NORTHWESTERN UNIVERSITY

Data-driven Functional Materials Design and Discovery

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemistry

By

Yiqun Wang

EVANSTON, ILLINOIS

June 2021

© Copyright by Yiqun Wang 2021

All Rights Reserved

ABSTRACT

Functional electronic materials have transformed modern society toward a highly digitized and interconnected global community. The ever-growing demand for electronic devices with superior functionality poses a great challenge to the state-of-the-art field-effect transistors owing to the limited charge density afforded by silicon. Materials scientists and chemists have been working closely to identify novel microelectronic materials, yet the design and discovery of these materials from the atomic-level is anything but trivial. With recent advances in machine learning algorithms as well as the advent of various crystalline materials databases with both experimental and simulated data, we are now able to exploit the strengths of data-driven methods in combination with *ab initio* simulations to efficiently and effectively discover novel materials with desired functionality. In this thesis, I employ a variety of techniques to address the electronic materials design challenge. Specifically, I focus on the lacunar spinel family, which exhibits a metal-insulator transition upon structural distortion, by applying density functional theory simulations to understand the phase-transition mechanisms and explore the materials phase space. Next, I introduce the adaptive optimization engine (AOE), a novel materials design workflow that learns directly from chemical compositions to realize multiple-property optimization. The AOE frees chemists from solely relying on their intuition in materials design. It also enables the co-design of functional materials, and is capable of efficiently identifying the compositions exhibiting superior functionality. Then, I present the `deepKNet`, a deep neural network which learns from the momentum-space crystal structure genome to make property classifications. The quantitative understanding of the structure-property relationship in crystalline materials is a key step towards efficient materials design where we optimize structure types and chemical compositions in a round-robin fashion. Lastly, I introduce the symbolic regression (SR) technique and its potential applications in materials science. This method is particularly helpful when we want to build a surrogate model mapping input features/descriptors to the output. SR will automatically search for the best function form generated

by genetic programming. Unlike other black-box machine learning models, SR offers improved interpretability and insight to the quantitative model, which is invaluable to materials researchers. I hope that my work can inspire more chemists and materials scientists with domain expertise, i.e., synthesis, characterization, theoretical simulation, and informatics, to work collaboratively to further unleash the power of data-driven materials design and discovery.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Dr. James Rondinelli, for his academic guidance and support. I learned a lot from him over the years, with professionalism being one of the most precious merits. I enjoyed my graduate studies at Northwestern, and have made some significant accomplishments here. None of them would be possible without the help from James.

I also would like to thank all members of the Materials Theory and Design Group, for being nice and supportive to me. The healthy and positive group dynamics is key to the success of each individual within our group. I enjoyed all casual conversations, group meeting discussions, “I scream” ghost pepper tasting sessions, and group BBQ events. Academic discussions on various topics, of course, also play an important role in my daily life. I specifically thank Dr. Nicholas Wagner, Dr. Liang-Feng Huang, Dr. Mingqiang Gu, Dr. Danilo Puggioni, Dr. Michael Waters, Dr. Alexandru Georgescu, Dr. Xuezeng Lu, Dr. Yongjin Shin, Nathan Koocher, Jaye Harada, Daniel Hickox-Young, Richard Saballos, Lauren Walters, Kyle Miller, Kathleen Mullin, Sean Koyama, and Peiwen Ren for insightful discussions and brainstorms.

Next, I am grateful to my collaborators for working closely with me to make scientific advances. I thank Dr. Danilo Puggioni for helping/guiding me with my first paper. I thank Dr. Nicholas Wagner for contributing to the paper on symbolic regression in materials science. I thank Akshay Iyer and Dr. Wei Chen for their work in the adaptive optimization engine paper. I thank Dr. Emily Schueller, Julia Zuo, Dr. Ram Seshadri, and Dr. Stephen Wilson for their contributions to the DM-REF project. And I thank Fei Xia, Dr. Xiao-Jie Zhang, Dr. Ram Seshadri, Dr. Stephen Wilson, Dr. Elsa Olivetti, and Dr. Danilo Puggioni for their input in the `deepKNet` work.

Thank you to my committee members, including Dr. Kenneth Poeppelmeier, Dr. Wei Chen, and Dr. Mark Hersam. Your valuable feedback to my work helped me get ready to be independent and face the outside world with confidence after graduation.

I also would like to give special thanks to Dr. Toru Shiozaki, my former advisor for the first

year and a half in graduate school at Northwestern. I learned a lot about how to become a rigorous scientist and constantly improve my work from Toru. That work attitude is vital for my future success.

I thank Dr. Anastassia Alexandrova for mentoring me during my summer research internship at the University of California, Los Angeles. I had a wonderful summer working in her research group, and I will always remember that lovely office filled with California sunshine. Anastassia introduced me into the academic research world and she literally “suggested” that I should apply to Northwestern for graduate school (and here I am). Her positive reference letters also helped me get into Northwestern and the Rondinelli group.

Last but not least, I would like to thank all my dearest family members, especially my mother, Mei Guan, for supporting my decision to pursue a Ph.D. degree in the United States. Whenever I met difficulties in either life or research, I think of my family, video chat with them over the Pacific Ocean, from which I could always get some relief and encouragement to cheer up. My girlfriend, Yiran (Maria) Xu, also offered me great support over the years, and we happily explored the Chicago area as well as many other part of the United States. I am grateful to everyone who appeared in my life, I look forward to marching into a happy future with you all.

My work is supported by the National Science Foundation (NSF) under award number DMR-1729303. *Ab initio* density functional theory simulations are performed on Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number ACI-1548562 and the DoD-HPCMP (Copper and Centennial cluster). I acknowledge partial support from the Predictive Science and Engineering Design (PS&ED) program at Northwestern University. The work is also funded in part by the Advanced Research Projects Agency- Energy (ARPA-E), U.S. Department of Energy, under Award No. DE-AR0001209.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Motivation

Functional electronic materials have dramatically transformed the society we live in over the recent decade by enabling materials platforms for powerful computing and storage units. These technical advances accelerate the digitization of the modern world, allowing the prevalence of smart phones, personal computers, as well as a variety of internet-based cloud computing services. As of October 2020, almost 59% of the global population has internet access. The ever-growing market demand for electronic devices with higher capacity and faster operation speed marks a golden era for the development of the semiconductor industry, yet also poses a great challenge for them to realize new devices on demand. According to Moore's law, the number of transistors in a dense integrated circuit doubles every two years. However, the exponential growth of computing power has come to a bottleneck owing to the physical limitations of silicon-based transistors—the already nanoscale transistors are not able to afford much higher charge densities, or further reduce their size before quantum effects begin to have a significant (and unwanted) impact on their performance. The need to find novel functional materials that outperform the current state-of-the-art silicon-based field-effect transistors has come to the attention of chemists and materials scientists, who have been working collaboratively to design and discover new materials from the atomic level.

However, effectively exploring the chemical space spanned by the multi-dimensional chemical compositions and crystal structure types, is anything but trivial. To date, most materials scientists rely much upon their chemical intuition as well as experience from years of hard work to find new materials. This intuition-driven discovery has its own success in identifying new electronic materials, yet it has become more challenging to keep up with the fast-growing market demand. It is usually highly time- and resource-consuming to propose, simulate, synthesize, and characterize the properties of a new material even in the most advanced laboratories in the world. Moreover, the time-lag between the discovery of a new material and its commercialization could be 20 years or more, owing to the complex design, synthesis, optimization, and production processes involved [1]. Therefore, there is an urgent need for scientists to find a more effective and efficient materials discovery workflow.

During the recent decade, as more general materials databases (e.g., AFLOWLIB [2], Materials Project [3], Open Quantum Materials Database [4], etc.) become available to the public, materials informatics gained popularity among materials scientists. The open-access databases alleviate the strong requirement of domain expertise in materials research, which allows young and less-experienced researchers to quantitatively understand the relationships between crystal structure, chemical composition, and materials properties, and make inferences to realize new materials by design. Meanwhile, the statistical models built to solve materials science problems offer us a novel way of understanding materials physics—from a statistical perspective. In fact, researchers have made remarkable achievements using statistical analysis to decode the structure-property relationship within a variety of materials families, and helped accelerate the discovery of novel functional materials [5, 6, 7].

In response to the 2011 Materials Genome Initiative (MGI) launched by former President Obama in an effort to double the pace of advanced-materials discovery, manufacture, and commercialization, my graduate research is focused on accelerating the design and discovery of novel functional electronic-transition materials. Specifically, I have been working towards the development of an integrated and data-driven materials discovery workflow incorporating *ab initio* theoretical simulations, statistical learning, as well as optimization theory, in order to effectively identify electronic materials with superior functionality by design.

1.2 Statement of Research Goals

The goal of my graduate studies is to understand, design, and discover new materials exhibiting metal-insulator transitions (MITs) using *ab initio* simulations and statistical learning approaches. MIT materials could switch between the metallic and insulating states controlled by external stimuli (e.g., temperature, strain, etc.), which are ideal candidate materials for novel transistors. The main hypotheses that my research is based on include (1) MITs can be reformulated into microscopic structure distortions and symmetry-breaking responses; and (2) we can use a combination of theoretical modeling and machine-learning methods to capture the intimate correlation between crystal structure, chemical compositions, and materials properties of interest, which could then inform us about regions of phase space where promising materials may exist. More specifically, my work focuses on:

(i) Understanding the metal-insulator transition mechanisms using *ab initio* simulations;

(ii) Identifying new MIT materials with superior functionality and synthesizability;

(iii) Decoding the crystal structure-property relationship using statistical learning methods; and

(iv) Developing data-driven materials discovery workflow and novel learning algorithms to understand materials physics.

1.3 Thesis Organization

This thesis is organized into eight chapters. Chapter 1 here introduces my motivation and research goals. Chapter 2 provides some basic research background information on concepts and terminology used throughout this thesis, including physics of metal-insulator transition materials—our primary target materials system; basics of first principles simulations from both theory development and computational deployment perspectives; and last, general concepts and commonly used machine learning models in materials informatics research. Starting from Chapter 3, I present some detailed research projects relevant to the thesis topic. Chapter 3 is about utilizing *ab initio* density functional theory simulations to understand the metal-insulator transitions within the lacunar spinel family with formula unit GaM_4Q_8 (M = V, Mo, Nb, Ta; Q = S, Se). Here I describe how different exchange-correlation functionals would impact property predictions on the lacunar spinels. Then, I discuss the complex phase space spanned by the multiple metastable transition-metal cluster geometries. In Chapter 4, I identify the most promising metal-insulator transition materials within the complex lacunar spinel family using featureless adaptive optimization. In this work, I not only identified 12 novel complex lacunar spinels, which simultaneously exhibit high resistive switching ratio and synthesizability, but also introduced a robust adaptive optimization engine (AOE). The AOE learns directly from chemical compositions to achieve multiple-property optimization tasks. This work enables the co-design of functional electronic materials from limited physical understanding and data availability. Chapter 5 presents a novel deep neural network, the `deepKNet`, which learns from the 3D crystal structure genome to classify multiple materials properties. This work reveals the intimate correlation between crystal structure and properties (including electronic band gap, elasticity, and thermodynamic stability). I also demonstrate that machine learning approaches could not only be used to make useful predictions or generate new crystal structures, but also help us gain more insights in materials physics from a statistical perspective. In Chapter 6, I introduce genetic programming-based symbolic regression (GPSR) and its potential applications in materials research. GPSR could automatically generate the function form, i.e., the mathematical expression, mapping the features to our target property. GPSR is free from the pre-defined function form or statistical distribution as in other conventional machine learning models, which makes it a helpful tool for materials scientists to understand the mathematical relationship between physical variables and system response. In Chapter 7, I briefly conclude my research projects by summarizing the scientific problems I met and how I resolved them through various approaches. Lastly, Chapter 8 is an exciting outlook where I discuss ideas for building an integrated materials discovery workflow by exploiting the strengths of multiple techniques I used and developed during my research. I hope that in the near future, we will be able to utilize this workflow to make more contributions to the materials research community.

CHAPTER 2
RESEARCH BACKGROUND

I start by introducing some fundamental background knowledge before proceeding to the detailed research projects. I describe three different domains most relevant to my research goals: (1) introduction to the metal-insulator transition materials, which are the primary target materials of my research; (2) density functional theory simulations on crystalline materials with periodic boundary conditions; and (3) basic ideas of machine learning and their applications in materials research.

2.1 Metal-Insulator Transition Materials

2.1.1 Definition

Metal-insulator transition materials, later referred to as MIT materials, belong to a unique family of condensed-matter crystalline materials exhibiting abrupt changes in its electrical conductivity upon various external stimuli (e.g., temperature change, electric pulse, applied pressure, etc.). Although the change in electrical resistivity upon temperature change is ubiquitous in almost all known materials, here we only consider materials with a large resistive switching ratio (e.g., $> 10^2$) and a change of sign in $d\rho/dT$ (i.e., temperature coefficient of resistance) associated with (structural or magnetic) phase transitions.

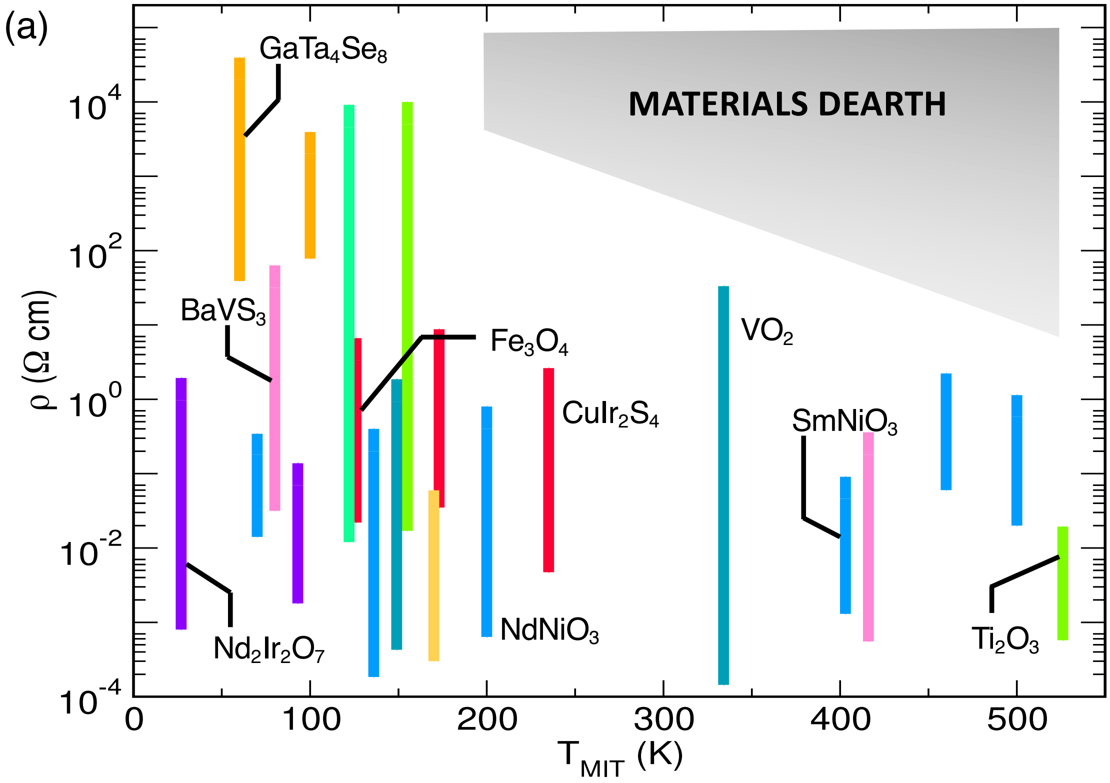
To date, we only have knowledge about $O(10^1)$ materials exhibiting MIT, most of them being transition-metal oxides and chalcogenides [8]. The wealth of electronic states afforded by these compounds originates from coupling interactions among atomic-scale structural, electronic, and magnetic degrees of freedom [9, 10, 11, 12]. As the d-orbital occupancy of the transition metal cation in these materials increases, one encounters band insulators (d^0 for titanates), Mott-Hubbard insulators with t_{2g} occupancy (d^1 titanates), competing t_{2g} – e_g orbital occupancies (for manganites and ferrites), and charge transfer insulators with e_g occupancy (nickelates) [13]. Cation valence, correlation effects, orbital physics, and dimensionality imprint structural distortions to the size, shape, and connectivity of the basic metal-ligand polyhedral building units [14]. These distortions alter the hybridization between the localized transition-metal d states and the highly polarizable ligand p orbitals, which ultimately determine the MIT characteristics. Figure 2.1(a) shows some selected MIT materials, where the vertical bars indicate the range of accessible resistivity from the metallic to insulating state. Figure 2.1(b) provides a few mechanisms behind the MITs.

2.1.2 Applications

MIT materials have garnered much research interest from the materials community over the recent decades owing to their potential to complement state-of-the-art silicon-based field-effect transistors. Metal-insulator transitions in d^n $(0 < n < 10)$ transition-metal compounds that are triggered by an applied gate bias, by charge accumulation or depletion, or by the application of a strain generated in a piezoelectric layer [15], enable fundamentally new alternatives to traditional switching devices [16, 17].

For instance, lacunar spinel $\mathrm{GaTa_4Se_8}$ and $\mathrm{GaV_4S_8}$ exhibit both volatile and non-volatile resistive switching behavior upon electric pulse [18], which makes them ideal candidates for Resistive Random Access Memory (RRAM) materials. $\mathrm{Pb(Zr,Ti)O_3}$ thin films are typically used in Ferroelectric Random Access Memory (F-RAM) [19]. With the capability to maintain the high resistivity state (off state) without supplying a voltage, these non-volatile memory materials could safely keep the digital information with lower energy consumption. Thermally-driven MIT materials such as $\mathrm{VO_2}$, on the other hand, respond to changes in the environmental temperature, and have additional applications in thermal camouflage [20, 21].

The collective response of a correlated-electron system results in device characteristics that cannot be achieved with traditional semiconductors. And the various responses of the MIT materials to different external stimuli enable us to develop materials adaptive to a variety of end environments.



(b)

Displacive Distortion		Main Interaction	Chemistries	Prototype Structures
Breathing		Valence instability	Fe, Ni, Ir, Bi	Perovskite, Spinel, Pyrochlore, Corrundum, Brownmillerites
First-Order Jahn-Teller		Orbital degeneracy	Mn, Ru, Cu	Ruddlesden-Popper, Spinel, Perovskite
Peierls-like		Metal-metal bonding	Ti, V, Nb, Ir	Rutile, NiAs, Magneli, Hollandite

Figure 2.1: (a) The range in resistivity accessible by switching (indicated by the vertical bar length) and transition temperature for a variety of MIT materials, emphasizing the chemical and compositional complexity. No materials operate above room temperature with a $\sim 10^5$ change in resistivity. (b) Relationship between local structural distortions and the physical interactions in transition-metal oxides exhibiting MITs. These structural signatures can serve as proxies from which we can design improved MIT features within the itemized chemistries and prototype structures.

2.1.3 Challenges

Although MIT materials may have a promising future in microelectronic devices as well as many other applications, it remains a grand challenge for us even to identify one novel MIT material, either from existing materials databases or from *de novo* design. This challenge mainly originates from the diversity of existing MIT materials across multiple structure types as well as chemical compositions. As shown in Figure 2.1(b), subtle structural distortions of various kinds associated with different elements determine the MITs together. These features make it challenging to formulate discovery models to predict these phase transitions.

Another key challenge for the practical realization of novel MIT materials is developing systems which display MITs that respond to realizable external stimuli, i.e., practically achievable strains or voltages. This will require going beyond existing materials and crystal structures to new classes of MIT compounds, which necessitates tackling more complex structures and/or chemistries. Such materials must possess the needed functionality of (i) high off-state resistivity to support low-power operation, and (ii) large change in resistivity at or above room-temperature for multi-state computation (Mott transistors) or for reconfigurable and purposefully transient electronics.

Moreover, although existing general materials databases have an order of 10^5 compounds, we still only have very limited knowledge about MIT materials. Typically, when a material is a priori unknown to be a MIT material, simulating the phase transitions using *ab initio* approaches to identify MIT will be challenging. Lastly, in practice, we are often faced with multiple requirements for the new material to fulfill, e.g., functionality, synthesizability, and stability. The co-design of multiple properties for MIT materials with desired properties also leaves us with much room for improvement.

2.2 *Ab Initio* Electronic Structure Theory

2.2.1 Theory

While most objects we see or interact with are mostly at visible length scale, e.g., cell phones, cars, etc., the systems that condensed-matter physicists and chemists typically focus on occur at the angstrom scale ($1\,\text{Å} = 10^{-10}\,\text{m}$). Many existing solutions to help us visualize and analyze these microscopic molecules or crystal structures utilize diffraction-based methods, e.g., X-ray diffraction, neutron scattering, etc., yet knowing their structures alone does not necessarily help us fully understand their physical properties. Therefore, theorists have approached this challenge by proposing first principles electronic structure methods to formally describe how electrons are organized in materials.

At the atomic scale, Newtonian mechanics is insufficient to describe the particle interactions. We have to take quantum mechanical effects into consideration. For any quantum system, the many-body Schrödinger equation (Equation 2.1) governs its wave function $|\Psi\rangle$.

$$\begin{aligned}\Bigg[-\sum_i \frac{1}{2}\nabla_i^2 - \sum_I \frac{1}{2M_I}\nabla_I^2 \;+\; \frac{1}{2}\sum_{i\neq j}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} \\ +\frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|} \;-\; \sum_{i,I}\frac{Z_I}{|\mathbf{r}_i-\mathbf{R}_I|}\Bigg]|\Psi\rangle = E\,|\Psi\rangle\end{aligned} \tag{2.1}$$

where i and I account for all electrons and nuclei in the system, respectively. Although there are many post–Hartree–Fock algorithms (e.g., configuration interaction [22], matrix product state [23], coupled cluster [24], etc.) that solve this second-order partial differential equation to obtain the many-body wave function, they are typically used only to accurately simulate small molecules. For extended systems like crystalline materials with periodic boundary conditions, the number of electrons becomes too large to be tractable.

An alternative way to solve for the ground state wave function is by using the density functional theory (DFT). In the DFT framework, instead of working with the coordinates of all N electrons, DFT only requires the total electron density function $\rho(\mathbf{r})$, which significantly reduces the number of variables from $3N$ to 3. Based on the two fundamental theorems proposed by Hohenberg and Kohn, which states that (1) the energy of the ground state is uniquely determined by the electron density function; and (2) by minimizing the system's energy according to the electron density, we can obtain the ground state energy E_0. The total energy can be expressed as

$$\begin{aligned} E[n] = \int d\mathbf{r} n(\mathbf{r}) V_n(\mathbf{r}) + \frac{1}{2} \int d\mathbf{r} d\mathbf{r}^{'} \frac{n(\mathbf{r}) n(\mathbf{r}^{'})}{|\mathbf{r} - \mathbf{r}^{'}|} \\ - \sum_i \int d\mathbf{r} \phi_i^*(\mathbf{r}) \frac{\nabla^2}{2} \phi_i(\mathbf{r}) + E_{xc}[n] \end{aligned} \tag{2.2}$$

where n represents the electron density, which is a function of the position vector $\mathbf{r}$, and ϕ is the single-particle wave function. The first term in Equation 2.2 accounts for the N-electron potential energy from the external field, the second term describes the electron-electron interactions, the N-electron kinetic energy comes as the third term, and the last term is the exchange-correlation energy (E_{xc}). Since the exact function form of the exchange-correlation potential (V_{xc}) is unknown, various approximations have been developed to describe E_{xc} in different solid-state systems. In Chapter 3, I discuss how different exchange-correlation potentials influence materials property simulations. By targeting the electron density function, this theory essentially seeks for the mapping between the electron density and the total energy, hence the name—density functional theory.

2.2.2 Computational simulation

Although DFT is elegant in its simplicity, there is still a gap between theory (Equation 2.2) and executable software for us to carry out the simulations. Admittedly, the development of quantum mechanical theory is quite fascinating, but the deployment and transformation from theory to computer programs is also a work of art, which reveals the beauty of human intelligence. Classic computer systems can always be broken down to operations on combination of bits (i.e., 1 and 0), but how do we manipulate the bits to solve Equation 2.2? We will not go too deep into the mathematics here, in a nutshell, we first define the Hilbert space spanned by the basis functions, e.g., Gaussian-/Slater-type atomic orbitals in a molecule or a set of plane waves in Fourier space for solid-state systems; then, we construct the Hamiltonian operator within this Hilbert space, and lastly solve for the eigenvalue and eigenvector pairs of the Hamiltonian, from which we will obtain the ground state energy and wave function (or electron density function). In computational simulations, all we consider is how electrons reside within the system, since once we know the ground state electron density function or electronic wave function, then in principle we can derive any ground state property we are interested in.

Let us take another look at Equation 2.2, which explicitly contains the electron density term $n(\mathbf{r})$. This leads to a paradox—we solve this equation to obtain the ground state electron density function and total energy, but we need to know the electron density function beforehand so as to solve this equation. The paradox leads to the famous self-consistent-field (SCF) method, where we start with an initial guess of the electron density distribution, plug it in to solve Equation 2.2, then obtain the updated electron density. This iterative procedure continues until the energy difference between two consecutive iterations are smaller than a pre-defined threshold. Equipped with modern supercomputer clusters, chemists and materials scientists are able to use the SCF-based algorithms to successfully simulate many different chemical systems from molecules to crystals.

However, the exact mathematical expression of the exchange-correlation energy term E_{xc} in Equation 2.2 is not known (i.e., no analytical expression), and scientists have developed many different functionals to achieve a balance between accuracy and computational efficiency. The effect of different functionals on materials properties is later discussed in detail in Chapter 3. To date, DFT is the most widely adopted theoretical approach to quantitatively study the electronic structure of crystalline materials.

2.2.3 Limitations

The major limitations of utilizing DFT to understand condensed-matter systems originate from both the theory itself as well as technical challenges. Owing to the fact that we do not know the exact mathematical form of the exchange-correlation term in Equation 2.2, various approximations have to be made for this significant interaction term. Generally speaking, there is a tradeoff between simulation accuracy and computational power consumption—the more accurate results we want to obtain, the more sophisticated functional we need to use (e.g., HSE06 hybrid functional), which will increase the computational expenses. Taking magnetism and relativistic effects into consideration, which are usually important for heavy-transition-metal compounds that we are interested in, would only make DFT simulations more complex. Physicists have been trying to develop new theoretical frameworks (e.g., the density matrix renormalization group algorithm [25], Green's function [26], quantum Monte Carlo [27], etc.) to overcome these challenges, although the search for new mathematics and physics to solve this problem is far from being an easy task.

With the advent of high-performance computing clusters equipped with start-of-the-art CPUs and GPUs, simulations of many large materials systems have been made possible. However, it is important to realize that theorists develop the numerical models (i.e., what approximations to make) based on the current computational power limit. If more advanced computing tools, e.g., quantum computers, become available in the near future, the community may develop new theories adapted to those systems, without having to consider many time complexity and storage constraints.

However, even if we have unlimited computational power as well as a perfect analytical solution to Equation 2.1 or Equation 2.2, it would still not be enough. We need to realize that there is a gap between theoretical models and real-world materials systems. In theoretical modeling, we typically assume a perfect and clean system. For instance, we assume that crystal structures are infinite periodic systems, without any local defect or inhomogeneity. In reality, experimentalists obtain finite-sized crystals (or even in powder form), which may contain many defects or impurities. Besides, it is almost impossible to find a truly isolated system that does not interact with the surrounding environment. Temperature and pressure, for instance, could have significant impacts on the physical properties of materials. While there are methods to incorporate these external effects, theoretical modeling will always be, strictly speaking, imperfect. Just as the statistician George Box said, "*all models are wrong, but some are useful*". It is important to keep in mind the theoretical model limitations while conducting research, and we should constantly seek for better solutions to existing problems.

2.3 Machine Learning in Materials Science

The research topics about machine learning (ML), or in general artificial intelligence (AI), first originated from the computer science research society. The astounding performance of convolutional neural networks on image classification tasks marked a new era of AI research. AI has achieved remarkable advances in a large variety of applications in the academic as well as industrial world. I will not give a comprehensive tutorial on ML or AI in this thesis, instead, I will present a brief tour in ML, followed by some of my ML-related projects to solve materials science and chemistry domain challenges in later chapters.

2.3.1 State-of-the-art machine learning algorithms

A widely quoted definition of ML by Dr. Tom Mitchell states, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [28] In other words, machine learning models learn from existing data to accomplish some tasks with some performance metric, and is supposed to do better as it becomes more experienced. The most commonly seen and widely used ML models belong to the supervised learning category [29, 30], where the goal of the model is to predict the results (target property) as accurately as possible from the input features (could be either numerical or categorical, or a mixture of both). Based on the nature of the task, supervised learning can be further categorized into regression and classification, where the goal of regression is to predict a continuous variable (typically denoted as $\hat{y}$), while that for classification models is to classify input data into different categories. Some commonly used regression models include linear regression, Gaussian processes regression, kernel ridge regression, etc. [31, 32]. Random forest, gradient boosting tree, and support vector machine, are widely used for classification tasks [33, 34]. A common practice for machine-learning practitioners is to apply multiple learning algorithms on the dataset, and choose the best-performing model for production.

While supervised learning is most widely used for industrial applications, the learning process of humans is more "unsupervised", especially at a young age. There are a variety of machine learning methods that learn directly from data without labels, then perform clustering or other density evaluation tasks. This family of algorithms belong to the unsupervised learning regime, and is also very useful for many real-world applications such as customer behavior analysis [35, 36]. Reinforcement learning, the third general category of ML algorithms, has gained popularity over the recent years owing to the capability of making consistent improvement over training [37]. The main idea behind this learning algorithm is to reward the model when it performs well for a given task, and let the model learn how to react to different scenarios based on previous experience. This is particularly useful when the state space of a problem is inaccessibly large, e.g., Go, where hard-coded instructions are difficult to establish. Through reinforcement learning, many computer robots are able to beat the top human teams in a number of well-known strategy video games. For instance, OpenAI Five [38] and AlphaStar [39] are computer bots that can beat top professional human players in Dota2 and StarCraft video games, respectively.

As the AI research society evolves over time, the deep learning community gradually started to play a dominant role while deeply influencing many other research domains. The great success of deep learning algorithms was first demonstrated by their superior performance in computer vision [40] and natural language processing applications [41, 42]. The former has already been widely used in modern autonomous vehicles, and the latter could help us analyze lengthy text files, or help make the search engine more intelligent. Deep learning is powerful for being able to learn from a huge amount of data and make generalized predictions on new input. In other words, it can achieve a balance between bias and variance through its complex non-linear transformation and decision-making layers. In fact, it has been shown that AI-driven models could outperform humans easily in a number of applications including Go (AlphaGo) [43], protein structure prediction (AlphaFold2) [44], and machine translation (GPT-3) [45].

2.3.2 Applications of machine learning in materials science

In response to the call from the Materials Genome Initiative (MGI) to integrate data-driven methods to double the pace of advanced-materials design and discovery, materials research has evolved and adapted to a more data-intensive environment. Chemists and materials scientists typically rely much upon their chemical intuition to determine which material family is more “interesting” and should be prioritized for careful examinations. This method has a strong requirement on domain expertise. Moreover, there are a thousand Hamlets in a thousand people’s eyes, it is hard to devise a standard metric to judge the chemical intuition. Machine learning models provide a viable solution to this problem. With the same dataset and model settings, two researchers from the opposite side of the Earth should be able to obtain the same statistical results. In recent scientific publications in both chemistry and materials science journals, we see an increasing amount of ML applications to predict materials properties and guide experimental synthesis or computational simulations. For example, machine-learned exchange-correlation functionals have been developed as an alternative to other physics-based DFT functionals [46]; the symbolic regression technique has been introduced to solve materials science problems with automatic function form generation (see Chapter 6). More applications of ML in chemistry and materials science can be found in the review paper in Ref [47].

Machine learning models could not only make predictions on either regression or classification tasks, but are also able to generate novel molecules or crystal structures. Materials discovery is in general a grand challenge without an effective solution that is universally agreed upon yet. Recent advances in generative adversarial network (GAN) and variational autoencoder (VAE) provide some feasible solutions to this problem [48, 49]. The central idea is to compress materials representation into a low-dimensional latent vector, by sampling from this latent space then reconstructing new structures through the decoder, we will potentially be able to identify novel structures. This method is fundamentally different from conventional intuition-driven methods where we only explore the vicinity of the known materials space. If the latent space could effectively represent the materials of interest, then the sampling technique could generate new structures "orthogonal" to the known materials space.

Another emerging research field in materials informatics involves the concept of active learning. Instead of adopting the conventional "frequentist statistics" view, active learning algorithms typically take a "Bayesian statistics" view. Specifically, rather than relying on a batch of existing training data ($> 10^2$), active learning models start from a relatively small number of initial data ($\approx 10^1$) and sequentially update the model as more data become available to the model. This sequential learning strategy is suitable for chemistry and materials research since humans also learn science in a sequential manner—we learn from previous experience to infer our next steps in a linearized scientific method. The Bayesian approach also has the advantage of uncertainty quantification, where we can have a better idea of how much we could trust the model predictions. Later in Chapter 4, I will present a project using multi-objective Bayesian optimization for functional materials design.

2.3.3 Challenges

The goal of data-centric chemists is typically not to develop new ML algorithms, instead, we mainly focus on transforming chemistry and materials science problems into the ML realm. Constructing the interface between two different research domains is non-trivial. Featurization of materials systems, for instance, requires much domain knowledge as well as feature engineering skills. Apart from directly using physical properties (e.g., atomic mass, cell volume) as features, (multi)graph representation of molecules and solid-state materials have become more prevalent since they could effectively capture the connectivity and complex interactions within the systems [5, 50]. In fact, selection of features or materials descriptors is key to achieving ML-solutions to our domain problems. The mutual information between the descriptors and target properties, i.e., how much information can be learned from the provided features to make the prediction, determines the upper limit of the statistical learning model performance. Just like using a higher kinetic energy cutoff in DFT simulations or including more atomic orbitals in Hartree–Fock simulations would yield more accurate results, selecting appropriate features to represent molecules or crystals, plays a deterministic role in materials informatics research.

Data availability is also a problem which prevents wider application of ML to materials science problems. Although more general materials databases have become publicly available, researchers typically have their unique target materials family. Information provided by the databases may be limited for some specific materials families (e.g., the lacunar spinel family). Conventional ML models exhibit better generalizability when more training data is available, which is quite often out of reach in chemistry research. Meanwhile, most existing materials databases use DFT-simulated physical properties using generalized gradient approximations with onsite Coulomb interactions for correlated systems. These simulated properties may not be accurate since DFT assumes 0 K environment under vacuum, with an approximated exchange-correlation functional.

Besides, the statistical nature of ML could also introduce some difficulty in model interpretations. Many ML models, especially deep neural networks, are famous black-box models whose interpretability is either vague or nearly impossible. After those non-linear activation functions and normalization operations, we could barely visualize or understand how the model connects the input features to the final output. To make it even worse, the classification model output is a statistical distribution over several classes, which is not a deterministic value like solving an analytical function. Therefore, we sometimes find it hard to understand how the model makes those predictions, and how much we can trust them. In Chapter 5, I will present some model interpretation work on a novel deep neural network for materials property classification tasks.

CHAPTER 3

UNDERSTANDING METAL-INSULATOR TRANSITIONS IN LACUNAR SPINELS

This chapter is composed of sections from Ref [51], which has been adapted with permission. Recent publications have also been updated throughout. This work was written in collaboration with Dr. Danilo Puggioni. © Copyright 2019 American Physical Society.

In this chapter, we perform systematic density functional theory (DFT) calculations to assess the performance of various exchange-correlation potentials V_{xc} in describing the chalcogenide $\mathrm{GaM_4Q_8}$ lacunar spinels (M=Mo, V, Nb, Ta; Q=S, Se). We examine the dependency of crystal structure (in cubic and rhombohedral symmetries), electronic structure, magnetism, optical conductivity, and lattice dynamics in lacunar spinels at four different levels of V_{xc}: the local density approximation (LDA), generalized gradient approximation (GGA), meta-GGA, and hybrid with fractional Fock exchange. We find that LDA underperforms the Perdew-Burke-Ernzerhof (PBE) and PBE revised for solids (PBEsol) GGA functionals in predicting lattice constants as well as reasonable electronic structures. The performance of LDA and GGAs can be improved both quantitatively and qualitatively by including an on-site Coulomb interaction (LDA/GGA+U) with a Hubbard U value ranging from 2 eV to 3 eV. We find that the PBE functional is able to produce a semiconducting state in the distorted polar $R3m$ phase without on-site Coulomb interactions. The meta-GGA functional SCAN predicts reasonable lattice constants and electronic structures; it exhibits behavior similar to the GGA+U functionals for small U values of 1 eV to 2 eV. The hybrid functional HSE06 is accurate in predicting the lattice constants, but leads to a band gap greater than the experimental estimation of 0.2 eV [18, 52] in this family. All of the lacunar spinels in the cubic phase are metallic at these levels of band theory, however, the predicted valence bandwidths are extremely narrow (≈0.5 eV). The DFT ground states of cubic vanadium chalcogenides are found to be highly spin-polarized, which contrast previous experimental results. With spin-orbit coupling (SOC) interactions and a Hubbard U value of 2 eV to 3 eV, we predict a semiconducting cubic phase in all compounds studied. SOC does not strongly impact the electronic structures of the symmetry-broken $R3m$ phase. We also find that these V_{xc} potentials do not quantitatively agree with the available experimental optical conductivity of GaV_4S_8; nonetheless, the LDA and GGA functionals correctly reproduce its lattice dynamical modes. Our findings suggest that accurate qualitative and quantitative simulations of the lacunar spinel family with DFT requires careful attention to the nuances of the exchange-correlation functional and considered spin structures. Last, we perform inelastic neutron scattering simulations on $GaNb_4Se_8$ and $GaTa_4Se_8$ to further explore their complex phase space spanned by the multiple geometries of the transition-metal cluster. Our simulated results qualitatively agree with recent experimental observations, where we find additional diffraction peaks appear after the symmetry-lowering structural phase transition. We also successfully identify the ground state (space group $P2_12_12_1$) with lowest energy and find it exhibit dynamic stability in both compounds. We also find there are multiple metastable competing phases accessible through minor lattice distortions from the cubic phase in both $GaNb_4Se_8$ and $GaTa_4Se_8$.

3.1 Introduction

The lacunar spinel family GaM_4Q_8 (M = Mo, V, Nb, Ta; Q = S,Se) have garnered attention for decades owing to their fascinating properties, which include metal-insulator transitions [53], the capability to host skyrmion lattices [54], and multiferroism [55]. GaV_4S_8 and $GaMo_4S_8$ are the most well studied materials in this family; they exhibit Jahn-Teller-type structural phase transitions at $\approx$40 K upon cooling, followed by spontaneous magnetic ordering below their Curie temperatures T_{C} [56]. The multiple phase transitions – metallic-to-insulating and paramagnetic-to-ferromagnetic – connecting distinct physical states make these transition-metal compounds ideal candidate materials for novel electronic platforms [52].

After decades of continuous studies on various properties of the lacunar spinel compounds, the mechanism of these phase transitions as well as the proper theoretical approaches to describe various electronic states are still unclear. For instance, while the vanadium and molybdenum compounds can undergo symmetry-lowering structural phase transitions at low temperature [57, 58, 59], the niobium and tantalum lacunar spinels remain in the high-symmetry cubic phase over a broad temperature range [60]. One of the possible reasons for this behavior in the family may be attributed to variations in the strength of electron-electron interactions [57], since electron-correlation effects are expected to be stronger in $3d$ rather than $5d$ transition metals. However, there is also evidence that local structural distortions in lacunar spinel compounds could lead to insulating states even in the absence of strong correlation [61, 62]. In addition, a number of members within the lacunar spinel family exhibit interesting resistive-switching behavior [18], making them potential materials for resistive random-access memory (RRAM) materials. Much of the literature attributes the aforementioned features to the special tetrahedral transition-metal clusters within the unit cell [63]; yet, how and why it supports all of these properties remains to be agreed upon [64, 65, 56]. In order to have a better understanding of the structure-property relationships among the lacunar spinels, a qualitative and possibly quantitative investigation of electron-correlation effects and structural distortions within these materials is needed.

Density functional theory (DFT) simulations are widely used in solid-state materials research owing to the efficiency and accuracy they achieve by replacing the original many-electron interaction problem with an auxiliary independent-particle problem through a suitably constructed exchange-correlation potential (V_{xc}). Because DFT simulations can capture the interplay of structural effects on electron-electron interactions and its dependence on determining the ground state, it is an ideal method to study the lacunar spinels with many internal atomic, spin, and orbital degrees-of-freedoms. However, no available V_{xc} can provide the exact description of exchange and correlation, which necessitates benchmarking both common and state-of-the-art density functionals against available experimental data. To that end, it becomes possible to identify the optimal functional for describing and predicting properties in the lacunar spinel family.

In this work, we systematically benchmark the performance of DFT V_{xc} functionals in describing the lacunar spinel family at four rungs of “Jacob’s ladder”, specifically the local density approximation (LDA), the generalized gradient approximation (GGA) as implemented by Perdew-Burke-Ernzerhof (PBE) and PBE revised for solids (PBEsol), the meta-GGA functional SCAN, and the hybrid functional HSE06. Our aim is to identify the best description of the lacunar spinel family from first-principles DFT simulations and where compromises on performance are made so as to facilitate future studies and predictions (see Chapter 4). To that end, we investigate the functional dependency of lattice parameters, magnetism, electronic structures, optical properties, and lattice dynamics in both the cubic and Jahn-Teller distorted rhombohedral phases. Our main conclusion is that GGA and higher level V_{xc} functionals are more reasonable than LDA in predicting almost all properties assessed. The GGA functionals with an on-site Coulomb interaction (GGA+U) value of $U \approx 2\,\mathrm{eV}$ quantitatively improves functional performance for the electronic structures of the rhombohedral phases. Spin-orbit coupling (SOC) interactions lift orbital degeneracies in the electronic structures of the cubic phases and enable a semiconducting phase to emerge with Hubbard U values ranging from 2 eV to 3 eV. However, SOC does not significantly impact the electronic structures of the rhombohedral phase, where orbital symmetry is already broken by lattice distortions. SCAN and HSE06 are able to predict accurate lattice parameters, but HSE06 leads to band gaps significantly larger than experimental estimations. Our findings suggest that the predicted physical properties of the lacunar spinel family are highly V_{xc} functional dependent. Therefore, it is important to benchmark different V_{xc} performance on properties of interest before further studies. It is likely that the coupling of internal degrees of freedom in lacunar spinels, e.g., local cluster distortion, intra- and inter-cluster electronic and magnetic interactions, underlie the observed fascinating behavior as well as our reported high sensitivity to V_{xc} in this materials family.

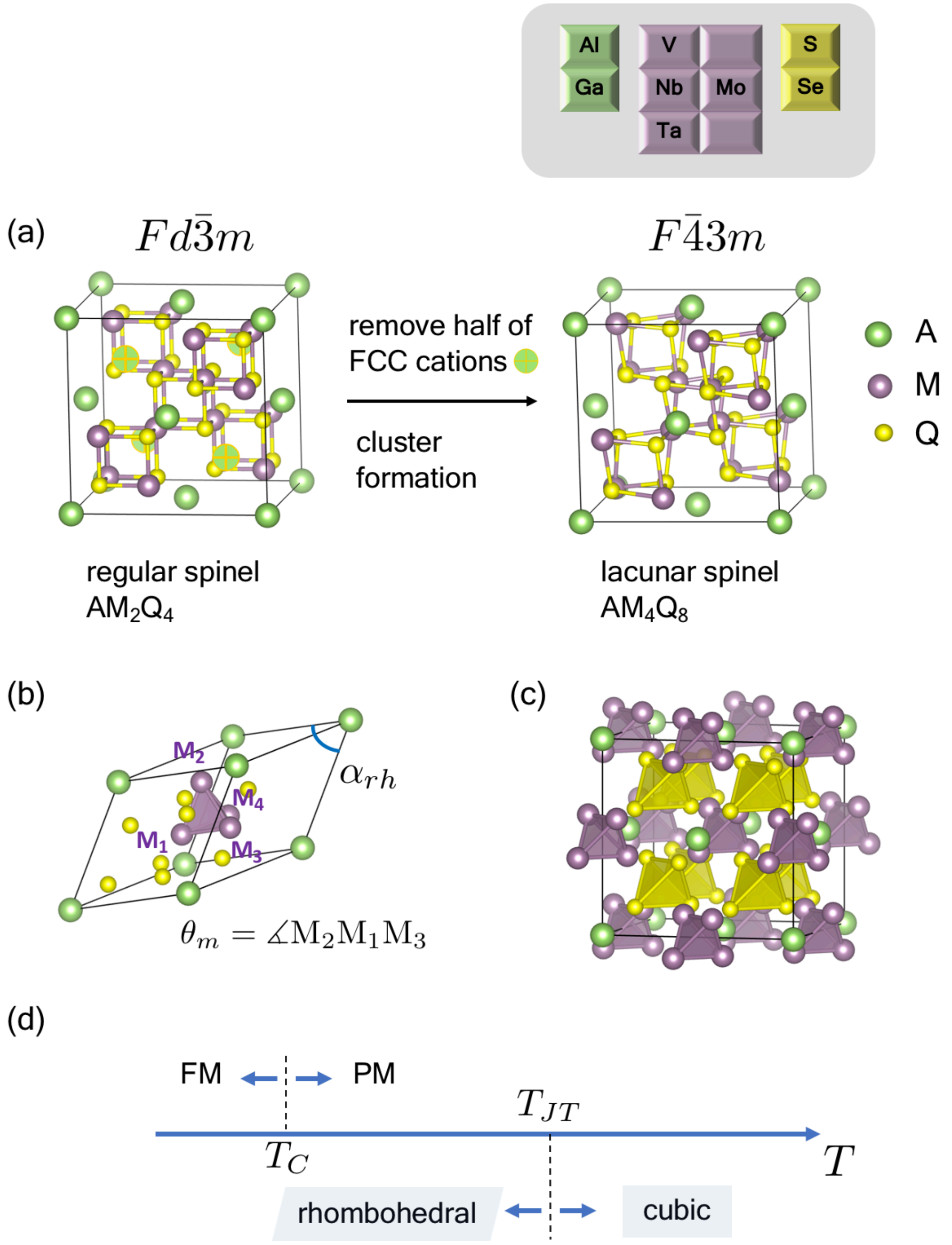


Figure 3.1: (a) Derivation of the cubic phase lacunar spinel AM_4Q_8 from an ideal spinel structure, some anions are hidden in the figure to facilitate visualization of the M_4 cluster formation. (b) The primitive cell of AM_4Q_8 in both cubic and rhombohedral phases, with the interaxial rhombohedral angle α_{rh}, intra-cluster metal-metal-bond angle θ_m. (c) M_4 cluster connectivity in the cubic phase, they occupy the four octahedral holes created by the A cations. (d) Schematic phase diagram of lacunar spinels exhibiting multiple phase transitions. (Key: FM=ferromagnetic, PM=paramagnetic).

