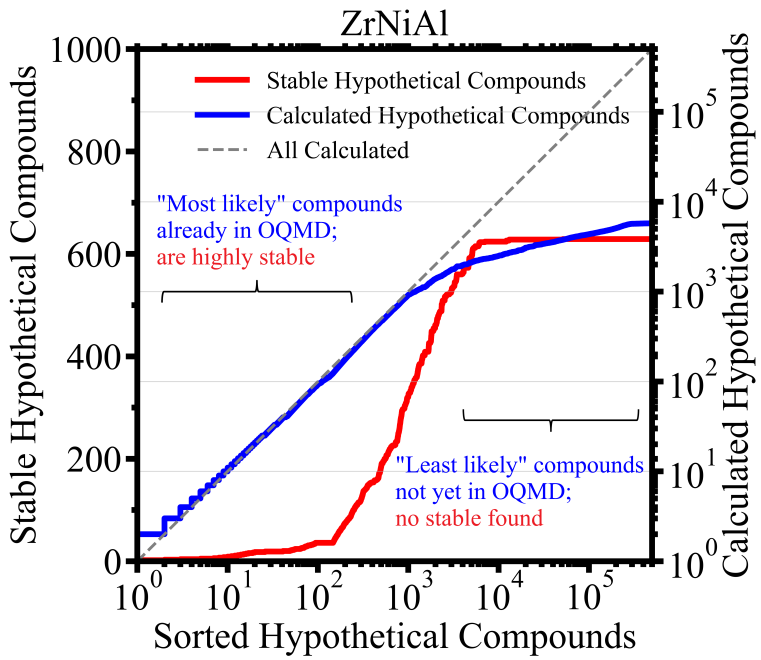

Figure 3.9: Discovered stable hypothetical compounds of type ZrNiAl, plotted against the sorting index according to ESP-based engine ($N = 1000$ for a, b, d and $N = 100$ for c). Also shown are the total number of hypothetical compounds (stable and unstable) calculated up to the sorting index, and a diagonal line representing all hypothetical compounds.

3.5 Conclusion

We have assessed the performance of automated recommendation engines in accelerating the discovery of stable inorganic compounds. We find that the iCGCNN, ESP, and ISP methods all perform strongly in sorting out the most-likely and less likely stable hypothetical compounds that are already calculated in OQMD. After implementing workflows to improve the recommendation engines, specifically optimal design of training set for iCGCNN and iterative feedback loop for ESP and ISP, we find that the iCGCNN broadly performs the best out of these methods. The DMSP method performed less strongly, partly due to the high computational expense of the Magpie implementation. Finally, we examined the status of recommendation engine-based materials searches

in OQMD, finding that a diverse range of compounds predicted to be stable by ESP have already been calculated but many more remain to be calculated.

CHAPTER 4
HIGH THROUGHPUT DISCOVERY OF STABLE INORGANIC COMPOUNDS

Some parts of this chapter are taken directly from our published paper. [31]

4.1 Background

One of the major advantages of the OQMD is its flexibility to allow researchers to explore hypothetical compounds (not previously known), which enhances our scientific understanding in providing candidate compounds for future applications. We are continually performing this exploration via high-throughput DFT calculations on a wide variety of hypothetical compounds. We often use "structure prototypes" as a blueprint for generating hypothetical compounds. We define the prototype of a structure as the combination of its stoichiometry, space group, and Wyckoff site occupancies. This blueprint is useful for two reasons. Reason #1 is that most known compounds share common prototypes, like NaCl-, CsCl-, and Heusler-type, among many others. In fact, 83% of all compounds in the ICSD share a prototype with another compound, and 27% of all compounds share a prototype with ≥ 50 other compounds. [33] Reason #2 is that we can use prototype to arrive at a stable or metastable crystal structure for hypothetical compounds. This is done by taking a known compound of that prototype, substituting in elements of the hypothetical composition, and using DFT to relax unit cell parameters and atomic coordinates along the symmetry directions.

It is natural to start with the most common prototypes as the blueprint for generating hypothetical compounds. In the OQMD, we used several common prototypes to conduct "exhaustive" high-throughput DFT of hypothetical compounds, *i.e.*, calculate nearly all possible compounds by

substitution of elements from the periodic table. The prototypes completed include binaries B1 (NaCl), B2 (CsCl), B3 (zincblende), B_h (WC), C15 ($MgZn_2$), $D0_3$ (BiF_3), $D0_{19}$ (Ni_3Sn), $D0_{22}$ (Al_3Ti), $L1_0$ (CuAu), $L1_1$ (CuPt), $L1_2$ (Cu_3Au); and ternaries $C1_b$ (Half Heusler), and $L2_1$ (Full Heusler). This has amounted to 393879 DFT calculations (54030 binary and 339849 ternary), and has been quite fruitful: 1973 of these hypothetical compounds (396 binary and 1577 ternary) are on the convex hull of stability and do not have an ICSD polymorph that is close in energy. On the other hand, most of the hypothetical compounds are above the convex hull and therefore much less likely to be synthesizable (see Figure 4.1). DFT calculations of unstable compounds are still useful to have for boosting training sets of machine learning (ML) models as well as general understanding. However, while we can continue these exhaustive high-throughput DFT calculations using other common prototypes, we cannot do this for all 10203 prototypes we have from ICSD. If we wanted to do exhaustive DFT for all prototypes up to five components using 76 elements in the periodic table ($Z \leq 83$ excluding noble gases, Tc, and Pm), then we would have to do trillions of DFT calculations.

Thankfully, algorithms based on data mining of materials databases as well as machine learning have been developed to accelerate the discovery of inorganic compounds. In Chapter 3, we discussed these methods at length, tested them, and developed protocols to exploit them most efficiently to discovery stable inorganic compounds. Importantly, we found that the ISP and ESP methods are highly efficient when conducted following an iterative feedback loop, where newly found stable compounds are added to the initial pool of stable compounds (which come from ICSD) and are subsequently used to re-computed the likelihood predictions. We used these two methods extensively to discover well over 10,000 new stable compounds and added them to OQMD. The ESP method has been used by other groups to discover tens of thousands of additional stable compounds, [84, 121] and we have added these to OQMD as well. Furthermore, we discovered more

than 10,000 additional compounds with a novel strategy we developed: by mixing two or more already-known stable compounds that differ by only one element. For example, the mixed compound $ZnCd_3S_4$ can hypothetically be formed by mixing Zn and Cd on their respective lattice site in the isostructural compounds ZnS and CdS. In this chapter, we provide a summary of the growth of OQMD and the new stable compounds that have been recently added, the majority of which are directly attributed to the aforementioned projects. We will also detail the mixing method and statistics of the mixed compounds we obtained from the method.

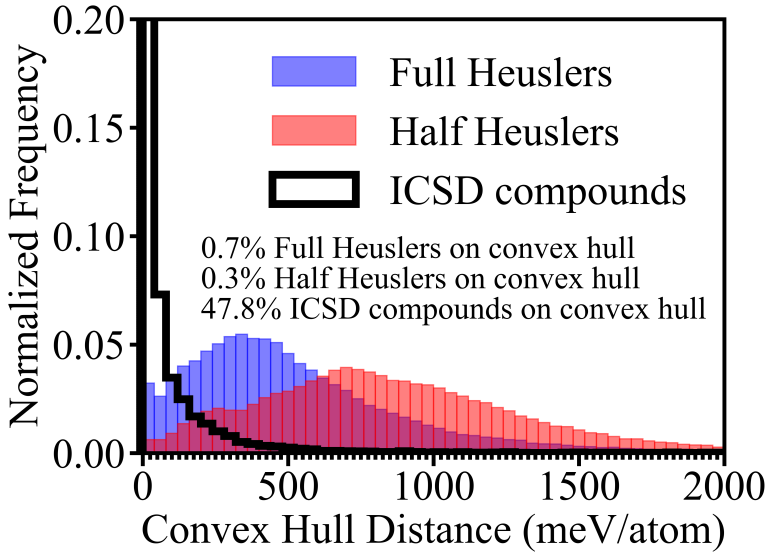

Figure 4.1: Distribution of convex hull distances of all hypothetical Full Heusler and Half Heusler compounds in the OQMD, compared to experimentally known ICSD compounds. Less than 1% of hypothetical Heusler compounds are stable under zero thermodynamic conditions, compared to 48% of ICSD compounds.

4.2 Overview of the Newly Discovered Compounds

The OQMD has grown tremendously since its introduction in 2013, and now consists of over 1,000,000 compounds. A growing fraction of the stable compounds in OQMD are hypothetical

(*i.e.*not experimental compounds from the ICSD). In Figure 4.2, we plot the fraction of stable compounds in OQMD that come from the ICSD. In this plot, all ICSD compounds are assumed to have been calculated prior to the year 2014, although many were calculated later than that. The point is to show that the fraction of stable compounds in OQMD that come from ICSD has been decreasing over time. Up until 2020, non-ICSD stable compounds were generated by various members of the Wolverton group as part of their research projects, as well as from exhaustive high-throughput DFT searches of common prototypes like NaCl, Heusler, etc. In 2020, a very sharp drop in the ICSD fraction occurred when we added compounds from the breakthrough paper that used the ESP method. [84]. We have also extensively used ISP and ESP to discover new stable compounds not reported in that paper from 2020 through today; this can be seen in Figure 4.3, where there is a sharp increase in the number of stable compounds discovered OQMD that occurred from 2020 onwards (the discovery dates from pre-2011 compounds are taken from the ICSD referenced papers). This sharp increase serves to show just how powerful our computational methods are in accelerating the discovery of new compounds. To obtain the new stable compounds, we followed the iterative protocol described in Chapter 3, where we started with all experimentally known stable compounds from ICSD to generate likely-stable hypothetical compounds.

The roughly 1,000,000 compounds in OQMD today are summarized in Figure 4.4. In Figures 4.4a and 4.4b, histograms of stabilities (*i.e.*convex hull distances) are plotted for all ICSD and all OQMD compounds, respectively. Since ICSD compounds are experimentally observed, naturally their stabilities tend toward zero, whereas hypothetical compounds ("non-ICSD" in 4.4b) generally have higher stabilities. However, largely due to data mining and ML based methods as previously discussed, many of the hypothetical compounds have small stabilities, *e.g.*below 20 meV/atom, indicating they might be synthesizable as metastable compounds.

In Figures 4.4c and 4.4d, histograms of the number of element types (binary, ternary, etc.)

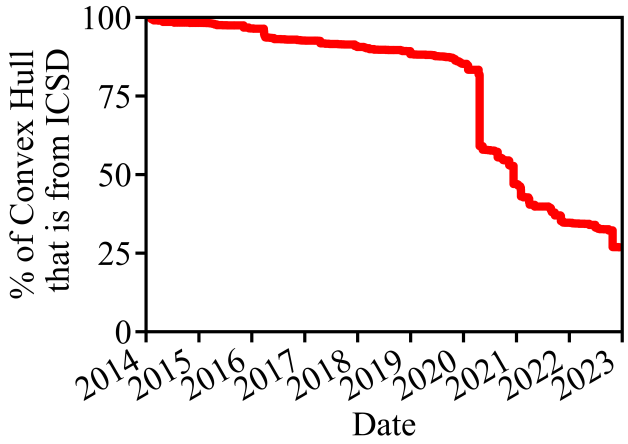

Figure 4.2: Percentage of all compounds on the OQMD convex hull that are sourced from the ICSD, plotted against the date of calculation. The OQMD started with ICSD compounds (100%), but ICSD compounds now make up just ∼30% of the OQMD convex hull. For this plot, all ICSD compounds are assumed to have been calculated prior to 2014, although many were calculated later than that.

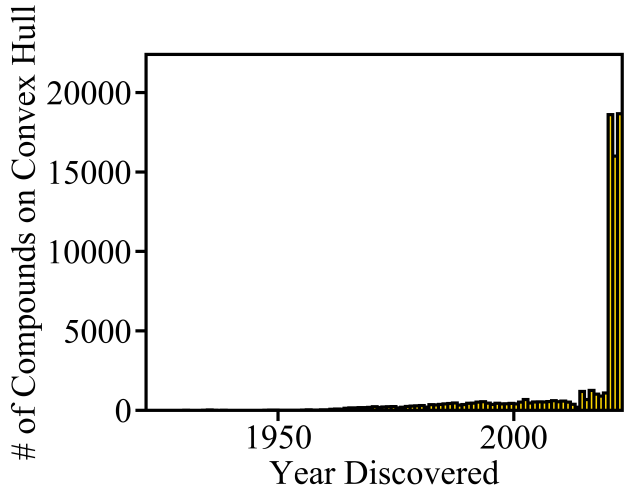

Figure 4.3: Number of unique compounds on OQMD convex hull, binned by the year they were discovered. If the compound came from ICSD, then the date of the original paper is used; otherwise, the date that the OQMD entry was created is used.

are plotted for all OQMD and all stable compounds, respectively. The "non-ICSD" compounds are more biased towards ternary and quaternary (3 and 4 components) than the ICSD compounds. This may be because much experimental effort has focused on perfecting the relatively small binary phase diagrams, whereas experimentalists have not yet had an opportunity to tap into the whole space of compounds that are possible based on combinatorial substitution of elements into ternary and quaternary prototypes. With the combined effort of high-throughput DFT and efficient data-mining and ML methods, we are now able to systematically evaluate and discover new compounds in the ternary, quaternary, and quinary space. Compositions greater than 5 components are certainly possible, but little exploration has been done thus far.

In Figure 4.4e, the histogram of the number of atoms per unit cell is plotted for all stable compounds. It is important to highlight that our high-throughput DFT efforts have thus far focused on relatively small unit cells (fewer than 15 atoms) because they are computationally cheap. Much work remains to be done in uncovering stable compounds with larger unit cells. In Figure 4.4f, band gaps are plotted for stable compounds; the ICSD and non-ICSD distributions are quite representative of one another, since our high-throughput DFT efforts have targeted a wide range of prototypes across the ICSD. In Figures 4.4g and 4.4h, distributions of space groups are plotted for ICSD and non-ICSD compounds, respectively; differences between these two distributions are due to the fact that we have combinatorially explored some common and small-unit-cell prototypes, such as Heusler, over other less common prototypes with larger unit cells.

The prototypes of the newly discovered compounds in OQMD are highly diverse. In Figure 4.5, we plot the distribution of the number of compositions per prototype that are stable in OQMD. The prototypes are clustered by crystal system, and the size of a bubble (prototype) is proportional to the number of compositions in that prototype. There are a handful of prototypes that are significantly larger than most others, since these are very common prototypes. In some cases, such as

"Quaternary ZrNiAl," the prototype rose to prominence by mixing compounds (see Section 4.3), *e.g.*by mixing two compositions with the ZrNiAl prototype. In 4.6, we show the distribution of prototypes in a different way, this time coloring the prototypes by the factor increase in the number of new hypothetical compounds over ICSD compounds, and grouping together prototypes with different stoichiometries but the same Wyckoff sites. This way, the example of ZrNiAl and quaternary ZrNiAl are treated as one prototype, which happens to be the most common prototype by far. Here, we can clearly see how many prototypes that are of interest for various applications have grown significantly in OQMD.

4.3 Mixed Ordered Compounds

A given lattice is often stabilized by compositions that are similar to one another; for example, it is no coincidence that ZnS and CdS both form in the zincblende structure, as Zn and Cd are both Group-XII cations. It would then be a fair question to ask whether Zn and Cd can mix together on the same sublattice. Often, such mixing is disordered under nonzero temperatures, and the entire lattice changes at a phase transition temperature. However, Zn and Cd is known to mix in an ordered way to form $ZnCd_3S_4$. In this case, the symmetry is broken from $F\bar{4}3m$ in ZnS and CdS to $P\bar{4}3m$ in $ZnCd_3S_4$. There are also experimentally reported ordered mixtures that retain the symmetry of their parent compounds; for example, in parent compounds Sr_2Si and Ca_2Si (Pnma), Sr and Ca occupy two Wyckoff orbits, and in the mixture SrCaSi (Pnma), Sr and Ca each occupy one of same two orbits.

We conducted a high-throughput DFT search for new ordered mixtures of known parent compounds. To start with, we identified all pairs of stable compounds with the same prototype and with compositions that differ by only one element, *e.g.*ZnS and CdS. To minimize computational expense, we considered only prototypes with 12 or fewer atoms in the primitive cell. We then con-

structed and calculated all structures in which the differing elements, in our example Zn and Cd, are mixed and ordered on the relevant sublattice, such that the number of atoms in the primitive cell remains 12 or fewer (no supercells were generated). In total, 88411 structures were calculated.

Since these compounds are simply ordered mixtures of the parent compounds, both in terms of composition and crystal structure, it would be reasonable to hypothesize that mixed compounds' properties "lie between" the corresponding parent compounds' properties. For example, the mixed compound's formation energy E_f (say, $ZnCd_3S_4$) might be a composition-weighted linear combination of the parent formation energies: $E_f(\text{ZnCd}_3\text{S}_4) = (1/4)E_f(\text{ZnS}) + (3/4)E_f(\text{CdS})$. To test this hypothesis, we plot in Figure 4.7 the mixed compounds' formation energies (4.7a), band gaps (4.7b), and magnetic moments (4.7c) against the composition-weighted linear combination of corresponding parent properties. The dashed black line represents the case in which the hypothesis is true.

The formation energies (Figure 4.7a) of mixed compounds differ from the corresponding parent formation energies with a MAE of 38 meV/atom. We will refer to this difference as a "mixing energy." The percentiles of the mixing energies, in order from 10th to 90th percentile in intervals of 10%, are -13, -3, 0, 2, 4, 8, 15, 28, and 69 meV/atom. Thus, the majority of mixed compounds have a positive mixing energy, indicating that DFT predicts they will decompose into the parents. The compounds with negative mixing energy are thermodynamically favored to mix, rather than decompose into the corresponding parent compounds. However, the negative mixing energies tend to be small (median of -4 meV/atom); for reference, the median decomposition energy of other stable compounds in OQMD is -23 meV/atom. Such a small mixing energy may indicate that the compound would be disordered at room temperature, rather than ordered at zero temperature as we've modeled them. Note also that the mixed compounds may not necessarily have the same crystal lattice as the parent compounds; there may be a different crystal structure at the same com-

position that is lower in energy than the parent lattice. Nevertheless, due to variational principle, DFT predicts that *some* structure is more stable than the parent structures at a composition between the parent compositions, which in itself is interesting and begs further investigation.

The band gaps (Figure 4.7b) of mixed compounds differ from parent band gaps to some degree. For the following statistics, we excluded cases where the mixed or the parent compounds are metals (band gap < 0.01 eV). The band gap differences have a MAE of 0.25 eV; the percentiles of the band gap differences, in order from 10th to 90th percentile in intervals of 10%, are -0.60, -0.31, -0.20, -0.13, -0.08, -0.04, -0.00, 0.05, and 0.12 eV. Thus, for the most part, the mixed compound band gaps are smaller than, but still relatively close to, the linear sum of parent band gaps. This has important implications for the field of "band engineering"; if a mixed compound is possible to make, it is likely to have a band gap that lies between the parent band gaps. In many applications, such as photovoltaics, the value of band gap is crucial for performance and much of the engineering focuses on finding materials within the appropriate window of band gap values. Interestingly, 1854 mixed compounds "lost" a band gap and become metals, meaning they don't have a band gap while the parents do; and 152 mixed compounds "gained" a band gap and become semiconductors/insulators, meaning they have a band gap while the parents don't. As can be seen in the plot, a significant number of these mixed compounds differ from the parents by upwards of 2 eV or more. This finding suggests that the mixing operation can, in some cases, greatly alter the properties of the compound.

The magnetic moments (Figure 4.7c) of mixed compounds differ from parent magnetic moments as well. For the following statistics, as we did for band gaps above, we excluded cases where the mixed or the parent compounds are nonmagnetic (magnetic moment < 0.01 μ_B/atom). The magnetic moment differences have a MAE of 0.05 μ_B/atom; the percentiles of the magnetic moment differences in order from 10th to 90th percentile in intervals of 10%, are -0.07, -0.01,

-0.01, 0.00, 0.00, 0.00, 0.01, 0.02, and 0.08 μ_B/atom. Thus, relatively speaking, the magnetic moments of mixed compounds are quite close to the corresponding parent magnetic moments. In addition, 918 of the mixed compounds "lost" magnetism, while 991 of the mixed compounds "gained" magnetism during the mixing operation. Thus, as with the band gaps, this phenomenon of losing or gaining magnetism is uncommon but does occur to a surprising degree. Note that in this study we only considered the nonmagnetic (no magnetic moment) and ferromagnetic (nonzero magnetic moment) states; we did *not* consider paramagnetic or antiferromagnetic states, due to computational complexity.

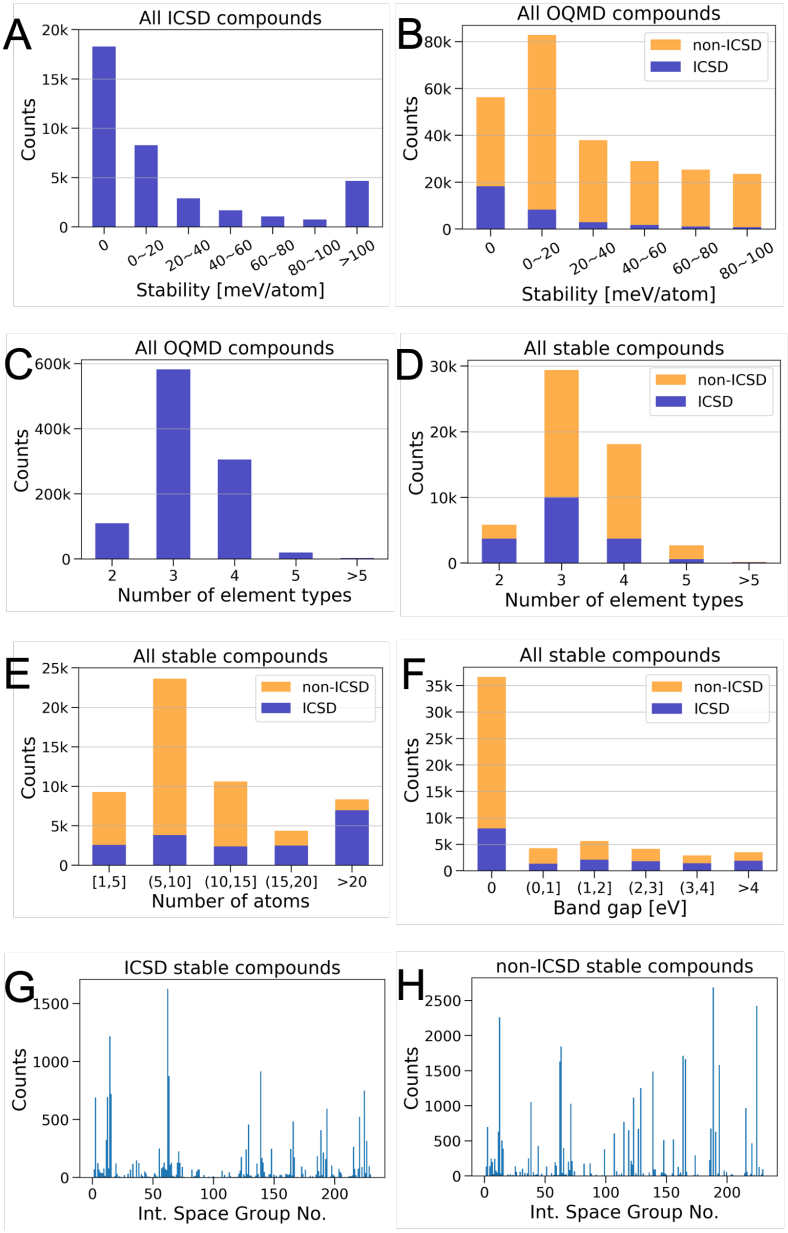

Figure 4.4: Statistical summary of current OQMD. (a) Stability (convex hull distance) of all ICSD compounds calculated in the OQMD. (b) Stability (convex hull distance) of all ICSD and non-ICSD compounds. (c) Number of elements types of all compounds in the OQMD. (d) Number of elements types, (e) number of atoms and (f) band gaps of all stable compounds in the OQMD. Space groups of (g) all stable ICSD compounds and (h) all stable non-ICSD compounds in the OQMD.

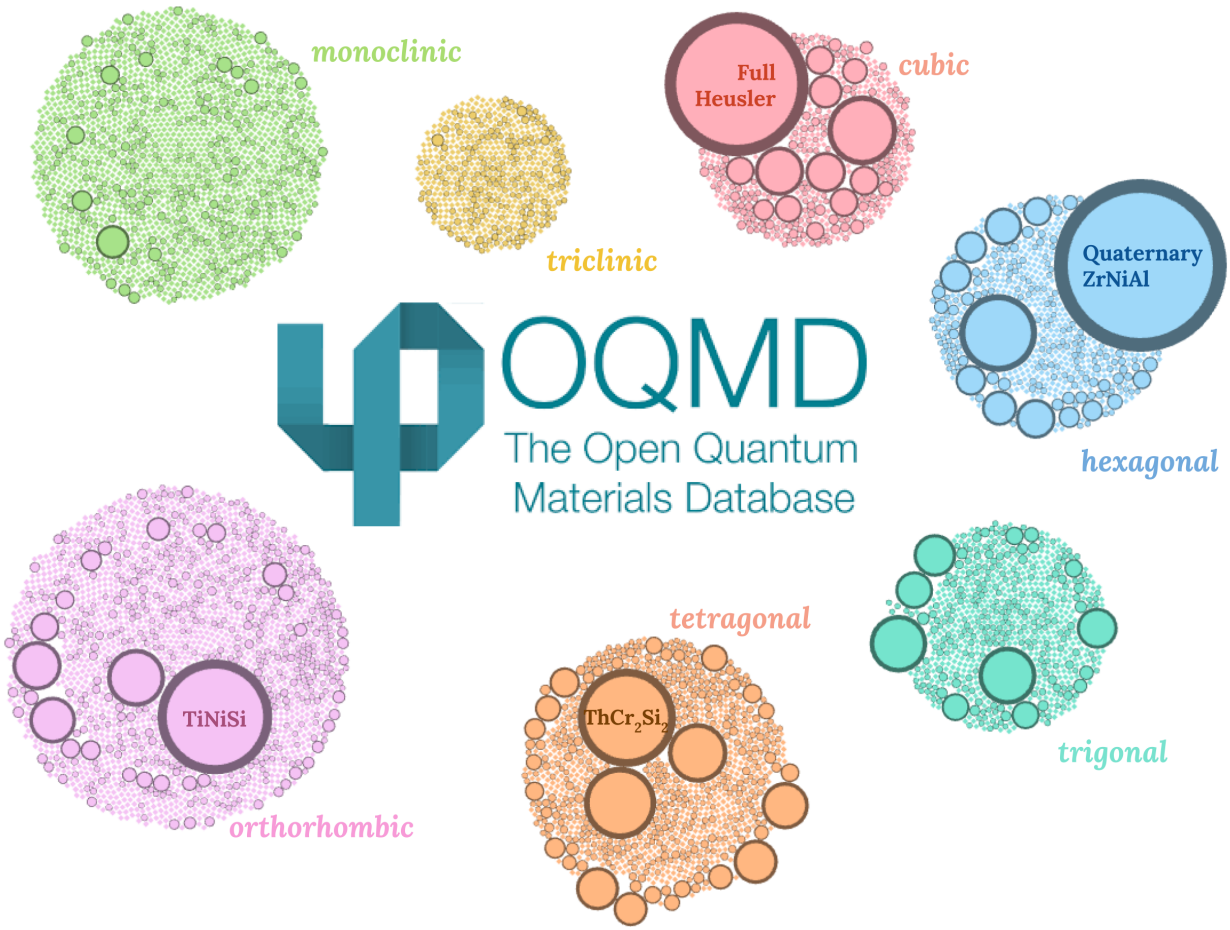

Figure 4.5: Distribution of all stable compounds in OQMD. Each circle indicates a distinct prototype, with the size of the circle being proportional to the number of stable compounds in the OQMD with such prototype. The prototypes are clustered according to the space group family, and the ones with the largest number of stable compounds are labeled.

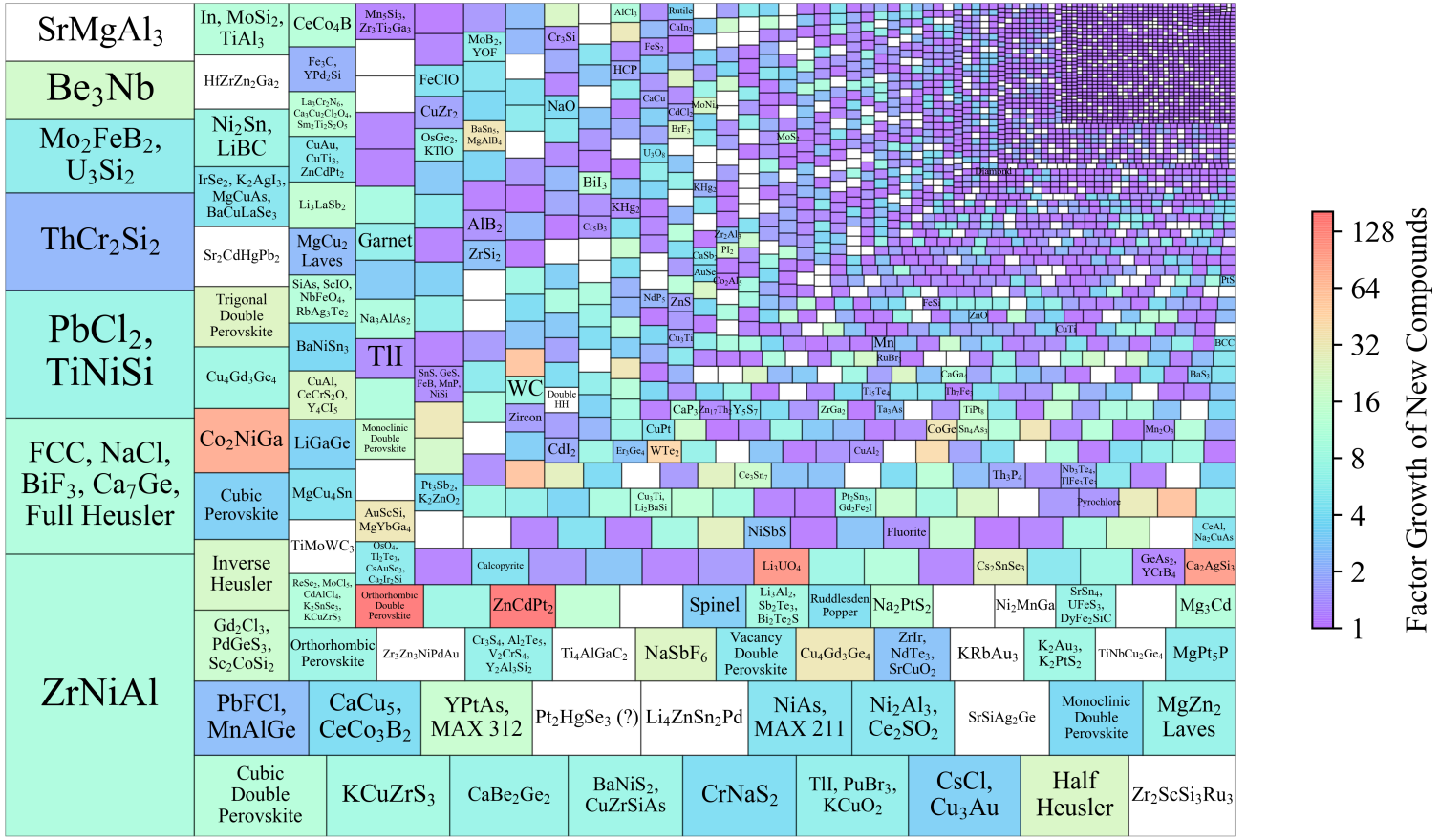

Figure 4.6: Box sizes are proportional to the number of stable compounds in OQMD belonging to a given prototype. The largest boxes and some well-known prototypes are labelled (the smallest labels can be viewed by zooming in on this .pdf document). The colors represent how many times the number of stable compounds has grown since the beginning, when OQMD contained only ICSD compounds.