Table 3.1: Experimental Jahn-Teller (T_{JT}) and Curie (T_{C}) transition temperatures and unit cell volumes (V) for the distorted $R3m$ vanadium and molybdenum lacunar spinels. The vanadium (molybdenum) chalcogenides exhibit acute (obtuse) angluar distortions away from the ideal cubic 60°. θ_m and α_{rh} are obtained at temperatures below T_{JT}, and no significant structural changes have been observed around T_{C}.

Compound	T_{JT} (K)	T_{C} (K)	α_{rh} (°)	θ_m (°)	$\mathrm{V}_{\mathrm{cell}}^{F43m}$ (Å^3)	$\mathrm{V}_{\mathrm{cell}}^{R3m}$ (Å^3)	Ref.
$\mathrm{GaV_4S_8}$	44	12.7	59.6	58.4	225.6	224.3	[55], [66]
$\mathrm{GaV_4Se_8}$	41	17.5	59.6	57.7	260.7	259.6	[67]
$\mathrm{GaMo_4S_8}$	45	19.5	60.5	61.6	230.1	230.0	[58], [66]
$\mathrm{GaMo_4Se_8}$	45	23	60.6	61.4	263.3	262.2	[59]

3.2 Materials and Methods

3.2.1 Crystal structure, electrical, and magnetic properties

The crystal structure of the lacunar spinel, also referred to as an A-site deficient spinel ($\mathrm{AM_4Q_8}$), is derived from the regular spinel ($\mathrm{AM_2Q_4}$ composition) by removing one of the interpenetrating FCC A-site sublattices as depicted in Figure 3.1(a). Upon removing half of the A-site cations occupying the tetrahedral holes in the regular spinel, the space group loses inversion symmetry, reducing in symmetry from $Fd\bar{3}m$ to $F\bar{4}3m$ (space group no. 216). The structure then also undergoes additional internal displacements and spontaneous strains: the previously equidistant M-M network breaks into isolated tetrahedral transition-metal clusters with chalcogenide ligands $[\mathrm{M_4Q_4}]^{5+}$. In order to quantitatively describe the internal degrees of freedom in the crystal structure, we define α_{rh} as the interaxial angle of the rhombohedral unit cell, and θ_m as the $\mathrm{M_2}$-$\mathrm{M_1}$-$\mathrm{M_3}$ angle centering the apical metal atom along the C_{3v} axis of the $\mathrm{M_4}$ cluster, as shown in Figure 3.1(b).

At room temperature, all lacunar spinels studied here exhibit cubic $F\bar{4}3m$ symmetry with $\alpha_{rh} = \theta_m = 60°$. $\mathrm{GaV_4S_8}$, $\mathrm{GaV_4S_8}$, $\mathrm{GaMo_4S_8}$, and $\mathrm{GaMo_4Se_8}$, however, undergo symmetry-lowering structural Jahn-Teller (JT) transitions at ≈40 K from $F\bar{4}3m$ to $R3m$ (space group no. 160) [57, 58, 59], followed by spontaneous magnetic ordering at a lower Curie temperature T_{C} (Figure 3.1(d)). These displacive distortions lead to a unit cell of slightly different volume, lattice parameters, and rhombohedral angle α_{rh}. The geometry of the metal cluster within the unit cell is also distorted away from its cubic structure and the θ_m angle deviates from the ideal cubic value (60°). The experimental crystallographic data for the distorted vanadium and molybdenum lacunar spinels are tabulated in Table 3.1, where we also provide the Jahn-Teller and Curie transition temperatures.

The lacunar spinels are reported to be narrow-bandwidth semiconductors with $\approx$0.2 eV band gaps that vary with temperature [56, 52, 55]. Early work showed that the valence bands mainly consist of transition-metal d orbitals [68]. Since the transition-metal M_4 clusters are relatively distant from each other with about 4 Å inter-cluster separation (Figure 3.1(c)), the low-energy valence electronic structure can be described using a molecular orbital (MO) diagram for the cluster (Figure 3.2). In the cubic phase, the valance bands are triply degenerate with t_2 symmetry. V, Nb, and Ta chalcogenide lacunar spinels all exhibit t_2^1 occupancies whereas Mo exhibits t_2^5 filling, indicating susceptibility to a first-order Jahn-Teller distortion. After the Jahn-Teller structural distortion, the triply degenerate t_2 orbital splits into two sets of orbitals, a_1 and e. The relative energy of the two sets of orbitals is occupancy-dependent; the a_1 orbital is more stable in the vanadium compounds whereas the e orbitals are preferentially stabilized in the molybdenum compounds.

In addition to these structural transitions, the vanadium and molybdenum compounds show spontaneous magnetic ordering at T_C when in the rhombohedral phase. GaV_4S_8 is also reported to have a complex magnetic phase diagram at low temperature [54]. The effective local magnetic moment in both the paramagnetic and ferromagnetic phases corresponds to approximately 1 unpaired electron per unit cell, and is mostly localized about the transition-metal cluster [56] rather than on the individual atomic sites comprising the cluster. $GaNb_4S_8$, $GaNb_4Se_8$, $GaTa_4Se_8$ are paramagnetic at ambient conditions with effective magnetic moments of 1.76 μ_B, 1.6 μ_B, and 0.7 μ_B per cluster [69, 60]. No structural phase transition or spontaneous magnetic ordering are reported in these compounds down to 1.6 K [60].

Last, we note that the family of materials is also often referred to as Mott insulators owing to the large distance between transition-metal clusters [56] and not typically because of strong electron-

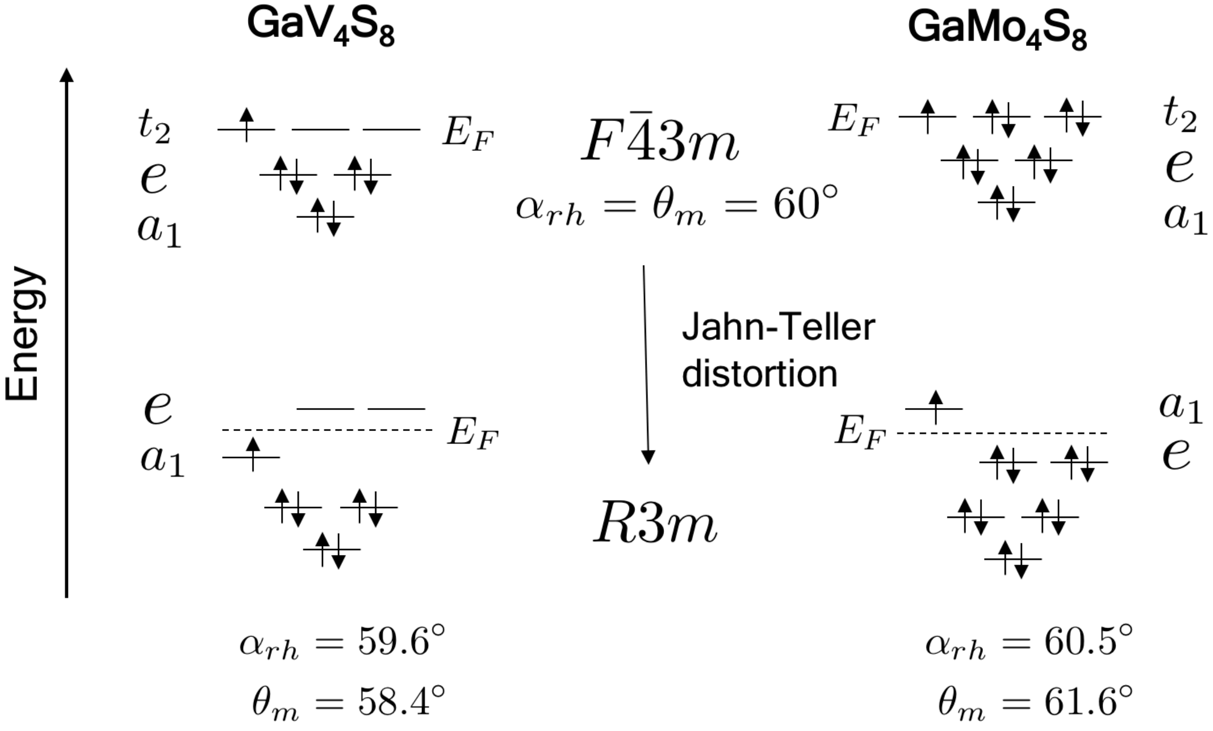


Figure 3.2: Valence molecular orbital diagrams of GaV$_4$S$_8$ and GaMo$_4$S$_8$. The valence t_2-symmetry orbitals are triply degenerate in the cubic phase. The orbital degeneracy is lifted by the accompanied Jahn-Teller distortion with a distortion sense that stabilizes and leads to filling of either the a_1 (GaV$_4$S$_8$) or e (GaMo$_4$S$_8$) orbitals based on orbital occupancy of the metals forming the cluster. The dotted lines indicate the Fermi level in the distorted phases, whereas in the cubic phase the Fermi level intersects the triply degenerate valence bands.

electron interactions [61] although they likely play some role. The semiconducting behavior is typically attributed to variable range hopping (VRH) conduction [57] among these separated metal clusters. Nonetheless, the microscopic mechanisms behind the semiconducting nature, as well as the multiple phase transitions, are still under active investigation [55, 70].

3.2.2 Exchange-correlation functionals

We use exchange-correlation potentials (V_{xc}) at four different levels of approximation to assess the structure and properties of the chalcognide lacunar spinels. The functionals examined include LDA, GGA as implemented by Perdew-Burke-Ernzerhof (PBE) [71], and PBE revised for solids (PBEsol) [72], meta-GGA functional SCAN as implemented by Sun et al. [73], and Heyd-Scuseria-Ernzerhof hybrid functional HSE06 [74]. The V_{xc} in LDA is not derived from first principles, but from Monte Carlo simulations of the uniform electron gas. The functional solely depends upon the local electron density in space and usually provides a good approximation for simple materials (including metals) with electronic states that vary slowly in space. However, the LDA potentials decay rapidly for finite systems while the true exchange-correlation potential has significant non-local contributions; this behavior often leads to overestimation of the binding energy [75] and underestimation of lattice constants in solids [76].

To improve on the LDA, GGA functionals that take the gradient of electron density $\nabla n(\mathbf{r})$ into consideration have been developed. The PBE and PBEsol functionals improve the binding energy by roughly an order of magnitude, but have a general tendency to overestimate lattice constants [76]. Since LDA and GGA functionals are well-known to be unable to predict the insulating state of Mott insulators [77] with strong correlations and nonlocal exchange, the beyond DFT method, DFT$+U$, is typically used to account for such interactions among the localized d electrons. The on-site Coulomb interaction term U favors the on-site occupancy matrix towards fillings that are fully occupied or fully unoccupied and hence a more localized electronic structure within the correlated manifold. Here, we use the GGA functionals PBE and PBEsol with on-site Coulomb interaction

(GGA+U) and U values of 1.0, 2.0 and 3.0 eV on the the M-metal sites using the formalism introduced by Dudarev et al. [78] to assess the effect of electron correlation in the M_4 clusters. The range of U values is based on results from previous computational studies [79, 80, 69] and our own preliminary assessments, where we focused on reasonable band gap and magnetic moment predictions.

It comes naturally from the previous two rungs of Jacob's ladder that the second-order derivative of the electron density should be considered. Meta-GGA functionals are essentially an extension to GGAs whereby the Laplacian of the electron density $\nabla^2 n(\mathbf{r})$ is also considered. In practice, the kinetic energy density $\tau(\mathbf{r}) = \sum_{i=1}^{N_{occ}} \frac{1}{2} |\nabla\psi_i(\mathbf{r})|^2$ is used, where the summation runs over the occupied Kohn-Sham orbitals $\psi_i(\mathbf{r})$. The recently developed meta-GGA functional SCAN (strongly constrained and appropriately normed semi-local density function) fulfills all known constraints required by the exact density functional, and is reported to have achieved remarkable accuracy for systems where the exact exchange-correlation hole is localized around its electron [73].

Hybrid DFT functionals incorporate a portion of exact exchange interaction from Hartree-Fock (HF) theory with that of a local or semi-local density functional. The semi-empirical hybrid functional B3LYP has been widely used for finite chemical systems and has shown more accurate results in thermochemical and electronic properties [81, 82]. In periodic solid state systems, one route to incorporate an exact exchange interaction is by means of range separation. In the range separated HSE06 hybrid functional, the short-range (SR) exchange interaction consists of partial contributions from exact exchange and the PBE functional. The long-range (LR) part of the Fock exchange term is replaced by that from the semi-local PBE functional. The correlation term from PBE is used in the HSE06 hybrid functional. The resulting exchange-correlation energy expression is

$$E_{xc}^{\mathrm{HSE06}} = \frac{1}{4}E_x^{\mathrm{HF,SR}} + \frac{3}{4}E_x^{\mathrm{PBE,SR}} + E_x^{\mathrm{PBE,LR}} + E_c^{\mathrm{PBE}}.$$

The inclusion of exact-exchange interactions in hybrid functionals also partly fixes the self-interaction problem in pure DFT functionals, and can provide accurate descriptions of lattice parameters, bulk moduli and band gaps in periodic systems [83, 84, 85].

3.2.3 Computational details

We perform DFT simulations as implemented in the Vienna Ab initio Simulation Package (VASP) [86, 87]. The projector augmented-wave (PAW) potentials [88] are used for all elements in our calculations with the following valence electron configurations: Ga ($3d^{10}4s^24p^1$), Mo ($4s^24p^64d^55s^1$), V ($3s^23p^63d^44s^1$), Nb ($4s^24p^64d^45s^1$), Ta ($5p^65d^46s^1$), S ($3s^23p^4$), and Se ($4s^24p^4$). Based on convergence test with respect to k-point meshes in reciprocal space and plane wave basis set cutoff energies, we use a Γ-centered $6 \times 6 \times 6$ mesh with a 500 eV kinetic energy cutoff. For HSE06 calculations, we use a $4 \times 4 \times 4$ k-point mesh and a 400 eV kinetic energy cutoff due to the high computational cost and convergence difficulties for the spin-polarized calculations. Since the lacunar spinels are small-gap semiconductors, we employ Gaussian smearing with a small 0.05 eV width. For density-of-state calculations, we use the tetrahedron method with Blöchl corrections [89].

We perform full lattice relaxations with different DFT functionals until the residual forces on an individual atom are less than 1.0 meVÅ^{-1}. The experimental crystal structures of the lacunar spinels GaV_4S_8, GaV_4Se_8, $GaMo_4S_8$, $GaMo_4Se_8$, $GaNb_4S_8$, $GaNb_4Se_8$, and $GaTa_4Se_8$ are obtained from the Inorganic Crystal Structure Database (ICSD) [90] and used as initial inputs for these geometry relaxations. $GaTa_4S_8$ is not included here since the structure is not experimentally reported. Both high-temperature cubic and low-temperature rhombohedral phases are investigated for all target compounds. Crystal structures of the rhombohedral phase of Nb and Ta compounds are obtained by making a small displacement to their cubic atomic positions along the symmetry-lowering pathway (i.e., from $F\bar{4}3m$ to $R3m$), followed by DFT structural relaxations. All experimental and DFT-relaxed crystal structures are available electronically at Ref. [91].

The effect of on-site Coulomb interactions on the crystal structures is also investigated at the LDA and GGA functional level. Since the lacunar spinels exhibit various magnetic properties, we also initialize the calculations with multiple possible magnetic configurations for the lattice relaxations. This is a necessary process owing to the multiple metastable spin configurations accessible. The magnetic configuration with the lowest energy is reported as the DFT ground state and used to compare with other functional results. Spin-orbit interactions are also considered in our electronic structure simulations owing to their potentially significant impact on the orbital structure of $4d$ and $5d$ transition metals [92]. For spin-orbit coupling (SOC) calculations, we use the fully-relaxed crystal structures from the aforementioned non-SOC simulations. The magnetic moment is set to be $1\,\mu_B$ per formula unit along the (111) direction for both the cubic and rhombohedral phases.

Zone center ($\mathbf{k} = \mathbf{0}$) phonon frequencies and eigendisplacements for both the cubic and rhombohedral phases of $\mathrm{GaV_4S_8}$ (within primitive cells) are obtained using the frozen-phonon method with pre- and post-processing performed with the `Phonopy` package [93]. Inelastic neutron scattering simulations are implemented using the dynamic structure factor simulator provided by the `Phonopy` package. 3.2 million randomly and independently generated sampling points with uniform distribution within the sampling space are used to simulate the experimentally observed inelastic neutron scattering patterns. These results are generated using our own in-house code at Ref. [94].

3.3 Results and Discussions

3.3.1 $F\bar{4}3m$ cubic phase

Lattice parameters

The crystal structures of the lacunar spinels with cubic symmetry are fully relaxed with DFT using the different V_{xc} potentials. Figure 3.3 shows the volumetric error for the DFT ground state unit cell volume relative to the experimental room temperature data. For molybdenum and vanadium compounds, we report the cell volumes of ferromagnetic spin structures with magnetic moments of $1\,\mu_B$ and $5\,\mu_B$ per formula unit, respectively. The niobium and tantalum compounds are nonmagnetic at all DFT functionals levels. See Section 3.3.1: *Magnetism* for the detailed descriptions

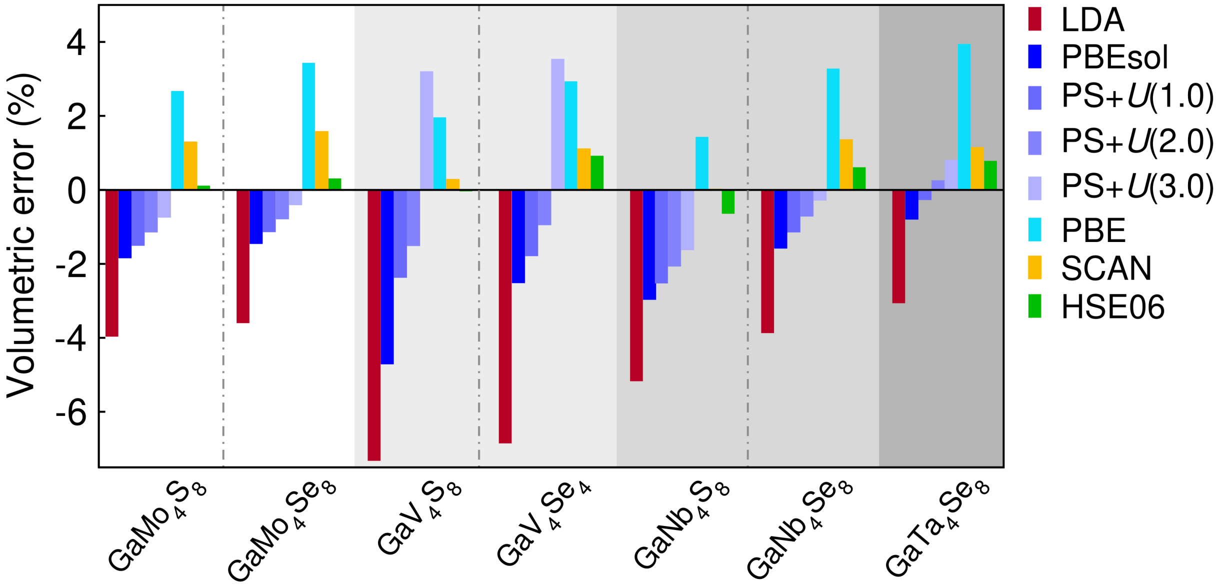


Figure 3.3: Relative error in the unit cell volume of the cubic phase at different levels of DFT. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

of the magnetic moment configurations.

In general, the LDA and PBEsol functionals underestimate the lattice parameters, while PBE predicts larger lattice constants compared with experimental data. LDA has relatively larger deviations (4 % or higher) compared with the GGA results, it is a well-known problem that LDA tends to underestimate the lattice constants. We also check the effect of on-site Coulomb interactions (LDA and GGA+U) on lattice parameters with U values up to 3 eV. With increasing on-site Coulomb interaction strength, we find the lattice parameters follow a monotonic increasing trend for both the LDA and GGA functionals. We only show the trend for PBEsol in Figure 3.3 owing to its similarity with the others. Therefore, a reasonable Hubbard U value could quantitatively improve the lattice parameter predictions in the LDA and PBEsol functionals.

Interestingly, the vanadium compounds exhibit cell volumes that are the most sensitive to the choice of the U value among the lacunar spinels. For instance, the difference in volumetric error induced by $U = 3.0$ eV is less than 2 % in $GaMo_4S_8$, but that difference is almost 8 % in GaV_4S_8. The highly spin-polarized electronic state used for the vanadium compounds may be a possible cause of the different sensitivity on the on-site Coulomb interactions.

GGA functionals with $U < 3.0\,\mathrm{eV}$ generally predict reasonable cubic lattice constants with less than 4 % error in the cell volumes. The meta-GGA functional SCAN and hybrid functional HSE06 have smaller errors in predicting lattice constants, which give less than 2% error for all 7 compounds studied here. Considering the high computational cost of structural relaxations with HSE06, SCAN should be preferred over HSE06 for lattice parameter estimation unless one requires a specific accuracy requirement or improved forces.

These results suggest that most of the DFT functionals are able to predict reasonable cubic phase crystal structures in the lacunar spinel family with less than 4 % error in the volumes. Generally, we recommend using GGA functionals with a tunable Hubbard U value of 1 eV to 3 eV for lattice parameter predictions. SCAN and HSE06 give more accurate lattice constants compared with lower-level functionals, while SCAN is preferable based on a compromise between accuracy and efficiency.

Magnetism

Experimentally, the vanadium and molybdenum compounds exhibit paramagnetism above their Curie temperatures and exhibit spontaneous magnetic ordering at low temperature [56]. The magnetically ordered phases can host multiple fascinating magnetic states, including ferromagnetism and complex spin textures (e.g., skyrmion lattices) [54]. Those complex magnetic structures are not considered here. The niobium and tantalum compounds show very weak magnetism and do not exhibit spontaneous magnetic ordering down to 1.6 K [60]. Since the transition-metal clusters are relatively far from each other with a distance of around 4 Å, the inter-cluster magnetic interactions are expected to be quite small. Here we use a ferromagnetic spin configuration on all metal sites within the cluster to model the magnetically ordered phases.

From our DFT simulations, different transition-metal clusters are able to hold various magnetic configurations. For the molybdenum compounds, we are only able to stabilize one ferromagnetic

Table 3.2: Energy differences (in eV/f.u.) of different magnetic configurations compared with non-magnetic calculations for the cubic phase. E_σ denotes the energy of the highly-polarized state with $5\,\mu_B$ or $7\,\mu_B$ per formula unit. An '–' indicates that the state was not stable. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

	GaV_4S_8		GaV_4Se_8	
	E_σ	$E_{\mu=1\mu_B}$	E_σ	$E_{\mu=1\mu_B}$
LDA	–	–	–	–
LDA+$U(1.0)$	–	0.005	-0.172	-0.007
LDA+$U(2.0)$	-0.404	-0.008	-0.68	0.06
LDA+$U(3.0)$	-0.946	-0.03	-1.264	-0.051
PBEsol	–	-0.001	-0.072	-0.008
PS+$U(1.0)$	-0.305	-0.014	-0.562	-0.028
PS+$U(2.0)$	-0.823	-0.035	-1.123	–
PS+$U(3.0)$	-1.641	–	-2.174	–
PBE	-0.095	-0.010	-0.327	-0.022
SCAN	-0.875	-0.049	-1.196	-0.064
HSE06	-1.355	0.005	-1.742	-0.092

configuration in the cubic phase which corresponds to $1\,\mu_B$ per primitive cell. The magnetic moments are evenly distributed about the four molybdenum atoms in the Mo_4 cluster with negligible contributions from other atomic species. In contrast, the vanadium compounds show numerous stable magnetic configurations (Table 3.2). Apart from the same ferromagnetic configuration as in the molybdenum compounds, we also find a highly spin-polarized state in the cubic phase. To the best of our knowledge, the electronic structures of this state has not been reported before. Recent neutron diffraction studies show that there is one single spin distributed across the V_4 cluster instead of residing on a single vanadium ion [95].

In the highly spin-polarized state, the magnetic moment could be $5\,\mu_B$ or $7\,\mu_B$ per formula unit (f.u.), depending on the DFT functional used. The spin-moments are evenly distributed about the transition-metal cluster, with approximately 1.25 μ_B localized on each vanadium atom. This state is significantly lower in energy than the ferromagnetic configuration with $1\,\mu_B$ per formula unit in our DFT simulations.

In several cases, we are not able to stabilize some of the magnetic configurations for vanadium compounds (indicated by '–' in Table 3.2). For example, LDA only converges to non-magnetic configurations, and PBE+$U = 2.0\,\mathrm{eV}$ cannot stabilize the state with $1\,\mu_B$ per cluster. In cases where both states can be stabilized, however, the more strongly spin-polarized state is always significantly more stable than the other two configurations (Table 3.2). We also observe a trend that the highly polarized state is more favored with larger on-site Coulomb interaction or with higher level DFT functionals. In addition, the $\mu = 1\,\mu_B$ state is usually energetically closer to the non-magnetic state than the highly spin-polarized state. These ground state magnetic configurations are also sensitive to the V_4 cluster volume, which we show varies with different levels of DFT functional (Figure 3.4). A larger V_4 cluster usually supports a higher magnetic moment, while a smaller volume leads to reduced or quenched moments. Our findings show that local structure and magnetic moments are correlated with each other and should be assessed carefully because both depend on the choice of exchange-correlation functional. A recent study utilizing dynamical mean-field theory simulations showed similar results, where the significance of electron correlations in describing the MO Mott physics and structural properties of $\mathrm{GaV_4S_8}$ is also reported [96].

There is also evidence that local cluster distortions still exist above the Jahn-Teller temperature [97], and that the symmetry-broken $\mathrm{V_4}$ cluster could lead to different magnetic configurations that are in better agreement with experimental results [79]. Why this occurs is attributed to the physics of the distorted phase described next. To that end, we suggest high-resolution detection methods (e.g., pair distribution function) be used to probe the local structures of cubic phase lacunar spinels.

Last, the niobium and tantalum compounds are always non-magnetic in our calculations, regardless of the initial magnetic configuration or choice of DFT functional. This may be a consequence of strong but geometrically frustrated antiferromagnetic interactions in the cubic $\mathrm{Nb_4}$ and $\mathrm{Ta_4}$ clusters [69] or due to a reduction in the on-site Hund's interactions, which drives moment formation, from the greater hybridization from the extended $4d$ and $5d$ orbitals.

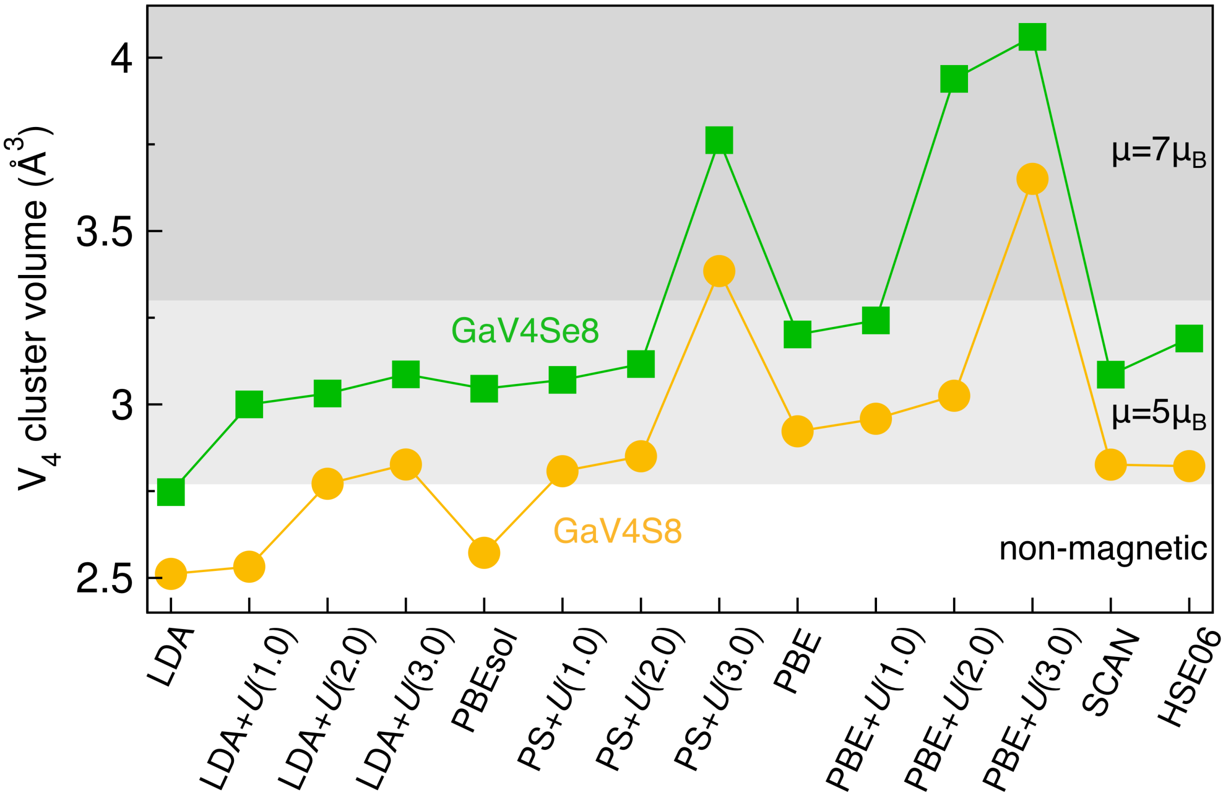


Figure 3.4: The volume of the tetrahedral V$_4$ cluster with different DFT functionals and their corresponding ground state magnetic moment per formula unit. The white area shows non-magnetic results. The light-shaded and dark-shaded areas correspond to states with $5\,\mu_B$ and $7\,\mu_B$ magnetic moments, respectively. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

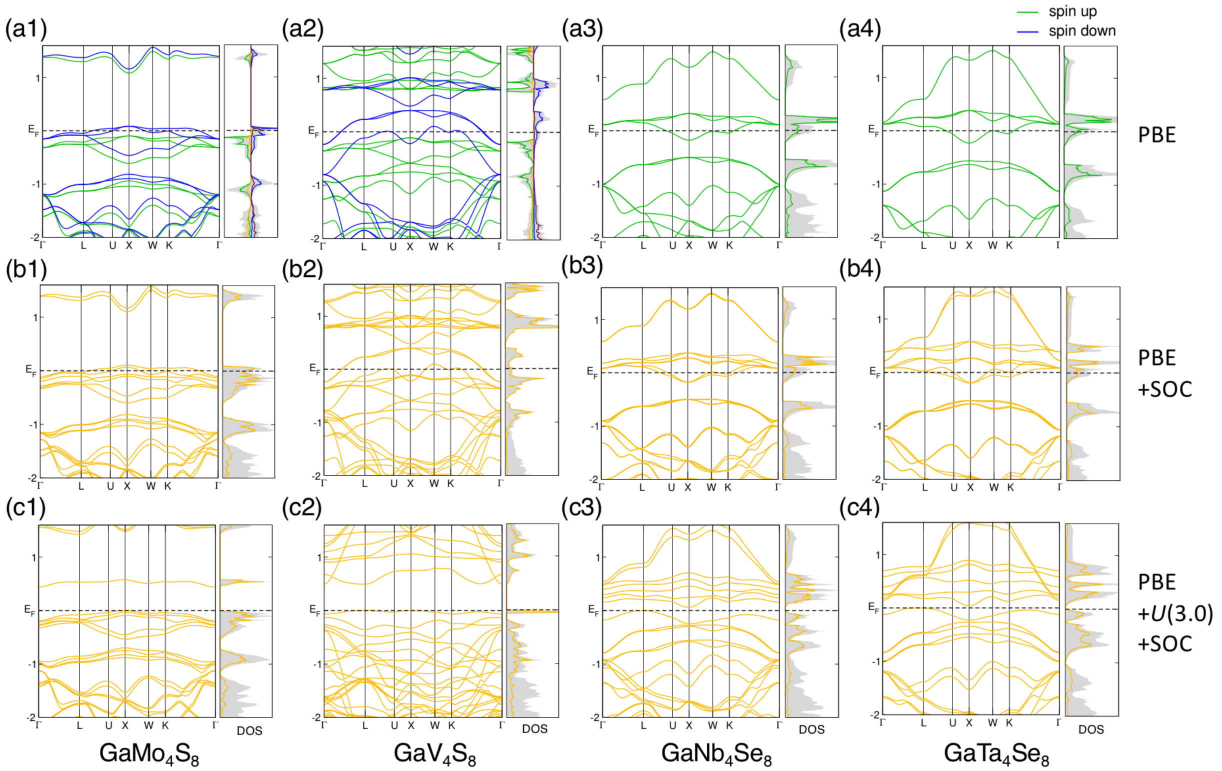


Figure 3.5: DFT-PBE ground state band structures and density of states (DOS) of the cubic lacunar spinels within a primitive cell. The Fermi level (E_F) is indicated by a broken line. The gray shaded areas in the DOS panels correspond to the total electronic density of states. The second and third rows show results with SOC included, as indicated in the rightmost column. The orange curve in the DOSs represent the contribution from the transition-metal cluster.

Electronic structures

We next use the relaxed cubic crystal structure and ground state magnetic configuration of each compound and examine the electronic structure (Figure 3.5). According to the idealized charge distribution in $Ga^{3+}[M_4X_4]^{5+}X_4^{2-}$, the number of electrons per V_4, Nb_4, and Ta_4 cluster is 7 (since they are in the same column of the periodic table) while there are 11 electrons for a Mo_4 cluster; these electrons fill the cluster orbitals depicted in Figure 3.2.

We selectively show the electronic structures of GaV_4S_8, $GaMo_4S_8$, $GaNb_4Se_8$, and $GaTa_4Se_8$ in Figure 3.5 because their S/Se counterpart compounds with the same transition-metal cluster exhibit similar band properties. From our PBE-DFT band structures and projected DOSs for the molybdenum, niobium, and tantalum compounds (Figure 3.5 a[1,3,4]), we find six valence bands mainly consisting of transition-metal d-orbital character with relatively small contribution from the anion p orbitals. The triply degenerate band (t_2 MO symmetry) is higher in energy than the doubly (e MO) and singly (a_1 MO) degenerate bands. The band degeneracy and ordering agree well with the cluster MO descriptions of the low-energy electronic structure in these compounds.

The DFT ground state electronic structures of the vanadium compounds, however, are significantly different from the other chalcogenides in the lacunar spinel family (Figure 3.5 a2). All six valence bands in the spin-up channel (green bands) are fully occupied, while only the lower part of the spin-down channel is partially occupied. The triply degenerate spin-down bands are shifted ≈1 eV above the Fermi level. Interestingly, the metastable magnetic state with 1 μ_B per cluster exhibits band dispersions that are more similar to the rest of the family (Figure 3.6) and the magnetic moment of 1 μ_B agrees better with experimental results. It remains unknown whether this DFT ground state in the cubic phase is stable and experimentally accessible; further low-temperature neutron-based scattering measurements, for example, could be used to probe the existence of this spin configuration.

We next quantitatively assess the impact of different V_{xc} as well as on-site Coulomb interactions on the electronic structures by defining two parameters, γ and Δ, as shown in Figure 3.6, which describe the key features in the band structure. γ corresponds to the energy difference between different spin-channels of the triply degenerate valence band at the Γ point. Δ quantifies the magnitude of the splitting among the triply-degenerate minority-spin bands at the X point, $k = (1/2, 0, 1/2)$, near E_F. The values of γ and Δ for the chalcogenide lacunar spinels at different levels of DFT theory are tabulated in Table 3.3. For the non-magnetic Nb and Ta compounds, we only report Δ.

All of the cubic phase lacunar spinels are metallic from band theory without considering spin-orbit interactions. Figure 3.5 shows that the Fermi level, E_F, is always located within the valence bands, regardless of the magnetic configuration or DFT functional. Specifically, the cubic phase

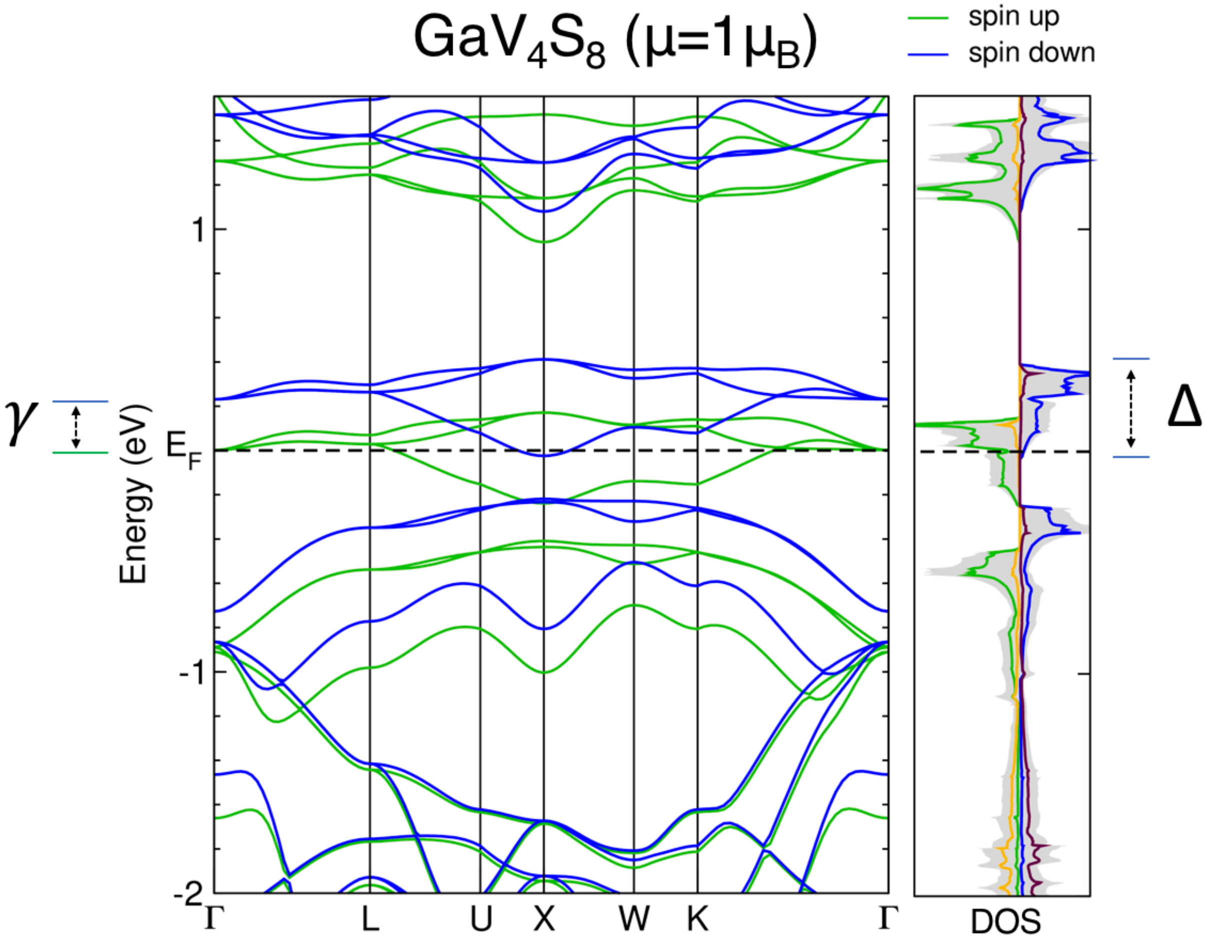


Figure 3.6: DFT-PBE band structure and DOS of metastable cubic GaV$_4$S$_8$ with 1 μ_B per formula unit. We define γ as the exchange splitting between different spin channels and Δ as the splitting of the three valence bands at the X-point. Here, those bands are located approximately at E_F and $E_F + 0.3\,\text{eV}$.

Table 3.3: Electronic band splitting of the triply degenerate bands in the cubic lacunar spinels at different levels of DFT. γ quantifies the splitting between the two spin channels. Δ is the value of band splitting among triply degenerate bands at the X point in momentum space near E_F. For the vanadium compounds, these values are tabulated for different spin-magnetic moment states separately. An '–' indicates that the state was not stable.

compound		LDA	PBEsol	PBE	PBE+$U(1.0)$	PBE+$U(2.0)$	SCAN	HSE06
$GaMo_4S_8$	γ	0.09	0.15	0.16	0.22	0.28	0.22	0.56
	Δ	0.69	0.65	0.55	0.55	0.55	0.59	0.67
$GaMo_4Se_8$	γ	0.11	0.14	0.16	0.22	0.29	0.22	0.52
	Δ	0.50	0.48	0.38	0.39	0.40	0.44	0.52
GaV_4S_8	γ	–	–	1.13	1.58	2.01	1.76	3.1
(5 μ_B)	Δ	–	–	0.48	0.44	0.33	0.43	0.46
GaV_4S_8	γ	–	0.12	0.23	0.32	–	0.37	0.95
(1 μ_B)	Δ	–	0.52	0.41	0.40	–	0.43	0.55
GaV_4Se_8	γ	–	1.07	1.17	1.65	2.60	1.82	3.2
(5 μ_B)	Δ	–	0.42	0.36	0.34	0.20	0.35	0.51
GaV_4Se_8	γ	–	0.20	0.23	0.34	–	0.40	0.81
(1 μ_B)	Δ	–	0.37	0.30	0.28	–	0.31	0.41
$GaNb_4S_8$	Δ	0.88	0.82	0.69	0.71	0.71	0.73	0.86
$GaNb_4Se_8$	Δ	0.63	0.59	0.50	0.52	0.53	0.55	0.66
$GaTa_4Se_8$	Δ	0.74	0.71	0.60	0.63	0.63	0.65	0.69

V and Mo compounds are predicted to be half-metals as only one spin channel crosses the Fermi level whereas the other spin-channel is fully gapped. In either group VB or VIB transition-metal compounds, there is an odd number of electrons in three degenerate bands (Figure 3.2). For the low spin-polarized states with 1 μ_B magnetic moment per formula unit, we then find that the Fermi level crosses this set of triply degenerate bands and metallicity is protected by the $F\bar{4}3m$ crystal symmetry. Here the splitting of these triply degenerate valence bands throughout the Brillouin zone is quite small; although the Δ value is functional dependent, it does not exceed 0.7 eV. There is also a small trend of increasing splitting between different spin channels (γ) with higher levels DFT functionals. We attribute this to the more accurate exchange interactions captured with the more advanced functionals.

The flat valence bands derived from these cluster orbitals lead to large effective masses, and these electrons should be highly localized in real space. This is in agreement with the fact that the transition-metal clusters are far from each other within the unit cell, and the electrons are highly localized within the cluster. One of the possible conduction mechanisms for the lacunar spinels is through variable-range hopping (VRH) [57]. It is for the same reason that these compounds have been called "Mott insulators" [56].

For the highly-polarized magnetic state in the vanadium compounds, γ is much larger than Δ, which makes it different from the rest of the family. In this case, the Δ term may not be that important since the triply degenerate band is no longer the highest occupied band. The two bands crossing the Fermi level are the a_1 and e orbitals in the spin-down channel. It is therefore possible to obtain a semiconducting state by shifting the e-symmetry orbitals to higher energy and fully occupying the a_1 orbital. Indeed, we find such a state in $\mathrm{GaV_4Se_8}$ using the SCAN functional, where the band gap is approximately 60 meV. Whether this highly-polarized state is experimentally accessible, however, remains unclear.

We next report results with SOC included in our simulations. The band structures and DOSs with the PBE functional are shown in Figure 3.5 b[1-4]. Orbital degeneracy is partly broken compared with the non-SOC band structures. The broken symmetry here is vital for reproducing a semiconducting state since it enables further orbital splitting by increasing electron-electron interactions. Figure 3.5 c[1-4] show the electronic structures with PBE+SOC and a U value of 3.0 eV, where all four compounds exhibit a small but finite band gap. It is interesting to note that both SOC and on-site Coulomb interactions are necessary in order to produce a semiconducting cubic phase for all compounds studied. Intuitively, SOC serves the purpose of symmetry-breaking in the highly-symmetric cubic phase while on-site Coulomb interactions localize electrons and increase repulsion between bands, which eventually lead to a semiconducting state in the cubic lacunar spinels. Although the electron-correlation effect (modeled by the Hubbard U) is typically considered more important in $3d$ transition metals, spin-orbit interactions are more significant in $5d$ transition metals. Indeed, the lacunar spinel compounds investigated, which include transition metals from the $3d$, $4d$, and $5d$ rows, exhibit similar yet non-identical behaviors. This behavior could be the outcome of competing SOC and on-site Coulomb interactions within these transition-metal cluster systems. It has been shown that spin-orbit coupling effect within the lacunar spinel system could lead to exciting physics (e.g., spin-orbital entangled molecular j_{eff} states [92, 98]).

Our findings in the cubic phase lacunar spinels indicate that different DFT functionals, as well as various internal electron-electron, spin-orbital interactions, can lead to qualitatively different interpretations of their electronic and magnetic properties. Therefore, extra care in the exchange-correlational functional selection should be taken before pursuing extensive DFT simulations on this family.

3.3.2 $R3m$ distorted phase

Lattice parameters

In this section, we investigate the DFT functional dependency of properties in the distorted rhombohedral phase. Since only molybdenum and vanadium compounds are reported to exhibit Jahn-Teller-type structural distortions, we benchmark the V_{xc} performance in predicting lattice param-

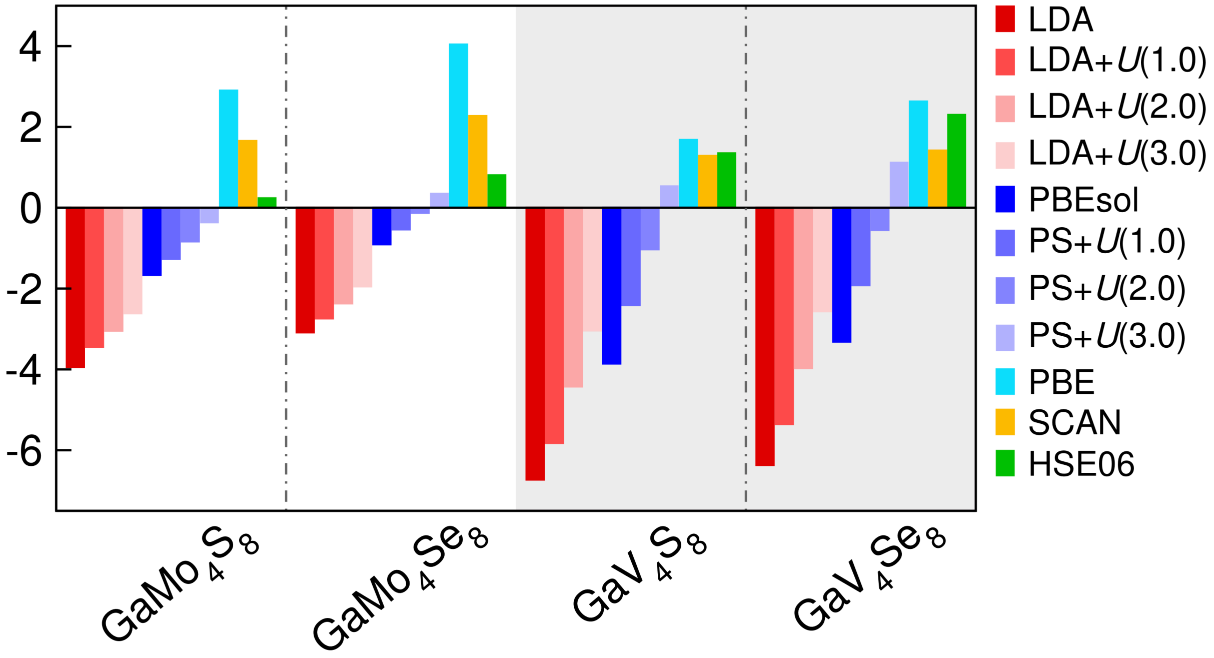


Figure 3.7: Relative error of the rhombohedral unit cell volume at different levels of DFT. PS is an abbreviation for the PBEsol functional and the number in parentheses is the value of the on-site Coulomb interaction used in the LDA/GGA+U method.

eters against available experimental data of GaV_4S_8, GaV_4Se_8, $GaMo_4S_8$, and $GaMo_4Se_8$. In all cases, we use a ferromagnetic spin configuration with $1\mu_B$ magnetic moment per unit cell in our structural relaxations; see Section 3.3.2: *Magnetism* for a detailed discussion of the magnetic moment configurations.

The Jahn-Teller structural phase transition reduces the crystal symmetry from space group $F\bar{4}3m$ to $R3m$ and occurs with a change in unit cell volume. The relative error of the fully relaxed unit cell volumes for the molybdenum and vanadium compounds are shown in Figure 3.7. Here, we observe a similar trend as found in the cubic phase. The LDA and PBEsol functionals underestimate the ground state lattice volume, while LDA shows larger deviations from the experimental data. Moreover, structural relaxations of the rhombohedral phases of GaV_4S_8 and $GaMo_4S_8$ with LDA converge to non-magnetic cubic structures, regardless of the initial magnetic moment configurations. LDA is able to stabilize a ferromagnetic configuration in the rhombohedral phase only with on-site Coulomb interactions (LDA+U).

PBE overestimates the lattice constants of all four compounds. With increasing value of the on-site Coulomb interactions, the lattice parameters also increase slightly. In the rhombohedral phase, the cell volume of vanadium compounds is not as sensitive to the Hubbard-U value as in the cubic phase, presumably because the electronic structure is semiconducting in the $R3m$ symmetry. SCAN and HSE06 functional again perform quite well with regard to the lattice parameters with less than 2% error.

Internal degrees of freedom

The occupied Wyckoff sites of the transition metals also split upon the transition into the rhombohedral phase, leading to one apical site [M_1 in Figure 3.1(b)] along the C_{3v} distortion axis and three basal atoms [M_2, M_3, M_4 in Figure 3.1(b)] forming a plane perpendicular to the C_{3v} axis. The Wyckoff positions of the transition metals in $GaMo_4S_8$ and GaV_4S_8 with $R3m$ symmetry (space group no. 160) after structural relaxation with different exchange-correlation functionals are tabulated in Table 3.4. The selenide compounds show similar functional dependencies and are not shown here. In general, the changes in Wyckoff positions with respect to functional are quite small. However, we find that the z_1 value in GaV_4S_8 has a significantly higher functional dependency over that in $GaMo_4S_8$ (Figure 3.8). Both increasing the value of U as well as going to higher levels of exchange-correlation functionals favor larger structural distortions in GaV_4S_8, i.e., keeping the apical V atom far away from the center of the tetrahedral transition-metal cluster. The z_1 Wyckoff position of the Mo atoms is also largely insensitive to the choice of the DFT functional, possibly owing to the reversed distortion in $GaMo_4S_8$, where steric effects might prohibit further distortion.

After the structural phase transition, both the rhombohedral angle α_{rh} and bond angle θ_m in the transition-metal cluster diverge from the cubic 60°, leading to a greater number of internal degrees of freedom in the distorted phase. The latter is correlated with the change in occupied Wyckoff sites of the transition metals. We record these internal bond angles of the four lacunar spinels after

Table 3.4: Wyckoff positions of the transition metals in rhombohedral $GaMo_4S_8$ and GaV_4S_8 after structural relaxation with different DFT functionals. The z value of the $3a$ and $9b$ sites in space group no. 160 are labeled z_1, z_2, respectively. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

	$GaMo_4S_8$			GaV_4S_8		
	$3a$ (z_1)	$9b$ (x)	$9b$ (z_2)	$3a$ (z_1)	$9b$ (x)	$9b$ (z_2)
experimental [66]	0.4014	0.1956	0.2023	0.3910	0.1937	0.2013
LDA	0.3982	0.1960	0.2022	0.3944	0.1946	0.1998
PBEsol	0.4012	0.1951	0.2012	0.3913	0.1969	0.2005
PS+$U(1.0)$	0.4014	0.1950	0.2011	0.3877	0.1966	0.2005
PS+$U(2.0)$	0.4015	0.1949	0.2010	0.3856	0.1958	0.2005
PS+$U(3.0)$	0.4016	0.1948	0.2010	0.3834	0.1951	0.2008
PBE	0.4020	0.1956	0.2009	0.3888	0.1972	0.2005
SCAN	0.4029	0.1962	0.2006	0.3852	0.1963	0.2004
HSE06	0.4025	0.1959	0.2007	0.3839	0.1956	0.2005

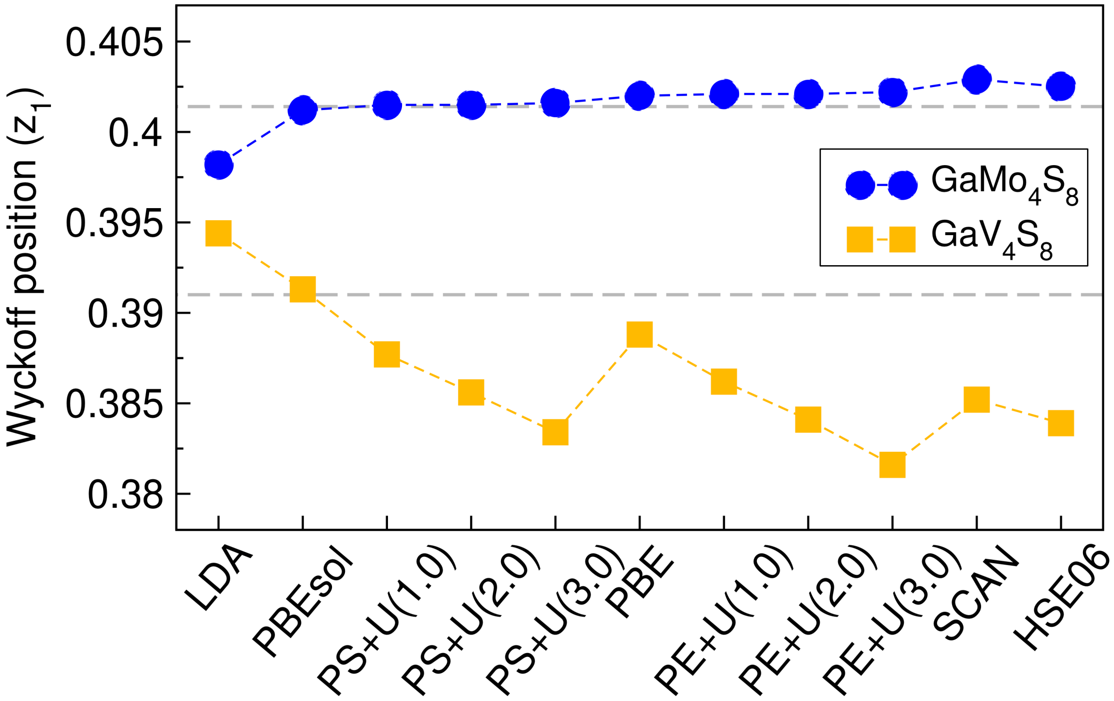


Figure 3.8: The $3a$ (z_1) Wyckoff position in rhombohedral $GaMo_4S_8$ and GaV_4S_8 at different V_{xc}. The gray dashed lines correspond to the experimental values. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

Table 3.5: The unit cell rhombohedral angle α_{rh} (in degrees) for the $R3m$ phases and the corresponding apical bond angle θ_m (in degrees) for the transition-metal cluster at different levels of DFT functional. PS is an abbreviation for the PBEsol functional and the number in parenthesis is the value of the on-site Coulomb interaction used in the GGA+U method.

	$GaMo_4S_8$		$GaMo_4Se_8$		GaV_4S_8		GaV_4Se_8	
	α_{rh}	θ_m	α_{rh}	θ_m	α_{rh}	θ_m	α_{rh}	θ_m
experimental	60.47	61.60	60.57	61.43	59.62	58.38	59.56	57.72
LDA	60.00	60.00	60.78	62.76	60.00	60.00	59.55	57.71
PBEsol	60.70	62.29	60.80	62.91	59.56	57.67	59.33	56.59
PS+$U(1.0)$	60.75	62.50	60.81	63.00	59.28	56.39	59.20	56.00
PS+$U(2.0)$	60.76	62.59	60.81	63.06	59.16	55.91	59.09	55.60
PS+$U(3.0)$	60.77	62.66	60.81	63.12	58.99	55.26	58.87	54.75
PBE	60.73	62.53	60.79	63.07	59.36	56.62	59.26	56.09
SCAN	60.76	62.80	60.84	63.27	59.21	55.72	59.12	55.47
HSE06	60.77	62.74	60.79	63.15	59.05	55.38	58.94	55.04

structural relaxation using different DFT functionals; the results are shown in Table 3.5.

Almost all DFT functionals (except for LDA) predict similar results for α_{rh} compared with the experimental data, but in general they give larger local metal-cluster distortions. θ_m values are $1\sim2$ degrees larger than experimentally reported in the molybdenum compounds, whereas the vanadium compounds show a similar but reversed trend in θ_m, i.e., $1\sim2$ degrees smaller. This agreement is reasonable, and the difference in internal coordinates compared with experiment might come from low-resolution experimental characterization [61]. It is also possible that the structural phase transition is incomplete at low temperature [58]. We also find a minor trend that higher-level functionals, as well as larger Hubbard-U values, favor larger structural distortions. Since our DFT simulations are performed at 0 K, while lab characterizations are performed at finite temperature, our results are more likely to capture the correct ground state structure where thermal expansion effects are small.

It is interesting to note that the vanadium and molybdenum compounds show reversed structural distortions across the phase transition. This can be explained from the valence MO diagram in Figure 3.2. In the cubic phase, there is either 1 electron (GaV_4S_8) or 5 electrons ($GaMo_4S_8$) in the valence t_2 orbitals. Such electronic configurations are Jahn-Teller active, whereby a structural distortion accompanied by orbital-degeneracy lifting could further stabilize the system. Owing to the different electron occupations in the vanadium and molybdenum compounds, their favored electronic configurations require a reversed ordering of the a_1 and e orbitals. Therefore, the spontaneous structural distortion permits each compound to lift its orbital degeneracy and achieve its favored electronic configuration.

We now summarize the structural benchmark assessment of the cubic and rhombohedral phases. We recommend using the GGA+U method for lattice structure relaxations with a Hubbard-U value of approximately 2 eV to 3 eV. LDA functional should be used with on-site Coulomb interactions for both cubic and rhombohedral phases. The SCAN functional is another reasonable choice that predicts accurate lattice structures. Structural relaxations with HSE06 give very accurate lattice constants, but its high computational costs may be prohibitive if only trying to obtain reasonable crystal structures.

Magnetism

We find that the DFT ground state of both the molybdenum and vanadium compounds in the distorted $R3m$ structure are ferromagnetic with $1\,\mu_B$ per formula unit. (Simulation of the complex magnetic phase diagram of lacunar spinels is out of the scope of this work, readers with interest should refer to Refs. [54, 55, 99, 100, 101, 102, 103, 104].) In the molybdenum compounds, the magnetic moment is evenly distributed about all four Mo atoms, which is the same as what we found in the cubic phase. In the vanadium compounds, however, the apical V atom along the C_{3v} symmetry axis has a large local magnetic moment. The other three basal V atoms have relatively smaller moments that are anti-aligned to the apical spin. This results in a ferrimagnetic configuration in the V_4 cluster, giving a net-magnetic moment of $1\,\mu_B$ per formula unit. For instance, in the rhombohedral phase of GaV_4S_8 with the PBE functional, the magnetic moment on the apical V atom is $1.3\,\mu_B$ while the three basal V atoms contribute each $-0.1\,\mu_B$. Thus, the net-magnetic moment in one formula unit of GaV_4S_8 is $1\,\mu_B$. Some DFT functionals (e.g., PBE+$U(1.0)$) are able to stabilize a ferromagnetic configuration in GaV_4S_8 similar to that of $GaMo_4S_8$, i.e., evenly distributed, but this magnetic configuration is less energetically favorable compared with the ferrimagnetic configuration. We report properties of the rhombohedral phase vanadium compounds using the ferrimagnetic configuration in the remainder of this chapter. A recent work that used random-phase approximation correctly reproduced the ground state of GaV_4Se_8 and explored the coupling between magnetism and structure [62].

Electronic structures

Next, we examine the electronic structures of rhombohedral GaV_4S_8, $GaMo_4S_8$, $GaNb_4Se_8$, and $GaTa_4Se_8$, because there is evidence that symmetry-breaking in the transition-metal clusters without distortion of the lattice parameters could lead to different magnetic configurations [79]. Such small local distortions may also be challenging to detect with low-resolution characterization techniques; for that reason, we hypothesize that the niobium and tantalum-based lacunar spinels could also exhibit a distorted rhombohedral phase. Therefore, we slightly distort the cubic niobium and tantalum lacunar spinel structures along the symmetry-breaking pathway, and use these geometries as the initial structure for structural relaxations. The structural relaxation settings using different DFT functionals are similar to those used for the cubic phase.

Figure 3.9 a[1-4] presents the electronic band structures and projected DOSs of these four compounds. The triply degenerate valence bands in the cubic phase split into two sets of orbitals with a_1 and e symmetry. In the molybdenum compounds, the minority spin a_1 orbital shifts to higher energy, above the Fermi level, such that five valence electrons occupy the three majority spin orbitals and the minority e orbitals. The vanadium, niobium, and tantalum compounds exhibit different orbital occupations and structural distortions; the a_1 orbital is further stabilized to lower energy relative to the other five orbitals, and the only one valence electron occupies the a_1 orbital. Remarkably, we find that the PBE functional is able to open up a small band gap without any on-site Coulomb interactions in the distorted phase GaV_4S_8 and $GaMo_4S_8$. For $GaNb_4Se_8$ and

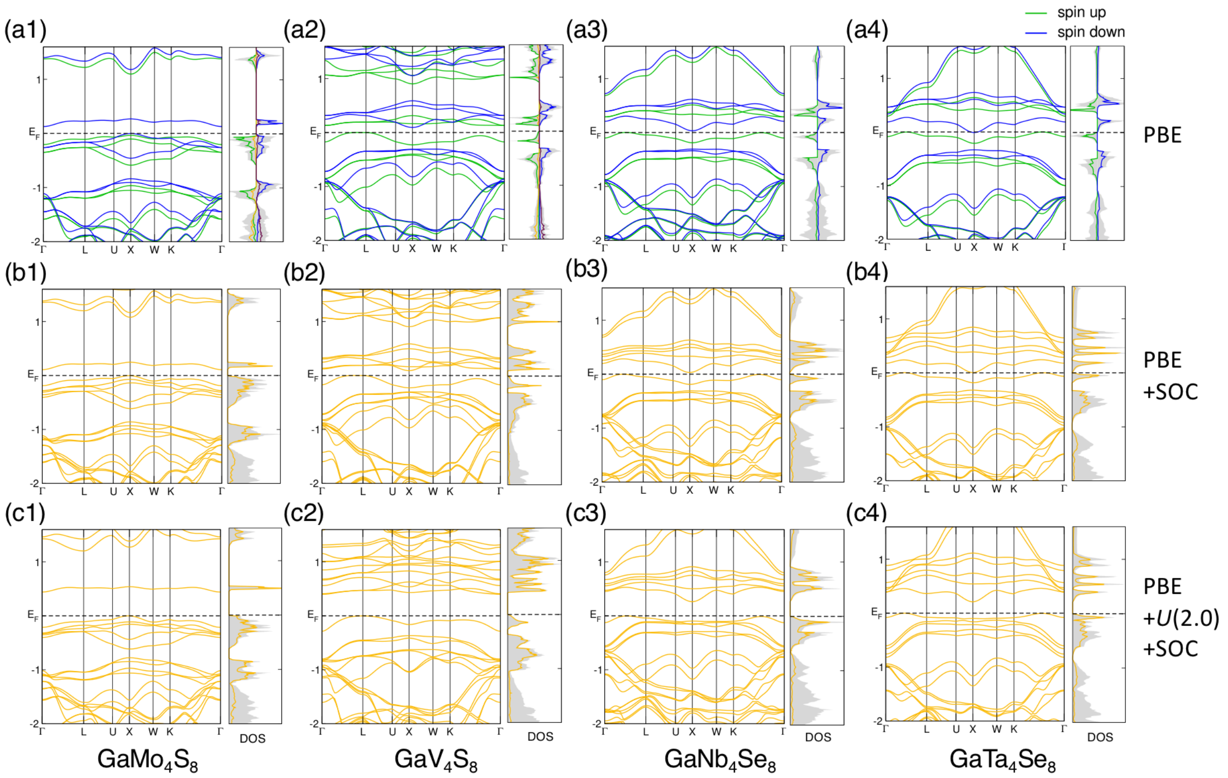


Figure 3.9: DFT-PBE ground state band structures and density of states (DOS) of the rhombohedral $R3m$ lacunar spinels. The second and third rows show simulation results with SOC. Color representation is the same as that in Figure 3.5.

$GaTa_4Se_8$, the lowest conduction band (minority a_1) barely touches the Fermi level. However, LDA predicts an unreasonable metallic ground state for these compounds. This finding indicates that the structural distortion alone is sufficient to lift the orbital degeneracy and open a semiconducting gap without strong electron-correlation effect—apparently the additional electron density gradient in V_{xc} through the enhancement factor provides an improved description. In other words, the rhombohedral lacunar spinels may not be strictly described as “Mott” insulators. Our findings confirm the importance of local structural distortions on electronic structures in lacunar spinels [61].

Meanwhile, different GGA functionals qualitatively describe the rhombohedral electronic structures differently. For instance in GaV_4S_8, PBEsol predicts a metallic state, while PBE opens a small gap of 0.09 eV. In $GaMo_4S_8$, PBEsol gives a very small band gap of 0.02 eV while PBE predicts a band gap of $0.13\,\mathrm{eV}$, which is much closer to the experimentally estimated value of $0.2\,\mathrm{eV}$. Therefore, DFT-GGA simulations on this family of compounds should be performed with extra caution with attention focused on the role of the enhancement factor in reducing the self-interaction error [105]. We recommended that when using the PBEsol functional to simulate the electronic structures of the lacunar spinels, a slightly larger ($\sim$1 eV) Hubbard U value is used than that for PBE.

Figure 3.9 b[1-4] shows the electronic structures with the PBE functional and SOC. The effect of including SOC is similar to that in the cubic phase; orbital degeneracy is lifted, but the overall band structures remain similar. Figure 3.9 c[1-4] shows the effect of now adding a Hubbard on-site Coulomb interaction of $2.0\,\mathrm{eV}$. All four compounds now exhibit a clear band gap, where the conduction band is pushed to higher energy owing to stronger electron-electron interactions. Interestingly, we find that SOC does not seem to play a decisive role in predicting reasonable electronic structures in the rhombohedral phase, whereas the GGA functional alone could predict qualitatively correct behavior. It is possible that SOC plays a less significant role here since crystal symmetry is already broken in the rhombohedral phase, unlike in the highly symmetric cubic phase. Our findings here support our previous hypothesis about the roles of SOC and electron-correlation effect in producing semiconducting phases.

Figure 3.10 shows the different band gaps predicted using different DFT functionals. The decreasing band gap in the V-Nb-Ta series from $3d$ to $5d$ agrees well with our physical intuition, where electron-correlation effects are expected to decrease. The reason why molybdenum compounds show large band gaps might be caused by different orbital occupations – more valence electrons lead to larger orbital repulsion, which pushes the conduction band to a higher energy level, leading to a larger band gap. We also observe a higher functional rather than compositional dependency on the band gap. With an increasing Hubbard-U value, the band gap increases monotonically. Since a larger on-site Coulomb interaction effectively increases the repulsion between bands with the same spin, it results in a larger gap between the highest occupied and lowest unoc-

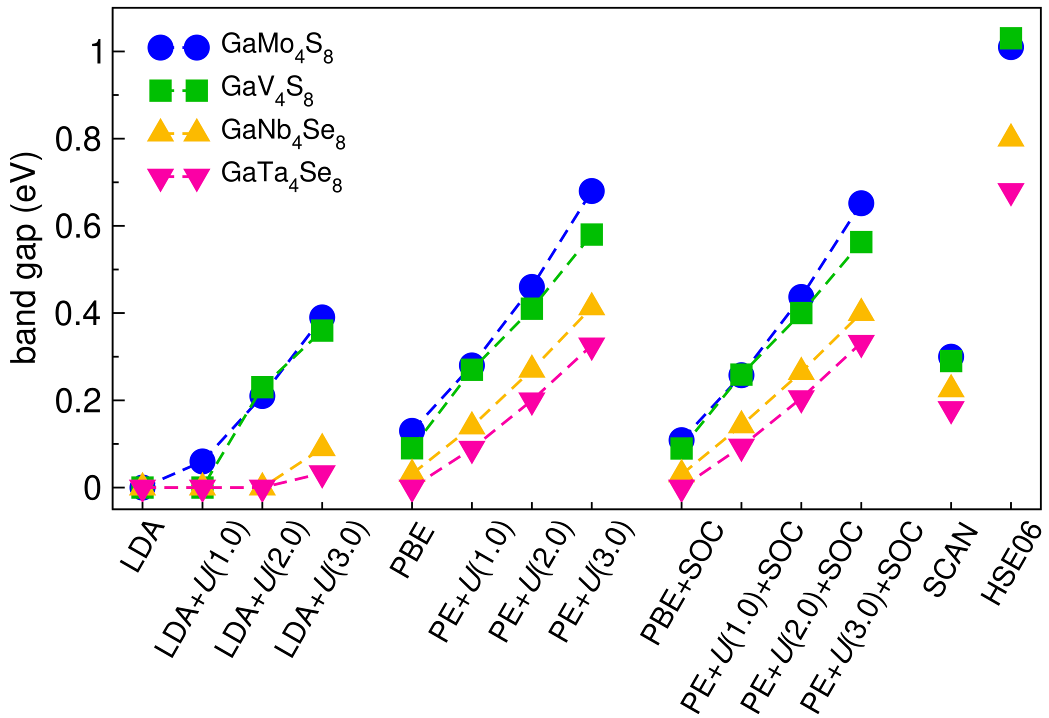


Figure 3.10: DFT-functional dependence of the electronic band gaps for lacunar spinels in the rhombohedral $R3m$ structure.

cupied bands.

We also observe an interesting similarity in Figure 3.8 and Figure 3.10, where the functional dependency of the Wyckoff position $3a$ (z_1) in GaV_4S_8 is similar to the trend in the band gap. A larger structural distortion in GaV_4S_8 also leads to a higher electronic band gap. $GaMo_4S_8$ however, does not exhibit such correlated properties. The distinct behaviors of the Mo and V compounds indicate rather different relationships between the structural distortion and the ground state electronic structures. Niobium and tantalum compounds have relatively smaller band gaps compared with vanadium and molybdenum ones, which is consistent with experimental estimation of band gaps. It is also clear that including SOC has a negligible effect on ground state band gap in all four compounds studied here.

SCAN predicts band gaps close to the experimentally suggested 0.2±0.1 eV value [52], whereas HSE06 finds approximately a 1.0 eV band gap for the vanadium and molybdenum chalcogenides,

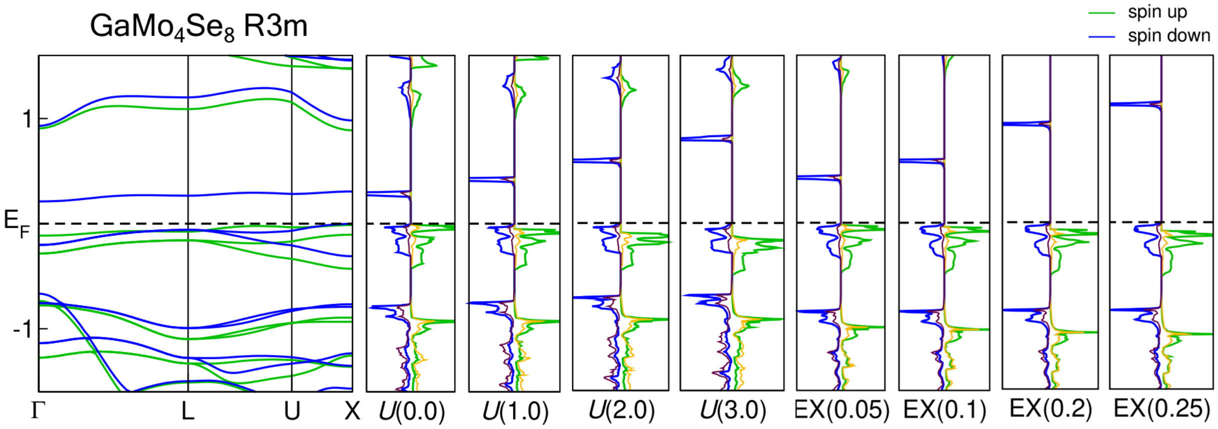


Figure 3.11: Effect of on-site Coulomb interactions (U values of 0.0, 1.0, 2.0, and 3.0 eV) and the amount of exact exchange interactions in the hybrid HSE functional (EX values of 0.05, 0.1, 0.2, and 0.25) on the electronic structures of rhombohedral $R3m$ GaMo$_4$Se$_8$. $\mathrm{EX} = 0.25$ corresponds to the standard amount of exact exchange in HSE06. The band structure panel on the left is obtained using the PBE functional.

and around 0.7 eV for the niobium and tantalum compounds. Since the hybrid functionals partially correct the self-interaction problem in DFT, it is expected to predict more accurate band gaps than lower rung functionals. The larger portion of non-local and range-separated exact exchange interactions included in HSE06, however, might also destroy the balance within the transition-metal cluster, causing the large deviations in the band gaps of the lacunar spinels. It has also been reported that the ferromagnetic ground state is determined by the symmetric exchange interactions [79], which could possibly explain the different behaviors of HSE06 from lower-level functionals. More careful experimental characterization of the distorted phase band gaps is required to have a better understanding of which V_{xc} performs the best.

We next examine the effect of the on-site Coulomb interactions and exact exchange interactions on the electronic structures of rhombohedral GaMo$_4$Se$_8$ (Figure 3.11). The band structure and DOS of GaMo$_4$Se$_8$ using the PBE functional with Hubbard-U values of 0.0, 1.0, 2.0, and 3.0 eV are shown in the first five panels. The four panels starting from the right of Figure 3.11 correspond to the DOS obtained using the HSE06 functional with different portions of exact exchange included as indicated in parenthesis. In general, we observe very similar DOS for the occupied bands.

The three valence bands in the spin-up channel are slightly shifted to lower energy relative to the Fermi level, E_F with either larger U values or larger amounts of exact exchange. The orbitals beneath these valence orbitals, approximately located at $-1\,\mathrm{eV}$, are always lower in energy in our HSE06 calculations. Because the HSE06 functional treats all orbitals on the same footing, these lower energy orbitals are also ‘corrected’ in a self-consistent manner, whereas the on-site Coulomb interaction through the $+U$ correction basically forces integer occupancy among only the correlated orbitals.

In addition, we find an increasing trend in the band gap with larger U values or greater contributions of exact exchange to V_{xc}. We find that $U = 1.0$-$2.0\,\mathrm{eV}$ leads to very similar electronic structures obtained with HSE06 with 5-10 % exact exchange. Our findings here suggest that a GGA+U functional could be used as an alternative method to study electronic structures in lacunar spinels by reproducing the low-energy electronic structure obtained from a hybrid functional but at lower computational cost. The limitation is that lower lying orbitals that may be of interest are not corrected and therefore cannot exactly reproduce the results of the hybrid functional. Based on our simulation results, we do not suggest using HSE06 functional for electronic structure simulations in the lacunar spinel family.

Optical conductivity

We compute the optical conductivity of the ferrimagnetic rhombohedral phase $\mathrm{GaV_4S_8}$ and compare our DFT results with the experimental data [70] in Figure 3.12. The experimental data shows the first optical transition occurs at ≈2,700 cm^{-1} (black symbols), corresponding to an approximate 0.33 eV optical band gap. The optical conductivity then plateaus at approximately 800 $\Omega^{-1}\,\mathrm{cm}^{-1}$ for higher frequencies. Our DFT simulations are able to semi-qualitatively capture the plateau structure, but do not quantitatively reproduce the optical conductivity. With increasing values of U, the plateau shifts to higher frequency and this behavior coincides with a larger optical gap as expected from the aforementioned band gap dependencies on the exchange-correlation functional.

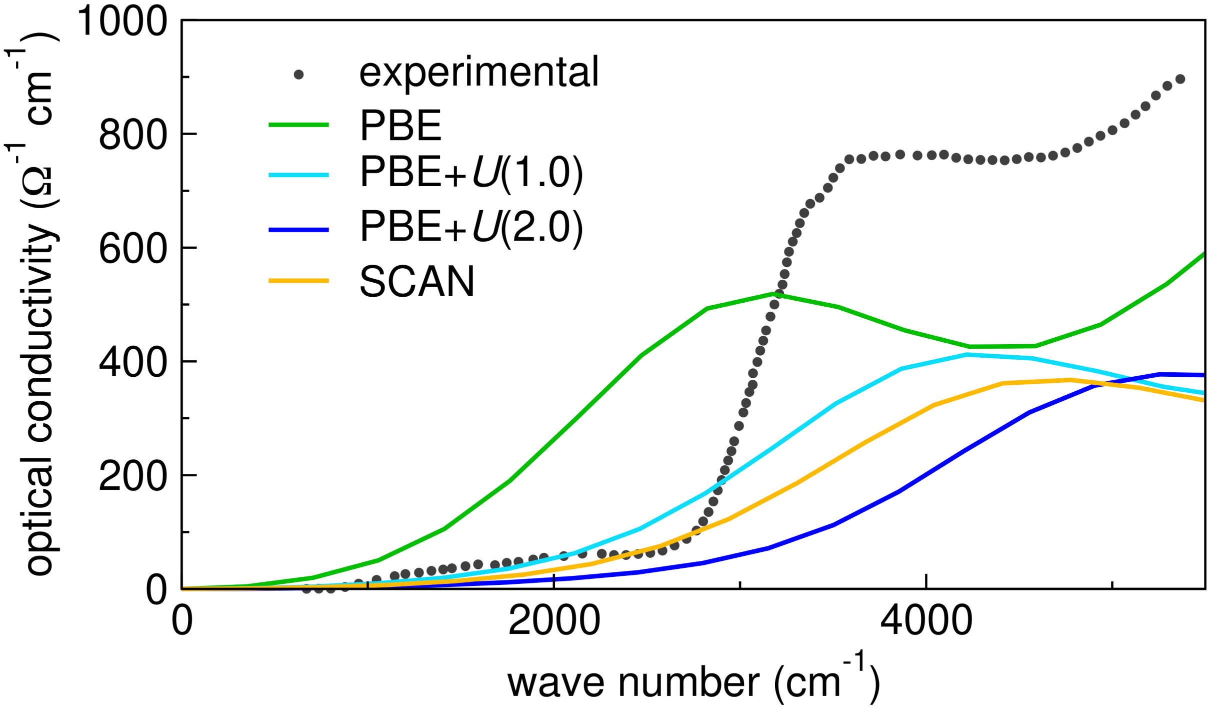


Figure 3.12: DFT calculated optical conductivity of GaV_4S_8 in the rhombohedral $R3m$ structure compared with the experimental values obtained from Ref. [70].

SCAN functional performs similar to PBE with $U = 1.0$-2.0 eV. PBE with $U = 2.0$ eV gives an optical gap closest to the experimental value. We note that because DFT is a single-particle ground state theory, it may not be the optimal tool to study excited state properties, such as optical conductivity. More accurate simulations, for example, could be pursued by solving the Bethe-Salpeter equation using the GW quasiparticle energies [106].

Lattice dynamics

Last, we investigate the exchange-correlation functional dependency on the phonon frequencies in GaV_4S_8. We present our computed normal mode frequencies for both the cubic and rhombohedral phase with different functionals in Figure 3.13: LDA, PBE, PBE+$U = 1.0$ eV, and SCAN. These calculated values are compared with the experimental Raman/IR frequencies at 80 K reported in Ref. [107], which appear in the first column of Figure 3.13. Note that the LDA results for the

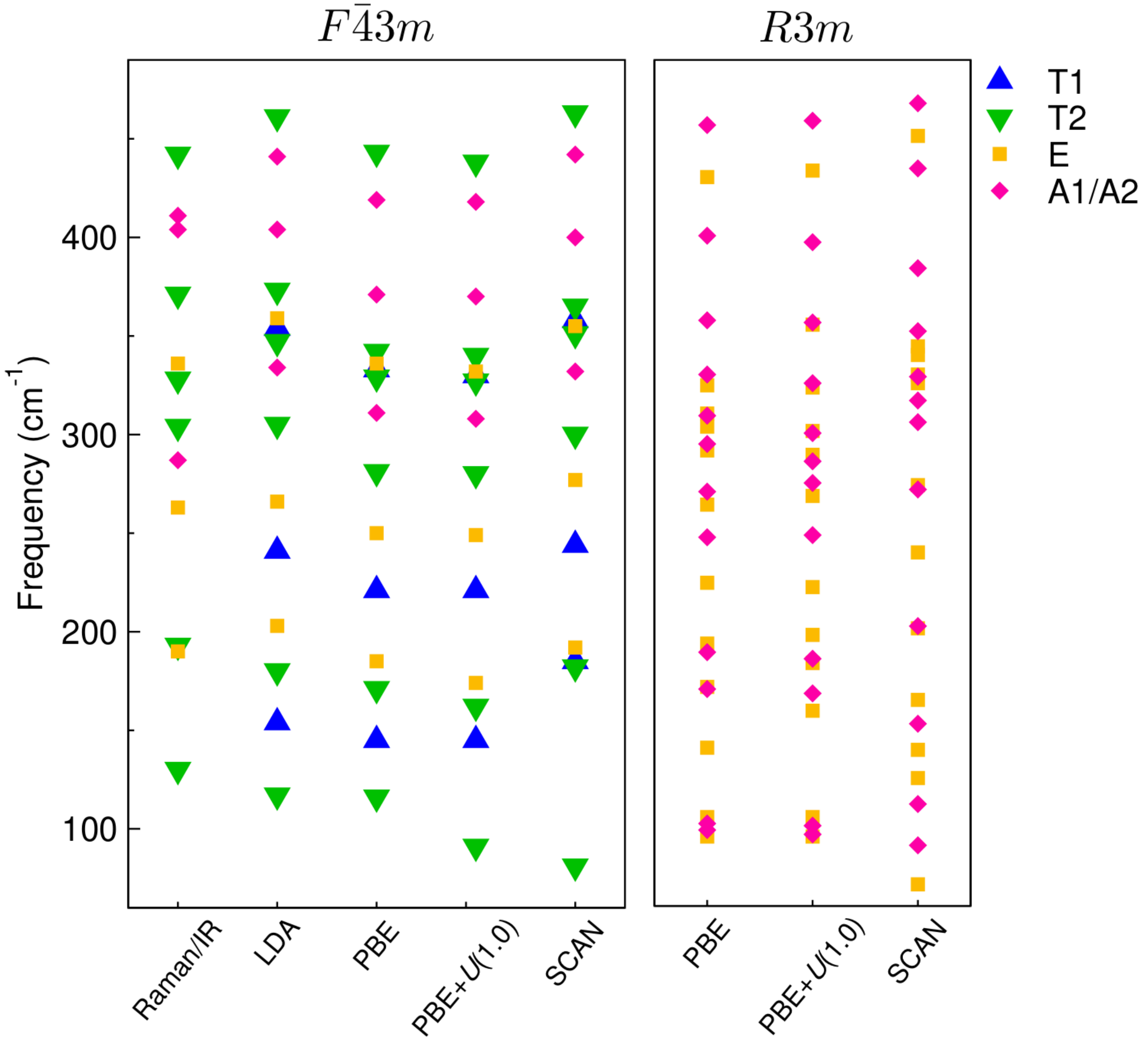


Figure 3.13: Phonon frequencies of GaV_4S_8 in the cubic (left panel) and rhombohedral (right panel) phase with different DFT functionals. The experimental data from Ref. [107] is reproduced in the first column labeled 'Raman/IR'.

rhombohedral phase are not shown, because the $R3m$ structure is unstable at the LDA level.

We find that the cubic phase phonon frequencies generally agree well with the experimental IR/Raman characterization data. LDA and PBE perform reasonably well in reproducing the phonon frequencies. However, with PBE+$U = 1.0\,\mathrm{eV}$, we find a significant decrease in the frequency of the lowest T_2 phonon mode. The same behavior is also obtained with the SCAN functional. Interestingly, this T_2 mode mainly corresponds to the distortion of the transition-metal cluster along the symmetry-breaking pathway. This could be evidence of electron-phonon coupling induced structure instability [64]. Although no Raman/IR data is available for the rhombohedral phase, we still see the same phonon mode softening with functional choice upon going from PBE to SCAN. The major difference between different functionals is at low-frequency, where the vibrational modes are mainly related to the transition-metal (V_4) clusters. Our finding here suggests that lattice dynamics in the lacunar spinels also have non-negligible functional dependency. More experimental (temperature-dependent) data, however, is required to ascertain the functional that best reproduces the lattice dynamical properties.

3.3.3 Exploring the phase space

In this section, we further explore the structural phase space of $GaNb_4Se_8$ and $GaTa_4Se_8$. From the previous analysis, we found that minor structural distortions of the transition-metal clusters could play a deterministic role on their physical properties. However, the interatomic distances (e.g., Ta-Ta distance) within the transition-metal clusters are approximately $3 \sim 4\,\text{Å}$, and only the average microscopic distortions are observed using conventional experimental scattering techniques (e.g., X-ray diffraction). To that end, we utilize a combination of experimental inelastic neutron scattering (INS) measurements and computational simulations to take a closer look at the local geometries. Compared with X-rays, the wavelengths of neutrons fit better with the length scale of interatomic distance, therefore they serve as a better detector to accurately probe the atomic distributions.

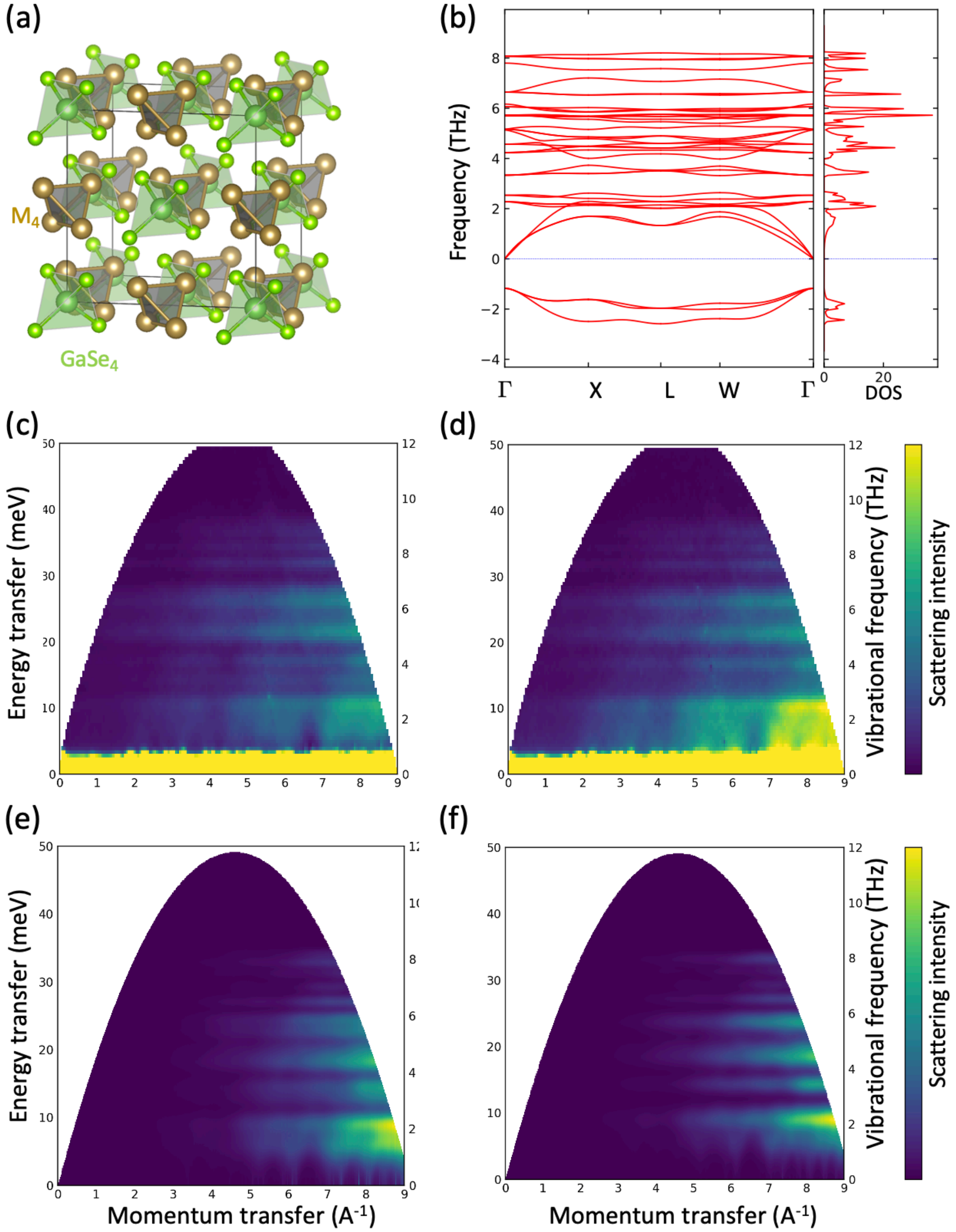


Figure 3.14: (a) The crystal structure of GaM_4Se$_8$ (M = Nb, Ta) with highlighted transition-metal cluster M_4. The ligands (Se anions) of the clusters are not shown here. (b) The GaTa$_4$Se$_8$ $F\bar{4}3m$ phase phonon dispersion and density of states (DOS) using the PBE functional. Phonon modes with imaginary eigenvalues are displayed within the negative frequency region. Experimental inelastic neutron scattering (INS) patterns of GaTa$_4$Se$_8$ collected at (c) 5 K, and (d) 100 K. DFT-simulated INS patterns of GaTa$_4$Se$_8$ with (c) space group $P2_12_12_1$, and (d) space group $F\bar{4}3m$.