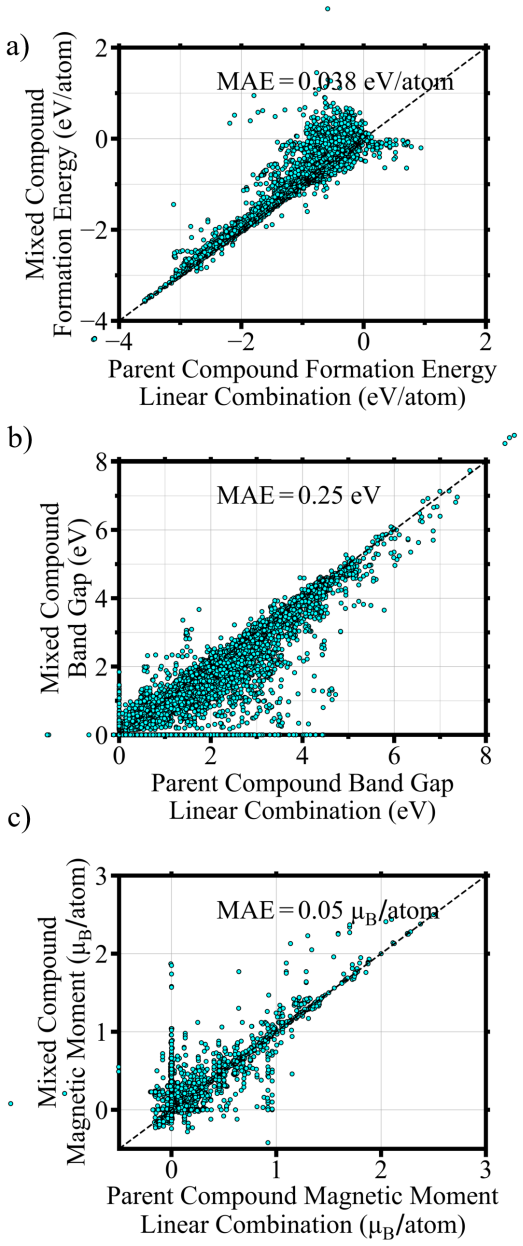

Figure 4.7: Comparison of mixed compound formation energies (a), band gaps (b), and magnetic moments (c) against the corresponding linear sum of parent compounds' properties. The black dashed line represents the case where the mixed and parent properties are equal. The mean absolute errors (MAE) are reported; for band gaps and magnetic moments, we include in the MAE calculation only cases where neither the child nor parent properties are zero.

CHAPTER 5
ARTIFICIAL INTELLIGENCE ACCELERATES THE PREDICTION OF STABLE MATERIALS

Most of this chapter is taken directly from our unpublished manuscript.

5.1 The Need for Artificial Intelligence to Predict Materials Stability

As the world recognizes the power of computation, so too is the field of materials synthesis undergoing a transition from laboratory to computer. Behind this transition is the increasing ability to predict, without experimental input, whether a material will form under specified environmental conditions. At the atomic level, compounds are predicted using their formation enthalpies, or energies of forming the compounds relative to their elemental reference states. If this energy is lower than that of any other possible compound or linear combination of compounds within the phase space, then this compound lies on the convex hull of stability and is therefore stable and predicted to exist at $T = 0$. Otherwise, the compound is predicted to be unstable and decompose into other phases; while a small convex hull distance ($\lesssim$ 100 meV/atom [58, 59]) can provide a clue about metastability (*i.e.*stability at nonzero temperature, pressure, etc.), it cannot guarantee metastability. It is important to consider metastability because many materials must operate at high temperature, pressure, etc., or must be synthesized at high temperature and rapidly cooled to room temperature for real-world applications. When nonzero temperature is applied, then we add an entropy term to predict the Gibbs free energy and then conduct the convex hull analysis using Gibbs free energies rather than formation enthalpies. An example illustration of a convex hull analysis at finite pressure is shown in Figure 5.1. A thorough explanation of the generalized Gibbs phase rule is provided in

Ref. 122.

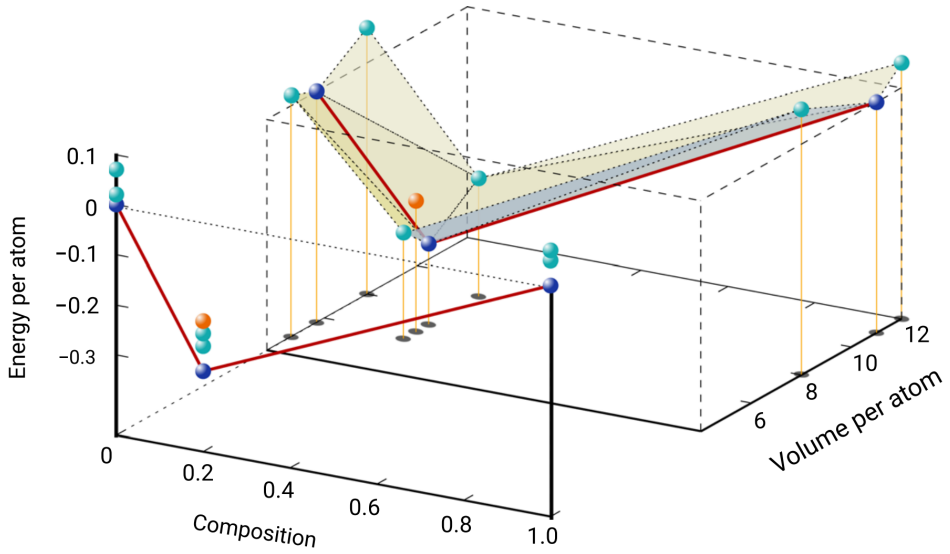

Figure 5.1: (Adapted from Ref. 57) Projection of a zero-pressure (composition-energy) convex hull (left) to various nonzero-pressure (composition-volume-energy) convex hulls (right). The convex hulls are marked by solid red (zero-pressure) and dotted black lines (nonzero-pressure). The convex hull, indicating stable phases, may contain different phases at zero pressure (dark blue spheres) versus nonzero pressures (light blue spheres). Some hypothetical phases are never stable at any pressure (orange spheres). One can perform a similar convex hull analysis to determine thermodynamic stability at nonzero temperature, surface/interfacial pressure, and other drivers.

For a given material, it is now possible to compute all quantities relating to thermodynamic stability with a high degree of accuracy using first-principles methods, especially DFT. [22, 23] For example, the total energy of a crystal can be computed with errors on the order of 1 meV/atom. Formation energies computed by DFT match experimental values with a MAE of just 91 meV/atom, which is within experimental error. [12] The corresponding convex hull stabilities can also be computed straightforwardly, since competing phase energies, experimentally known (*e.g.*from ICSD) and hypothetical, can be readily obtained from large-scale and rapidly growing DFT databases like Materials Project (MP), [8], OQMD [7, 31] AFLOW [9], JARVIS, [10] NOMAD, [82, 83] and OPTIMADE [11]. In first principles calculations of nonzero temperature ($T > 0$) materials,

one must add configurational and vibrational contributions to the free energy, as these can stabilize materials that are unstable at $T = 0$ and vice versa. Disorder and hence configurational entropy can be treated by converting the random structure to a "special quasirandom structure" (SQS) whose cluster correlation functions match those of the disordered structure; [123] or, to investigate a range of compositions, one can do a full cluster expansion study. DFT can also be used to handle vibrational entropy by investigating phonons (or atomic vibrations), which are approximated by harmonic terms to the spring force (and in some cases, anharmonic terms). [124] The full investigation requires dozens of DFT calculations of displaced-atom structures.

Although there are well known failures of DFT due to the use of approximated exchange correlation functionals, DFT remains the state-of-the-art method for highly reliable computations of solid state properties. [125, 126] The more significant challenge inhibiting computational materials discovery is the high computational cost of DFT calculations. The cost of DFT scales cubically with the number of atoms, and crystal structures that contain dozens of atoms typically cost hundreds of CPU hours each. While a large proportion of all experimentally known compounds have been calculated, it is much more expensive to predict new compounds. The configurational space of possible materials is infinitely large, and it is unfeasible to run DFT on all of them. The configurational space can be restricted to roughly 10000 known structure prototypes [33, 127] and around 80 technologically relevant elements, but there are still billions of possible compounds up to 5 elements that can be made from these prototypes. Supercells and $T > 0$ compounds are even more difficult, since they cost orders of magnitude more CPU hours than a single $T = 0$ structure. While high throughput calculations of $T > 0$ properties have been conducted for some material classes recently, [128, 129] there is no existing first principles study of the temperature effect for most known compounds today.

The question then arises whether a cheaper surrogate model can replace DFT in predictions of

stable materials. It is challenging to come up with a model that captures the extreme nonlinearity of the relationship between material and stability. Machine learning (ML) is well suited for this task because it is drastically cheaper than DFT and demonstrated to effectively learn complex relationships that are not necessarily well understood by experts (and can even improve understanding). The diversity of ML techniques, many of which have emerged as recently as the last few years, have inspired a flurry of studies into how to effectively represent materials and model their stability. To predict $T = 0$ stability, there have been numerous highly successful ML attempts that have led to the discovery of new stable materials, thanks in no small part to the coincident rapid growth and accessibility of DFT-computed formation energy data. ML predictions of $T > 0$ stability are an even newer area of exploration, as researchers are just beginning to tackle the enormous challenges of limited available DFT data (due to high expense) and complexity of the underlying theory.

In this Review, we cover the most recent ML advances in predicting materials stability at zero and nonzero temperatures. Our Review is organized as follows. First, we provide an overview of the state-of-the-art ML methods developed to predict materials stability, with much emphasis placed on the prediction of $T = 0$ formation energy. Next, we explore how such ML methods have been used to accelerate the discovery of $T = 0$ stable compounds (*i.e.*lying on the convex hull). Finally, we envision the future of materials stability predictions, which can incorporate other thermodynamic drivers such as pressure and surface/interfacial pressure.

5.2 Overview of Machine Learning Frameworks and the Prediction of Zero Temperature Formation Energy

In this section, we provide an overview of the state-of-the-art ML frameworks for materials predictions, including a special emphasis on formation energy as the target property since this is one of the most extensively calculated properties. As materials science is a highly nonlinear and mul-

tivariate problem, a wide spectrum of machine learning frameworks have been developed to target different aspects of the problem. One of the most crucial aspects of every ML framework is the featurization used to represent the material. As always, the features must represent some attribute of the material that correlates with the target property. For example, a simple way to obtain material attributes is to extract any properties having to do with its composition, *e.g.*its constituent elements and their relative amounts. A special advantage of this approach is that it does not require knowledge of the compound's crystal structure or physical properties, which require measurement or first principles simulations to obtain. This kind of feature extraction has been streamlined [111, 130, 131] and, with simple regression or classification models, has been shown to predict formation energy with errors on the order of 200 meV/atom as well as other properties. [111]

Beyond just the material's composition, its crystal structure (relating to the arrangement of atoms) is unquestionably relevant to its properties. In first principles calculations, the only necessary inputs are the unit cell dimensions and a list of atoms and their position coordinates within the unit cell. However, crystal structure representation for machine learning is much trickier since it is key to engineer features with physical significance to guide learning. For example, Ward *et al.*developed a set of features encoding averaged structural information, including coordination number, heterogeneity, chemical ordering, packing efficiency, and local environment, by extracting Voronoi tessellations of atoms in the unit cell. This method was found to exhibit lower errors for all training set sizes up to 3000 entries, compared to two earlier crystal structure representations: Coulomb matrix (CM) and partial radial distribution function (PRDF). [112, 132, 133] In another example, Seko *et al.*[113] developed a set of structural features emphasizing the relationships between elemental properties and structure prototype, on top of PRDF, "generalized" GRDF, bond-orientational order parameter, [134] and angular Fourier series, [135] achieving excellent formation energy prediction error of 41 meV/atom on a set of 18000 compounds, although the limited

configurational range of chemistries and prototypes in these compounds may lead to lower errors. [136] Kajita *et al*.devised a 3D voxel representation encoding the electron distribution inside a unit cell and demonstrated a MAE of $\sim$400 meV/atom on predictions of 680 randomly selected oxides from ICSD. [114] Jiang *et al*.achieved an impressive MAE of 61 meV/atom across a wide range of prototypes and compositions by extracting interpretable features based on topological representation derived from persistent homology. [115]

There is clearly an abundance of features to consider, and it is generally not clear which features relate to the target property and how. While domain knowledge can help, this requires an expert and even experts may lack certain relevant knowledge. In addition, there can be combinations of features, in the form of analytic expressions, that relate to the target property better than features alone. Although the feature combinations must be constrained to have physical meaning, *e.g*.having consistent units, there could conceivably be millions of features to choose from. Furthermore, it is desirable to keep only a small number of candidate features to relate to the target property in order to avoid overfitting. A well known ML technique to tackle the feature selection problem is LASSO, or "least absolute shrinkage and selection operator", where the objective function of minimization includes the l_1 norm of the number of nonzero features. [137] Compressed sensing is also helpful here, because it is necessary to reconstruct a signal from a set of observations that is far smaller than the feature space size. Ghiringhelli *et al*.adapted the LASSO approach with compressed sensing to identify physical descriptors relating to the energy differences of binary compound crystal structures. [138] Building on this method, Ouyang *et al*.developed the SISSO method, or "sure independence screening and sparsifying operator". Designed to efficiently handle enormous feature spaces ($\gg 10^4$ features), the SISSO approach first conducts a dimensionality reduction by screening candidate features most relevant to the target property (based on inner product between feature and property), [139] and then conducts LASSO on the smaller feature space.

[140]

Another strategy to deal with large feature spaces in materials prediction is to circumvent the feature engineering step and use deep learning ML techniques. A relatively recent ML advance, deep learning has a well-documented track record of learning complex and hierarchical relationships from massive amounts of data, *e.g.*identifying objects among millions of images, without the need for engineered, domain-specific features. [141] In one demonstration of deep learning in materials predictions, Zhou *et al.*used a single-hidden-layer fully connected neural network to predict formation energies of elpasolite (ABC_2D_6 double perovskite) compounds from their "Atom2Vec" atom feature vectors. [116] Jha *et al.*found that their ElemNet model, a deep neural network of up to 17 layers using elemental composition vectors as input, significantly outperformed a Random Forest (RF) model trained on either similar elemental composition vectors or physics-informed composition features. [117] As this model exhibited performance degradation beyond 17 layers, Jha *et al.*[118] developed a novel deep regression network architecture with individual residual learning, or IRNet, where shortcut connections are placed between every sequence (fully connected layer, batch normalization, and nonlinear activation) to allow gradient propagation across all layers and resolve the vanishing gradient problem. The authors showed that the IRNet exhibited steadily decreasing MAE in formation energy predictions through 20000 training iterations in networks of up to 48 layers, and significantly better performance compared to a stacked residual network (shortcut connections between blocks of 4 sequences) and a plain network with no shortcut connections. Attention networks have been explored to better represent stoichiometry in the context of corresponding properties; [109, 142] for example, doping a material with small amounts of another element can lead to large changes in its properties in a way that traditional representations of stoichiometry would not capture.

A significant challenge in incorporating ML methods into materials predictions is that most ML

methods require fixed length vectors as input, whereas materials can vary widely in compositional and structural complexity. As a result, applications of these ML methods are restricted to scenarios in which the materials under study are described with the same number of features. For example, materials belonging to the same structure prototype, like ABC_2D_6 elpasolites, can be described in terms of the structural attributes that are the same for any elpasolite. Or, when the materials belong to different prototypes or stoichiometries, the features must either be attributes that are common to all materials or "summarized" attributes, *e.g.*average coordination number. However, materials can exhibit complex and variable symmetries, local geometry, and hierarchy that are difficult to simultaneously account for with fixed length feature vectors. Xie and Grossman [101] developed an ML framework, which they called "crystal graph convolutional neural network" (CGCNN), that can handle arbitrary sized input vectors. In the CGCNN framework, crystal structures are modeled as crystal graphs where nodes are feature vectors of atom properties and edges encode bond information between atom pairs. Unlike molecular graphs, [143] crystal graphs allow multiple edges between the same node pair so that periodicity of the unit cell is fully accounted for. The convolutional layers are used to convolve the atom feature vectors with their edge and node connections (up to 12 nearest neighbors), as well as hidden layers to capture more complex structure-property relationships. The authors demonstrated the ability of CGCNN to predict formation energy and other properties for a wide variety of materials from MP, [8] with MAEs comparable to DFT MAEs with respect to experimental values. [12] Park and Wolverton [102] developed an "improved" version of CGCNN, which they called "iCGCNN", demonstrating significantly lower MAE in formation energy predictions compared to the original CGCNN. The iCGCNN framework encodes Voronoi neighbors rather than 12 nearest neighbors, and included Voronoi polyhedral information (such as solid angles, areas, and volumes) as part of the edge embeddings. In addition, 3-body correlations were included in the convolution function on top of 2-body correlations, and edge embeddings

were optimized in addition to nodes. Pal *et al.*[103] further improved the iCGCNN framework by introducing a scalar associated with the minimum interatomic distance in the crystal structure and optimized so as to predict the relaxed volume of the crystal and simultaneously its formation energy. This way, the predicted formation energy corresponds to the relaxed crystal structure, which is not initially known prior to a DFT calculation and can be significantly different from the unrelaxed (initial guess) crystal structure. This problem regarding relaxed vs. unrelaxed crystal structure was also addressed by Schmidt *et al.*by developing a crystal graph attention network in which bond distances are replaced by graph distances in the edge embeddings. [85] Although neural networks are a strong choice of ML method for crystal graph network architectures, they are not strictly required; Chen *et al.*[104] developed a generalized crystal graph network architecture called MatErials Graph Network (MEGNet). An additional advantage of MEGNet is the inclusion of global state attributes so that materials under variable thermodynamic conditions like temperature, pressure, and entropy can be included. Banjade *et al.*[105] added structural motifs (e.g. polyhedra) in addition to atomic bonds in their atom-motif dual graph network (AMDNet), demonstrating improved performance compared to MEGNet.

While crystal graph networks have been shown to greatly outperform non-deep-learning methods like RF when data sets like DFT formation energies exceed 10^4 data points, their performance is significantly weaker when data sets are smaller than 10^4 points. [110] Indeed, some material properties are less available than others due to labor or computational costs, making them more challenging to predict with ML methods. For example, vibrational entropy is highly expensive to compute from first principles, but crucial to predicting $T > 0$ stability. An increasingly popular strategy to build better ML models on small data sets is transfer learning, where model weights trained on larger data sets are transferred to a model of smaller data set. Jha *et al.*[106] demonstrated that by transferring weights from an ElemNet model trained on large DFT data sets to an

ElemNet model on smaller DFT and experimental data sets, they achieved MAEs much lower than when training a model on the corresponding smaller data sets from scratch. Other groups have similarly demonstrated transfer learning on materials formation energy predictions. [104, 109] Chen and Ong developed the AtomSets framework to extract MEGNet compositional and structural features for transfer learning, and successfully achieved strong performance on models of small data sets of bulk moduli, band gap, phonon density of states, and formation energy. [107] There are also recent methods to learn multiple properties simultaneously with a single ML model. The SISSO method has been adapted for multi-task learning and demonstrated on predicting the relative stability of binary compounds across a range of crystal prototypes with sparse or limited data. [144] De Breuck, Hautier, and Rignanese [145] developed an architecture called "material optimal descriptor network" (MODNet), where feature selection using Matminer features [130] is followed by joint learning to learn multiple properties simultaneously. The authors predicted vibrational entropy of crystals with MAE of just 0.009 meV/K/atom. CGCNN, SISSO, and MODNet frameworks are illustrated in Figure 5.2 as a sample.

5.3 Application of Machine Learning to Predict Zero Temperature Stable Compounds

Although formation energy is the most frequent target property estimated by ML models, the quantity most relevant to stability prediction is the material's convex hull distance ("CHD"). It is important to note that the convex hull here includes only already-DFT-calculated phases; this is necessary for ML predictions but is in fact artificial because if the material in question is confirmed DFT-stable, then according to thermodynamics the true convex hull must include this material and a true CHD (not the one we predict) must be zero. As CHD is merely a subtraction of formation energies of the material and its competing phases, ML can predict CHD with approximately the same MAE as formation energy. [102] However, ground state compounds and their polymorphs

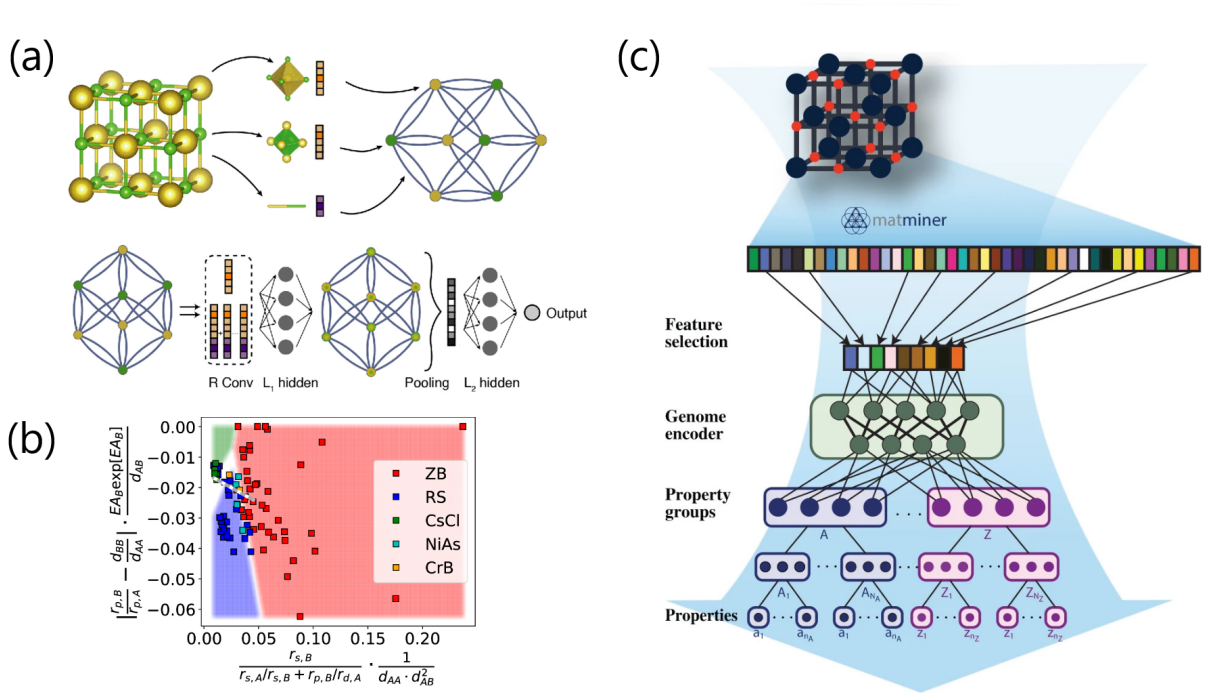

Figure 5.2: Sample of recent ML frameworks for materials discovery. (a) (Adapted from Ref. 101) CGCNN framework, where crystal structure is encoded as a graph and passed through a convolution neural network to predict a property such as formation energy. (b) (Adapted from Ref. 144) SISSO applied to construct a structure map as a multivariate function of several intuitive material properties. (c) (From Ref. 145) MODNet architecture for joint learning of multiple material properties, especially useful for small datasets like phonon-calculated vibrational entropy.

can differ by <10 meV/atom, which is still lower than the best ML MAE's to date. Although formation energies calculated by DFT have a MAE of 91 meV/atom compared to experimental values, [12] relative stabilities calculated by DFT benefit from subtraction of errors [146, 147] whereas ML stabilities do not. [119]

Despite this caveat of formation energy ML models, there have been numerous reports of stable compound discoveries. [39, 74, 85, 95, 101–103, 117, 148–151] Even when MAE's are high, the ML model can still be successful if only the compounds with very lowest predicted CHD are selected for DFT evaluation, especially when the predicted CHD is highly negative with magni-

tude greater than the MAE (although MAE can vary as a function of CHD [74, 132]), and if the training sets have a good balance of ground-state and higher-energy structures. [149] For example, Kim *et al*.[151] used the random forest model with composition and Voronoi structural features to accelerate the discovery of new quaternary Heusler (QH) compounds (stoichiometry $XX'YZ$ where X, Y, Z are metal elements). Their search space was a daunting 3.2 million compounds, but the training set was a rather healthy 96000 QH's from a previous high-throughput DFT study [91] in addition to 184000 ternary Heuslers (TH) and 146000 diverse non-Heusler (NH) compounds from OQMD. An important point they made is that their model achieved much lower MAE when the training set consisted of TH and NH compounds but very few QH's than when the training set consisted of only very few QH's, but when the training set had $> 10^4$ QH's then the MAE's were the same; this result suggests that already-done DFT calculations of different materials can be used to boost training sets for ML models targeting an unexplored class of materials. With a final MAE of 37 meV/atom, their model predicted 909 stable candidates, of which 55 were confirmed by DFT for a success rate of $55/909 = 6\%$ that was 30 times higher than the previous DFT study not using ML. [91] In another success story, Schmidt *et al*.[85] applied their crystal graph attention network to discover 325 stable mixed perovskites (stoichiometry ABX_2Y where A, B are cations and X, Y are anions) among 15 million candidates. Their success was aided by transfer learning, where they found that a mixed perovskite model pretrained on 2 million non-mixed-perovskite compounds had much lower MAE than a non-pretrained model. Noh *et al*.[152] introduced uncertainty quantification using hyperbolic tangent function to the CGCNN framework, thus enabling individual predicted-stable compounds to be scrutinized by the model uncertainty. Singh *et al*.[153] used SISSO to identify meaningful elemental features and construct an algebraic formula for the formation energy of rare earth compounds [RE]X_2 (where X is a transition metal). They then used their formula to learn about how the alloying effects between cerium-based compounds CeX_2-CeX_2'

can be explained by elemental properties. Bartel *et al*.similarly used SISSO to predict the stability of perovskites. [154]

Predicting formation energy is not the only way to discover new materials with ML. For example, the use of chemical rules or data mining of materials databases to map composition to prototype has been around for decades and continues to be a popular approach. [62, 81, 97–99, 155] In addition, one can take existing compounds and substitute in elements [56, 156] or ionic species [55] that tend to play similar roles in chemical environments. Wang, Botti, and Marques [84] used the elemental substitution method (illustrated in Figure 5.3) to discover 18479 stable compounds out of 189981 likely candidates, an impressive number given that today's DFT databases contain on the order of tens of thousands of experimentally known compounds.

When the search space is very large, another question is how to sample it more efficiently for effective model training. [157] Common search spaces in the materials context are the following:

1. Fixed composition, variable prototypes (formally defined [33]) and their corresponding structure variables (commonly called "crystal structure prediction").

2. Fixed prototype, variable compositions, each of which having different values for the structure variables.

3. Fixed combination of elements, variable compositions (and optionally their structures, i.e. "phase space").

4. Variable combinations of elements, compositions, and structures.

With a well defined search space and suitable ML method, one can then pursue active learning strategies. [158] There is much work being done in the incorporation of active learning strategies into autonomous, robotic experimental synthesis. [159, 160] In the area of CSP, where ML

models must be trained to predict unphysical structures at a fixed composition, active learning techniques have enabled CSP to replace expensive DFT calculations with graph network [161] and ML interatomic potential. [162] For broader search spaces, Montoya *et al*.developed a framework of autonomous intelligent agents for materials discovery that can be used for not just synthesis but also prediction of stable materials. [108] The framework, illustrated in Figure 5.4, allows the user to set up and deploy a flexible campaign with specified objective property, constraints, search space, choice of agent (input data, ML model, search strategy, etc.) and user feedback data (*e.g*.DFT, experiment). As a demonstration, they ran a simulated campaign of stable compound searches in the space of iron-containing binaries (Fe-X) and metal oxides (M-O) (following the form #3 above), using various input seeds (*e.g*.experimental DFT calculations, randomly sampled Fe-X or M-O), a neural network with/without adaptive boosting and composition/Voronoi structure features as the ML model, and various search strategies (*e.g*.lower confidence bound, ϵ-greedy, query-by-committee, Bayesian optimizers like Gaussian process). Their best performing agent was a neural network with adaptive boosting, uncertainty estimation based on the adaptive boosting, and a simple greedy search with lower confidence bound. They then executed searches for various other material classes and reported 383 new compounds with CHD < 200 meV/atom. In a later work, Ye *et al*.[163] used the same agent and reported 894 new compounds with CHD < 1 meV/atom (*i.e*.stable).