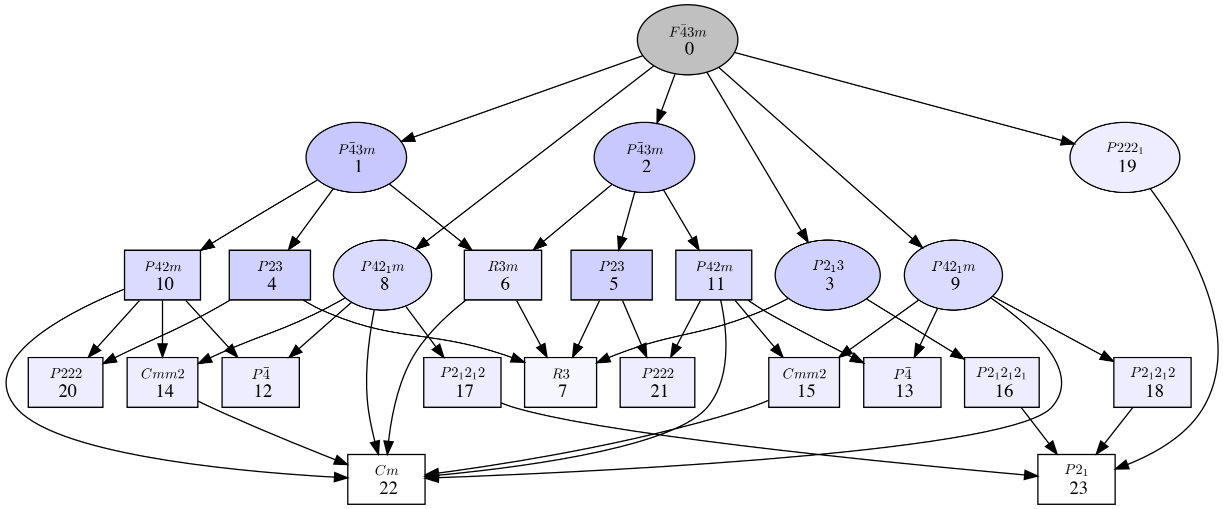


Figure 3.15: Subgroups of the space group $F\bar{4}3m$ accessible through a single irreducible representation order parameter.

From the phonon analysis results discussed in the previous section, we found multiple phonon modes with imaginary eigenvalues (shown in the negative frequency region in Figure 3.14(b)) throughout the entire first Brillouin zone. Since the average ground state crystal structure of $GaNb_4Se_8$ and $GaTa_4Se_8$ is still unclear (unlike GaV_4S_8 or $GaMo_4S_8$ with the $R3m$ ground state), we distort the $F\bar{4}3m$ phase crystal structure along the imaginary phonon modes at selected k-points. For degenerate modes, we also searched through some linear combinations of the modes to sample as many reasonable subgroup structures as possible, derived from the cubic phase. With the distortions initialized from the phonons modes, we use DFT simulations to fully relax the crystal structures. Most of the candidate structures are able to maintain their space group after the relaxation; the results are tabulated in Table 3.6. The subgroups of the space group $F\bar{4}3m$ accessible through a Landau-type transition (single irreducible representation order parameter) are illustrated in Figure 3.15.

Interestingly, both $GaNb_4Se_8$ and $GaTa_4Se_8$ adopt $P2_12_12_1$ symmetry as the ground state structure, which is significantly more stable than the cubic phase. This phase was suggested to be the ground state by our experimental collaborators Julia Zuo and Dr. Stephen Wilson at the University of California, Santa Barbara; however, we were not able to identify this phase from phonon dis-

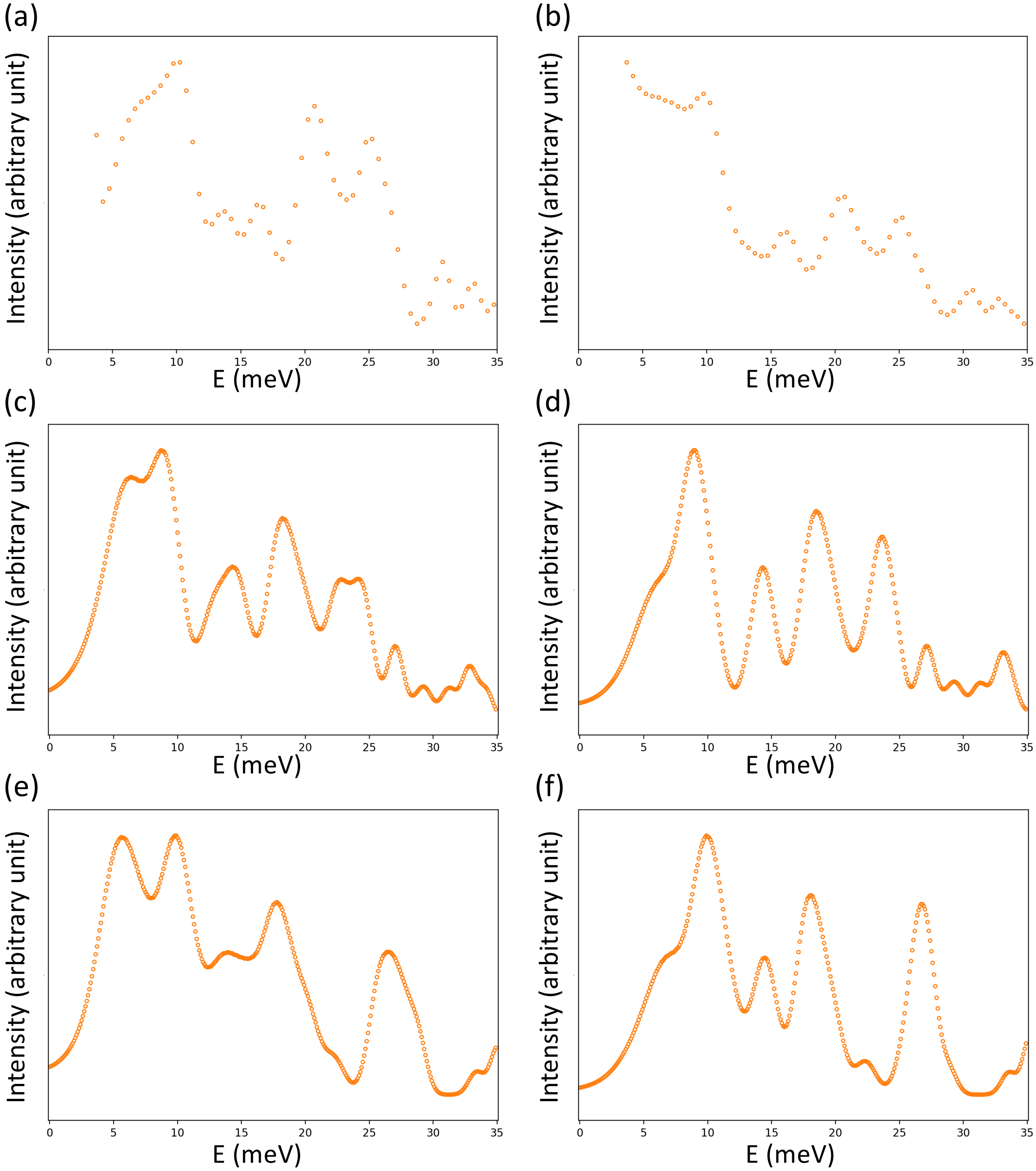


Figure 3.16: Inelastic neutron scattering (INS) patterns collected within $|Q|$ range 7–9 Å^{-1}. GaTa$_4$Se$_8$ experimental INS patterns collected at (a) 5 K, and (b) 100 K. DFT-simulated INS patterns of GaTa$_4$Se$_8$ with (c) space group $P2_12_12_1$, and (d) space group $F\bar{4}3m$. DFT-simulated INS patterns of GaNb$_4$Se$_8$ with (e) space group $P2_12_12_1$, and (f) space group $F\bar{4}3m$.

Table 3.6: A summary of phase space exploration results in $GaNb_4Se_8$ and $GaTa_4Se_8$ using DFT simulations. "Experimental" means this phase is suggested by our experimental collaborators, not identified from distorting soft phonon modes. ΔE is the energy difference between the specified phase and the cubic phase (a negative number means this phase is more stable). In the "Phonon" column, "stable" means there are no imaginary phonons (either acoustic or optical) at 0 K. An "–" means data is not available.

Compound	Space group	$k-$point	ΔE (meV/f.u.)	Phonon
$GaNb_4Se_8$	$F\bar{4}3m$	–	0	unstable
	$P2_12_12_1$	Experimental	-29	stable
	$C222_1$	X (1/2, 0, 1/2)	-21	unstable
	$Pmn2_1$	X (1/2, 0, 1/2)	-20	stable
	$P\bar{4}2_1m$	Experimental	-20	unstable
	$Imm2$	Γ (0, 0, 0)	-12	–
	$P\bar{4}m2$	X (1/2, 0, 1/2)	-9	–
	$R3m$	L (1/2, 1/2, 1/2)	50	–
$GaTa_4Se_8$	$F\bar{4}3m$	–	0	unstable
	$P2_12_12_1$	Experimental	-42	stable
	$P\bar{4}2_1m$	Experimental	-38	stable
	Cm	L (1/2, 1/2, 1/2)	-33	–
	$Pmn2_1$	X (1/2, 0, 1/2)	-31	stable
	Cc	L (1/2, 1/2, 1/2)	-27	–
	$C222_1$	X (1/2, 0, 1/2)	-25	unstable
	$P\bar{4}m2$	X (1/2, 0, 1/2)	-14	–
	$R3m$	Γ (0, 0, 0)	-1	–
	$R3m$	L (1/2, 1/2, 1/2)	49	–

tortions. We later found that $P2_12_12_1$ is a subgroup of $F\bar{4}3m$ associated with the X (1/2, 0, 1/2) reciprocal point, and the suitable phonon should transform as irreducible representation X_5 with order parameter (a,0,b,0,c,0). In order to validate this, we simulate the inelastic neutron scattering patterns of $GaNb_4Se_8$ and $GaTa_4Se_8$ in both $F\bar{4}3m$ and $P2_12_12_1$ phases. The corresponding experimental data is collected at 5 K and 100 K, respectively. The results are shown in Figure 3.14 and Figure 3.16.

In general, the simulated INS agree qualitatively with the experimental data. Since it is well-known that the PBE functional overestimates the lattice parameters, it is reasonable to see that the phonon frequencies are overall red-shifted (i.e., having lower frequencies than the experimental values). The most significant discovery is the peak splitting shown in Figure 3.16[a-d]. The $F\bar{4}3m$ phase on the right has higher symmetry, and we observe one single peak around 15 meV. However, on the left side with $P2_12_12_1$ symmetry, the simulated INS (Figure 3.16(c)) reveals two peaks (the one on the left is slightly merged into the right one). Our current understanding is that the peak splitting could be attributed to the breaking symmetry from the cubic phase to the $P2_12_12_1$ phase, where the interatomic distances of Nb_4 and Ta_4 clusters change after the distortions. The peak splitting is more obvious in the experimental data, which supports the structural phase transitions observed in these two compounds upon cooling.

The INS patterns of $GaNb_4Se_8$ are also shown in Figure 3.16[e, f], we notice that the relative intensities varies a lot from the Ta compound. The origin of these differences is still under active investigation by our team.

3.4 Conclusions

In conclusion, LDA underperforms the other functionals and we recommended to use it only with on-site Coulomb interactions added. The GGA functionals (PBE and PBEsol) perform reasonably well, and the results can be quantitatively improved with on-site Coulomb interactions explicitly added. The meta-GGA functional SCAN is another alternative choice that works well and does not require extra parameterization. Last, the hybrid functional HSE06 predicts accurate lattice structures, but leads to a large electronic band gap in the low-temperature rhombohedral phase. Owing to its high computational cost as well as large deviation in electronic structure predictions, we do not recommend using this hybrid functional for the lacunar spinel family.

All exchange-correlation functionals predict reasonable lattice constants in both the cubic and rhombohedral polymorphs of the chalcogenide lacunar spinels. For electronic structure simulations, the cubic phase is always metallic from band theory and exhibits a narrow transition-metal-derived bandwidth at the Fermi level. Spin-orbit interactions are necessary to predict a semiconducting state in the cubic phase, but not in the rhombohedral phase, at the DFT level. At the LDA and GGA level, on-site Coulomb interactions of 2 eV to 3 eV are recommended to obtain quantitatively improved results. We also found that the PBE functional without on-site Coulomb interactions could predict stable semiconducting states for the rhombohedral phase. Our results obtained with SCAN are similar to PBE+$U(2.0)$ and thus can be safely used in simulations. We also find a highly spin-polarized DFT ground state in $\mathrm{GaV_4S_8}$, which differs from available experimental data, motivating additional investigations of the magnetic order. We found that the single-particle DFT simulations of the optical conductivity do not give a quantitatively satisfying description of $\mathrm{GaV_4S_8}$; more sophisticated methods such as with the GW method may be necessary to treat the excited state properties in the lacunar spinels. The LDA and PBE functional, however, perform well in predicting cubic phase phonon frequencies in $\mathrm{GaV_4S_8}$.

Our INS simulations qualitatively agree with the experimental data, supporting the view that the transition-metal clusters adopt multiple configurations when compete to give the global ground state ($P2_12_12_1$) for $\mathrm{GaNb_4Se_8}$ and $\mathrm{GaTa_4Se_8}$. The cluster geometry could influence the intrinsic electronic structures and magnetic interactions, hence determine the different physical properties observed under different conditions. Our findings here provide a possible explanation to some of the fascinating physical properties observed in the lacunar spinel family (e.g., Skyrmion lattice, resistive-switching, etc.). Since INS is a diffraction-based characterization technique, whose results are ensemble averaged, we suggest using a local probing method (e.g., pair distribution function) to reveal more details about the interatomic distances of the transition-metal clusters.

CHAPTER 4
FEATURELESS ADAPTIVE OPTIMIZATION ACCELERATES FUNCTIONAL ELECTRONIC MATERIALS DESIGN

This chapter is adapted with permission from Ref. [108]. The work was performed and written in collaboration with Akshay Iyer. © Copyright 2020 American Institute of Physics.

Electronic materials exhibiting phase transitions between metastable states (e.g., metal-insulator transition materials with abrupt electrical resistivity transformations) are challenging to decode. For these materials, conventional machine learning methods display limited predictive capability due to data scarcity and the absence of features impeding model training. In this chapter, we demonstrate a discovery strategy based on multi-objective Bayesian optimization to directly circumvent these bottlenecks by utilizing latent-variable Gaussian processes combined with high-fidelity electronic structure calculations for validation in the chalcogenide lacunar spinel family. We directly and simultaneously learn phase stability and band gap tunability from chemical composition alone to efficiently discover all superior compositions on the design Pareto front. Previously unidentified electronic transitions also emerge from our featureless adaptive optimization engine. Our methodology readily generalizes to optimization of multiple properties, enabling co-design of complex multifunctional materials, especially where prior data is sparse.

4.1 Introduction

Upon traversing a critical temperature, the electrical resistivity of a metal-insulator transition (MIT) material can change by orders of magnitude [109]. Athermal approaches may also trigger the electronic transitions, including (chemical) pressure, variable carrier-densities, and applied electromagnetic fields. The transformations can be used to encode, store, and process information for beyond von-Neumann microelectronics and overcome performance limits of conventional field-effect transistors [110] for advanced logic/memory technologies [111]. Because macroscopic MITs occur in materials with diverse chemistries and structures (Figure 4.1(a)), various microscopic mechanisms – electron-lattice interactions, electron-electron interactions, or a combination thereof – lead to large variations in critical temperatures and accessible resistivity changes [112, 113]. This diversity exacerbates the efficient discovery and optimization challenge of achieving multiple property requirements to outperform silicon-based devices [114], including stability, large reversible resistivity changes ($\approx 10^5$), and above room-temperature operation.

The aforementioned complexity is ubiquitous in formulating atomic scale materials chemistry and macroscopic functionality relationships to guide property optimization. Presently, the principal solution relies on a better understanding of the underlying materials physics. Numerous data-driven machine learning models, however, have shown promising results in deciphering nonlinear relationships between materials structure and properties when sufficient training data is available [47, 115, 5, 116, 117]. The predictive performance (error and efficiency) of these approaches is limited by the quality and quantity of the data, typically $> \mathcal{O}(10^2)$, which poses a severe challenge to MIT materials design owing to the relatively small size of available dataset of $\approx \mathcal{O}(10^1)$. The suitability of the machine learning model is determined by the input dimensionality and dataset size, which for high dimensional inputs necessitates large datasets and complex models for good predictive performance.

A number of sequential materials design strategies have recently emerged [118, 119, 120, 121] to rescue the lack of data problem. Mostly being based on the Bayesian approach, these methods utilize knowledge extracted from existing data to infer properties of unknown materials following a step-by-step discovery manner. This sequential optimization method fits well with the regular materials discovery procedure both experimentally and computationally, since property evaluations are usually time and effort consuming (e.g., synthesis and simulations). Nevertheless, these sequential learning models typically rely on numerical materials descriptors (features) whose selection may be informed by domain knowledge or trial-and-error approaches. For MIT materials

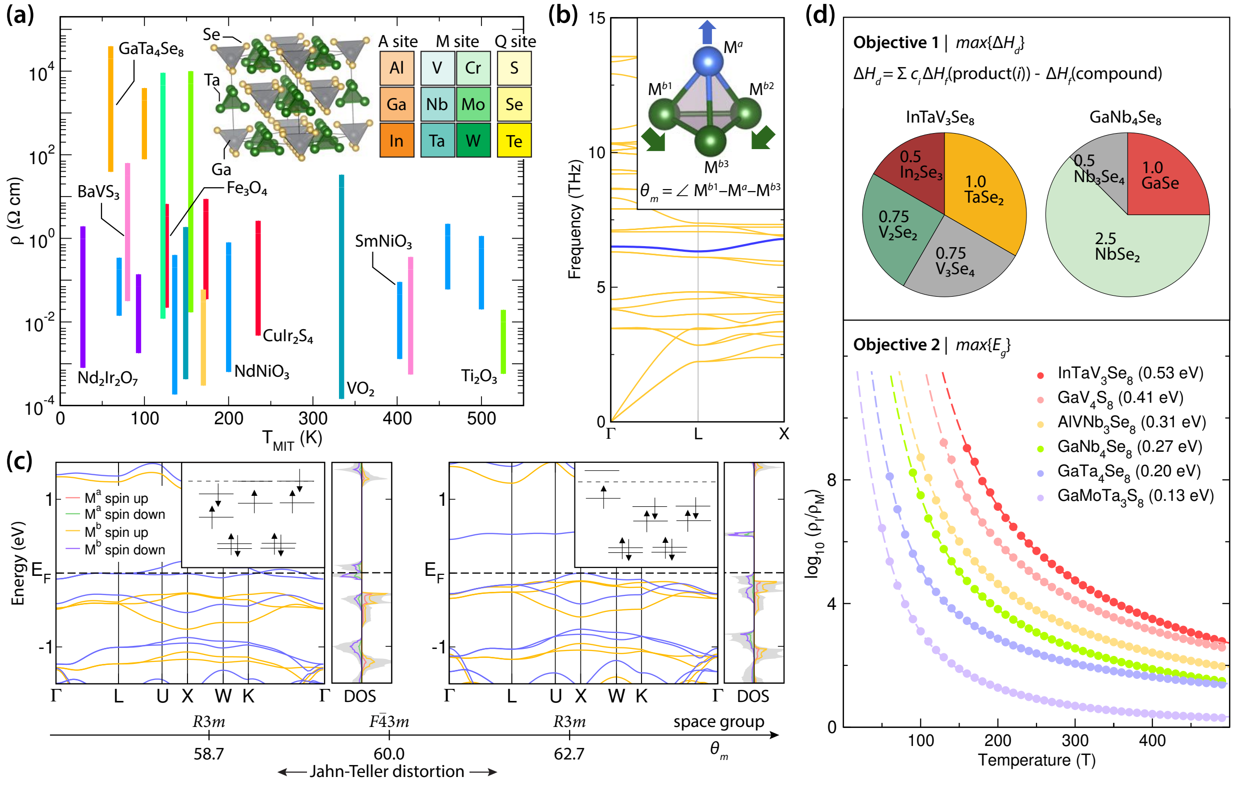


Figure 4.1: Metal-insulator transition materials and design objectives for the lacunar spinel family. (a) The range in resistivity accessible (length of bar) across the MIT and transition temperature for a variety of MIT materials. (left inset) The crystal structure of $GaTa_4Se_8$. (right inset) Candidate elements on each site of the lacunar spinel structure. (b) DFT-simulated phonon dispersion curves of $GaMo_4S_8$ in the rhombohedral ground state, the blue curve corresponds to the Jahn-Teller active cluster distortion mode. (inset) The transition-metal cluster with a single apical M^a atom and three basal M^b atoms. The arrows indicate displacements characterizing the Jahn-Teller active phonon mode. The intra-tetrahedral cluster angle θ_m formed by M^{b1}-M^a-M^{b2}. (c) Electronic band structures and projected density of states (DOS in units of states/eV/spin/f.u.) of $GaMo_4S_8$ in its (right) semiconducting ground state and (left) metallic metastable phase with θ_m. The two $R3m$ phases are connected by the Jahn-Teller-type structural distortion with a $F\bar{4}3m$ intermediate state. (insets) Molecular orbital diagrams of the Mo_4 cluster with different local geometries. (d) Design Objective 1 with the definition of decomposition enthalpy change and the graphical decomposition pathways of two lacunar spinels for demonstration. The DFT-simulated temperature-dependent log ratio of the resistivity in the insulating and metallic phases of lacunar spinels, including experimentally known compounds and newly discovered compositions, serves as design Objective 2. DFT band gaps specified in parentheses.

systems which lack of microscopic understanding in how different compositions influence the phase transitions, this leads to ambiguity in feature formulation for discovery of MIT materials from structure and composition alone rather than through effective Hamiltonians [113].

What could we do when there is little data available while the governing materials physics is not abundantly clear? Here we demonstrate a generic strategy to overcome the data scarcity as well as the feature engineering problems. We utilize multi-objective Bayesian optimization (MOBO) with latent-variable Gaussian processes (LVGP) to simultaneously optimize the band gap tunability and thermal stability in a family of candidate MIT materials – the lacunar spinels (introduced in the next section). With the goal to identify the optimal compositions, among hundreds of possible chemical combinatorics with both high functionality as well as synthesizability, we successfully retrieved all 12 superior compositions on the Pareto front by searching through a small fraction of the total design space. Notably, the chemical compositions (i.e., element on each crystallographic site) are all the model requires to guide this discovery procedure. No handcrafted features are required in this method, hence featureless learning, making our methodology easily generalizable to other materials design problems. We also showcase how this model could offer helpful guidance on making better decisions towards the optimal design—selecting the next candidate compound to synthesize or simulate. Our adaptive optimization engine (AOE) frees researchers from exclusively relying on their chemical intuition, which can require an entire career to accumulate, and is particularly valuable when the research budget is limited.

4.2 Materials Design Objectives

The complex lacunar spinel family $A\mathrm{M}^a\mathrm{M}^b_3Q_8$ with trivalent main group A, transition metal M, and chalcogenide Q ions demonstrate the complexity active in MIT materials design. The structure comprises transition-metal clusters (TMC) with M^a and M^b cations at the apical and basal positions of the tetrahedra (Figure 4.1(b) inset). Although there are hundreds of possible elemental combinations on the four lattice sites in the crystal structure (Figure 4.1(a)), only tens of the lacunar spinels have been experimentally reported [52, 66]. For example, GaV_4S_8 ($M^a = M^b = V$) exhibits a MIT [53], exotic spin textures [54], and multiferroism [97] while $GaVTi_3S_8$ shows negative magnetoresistance and half-metallic ferromagnetism [122]. Most lacunar spinels are narrow-bandwidth semiconductors in their ground states [56, 52]; these electronic properties are governed by distortions of the local TMC from the ideal T_d geometry [61], which manifest as low-frequency phonons as shown for $GaMo_4S_8$ (Figure 4.1(b), blue curve). Jahn-Teller-type distortions, which correspond to elongation along the [111] direction alter the TMC geometry, are particularly important; they transform the insulating $GaMo_4S_8$ ground state into a metastable metallic phase (Figure 4.1(c)). The MIT arises from a redistribution of electrons among the structure-driven orbital hierarchy (Figure 4.1(c) insets). We can further use *ab initio* molecular dynamics (AIMD) simulations to determine the MIT temperature and validate the phase transition. Furthermore, these phases host low energy electronic structures, discernible from the projected density of states (pDOS) in Figure 4.1(c), that arise from the different M^a and M^b sites. This capability to exhibit distinct and tunable electronic phases poses a challenge in the design of lacunar spinels from physics-based models while also making them an ideal system for MIT performance optimization.

In pursuit of novel MIT materials with superior performance, we specifically seek lacunar spinels that exhibit high thermodynamic stabilities and large resistivity-switching ratios, which we formulate as two design objectives for our materials discovery task. We reduce the approximately $\mathcal{O}(10^3)$ compositional space to 270 candidates that maintain a $1\,M^a$ to $3\,M^b$ ratio. ($AM^a_2M^b_2Q_8$ compositions are excluded as they remove the C_{3v} symmetry fundamental to the MIT; Cr is also excluded from occupying the M^b site, because it destabilizes the cluster [123].) In addition, although there have been several attempts to make mixed-anion lacunar spinels [124, 99], we only focus on the transition-metal clusters in this work since they dominate the valence bands. This design space extends the known composition space that have been experimentally synthesized; therefore, it is important to determine the crystal stability, i.e., whether the selected chemical combination forms a thermodynamically stable lacunar spinel structure. To that end, we define the first design objective as the decomposition enthalpy change (ΔH_d, Figure 4.1(d)), and use density functional theory (DFT) simulations to evaluate formation energies (see Appendix Section A.1). Materials with larger ΔH_d are expected to be more synthesizable [125] and stable during operation, making it a useful filter to prioritize compounds for subsequent theoretical analysis and synthetic processing. The second design objective is the ground state band gap (E_g). We use it as a proxy for the resistivity-switching ratio since E_g is positively correlated with the resistivity change between different electronic states (Figure 4.1(d)). A larger E_g also allows for greater band-gap tunability through control over the C_{3v} distortion, which is a desirable feature for programmable electronics. Importantly, because E_g is small for most MIT materials, stability is expected to be lower and more difficult to achieve than that of nonpolymorphous compounds with majority ionic or covalent bonding [126].

4.3 Adaptive Optimization Engine (AOE)

The nonlinear responses of both design objectives bring severe challenges to compound optimization beyond those amplified by chemical combinatorics using data-driven models. We overcome these obstacles by implementing a cyclic adaptive optimization engine shown in Figure 4.2, which consists of four iterative tasks (*vide infra*): property evaluation, aggregation of data (in a repository), featureless learning, and composition optimization. Beyond returning a predictive model capable of predicting properties from compositions alone, our iterative AOE leverages earlier approaches [118, 119, 120] to deliver materials with superior performance by design of composition-based solutions. In contrast to single objective design which often has a unique solution, multiobjective design aims to uncover the Pareto front—a set of non-dominated designs where no individual objective can be improved without deterioration in other objectives. In other words, the Pareto front represents the optimal trade-offs that can be achieved amongst competing objectives. There is no relative importance of multiple objectives in the process of identifying the Pareto front, which simply offers the designer several options from which to select the subset of compositions

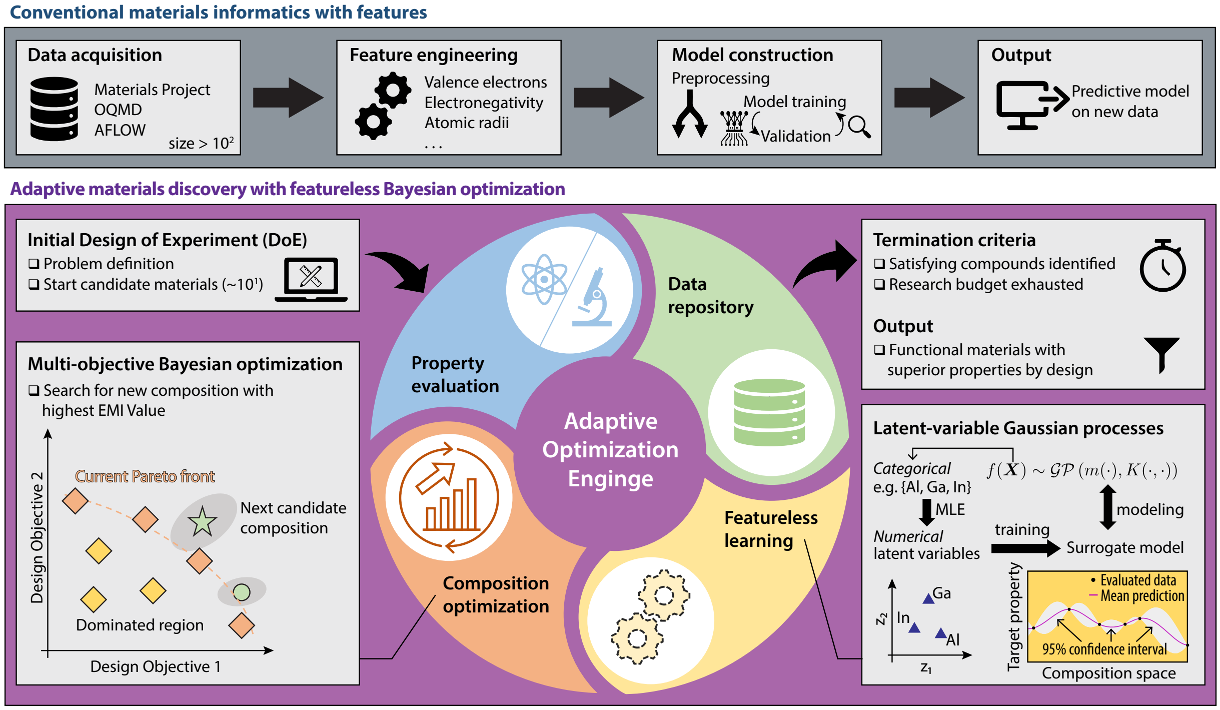


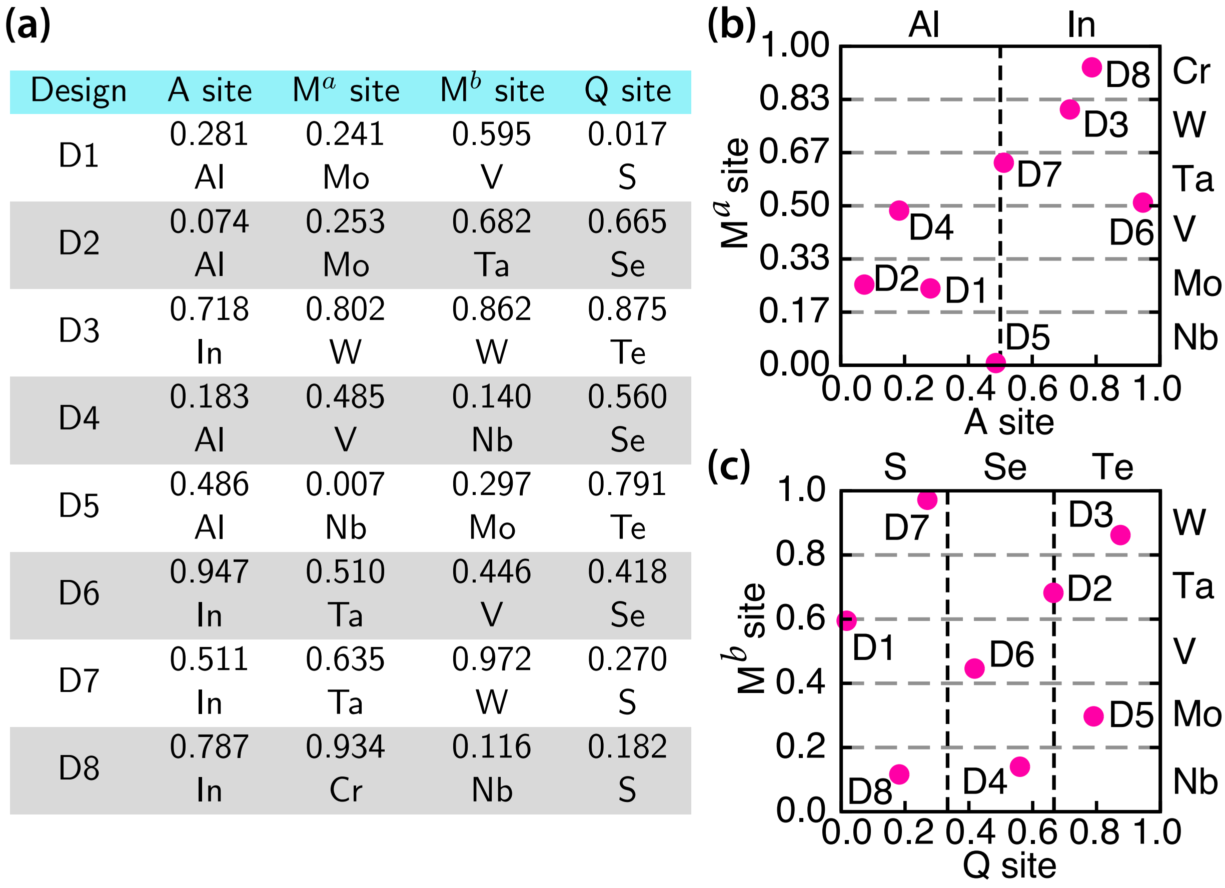
Figure 4.2: Comparison of conventional (feature-required) machine learning with the featureless adaptive optimization engine. Upper panel: the workflow of a conventional feature-based machine learning model typically involves data acquisition, feature engineering, model construction, and property prediction. Lower panel: the adaptive materials discovery scheme starts from an initial set of design of experiments (DoE), where system variables, design objectives, and design space are first defined for the problem, providing a few $\mathcal{O}(10^1)$ candidate materials to initialize the discovery procedure. Property evaluation: the target material properties (design objectives) are evaluated either by experimental measurement or theoretical simulations. Candidate composition and its evaluated properties are then added to a data repository, which initially may either be empty or only contains entries for existing materials within the design space. Its size grows as more candidate materials are evaluated during the adaptive optimization process. Featureless learning involves directly learning from the chemical composition of materials comprising the data repository by mapping each compositional variable into a two-dimensional latent space (spanned by z_1 and z_2) using maximum likelihood estimation, which enables the construction of a latent-variable Gaussian process (LVGP) surrogate model. One surrogate model is constructed for each design objective using all currently available knowledge within the data repository. Composition optimization through multi-objective Bayesian optimization is then performed with the LVGP models to obtain the next candidate material composition with the highest expected maximin improvement (EMI) value. The model accounts for uncertainty with the 95% confidence interval shown as the shadowed area around the new compositions (the green symbols). In the lower left inset, the green star composition outperforms the green circle composition, and will be passed to the next property evaluation procedure. The iterative optimization step continues until all compounds satisfying the objectives are discovered, forming the Pareto front, or computational resources expire.

for further investigation and development. Since the designer's preference may be subjective or informed by other criteria, herein we present only the framework for Pareto front discovery and its comprising compositions.

The AOE has the important advantage of bypassing the feature engineering procedure as in conventional ML methods; it learns properties directly from the chemical composition at each site (i.e., A, M^a, M^b, Q). Gaussian Process (GP) is ideally suited for this problem, because (a) it interpolates data and hence is ideal for surrogating deterministic responses such as DFT results, and (b) it provides a principled statistical representation for uncertainty quantification, which is essential for Bayesian optimization. Latent-variable methods provide a fundamentally different approach to modelling categorical design variables by alleviating the need for handcrafted features (see Appendix Section A.2). It transforms categorical variables (i.e., elemental compositions) into a continuous numerical space. Utilizing these approaches in the AOE, we achieve featureless learning and then perform composition optimization under the multiple objectives through latent-variable Gaussian processes (LVGP).

We start the MIT-materials AOE for the lacunar spinel family through an initial design of experiment (DoE) consisting of four experimentally known compounds within the family (i.e., $\mathrm{GaMo_4S_8}$, $\mathrm{GaV_4S_8}$, $\mathrm{GaNb_4Se_8}$, and $\mathrm{GaTa_4Se_8}$) and eight new compositions generated by discretized Latin Hypercube Design (LHD) [127] (Figure 4.3). This procedure ensures a variety of elemental combinations within the initial DoE set, where each candidate element will appear at least once, so that the model has knowledge about different elemental contributions to the design objectives.

Next, we use high-fidelity DFT simulations to evaluate ΔH_d and E_g (see Appendix Section A.1). This is the most resource-intensive step among the four tasks; therefore, it is desirable to iterate through the AOE (property evaluation) step as few times as possible. Although it is application dependent, AOE can be terminated if a compound with target properties is discovered or the budget (computational/experimental) has been exhausted. Then, we create a data repository that



(a)

Design	A site	M^a site	M^b site	Q site
D1	0.281 Al	0.241 Mo	0.595 V	0.017 S
D2	0.074 Al	0.253 Mo	0.682 Ta	0.665 Se
D3	0.718 In	0.802 W	0.862 W	0.875 Te
D4	0.183 Al	0.485 V	0.140 Nb	0.560 Se
D5	0.486 Al	0.007 Nb	0.297 Mo	0.791 Te
D6	0.947 In	0.510 Ta	0.446 V	0.418 Se
D7	0.511 In	0.635 Ta	0.972 W	0.270 S
D8	0.787 In	0.934 Cr	0.116 Nb	0.182 S

Figure 4.3: Design of experiment (DoE) for the complex lacunar spinel family. (a) A four-dimensional Latin Hypercube Design of size eight is generated, where each dimension corresponds to a crystal site (e.g., A, M^a, etc.). Since the four known compounds are all gallium-based, we only consider Al and In for the A site design. (b, c) Each dimension is evenly divided into a number of grids, each grid represents one candidate elemental composition at that crystal site. For instance, the Q site is divided into three grids because there are three candidate elements (S, Se, Te) on that site. The designed composition could then be determined using the grid-composition correspondence. For example, Design ID Number 1 (D1) resides in the grid corresponding to {Al, Mo, V, S}; therefore, its composition is $AlMoV_3S_8$.

contains entries for both composition and the evaluated properties. Unlike other ML methods, we do not rely on a large number of existing data at either the onset or later in the learning process.

We then construct a LVGP model by mapping the elemental compositions (e.g., Al, Ga, In) into a two-dimensional (2D) latent space (Figure 4.2, lower right inset) where the relative positions of elements are obtained using maximum likelihood estimation (MLE). This latent space representation enables us to construct Gaussian process surrogate models for the unknown underlying design objectives, ΔH_d and E_g, as a function of composition. The MOBO step then begins and we use the LVGP models to predict ΔH_d and E_g of the *unexplored* compositions in our design space; we choose the next candidate composition for evaluation using the expected maximin improvement (EMI, see Appendix Section A.2) as the acquisition function, which quantitatively describes the performance gain compared against the compositions at the current Pareto front. The EMI is defined in such a way that both objectives have equal weighting, and the objective properties are normalized with respect to the current min-max values (see Appendix Section A.2 for details). This acquisition function considers both exploration of compositions with high uncertainty (Figure 4.2, shaded ellipses, lower left inset) as well as exploitation of candidates with high performance gain. The composition with highest EMI is then selected for DFT simulation (property evaluation), at which point another AOE cycle commences.

The aforementioned iterative optimization procedure progresses and explores the available design space. One new lacunar spinel composition is evaluated and added to repository after each AOE iteration. The LVGP models are also updated in each iteration as more knowledge becomes available. Owing to the high computational cost of the property evaluation process, we terminate the optimization process after searching through 1/3 of the entire design space. In order to validate the effectiveness of this method, we ultimately evaluated ΔH_d and E_g with DFT calculations of all 270 compositions within the design space by expending approximately 3×10^6 CPU hours.

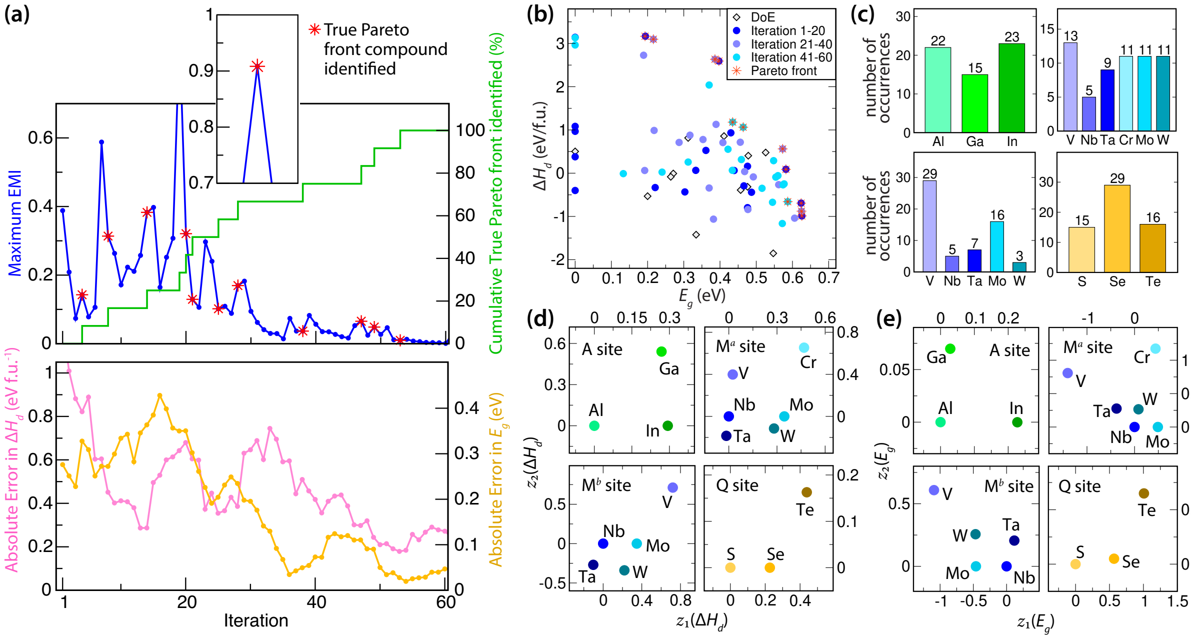


Figure 4.4: The results of adaptive optimization on the lacunar spinel family. (a), Upper panel: Evolution of the highest expected maximin improvement (EMI, blue line) and percentage of true Pareto front compounds identified (green line) as a function of iteration number. Results of the first 60 iterations are shown here. The red asterisks represent sampling points where a true Pareto front design is successfully identified. Lower panel: The moving average of absolute error in the predicted E_g and ΔH_d values for a compound selected by the acquisition function for property evaluation. (b), The distribution of initial design of experiment and the first 60 evaluated compounds. Compounds evaluated in earlier stages have darker colors. True Pareto front designs are marked with red stars. (c), Distribution of Bayesian optimization-sampled elemental compositions for the first 60 iterations. (d, e), Latent space representation of elemental composition at different crystal structure sites in the ΔH_d and E_g surrogate model, respectively. Results obtained after 60 iterations.

4.3.1 AOE performance

Figure 4.4(a) displays the results of the AOE. We successfully identify all 12 materials at the true Pareto front within 53 iterations (red asterisks, upper panel)—compositions and objective-related properties are enumerated in Table 4.1. Combined with the 12 compounds from our initial DoE, we explored less than 25% of the entire design space before identifying all lacunar spinels on the Pareto front. Interestingly, Pareto-front compositions are mostly found with high EMI values, showing that our model makes beneficial recommendations on which composition to evaluate next. High prediction uncertainty likely explains why a Pareto-front composition is not identified for some iterations with a large EMI. The EMI values reduce to nearly zero after all Pareto front compositions are identified (blue, upper panel) since all candidates not sampled are dominated by the Pareto front compounds. We also show the absolute error in the LVGP-predicted ΔH_d (pink) and E_g (orange) values of the evaluated composition at each iteration to further demonstrate the effectiveness of our model (Figure 4.4(a)). We find a general decreasing trend in error and therefore better model predictability as it becomes aware of more composition-property knowledge.

Figure 4.4(b) shows the history of composition explored by the AOE for the first 60 iterations. The initial DoE sets are relatively scarcely distributed away from the true Pareto front (marked as red asterisks), yet the model explores regions far from that covered by the DoE sets and is able to identify 75% of Pareto front compositions within the first 40 iterations. First, we begin to understand this performance by examining the distribution of elements sampled by the MOBO (Figure 4.4(c)). Our model does not exhibit much compositional bias upon sampling elements for the A site; however, it shows clear preferences for choosing certain elements on other sites. V and Mo are sampled more frequently on the basal M^b site, while Nb and Ta are less favored on the apical M^a site. Se is also preferred over S and Te for the Q site.

Then we examine the 2D latent space representations for both design objectives obtained after 60 iterations of AOE (Figure 4.4(d) and (e)). The relative positioning of elements in the latent space reflects correlations in their influence on properties; elements in close proximity exhibit

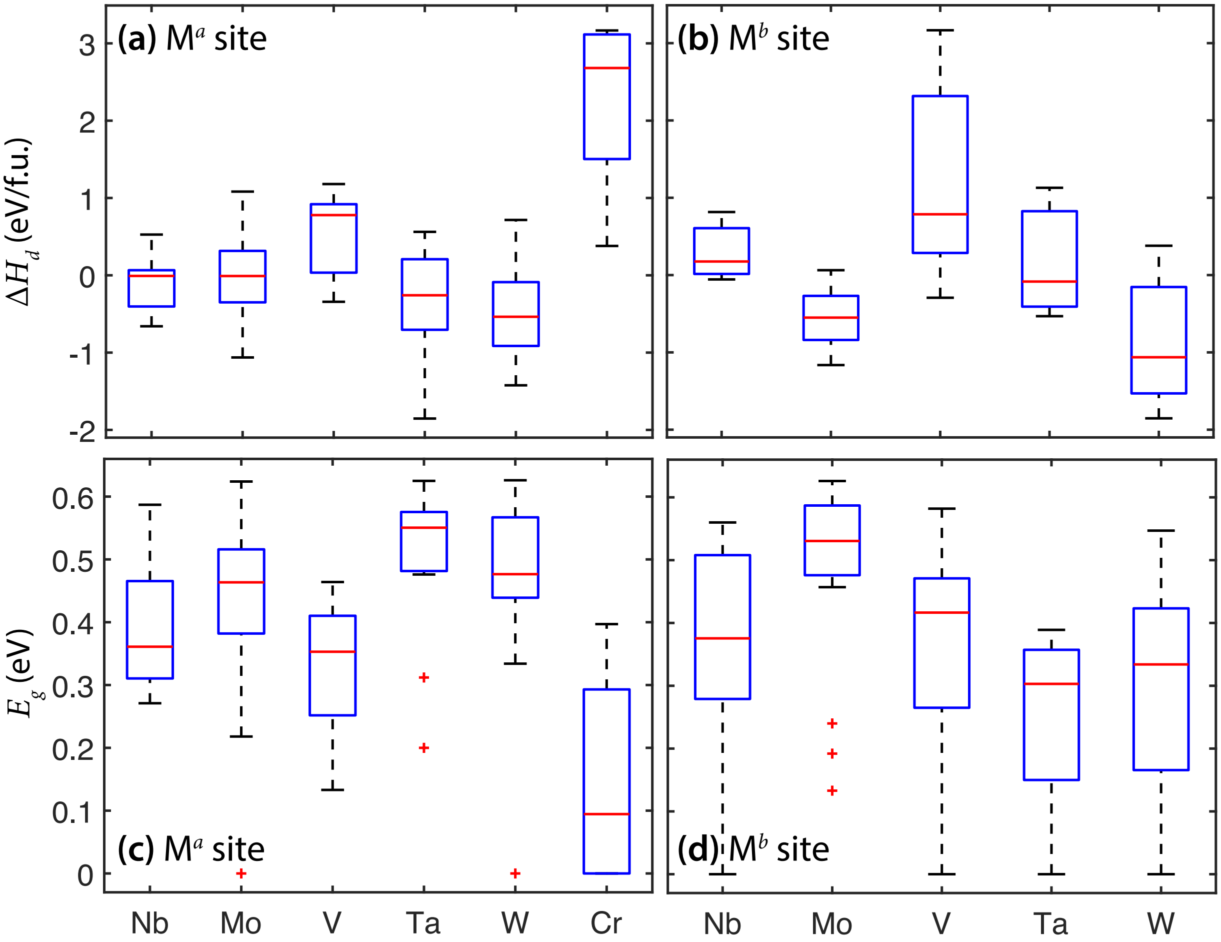


Figure 4.5: Composition-property relationships at the transition-metal sites. Distribution of DFT-evaluated properties of the complex lacunar spinel family with 12 initial DoE sets and 60 iterations of AOE. This data presents the impact different elemental compositions at the transition-metal sites (i.e., M^a and M^b) have on the two design objectives (i.e., ΔH_d and E_g). (a, b) decomposition enthalpy change distribution at M^a, M^b site. (c, d) band gap distribution at M^a, M^b site.

similar impact. Interestingly, different transition metals exhibit distinct correlation patterns across various sites and objective properties. This variation leads us to conclude that (i) the transition metals contribute to stability and band gap in different and unexpected ways, and (ii) the lack of any resemblance in element positioning in the site-dependent latent spaces, except for the M^a site, to the periodic table indicates that chemical-intuition-based MIT design within the lacunar spinels is highly nontrivial. For example, chromium is located far from the other elements in the M^a latent space, indicating that its influence on properties is distinct. Indeed, Cr-containing compounds have significantly lower E_g and higher ΔH_d (Figure 4.5).

The aforementioned performance is robust as revealed by our multi-trial results (Figure 4.6(a)), where we find the AOE successfully identifies 90 % of the true Pareto-front compositions by exploring 30 % of the design space with different initial DoE sets. Since LHD is inherently random, repeating the DoE procedure will lead to another randomly generated DoE set. Therefore, we use this method to run multiple trials of AOE with different DoE sets. The size of DoE is another parameter for the designer to select in the AOE framework. Since the computational budget is often the bottleneck in discovery, the designer must allocate it wisely between the DoE and AOE. We investigated this problem using a set of four DoE sizes: 6, 12, 18, and 24, because there are six elements admissible at the M^a site (Figure 4.6(b)). In each case, the computational budget is fixed to 40 and 60 simulations and they are split between DoE size and AOE iterations. For example, 40 simulations can be split into DoE of size 6 and 34 iterations of AOE whereas a DoE of size 12 corresponds to 28 iterations of AOE, etc. Here, the four known gallium based compounds were not explicitly included in the DoE. We find that using a small DoE to initialize AOE (conversely, allocating more simulations to the AOE) is advisable, as its uncertainty guided exploration is more likely to discover Pareto compositions (Figure 4.6(b)).

Single-objective Bayesian optimization on both band gap (E_g) and stability (ΔH_d) are also performed using Expected Improvement acquisition criterion [128], as shown in Figure 4.6(b, c), respectively. Unsurprisingly, the model shows much higher efficiency in identifying the optimal

Table 4.1: DFT-evaluated ground state properties of the Pareto front compounds. NOI is the number of iterations taken to discover the compound during the adaptive optimization process. Values of $\Delta H_d > 0$ (units of eV f.u.$^{-1}$) indicate an endothermic reaction occurs and the stable compound disfavors decomposition. E_g is the DFT band gap in eV. ν_{JT} is the frequency (THz) of the Jahn-Teller-type phonon involving the TMC. P is the electric polarization in μC cm^{-2}. The value of θ_m in the insulating ground state and transition type, Type I (MIT) or Type II (SIT), are also specified.

Compound	NOI	ΔH_d	E_g	ν_{JT}	P	θ_m	Type
$InWV_3S_8$	4	0.09	0.58	5.83	0.56	65.0	II
$AlCrV_3Se_8$	8	3.17	0.19	3.77	1.87	56.4	II
$InMo_4Se_8$	14	-0.69	0.62	4.55	1.08	63.4	I
$InWMo_3Se_8$	19	-0.99	0.63	4.43	0.24	63.8	I
$InCrV_3S_8$	20	2.59	0.40	4.75	0.28	56.6	II
$AlCrV_3S_8$	21	2.63	0.39	5.81	1.02	57.0	II
$InCrV_3Se_8$	25	3.10	0.22	3.45	0.58	56.0	II
$InTaMo_3Se_8$	28	-0.88	0.62	4.25	1.38	54.8	II
$AlTaV_3Se_8$	38	0.56	0.56	3.90	0.15	57.3	I
AlV_4Se_8	47	1.06	0.46	4.08	2.80	54.9	I
$InNbMo_3Se_8$	49	-0.66	0.59	4.44	0.75	55.2	II
GaV_4Se_8	53	1.18	0.44	4.09	2.37	55.0	I

composition than in the multi-objective task, where less than 10 % of the entire design space is explored. We also notice that the model is always able to quickly infer the compound with highest stability, as depicted by the steep curve in Figure 4.6(c). Intuitively, thermodynamic stability is straightforward to linearize from elemental reference states whereas the band gap is determined by the valence electronic structure and multiple interactions. Therefore, it might be easier for the model to decode the relationship between composition and stability, while learning the band gap dependency requires accumulating more knowledge.

4.3.2 Pareto compound analysis

We use DFT simulations to examine the properties of the identified Pareto-front compositions, focusing on ΔH_d, E_g, and the Jahn-Teller active phonon ν_{JT} involved in the MIT (Table 4.1). We find most Pareto-front compositions consist of two different cations on the M^a and M^b site, only three have $\mathrm{M}^a = \mathrm{M}^b$, with 75 % of the optimized materials being selenides. GaV_4Se_8 is the only Pareto

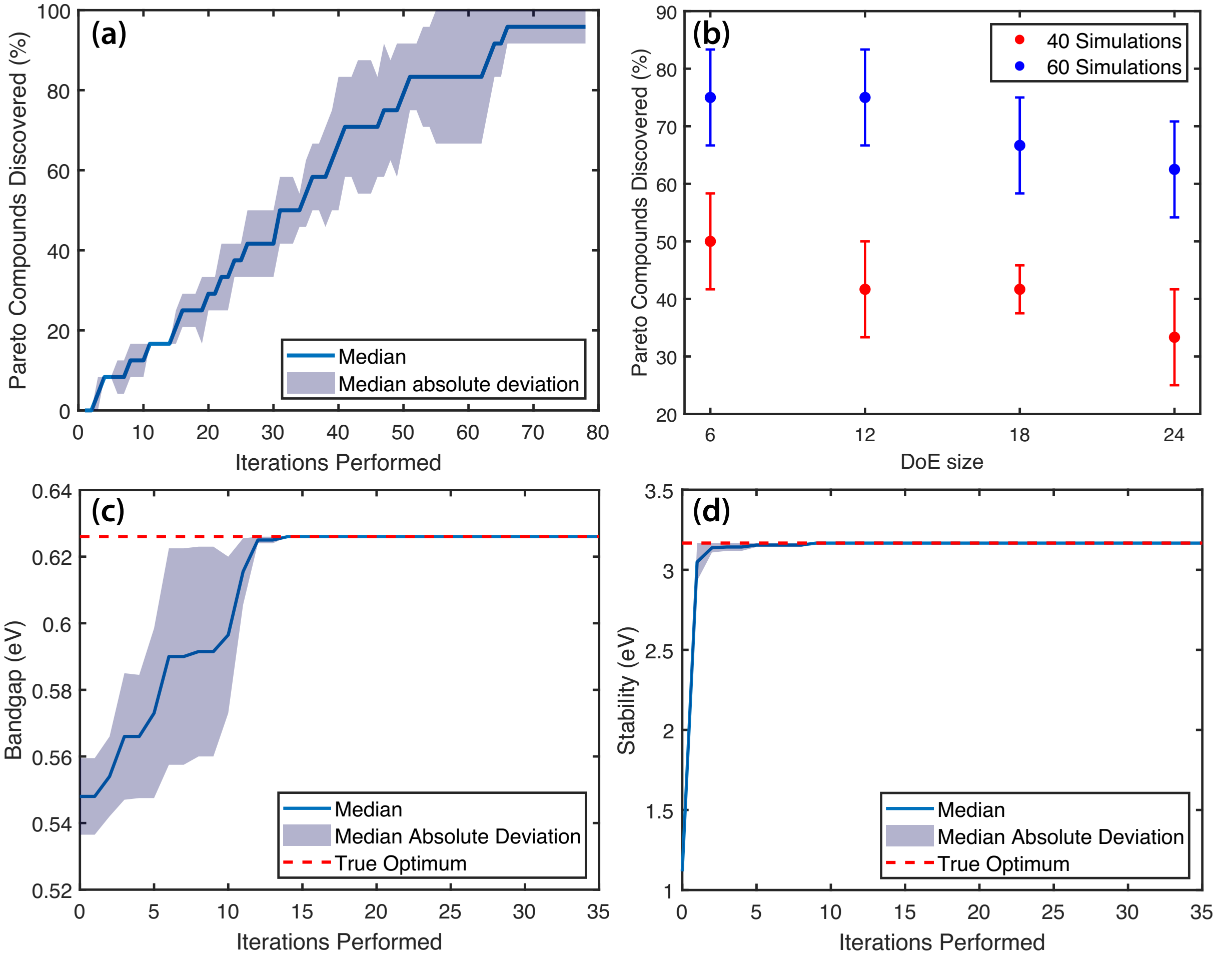


Figure 4.6: Robustness of the Adaptive Optimization Engine (AOE). (a) The optimization history for 10 replicates of AOE, each initialized with a distinct set of 12 initial DoE compounds. In each trial, the initial DoE set consists of the same four known lacunar spinel compounds and eight new compositions designed by the DoE procedure. Solid line shows the median percentage of true Pareto front compounds discovered at each iteration. The shaded area represents the median absolute deviation across 10 trials. (b) The fraction of Pareto front compounds discovered when the computational budget is fixed to 40 and 60 simulations. Filled circles and their corresponding error bars represent the median and median absolute deviation respectively. (c,d) The optimization history of 10 replicates of single-objective Bayesian optimization, targeting maximum band gap (E_g) and stability (ΔH_d^*), respectively. The initialization method is the same as described in (a). Global optimum ($E_g^* = 0.626\,\text{eV}$, $\Delta H_d^* = 3.167\,\text{eV}$) is identified within $10\,\%$ exploration of design space.

front compound previously synthesized, and verified to exhibit resistive-switching behavior under an applied electric pulse [18]. All compounds exhibit $R3m$ symmetry and are dynamically stable in their ground state ($\nu_{\text{JT}} > 0$). The phonon frequencies of the selenides, including ν_{JT} are lower than those of the sulfides. All of the designed lacunar spinels also exhibit semiconducting gaps with semilocal exchange-correlation and static Coulomb interactions and exhibit nonzero electric polarizations. Compositions with larger band gaps tend to have lower stability as determined by ΔH_d: 2/3 are stable ($\Delta H_d > 0$, indicating decomposition is endothermic), whereas four of the 12 compounds comprising Mo have small values of $\Delta H_d < 0$, which could nonetheless be stable and synthesizable [125, 129]. Typically, highly ionic materials with large electronic band gaps are also quite stable (e.g., NaCl). However, we find a clear trade-off between these two properties for the Pareto front compositions. One possible reason is because all of these candidate materials are small-gap semiconductors (with $E_g < 0.65\,\text{eV}$) due to metal-metal and semiconvalent bonding while also being polymorphous; therefore, these lacunar spinels are unlikely to follow the general trend. In addition, Figure 4.5 shows that the transition metals contribute to E_g and ΔH_d in quite different ways, which could lead to this functionality-stability trade-off. The AOE, however, does not posses knowledge of chemistry beyond the lacunar spinel family; yet, it is able to resolve the ΔH_d-E_g relationship regardless of whether there is a trade-off or positive correlation. These findings reinforce the effectiveness of this model.

Although the ground states of these materials are all semiconducting, we find two different electronic transitions upon traversing the ideal TMC geometry ($\theta_m = 60°$): the expected (Type I) metal-to-insulator transition and an unexpected (Type II) semiconductor-to-insulator transition (SIT). Figure 4.7(a) shows the changes to the electronic structure for the MIT lacunar spinels AlTaV$_3$Se$_8$ and InWMo$_3$Se$_8$ with the insulating state (lower panel) always lower in energy than the metastable metallic phase (upper panel) after the Jahn-Teller-type distortion ($\theta_m \neq 60°$, Table 4.1). The pDOS of these compounds show that the metallic state in the Type I transition arises from cluster distortion-triggered orbital ordering and occupancy changes, similar to the mecha-

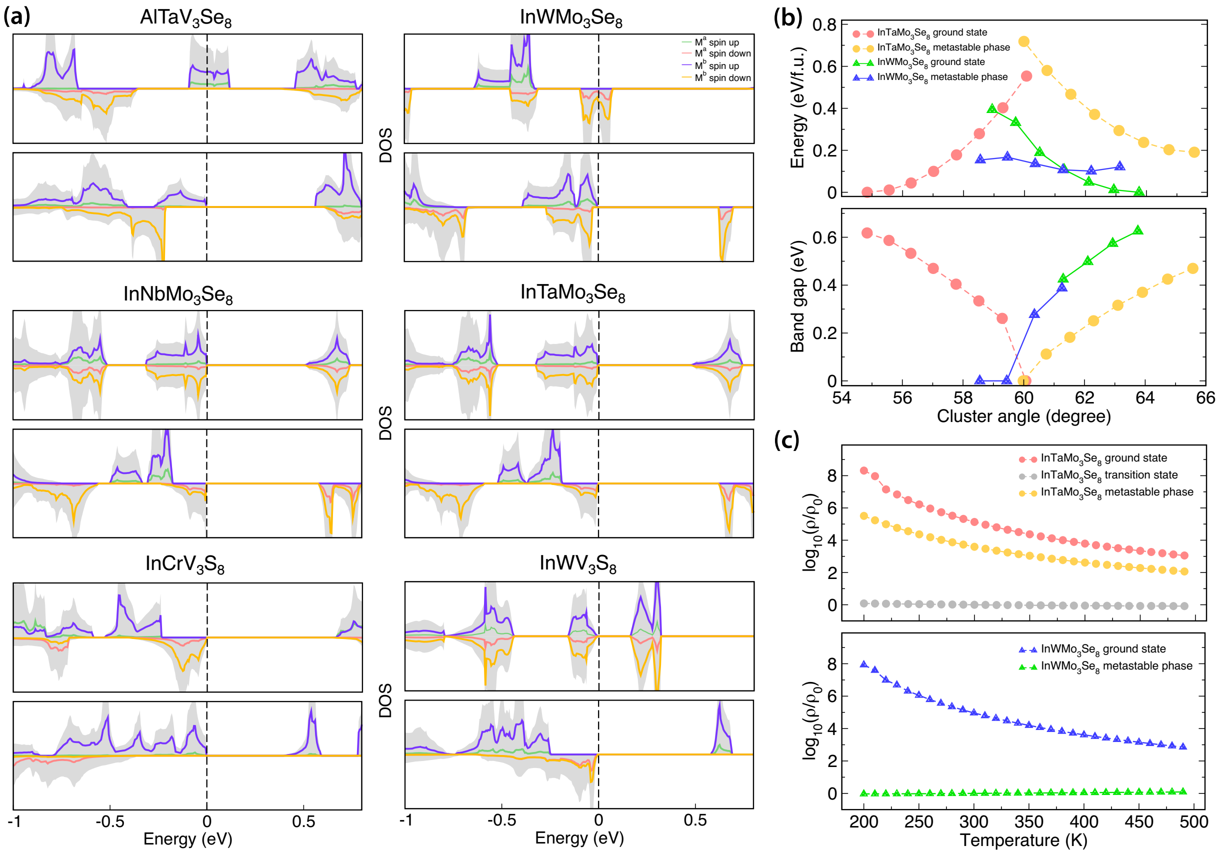


Figure 4.7: DFT-simulated electronic properties of selected lacunar spinel compositions at the Pareto front. (a) The projected electronic density-of-states (DOS) of $AlTaV_3Se_8$, $InWMo_3Se_8$, $InNbMo_3Se_8$, $InTaMo_3Se_8$, $InCrV_3S_8$, and $InWV_3S_8$. The lower panel of each composition shows the ground state electronic structure and the upper panel shows the DOS of the metastable phase after the Jahn-Teller distortion. Both panels are normalized and span a range of 15 states per formula unit for each spin channel (vertical axis). $AlTaV_3Se_8$, $InWMo_3Se_8$ exhibit metal-insulator transitions whereas the other compounds show semiconductor-to-insulator transitions. (b) The DFT relative energies and band gaps of $InWMo_3Se_8$ and $InTaMo_3Se_8$ as a function of the cluster distortion angle θ_m. $InTaMo_3Se_8$ undergoes a semiconductor-to-insulator transition with a metallic intermediate state for $\theta_m \approx 60°$. (c) Simulated DC resistivity of the compounds in (b) for their corresponding metallic, semiconducting, and intermediate states.

nism depicted in Figure 4.7(b). However, the metallic states are different owing to the chemistry of the metals comprising the TMCs. We also find that the basal M^b site plays a more decisive role near the Fermi level with minor contribution from the apical M^a site. The M^a site on the other hand, plays an active role in the Jahn-Teller-active phonon owing to differences in atomic mass (Table 4.1). The remaining lacunar spinels in Figure 4.7(a), $\mathrm{InNbMo_3Se_8}$, $\mathrm{InTaMo_3Se_8}$, $\mathrm{InCrV_3S_8}$, and $\mathrm{InWV_3S_8}$, exhibit a Type II transition. The lower and upper panel show their ground and metastable state pDOS, respectively. Interestingly, some compounds undergo singlet formation and transform into a nonmagnetic phase (e.g., $\mathrm{InNbMo_3Se_8}$) while others remain ferromagnetic after the cluster distortion (e.g., $\mathrm{InCrV_3S_8}$) owing to competition between spin-pairing and magnetic interactions [130].

Last, we model the switching process and resistivity upon structural distortion for $\mathrm{InWMo_3Se_8}$ (Type I) and $\mathrm{InTaMo_3Se_8}$ (Type II) by modulating the amplitude of the ν_{JT} atomic displacements for each material in both the (insulating) ground and (metallic or semiconducting) metastable states. The DFT-simulated energy and corresponding band gap at different cluster angles (θ_m) are shown in Figure 4.7(b). Both compounds show first-order transitions. Owing to the small changes in the TMC geometry required for switching, readily available external stimuli could be used to trigger the transitions [131, 53, 132]. The simulated DC resistivity of $\mathrm{InWMo_3Se_8}$ and $\mathrm{InTaMo_3Se_8}$ clearly shows the promising functionality of these newly discovered compositions in the lacunar spinel family (Figure 4.7(c)). Since we successfully identify all 12 Pareto-front compositions by searching through less than 25% of the design space, our work demonstrates the efficiency of featureless adaptive materials discovery for electronic materials design. The featureless AOE is particularly useful when data availability and physical understanding of the target materials system is limited at either the atomic or microstructural scale.

4.4 Discussion and Outlook

Our multiple property objectives of high stability and large insulating band gaps were achieved by using Bayesian optimization (BO) for MIT materials-composition design without explicitly constructing features (descriptors) via latent-variable Gaussian process implemented in our adaptive optimization engine. We successfully identified all 12 Pareto-front lacunar spinel compositions by searching through less than 25% of the design space. Since the Utopian composition with both high functionality and stability (i.e., the upper right corner of Figure 4.4(b)) cannot be realized, the Pareto front illustrates the trade-offs among objectives. This information is beneficial to materials scientist as it aids in the selection of candidate materials to further investigate or deploy. The selection rules will depend on the designer's preferences and whether to favor one property over others as well as their willingness to compromise. Specifically, for the lacunar spinel family, it is known that experimental synthesis of high-quality single crystals is challenging. For those with different transition metals on the apical M^a and basal M^b site compounds, it is hard to guarantee uniform orientation of the apical atom across the entire crystal (i.e., they might be randomly oriented). Therefore, we report the steps needed to identify all Pareto designs to quantify our model efficiency. Because these materials have garnered much research attention in recent years owing to the richness of their fascinating physical behaviors (e.g., MITs, skyrmion lattices, and superconductivity), we anticipate the newly identified lacunar spinels will be pursued experimentally in search of these phenomena. It is more reasonable to starting with making Pareto designs with 1:4:8 stoichiometry (e.g., $InMo_4Se_8$) which are easier to synthesize and have large insulating phase band gaps.

Although we have seen an increasing emphasis on using Bayesian optimization for materials design, previous work relied heavily upon handcrafted features, which is a challenging task, or single objective optimization. The former usually requires either knowledge of influential features based on theory and literature or large datasets to perform sensitivity analysis and correlation analysis to identify features that influence properties of interest. In the lacunar spinel MIT materials design, the scientific community is limited by chemical intuition as well as large datasets to identify appropriate features. This hinders the application of traditional BO implementations for MIT design. The propensity to use features arises mainly due to a lack of accurate and efficient machine learning methods to model categorical inputs. Here we showed LVGP can circumvent feature identification by directly modelling elements as categorical variables. The mapping of the categorical variables into low-dimensional quantitative latent variables provides an inherent ordering for the categories and physics-based dimensionality reduction. Like conventional Gaussian process models, the LVGP model provides uncertainty quantification, which is crucial for employing the BO strategy for material composition optimization. LVGP enables featureless learning and subsequently featureless BO, making it a generic step forward in machine learning and materials design.

The AOE we demonstrated is theoretically more efficient than evolutionary algorithms for identifying the Pareto frontier in a complex, combinational design space. Although designing materials under a single criterion is more efficient, such efforts may not meet the requirements of deployment. For lacunar spinels investigated here, maximizing E_g exclusively leads to an unstable composition while maximizing ΔH_d exclusively leads to a composition with a small bandgap. In contrast, MOBO identifies the Pareto front to delineate the trade-off between materials properties and allows the designer to choose compositions for detailed study. In this context, the need to perform more iterations of MOBO within the AOE is justified. Indeed, it is typically not the sole goal to find all Pareto front designs, but rather to identify the best candidates within a limited research budget. The AOE clearly provides an efficient way to minimize the effort towards a better design by suggesting the next experimental design.

Similar to forward materials design demonstrated here, inverse materials design [133] can be cast as an optimization problem and tackled via the AOE framework. Although forward design is achieved with the objective of maximizing the desired properties, inverse design can be accomplished by redefining the objective as the minimization of the difference between the predicted and target properties. The design space, i.e., the choice of admissible elements, must be defined appropriately to ensure the target properties are achieved. To that end, our work advances materials innovation for forward and inverse design of both inorganic (as shown herein) and organic materials, such as identification of new quantum materials, design of protein sequence in biomaterials, and monomer sequence in polymeric materials. It is particularly useful when data availability and physical understanding of the target materials system is limited at either the atomic or microstructural scale. This methodology could be further extended to mixed-variable optimization problems, e.g., co-design of composition and chemical stoichiometry through doping, which we are now actively developing.

CHAPTER 5
LEARNING THE CRYSTAL STRUCTURE GENOME FOR PROPERTY CLASSIFICATION

Materials property predictions have improved from advances in machine learning algorithms, delivering materials discoveries and novel insights through data-driven models of structure-property relationships. Nearly all available models rely on featurization of materials composition, however, whether the exclusive use of structural knowledge in such models has the capacity to make comparable predictions remains unknown. Here we employ a deep neural network (DNN) model, `deepKNet`, to learn structure-property relationships in crystalline materials without explicitly considering chemical compositions. The focus is on classification of crystal systems, mechanical elasticity, electrical behavior, and phase stability. The `deepKNet` model utilizes a three-dimensional (3D) momentum space representation of structure from elastic X-ray scattering theory in a manner that includes rotation and permutation invariance. We find that the spatial symmetry of the 3D point cloud, which reflects crystalline symmetry operations, is more important than the point intensities contained within for making a successful metal-insulator classification. In contrast, the intensities are more important for predicting bulk moduli. Our findings here are also supported by learning from simulated neutron diffraction patterns, in comparison. We find learning the materials structure genome in the form of a chemistry-agnostic DNN demonstrates that some crystal structures inherently host high propensities for optimal materials properties, which enables the decoupling of structure and composition for future co-design of multifunctionality.

5.1 Introduction

One of the most frequently used phrases in materials research is “structure-property relationships.” It forms the cornerstone of forward and inverse system-level-based materials design [134, 135],

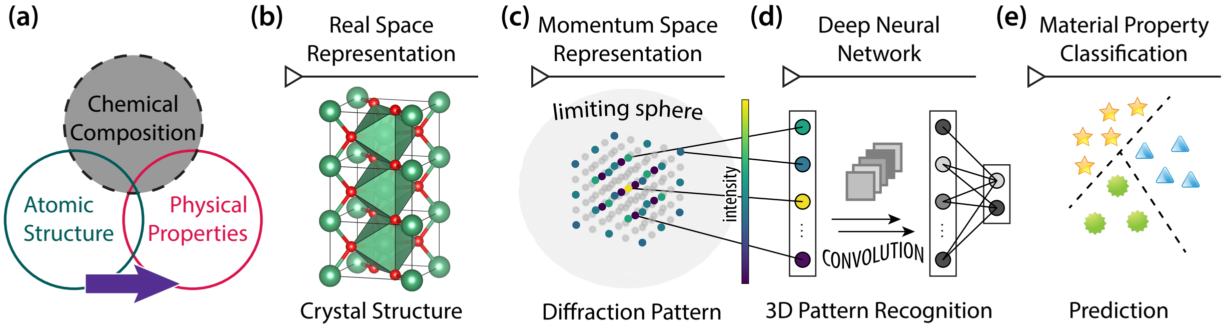


Figure 5.1: Workflow for constructing the `deepKNet` model to learn (a) structure-property relations without featurization of chemical composition. The process begins with (b) the real-space crystal structure representation (in either the conventional cell or primitive cell), which is transformed into a (c) momentum space representation by simulating the 3D X-ray diffraction pattern, which is represented as point cloud. Only diffraction points within the limiting sphere are physically observable. (d) The DNN model is then constructed to learn directly from the point cloud data to accomplish (e) property-classification tasks.

and it is principally used in two modalities: (1) to exclusively describe relationships for a single material family, such that the composition is fixed, and dependencies arise from processing-based microstructural changes, or (2) to explicitly describe effects arising from changes in composition, which inadvertently contracts the full "structure-composition-property" relationship phrase despite chemical dependencies dominating structural changes. Admittedly, both atomic structure and chemistry mutually determine materials properties (Figure 5.1(a)). The intimate interwoven description of what defines a material – the elemental species involved and the crystallographic structure the atoms adopt once bonded together given a fixed ratio – and which physical properties can "live" in various structures pose a challenge for novel materials design and discovery. With the absence of theoretical or statistical guidance, materials scientists need to search through a combinatorial space spanned by both chemical compositions as well as structure types [136].

Despite the key role chemistry plays in physical properties, condensed-matter physicists have harnessed effective theoretical models, e.g., Hubbard, Heisenberg, and Fu-Kane models, etc., based on different interactions, orbital symmetries, and topologies to describe the electronic and magnetic phases of materials without explicitly encoding material composition. The premise relies on recognition that the low-energy electrons comprising atoms interact on a lattice, which may map onto a (portion of a) known crystal structure. Even with modern computational simulations, e.g., those based on density functional theory (DFT), chemical information is only included in the form of atomic orbitals at each crystallographic site and their corresponding atomic numbers to provide a potential for the electrons to interact. To that end, we pose the following question: *Is it possible to marginalize compositional information and understand to what extent crystal structure exclusively determines materials properties?*

In this chapter, we address this question using a statistical learning-based method, leveraging open access to numerous materials databases [3, 4, 2, 137] and recent advances in materials informatics tools [47, 138, 115, 139]. Many machine learning (ML) models exploiting these data have successfully predicted materials properties: local connectivity-based models [140] and graph neural networks [5, 50, 141] have achieved DFT-level performance, and have helped accelerate the discovery of novel functional materials [7, 142]. Here, we learn the materials structure-property relationship from crystal structure alone – without use of chemical composition as illustrated in Figure 5.1(a) – to predict a variety of properties including crystal system, elasticity, metallicity, and stability. This approach is unique from existing materials informatics models, which typically utilize both structural and compositional information as features. We use a momentum-space representation of crystal structures in the form of simulated X-ray diffraction (XRD) patterns to generate a three-dimensional (3D) point cloud, which serves as a unique structural fingerprint of each material. We then construct and train a deep neural network (DNN), which is invariant under rotation and permutation operations on the input 3D XRD patterns, to learn different materials properties. By concealing and perturbing information in the 3D point cloud fed to the DNN, we ascertain that crystal structure plays a decisive role in materials elasticity and metallicity, but it is comparatively less important in determining phase stability. Our findings reveal the correlations among crystal structures and different materials properties, which could enable co-design of material function by prioritizing optimization of crystal structure or composition to achieve desired performances.

5.2 Methodology

5.2.1 Materials representation

A perfect crystal under periodic boundary conditions in real space is mathematically described as the convolution of its Bravais lattice (BL) and the atomic structure of the asymmetric unit (motif) within the unit cell (Figure 5.1(b)). Owing to the periodicity in real space, materials scientists typically use diffraction-based methods (e.g., X-ray or neutron scattering) to determine the crystal structures. The process of X-ray diffraction is the mathematical equivalence of a Fourier transform ($\mathcal{F}$); it converts the real-space crystal structure into momentum space and forms a new reciprocal-space lattice exhibiting intensities dependent on the so-called structure factor (F) as:

$$\begin{aligned}\mathcal{F}(\text{BL} * \text{motif}) &= \mathcal{F}(\text{BL}) \cdot \mathcal{F}(\text{motif}) \\ &= (\text{reciprocal lattice}) \cdot F_{hkl} \qquad (5.1)\end{aligned}$$

where $*$ and $\cdot$ are the convolution and product operations, respectively, and $h\ k\ l$ are integer labels of the reciprocal lattice points that correspond to the Miller indices for lattice planes in real space. The aforementioned real-space convolution relationship then becomes a product between the reciprocal lattice and structure factor F_{hkl}. The physical observable from XRD is the diffraction intensities I_{hkl} (real), not the structure factors F_{hkl} (complex). Rather, I_{hkl} is proportional to the square modulus of the structure factor $|F|^2 = F^*_{hkl} \cdot F_{hkl}$, where $*$ is the complex conjugate, and

$$F_{hkl} = \frac{1}{V_{\text{cell}}} \sum_{j=1}^{N} f_j(\mathbf{g}_{hkl}) e^{2\pi i(\mathbf{g}_{hkl} \cdot \mathbf{r}_j)} , \qquad (5.2)$$

which serves as the Fourier series coefficients of the real space periodic electron density $\rho(\mathbf{r})$ derived from atoms located at $\mathbf{r}_j$ in the unit cell. The atomic scattering factors for atom j at reciprocal point $\mathbf{g}_{hkl}$ are

$$f_j(\mathbf{g}_{hkl}) = \int d\mathbf{r}_j \, \rho(\mathbf{r}_j) e^{2\pi i(\mathbf{g}_{hkl} \cdot \mathbf{r}_j)} . \qquad (5.3)$$

Given the intensity I_{hkl} encodes atomic structure and electron density information, we propose to utilize it as a 3D momentum space representation for predicting physical properties of crystalline materials without explicit compositional features. The diffraction intensity values reflect the number of electrons associated with an ion or element in a material. Owing to the phase problem in crystallography – the complex phase factor is lost upon calculating the square modulus of F_{hkl} – reconstructing the original electron density function through a direct inverse Fourier transform, however, is not feasible. Chemical composition identification is then nearly impossible for our model. The spatial distribution of diffraction intensities, however, are unique to each material as they depend on crystal symmetries of the atomic structure[1]. Therefore, we use the intensity distribution as the structural signature from which to learn materials properties. Since the mapping function from the diffraction intensity I_{hkl} to the target materials properties is unknown (Figure 5.1a, purple arrow), we use DNNs to decode the structure-property relationship as they are ideal candidates for function approximation. Owing to the fact that existing experimental methods typically access a 2D slice of the full 3D diffraction patterns, and not all experimental XRD patterns are readily available in open databases, we simulate the full 3D patterns using a modified version of the XRD calculator implemented in `Pymatgen` [143].

We retrieve materials data from the Materials Project database [3] (using data retrieved on January 20, 2021). In order to ensure the quality of data, we consider only materials with cross-reference labels in the Inorganic Crystal Structure Database (ICSD) database [144]. After filtering based on this constraint, we obtained a dataset comprising 48,524 materials with the following specified properties: crystal system, bulk modulus (B), shear modulus (G), electronic band gap (E_g), and energy above the convex hull (E_H). All materials properties utilized herein were simulated using DFT by the Materials Project. Since not all properties are available for every compound in the database, the total number of materials for each classification task differs (Table 5.1). We assigned thresholds in Table 5.1 for the different classification tasks to ensure physically meaningful

[1] It is possible to artificially make two materials exhibit identical diffraction patterns, but we only consider materials in equilibrium states

Table 5.1: The DNN `deepKNet` is trained on the 3D point cloud representation of materials from the Materials Project database, but the number of materials used for different classification tasks varies due to data availability. The classification boundary values are chosen so as to ensure class balance.

Classification Task	Total Compounds	Class Distribution
Electrical Response	38,917	18,461 with $E_g = 0\,\mathrm{eV}$ (metal) and 20,456 with $E_g > 0\,\mathrm{eV}$ (insulator)
Elasticity	8,804	3,849 with $B > 100\,\mathrm{GPa}$ and 3,246 with $G > 50\,\mathrm{GPa}$
Thermodynamic Stability	48,524	28,512 stable compounds with $E_H < 10\,\mathrm{meV\,atom^{-1}}$

class boundaries (e.g., metal and insulator), and to maintain a balanced dataset.

For each material, we first construct its conventional standard cell using the DFT-relaxed crystal structure reported by the Materials Project (Figure 5.1(b))[2], and then simulate its 3D XRD pattern using Cu K_α radiation ($\lambda = 1.5418\,\text{Å}$). Under our kinematic approximation, only reciprocal lattice points ($h\,k\,l$) within the limiting sphere of radius $4\pi/\lambda$ exhibit finite diffraction intensity while the intensity in the remainder of momentum space is strictly zero (Figure 5.1(c)). The initial features for each material then comprise a set of $\{[h_i, k_i, l_i, I_i] \mid i \in [1, n]\}$ diffraction points, where n is the total number of points within the limiting sphere. Since the shape and size of the reciprocal lattice vary from material to material, as they are dependent on the crystalline symmetry and real space lattice constants, each compound exhibits (1) a unique diffraction point (g_{hkl}) density, (2) configuration of these points within the limiting sphere, and (3) intensity values of these points. Therefore, we further convert the $h\,k\,l$ indices of the diffraction pattern to Cartesian coordinates using the reciprocal lattice vectors. We also take the natural log of the intensity values, $\ln(1+I)$, to bring all features to a similar scale. Implementation details are available in Appendix Section B.1.

Since each material has a different diffraction point density within the limiting sphere, we define a fixed number of g_{hkl} points n to featurize all compounds. We discuss the impact of n on model performance later. Note that n is a variable from which we can learn materials physics; it is not a machine learning hyperparameter. We specifically consider four different n values, which is determined by the range of Miller indices included in the feature set:

- Reciprocal basis vectors $(1\,0\,0), (0\,1\,0), (0\,0\,1)$, $n = 3$ points;
- Miller indices $h\,k\,l \in \{\bar{1}, 0, 1\}$, $n = 27$ points;
- Miller indices $h\,k\,l \in \{\bar{2}, \bar{1}, 0, 1, 2\}$, $n = 125$ points; and
- Miller indices $h\,k\,l \in \{\bar{3}, \bar{2}, \bar{1}, 0, 1, 2, 3\}$, $n = 343$ points.

[2] A primitive standard cell can also be used with comparable performance, however, we report results using the conventional cell because it is easier for symmetry analysis from a human perspective

For instance, in the $n = 125$ case, we include all combinations of $h\,k\,l$ within $\{\bar{2}, \bar{1}, 0, 1, 2\}$, for a total of 125 points, into the feature set. The point cloud representation of some common crystals are shown in Figure 5.2, from which we can see the diversity in point density, shape, and diffraction intensity across different materials. All diffraction points beyond the considered index range are eliminated, and hence invisible to the model. For materials with less than n diffraction points available within the limiting sphere, which occurs for a compound with a small unit cell, we pad the 3D point cloud with dummy points of all zeros to match the size. After this data pre-processing step, all materials should have a feature set defined by an $n \times 4$ array, with n rows and 4 columns: $[x, y, z, I]$, which represent the Cartesian coordinates and the log diffraction intensity, respectively.

This crystalline material representation is in the form of point cloud—an unordered set of points distributed in high-dimensional space. Since the orientation of the reciprocal lattice basis is arbitrary, and the set of points do not follow a specific order, swapping the order of two points should not have any impact on material properties. This behavior is different from pixels in an image. Therefore, our model should be invariant under both 3D rotation and permutation operations on the input points. In order to enforce the rotation and permutation invariance of our model, we apply random 3D rotation and random shuffling of the point sequence of each material before feeding them to the model. Specifically, we use 3 randomly and independently generated Euler angles within the range $[-\frac{1}{4}\pi, \frac{1}{4}\pi]$ for the crystal system classification task, while we use $[-\pi, \pi]$ for all physical property classification tasks. The justification for selecting different ranges of the Euler angles is explained later (*vide infra*). To make the classification tasks more challenging, we not only apply the aforementioned data augmentation to the training set, but also to the validation and test sets to demonstrate the robustness of the model. Therefore, the model never sees the same representation of a material twice, yielding an effectively infinitely sized dataset. In addition, we show later that the performance of the model for property predictions on the test dataset is independent of the random 3D rotations and point permutations.

We split the dataset into training, validation, and test sets, with ratios of 0.6, 0.2, and 0.2,

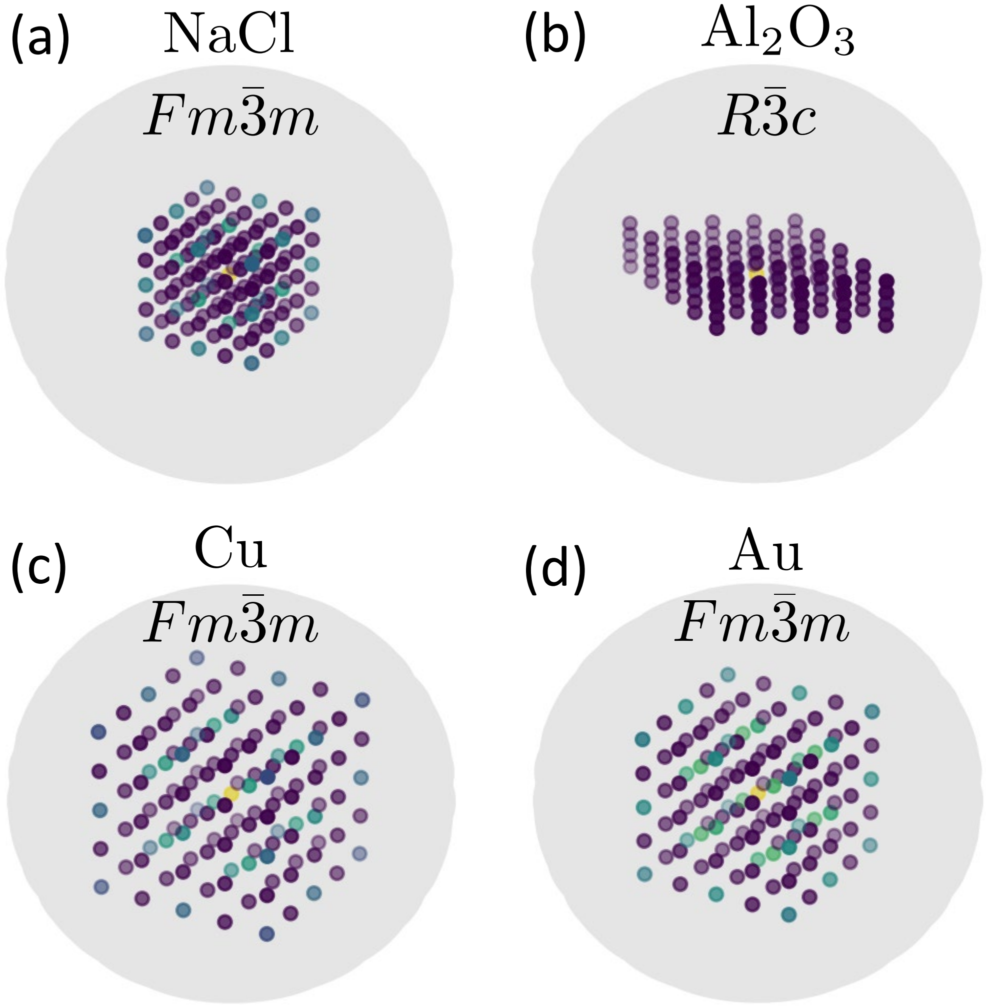


Figure 5.2: The simulated X-ray diffraction patterns of select crystals with corresponding space group. The gray spheres represent the limiting sphere of radius $\frac{4\pi}{\lambda}$. Only diffraction points with Miller indices within $\{\bar{2}, \bar{1}, 0, 1, 2\}$ are shown here. The intensity of the origin (0 0 0) is calculated as the total electron density within cell. This point cloud representation of crystal structures simultaneously displays rotation and permutation invariance.

respectively. The validation set is used to select the optimal combination of hyperparameters. We report the model performance on the test set containing materials that the model has never seen. Since our goal is to understand materials physics using a DNN as an information extractor, we train each model on 3 randomly and independently generated training-validation-test datasets, and report the mean value performance metric on the test set to reduce the variance of results.

5.2.2 Network architecture

Learning from 3D point-cloud data is an active area of computer-vision research. Owing to the rotation and permutation invariance requirements of our I_{hkl} point-cloud representation, most conventional ML models cannot be directly applied to our learning problem. For instance, conventional 2-dimensional convolutional neural networks (CNNs), which are the most prevalent network structure for 2D image classification tasks [145], are robust against object translations; however, permutation of the input data (e.g., swapping pixels of an image) could break down the network. Existing solutions to this problem include PointNet [146], multi-view CNN [147], and some other CNN variants [148, 149]; however, these tend to focus on object detection/classification and segmentation learning tasks.

Here we demand more from the neural network model, which goes beyond the 3D computer vision problem—the analogue of which would be identifying the 1 among 7 crystal systems a material belongs to by knowing how atoms are arranged in a unit cell. The features we use for the materials-property classification tasks include not only positional data (i.e., Cartesian coordinates), but also the diffraction intensity as the fourth dimension. Thus, the input features together contain information about the cell shape, cell size, symmetry, and electron density. This information is all simultaneously embedded within the sparse distribution of diffraction points in momentum space. To that end, the DNN needs to learn the patterns of different material properties (e.g., metals and insulators) using their structural fingerprints, and not only identify structural patterns given structural features [150].

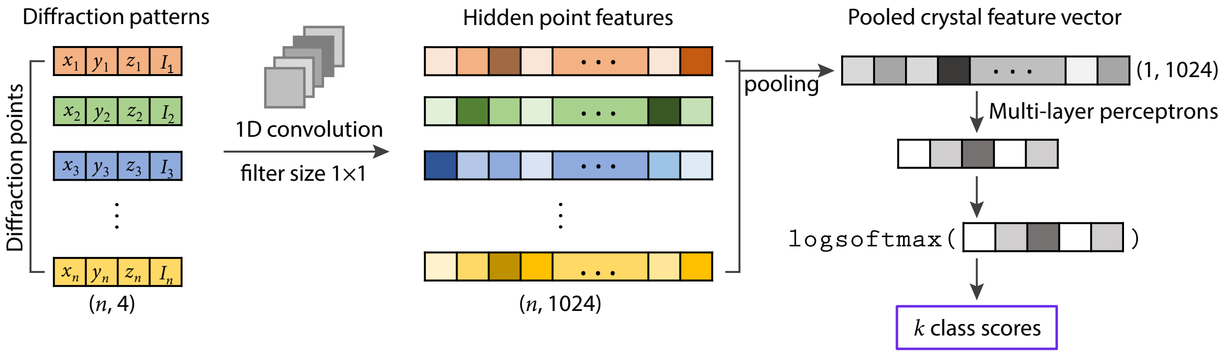


Figure 5.3: The `deepKNet` architecture. Multiple 1D convolutional layers with filter size 1×1 are applied to extract the position-intensity relationship from the simulated diffraction data. The shape of feature and intermediate tensors are indicated in parenthesis and n is the number of diffraction points considered. Operations in this step do not involve point-point communications; therefore, permutation invariance is preserved. Then, a symmetric function is used to pool the crystal feature vector from all diffraction points. Here, the `max` pooling function is used, but others also work. Lastly, multi-layer perceptrons are used to eventually make the classification decision. See Appendix Section B.2 for details of model hyperparameter selection.