5.4 Outlook and Potential Opportunities

In this Review, we covered the many tremendous ML advancements over the last few years in the computational prediction of stable materials. Predicting stable $T = 0$ materials is becoming easier and faster than ever, thanks to numerous ML models targeting formation energy and the rapidly growing availability of DFT data. As a result, it is now possible to predict, with a high degree

of confidence, which compounds are stable at $T = 0$ using the convex hull construction. While predicting $T > 0$ stable materials is more challenging due to computational expense, emerging ML techniques are proving remarkably successful at predicting vibrational free energy and disorder with limited training data. As these methods mature, we envision the ability to predict, without any experimental input, Gibbs free energy of any material using a combination of first principles and ML methods. With accurate Gibbs free energy predictions, we can then construct a convex hull at any temperature to indicate which compounds are stable. With this ability to predict the stability of materials, we could proceed with "inverse" design, where a desired functionality along with stability are targets for materials discovery algorithms without any compositional or structural constraints. [164]

Aside from formation energy, vibrational free entropy and disorder, there are many other contributions to Gibbs free energy that deserve more exploration. [122] For one, applied hydrostatic pressure is a well-known knob for synthesis of metastable materials. When interested in just one specific value of pressure, one can straightforwardly add this external pressure during the DFT calculation, and hence construct a convex hull of formation enthalpies for all materials at this pressure. However, there may be presently unknown phases on the convex hull, and it is also much less computationally feasible to construct convex hulls over a wide range of pressures this way. When the pressures are on the order of GPa, typical of high-pressure synthesis conditions, the pressure effect can be approximated as linear with volume, and the dV term can be readily obtained from materials databases already. [57] However, the computational prediction of new high-pressure phases is still a largely unexplored concept. MLIP's have shown promise in this direction for limited chemical systems, like elemental boron, [165] black phosphorus, [166] and aluminum nitride. [167] In addition, the application of anisotropic pressure could lead to the formation of phases that are otherwise not possible with hydrostatic pressure. Another highly relevant knob is interfacial/surface energy,

[168] which could destabilize and prevent the nucleation of a phase that is otherwise predicted to coexist with other phases (*i.e.*sharing tie lines) or stabilize an otherwise unstable phase (especially in nanomaterials and thin films where the surface-area-to-volume ratio is high). Interfacial/surface energies are difficult to accurately calculate with DFT, due to computational expense, millions of predicted tie lines, [37] and the possibility that interfacial structure can differ from the bulk. Other knobs that can stabilize certain phases include electromagnetic fields and chemical potential. ML may one day allow us to predict the equations of state for any material, [169] but this ambitious endeavor is currently hindered by a lack of data.

Lastly, we point out subtle but important differences between materials stability versus synthesizability. Stability is the property that a compound has the lowest free energy out of all possible competing compounds under a non-empty set of environmental conditions. On the other hand, synthesizability refers to the ability to synthesize a material with a non-empty set of recipes. Much of the time, synthesizable materials that have been made before are known to be stable at $T = 0K$ in DFT databases, or can be shown to be stable at finite temperature, pressure, *etc.* Similarly, unstable materials, including "fantasy materials" with useful properties but are far above the convex hull, [170] are likely not synthesizable. However, there can be exceptions, such as materials that have been synthesized via non-equilibrium formation pathways [171] or materials with DFT-predicted stability that are not accessible in laboratory synthesis, perhaps due to high nucleation barriers, slow kinetics, or unknown recipe. Artificial intelligence may one day be of great help in predicting these stable-but-not-synthesizable materials, as well as in predicting effective recipes for synthesizable materials. [172] Some have tried to use machine learning to predict synthesizability as an empirical property rather than first-principles stability; [173–179] while such approaches are limited by biases in reported synthesis attempts, they may prove helpful for synthetic chemists when stability has not.

5.5 Conclusion

We have summarized the recent and growing body of work in developing ML models to predict the stability of materials with minimal experimental input. First, we covered the variety of ML frameworks for the prediction of material properties. The ML frameworks vary in their crystal representation, some based on intuitive composition and structural features and others based on whole structure representations, and their algorithm choice, from shallow learning to deep learning. Next, we focused on the applications of ML frameworks on predicting the stability of materials at $T = 0$. Many ML models target formation energy or convex hull distance, while others can predict the existence of materials based on data mining of literature for knowledge of other existing materials. We then covered how ML can aid the prediction of $T > 0$ stability parameters like disorder and vibrational entropy. Finally, we suggested that ML could predict whether other environmental conditions like pressure and surface/interfacial energy will stabilize new materials, and that ML models bringing together these conditions are needed.

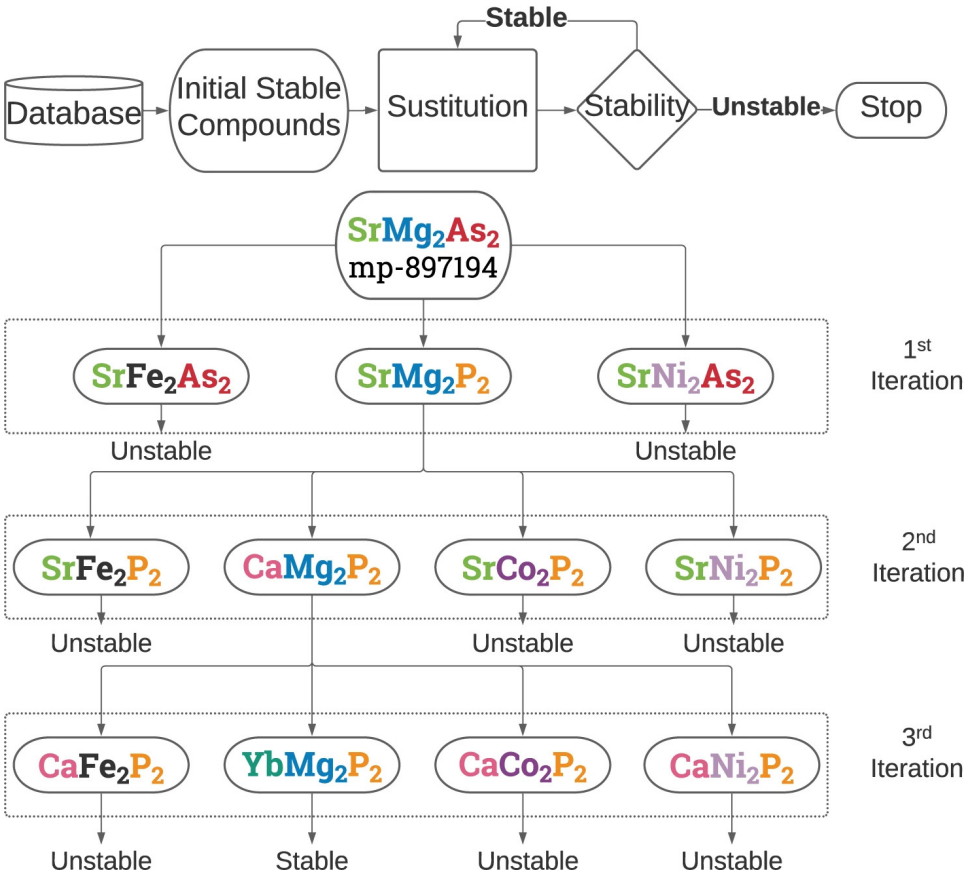

Figure 5.3: From Ref. 84) Workflow to discover new stable compounds using elemental substitution. Likely stable compounds are generated by substituting chemically similar elements into already-known stable compounds from materials databases. After DFT confirms which candidate compounds are stable, then these can be used to generate more likely stable compounds, and the process can be reiterated.

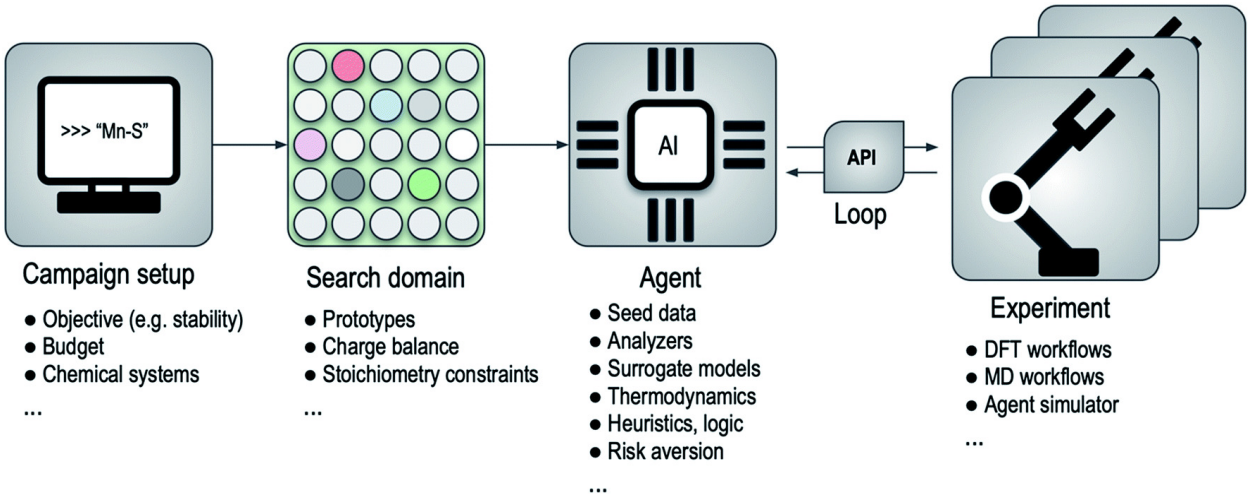

Figure 5.4: From Ref. 108) Framework for computational autonomy for materials discovery (CAMD). Starting with a user-defined search campaign with objective properties, budget constraints, search domain with chemical and structural criteria, and input data, an autonomous agent predicts materials with optimized properties using active learning strategies. The agent can request data from experiment or simulations as needed to validate ML-predicted properties and reduce model uncertainty.

CHAPTER 6
CONCLUSION AND FUTURE WORK

Some parts of this chapter are taken from our unpublished manuscript.

6.1 Summary of Work

In this document, we present a body of work concerning the high-throughput discovery of inorganic compounds using computational methods, especially involving DFT and DFT database (OQMD).

Following an introduction in Chapter 1, we first describe in Chapter 2 a novel method capable of rapidly (and in a high-throughput fashion) solving the crystal structures of compounds with existing experimental diffraction data. This method makes use of the fact that most inorganic compounds share common crystal structures (or "prototypes") with other compounds. Using partial data from diffraction analysis (stoichiometry, space group, and number of atoms per unit cell), we search a structure database for prototypes that share these characteristics. We then decorate the candidate prototypes with elements from the target composition, and compute DFT stability (leveraging convex hulls from a comprehensive DFT database) as well as the match to diffraction pattern. We then take the best-performing candidate and validate whether it is a plausible structure. As this method is cheap, fast, and effective, we employed it to solve 521 crystal structures from the PDF, and expanded the OQMD to these materials (a 1% expansion of experimentally known materials), enabling them to be further studied and considered for a variety of applications.

In Chapter 3, we explore DFT-based methods to accelerate the discovery of new, not-yet-synthesized compounds. Over the last ten years, search algorithms were previuosly developed based on data-mined substitution of chemically similar elements into known materials [55, 56] and

machine learning of formation energies. [102] We performed a systematic comparison of these methods using a data set of over 100,000 already-calculated Full Heusler compounds (of which $\sim$1,000 are stable). We also developed ways to improve each of the search methods. For the chemical similarity methods, we found that iteratively re-training the likelihood values of hypothetical compounds after successive iterations with newly acquired stable compounds led to greatly improved performance, as much as 1 order of magnitude in success rate in finding the stable compounds. For the machine learned formation energy method, we found that building a training set of 1% randomly sampled compounds in the search space in addition to experimental compounds greatly reduced the mean absolute error and mean error of formation energy predictions within the search space, and greatly improved performance by up to 1 order of magnitude. Comparing the search methods side-by-side, we found that chemical similarity method worked better for predicting the highest-likelihood compounds, while the machine learning method worked better for lower-likelihood compounds.

In Chapter 4, we use the above materials search methods to conduct high-throughput DFT discovery of over ten thousand new stable compounds, and summarize the new stable compounds in the OQMD. New compounds span nearly the whole range of inorganic compositions and structure types, many of which are being explored for future applications. The number of compounds in OQMD that come from experimental sources has reduced from 100% at the beginning to about a third today. We also describe another method of producing stable compounds by ordered mixing of two or more stable compounds with the same structure and one different element (for example, $0.5Ca_2Si + 0.5Sr_2Si = SrCaSi$). Since the parent compounds are stable, the mixed compound need only be slightly lower in energy than the linear combination of the parents in order to be stable. We find that, out of thousands of mixed compounds produced by our method, about 20% are indeed lower in energy. Most of the mixed compounds have formation energies, band gaps, and magnetic

moments that are close to or in between the corresponding linear combination of the parents.

In Chapter 5, we provide a review of recent developments in artificial intelligence to accelerate the prediction of stable materials. The computational materials community is making strides in predicting large numbers of new stable compounds at zero temperature by leveraging recently developed search algorithms based on data mining and machine learning. However, predicting materials at finite temperature remains highly expensive and there is great potential for artificial intelligence to accelerate this process. Furthermore, there are other thermodynamic factors that can stabilize compounds, such as pressure and surface/interfacial pressure.

6.2 Limitations and Opportunities

Here, we discuss the limitations of our methods and propose future opportunities to build upon our work.

6.2.1 Solving Structures and Discovering Materials with Unknown Prototypes

In our crystal structure solution work from Chapter 2, one of the main limitations of our work had to do with the computational cost of DFT calculations. As a result of this limitation, we limited our structure search only to prototypes from OQMD that matched all structural characteristics of the target compound. However, it is likely the case for some of compounds that the correct structure does not have the same prototype as any other structure in OQMD, or that the structure has a different space group or number of atoms per unit cell from what was reported in the PDF database. To consider other possible prototypes in other space groups would require many more DFT calculations and would be intractable. This limitation also comes up in Chapters 3 and 4, where we discuss materials discovery methods and employed them in high throughput to discover new compounds. Our search methods are limited to prototype candidates that already exist in

materials databases, and will fail to find stable structures that have unknown prototypes. Often, experimentalists discover new prototypes, which then become seeds for high-throughput DFT discovery projects; for example, a quaternary MAX phase was recently discovered and a subsequent high throughput study was conducted. [180–183] However, to predict a new prototype computationally would either involve some clever intuition or would involve a crystal structure prediction algorithm. We generated new prototypes by mixing stable compounds in Section 4.3, but in this case the lattice was preserved and only the stoichiometry changed. Crystal structure prediction using DFT is extremely computationally expensive. To make this problem more tractable, one might use machine learning to predict formation energies using any of the ML models that have been developed for this purpose (see Chapter 5 for an overview of ML methods). This concept has been explored recently for crystal structure prediction. [161] However, this will require the ML model to be highly accurate (errors in the tens of meV/atom) in order to properly sort candidate structures by energy. This is a challenge, not just because ML models are hard to design, but also because ML models have difficulty extrapolating to unknown structures when the training set doesn't sample them enough. What today's state-of-the-art ML models may be able to do fairly well is eliminate the highly unphysical candidates and present a list of possible low-energy candidates for DFT confirmation.

6.2.2 Discovering Disordered Compounds

The application of finite temperature can induce a phase transition from a fully ordered structure at low temperature to a disordered structure at high temperature. In addition, alloying, doping, and/or vacancies can be created in a material with a stoichiometry that, for thermodynamic reasons, would not otherwise form ordered phases. The ICSD contains on the order of 100,000 disordered compounds, or around half of the entire database. Unfortunately, structures for DFT

simulation are necessarily fully ordered. At zero temperature, this is not a problem; materials are fully ordered, and we don't need to include disordered candidate structures when discovering stable compounds. On the other hand, at finite temperature, we must always consider the possibility that the computationally-predicted structure is disordered.

Disordered structures can be modelled in first principles by performing a cluster expansion. In this procedure, we study a continuous range of compositions between elemental components, *e.g.*Ag and Au. The energy of a disordered fcc lattice containing some combination of Ag and Au can be expressed as an expansion of ordered clusters, whose energies are computed by DFT. Although there are an infinite number of clusters, only the smallest-degree clusters need to be computed for sufficient convergence of the total energy. Following this, the zero-temperature energy can be obtained for any composition and then the configurational entropy contribution to free energy can be simply added. Configurational entropy assumes perfectly random mixing, however in reality short-range order occurs in the mixture. In principle, cluster expansion can describe the short-range order of the system. If one is concerned with a single compound, such as one reported in the ICSD, the standard technique is to generate an SQS, which is an ordered structure whose cluster coefficients closely match those of the disordered structure. [123] Although the cluster expansion and SQS methods are used frequently, to date there is no standardized procedure to conduct them (as we do for DFT calculations, by using calculation settings in OQMD, Materials Project, etc. that are designed to ensure efficiency and convergence). If we can streamline such methods, then it may be feasible to conduct high throughput DFT studies of disordered materials and even discover new disordered materials and the temperature ranges that stabilize them.

In addition, many research teams have explored the use of ML to model disordered materials. A particular class of disordered compounds that has attracted much interest recently is high entropy alloys and ceramics, which are many-component (5 or more component) single-phase compounds

that are predominately stabilized by configurational entropy. [184, 185] This is a rich compositional space that may be enriched with useful properties, such as high strength and ductility. [186, 187] A common target for ML models is to predict whether a many-component compound will form a single-phase or a distribution of phases. [188, 189] Another strategy is to predict a so-called "entropy forming ability" (EFA) rather than directly computing free energy; this is a quantity that represents the energy distribution of structures with energy close to the ground state structure, and is meant to signify the ability for a composition to disorder. [190, 191] A more direct treatment of disorder that has been explored involves the ML prediction of force fields, using representations and ML architectures that capture the structure and physical interactions that govern disorder, such as many-body tensor representation (MBTR), [192] smooth overlap of atomic positions (SOAP), [135] Behler-Parrinello neural network potential (NNP), [193] Gaussian approximation potential (GAP), [194] spectral neighbor analysis potential (SNAP), [195] and moment tensor potential (MTP). [196, 197] As such ML models require problem-specific training data to reduce extrapolation, there have been successful attempts to develop an active learning framework with ML models. [198, 199]

6.2.3 Discovering Phonon-Stabilized Compounds

In addition to disorder, the application of temperature will cause atoms in crystals to vibrate about their equilibrium positions. These vibrations are called phonons and they contribute to the Gibbs free energy. Although there are materials, particularly composed of light elements like boron, [200] where the zero-point (zero temperature) vibrations can stabilize different structures, vibrational entropy generally only affects stability at finite temperature, as does disorder. Phonon contributions can in principle be calculated from first principles, but they typically involve DFT calculations of many displaced-atom supercell structures to compute interatomic force constants, with varying

degrees of complexity (such as lower complexity with harmonic approximation or higher complexity with quasi-harmonic approximation (QHA) or effective harmonic Hamiltonian (EHH)), or involve *ab-initio* molecular dynamics and thermodynamic integration. There are numerous phonon data with varying degrees of theoretical complexity in the literature, including high throughput databases. [129]

In any case, the calculations can range from highly to extraordinarily expensive, depending on the degree of accuracy one needs. To tackle this expense, research teams are replacing DFT with ML, especially machine learned interatomic potentials, in predictions of interatomic force constants [201] and in thermodynamic integration. [202–204] However, even with ML, obtaining accurate vibrational entropy using these methods is very expensive and not generally employed in high-throughput search frameworks. Much cheaper ML solutions have involved models relying on descriptors and deep learning to directly predict vibrational or Gibbs free energy. [145, 205, 206]

Phonon calculations are generally only performed for already-synthesized materials, especially ones that are expected to have useful properties. However, as phonon calculations leveraging DFT or ML become more computationally feasible, we can begin to predict hypothetical materials that can be stabilized by high temperature, and the ranges of temperature that stabilize them. The hope is that T vs. composition phase diagrams can be implemented in DFT databases, complete with both experimentally observed and hypothetical predicted materials.

6.2.4 Incorporating DFT Simulations into CALPHAD

CAlculation of PHAse Diagrams (CALPHAD) [207] is a term to describe methods to fit models of Gibbs free energy based entirely on theoretical or empirical formulas consistent with thermodynamics, and use these models to construct phase diagrams and predict stability and properties of highly complex alloys. The kinds of alloys that are of interest in the metallurgical field often have

numerous component elements, each with its own function, such as Cr to improve oxidation resistance. For example, "superalloys," a class of Ni- or Co-based alloys that exhibit ultrahigh strength and temperature resistance for applications in the aerospace industry, [208] may contain varying amounts of Co, Ni, Cr, W, Al, Ti, V, Ru, Re, and Ta. For materials with this many components, it is practically impossible to compute phase stability from first principles or even with ML. On the other hand, CALPHAD formalisms have been designed to handle multicomponent systems. The compound energy formalism (CEF) [209] and extended CEF [210] enable the modelling of multicomponent systems by utilizing DFT calculations of all binary end-member (*e.g.*Co+Cr) occupations on the sublattices of a phase of interest; for example, there are 2^5 end-members, or ways to arrange Co and Cr on the 5 sublattices of the σ phase. As such formalisms were found to be inadequate when only binary end-members (not ternary, etc.) are used, the effective bond energy formalism (EBEF) was developed to improve extrapolation to multicomponent systems. [211] Now that DFT is readily available and inexpensive, one can employ such CALPHAD formalisms with only a small number of DFT calculations in order to effectively model phase formations in real-world many-component alloys.

REFERENCES

[1] L.-D. Zhao *et al.*, "Ultralow thermal conductivity and high thermoelectric figure of merit in snse crystals," *Nature*, vol. 508, no. 7496, pp. 373–377, Apr. 2014.

[2] G. Tan, M. Ohta, and M. G. Kanatzidis, "Thermoelectric power generation: From new materials to devices," *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 377, no. 2152, p. 20 180 450, Aug. 2019.

[3] S. Kirklin, B. Meredig, and C. Wolverton, "High-throughput computational screening of new li-ion battery anode materials," *Advanced Energy Materials*, vol. 3, no. 2, pp. 252–262, 2013. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aenm.201200593.

[4] M. Aykol, S. Kim, V. I. Hegde, S. Kirklin, and C. Wolverton, "Computational evaluation of new lithium-3 garnets for lithium-ion battery applications as anodes, cathodes, and solid-state electrolytes," *Phys. Rev. Materials*, vol. 3, p. 025 402, 2 Feb. 2019.

[5] Z. Yao, V. I. Hegde, A. Aspuru-Guzik, and C. Wolverton, "Discovery of calcium-metal alloy anodes for reversible ca-ion batteries," *Advanced Energy Materials*, vol. 9, no. 9, p. 1 802 994, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aenm.201802994.

[6] J. K. Harada, N. Charles, K. R. Poeppelmeier, and J. M. Rondinelli, "Heteroanionic materials by design: Progress toward targeted properties," *Advanced Materials*, vol. 31, no. 19, p. 1 805 295, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201805295.

[7] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *JOM*, vol. 65, pp. 1501–1509, 11 Nov. 2013.

[8] A. Jain *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011 002, 2013. eprint: https://doi.org/10.1063/1.4812323.

[9] S. Curtarolo *et al.*, "Aflow: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012.

[10] K. Choudhary *et al.*, "The joint automated repository for various integrated simulations (jarvis) for data-driven materials design," *npj Computational Materials*, vol. 6, p. 173, 1 Nov. 2020.

[11] C. W. Andersen *et al.*, "Optimade, an api for exchanging materials data," *Scientific Data*, vol. 8, no. 1, p. 217, Aug. 2021.

[12] S. Kirklin *et al.*, "The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies," *npj Comput. Mater.*, vol. 1, p. 15 010, 1 Dec. 2015.

[13] A. R. Oganov and C. W. Glass, "Crystal structure prediction using ab initio evolutionary techniques: Principles and applications," *The Journal of Chemical Physics*, vol. 124, no. 24, p. 244 704, 2006. eprint: `https://doi.org/10.1063/1.2210932`.

[14] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "Crystal structure prediction via particle-swarm optimization," *Phys. Rev. B*, vol. 82, p. 094 116, 9 Sep. 2010.

[15] D. C. Lonie and E. Zurek, "Xtalopt: An open-source evolutionary algorithm for crystal structure prediction," *Computer Physics Communications*, vol. 182, no. 2, pp. 372–387, 2011.

[16] C. J. Pickard and R. J. Needs, "Ab initio random structure searching," *Journal of Physics: Condensed Matter*, vol. 23, no. 5, p. 053 201, Jan. 2011.

[17] B. Meredig and C. Wolverton, "A hybrid computational–experimental approach for automated crystal structure solution," *Nature Materials*, vol. 12, no. 2, pp. 123–127, Feb. 2013.

[18] L. Ward, K. Michel, and C. Wolverton, "Three new crystal structures in the Na–Pb system: solving structures without additional experimental input," *Acta Crystallographica Section A*, vol. 71, no. 5, pp. 542–548, Sep. 2015.

[19] L. Ward, K. Michel, and C. Wolverton, "Automated crystal structure solution from powder diffraction data: Validation of the first-principles-assisted structure solution method," *Phys. Rev. Mater.*, vol. 1, p. 063 802, 6 Nov. 2017.

[20] G. L. Hart *et al.*, "Revisiting the revised ag-pt phase diagram," *Acta Materialia*, vol. 124, pp. 325–332, 2017.

[21] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," *Advanced Materials*, vol. 31, no. 46, p. 1 902 765, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201902765.

[22] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, B864–B871, 3B Nov. 1964.

[23] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, A1133–A1138, 4A Nov. 1965.

[24] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, pp. 3865–3868, 18 Oct. 1996.

[25] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, "Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids," *Phys. Rev. Lett.*, vol. 91, p. 146 401, 14 Sep. 2003.

[26] J. P. Perdew *et al.*, "Restoring the density-gradient expansion for exchange in solids and surfaces," *Phys. Rev. Lett.*, vol. 100, p. 136 406, 13 Apr. 2008.

[27] J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," *The Journal of Chemical Physics*, vol. 118, no. 18, pp. 8207–8215, 2003. eprint: https://doi.org/10.1063/1.1564060.

[28] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Phys. Rev. B*, vol. 54, pp. 11 169–11 186, 16 Oct. 1996.

[29] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational Materials Science*, vol. 6, no. 1, pp. 15–50, 1996.

[30] P. E. Blöchl, "Projector augmented-wave method," *Phys. Rev. B*, vol. 50, pp. 17 953–17 979, 24 Dec. 1994.

[31] J. Shen *et al.*, "Reflections on one million compounds in the open quantum materials database (OQMD)," *Journal of Physics: Materials*, vol. 5, no. 3, p. 031 001, Jul. 2022.

[32] A. R. Akbarzadeh, V. Ozoliņš, and C. Wolverton, "First-principles determination of multicomponent hydride phase diagrams: Application to the li-mg-n-h system," *Advanced Materials*, vol. 19, no. 20, pp. 3233–3239, 2007. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.200700843.

[33] S. D. Griesemer, L. Ward, and C. Wolverton, "High-throughput crystal structure solution using prototypes," *Phys. Rev. Mater.*, vol. 5, p. 105 003, 10 Oct. 2021.

[34] E. Perim *et al.*, "Spectral descriptors for bulk metallic glasses based on the thermodynamics of competing crystalline phases," *Nature Communications*, vol. 7, no. 1, p. 12 315, Aug. 2016.

[35] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nature Materials*, vol. 12, no. 3, pp. 191–201, Mar. 2013.

[36] R. Woods-Robinson, D. Broberg, A. Faghaninia, A. Jain, S. S. Dwaraknath, and K. A. Persson, "Assessing high-throughput descriptors for prediction of transparent conductors," *Chemistry of Materials*, vol. 30, no. 22, pp. 8375–8389, 2018. eprint: https://doi.org/10.1021/acs.chemmater.8b03529.

[37] V. I. Hegde, M. Aykol, S. Kirklin, and C. Wolverton, "The phase stability network of all inorganic materials," *Science Advances*, vol. 6, no. 9, eaay5606, 2020. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aay5606.

[38] E. B. Isaacs and C. Wolverton, "Inverse band structure design via materials database screening: Application to square planar thermoelectrics," *Chemistry of Materials*, vol. 30, no. 5, pp. 1540–1546, 2018. eprint: https://doi.org/10.1021/acs.chemmater.7b04496.

[39] M. Amsler, L. Ward, V. I. Hegde, M. G. Goesten, X. Yi, and C. Wolverton, "Ternary mixed-anion semiconductors with tunable band gaps from machine-learning and crystal structure prediction," *Phys. Rev. Mater.*, vol. 3, p. 035 404, 3 Mar. 2019.

[40] A. Gindhart, T. Blanton, J. Blanton, and S. Gates-Rector, "The power of electron diffraction phase analysis and pattern simulations using the icdd® powder diffraction file™ (pdf-4)," *Microscopy and Microanalysis*, vol. 24, no. S1, pp. 1154–1155, 2018.

[41] V. K. Pecharsky and P. Y. Zavalij, *Fundamentals of powder diffraction and structural characterization of materials*, 2nd ed. New York : Springer, 2009.

[42] H. Putz, J. C. Schön, and M. Jansen, "Combined method for *ab initio* structure solution from powder diffraction data," *Journal of Applied Crystallography*, vol. 32, no. 5, pp. 864–870, Oct. 1999.

[43] Y. Zhong, C. Wolverton, Y. Austin Chang, and Z.-K. Liu, "A combined calphad/first-principles remodeling of the thermodynamics of al–sr: Unsuspected ground state energies by "rounding up the (un)usual suspects"," *Acta Materialia*, vol. 52, no. 9, pp. 2739–2754, 2004.

[44] O. M. Løvvik and O. Swang, "Structure and stability of possible new alanates," *Europhysics Letters*, vol. 67, no. 4, p. 607, Aug. 2004.

[45] C. Wolverton and V. Ozoli ņ š, "Hydrogen storage in calcium alanate: First-principles thermodynamics and crystal structures," *Phys. Rev. B*, vol. 75, p. 064 101, 6 Feb. 2007.

[46] C. Wolverton, D. J. Siegel, A. R. Akbarzadeh, and V. Ozoliņš, "Discovery of novel hydrogen storage materials: An atomic scale computational approach," *Journal of Physics: Condensed Matter*, vol. 20, no. 6, p. 064 228, Jan. 2008.

[47] L. Ward and K. Michel, *Materials/mint: Initial release*, version 11Nov16, Nov. 2016.

[48] W. A. Dollase, "Correction of intensities for preferred orientation in powder diffractometry: application of the March model," *Journal of Applied Crystallography*, vol. 19, no. 4, pp. 267–272, Aug. 1986.

[49] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[50] M. J. Mehl *et al.*, "The aflow library of crystallographic prototypes: Part 1," *Computational Materials Science*, vol. 136, S1–S828, 2017.