The network architecture capable of solving this problem is elegant in its simplicity as depicted in Figure 5.3. Inspired by PointNet, we use 1-dimensional (1D) convolutional layers with filter size 1×1 to extract features from the primitive point cloud data. Each feature column, i.e., Cartesian coordinates and intensity, is treated as one input channel, and the filters convolve over all points in each channel, then are summed over the input channels to obtain newly learned output channels. After a few layers of 1D convolution, the model learns the position-intensity relationship of different points, whose output features should be invariant to rotation of the Cartesian coordinates of input points (e.g., distance to origin). This step only involves operations within each individual point. No point-point communications are made (owing to the size of filter being 1×1), hence preserving permutation invariance. Now, the learned material representation becomes a tensor of shape (n, m), where m is a hyperparameter indicating the number of embedding dimension. (We use $m = 1024$ for all classification tasks.)

After obtaining the hidden point features from the 1D convolutional layers, we apply a symmetric function to aggregate information from all points. We find that the `max` pooling function works well in all our tasks, and this operation safely preserves permutation invariance, because it does not involve point indexing. In addition, we also tried a self-attention-based pooling algorithm, and found that the performance gain is negligibly small (e.g., ROC-AUC value from 0.910 to 0.915 for metal-insulator classification, and from 0.950 to 0.957 for bulk modulus classification) while the model size becomes several times larger than using the `max` pooling function. Therefore, although knowing that `max` pooling is not the only working method for information aggregation, we use this pooling function for all our classification tasks. It also enables physically meaningful model interpretation since it allows us to know which points contribute to the pooled crystal feature vector (*vide infra*). Multi-layer perceptrons are then used after the pooling layer and eventually the model will make a multi-class prediction from the input point cloud representation. We apply batch normalization to all convolutional and fully connected layers. Other network structures that can deal with 3D equivariance[151] are also viable solutions to our problem, but we find that the performance bottleneck mainly originates from the input features rather than the network.

To compare the physical knowledge learned by the network, we use the same network structure (with different parameters) to learn all target properties. Details of hyperparameter selection are given in Appendix Section B.2. Model performance in all classification tasks is based on averaging over three independent runs with different data splits.

5.3 Results and Discussion

5.3.1 Learning crystal systems

We begin our initial assessment of the learning capability of `deepKNet` using a simple computer vision task: crystal-system classification. The objective is to predict the correct crystal system for a material given only the XRD pattern. Because hexagonal and trigonal cells have identical conventional cell shapes, i.e., $a = b \neq c; \alpha = \beta = 90°$ and $\gamma = 120°$, we combine these classes together as one, which leads to a total of 6 classes: cubic, tetragonal, orthorhombic, hexagonal/trigonal, monoclinic, and triclinic. Since the crystal systems are uniquely defined by the real space lattice vectors, we should only need to provide the model with $n = 3$ diffraction points, corresponding to the reciprocal lattice basis vectors. We also mask the diffraction intensity information for this task by removing the fourth dimension of each point, making it invisible to the model.

The `deepKNet` model achieves excellent performance with an accuracy of 0.98 on the test set. We find that many of the misclassifications are caused by the difference in threshold of "equivalence." For instance, the model has difficulty differentiating tetragonal from orthorhombic cells when the ratio of two lattice parameters are approximately unity. See Appendix Section B.3 for additional analysis of the crystal-system classification. Furthermore, we tolerate the less-than-perfect accuracy after recognizing the network is not fully rotation-invariant for the crystal-system-classification task. Here, the Euler angles are constrained between $[-\frac{1}{4}\pi, \frac{1}{4}\pi]$ rather than using completely arbitrary rotation angles spanning 2π, because the network architecture we use works best with certain spatial orientations of the points. The PointNet-like model in `deepKNet` has difficulty in finding a principal axis and canonicalizing the input when utilizing large rotation angles [146]. Nonetheless, `deepKNet` is able to "visualize" the shape of the 3D point cloud representation regardless of random 3D rotations. Interestingly, the physical properties considered in the next section are completely immune to such random 3D rotations, which is reasonable as the properties are scalar quantities.

Next, we ask a more challenging question—is it possible to distinguish between materials exhibiting trigonal and hexagonal cells? We find that given only the three reciprocal lattice basis vectors without diffraction intensity values, the model achieves an area under the receiver operating characteristic curve (ROC-AUC, later referred to as AUC) of 0.87. However, once we unmask the diffraction intensity of the three points, the AUC value increases to 0.94. As we further increase the number of diffraction points (with intensity) from $n = 3 \to 27$, `deepKNet` performance significantly improves. It distinguishes between the trigonal and hexagonal systems with an $\mathrm{AUC} = 0.97$. The results here primarily show that the diffraction intensity I plays a significant role for our classification model, which is an advantage of using 3D features over projected

2D patterns [150]. The amount of momentum space knowledge (n) plays a secondary role and is explored in more detail next. Therefore, we always include the diffraction intensity information in the remaining classification tasks.

5.3.2 Learning properties

We next train `deepKNet` to learn materials properties by learning hidden patterns within the 3D point cloud data based on crystal structure. The four materials properties we target are metallicity, bulk modulus, shear modulus, and thermodynamic stability. The classifications involve: separating compounds without (metals) from those with (insulators) a 0 K gap E_g in the electronic structure at the DFT level, distinguishing stiff compounds with bulk modulus (B) greater than 100 GPa, or shear modulus (G) larger than 50 GPa, from flexible compounds, and identifying thermodynamically stable materials with $E_H < 10\,\mathrm{meV\,atom^{-1}}$, respectively.

First, we examine the impact of the total number of diffraction points (n) on model performance for each classification task (Figure 5.4(a)). For all tasks, we find that as more diffraction points become visible to the model, the performance of the classifier initially improves significantly (from $n = 3 \rightarrow 27 \rightarrow 125$). The performance then plateaus after 125 points with negligible performance gain using 343 diffraction points. Figure 5.4(a) also reveals that the electrical and mechanical properties are predicted with better quality than the thermodynamic properties. This behavior is reasonable given the importance of composition and chemical identity to material stability [152]. Thus we conclude that limited stability information can be learned from crystal structure alone.

Figure 5.4(a) also makes it clear that the reciprocal lattice vectors are available ($n = 3$) produce distinct baseline performances among the properties examined. Specifically, the metal-insulator classifier achieves an AUC of 0.80, a value often considered as an "effective" model performance. These 3 diffraction points indicate the model only has knowledge about the crystal system and cell volume, which we validated using a simple random forest model (Appendix Section B.4). Although we typically compare the AUC value of a binary classifier with 0.5 as baseline, here we

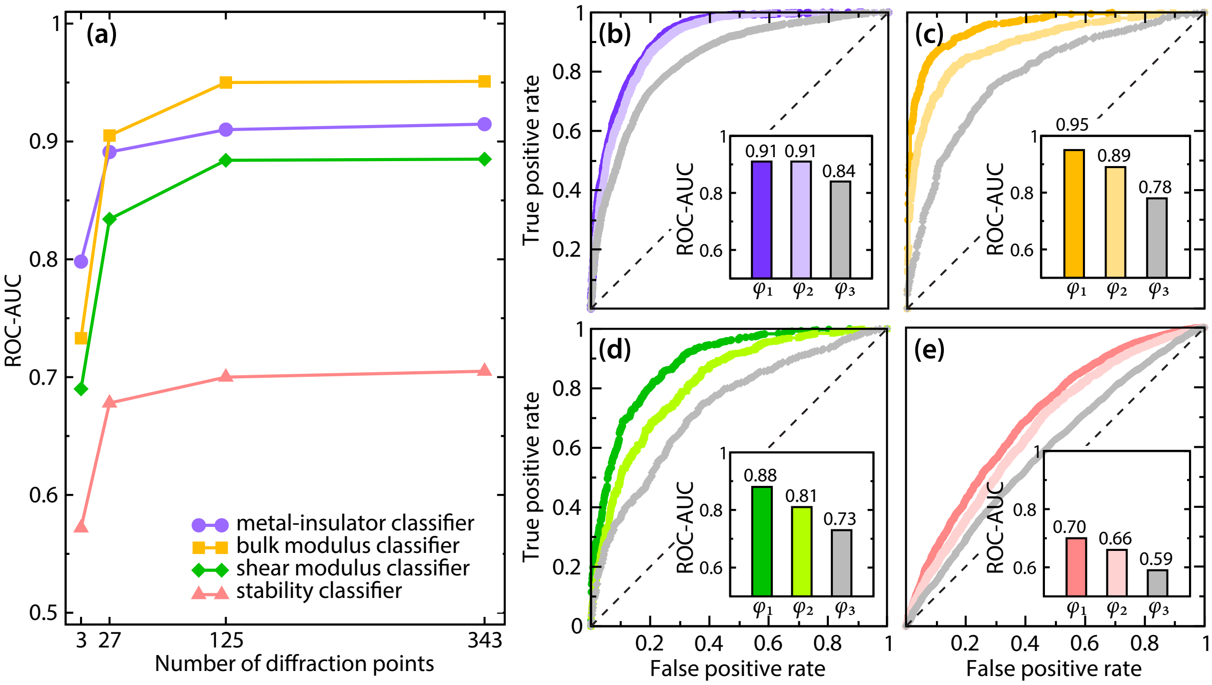


Figure 5.4: Model performance in multiple classification tasks. (a) ROC-AUC values in four binary classification tasks with a different total number n of diffraction points visible to the model as described in the Methods. ROC curves with $n = 343$ for the (b) metal-insulator classification, (c) bulk modulus classification, (d) shear modulus classification, and (e) thermodynamic stability classification. Model performance using the original diffraction dataset (dark coloring, φ_1), randomly scaled intensity (light coloring, φ_2), and only systematic absence information (gray, φ_3) are shown in the insets.

emphasize in the case of metal-insulator classification one should assess the performance of the model with 0.8 rather than 0.5. This comparison with an AUC of 0.8 is what is expected based on minimal knowledge fed to the ML models, and it is unlikely to provide significant insights to facilitate materials design. Moreover, in most ML work, the baseline (i.e., worst-case model performance) is rarely discussed, yet it is quite important for researchers to understand the difficulty of such predictive tasks.

Next, we focus on understanding the model performance on the metal-insulator and bulk and shear moduli classifications—what exactly does the model learn from the diffraction patterns? DNN model interpretability is a known problem owing to the nonlinear activation functions and complex network structures. To that end, we choose another route to understand the model performance. Instead of "opening the black-box", we make perturbations to the input features to form new datasets φ_i, and examine the response as quantified with the true and false positive rates and ROC-AUC values for each classification task using the same DNN architecture (Figure 5.4(b-e)). We assign the original diffraction data as φ_1. It contains information pertaining to the crystal lattice parameters (position of diffraction points), crystal symmetry (spatial distribution of relative diffraction intensity), and electron density (diffraction intensity values). These are the input features from which we determine the relative contributions in the final decision-making of the `deepKNet` model.

To separate the diffraction intensity values from their spatial symmetry, we generate a random multiplier uniformly sampled within the range (0, 1] for each material during each training epoch, and then scale all of its diffraction intensity values with this multiplier before feeding them to the model. Different materials will have different random multipliers, but all diffraction points within the same material will be scaled by the same multiplier. The randomly scaled diffraction patterns correspond to the dataset φ_2, and would preserve the spatial symmetry (i.e., relative intensities) of the diffraction points, but the model would not be able to rely on the absolute values of the intensities, which are related to the electron density and atomic numbers. In addition, we also examined whether the model is learning from systematic absence information in the dataset, i.e., $h\,k\,l$ combinations that have zero intensity, to make predictions. Dataset φ_3 is obtained by replacing all non-zero diffraction intensity values with unit intensity, $I_{hkl} = 1$, while all others remain $I_{hkl} = 0$.

Figure 5.4(b-e) present the model performance with different perturbations to the input diffraction patterns. We find that the metal-insulator classifier is significantly more robust against random scaling of the intensity values than other classifiers, where it is still able to achieve $\mathrm{AUC} = 0.91$ with random intensities (see φ_2 in Figure 5.4(b)). The performance of the bulk modulus and shear modulus classifiers reduce from 0.95 to 0.89, and from 0.88 to 0.81, respectively. These changes are statistically meaningful (Appendix Section B.5). Notably, we achieve a truly composition-free model after random scaling of the materials diffraction intensity values. The model completely loses information about atomic number and electron density in this case, but it is still aware of which $\mathbf{g}_{hkl}$ points are symmetric and their spatial distributions. Our findings here suggest that the metal-insulator classifier relies mostly on the spatial symmetry of the diffraction patterns, while the elasticity-property classifiers depend more on the absolute intensities, which encode the electron density.

All models exhibit inferior performance with only systematic absence information, as indicated by the gray curves (φ_3 in Figure 5.4(b-e)). The results here are reasonable, because we lose some symmetry information as all finite diffraction intensity values become unit intensity. We conclude that the model learns distinct patterns for different target materials properties, and is able to capture the physically meaningful features (e.g., spatial symmetry of diffraction patterns) to learn the materials structure genome and make property predictions.

5.3.3 Model interpretation

We now partially open the black box of the DNN model to further understand how it classifies metals from insulators. We plot the distribution of critical points both with normalized interplanar d_{hkl} spacings and in the limiting sphere that contribute to the final crystal feature vector of 6 well-known materials (Figure 5.5). In order to facilitate visualization, we choose a small model which uses $n = 125$ diffraction points as input and 32-dimensional crystal feature embeddings. This small model has $\mathrm{AUC} = 0.89$, which is acceptable for use in model interpretation. Larger models will have better performance, yet more complicated classification rules. The model correctly predicts the metallicity of all 6 crystals with high confidence. The complete list of Miller indices of the critical points are provided in Appendix Section B.6.

We consistently find an important critical point at large d-spacing, which corresponds to the (001) reciprocal basis vector and for the cubic systems presented defines the lattice shape and cell volume. The model also requires more information from the lattice planes with smaller interplanar

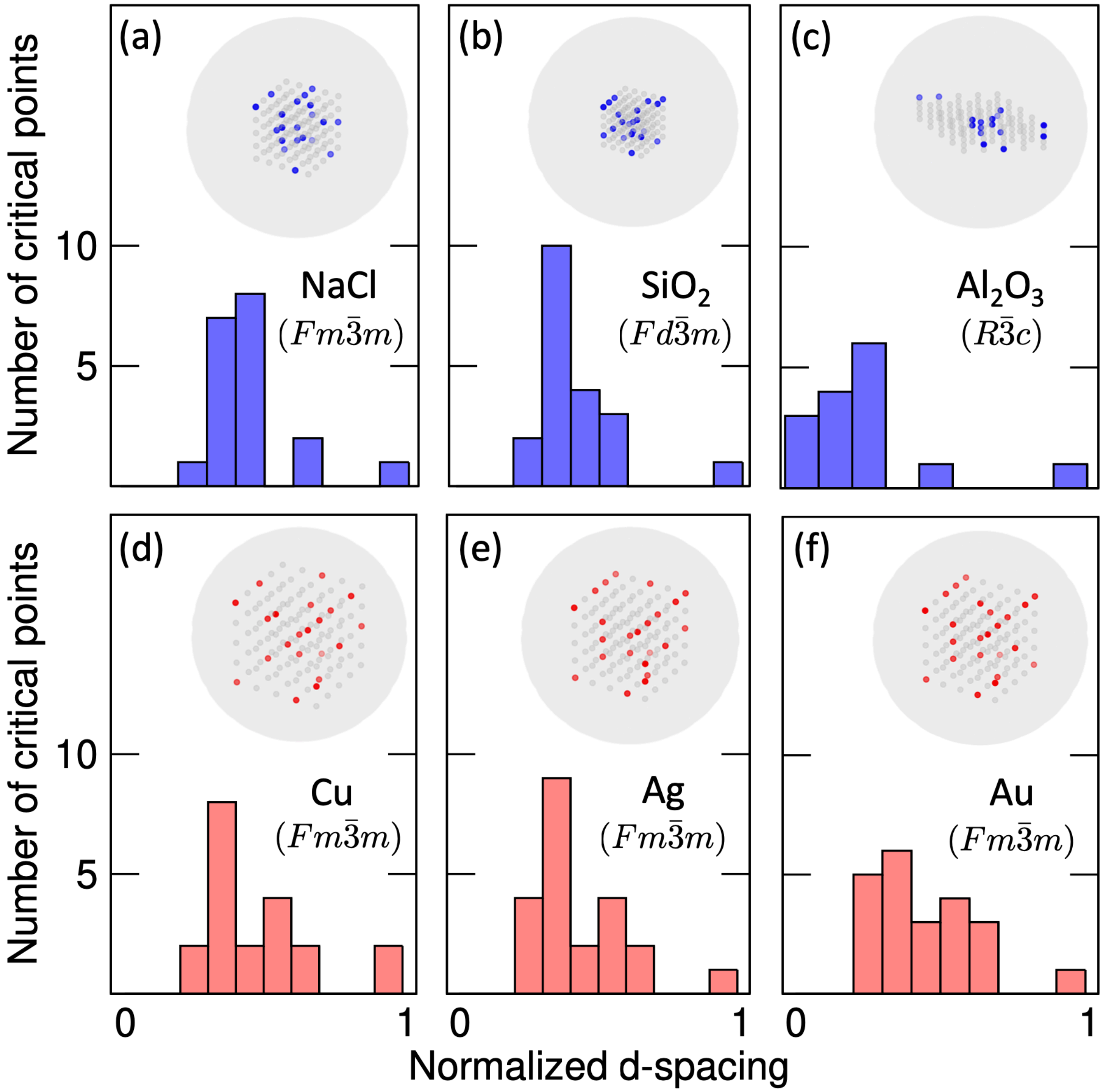


Figure 5.5: Distribution of critical diffraction points with normalized interplanar d_{hkl} spacings of a few common insulators (NaCl, SiO_2, Al_2O_3) and metals (Cu, Ag, Au). The d-spacings are normalized to facilitate comparison across different materials. The critical points in the limiting sphere (the gray sphere) are those that contribute to the final crystal feature vector after `max` pooling, and are marked with blue for insulators, and red for metals, respectively. Non-critical points are represented with light gray points.

distances, which correspond to higher $h\,k\,l$ indices, as seen by the clustering of critical points. This is reasonable because such points provide information about interplanar interactions as these distances, which are governed by orbital hybridization and attractive and repulsive electostatic contributions. In addition, we find all 6 materials exhibit at least one "gap" in the d-spacing distribution. The critical point distribution of Cu and Ag are almost identical. Although Au and NaCl exhibit the same space group as Cu and Ag (i.e., $Fm\bar{3}m$), their distributions are different. A thorough understanding of the model prediction mechanism remains difficult at this time owing to the complicated decision rules underlying the deep neural network. Interestingly, the model learns the operation of spatial parity. It recognizes inversion symmetry inherent to the XRD patterns (Friedel's law), since it only contains an average of 2 duplicate points with inversion symmetry in the final critical point set, e.g., $(2\,2\,2)$ and $(\bar{2}\,\bar{2}\,\bar{2})$.

5.3.4 Model limitations

Since we do not explicitly have elemental composition information in the XRD patterns, we expect the model to have difficulty making predictions on materials from the same family, i.e., with similar crystal structures yet different compositions and various properties. To that end, we examine the model performance on the $AB\mathrm{O}_3$ perovskite family (Table 5.2). All compounds listed here were removed from the training and validation dataset for this classification task.

Overall the model performs poorly in classifying metals from insulators in the perovskite family. We find the model tends to predict all trigonal ($R3c$ and $R\bar{3}c$) and orthorhombic $Pnma$ compounds to be insulators. The model in general exhibits low confidence scores in predicting most of the perovskite materials, which is reasonable since minor structural distortions in these materials could drive metal-to-insulator transitions [153], while the change in diffraction patterns might be indistinguishable to the model. The model also makes significantly more insulator predictions than metals in this family, whereas the true labels are more balanced. The model performance in the perovskite family is reasonable since undoubtedly chemistry and interactions among different

Table 5.2: Model performance for select materials in the perovskite family. ‘M’ and ‘I’ labels indicate metal and insulator, respectively. The score is the probability associated with the predicted class, indicating how confident the model is on that prediction.

Compound	Space group	True label	Prediction	Score
$LiNbO_3$	$R\bar{3}c$	I	I	0.72
$LiOsO_3$	$R\bar{3}c$	M	I	0.56
$LaNiO_3$	$R\bar{3}c$	M	I	0.56
$LaCoO_3$	$R\bar{3}c$	M	I	0.57
$LiNbO_3$	$R3c$	I	I	0.70
$LiOsO_3$	$R3c$	M	I	0.56
$LiTaO_3$	$R3c$	I	I	0.71
$NdNiO_3$	$Pnma$	M	I	0.61
$YNiO_3$	$Pnma$	M	I	0.54
$CaFeO_3$	$Pnma$	M	I	0.75
$SrRuO_3$	$Pnma$	M	I	0.62
$CaTiO_3$	$Pnma$	I	I	0.78
$NdNiO_3$	$P2_1/c$	I	M	0.83
$YNiO_3$	$P2_1/c$	I	I	0.54
$CaFeO_3$	$P2_1/c$	I	I	0.77
$SrFeO_3$	$Pm\bar{3}m$	M	M	0.73
$SrTiO_3$	$Pm\bar{3}m$	I	M	0.64

microscopic electronic, spin, and orbital degrees-of-freedom play a significant role in determining materials properties. Although the perovskite family poses a challenge to `deepKNet`, the poor performance is expected since we designed this task to reveal the limitations of only using structural information to predict materials properties. The aforementioned model performance across many structure types still uncovers that metals and insulators exhibit distinct XRD patterns, and our model is able to capture those difference effectively.

5.3.5 Learning from neutron diffraction data

We now use neutron scattering (ND) patterns to represent crystal structures, rather than the XRD patterns to further assess whether the model can distinguish atomic species from diffraction intensity information. Neutrons interact with the nuclei via the nuclear strong force, whose interaction can be approximated by a short-ranged Fermi pseudopotential. Since the Fermi pseudopotential is a delta function, whose strength is parameterized by the scattering length b, the neutron form factor is Q-independent in momentum space, which is the main difference between neutron and X-ray scattering (Equation 5.3). In addition, the neutron scattering lengths are nonmonotonic across the periodic table and differ even between isotopes of the same element. Therefore, the model cannot learn the total electron density in the same way as from the XRD patterns.

We perform the same series of classification tasks and make identical perturbations to the intensities, corresponding to data φ_i, described before using the ND patterns as input (Figure 5.6). The overall performance using ND and XRD features are quite similar for all four classification tasks, although the model learns relatively less from the ND patterns. A significant performance loss occurs for φ_1 and φ_2 for both bulk and shear moduli (e.g., AUC drops from 0.95 to 0.9 for bulk modulus) when going from the XRD to ND models, which supports our previous hypothesis that the electron density plays an active role in determining elastic properties. This suggests the model is exclusively making predictions based on the cell shape, volume, and crystal symmetry information. That being said, we learn that metals and insulators do look different from a crystal structure perspective. We also notice that φ_3 from the XRD and ND data remains identical throughout all classification tasks. This behavior is expected because with φ_3, the model only has information about whether a point has finite diffraction intensity due to crystal symmetry constraints, regardless of the scattering probe used. By comparing the results from XRD and ND, we are more confident that crystal structure alone plays a significant role in determining the electronic and elastic properties of crystalline materials, while it is less important for thermodynamic stability.

5.4 Conclusions and Outlook

In conclusion, we use DNN models to show the intimate correlation between crystal structure and materials metallicity and elasticity. We learn from both XRD and ND patterns that crystal symmetry plays a significant role in determining electronic band gaps, while electron density contributes more to elastic properties. Stability, however, is strongly composition-dependent and therefore

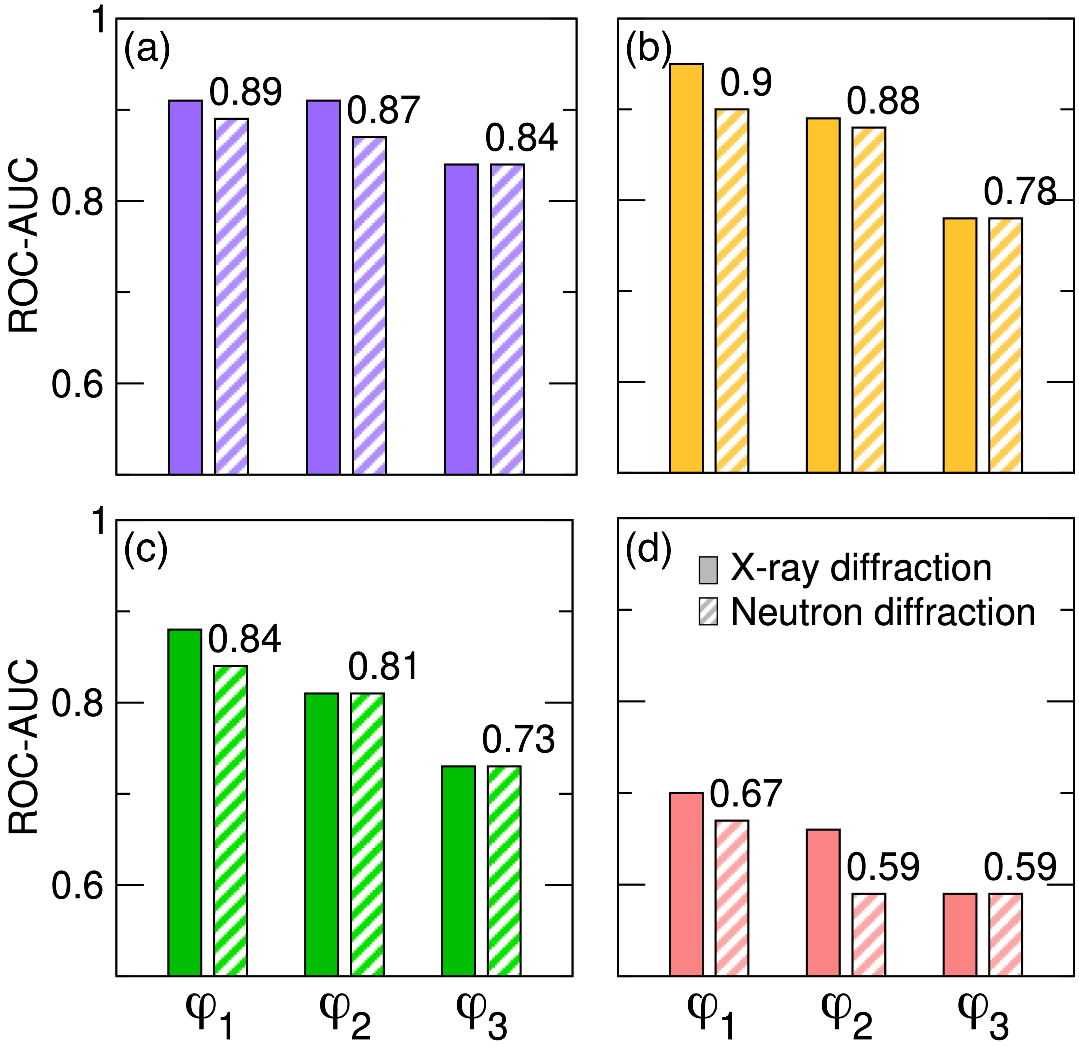


Figure 5.6: Comparison of model performances using X-ray diffraction (fully shaded bars) and neutron diffraction (striped bars) patterns as input features for (a) metal-insulator, (b) bulk modulus, (c) shear modulus, and (d) thermodynamic stability classification tasks. ROC-AUC curves with $n = 343$. φ_1 through φ_3 represent the same input perturbations as in Figure 5.4. Only neutron diffraction data is annotated since the XRD values are the same as reported in Figure 5.4. The standard deviation of the reported AUC values are tabulated in Appendix Section B.5.

our model exhibits poor performance in predicting this thermodynamic response. These findings impart a better understanding of the role of crystal structures in functional properties.

The motivation and objective behind our structure-based DNN is fundamentally different from other existing materials informatics models (e.g., the crystal graph convolutional neural network [5]). For conventional informatics or machine learning tasks, researchers typically first obtain some data, then construct a learning model, and at last make some predictions using the trained model. This is an engineering-driven task, whose goal is to predict some target properties as accurate as possible. In other words, a "perfect model" is expected. However, we made a new attempt here – we make perturbations on physically meaningful input data, and use the neural network as information extractor, so that we can learn some distilled materials physics from different system responses. Instead of focusing on prediction accuracy, our goal here is to find the upper and lower bounds of system response and understand the governing factors of materials systems. The motivation behind this is more about understanding materials physics rather than building a predictive model. The question still remains whether we could apply similar input perturbations (to decouple these entangled factors such as structure and composition) on other materials descriptors in order to learn more materials physics. It is worth more investigation.

Moreover, if we have the exact Fourier series expansion of the periodic electron density function in real space, it would be possible to construct a sophisticated enough DNN model to learn the functional that maps ground state electron density to materials properties. However, this would require us to obtain orders of magnitude more number of points instead of only a few hundred, which is currently impractical. Based on our current understanding of the `deepKNet`, the network architecture is not learning the functional mapping, but mainly making predictions based on spatial symmetry and electron density information hidden in the diffraction patterns. In other words, it is performing complex pattern recognition rather than learning the underlying functional relationship and mathematical structure of materials. This fact may be a result of performing classification tasks rather than regression modeling. We suspect that learning the density functional mapping using a regression DNN model is possible, but requires a large neural network of unknown architecture.

Lastly, our work here not only reveals some interesting correlation between crystal structure and materials properties, but also demonstrates the capability of DNNs beyond making accurate property predictions. They are also valuable in advancing our materials-physics understanding through statistical analysis. This makes DNNs complementary methods to theoretical modeling and physics-based simulations.

CHAPTER 6
SYMBOLIC REGRESSION IN MATERIALS SCIENCE

This chapter is adapted with permission from Ref. [154]. The work was performed and written in collaboration with Dr. Nicholas Wagner. © Copyright 2019 Materials Research Society.

In this chapter, we showcase the potential of symbolic regression as an analytic method for use in materials research. First, we briefly describe the current state-of-the-art method, genetic programming-based symbolic regression (GPSR), and recent advances in symbolic regression techniques. Next, we discuss industrial applications of symbolic regression and its potential applications in materials science. We then present two GPSR use-cases: formulating a transformation kinetics law and showing the learning scheme discovers the well-known Johnson-Mehl-Avrami-Kolmogorov (JMAK) form, and learning the Landau free energy functional form for the displacive tilt transition in perovskite $LaNiO_3$. Finally, we propose that symbolic regression techniques should be considered by materials scientists as an alternative to other machine-learning-based regression models for learning from data.

6.1 Motivation

6.1.1 Era of big data in materials science

Modern scientists perpetuate the scientific process embodied by the works of Tyco Brahe, Johannes Kepler, and Isaac Newton in the heliocentric revolution. Brahe was the observationalist. He took extensive, precise measurements of the position of planets over time. Kepler was the phenomenologist. From Brahe's measurements, he derived concise analytical expressions that describe the motion of the solar system in a succinct manner. Last, Newton was the theorist. He realized the mechanism behind the apple falling from the tree is the same as that underlying planets traveling around the sun, which could be formulated into a universal law (Newtonian gravitational law). All three scientific modalities are vital in making scientific discoveries: data acquisition (Brahe), data analysis (Kepler), and derivation from first-principles (Newton).

With recent advances in computer science, theoretical modelling, and experimental instrumentation, materials scientists have in many ways created a “mechanical Brahe” and marched into a new era of big data. Datasets of materials information, obtained from advanced characterization techniques [155, 156, 157], combinatorial experiments [158, 159, 160], high-throughput first-principles simulations [161, 162], literature mining [163, 164], and other techniques, are created at a faster rate every day with less and less human labor. All of this data enables new opportunities to construct novel laws of phenomenological behavior for systems that previously lacked them.

Inspired by the Materials Genome Initiative (MGI) [1], the materials community is working collaboratively towards making digital materials data accessible to others. Multiple materials databases such as Materials Project [3], OQMD [4], AFLOWLIB [2], OMDB [137], AiiDA [165], Citrination and NOMAD, provide public access to millions of materials data points. Accessibility to an immense amount of materials data paves way for the next step of “automating Kepler” in the discovery of governing laws in materials processing-structure-properties-performance relationships, which could advance materials discovery, development, and technology innovation.

Since one of the fundamental research objectives of materials science and engineering is to deliver new materials with optimal performance under specified constraints, it is essential to understand how and which features govern the functionality. In other words, which degrees-of-freedom (or parameters) and their corresponding intrinsic relationships (or dependencies) to the material properties should be optimized. However, the multi-scale nature of materials science [159], e.g., from atomic-scale crystal structure to complex mesoscale domain structures and bulk mechanical properties or from femotosecond laser probes to hour-long recrystallization reactions, makes it particularly challenging to study many hierarchical relationships of different materials families. Given such a high-dimensional parameter space (e.g., chemical composition, crystal structure, external conditions, etc.), materials scientists often explore a finite subspace of all the factors that govern materials properties and performance. In addition, the available data is typically sparsely distributed. Although, access to a large materials database relieves, in part, the limited-data problem, there is an urgent need for a robust data-processing protocol to help discern governing laws in materials science and to deliver designer materials and synthesis/processing procedures.

6.1.2 An alternative to machine-learning methods

Much of the burgeoning field of materials informatics focuses on the aforementioned challenges. Machine learning (ML) models are currently the tools of choice for uncovering these physical laws. Although they have shown some promising performance in predicting materials properties [166], typical parameterized machine learning models are not conducive to the next stage of generalizing across domains—the ultimate goal of "automating Newton."

It is important to note that Newton's challenge was somewhat made easier, because Kepler's laws were parsimonious yet predictive. In a modern context, ML models can be predictive but their descriptions are often too verbose (e.g., deep-learning models with thousands of parameters) or mathematically restrictive (e.g., assuming the target variable is a linear combination of input features). Such black-box models have become more prevalent in modern materials science research; however, the interpretability of such models have always been a problem. Although there is a large body of work on data visualization and model understanding to address these issues, those subjects will be out of the scope of this perspective (see Ref. [167] for a review).

In this chapter, we focus on an alternative to machine-learning models: symbolic regression. Symbolic regression simultaneously searches for the optimal form of a function and set of parameters to the given problem, and is a powerful regression technique when little if any a-priori knowledge of the data structure/distribution is available.

Figure 6.1 shows the relative popularity of machine learning and symbolic regression in different research domains. We use data from the "Web of Science Core Collection" database in this analysis [168]. Among all publications whose topics are related to machine learning or symbolic

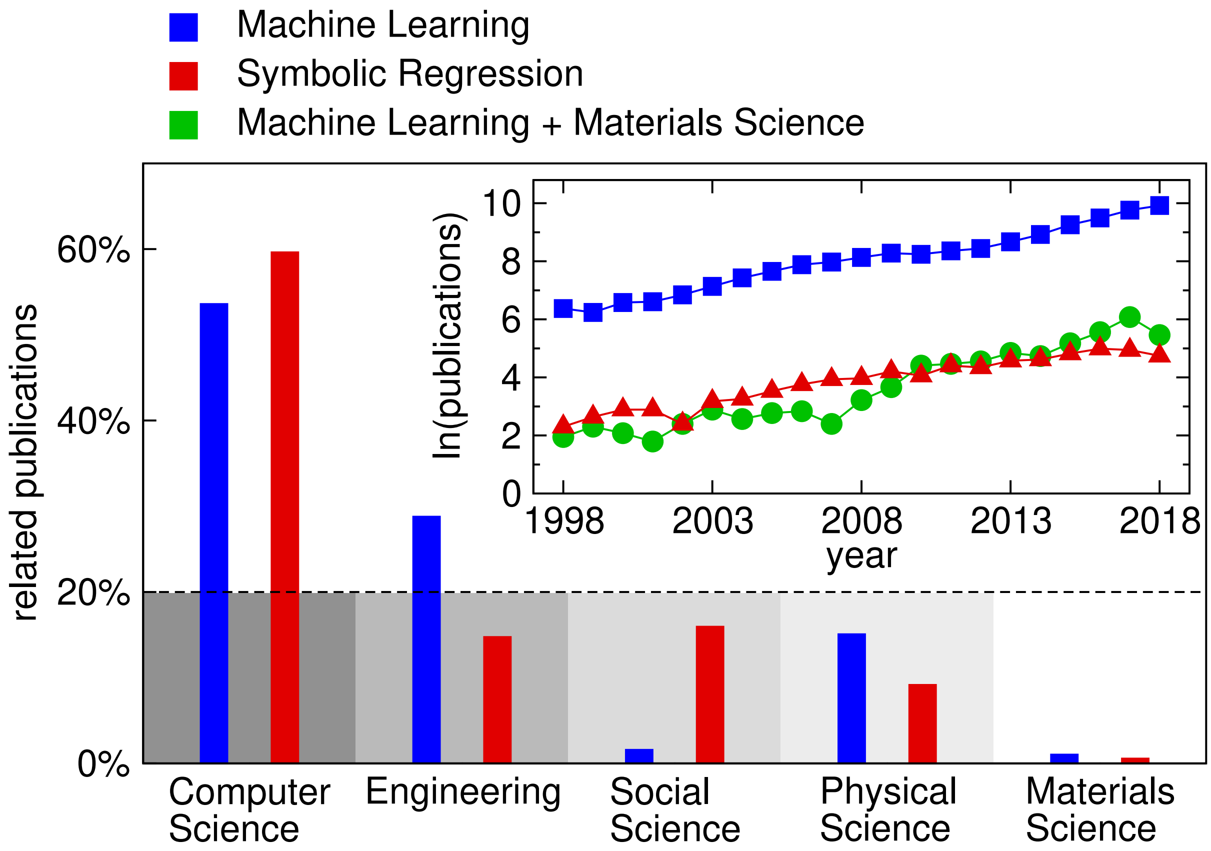


Figure 6.1: Relative contribution from different research domains to scientific journals related to machine learning and symbolic regression. Shaded panels indicate a 20% level of the research domain, emphasizing an opportunity in materials science. (inset) The trend in number of related publications on a natural logarithmic scale (ordinate) related to machine learning, machine learning and materials science, and symbolic regression, with respect to time.

regression, over 50% of the contributions come from the computer science research community, while multidisciplinary engineering is second. Social science and physical science each makes less than 20% of the contribution to the total number of publications. These two techniques are not so popular in materials science research, as the relative contribution is almost negligible compared to other research domains.

It is not surprising to see a dominant contribution from computer science in both the machine learning and symbolic regression communities, since it is where these techniques were born. It is interesting to notice that symbolic regression is relatively more popular than machine learning in social science research. One possible reason for this trend is that social science problems typically do not have a (known) physically motivated governing equation as in many physical sciences, where for example, Newtonian equations-of-motion, Schrodinger equation, etc. can be written formally. Symbolic regression arises naturally as a problem solver since it has the potential to find an appropriate functional form from social science data sets, e.g., questionnaire results, behavior patterns, etc.

We also report the trend in the number of publications (in a natural logarithm scale) in the following research domains [Figure 6.1(inset)]: machine learning, application of machine learning in materials science, and symbolic regression. All three domains exhibit a rapid (almost exponential) growth rate, whereas the number of machine-learning-related publications is orders of magnitude larger than the other two. The trend of symbolic regression applications in materials science is not shown here since the base number is too small; nonetheless, it also reveals a potential previously underappreciated research domain. For materials science problems, one is often also presented with the problem of unknown relationships among many variables. Symbolic regression presents an opportunity then to help in the formulation of structure-property relationships derived from these variables.

In this chapter, we encourage materials scientists and engineers to utilize symbolic regression techniques in solving their domain problems. To facilitate a better understanding of the utility and application of symbolic regression, we next introduce the genetic programming-based symbolic regression (GPSR) method and describe current research frontiers in symbolic regression. Next, we discuss several industrial applications of symbolic regression and propose potential uses in materials science. In addition, we present how GPSR can learn the Johnson-Mehl-Avrami-Kolmogorov (or Avrami) equation to describe recrystallization kinetics, as well as the Landau free energy expansion describing the structural phase transition in $\mathrm{LaNiO_3}$. Last, we conclude with some open challenges in materials research that may benefit from symbolic-regression methods.

6.2 Symbolic Regression and Current State-of-the-art Methods

6.2.1 Genetic programming-based symbolic regression (GPSR)

Symbolic regression is a method of finding a suitable mathematical model to describe observed data [169]. In conventional regression techniques, one optimizes parameters for a particular model provided as a starting point to the algorithm. For instance, a linear regression model is based on the assumption that the relationship of the dependent variables and regressor is linear [170]; an artificial neural network (ANN) is a nonlinear model which relies on a predefined network infrastructure such as neuron connections and activation function (e.g., sigmoid, softmax function).

In symbolic regression, however, no such a-priori assumptions on the specific form of the function is required. Instead, one provides a mathematical expression space containing candidate function building blocks, e.g., mathematical operators, state variables, constants, analytic functions, and then symbolic regression searches through the space spanned by these primitive building blocks to find the most appropriate solution. In other words, both model structures and model parameters are optimized in symbolic regression. Since there is no need for a predefined function form, optimization algorithms used in symbolic regression are different from conventional analytical/numerical optimization methods (e.g., conjugate gradient, Newton-Raphson method). In this section, we briefly introduce one of the most prevalent methods used in symbolic regression by means of genetic programming.

Genetic programming (GP) was developed by J.R. Koza [171] as a specific implementation of genetic algorithms (GA) [172], which are often utilized in the materials community for atomic structure prediction [173, 174, 175]. The idea is to evolve the solution of a given problem following Darwin's theory of evolution and to find the fittest solution after a number of generations. Instead of using strings of binary digits to represent chromosomes as in GA, solutions in GP are represented as tree-structured chromosomes with nodes and terminals. Figure 6.2a shows a chromosome example of the mathematical function $1 + \exp(-x_1)$. The tree consists of a set of interior nodes with

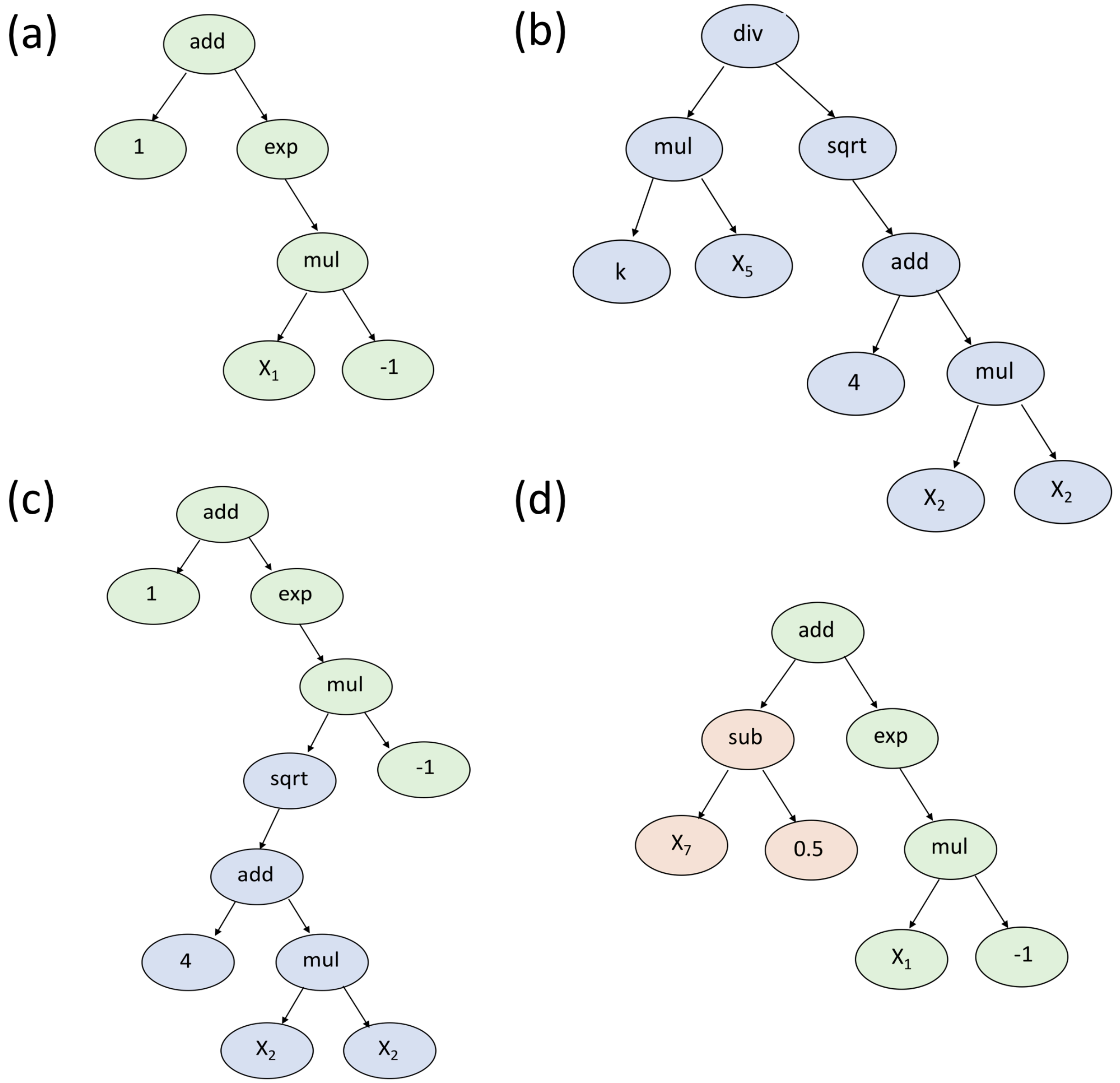


Figure 6.2: Tree-structure chromosome representation of computer programs in genetic programming. (a) parent1 $(1 + \exp(-x_1))$; (b) parent2 $(kx_5/\sqrt{x_2^2 + 4})$; (c) child of genetic crossover operation $(1+\exp(-\sqrt{x_2^2 + 4}))$; and (d) child of subtree mutation operation $(x_7-0.5+\exp(-x_1))$.

mathematical operations ($+$, $\times$, $\exp$) and terminal nodes with variables (x_1) and constants (± 1). A depth-first search can be used to traverse the tree to get the final mathematical expression of each individual solution.

The structure of a chromosome tree is not necessarily binary; its structure depends on the number of arguments the mathematical operator takes. For demonstration purposes, we only introduce simple operators that are either unary or binary. Users of GP could include a variety of functions suitable for their target problems. A large number of trees will be generated based on specified user settings and evaluated throughout the GP process. Each tree represents a potential solution of the problem. The way new trees are generated from the initial mathematical building blocks is a unique feature of GP since it mimics the natural evolution of Earth's ecosystem, i.e. through artificial sexual recombination and a natural selection process.

Figure 6.3 illustrates the process by which a solution of the symbolic regression problem is obtained using genetic programming. The procedure starts with a set of randomly generated initial terminal nodes (variables, constants) and functions, forming individual trees with different sizes and structures (Figure 6.2(a-d)). These fundamental building blocks come from a user-defined input set. This starting population typically has a large variety of tree structures due to the random process, which facilitates further exploration of the variable space and reduces the potential risk of being trapped in local minima. The initialization process terminates once the number of individuals reaches a user-defined population size, where the natural selection process then comes into play. The “fitness” of each individual solution in the initial population is then evaluated by comparing their function output with the true value from the data set. This fitness value describes how well the program performs in terms of solving the problem. The common error metrics used include mean squared error (MSE), root-mean squared error (RMSE), etc. Then GP evolves the current generation by randomly applying genetic operations to individuals, e.g., crossover and mutation. One or more individuals from the current generation will be selected as parent(s) based on the fitness score, typically the higher the score, the larger probability to be selected for reproduction. Such

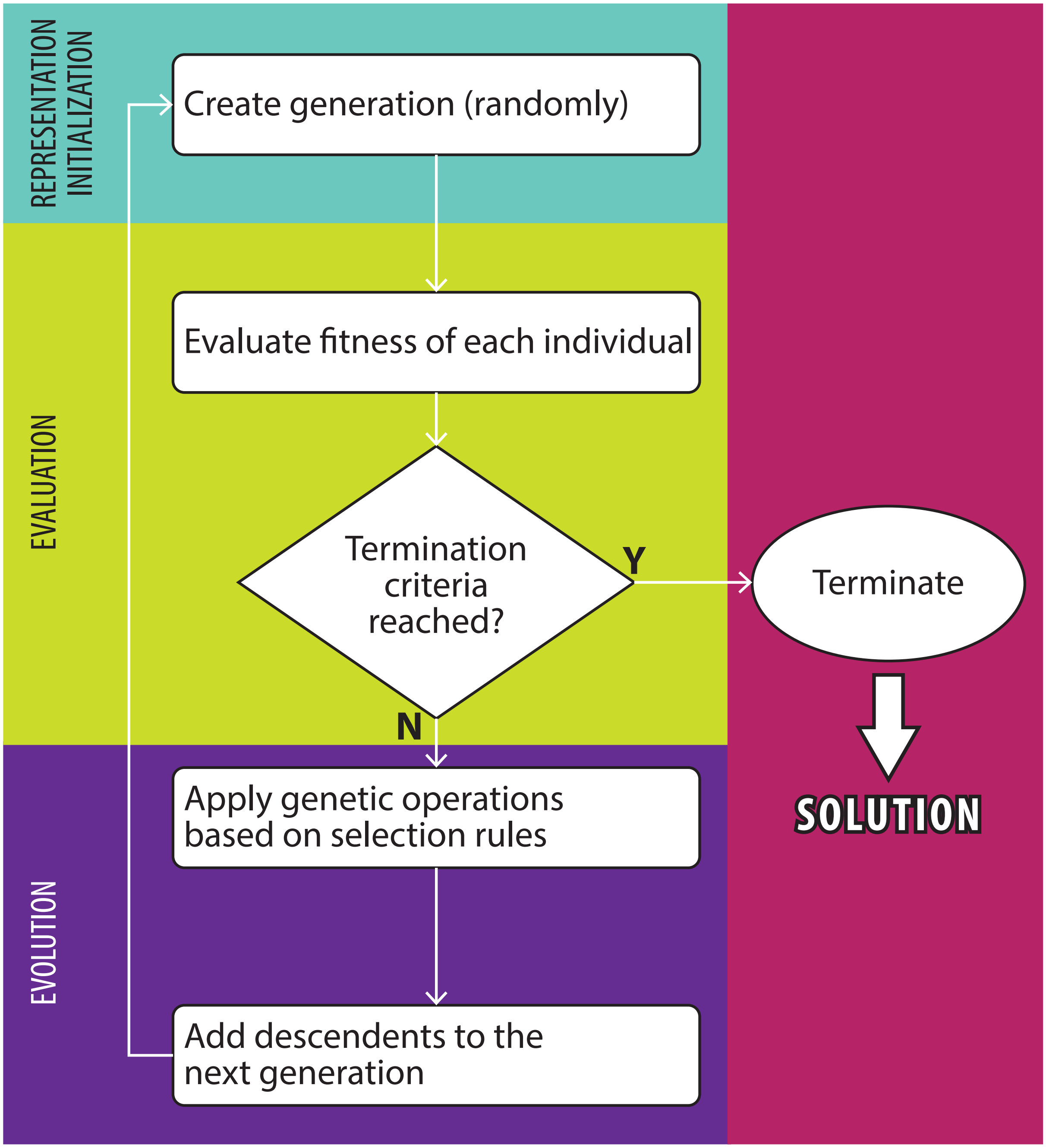

REPRESENTATION
INITIALIZATION
Create generation (randomly)
EVALUATION
Evaluate fitness of each individual
Termination criteria reached?
Y
Terminate
N
SOLUTION
EVOLUTION
Apply genetic operations based on selection rules
Add descendents to the next generation

Figure 6.3: Genetic programming flowchart depicting the iterative solution-finding process.

a selection rule agrees with the “survival-of-the-fittest” rule since good features are more likely to be inherited by the next generation, which is the essential step towards the optimal solution.

The genetic crossover operation takes two winners of the selection process as parents to breed their offspring. For instance, the two structures in Figure 6.2(a) and (b) are taken as parents. The crossover operator then randomly takes a subtree from parent (b) and substitutes another random subtree in parent (a) with that from (b). One possible offspring from the crossover operator is illustrated in Figure 6.2(c). Crossover is usually the dominant operation in the recombination process. Figure 6.2(d) is an example of an offspring from the mutation operation. The mutation operator only takes one parent structure, and randomly substitutes a subtree with another randomly generated structure; in case (d), the constant 1 is mutated to $(x_7 - 0.5)$. Although this operation is more aggressive compared to the crossover operation, since it adds randomness to the system, it is important to have a finite chance of mutation to introduce new variations, e.g., new constants and new features, and avoid being trapped in local minima.

The third category of genetic operations is reproduction, which duplicates the selected program and directly inserts its offspring to the next generation. It guarantees that some of the current generation will be preserved by the next generation, and partially protects the similarity between two generations. Detailed definitions and implementations of each genetic operation can vary from case to case, but the main features should be the similar to what we described here.

The newborns are then added to the next generation after each genetic operation, until the new population size reaches the specified set number. Then the new generation goes through the fitness evaluation and natural selection process again until the fitness value reaches a certain criteria or the maximum number of generations is reached. After termination of GP, the surviving individuals are expected to be highly evolved to adapt to the problem-dependent selection rule. More comprehensive descriptions of GP can be found in Koza’s original paper [171].

6.2.2 Advances in symbolic regression

Since Koza introduced the idea of GP in 1992, there have been significant efforts made to improve the performance of the original GPSR algorithm. The major problems to overcome in GPSR include:

(i) *Non-deterministic optimization.* It is not guaranteed that the performance of the descendent generations will be better than their parents.

(ii) *Difficulty in finding the proper constants.* Since the way GP generates constants is random, either in the initial input set or those brought into the population by mutations, there is no effective way to obtain the ideal coefficients as in other numerical regression methods.

(iii) *Limited capability to preserve good components of the equation* due to the fitness evaluation method. The fitness is evaluated based on the complete structure of an individual. Having a good feature in a subbranch does not necessarily lead to better individual performance, thus good equation components may get lost in the next generation.

We summarize some of the most popular alternative methods to conventional GPSR in Table 6.1 and discuss their similarities as well as the differences in four aspects, namely program representation, fitness evaluation, optimization method, and the solution form.

Multiple regression genetic programming (MRGP) [176] improves the program evaluation process by performing multiple regression on subexpressions of the solution functions. Instead of evaluating the fitness of each individual solution as a whole, MRGP decouples its mathematical expression tree into subtrees. The fitness of the solution is evaluated based on the best linear combination of these subtree structures. A least angle regression (LARS) algorithm is used to solve the linear regression problem here. Such a fitness evaluation scheme places more emphasis on finding good components even though it might only be a partial solution. For instance, the individuals that contain a correct form of a subtree structure of the correct solution (if known) are more likely to survive the natural selection process and pass these good features to the descendents. MRGP

Table 6.1: Comparison of genetic programming-based symbolic regression (GPSR) and its alternative methods: multiple regression genetic programming (MRGP), geometric semantic genetic programming (GSGP), Cartesian genetic programming (CGP), genetic programming-based relevance vector machine (GP-RVM), evolutionary polynomial regression (EPR), and fast function extraction (FFX).

	Program representation	Fitness evaluation	Optimization	Solution form
GPSR	rooted-tree	individual program	genetic evolution	rooted-tree
MRGP	rooted-tree	subexpressions	genetic evolution + linear regression	linear combination of subexpressions
GSGP	rooted-tree	distance in semantic space	genetic evolution in semantic space	rooted-tree/ semantic vector
CGP	acyclic graph	individual program	genetic evolution	2D grid of nodes
GP-RVM	rooted-tree	group of GP individual	genetic evolution + RVM	linear combination of GP individuals
EPR	vector of integers	individual program	genetic evolution + linear regression	polynomial function
FFX	basis functions	individual program	pathwise regularized learning	linear combination of basis functions

essentially decouples the current basis functions to find the best solution in an enlarged space at the vicinity of the original GP space. Indeed, this feature may be well-suited for multi-scale materials problems where modeling of systems across different length/time-scale is desired [177]. While some subexpressions capture relationship among variables within each scale, the final symbolic regression solution assembles models across the scale and returns the multi-scale model.

Geometric semantic genetic programming (GSGP) [178] evaluates the semantic performance of a computer program instead of the syntax performance as in conventional GPSR. While still using a rooted-tree structure to represent computer programs, GSGP focuses on its semantics, i.e. the behavior of a program. For instance, $add(x_1, x_1)$ is equivalent to $mul(2, x_1)$ in semantic space, but quite different in terms of syntax. It is reasonable to care more about the behavior of the program than how the function appears. By representing each program in a high-dimensional semantic space, the fitness evaluation is rather straightforward; one only needs to measure the distance of the program from the target point in that space. The closer a program is to the target point, the better performance it has in solving the given problem. Interestingly, the offspring of two parent vectors in semantic space lies between its parents in the semantic space; therefore, the offspring should be at least no worse performing than the poor-performing parent. Optimizing program semantics rather than syntax further frees symbolic regression from specific function forms, potentially making SR more efficient [178].

Cartesian genetic programming (CGP) [179] has a more sophisticated design than conventional GP. Here, a computer program is represented as a directed acyclic graph, which may be visualized as a two-dimensional grid of nodes. Each node owns a set of genes that determines the input-output and mathematical function that the node performs; the whole set of genes of the computer programs form its genotype. Decoding the genotype leads to the phenotype, i.e., the function form of the computer programs. The genotype-phenotype mapping is a unique feature of CGP which makes it closer to the real natural process.

GP-RVM [180] is an alternative GP method that combines Kaizen programming and a relevance vector machine (RVM) algorithm to solve symbolic regression problems. Kaizen programming (KP) is a collaborative version of genetic programming, where individuals work together with each other to solve the problem. The solution of a Kaizen process is a linear combination of GP individuals, and thus the fitness evaluation is based on a group of individual partial solutions instead of an individual program as a complete solution. RVM is a Bayesian kernel method that could extract important basis functions from the basis set without the prior knowledge to set a threshold and automatically deals with singularity. GP-RVM leverages advantages from both evolutionary algorithm and Bayesian kernel methods: the former mainly explores the parameter space while the latter extracts basis functions to build and solve for the optimal solution function within that space.

Evolutionary polynomial regression (EPR) [181] hybridizes the parameter estimation used in conventional numerical regression methods with the evolutionary optimization scheme in GPSR. EPR first explores the function space using genetic algorithms, then performs linear regression (e.g., least squares) to optimize the coefficients of each mathematical building block. Although EPR specifically uses polynomial expansions for the form of the functions, the solution is not necessarily a simple polynomial function since the transformed variables used in the polynomial expansion could be nonlinear functions of independent input variables. Such a hybrid method improves the stochastic GPSR method moving it towards a more deterministic approach although the computational cost may be relatively higher. In fact, the polynomial form of the expressions could make EPR suitable for materials design or multiobjective optimization purposes. Since the analytical gradient and Hessian of the solution can be evaluated, materials scientists may have more insights regarding the system and know what parameters to tune in order to achieve optimal design.

Fast function extraction (FFX) [182] is an efficient way to find good basis functions and solve for the best solution within the space it spans. The first step in FFX is to generate a large number of candidate basis functions built from input variables and other predefined variables. The evolutionary optimization scheme is not involved in FFX, instead, a pathwise regularized learning technique is used to identify the best coefficients and basis functions for the solution. Then, models obtained from the previous step are assessed based on the validation data set as well as their model complexity in order to identify the best solution. FFX is more efficient compared to other GP-based methods due to the deterministic optimization technique. Materials scientists could first use FFX to see whether the input function/variable basis is sufficient for their research problem, before further investigation using symbolic regression methods (either FFX or other variants).

The performance of some of the recently developed symbolic regression techniques has been assessed against popular machine learning methods [183], and it is reported that symbolic regression performs considerably well compared to state of the art ML algorithms with regards to predictive accuracy. However, the two methods do not simply exist in competition to one another. We also observe a trend of more hybridization between conventional ML algorithms and genetic programming in symbolic regression solvers [184, 185, 186, 187]. These advances have enabled symbolic regression to be used for solving real-world problems, which we will discuss in the following section.

6.3 Applications of Symbolic Regression

Although it seems that equations obtained from first principles (e.g., the Schrödiner equation) and empirical observations (e.g., the 18-electron rule [188]) are quite contradictory to each other, we see quite often that they symbiotically work together in solving real-world problems. For instance, both *ab initio* and experimental data have been used to develop effective interatomic force fields [189] or exchange-correlation functionals [190]. In fact, symbolic regression has the potential to serve as the bridge connecting experimental data to first principles. Schmidt *et al.* demonstrated that symbolic regression is capable of predicting connections between dynamics of subcomponents of the system and distill natural laws from experimental data [191]. Moreover, symbolic regression provides researchers with analytic equations, which expectably would have better interpretability over the raw data and potentially other black-box models. The equations could reveal how the dependent variable (system output) responds to multiple independent variables (system input), as well as the relationships between independent variables of the underlying function. We show this later in Section 6.3.3.

Common motivations underlying the use of GPSR for complex problem solving include when the system in question is not effectively modelled by a linear model. Existing multiple linear regression models are much faster and are already easy to interpret. GPSR is best used for systems with complex interactions between observable variables for which the form of which is not known beforehand—a situation common in materials science and engineering.

In addition, a GPSR approach could be useful for design optimization purposes. Although the evolutionary search process is a black box, the final solution is analytical, which potentially contains important information (e.g., regarding the gradient or Hessian) about relationships between the design variables and objectives. There is also need for multi-objective optimization such as finding the Pareto optimal combination of model performance and complexity in various domains—it is here that the symbolic regression technique has shown to be effective and interpretable [192]. We next describe some applications of symbolic regression in various science and technology domains.

6.3.1 Industrial applications

GPSR has been applied to a wide variety of problems in fields outside of materials science and chemistry. Most prominently featured in the popular press was work published by Schmidt and Lipson in *Science* [191], which showed GPSR could discover Hamiltonians and Lagrangians for systems of simple harmonic oscillators and double pendulums. Reports of using GPSR for real world systems, however, have been published since Koza's origination of the idea in the early 1990s and continue today.

Arkov *et al.* [193] used GPSR to identify equations governing gas turbine engines under multiple optimization conditions. Berardi *et al.* [194] used GPSR to find easy to interpret models for pipe failures in a UK water distribution system. Bongard and Lipson [195] applied GPSR to generate symbolic equations for nonlinear coupled dynamical systems in mechanics, ecology, and systems biology. The authors also emphasized that their symbolic models are easier to interpret than numerical models, which makes understanding more complex systems easier for future applications.

Cai *et al.* [196] identified correlation equations from experimental heat transfer measurements using GPSR with a sparsifying constraint. The authors' predicted correlations had lower percentage error than models developed graphically and numerically, albeit with more formula complexity than those traditional methods. Can and Heavey [197] applied GPSR to develop metamodels for predicting throughput rates in industrial serial production lines. McKay, Willis and Barton developed steady-state models for a vacuum distillation column and a chemical reactor [198].

La Cava *et al.* [199] applied GPSR to identify nonlinear governing equations of wind turbines. The Pareto front from their paper is reproduced in Figure 6.4. The Pareto front illustrates the trade-off between their model complexity as defined by the number and type of operations in the equation and the normalized variance in the prediction error. La Cava and other authors [200] also tested modifying standard GPSR with features from epigenetics, such as passive structure, phenotypic plasticity, and inheritable gene regulation. These researchers demonstrated their modifications improved the performance over standard GPSR by finding compact dynamic equations for synthetic data from nonlinear ordinary differential equations as well as real-world systems, e.g., cascaded tanks, a chemical distillation tower, and an industrial wind turbine. GPSR has also been applied to testing the efficient market hypothesis [201], formulating the synchronization control in oscillator networks [192], identifying the structure of helicopter engine dynamics [202], real-time runoff forecasting in France [203] and Singapore [204], designing circuits [179], predicting solar power production [205], finding dynamical equations for metabolic networks [206] in both cases where a starting model was known and from scratch, modelling global temperature changes [207],

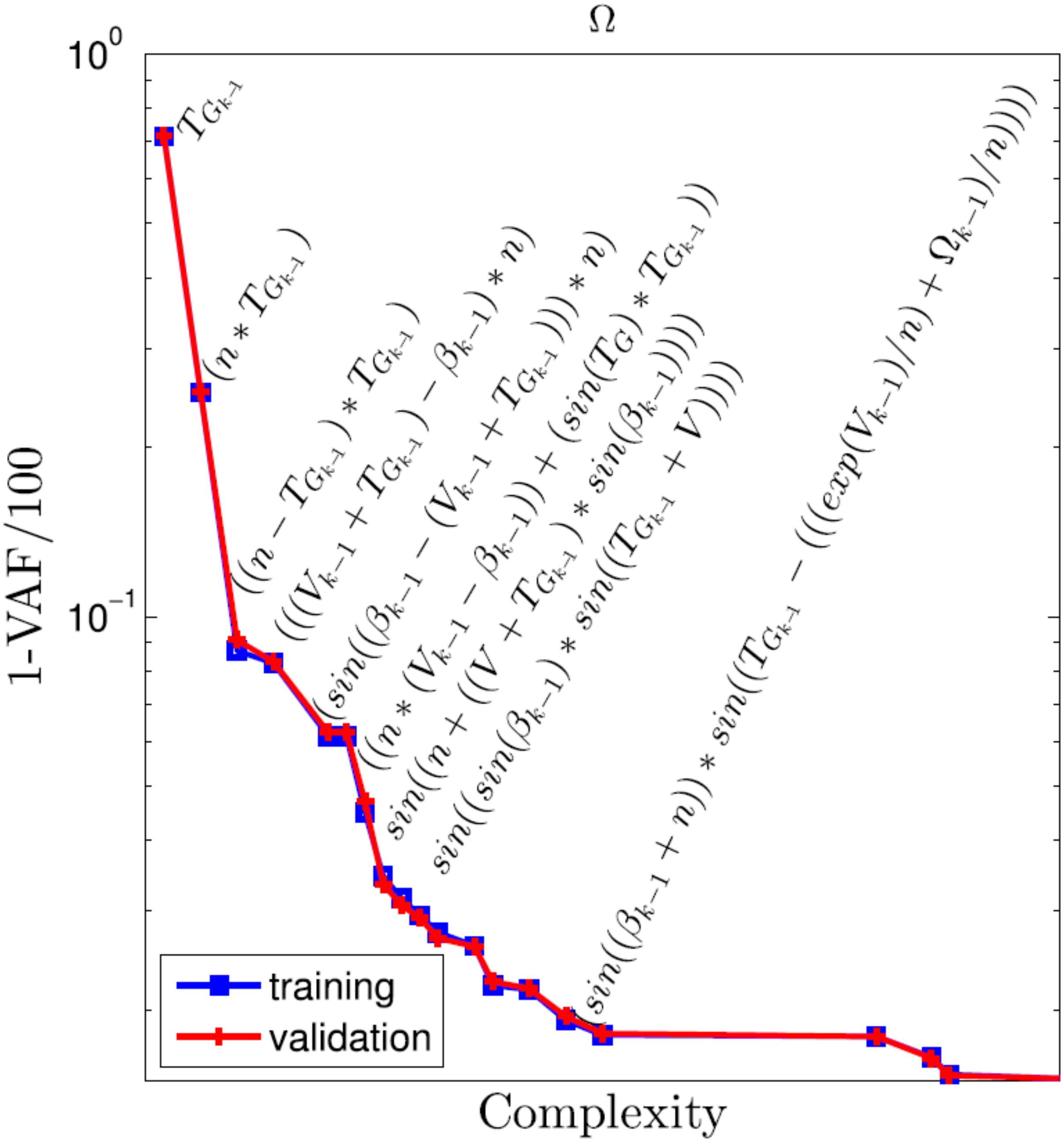


Figure 6.4: Example Pareto front showing trade-off between solution complexity and variance accounted for (VAF). Reproduced with permission from La Cava *et al.* [199].

and synthesizing second-order coefficient insensitive digital filter structures [208].

The existing uses of GPSR within chemistry are more extensive than that for materials science. We refer the reader to the review by Vyas, Goel, and Tambe [209] for further details. Some key studies with relevance to materials science are summarized here: Langdon and Barrett developed a model for oral bioavailability of a small molecule given a few hundred data points from expensive experiments [210]. Their model based on chemical descriptors showed promise for rapid drug screening, but had difficulty generalizing to novel molecules. Vyas *et al.* [211] also discovered structure-property relationships for drug absorption using GPSR. Here, they demonstrated R^2 values comparable to those achieved with artificial neural networks and support vector regression. Barmpalexis *et al.* [212] performed a multiobjective optimization using GPSR. They found a function mapping levels of 4 polymers to three different properties of a pharmaceutical release tablet that was more predictive than a shallow neural network. Last, Muzny, Huber, and Kazakov built a correlation model for the viscosity of hydrogen as a function of temperature and pressure [213].

6.3.2 Opportunities in materials science

Materials science has many potential areas where GPSR can be applied for the same reasons it find use in other disciplines. Nonlinear systems are abundant in materials science. Changes in materials properties occuring in response to structural, composition, and other external perturbations are frequently nonlinear in proximity to phase transitions or for large stimuli. For instance, changing the concentration of oxygen vacancies in a transition metal compound by an atomic percent can alter its ionic or electronic conductivity by orders of magnitude [214, 215]. The dynamical behavior of materials performance as a function of time is also of broad interest and technological importance, e.g., corrosion of nickel cathodes under different conditions [216]. These are the areas where a dynamical multivariable model would be ideal to understand the correlation among the variables and assist optimization of materials properties, e.g., corrosion resistance.

Frequently, materials scientists look for relationships (f) among multiple variables with the aim to find some closed-form expression such as $y = f(X)$, where y is the objective value and X are a set of variables. These equations are typically expressed in differential form, e.g., the Schrödinger equation ($i\hbar\frac{d}{dt}\left|\Psi(t)\right\rangle = \hat{H}\left|\Psi(t)\right\rangle$) or Newton's second law ($\mathbf{F} = m\frac{d\mathbf{v}}{dt}$). It has been shown that symbolic regression can generate ordinary nonlinear partial differential equations for nonlinear coupled dynamical systems [195, 217] as well as approximate ordinary differential equations [218]. Meanwhile, it is also often of desire to find conservation laws in physical systems. The ability to unearth conservation laws with symbolic regression goes beyond the aim of materials property predictions and helps researchers establish insight into the materials systems they study [191, 219]. In fact, we do not necessarily need a rigorous expression of natural laws in every case; sometimes an approximation with a simple yet effective expression serves well for the research purpose [220]. Symbolic regression could potentially balance the trade-off between model accuracy and simplicity, and might even help scientists discover new equations that redefine our understanding of functional materials in the same way those of Hall and Petch and Harper and Dorn changed our understanding of the mechanical properties of metals or as more recently how Berry phases and topological band theory changed our understanding of electronic structures.

As we mentioned earlier, materials properties and performance are affected by phenomena that involve multiple length scales. Most theoretical models are formulated to be optimal at a particular length scale. However, recent emphasis has been placed on multiscale and hiearchical modeling in the materials science community [221, 222, 177], and there is an increased need for effective, descriptive and predicative, multiscale models. Symbolic regression techniques are potential solutions to this challenge by directly searching for the interactions among variables operating and passing between multiple spatial and temporal scales. Another possible approach is to utilize existing simulation methods within each length scale, while using symbolic regression to find the suitable coupling interactions between scales, i.e. connecting models of different scales.

Other applications of GPSR in materials and molecular systems are in areas where supervised machine learning has already demonstrated usefulness in providing new insight or solutions. While machine learning has produced many impressive results [166, 159, 223], it is commonly understood that ML models exhibit a trade-off between performance on prediction metrics and the ability to explain the predictions of a model due to the complexity of state-of-the-art models like deep neural networks or gradient boosted decision trees. GPSR offers a middle ground with comparable performance but with the added ability to read and directly interpret the output function.

GPSR is also well-suited to the development of new descriptors [224] for materials properties. By combining features in a manner best suited to fitting data, new features are created that can be used as proxies for the property in question. This is also a common application of compressed sensing [225, 226]. Compressed sensing differs from GPSR in that while GPSR uses GP to iteratively evolve a solution, compressed sensing tries to enumerate as many combinations of primary features as possible and then use sparsifying operators to find a small dimensional subset that correlates with the target.

Materials scientists are not just interested in making predictions. They also want to identify the controlling features of a property and what role each feature plays; they want to understand which degrees-of-freedom to design or optimize to achieve targeted properties. Given some desired properties as objectives and the relevant variables, there are a number of numerical algorithms available for performing optimization [227], but they typically require some a-priori knowledge of the mathematical relationships within the system. The symbolic equations derived from GPSR offer insight into which microscopic or macroscopic knobs to turn for the design of desired functionality, such as corrosion resistance in steels [228].

Some novel ideas include targeting physical variables for which we do not know the proper mathematical expressions. For instance, the exchange-correlation functional used in density functional theory, or the correlation function for the viscosity of normal hydrogen [213]. Contraindicated material property pairs such as ferromagnetism and ferroelectricity or optical transparency and electrical conductivity would be interesting areas to pursue in search for routes to decouple or circumvent perceived coexistence incompatibilities. In these cases, GPSR could provide candidate representations of these terms, where our physical knowledge can be used to filter meaningful results. One advantage of using symbolic regression is that the balance between model accuracy and complexity is tunable with GPSR settings, e.g., stopping criteria, penalty on individual size, etc. This advantage is particularly evident when designing with constraints or optimizing for performance. The recommendations coming from the learned symbolic equations are more actionable than learned functions only optimized for test set accuracy since they are rigorously made to use fewer terms.

6.3.3 Use cases in materials science

Discovering the Johnson-Mehl-Avrami-Kolmogorov equation

We now show how to use genetic programming-based symbolic regression to learn the Johnson-Mehl-Avrami-Kolmogorov (JMAK), hereafter, Avrami equation. The Avrami equation quantitatively describes the growth kinetics of phases in materials at constant temperature. Here, we specifically study the recrystallization process of copper, the original experimental data was obtained from Ref. [229]. We expect the form of the function to be

$$y = 1 - \exp(-kt^n)\,,$$

where the phase fraction of transformation y is a function of time t. The coefficients k and n are unknown and change with respect to temperature and other environmental conditions.

We use GPSR as implemented in `gplearn` [230] to predict the relationship between fraction transformed y and time t. The hyper-parameters used in GPSR are listed in Table 6.2. When the population size is divided by the tournament factor, one obtains the number of individuals competing for reproduction each round. The parsimony coefficient regularizes the size of individuals by penalizing over-sized structures. We include operations of addition, subtraction, multiplication, negation, and the natural exponential function into the function set. The power function is not

Table 6.2: List of hyper-parameters used in GPSR to learn the Avrami equation. Grid search method is used to find the optimal hyper-parameters from the top three parameter sets.

Parameter	Values
population size	{2000, 5000}
tournament factor	{100, 500}
parsimony coefficient	{0.001, 0.005}
max generation	20
constant range	(-1, 1)
function set	{add, sub, mul, neg, exp}
crossover probability	0.7

included since it easily causes numerical overflow or invalid operations (e.g., power(-1, 0.5)). In general, this can be an issue when evaluating power-law dependent phenomena such as electrical transport equations. Working with log transforms of the original variables may be a more stable approach. Crossover operations dominate the genetic operations with a 70% probability to be applied; the other 30% chance corresponds to mutation operations (e.g., point mutation, subtree mutation, etc.). Additional details concerning the usage of `gplearn` are given in its documentation.

Data preprocessing is an essential step in machine learning before feeding data into the solver. Conventional preprocessing methods include shifting data to be zero-centered, and scaling data to unit standard deviation. However, the conventional preprocessing methods are not ideal choices in our case. Zero-centered shifting is not applicable to either the phase fraction transformed (y) or time (t) since we want to obtain the exact function form of the Avrami equation. Furthermore, scaling the time frame would only change the constant k in the Avrami equation. Here, we scale time from 0 to 10 for all data sets so that the constant t lies in the range [-1, 1]. The y values remain unchanged as the experimental data range over [0, 1].

We take the experimental copper recrystallization data at temperatures 135°C, 113°C, and 102°C as input and scale the time variable before performing regression. Since the experimental data only contains several points at each temperature (less than 10), we also take data points from interpolated lines and perform symbolic regression on the interpolated data for comparison.

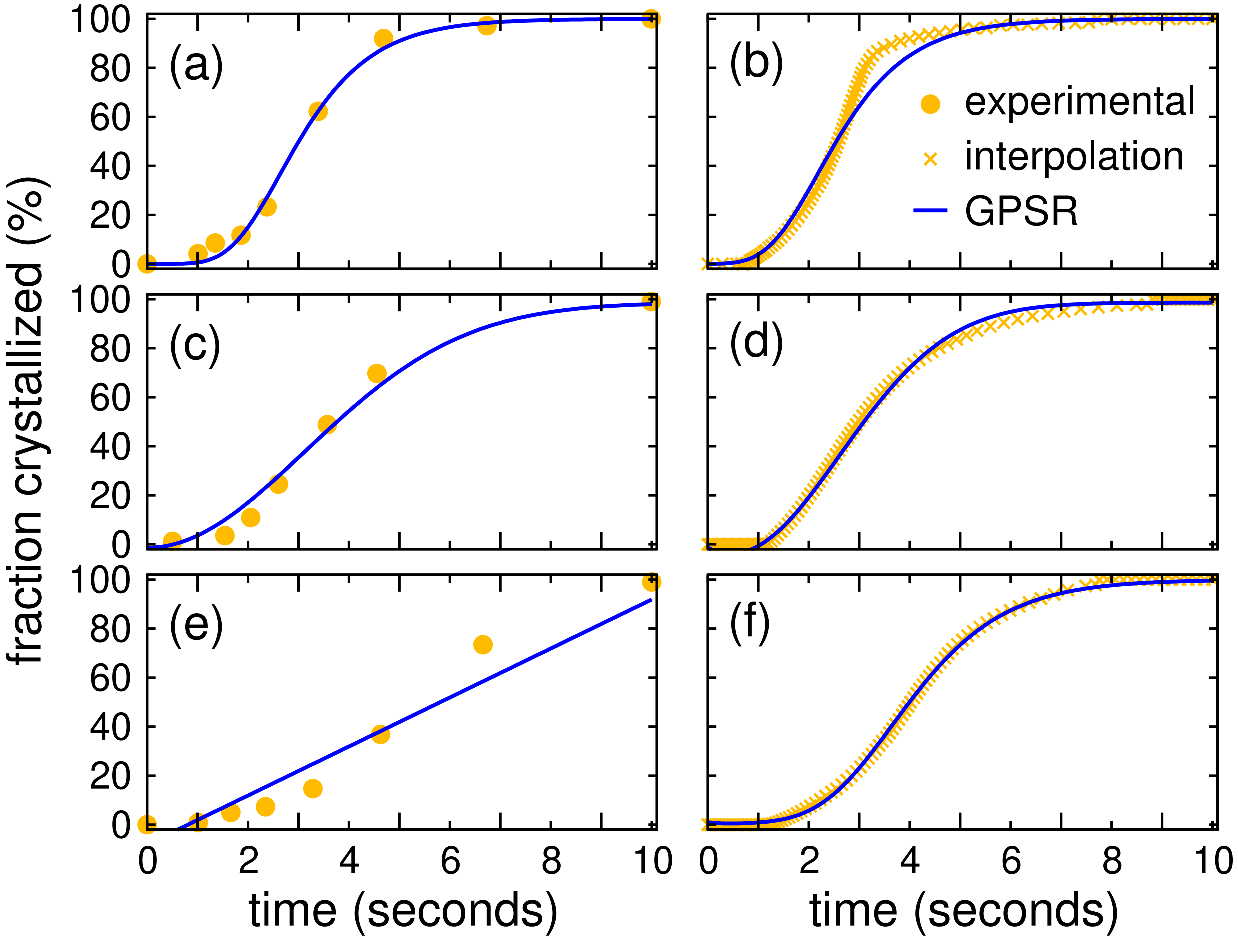


Figure 6.5: GPSR prediction and performance with different data sets. Left panels correspond to the direct experimental data while right panels use interpolated experimental data (a) 135°C experimental data, (b) 135°C extrapolated data, (c) 113°C experimental data, (d) 113°C extrapolated data, (e) 102°C experimental data, and (f) 102°C extrapolated data.

An ideal dataset is also generated, where y values are directly calculated as $1 - \exp(-0.6t^2)$. The optimal k and n values in each data set are obtained from numerical fitting using `Scipy`, given the form of the Avrami function. Ideally GPSR should be able to recover the correct function form as well as k and n constants for all data sets.

The best individual after 20 generations of evolution is collected for each hyper-parameter setting. Finally we manually pick the optimal individual with closest function form and constants to the Avrami equation within each data set. Our parameter fitting and GPSR evolution results are shown in Table 6.3 and Figure 6.5.

Table 6.3: Functions predicted by GPSR and numerical fitting results for the Avarmi equation describing copper recrystallization.

Dataset	Numerical fitting result	GPSR result	GPSR (transformed y) result
ideal	$y = 1 - \exp(-0.6t^2)$	$y = 0.994 - \exp(-0.58t^2)$	$1 - y = \exp(-0.58t^2)$
135°C experimental	$y = 1 - \exp(-0.019t^{3.2})$	$y = \exp[-\exp(-t + 2.64)]$	$1 - y = \exp(-0.024t^3)$
135°C interpolated	$y = 1 - \exp(-0.17t^{3.6})$	$y = \exp[-\exp(-t + 2.18)]$	$1 - y = \exp[-0.065t^2(t - 0.851)]$
113°C experimental	$y = 1 - \exp(-0.018t^{2.8})$	$y = 0.986 - \exp(-0.051t^2)$	$1 - y = \exp(-0.014t^3)$
113°C interpolated	$y = 1 - \exp(-0.042t^{2.5})$	$y = 0.986 - \exp[-0.108 \times t \times (t - 0.934)]$	$1 - y = \exp[-0.097t(t - 0.764)]$
102°C experimental	$y = 1 - \exp(-0.0054t^{2.9})$	$y = 0.1t - 0.081$	$1 - y = \exp(-0.0045t^3)$
102°C interpolated	$y = 1 - \exp(-0.01t^{3.1})$	$y = \exp[-(1.11 + 2t) \times \exp(-t + 1.42)]$	$1 - y = \exp(-0.01t^3)$

`Gplearn` successfully recovers the relationship between time (t) and fraction transformed (y) in most cases. With the ideal input data, `gplearn` evolves almost a perfect form of the Avarmi function as well as the constants (see the first row of Table 6.3). The performance on the raw experimental and interpolated data are generally worse than the ideal case but still reasonable. In most cases the exponential function form and polynomial function of t are both recovered. One source of error is the lack of the power function in function set, which was intentionally omitted as non-integer powers of variables cannot be correctly represented here. Rather, polynomial functions are used as an approximation. With the limited choice of mathematical functions, GP produces results that are very close semantically but exhibit different syntax.

We see in multiple cases GPSR produces functions of the $\exp[-\exp(\Theta)]$ form, where the expected function form is $1 - \exp(\Theta)$. Here we introduce a mathematical trick to show their equivalence. The functions e^x and $1 + x$ are called equivalent infinitesimals because

$$\lim_{x \to 0} e^x \sim 1 + x \, .$$

In our case, the exponent $-\exp(\Theta)$ quickly reaches zero when Θ (ideally having the form $-kt^n$) becomes more negative; therefore, the equivalent infinitesimal relationship holds and the predicted form of the function is equivalent to the expected one appearing in the Avrami expression. Based on this example, we encourage professional materials researchers, who are novice data scientists, to perform careful analysis of GPSR results when making the final interpretation of the obtained model.

In real-world applications where the actual function form in unknown, the exact syntax of GPSR results may not matter. There are potentially many cases where GPSR would produce semantically similar/equivalent functions that vary in their syntax, i.e. different function form. Considering the trade-off between model accuracy and complexity, in some cases it might be a virtue to find a simple approximated solution instead of using rigorous but complex relationships among multiple variables.

The GPSR result from the 102°C experimental data set (Figure 6.5(e)), turns out to be a linear relationship and deviates significantly from the original data. The poor result may be caused by the relatively small slope of the original data and the insufficient number of available data points. Taking the model performance and complexity into consideration, we find that the linear relationship survives due to its simple form. With interpolated data as input, functions evolved from GPSR agree better with the Avrami equation, as shown in Figure 6.5(f). Therefore, having more training data could improve the performance of GPSR, which also applies to other data-driven methods.

Interestingly, the performance of GPSR can also be improved by transforming (or simplifying) the mathematical expression. We compare results of directly training $y \sim t$ relationships with $(1-y) \sim t$, where in the latter case the target value $1-y$ is the percentage of copper untransformed. The results are shown in the last column of Table 6.3: The transformed functions show improved performance since we have already performed the subtraction function for the model. GPSR not only successfully recovers the exponential form of the equation, but also finds constants closer to the numerical fitting values.

Learning Landau free energy expansion

Next, we present a slightly more complicated case with two variables. The model we studied is the Landau free energy expansion for the cubic-to-rhombohedral structural phase transition in perovskite $\mathrm{LaNiO_3}$, where the free energy G of the system is expanded in powers of an order parameter θ as

$$G(\theta, T) = G_0(T) + \kappa(T - T_C)\theta^2 + \lambda\theta^4 \,, \tag{6.1}$$

where κ and λ are temperature-independent coefficients and θ is the angle of rotation about the [111] direction. This rotation angle of the corner-connected $\mathrm{NiO_6}$ is the order parameter for the displacive transition. The parameters we used are obtained from *ab initio* DFT simulations [231], where $\kappa = 1.696\,\mathrm{meV}\,10^{-3}\,\mathrm{K}/(^\circ)^2$, $\lambda = 0.0171\,\mathrm{meV}/(^\circ)^4$, and T_C is estimated to be 2.057 $(10^3\,\mathrm{K})$. We use $10^3\,\mathrm{K}$ as the unit for temperature so that the constants are brought into a smaller range.

In this case, we set the temperature T and order parameter θ as the input variables, and the free energy change $G(\theta, T) - G_0(T)$ as the output. We uniformly sampled 11 temperature points between [0, 1] (10^3 K), and 100 order parameter points within range [-20°, 20°]. The corresponding value for the change in free energy is calculated from Equation 6.1 using the parameters reported in the literature. A population size of 10,000 is used, with tournament size 25, parsimony coefficient 0.02, and constant range [-2.0, 2.0]. In order to simplify the problem, we only consider addition, subtraction, and multiplication operations in our search. Other settings are the same to those in the Avrami case.

The best individual after 15 generations of evolution has the function form

$$G(\theta, T) = G_0(T) + 1.983(T - T_C^{'})\theta^2 + 0.0165\theta^4 + \xi\,, \tag{6.2}$$

where $T_C^{'} = 1.894 + 8.32 \times 10^{-3}T^2$, and $\xi = (-1.72T^2 + 1.214T - 0.24)\theta - 1.334$. The GPSR-learned coefficients are quite close to the reported values, especially for the quartic term. This is probably because the penalty for a larger deviation in the leading (quartic) term is much higher than others. For the quadratic term of θ, GPSR successfully captured the coupling term $T\theta^2$ and its coefficient, but also an unexpected biquadratic $T^2\theta^2$. It should not have a strong impact on the function owing to its small coefficient (8.32×10^{-3}).

The Landau free energy with respect to the order parameter is plotted in Figure 6.6. We find that GPSR results (filled symbols) agree well with the DFT-derived Landau free energy function not only within the training region, i.e., with $T = 0, 500$ and $1,000\,\mathrm{K}$, but also reasonably well beyond it. The dashed red line in Figure 6.6 with $T = 3,000\,\mathrm{K}$ is not included in training the GPSR, yet the model reproduces both the shape and the correct global minimum position very well. However, the predicted function contains other coupling terms not present in Equation 6.1. These extra terms in ξ destroy the symmetry of the function, i.e., the free energy expansion is an even function by symmetry. This would become an issue especially when T is large, as we can

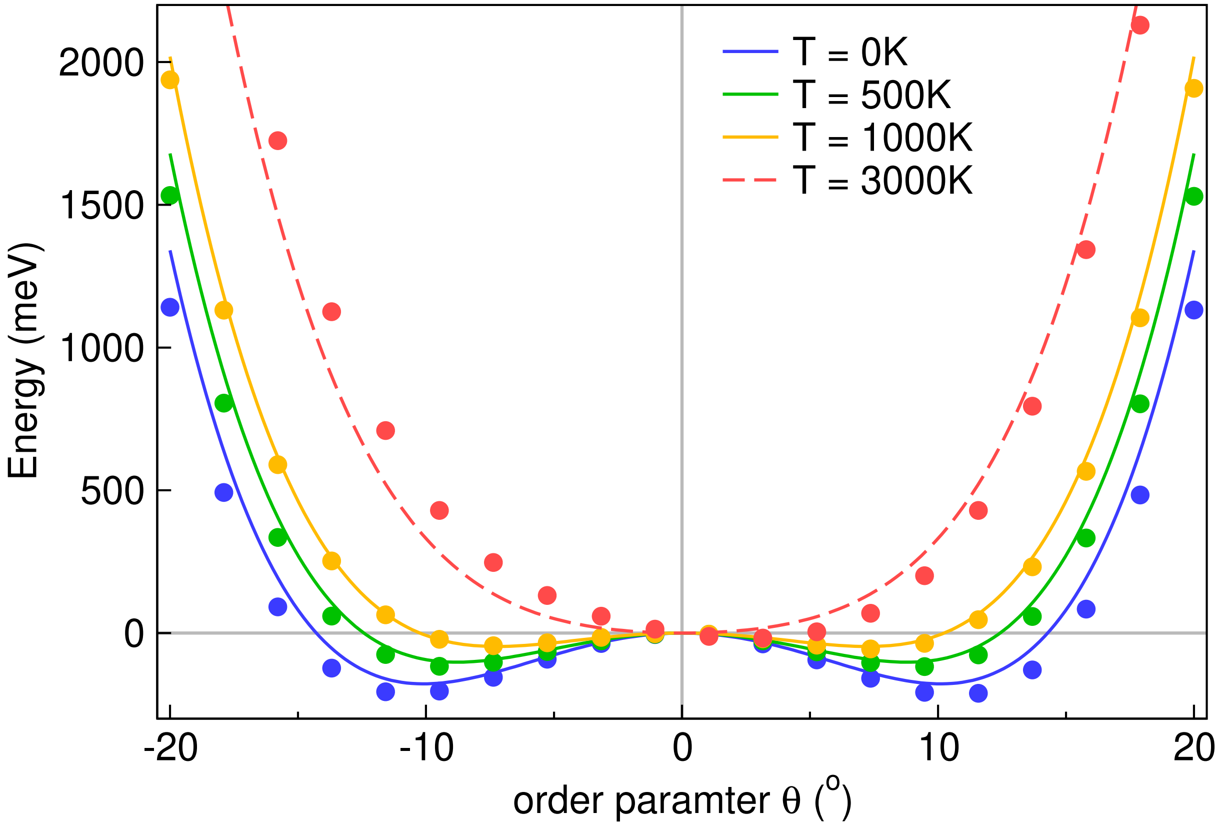


Figure 6.6: Landau free energy G for perovksite LaNiO$_3$ as a function of the order parameter θ at different temperatures. Both solid and dashed lines are calculated using Equation 6.1 with coefficients reported in Ref. [231]. Only solid lines are used during the training. The filled symbols correspond to GPSR predicted results using Equation 6.2.

see a minor shift to positive θ values occurs for the red filled symbols in Figure 6.6. Again, the model may need more data to learn the correct even function form. In general, GPSR does an excellent job in learning the relationship between temperature and the order parameter without any knowledge of the physical system. Our results here also reveal the potential to perform effective feature selection using GPSR, where the insignificant variables (features) could be approximated or ignored in post-processing (e.g., terms with negligibly small coefficients).

The purpose of these two use cases for learning the Avrami equation and Landau free energy expansion with GPSR is to show the potential of its application in materials science problems. Although these are relatively simple examples, the understanding and approaches applied could be generalized and utilized to solve real-world challenges. Apart from both function analysis and data pre-processing mentioned previously, hyper-parameter tuning is also an essential step to obtain the optimal solution. A grid-search scheme to search the hyper-parameter space (e.g., population size, number of generations, regularization, etc.) is recommended since the optimal settings differ from case to case. The grid-searching process can be exhausting, but it might also be rewarding. Comparing the results from different hyper-parameter settings could provide insight into the functional form of the optimal solution, especially if components of a particular function appears multiple times in the solution set.