[51] D. Hicks *et al.*, "The aflow library of crystallographic prototypes: Part 2," *Computational Materials Science*, vol. 161, S1–S1011, 2019.

[52] D. Hicks *et al.*, "The aflow library of crystallographic prototypes: Part 3," *Computational Materials Science*, vol. 199, p. 110 450, 2021.

[53] H. C. Kandpal, C. Felser, and R. Seshadri, "Covalent bonding and the nature of band gaps in some half-heusler compounds," *Journal of Physics D: Applied Physics*, vol. 39, no. 5, p. 776, Feb. 2006.

[54] J. He, S. S. Naghavi, V. I. Hegde, M. Amsler, and C. Wolverton, "Designing and discovering a new family of semiconducting quaternary heusler compounds based on the 18-electron rule," *Chemistry of Materials*, vol. 30, no. 15, pp. 4978–4985, 2018. eprint: https://doi.org/10.1021/acs.chemmater.8b01096.

[55] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, "Data mined ionic substitutions for the discovery of new compounds," *Inorganic Chemistry*, vol. 50, no. 2, pp. 656–663, 2011, PMID: 21142147. eprint: https://doi.org/10.1021/ic102031h.

[56] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, "The optimal one dimensional periodic table: A modified pettifor chemical scale from data mining," *New Journal of Physics*, vol. 18, no. 9, p. 093 011, Sep. 2016.

[57] M. Amsler, V. I. Hegde, S. D. Jacobsen, and C. Wolverton, "Exploring the high-pressure materials genome," *Phys. Rev. X*, vol. 8, p. 041 021, 4 Nov. 2018.

[58] W. Sun *et al.*, "The thermodynamic scale of inorganic crystalline metastability," *Science Advances*, vol. 2, no. 11, e1600225, 2016. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.1600225.

[59] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, "Thermodynamic limit for synthesis of metastable inorganic materials," *Science Advances*, vol. 4, no. 4, eaaq0148, 2018. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aaq0148.

[60] W. Sun *et al.*, "Thermodynamic routes to novel metastable nitrogen-rich nitrides," *Chemistry of Materials*, vol. 29, no. 16, pp. 6936–6946, 2017. eprint: https://doi.org/10.1021/acs.chemmater.7b02399.

[61] J. Odahara *et al.*, "Self-combustion synthesis of novel metastable ternary molybdenum nitrides," *ACS Materials Letters*, vol. 1, no. 1, pp. 64–70, 2019. eprint: https://doi.org/10.1021/acsmaterialslett.9b00057.

[62] W. Sun *et al.*, "A map of the inorganic ternary metal nitrides," *Nature Materials*, vol. 18, no. 7, pp. 732–739, Jul. 2019.

[63] C. G. Richter and W. Jeitschko, "Preparation and crystal structure of the titanium and hafnium bismuthides Ti_8Bi_9 and Hf_8Bi_9," *Journal of Solid State Chemistry France*, vol. 134, no. 1, pp. 26–30, Nov. 1997.

[64] H. F. McMurdie *et al.*, *Standard X-ray Diffraction Powder Patterns: Monograph 25, Section 12. Data for 57 Substances*. Washington, D.C.: Institute for Materials Research, National Bureau of Standards, Feb. 1975.

[65] C. W. Pistorius, "A_2BF_5 phases in the systems AF-BF_3 (a=k, rb, cs, tl; b=al, fe, cr, ga, v, tl, ln)," *Materials Research Bulletin*, vol. 10, no. 10, pp. 1079–1084, 1975.

[66] Y. Sakurai, H. Arai, S. Okada, and J.-i. Yamaki, "Low temperature synthesis and electrochemical characteristics of lifeo2 cathodes," *Journal of Power Sources*, vol. 68, no. 2, pp. 711–715, 1997, Proceedings of the Eighth International Meeting on Lithium Batteries.

[67] A. D. Weeks, E. A. Cisney, and A. M. Sherwood, "Montroseite, a new vanadium oxide from the colorado plateaus," *Trace Elements Investigations*, no. 335, pp. 1–15, 1953.

[68] J. Evans Howard T. and S. Block, "The crystal structure of montroseite, a vanadium member of the diaspore group," *American Mineralogist*, vol. 38, no. 11-12, pp. 1242–1250, Dec. 1953. eprint: `https://pubs.geoscienceworld.org/msa/ammin/article-pdf/38/11-12/1242/4245700/am-1953-1242.pdf`.

[69] C. Milton *et al.*, "Merumite - a complex assemblage of chromium minerals from guyana," *Geological Survey Professional Paper*, no. 887, 1976.

[70] I. E. Nemirovskaya, A. N. Grechenko, A. M. Alekseev, and V. V. Lunin, "Phase transformations in hydrogen sorption-desorption by hydrides of intermetallic compounds of the crb structural type," *Journal of Structural Chemistry*, vol. 32, no. 5, pp. 680–686, Sep. 1991.

[71] R. Van Essen and K. Buschow, "Hydrogen absorption in various zirconium- and hafnium-based intermetallic compounds," *Journal of the Less Common Metals*, vol. 64, no. 2, pp. 277–284, 1979.

[72] K. Ćirić, V. Koteski, D. Stojić, J. Radakovic, and V. Ivanovski, "Hfni and its hydrides – first principles calculations," *International Journal of Hydrogen Energy*, vol. 35, no. 8, pp. 3572–3577, 2010.

[73] Peterson, S.W., Sadana, V.N., and Korst, W.L., "Neutron diffraction study of nickel zirconium hydride," *J. Phys. France*, vol. 25, no. 5, pp. 451–453, 1964.

[74] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals," *Phys. Rev. Lett.*, vol. 117, p. 135 502, 13 Sep. 2016.

[75] G. Trimarchi, X. Zhang, M. J. DeVries Vermeer, J. Cantwell, K. R. Poeppelmeier, and A. Zunger, "Emergence of a few distinct structures from a single formal structure type during high-throughput screening for stable compounds: The case of rbcus and rbcuse," *Phys. Rev. B*, vol. 92, p. 165 103, 16 Oct. 2015.

[76] E. Parthé, *Elements of Inorganic Structural Chemistry: A Course on Selected Topics*. K. Sutter Parthé, 1990, ISBN: 9782950492401.

[77] A. R. Oganov and M. Valle, "How to quantify energy landscapes of solids," *The Journal of Chemical Physics*, vol. 130, no. 10, p. 104 504, 2009. eprint: `https://doi.org/10.1063/1.3079326`.

[78] H. Burzlaff and Y. Malinovsky, "A Procedure for the Clasification of Non-Organic Crystal Structures. I. Theoretical Background," *Acta Crystallographica Section A*, vol. 53, no. 2, pp. 217–224, Mar. 1997.

[79] J. A. Chisholm and S. Motherwell, "*COMPACK*: a program for identifying crystal structure similarity using distances," *Journal of Applied Crystallography*, vol. 38, no. 1, pp. 228–231, Feb. 2005.

[80] L. Zhu *et al.*, "A fingerprint based metric for measuring similarities of crystalline structures," *The Journal of Chemical Physics*, vol. 144, no. 3, p. 034 203, 2016. eprint: `https://doi.org/10.1063/1.4940026`.

[81] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," *Nature Materials*, vol. 5, no. 8, pp. 641–646, Aug. 2006.

[82] C. Draxl and M. Scheffler, "Nomad: The fair concept for big data-driven materials science," *MRS Bulletin*, vol. 43, no. 9, pp. 676–682, 2018.

[83] L. M. Ghiringhelli *et al.*, "Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats," *npj Computational Materials*, vol. 3, no. 1, p. 46, Nov. 2017.

[84] H.-C. Wang, S. Botti, and M. A. L. Marques, "Predicting stable crystalline compounds using chemical similarity," *npj Computational Materials*, vol. 7, no. 1, p. 12, Jan. 2021.

[85] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, and M. A. L. Marques, "Crystal graph attention networks for the prediction of stable materials," *Science Advances*, vol. 7, no. 49, eabi7948, 2021. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abi7948.

[86] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, "High-throughput computational screening of perovskites for thermochemical water splitting applications," *Chemistry of Materials*, vol. 28, no. 16, pp. 5621–5634, Aug. 2016.

[87] A. A. Emery and C. Wolverton, "High-throughput dft calculations of formation energy, stability and oxygen vacancy formation energy of abo3 perovskites," *Scientific Data*, vol. 4, no. 1, p. 170 153, Oct. 2017.

[88] S. Anand, M. Wood, Y. Xia, C. Wolverton, and G. J. Snyder, "Double half-heuslers," *Joule*, vol. 3, no. 5, pp. 1226–1238, 2019.

[89] J. He *et al.*, "Computational discovery of stable heteroanionic oxychalcogenides abxo (a, b = metals; x = s, se, and te) and their potential applications," *Chemistry of Materials*, vol. 32, no. 19, pp. 8229–8242, Oct. 2020.

[90] J. Shen, V. I. Hegde, J. He, Y. Xia, and C. Wolverton, "High-throughput computational discovery of ternary mixed-anion oxypnictides," *Chemistry of Materials*, vol. 33, no. 24, pp. 9486–9500, Dec. 2021.

[91] J. He, S. S. Naghavi, V. I. Hegde, M. Amsler, and C. Wolverton, "Designing and discovering a new family of semiconducting quaternary heusler compounds based on the 18-electron rule," *Chemistry of Materials*, vol. 30, no. 15, pp. 4978–4985, Aug. 2018.

[92] K. Pal *et al.*, "Accelerated discovery of a large family of quaternary chalcogenides with very low lattice thermal conductivity," *npj Computational Materials*, vol. 7, no. 1, p. 82, Jun. 2021.

[93] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding nature's missing ternary oxide compounds using machine learning and density functional theory," *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, Jun. 2010.

[94] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," *Chemistry of Materials*, vol. 29, no. 12, pp. 5090–5103, Jun. 2017.

[95] J. Schmidt, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the stability of ternary intermetallics with density functional theory and machine learning," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241 728, 2018. eprint: https://doi.org/10.1063/1.5020223.

[96] H.-C. Wang, J. Schmidt, S. Botti, and M. A. L. Marques, "A high-throughput study of oxynitride, oxyfluoride and nitrofluoride perovskites," *J. Mater. Chem. A*, vol. 9, pp. 8501–8513, 13 2021.

[97] A. Zunger, "Structural stability of 495 binary compounds," *Phys. Rev. Lett.*, vol. 44, pp. 582–586, 9 Mar. 1980.

[98] P. Villars, "A three-dimensional structural stability diagram for 998 binary ab intermetallic compounds," *Journal of the Less Common Metals*, vol. 92, no. 2, pp. 215–238, 1983.

[99] D. G. Pettifor, "The structures of binary compounds. i. phenomenological structure maps," *Journal of Physics C: Solid State Physics*, vol. 19, no. 3, p. 285, Jan. 1986.

[100] B. Meredig *et al.*, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Phys. Rev. B*, vol. 89, p. 094 104, 9 Mar. 2014.

[101] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Phys. Rev. Lett.*, vol. 120, p. 145 301, 14 Apr. 2018.

[102] C. W. Park and C. Wolverton, "Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery," *Phys. Rev. Materials*, vol. 4, p. 063 801, 6 Jun. 2020.

[103] K. Pal, C. W. Park, Y. Xia, J. Shen, and C. Wolverton, "Scale-invariant machine-learning model accelerates the discovery of quaternary chalcogenides with ultralow lattice thermal conductivity," *npj Computational Materials*, vol. 8, no. 1, p. 48, Mar. 2022.

[104] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, May 2019.

[105] H. R. Banjade *et al.*, "Structure motif–centric learning framework for inorganic crystalline systems," *Science Advances*, vol. 7, no. 17, eabf1754, 2021. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abf1754.

[106] D. Jha *et al.*, "Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning," *Nature Communications*, vol. 10, no. 1, p. 5316, Nov. 2019.

[107] C. Chen and S. P. Ong, "Atomsets as a hierarchical transfer learning framework for small and large materials datasets," *npj Computational Materials*, vol. 7, no. 1, p. 173, Oct. 2021.

[108] J. H. Montoya, K. T. Winther, R. A. Flores, T. Bligaard, J. S. Hummelshøj, and M. Aykol, "Autonomous intelligent agents for accelerated materials discovery," *Chem. Sci.*, vol. 11, pp. 8517–8532, 32 2020.

[109] R. E. A. Goodall and A. A. Lee, "Predicting materials properties without crystal structure: Deep representation learning from stoichiometry," *Nature Communications*, vol. 11, no. 1, p. 6280, Dec. 2020.

[110] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm," *npj Computational Materials*, vol. 6, no. 1, p. 138, Sep. 2020.

[111] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, p. 16 028, 1 Aug. 2016.

[112] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1094–1101, 2015. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24917.

[113] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, "Representation of compounds for machine-learning prediction of physical properties," *Phys. Rev. B*, vol. 95, p. 144 110, 14 Apr. 2017.

[114] S. Kajita, N. Ohba, R. Jinnouchi, and R. Asahi, "A universal 3d voxel descriptor for solid-state material informatics with deep convolutional neural networks," *Scientific Reports*, vol. 7, no. 1, p. 16 991, Dec. 2017.

[115] Y. Jiang, D. Chen, X. Chen, T. Li, G.-W. Wei, and F. Pan, "Topological representations of crystalline compounds for the machine-learning prediction of materials properties," *npj Computational Materials*, vol. 7, no. 1, p. 28, Feb. 2021.

[116] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, "Learning atoms for materials discovery," *Proceedings of the National Academy of Sciences*, vol. 115, no. 28, E6411–E6417, 2018. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1801181115.

[117] D. Jha *et al.*, "Elemnet: Deep learning the chemistry of materials from only elemental composition," *Scientific Reports*, vol. 8, no. 1, p. 17 593, Dec. 2018.

[118] D. Jha *et al.*, "Irnet: A general purpose deep residual regression framework for materials discovery," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19, Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2385–2393, ISBN: 9781450362016.

[119] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, "A critical examination of compound stability predictions from machine-learned formation energies," *npj Computational Materials*, vol. 6, no. 1, p. 97, Jul. 2020.

[120] S. P. Ong *et al.*, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, 2013.

[121] J. Schmidt *et al.*, *Large-scale machine-learning-assisted exploration of the whole materials space*, 2022.

[122] W. Sun and M. J. Powell-Palm, *Generalized gibbs' phase rule*, 2021.

[123] A. Zunger, S.-H. Wei, L. G. Ferreira, and J. E. Bernard, "Special quasirandom structures," *Phys. Rev. Lett.*, vol. 65, pp. 353–356, 3 Jul. 1990.

[124] A. Togo and I. Tanaka, "First principles phonon calculations in materials science," *Scr. Mater.*, vol. 108, pp. 1–5, Nov. 2015.

[125] P. Verma and D. G. Truhlar, "Status and challenges of density functional theory," *Trends in Chemistry*, vol. 2, no. 4, pp. 302–318, 2020, Special Issue - Laying Groundwork for the Future.

[126] N. Marzari, A. Ferretti, and C. Wolverton, "Electronic-structure methods for materials design," *Nature Materials*, vol. 20, no. 6, pp. 736–749, Jun. 2021.

[127] D. Hicks *et al.*, "Aflow-xtalfinder: A reliable choice to identify crystalline prototypes," *npj Computational Materials*, vol. 7, no. 1, p. 30, Feb. 2021.

[128] Y. Xia *et al.*, "High-throughput study of lattice thermal conductivity in binary rocksalt and zinc blende compounds including higher-order anharmonicity," *Phys. Rev. X*, vol. 10, p. 041 029, 4 Nov. 2020.

[129] G. Petretto *et al.*, "High-throughput density-functional perturbation theory phonons for inorganic materials," *Scientific Data*, vol. 5, no. 1, p. 180 065, May 2018.

[130] L. Ward *et al.*, "Matminer: An open source toolkit for materials data mining," *Computational Materials Science*, vol. 152, pp. 60–69, 2018.

[131] E. Gossett *et al.*, "Aflow-ml: A restful api for machine-learning predictions of materials properties," *Computational Materials Science*, vol. 152, pp. 134–145, 2018.

[132] L. Ward *et al.*, "Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations," *Phys. Rev. B*, vol. 96, p. 024 104, 2 Jul. 2017.

[133] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," *Phys. Rev. B*, vol. 89, p. 205 118, 20 May 2014.

[134] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," *Phys. Rev. B*, vol. 28, pp. 784–805, 2 Jul. 1983.

[135] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B*, vol. 87, p. 184 115, 18 May 2013.

[136] Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," *npj Computational Materials*, vol. 4, no. 1, p. 25, May 2018.

[137] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x`.

[138] L. M. Ghiringhelli *et al.*, "Learning physical descriptors for materials science by compressed sensing," *New Journal of Physics*, vol. 19, no. 2, p. 023 017, Feb. 2017.

[139] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00674.x.

[140] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L. M. Ghiringhelli, "Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates," *Phys. Rev. Materials*, vol. 2, p. 083 802, 8 Aug. 2018.

[141] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[142] A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, "Compositionally restricted attention-based network for materials property predictions," *npj Computational Materials*, vol. 7, no. 1, p. 77, May 2021.

[143] D. K. Duvenaud *et al.*, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[144] R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, "Simultaneous learning of several materials properties from incomplete databases with multi-task SISSO," *Journal of Physics: Materials*, vol. 2, no. 2, p. 024 002, Mar. 2019.

[145] P.-P. De Breuck, G. Hautier, and G.-M. Rignanese, "Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet," *npj Computational Materials*, vol. 7, no. 1, p. 83, Jun. 2021.

[146] G. Hautier, S. P. Ong, A. Jain, C. J. Moore, and G. Ceder, "Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability," *Phys. Rev. B*, vol. 85, p. 155 208, 15 Apr. 2012.

[147] C. J. Bartel, A. W. Weimer, S. Lany, C. B. Musgrave, and A. M. Holder, "The role of decomposition reactions in assessing first-principles predictions of solid stability," *npj Computational Materials*, vol. 5, no. 1, p. 4, Jan. 2019.

[148] Y. Zhao, E. M. D. Siriwardane, Z. Wu, M. Hu, N. Fu, and J. Hu, *Physics guided generative adversarial networks for generations of crystal materials with symmetry constraints*, 2022.

[149] S. Pandey, J. Qu, V. Stevanović, P. St. John, and P. Gorai, "Predicting energy and stability of known and hypothetical crystals using graph neural network," *Patterns*, vol. 2, no. 11, p. 100 361, 2021.

[150] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, and A. A. Lee, "Rapid discovery of stable materials by coordinate-free coarse graining," *Science Advances*, vol. 8, no. 30, eabn4117, 2022. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.abn4117`.

[151] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds," *Phys. Rev. Materials*, vol. 2, p. 123 801, 12 Dec. 2018.

[152] J. Noh, G. H. Gu, S. Kim, and Y. Jung, "Uncertainty-quantified hybrid machine learning/density functional theory high throughput screening method for crystals," *Journal of Chemical Information and Modeling*, vol. 60, no. 4, pp. 1996–2003, Apr. 2020.

[153] P. Singh, T. Del Rose, G. Vazquez, R. Arroyave, and Y. Mudryk, "Machine-learning enabled thermodynamic model for the design of new rare-earth compounds," *Acta Materialia*, vol. 229, p. 117 759, 2022.

[154] C. J. Bartel *et al.*, "New tolerance factor to predict the stability of perovskite oxides and halides," *Science Advances*, vol. 5, no. 2, eaav0693, 2019. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.aav0693`.

[155] Y. Zhang *et al.*, "The role of hume-rothery's rules play in the max phases formability," *Materialia*, vol. 12, p. 100 810, 2020.

[156] A. Vasylenko *et al.*, "Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry," *Nature Communications*, vol. 12, no. 1, p. 5561, Sep. 2021.

[157] Y. Kim, E. Kim, E. Antono, B. Meredig, and J. Ling, "Machine-learned metrics for predicting the likelihood of success in materials discovery," *npj Computational Materials*, vol. 6, no. 1, p. 131, Aug. 2020.

[158] D. P. Tabor *et al.*, "Accelerating the discovery of materials for clean energy in the era of smart automation," *Nature Reviews Materials*, vol. 3, no. 5, pp. 5–20, May 2018.

[159] N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, and G. Ceder, "Toward autonomous design and synthesis of novel inorganic materials," *Mater. Horiz.*, vol. 8, pp. 2169–2198, 8 2021.

[160] A. G. Kusne *et al.*, "On-the-fly closed-loop materials discovery via bayesian active learning," *Nature Communications*, vol. 11, no. 1, p. 5966, Nov. 2020.

[161] G. Cheng, X.-G. Gong, and W.-J. Yin, "Crystal structure prediction by combining graph network and optimization algorithm," *Nature Communications*, vol. 13, no. 1, p. 1492, Mar. 2022.

[162] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning," *Phys. Rev. B*, vol. 99, p. 064 114, 6 Feb. 2019.

[163] W. Ye, X. Lei, M. Aykol, and J. H. Montoya, "Novel inorganic crystal structures predicted using autonomous simulation agents," *Scientific Data*, vol. 9, no. 1, p. 302, Jun. 2022.

[164] A. Zunger, "Inverse design in search of materials with target functionalities," *Nature Reviews Chemistry*, vol. 2, no. 4, p. 0121, Mar. 2018.

[165] V. L. Deringer, C. J. Pickard, and G. Csányi, "Data-driven learning of total and local energies in elemental boron," *Phys. Rev. Lett.*, vol. 120, p. 156 001, 15 Apr. 2018.

[166] V. L. Deringer, D. M. Proserpio, G. Csányi, and C. J. Pickard, "Data-driven learning and prediction of inorganic crystal structures," *Faraday Discuss.*, vol. 211, pp. 45–59, 0 2018.

[167] K. Miwa, "Multibaric sampling for machine learning potential construction," *Phys. Rev. B*, vol. 103, p. 144 106, 14 Apr. 2021.

[168] C. B. Wahl, M. Aykol, J. H. Swisher, J. H. Montoya, S. K. Suram, and C. A. Mirkin, "Machine learning-accelerated design and synthesis of polyelemental heterostructures," *Science Advances*, vol. 7, no. 52, eabj5505, 2021. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.abj5505`.

[169] S. Srinivasan *et al.*, *Machine learning the metastable phase diagram of materials*, 2020.

[170] A. Zunger, "Beware of plausible predictions of fantasy materials," *Nature*, vol. 566, no. 7745, pp. 447–449, 2019.

[171] M. Bianchini *et al.*, "The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides," *Nature Materials*, vol. 19, no. 10, pp. 1088–1095, Oct. 2020.

[172] T. He, H. Huo, C. J. Bartel, Z. Wang, K. Cruse, and G. Ceder, *Inorganic synthesis recommendation by machine learning materials similarity from scientific literature*, 2023. arXiv: 2302.02303 [cond-mat.mtrl-sci].

[173] J. Jang, G. H. Gu, J. Noh, J. Kim, and Y. Jung, "Structure-based synthesizability prediction of crystals using partially supervised learning," *Journal of the American Chemical Society*, vol. 142, no. 44, pp. 18 836–18 843, Nov. 2020.

[174] A. Davariashtiyani, Z. Kadkhodaie, and S. Kadkhodaei, "Predicting synthesizability of crystalline materials via deep learning," *Communications Materials*, vol. 2, no. 1, p. 115, Nov. 2021.

[175] A. O. Oliynyk *et al.*, "High-throughput machine-learning-driven synthesis of full-heusler compounds," *Chemistry of Materials*, vol. 28, no. 20, pp. 7324–7331, Oct. 2016.

[176] G. Pilania, P. V. Balachandran, C. Kim, and T. Lookman, "Finding new perovskite halides via machine learning," *Frontiers in Materials*, vol. 3, 2016.

[177] P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, "Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory," *Phys. Rev. Materials*, vol. 2, p. 043 802, 4 Apr. 2018.

[178] A. Lee *et al.*, "Machine learned synthesizability predictions aided by density functional theory," *Communications Materials*, vol. 3, no. 1, p. 73, Oct. 2022.

[179] P. Raccuglia *et al.*, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, pp. 73–76, May 2016.

[180] L. Chen *et al.*, "Theoretical prediction and synthesis of (cr2/3zr1/3)2alc i-max phase," *Inorganic Chemistry*, vol. 57, no. 11, pp. 6237–6244, 2018, PMID: 29749734. eprint: https://doi.org/10.1021/acs.inorgchem.8b00021.

[181] Q. Tao *et al.*, "Two-dimensional mo1.33c mxene with divacancy ordering prepared from parent 3d laminate with in-plane chemical ordering," *Nature Communications*, vol. 8, no. 1, p. 14 949, Apr. 2017.

[182] M. Dahlqvist, J. Lu, R. Meshkian, Q. Tao, L. Hultman, and J. Rosen, "Prediction and synthesis of a family of atomic laminate phases with kagomé-like and in-plane chemical ordering," *Science Advances*, vol. 3, no. 7, e1700642, 2017. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.1700642`.

[183] M. Dahlqvist and J. Rosen, "The rise of max phase alloys – large-scale theoretical screening for the prediction of chemical order and disorder," *Nanoscale*, vol. 14, pp. 10 958–10 971, 30 2022.

[184] E. P. George, D. Raabe, and R. O. Ritchie, "High-entropy alloys," *Nature reviews materials*, vol. 4, no. 8, pp. 515–534, 2019.

[185] C. Oses, C. Toher, and S. Curtarolo, "High-entropy ceramics," *Nature Reviews Materials*, vol. 5, no. 4, pp. 295–309, Apr. 2020.

[186] Z. Li, K. G. Pradeep, Y. Deng, D. Raabe, and C. C. Tasan, "Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off," *Nature*, vol. 534, no. 7606, pp. 227–230, 2016.

[187] Z. Lei *et al.*, "Enhanced strength and ductility in a high-entropy alloy via ordered oxygen complexes," *Nature*, vol. 563, no. 7732, pp. 546–550, 2018.

[188] W. Huang, P. Martin, and H. L. Zhuang, "Machine-learning phase prediction of high-entropy alloys," *Acta Materialia*, vol. 169, pp. 225–236, 2019.

[189] Z. Pei, J. Yin, J. A. Hawk, D. E. Alman, and M. C. Gao, "Machine-learning informed prediction of high-entropy solid solution formation: Beyond the hume-rothery rules," *npj Computational Materials*, vol. 6, no. 1, pp. 1–8, 2020.

[190] P. Sarker *et al.*, "High-entropy high-hardness metal carbides discovered by entropy descriptors," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[191] K. Kaufmann *et al.*, "Discovery of high-entropy ceramics via machine learning," *Npj Computational Materials*, vol. 6, no. 1, pp. 1–9, 2020.

[192] H. Huo and M. Rupp, "Unified representation for machine learning of molecules and crystals," *arXiv preprint arXiv:1704.06439*, vol. 13754, 2017.

[193] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, p. 146 401, 14 Apr. 2007.

[194] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Physical review letters*, vol. 104, no. 13, p. 136 403, 2010.

[195] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," *Journal of Computational Physics*, vol. 285, pp. 316–330, 2015.

[196] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016.

[197] C. Nyshadham *et al.*, "Machine-learned multi-system surrogate models for materials prediction," *npj Computational Materials*, vol. 5, no. 1, pp. 1–6, 2019.

[198] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," *Computational Materials Science*, vol. 156, pp. 148–156, 2019.

[199] C. W. Rosenbrock *et al.*, "Machine-learned interatomic potentials for alloys and alloy phase diagrams," *npj Computational Materials*, vol. 7, no. 1, pp. 1–9, 2021.

[200] M. J. van Setten, M. A. Uijttewaal, G. A. de Wijs, and R. A. de Groot, "Thermodynamic stability of boron: the role of defects and zero point motion," *Journal of the American Chemical Society*, vol. 129, no. 9, pp. 2458–2465, 2007, PMID: 17295480. eprint: `https://doi.org/10.1021/ja0631246`.

[201] F. Legrain, A. van Roekeghem, S. Curtarolo, J. Carrete, G. K. Madsen, and N. Mingo, "Vibrational properties of metastable polymorph structures by machine learning," *Journal of chemical information and modeling*, vol. 58, no. 12, pp. 2460–2466, 2018.

[202] B. Grabowski *et al.*, "Ab initio vibrational free energies including anharmonicity for multicomponent alloys," *npj Computational Materials*, vol. 5, no. 1, pp. 1–6, 2019.

[203] R. Jinnouchi, F. Karsai, and G. Kresse, "Making free-energy calculations routine: Combining first principles with machine learning," *Physical Review B*, vol. 101, no. 6, p. 060 201, 2020.

[204] A. I. Duff *et al.*, "Improved method of calculating ab initio high-temperature thermodynamic properties with application to zrc," *Physical Review B*, vol. 91, no. 21, p. 214 311, 2015.

[205] M. Krynski and M. Rossi, "Efficient gaussian process regression for prediction of molecular crystals harmonic free energies," *npj Computational Materials*, vol. 7, no. 1, pp. 1–10, 2021.

[206] C. J. Bartel *et al.*, "Physical descriptor for the gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[207] N. Saunders and A. P. Miodownik, *CALPHAD (calculation of phase diagrams): a comprehensive guide*. Elsevier, 1998.

[208] J. Sato, T. Omori, K. Oikawa, I. Ohnuma, R. Kainuma, and K. Ishida, "Cobalt-base high-temperature alloys," *Science*, vol. 312, no. 5770, pp. 90–91, 2006. eprint: `https://www.science.org/doi/pdf/10.1126/science.1121738`.

[209] M. Hillert, "The compound energy formalism," *Journal of Alloys and Compounds*, vol. 320, no. 2, pp. 161–176, 2001, Materials Constitution and Thermochemistry. Examples of Methods, Measurements and Applications. In Memoriam Alan Prince.

[210] I. Ansara *et al.*, "Models for composition dependence," *Calphad*, vol. 24, no. 1, pp. 19–40, 2000.

[211] N. Dupin, U. R. Kattner, B. Sundman, M. Palumbo, and S. G. Fries, "Implementation of an effective bond energy formalism in the multicomponent calphad approach," *Journal of research of the National Institute of Standards and Technology*, vol. 123, p. 1, 2018.

VITA

Sean Darius Griesemer was born in Plano, Texas, on November 10, 1992. He attended grade schools in the Plano Independent School District and graduated from Plano East Senior High School in May 2011. The following September he entered the University of Chicago and in June 2015 received the degree of Bachelor of Arts in Physics with honors. He entered Northwestern University in September 2016 as a graduate student in the Department of Materials Science and Engineering.