The real challenge, however, is that very often materials science problems cannot be represented using regular functions (analytic and single-valued). A simple example would be to understand how different chemical compositions affect materials properties [232]. This problem originates from the inability to differentiate in the chemical space and is out of the scope of this prospective. It remains an open question in the materials research community.

6.4 Summary

Symbolic regression has shown competitive performance to other machine learning-based regression models in various research domains. While there are some shortcomings of the current state-of-the-art GPSR, e.g., high computational cost, non-deterministic optimization, there are numerous active research efforts focusing on improving the performance of symbolic regression to expand its use in real-world applications. The ability of symbolic regression to distill natural laws from data sets with high-dimensional parameter space makes it an ideal technique for materials science research, since these researchers typically face sparse data sets with multiple variables. Freed from having a fixed form of equations, symbolic regression can potentially reveal the significant interactions among physical variables.

CHAPTER 7
CONCLUSIONS

In this thesis, I employed a variety of computational and data-driven techniques to address the functional electronic-transition materials design challenges. Specifically, in Chapter 3, I focused on the (complex) lacunar spinel family, which is an ideal materials platform for next-generation electronics. However, the origin behind these fascinating physical properties remains unclear. Although there has been numerous experimental measurements performed to understand the mechanism governing the phase transition, there are relatively few computational/theoretical studies on these materials. Moreover, owing to the existence of transition-metal clusters in the lacunar spinels, the quantum state of the system is likely to be sensitive to local structural changes as well as inter/intra-cluster electronic interactions. I systematically investigated with density functional theory (DFT) the exchange-correlation functional dependency of several physical properties in the lacunar spinel family: crystal structures, electronic structures, magnetism, optical conductivity, and lattice dynamics. Our findings show that the GGA functional with on-site Coulomb interaction (GGA+U) of $1{\sim}2$ eV could quantitatively describe the lacunar spinels—no dynamical strong correlation effects are required. The meta-GGA functional SCAN and hybrid functional HSE06 also give results consistent with experimental data, but at much higher computational cost. The LDA functional is not recommended owing to its relatively large error. We also find the vanadium and molybdenum compounds are not Mott insulators in the low-temperature phase, and all compounds in the cubic phase are metallic from band theory. From the cubic phase phonon analysis, we started to understand the complex phase space of lacunar spinels spanned by the multiple metastable transition-metal cluster geometries. My findings in this project pave the way for future computational studies in transition-metal cluster compounds, and facilitate the design of novel compositions within the lacunar spinel family. In fact, I made a hypothesis that by controlling the geometry of the transition-metal cluster through external stimuli, one could effectively trigger an MIT within the complex lacunar spinel family. This idea led to the project described in Chapter 4.

In Chapter 4, I presented a novel featureless adaptive optimization engine (AOE). The AOE learns directly from chemical compositions alone to predict target materials properties for optimization, hence bypassing the feature construction step necessary to conventional machine learning models. This method is particularly suitable when prior knowledge of a particular material family is scarce, and for limited research budgets. I demonstrated the effectiveness of this new methodology on the complex lacunar spinel family, where all superior compositions on the design Pareto front have been identified by searching over less than 25% of the entire design space. The AOE could be easily generalized to other materials design tasks owing to its featureless nature, which also enables the co-design of functional materials. Hopefully, the AOE could help accelerate the pace of functional materials design and discovery in the near future.

Chapter 5 presented the `deepKNet`, a deep neural network (DNN) which learns from the crystal structure alone to make property classifications. “Structure-property relationships” is one of the most frequently used phrases in the materials research community, yet in most cases we are actually discussing “structure-composition-property relationships.” However, having a quantitative understanding to what extent crystal structures alone determine the materials properties, could facilitate novel materials design and discovery by optimizing structure and compositions separately. Although numerous statistical learning models have been developed to decode the structure-property relationship of crystalline materials, most of them explicitly include chemical composition in the feature set. Would it be possible to marginalize compositional information for generic solid-state materials and quantitatively study the correlation between crystal structure and materials properties? I presented a feasible solution to this question by utilizing a novel DNN that learns directly from the momentum space structure genome to predict multiple materials properties. Specifically, X-ray diffraction (XRD) patterns in the form of discrete 3-dimensional (3D) scattering points within momentum space were used as the only input features for the model to successfully accomplish multiple tasks: crystal system, elasticity, metallicity, and stability classifications. I designed the neural network architecture to be robust against multiple invariance requirements inherent in the 3D XRD patterns. I found that different materials properties have various dependencies on crystal structures; I learned that crystal symmetry plays a significant role in determining the metallicity of a material, whereas electron density information contributes more to elastic properties. Materials stability prediction, on the other hand, is more chemical-composition relevant; thus, the structure-based model is inferior to other DNNs that learn from compositional features. I also visualize the decision-making process of the metal-insulator classifier, and identified some trends for materials with similar crystal structures. This work demonstrates the feasibility to use DNN models to help scientists understand materials physics (more science oriented) rather than only building predictive models (more engineering oriented). Our findings here also emphasize the significance of crystal structures to certain materials properties, which could potentially help decouple the structural and compositional optimization processes in functional materials design tasks.

Lastly, in Chapter 6, I introduced the symbolic regression (SR) technique and its potential applications in materials science. The critical problem that we solve using SR is to find the mathematical function form mapping the experimental observations to materials properties of interest. Connecting a quantitative numerical relationship with interpretability is vital for chemists and materials scientists to understand the underlying materials systems. SR does not require any predefined function form or numerical relationship, yet it can automatically learn the most appropriate expression through genetic-programming based algorithms.

CHAPTER 8
OUTLOOK — TOWARDS AN INTEGRATED MATERIALS DISCOVERY WORKFLOW

In the last chapter, I would like to share some ideas about automating an integrated materials design and discovery workflow. From the previous chapters, one can learn some basic ideas regarding the strength and weakness of various materials research techniques—*ab initio* simulations, Bayesian optimization, deep neural networks, and symbolic regression. A nice strategy should fully exploit their strength while avoiding the weakness. To that end, I present an iterative materials discovery workflow as depicted in Figure 8.1.

We learn from both Chapter 4 and Chapter 5 that different materials properties have various dependencies on crystal structure and compositions. For instance, thermodynamic stability is more composition-dependent, while metallicity is more relevant to the crystal symmetry. Therefore, we avoid the simultaneous optimization of both composition and structures, instead, we can iteratively optimize these two variables. Specifically, the `deepKNet` could be applied to a large number of candidate crystal structures (e.g., artificially generated structures from various databases), it will be able to provide a reasonable (e.g., $O(10^1)$) number of top candidate materials that could exhibit metal-insulator transitions (MITs) in my research context. Then, we can use high-fidelity *ab initio* density functional theory simulations to validate whether these top candidates could exhibit MITs as predicted. If we identify a new MIT material, we can search through different chemical compositions within the same crystal structure (since we know crystal structure plays a decisive role in metallicity) for more MIT materials, possibly with better functionality.

Next, the adaptive optimization engine (AOE) proposed in Chapter 4 could identify the most promising materials compositions from within the design space, so that we do not have to enumerate all possible chemical compositions for the target crystal system of interest. With the new crystal compositions fed to the structure-based `deepKNet`, the network can give each candidate a

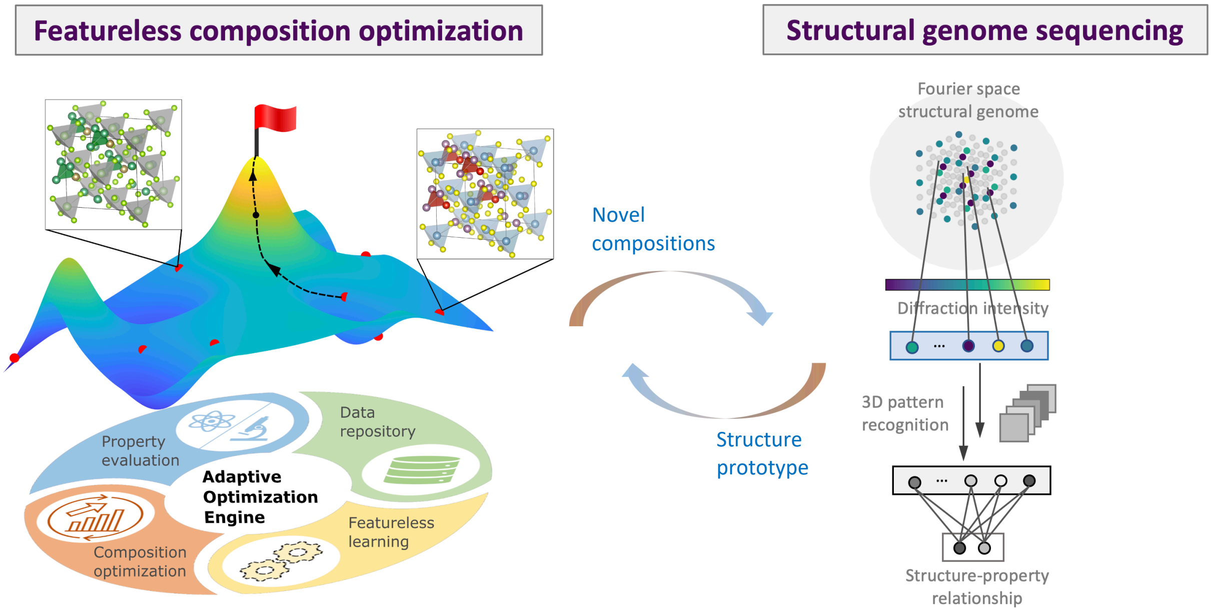


Figure 8.1: An iterative functional materials discovery workflow consists of featureless composition optimization and structural genome sequencing subroutines. We sequentially update the compositions and crystal structures in a round-robin fashion.

classification score (e.g., probability of being a metal-insulator transition material), and we could further filter out the promising candidates. In fact, this workflow could be easily generalized to solve materials discovery challenges beyond the metal-insulator transition compounds.

Of course, experimental synthesis, characterizations and validations are the last yet most vital step before we announce the success of any computational materials discovery project. Throughout the years, I witnessed more chemists and materials scientists collaborating with computational scientists and statisticians to utilize data-driven models for their domain challenges. These projects typically involve experimentation, theoretical modeling, computational simulation, and statistical learning. I believe this is in general a good trend, that we have started using sophisticated techniques to deal with complicated materials discovery challenges. I am honored to be able to make novel contributions to this research endeavor.

BIBLIOGRAPHY

1. Government, U. Materials Genome Initiative National Science and Technology Council Committee on Technology Subcommittee on the Materials Genome Initiative JUNE 2014. *Whitehouse.Gov* (2014).

2. Curtarolo, S. *et al.* AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **58,** 227–235 (2012).

3. Jain, A. *et al.* The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1,** 011002 (2013).

4. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *Jom* **65,** 1501–1509 (2013).

5. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **120** (Apr. 2018).

6. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J. & Lookman, T. Adaptive strategies for materials design using uncertainties. *Scientific reports* **6,** 1–9 (2016).

7. Ahmad, Z., Xie, T., Maheshwari, C., Grossman, J. C. & Viswanathan, V. Machine learning enabled computational screening of inorganic solid electrolytes for suppression of dendrite formation in lithium metal anodes. *ACS central science* **4,** 996–1006 (2018).

8. Georgescu, A. B. *et al.* A Database and Machine Learning Model to Identify Thermally Driven Metal-Insulator Transition Compounds. *arXiv preprint arXiv:2010.13306* (2020).

9. Tokura, Y. & Hwang, H. Y. Complex oxides on fire. *Nature Materials* **7,** 694–695 (2008).

10. Zubko, P., Gariglio, S., Gabay, M., Ghosez, P. & Triscone, J.-M. Interface physics in complex oxide heterostructures. *Annu. Rev. Condens. Matter Phys.* **2,** 141–165 (2011).

11. Chakhalian, J., Millis, A. & Rondinelli, J. Whither the oxide interface. *Nature Materials* **11,** 92–94 (2012).

12. Hwang, H. Y. *et al.* Emergent phenomena at oxide interfaces. *Nature Materials* **11,** 103–113 (2012).

13. Fujimori, A. Electronic structure of metallic oxides: band-gap closure and valence control. *Journal of Physics and Chemistry of Solids* **53,** 1595–1602 (1992).

14. Rondinelli, J. M., May, S. J. & Freeland, J. W. Control of octahedral connectivity in perovskite oxide heterostructures: An emerging route to multifunctional materials discovery. *MRS Bulletin* **37,** 261–270 (2012).

15. Newns, D., Elmegreen, B., Hu Liu, X. & Martyna, G. A low-voltage high-speed electronic switch based on piezoelectric transduction. *Journal of Applied Physics* **111,** 084509 (2012).

16. Newns, D. *et al.* Mott transition field effect transistor. *Applied Physics Letters* **73,** 780–782 (1998).

17. Chudnovskiy, F., Luryi, S. & Spivak, B. Switching device based on first-order metal-insulator transition induced by external electric field. *Future Trends in Microelectronics: the Nano Millennium* **148** (2002).

18. Corraze, B. *et al.* Electric field induced avalanche breakdown and non-volatile resistive switching in the Mott Insulators AM_4Q_8. *The European Physical Journal Special Topics* **222,** 1046–1056 (2013).

19. Arimoto, Y. & Ishiwara, H. Current status of ferroelectric random-access memory. *Mrs Bulletin* **29,** 823–828 (2004).

20. Xiao, L. *et al.* Fast adaptive thermal camouflage based on flexible VO_2/graphene/CNT thin films. *Nano Letters* **15,** 8365–8370 (2015).

21. Lee, S. *et al.* Anomalously low electronic thermal conductivity in metallic vanadium dioxide. *Science* **355,** 371–374 (2017).

22. Knowles, P. J. & Handy, N. C. A new determinant-based full configuration interaction method. *Chemical Physics Letters* **111,** 315–321 (1984).

23. Perez-Garcia, D., Verstraete, F., Wolf, M. M. & Cirac, J. I. Matrix product state representations. *arXiv preprint quant-ph/0608197* (2006).

24. Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics* **79,** 291 (2007).

25. Schollwöck, U. The density-matrix renormalization group. *Reviews of Modern Physics* **77,** 259 (2005).

26. Csanak, G., Taylor, H. & Yaris, R. in *Advances in atomic and molecular physics* 287–361 (Elsevier, 1971).

27. Nightingale, M. P. & Umrigar, C. J. *Quantum Monte Carlo methods in physics and chemistry* **525** (Springer Science & Business Media, 1998).

28. Mitchell, T. M. *et al.* Machine learning (1997).

29. Caruana, R. & Niculescu-Mizil, A. *An empirical comparison of supervised learning algorithms* in *Proceedings of the 23rd international conference on Machine learning* (2006), 161–168.

30. Hastie, T., Tibshirani, R. & Friedman, J. in *The elements of statistical learning* 9–41 (Springer, 2009).

31. Rasmussen, C. E. *Gaussian processes in machine learning* in *Summer school on machine learning* (2003), 63–71.

32. An, S., Liu, W. & Venkatesh, S. *Face recognition using kernel ridge regression* in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), 1–7.

33. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30,** 3146–3154 (2017).

34. Noble, W. S. What is a support vector machine? *Nature biotechnology* **24,** 1565–1567 (2006).

35. Barlow, H. B. Unsupervised learning. *Neural computation* **1,** 295–311 (1989).

36. Xia, G.-e. & Jin, W.-d. Model of customer churn prediction on support vector machine. *Systems Engineering-Theory & Practice* **28,** 71–77 (2008).

37. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

38. Berner, C. *et al.* Dota 2 with large scale deep reinforcement learning. *arXiv:1912.06680* (2019).

39. Arulkumaran, K., Cully, A. & Togelius, J. *Alphastar: An evolutionary computation perspective* in *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (2019), 314–315.

40. LeCun, Y., Bengio, Y., *et al.* Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361,** 1995 (1995).

41. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9,** 1735–1780 (1997).

42. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

43. Wang, F.-Y. *et al.* Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica* **3,** 113–120 (2016).

44. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35,** 4862–4865 (2019).

45. Brown, T. B. *et al.* Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

46. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Physical Review Letters* **108,** 253002 (2012).

47. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559,** 547–555 (July 2018).

48. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361,** 360–365 (2018).

49. Nouira, A., Sokolovska, N. & Crivello, J.-C. Crystalgan: learning to discover crystallographic structures with generative adversarial networks. *arXiv preprint arXiv:1810.11203* (2018).

50. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **31,** 3564–3572 (2019).

51. Wang, Y., Puggioni, D. & Rondinelli, J. M. Assessing exchange-correlation functional performance in the chalcogenide lacunar spinels $\mathrm{GaM_4Q_8}$ (M= Mo, V, Nb, Ta; Q= S, Se). *Physical Review B* **100,** 115149 (2019).

52. Cario, L., Vaju, C., Corraze, B., Guiot, V. & Janod, E. Electric-Field-Induced Resistive Switching in a Family of Mott Insulators: Towards a New Class of RRAM Memories. *Advanced Materials* **22,** 5193–5197 (2010).

53. Camjayi, A *et al.* First-order insulator-to-metal Mott transition in the paramagnetic 3D System $\mathrm{GaTa_4Se_8}$. *Physical Review Letters* **113,** 086404 (2014).

54. Kézsmárki, I *et al.* Néel-type skyrmion lattice with confined orientation in the polar magnetic semiconductor GaV_4S_8. *Nature Materials* **14,** 1116 (2015).

55. Widmann, S. *et al.* On the multiferroic skyrmion-host GaV_4S_8. *Philosophical Magazine* **97,** 3428–3445 (2017).

56. Pocha, R., Johrendt, D. & Pöttgen, R. Electronic and structural instabilities in GaV_4S_8 and $GaMo_4S_8$. *Chemistry of Materials* **12,** 2882–2887 (2000).

57. Sahoo, Y & Rastogi, A. Evidence of hopping conduction in the V_4-cluster compound GaV_4S_8. *Journal of Physics: Condensed Matter* **5,** 5953 (1993).

58. Francois, M *et al.* Structural phase transition in $GaMo_4S_8$ by X-ray powder diffraction. *Zeitschrift für Kristallographie-Crystalline Materials* **196,** 111–128 (1991).

59. François, M *et al.* Structural phase transition in $GaMo_4Se_8$ and $AlMo_4S_8$ by X-ray powder diffraction. *Zeitschrift für Kristallographie-Crystalline Materials* **200,** 47–56 (1992).

60. Abd-Elmeguid, M. *et al.* Transition from Mott Insulator to Superconductor in $GaNb_4Se_8$ and $GaTa_4Se_8$ under High Pressure. *Physical Review Letters* **93,** 126403 (2004).

61. Sieberer, M, Turnovszky, S, Redinger, J & Mohn, P. Importance of cluster distortions in the tetrahedral cluster compounds GaM_4X_8 (M= Mo, V, Nb, Ta; X= S, Se): Ab initio investigations. *Physical Review B* **76,** 214106 (2007).

62. Schueller, E. C. *et al.* Modeling the structural distortion and magnetic ground state of the polar lacunar spinel GaV_4Se_8. *Physical Review B* **100,** 045131 (2019).

63. Le Beuze, A, Loirat, H, Zerrouki, M. & Lissillour, R. Tetrahedral Clusters of $GaMo_4S_8$-Type Compounds: A Metal Bonding Analysis. *Journal of Solid State Chemistry* **120,** 80–89 (1995).

64. Rastogi, A. *et al.* An electron-phonon contribution to the stoner enhancement in $GaMo_4X_8$ compounds. *Journal of Low Temperature Physics* **55,** 551–568 (1984).

65. Rastogi, A. & Wohlfarth, E. Magnetic Field-Induced Transitions in the Mo_4 Cluster Compounds $GaMo_4S_8$ and $GaMo_4Se_8$ Showing Heavy Fermion Behaviour. *Physica Status Solidi (b)* **142,** 569–573 (1987).

66. Powell, A. V. *et al.* Cation Substitution in Defect Thiospinels: Structural and Magnetic Properties of $GaV_{4-x}Mo_xS_8$ $(0 \leq x \leq 4)$. *Chemistry of Materials* **19,** 5035–5044 (2007).

67. Fujima, Y, Abe, N, Tokunaga, Y & Arima, T. Thermodynamically stable skyrmion lattice at low temperatures in a bulk crystal of lacunar spinel GaV_4Se_8. *Physical Review B* **95,** 180410 (2017).

68. Shanthi, N & Sarma, D. Electronic structure of vacancy ordered spinels, $GaMo_4S_8$ and GaV_4S_8, from ab initio calculations. *Journal of Solid State Chemistry* **148,** 143–149 (1999).

69. Pocha, R., Johrendt, D., Ni, B. & Abd-Elmeguid, M. M. Crystal structures, electronic properties, and pressure-induced superconductivity of the tetrahedral cluster compounds $GaNb_4S_8$, $GaNb_4Se_8$, and $GaTa_4Se_8$. *Journal of the American Chemical Society* **127,** 8732–8740 (2005).

70. Reschke, S *et al.* Optical conductivity in multiferroic GaV_4S_8 and GeV_4S_8: Phonons and electronic transitions. *Physical Review B* **96,** 144302 (2017).

71. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77,** 3865 (1996).

72. Csonka, G. I. *et al.* Assessing the performance of recent density functionals for bulk solids. *Physical Review B* **79,** 155107 (2009).

73. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Physical Review Letters* **115,** 036402 (2015).

74. Heyd, J., Peralta, J. E., Scuseria, G. E. & Martin, R. L. Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional. *The Journal of Chemical Physics* **123,** 174101 (2005).

75. Van de Walle, A & Ceder, G. Correcting overbinding in local-density-approximation calculations. *Physical Review B* **59,** 14992 (1999).

76. Haas, P., Tran, F. & Blaha, P. Calculation of the lattice constant of solids with semilocal functionals. *Physical Review B* **79,** 085104 (2009).

77. Paier, J. *et al.* Screened hybrid density functionals applied to solids. *The Journal of Chemical Physics* **124,** 154709 (2006).

78. Dudarev, S., Botton, G., Savrasov, S., Humphreys, C. & Sutton, A. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+ U study. *Physical Review B* **57,** 1505 (1998).

79. Zhang, J. *et al.* Magnetic properties and spin-driven ferroelectricity in multiferroic skyrmion host GaV_4S_8. *Physical Review B* **95,** 085136 (2017).

80. Müller, H., Kockelmann, W. & Johrendt, D. The magnetic structure and electronic ground states of Mott insulators GeV_4S_8 and GaV_4S_8. *Chemistry of materials* **18,** 2174–2180 (2006).

81. Tirado-Rives, J. & Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. *Journal of Chemical Theory and Computation* **4,** 297–306 (2008).

82. Di Valentin, C., Pacchioni, G. & Selloni, A. Electronic structure of defect states in hydroxylated and reduced rutile TiO_2 (110) surfaces. *Physical Review Letters* **97,** 166803 (2006).

83. He, J. & Franchini, C. Screened hybrid functional applied to $3d^0 \rightarrow 3d^8$ transition-metal perovskites LaMO$_3$ (M= Sc–Cu): Influence of the exchange mixing parameter on the structural, electronic, and magnetic properties. *Physical Review B* **86,** 235117 (2012).

84. Hummer, K., Harl, J. & Kresse, G. Heyd-Scuseria-Ernzerhof hybrid functional for calculating the lattice dynamics of semiconductors. *Physical Review B* **80,** 115205 (2009).

85. Ramzan, M., Li, Y., Chimata, R. & Ahuja, R. Electronic, mechanical and optical properties of Y_2O_3 with hybrid density functional (HSE06). *Computational Materials Science* **71,** 19–24 (2013).

86. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **54,** 11169 (1996).

87. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **59,** 1758 (1999).

88. Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **50,** 17953 (1994).

89. Blöchl, P. E., Jepsen, O. & Andersen, O. K. Improved tetrahedron method for Brillouin-zone integrations. *Physical Review B* **49,** 16223 (1994).

90. Bergerhoff, D., Brown, I. D. & Allen, F. Crystallographic databases. *International Union of Crystallography, Chester* **360,** 77–95 (1987).

91. Crystal structures available at our group GitHub Page https://github.com/MTD-group/lacunar_spinel_structures.

92. Kim, H.-S., Im, J., Han, M. J. & Jin, H. Spin-orbital entangled molecular j_{eff} states in lacunar spinel compounds. *Nature Communications* **5,** 3988 (2014).

93. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scripta Materialia* **108,** 1–5 (2015).

94. The inelastic neutron diffraction simulator is available at GitHub Page https://github.com/raymond931118/INS_simulator.

95. Jynn, J., Ratcliff, W., Bleuel, M., Zhang, L. & Cheong, S.-W. *Neutron Investigation of the Magnetic Structures and Phase Transitions in Multiferroic GaV_4S_8* Conference abstract available at here.

96. Kim, H.-S., Haule, K. & Vanderbilt, D. Molecular Mott state in the deficient spinel $\mathrm{GaV_4S_8}$. *arXiv preprint arXiv:1810.09495* (2018).

97. Wang, Z. *et al.* Polar Dynamics at the Jahn-Teller Transition in Ferroelectric $\mathrm{GaV_4S_8}$. *Physical Review Letters* **115,** 207601 (2015).

98. Jeong, M. Y. *et al.* Direct experimental observation of the molecular $\mathrm{j_{eff}} = 3/2$ ground state in the lacunar spinel $\mathrm{GaTa_4Se_8}$. *Nature Communications* **8,** 782 (2017).

99. Franke, K. J. A. *et al.* Magnetic phases of skyrmion-hosting $\mathrm{GaV_4S_{8-\mathit{y}}Se_\mathit{y}}$ $(y = 0, 2, 4, 8)$ probed with muon spectroscopy. *Physical Review B* **98,** 054428 (5 2018).

100. Butykai, Á. *et al.* Characteristics of ferroelectric-ferroelastic domains in Néel-type skyrmion host $\mathrm{GaV_4S_8}$. *Scientific Reports* **7,** 44663 (2017).

101. Ruff, E. *et al.* Multiferroicity and skyrmions carrying electric polarization in $\mathrm{GaV_4S_8}$. *Science advances* **1,** e1500916 (2015).

102. Zhang, H.-M. *et al.* Possible emergence of a skyrmion phase in ferroelectric $\mathrm{GaMo_4S_8}$. *Physical Review B* **99,** 214427 (2019).

103. Ehlers, D *et al.* Skyrmion dynamics under uniaxial anisotropy. *Physical Review B* **94,** 014406 (2016).

104. Leonov, A. & Kézsmárki, I. Skyrmion robustness in noncentrosymmetric magnets with axial symmetry: The role of anisotropy and tilted magnetic fields. *Physical Review B* **96,** 214413 (2017).

105. Perdew, J. P., Kurth, S., Zupan, A. & Blaha, P. Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation. *Physical Review Letters* **82,** 2544 (1999).

106. Liu, P. *et al.* Relativistic GW+BSE study of the optical properties of Ruddlesden-Popper iridates. *Physical Review Materials* **2,** 075003 (2018).

107. Hlinka, J. *et al.* Lattice modes and the Jahn-Teller ferroelectric transition of $\mathrm{GaV_4S_8}$. *Physical Review B* **94,** 060104 (6 2016).

108. Wang, Y., Iyer, A., Chen, W. & Rondinelli, J. M. Featureless adaptive optimization accelerates functional electronic materials design. *Applied Physics Reviews* **7,** 041403 (2020).

109. Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Reviews of Modern Physics* **70,** 1039–1263 (Oct. 1998).

110. Shukla, N. *et al.* A steep-slope transistor based on abrupt electronic phase transition. *Nature Communications* **6** (Aug. 2015).

111. Zhou, Y. & Ramanathan, S. Mott Memory and Neuromorphic Devices. *Proceedings of the IEEE* **103,** 1289–1310 (Aug. 2015).

112. Yang, Z., Ko, C. & Ramanathan, S. Oxide Electronics Utilizing Ultrafast Metal-Insulator Transitions. *Annual Review of Materials Research* **41,** 337–367 (Aug. 2011).

113. Zhang, W., Liu, J. & Wei, T.-C. Machine learning of phase transitions in the percolation and XY models. *Physical Review E* **99** (Mar. 2019).

114. Coll, M. *et al.* Towards Oxide Electronics: a Roadmap. *Applied Surface Science* **482,** 1–93 (July 2019).

115. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *APL Materials* **4,** 053208 (Apr. 2016).

116. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5** (Aug. 2019).

117. Noh, J. *et al.* Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **1,** 1370–1384 (Nov. 2019).

118. Ling, J., Hutchinson, M., Antono, E., Paradiso, S. & Meredig, B. High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates. *Integrating Materials and Manufacturing Innovation* **6,** 207–217 (July 2017).

119. Seko, A. *et al.* Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization. *Physical Review Letters* **115** (Nov. 2015).

120. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **5,** 1–17 (2019).

121. Gopakumar, A. M., Balachandran, P. V., Xue, D., Gubernatis, J. E. & Lookman, T. Multi-objective optimization for materials discovery via adaptive design. *Scientific Reports* **8,** 3738 (2018).

122. Dorolti, E. *et al.* Half-Metallic Ferromagnetism and Large Negative Magnetoresistance in the New Lacunar Spinel $\mathrm{GaTi_3VS_8}$. *Journal of the American Chemical Society* **132,** 5704–5710 (Apr. 2010).

123. Bichler, D. & Johrendt, D. Tuning of Metal-Metal Bonding and Magnetism via the Electron Count in $\mathrm{Ga}_x\mathrm{V}_{4-y}\mathrm{Cr}_y\mathrm{S}_8$. *Chemistry of Materials* **19,** 4316–4321 (Aug. 2007).

124. Guiot, V. *et al.* Avalanche breakdown in GaTa 4 Se 8- x Te x narrow-gap Mott insulators. *Nature communications* **4,** 1–6 (2013).

125. Bartel, C. J. *et al.* Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nature Communications* **9** (Oct. 2018).

126. Burdett, J. K., Coddens, B. A. & Kulkarni, G. V. Band gap and stability of solids. *Inorganic Chemistry* **27,** 3259–3261 (Sept. 1988).

127. McKay, M. D., Beckman, R. J. & Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **21,** 239 (May 1979).

128. Mockus, J., Tiesis, V. & Zilinskas, A. The application of Bayesian methods for seeking the extremum. *Towards global optimization* **2,** 2 (1978).

129. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Science Advances* **4,** eaaq0148 (Apr. 2018).

130. Streltsov, S. V. & Khomskii, D. I. Covalent bonds against magnetism in transition metal compounds. *Proceedings of the National Academy of Sciences* **113,** 10491–10496 (Sept. 2016).

131. Vaju, C. *et al.* Electric-Pulse-driven Electronic Phase Separation, Insulator-Metal Transition, and Possible Superconductivity in a Mott Insulator. *Advanced Materials* **20,** 2760–2765 (July 2008).

132. Juraschek, D., Fechner, M. & Spaldin, N. Ultrafast Structure Switching through Nonlinear Phononics. *Physical Review Letters* **118** (Jan. 2017).

133. Zunger, A. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry* **2,** 1–16 (2018).

134. Olson, G. B. Designing a New Material World. *Science* **288,** 993–998 (May 2000).

135. Xiong, W. & Olson, G. B. Cybermaterials: materials by design and accelerated insertion of materials. *npj Computational Materials* **2** (Feb. 2016).

136. Rondinelli, J. M., Poeppelmeier, K. R. & Zunger, A. Research Update: Towards designed functionalities in oxide-based electronic materials. *APL Materials* **3,** 080702 (Aug. 2015).

137. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: An open-access online database for data mining. *PLOS ONE* **12** (ed Ostroverkhova, O.) e0171501 (2017).

138. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3,** 1–13 (2017).

139. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science* **1,** 46–53 (2021).

140. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8,** 1–12 (2017).

141. Karamad, M. *et al.* Orbital graph convolutional neural network for material property prediction. *Physical Review Materials* **4,** 093801 (2020).

142. Ren, Z. *et al.* Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609* (2020).

143. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (2013).

144. Hellenbrandt, M. The inorganic crystal structure database – present and future. *Crystallography Reviews* **10,** 17–22 (2004).

145. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

146. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. *Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 652–660.

147. Qi, C. R. *et al. Volumetric and Multi-view CNNs for Object Classification on 3D Data* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 5648–5656.

148. Wu, Z. *et al. 3D ShapeNets: A deep representation for volumetric shapes* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), 1912–1920.

149. Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* (2013).

150. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nature Communications* **9,** 1–10 (2018).

151. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. SE (3)-transformers: 3D roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503* (2020).

152. Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Science Advances* **2,** e1600225 (2016).

153. Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70,** 1039–1263 (4 1998).

154. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS Communications* **9,** 793–805 (2019).

155. Deelman, E. *et al.* PANORAMA: An approach to performance modeling and diagnosis of extreme-scale workflows. *The International Journal of High Performance Computing Applications* **31,** 4–18 (2017).

156. Lupini, A. R., Oxley, M. P. & Kalinin, S. V. Pushing the limits of electron ptychography. *Science* **362,** 399–400 (2018).

157. Ren, F., Pandolfi, R., Van Campen, D., Hexemer, A. & Mehta, A. On-the-Fly Data Assessment for High-Throughput X-ray Diffraction Measurements. *ACS Combinatorial Science* **19,** 377–385 (2017).

158. Stein, H. S., Guevarra, D., Newhouse, P. F., Soedarmadji, E. & Gregoire, J. M. Machine learning of optical properties of materials – predicting spectra from images and images from spectra. *Chemical Science* **10,** 47–55 (2019).

159. Alberi, K. *et al.* The 2019 materials by design roadmap. *Journal of Physics D: Applied Physics* **52,** 013001 (2019).

160. Green, M. L. *et al.* Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews* **4,** 011105 (2017).

161. Ye, W. *et al.* Harnessing the Materials Project for machine-learning and accelerated discovery. *MRS Bulletin* **43,** 664–669 (2018).

162. Tanaka, I., Rajan, K. & Wolverton, C. Data-centric science for materials innovation. *MRS Bulletin* **43,** 659–663 (2018).

163. Kim, E. *et al.* Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials* **29,** 9436–9444 (2017).

164. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. *Information retrieval and text mining technologies for chemistry* 2017.

165. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science* **111,** 218–230 (2016).

166. Zhuo, Y., Mansouri Tehrani, A., Oliynyk, A. O., Duke, A. C. & Brgoch, J. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nature Communications* **9,** 4377 (2018).

167. Hall, P. & Gill, N. *An introduction to machine learning interpretability* (O'Reilly Media, Incorporated, 2019).

168. Analytics, C. Web of science. *Trust the Difference. Web of Science Fact Book.* (2017).

169. Augusto, D. A. & Barbosa, H. J. *Symbolic regression via genetic programming* in *Proceedings. Vol. 1. Sixth Brazilian Symposium on Neural Networks* (2000), 173–178.

170. Seber, G. A. & Lee, A. J. *Linear regression analysis* (John Wiley & Sons, 2012).

171. Koza, J. R. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing* **4,** 87–112 (1994).

172. Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **261,** 872–878 (1993).

173. Meredig, B. & Wolverton, C. A hybrid computational–experimental approach for automated crystal structure solution. *Nature Materials* **12,** 123–127 (2013).

174. Chua, A. L.-S., Benedek, N. A., Chen, L., Finnis, M. W. & Sutton, A. P. A genetic algorithm for predicting the structures of interfaces in multicomponent systems. *Nature Materials* **9,** 418–422 (2010).

175. Mohn, C. E., Stølen, S. & Kob, W. Predicting the Structure of Alloys Using Genetic Algorithms. *Materials and Manufacturing Processes* **26,** 348–353 (2011).

176. Arnaldo, I., Krawiec, K. & O'Reilly, U.-M. *Multiple regression genetic programming* in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation* (2014), 879–886.

177. Moore, J. A., Ma, R., Domel, A. G. & Liu, W. K. An efficient multiscale model of damping properties for filled elastomers with complex microstructures. *Composites Part B: Engineering* **62,** 262–270 (2014).

178. Castelli, M., Silva, S. & Vanneschi, L. A C++ framework for geometric semantic genetic programming. *Genetic Programming and Evolvable Machines* **16,** 73–81 (2015).

179. Miller, J. F., Job, D. & Vassilev, V. K. Principles in the Evolutionary Design of Digital Circuits—Part I. *Genetic Programming and Evolvable Machines* **1,** 7–35 (2000).

180. Rad, H. I., Feng, J. & Iba, H. GP-RVM: Genetic programing-based symbolic regression using relevance vector machine. *arXiv preprint arXiv:1806.02502* (2018).

181. Giustolisi, O. & Savic, D. A. Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics* **11,** 225 (2009).

182. McConaghy, T. in *Genetic Programming Theory and Practice IX* 235–260 (Springer, 2011).

183. Orzechowski, P., La Cava, W. & Moore, J. H. *Where are we now? A large benchmark study of recent symbolic regression methods* in *Proceedings of the Genetic and Evolutionary Computation Conference* (2018), 1183–1190.

184. Icke, I. & Bongard, J. C. *Improving genetic programming based symbolic regression using deterministic machine learning* in *2013 IEEE Congress on Evolutionary Computation* (2013), 1763–1770.

185. Krawiec, K. Genetic programming-based construction of features for machine learning and knowledge discovery tasks. *Genetic Programming and Evolvable Machines* **3,** 329–343 (2002).

186. Lu, Q., Ren, J. & Wang, Z. Using genetic programming with prior formula knowledge to solve symbolic regression problem. *Computational Intelligence and Neuroscience* **2016,** 1 (2016).

187. Li, L., Fan, M., Singh, R. & Riley, P. Neural-Guided Symbolic Regression with Semantic Prior. *arXiv preprint arXiv:1901.07714* (2019).

188. Tolman, C. The 16 and 18 electron rule in organometallic chemistry and homogeneous catalysis. *Chemical Society Reviews* **1,** 337–353 (1972).

189. Van Beest, B., Kramer, G. J. & Van Santen, R. Force fields for silicas and aluminophosphates based on ab initio calculations. *Physical Review Letters* **64,** 1955 (1990).

190. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters* **393,** 51–57 (2004).

191. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324,** 81–85 (2009).

192. Gout, J., Quade, M., Shafi, K., Niven, R. K. & Abel, M. Synchronization control of oscillator networks using symbolic regression. *Nonlinear Dynamics* **91,** 1001–1021 (2018).

193. Arkov, V. *et al.* System identification strategies applied to aircraft gas turbine engines. *Annual Reviews in Control* **24,** 67–81 (2000).

194. Berardi, L., Giustolisi, O., Kapelan, Z. & Savic, D. Development of pipe deterioration models for water distribution systems using EPR. *Journal of Hydroinformatics* **10,** 113–126 (2008).

195. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **104,** 9943–9948 (2007).

196. Cai, W., Pacheco-Vega, A., Sen, M. & Yang, K. T. Heat transfer correlations by symbolic regression. *International Journal of Heat and Mass Transfer* **49,** 4352–4359 (2006).

197. Can, B. & Heavey, C. Comparison of experimental designs for simulation-based symbolic regression of manufacturing systems. *Computers and Industrial Engineering* **61,** 447–462 (2011).

198. McKay, B., Willis, M. & Barton, G. Steady-state modelling of chemical process systems using genetic programming. *Computers and Chemical Engineering* **21,** 981–996 (1997).

199. La Cava, W., Danai, K. & Spector, L. Inference of compact nonlinear dynamic models by epigenetic local search. *Engineering Applications of Artificial Intelligence* **55,** 292–306 (2016).

200. La Cava, W. *et al.* Automatic identification of wind turbine models using evolutionary multiobjective optimization. *Renewable Energy* **87,** 892–902 (2016).

201. Chen, S.-H. & Yeh, C.-H. Toward a computable approach to the efficient market hypothesis: An application of genetic programming. *Journal of Economic Dynamics and Control* **21,** 1043–1063 (1997).

202. Gray, G. J., Murray-Smith, D. J., Li, Y., Sharman, K. C. & Weinbrenner, T. Nonlinear model structure identification using genetic programming. *Control Engineering Practice* **6,** 1341–1352 (1998).

203. Khu, S. T., Liong, S. Y., Babovic, V., Madsen, H. & Muttil, N. Genetic programming and its application in real-time runoff forecasting. *Journal of the American Water Resources Association* **37,** 439–451 (2001).

204. Liong, S.-Y. *et al.* GENETIC PROGRAMMING: A NEW PARADIGM IN RAINFALL RUNOFF MODELING. *Journal of the American Water Resources Association* **38,** 705–718 (2002).

205. Quade, M., Abel, M., Shafi, K., Niven, R. K. & Noack, B. R. Prediction of dynamical systems by symbolic regression. *Physical Review E* **94,** 012214 (2016).

206. Schmidt, M. D. *et al.* Automated refinement and inference of analytical models for metabolic networks. *Physical Biology* **8,** 055011 (2011).

207. Stanislawska, K., Krawiec, K. & Kundzewicz, Z. W. Modeling global temperature changes with genetic programming. *Computers & Mathematics with Applications* **64,** 3717–3728 (2012).

208. Uesaka, K. & Kawamata, M. Synthesis of low-sensitivity second-order digital filters using genetic programming with automatically defined functions. *IEEE Signal Processing Letters* **7,** 83–85 (2000).

209. Vyas, R., Goel, P. & Tambe, S. S. in *Handbook of genetic programming applications* 99–140 (Springer, 2015).

210. Langdon, W. B. & Barrett, S. J. in *Evolutionary Computation in Data Mining* 211–235 (Springer-Verlag, Berlin/Heidelberg, 2005). ISBN: 3-540-22370-3.

211. Vyas, R., Goel, P., Karthikeyan, M., Tambe, S. & Kulkarni, B. Pharmacokinetic Modeling of Caco-2 Cell Permeability Using Genetic Programming (GP) Method. *Letters in Drug Design & Discovery* **11,** 1112–1118 (2014).

212. Barmpalexis, P., Kachrimanis, K., Tsakonas, A. & Georgarakis, E. Symbolic regression via genetic programming in the optimization of a controlled release pharmaceutical formulation. *Chemometrics and Intelligent Laboratory Systems* **107,** 75–82 (2011).

213. Muzny, C. D., Huber, M. L. & Kazakov, A. F. Correlation for the Viscosity of Normal Hydrogen Obtained from Symbolic Regression. *Journal of Chemical & Engineering Data* **58,** 969–979 (2013).

214. Markov, A. A. *et al.* Oxygen Nonstoichiometry and Ionic Conductivity of $Sr_3Fe_{2-x}Sc_xO_{7-\delta}$. *Chemistry of Materials* **19,** 3980–3987 (2007).

215. Nakamura, A. & Wagner Jr, J. B. Defect structure, ionic conductivity, and diffusion in yttria stabilized zirconia and related oxide electrolytes with fluorite structure. *Journal of the Electrochemical society* **133,** 1542 (1986).

216. Daza, L *et al.* Modified nickel oxides as cathode materials for MCFC. *Journal of Power Sources* **86,** 329–333 (2000).

217. Maslyaev, M, Hvatov, A & Kalyuzhnaya, A. Data-driven PDE discovery with evolutionary approach.(2019). *arXiv preprint arXiv:1903.08011.*

218. Gaucel, S., Keijzer, M., Lutton, E. & Tonda, A. *Learning dynamical systems using standard symbolic regression* in *European Conference on Genetic Programming* (2014), 25–36.

219. Schmidt, M. & Lipson, H. Symbolic regression of implicit equations. *Genetic Programming Theory and Practice* **7,** 73–85 (2009).

220. Von Barth, U. & Hedin, L. A local exchange-correlation potential for the spin polarized case. i. *Journal of Physics C: Solid State Physics* **5,** 1629 (1972).

221. Voorhees, P, Spanos, G, *et al.* Modeling across scales: a roadmapping study for connecting materials models and simulations across length and time scales. *TMS, Warrendale, PA* **14** (2015).

222. Yadollahi, A., Shamsaei, N., Thompson, S. M. & Seely, D. W. Effects of process time interval and heat treatment on the mechanical and microstructural properties of direct laser deposited 316L stainless steel. *Materials Science and Engineering A* **644,** 171–183 (2015).

223. Ward, L. & Wolverton, C. Atomistic calculations and materials informatics: A review. *Curr. Opin. Solid State Mater. Sci.* **21,** 167–176 (2017).

224. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Physical Review Letters* **114,** 105503 (2015).

225. Ghiringhelli, L. M. *et al.* Learning physical descriptors for materials science by compressed sensing. *New Journal of Physics* **19,** 023017 (2017).

226. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2,** 083802 (2018).

227. Vanderplaats, G. N. *Numerical optimization techniques for engineering design* (Vanderplaats Research and Development, Incorporated, 2001).

228. Shimada, M, Kokawa, H, Wang, Z., Sato, Y. & Karibe, I. Optimization of grain boundary character distribution for intergranular corrosion resistant 304 stainless steel by twin-induced grain boundary engineering. *Acta Materialia* **50,** 2331–2341 (2002).

229. Decker, B. F. & Harker, D. Activation energy for recrystallization in rolled copper. *JOM* **2,** 887–890 (1950).

230. Stephens, T. *gplearn* https://gplearn.readthedocs.io/en/stable. 2016.

231. Gou, G., Grinberg, I., Rappe, A. M. & Rondinelli, J. M. Lattice normal modes and electronic properties of the correlated metal $LaNiO_3$. *Physical Review B* **84,** 144101 (2011).

232. Yu, H. *et al.* Electronic, crystal chemistry, and nonlinear optical property relationships in the dugganite $A_3B_3Cd_2O_{14}$ family. *Journal of the American Chemical Society* **138,** 4984–4989 (2016).

233. Rastogi, A. K. *et al.* Itinerant electron magnetism in the Mo4 tetrahedral cluster compounds $GaMo_4S_8$, $GaMo_4Se_8$, and $GaMo_4Se_4Te_4$. *Journal of Low Temperature Physics* **52,** 539–557 (Sept. 1983).

234. Resta, R. Macroscopic polarization in crystalline dielectrics: the geometric phase approach. *Reviews of Modern Physics* **66,** 899 (1994).

235. Crystal structures available at our group GitHub Page https://github.com/MTD-group/Pareto_font_lacunar_spinel.

236. Kirklin, S., Meredig, B. & Wolverton, C. High-throughput computational screening of new Li-ion battery anode materials. *Advanced Energy Materials* **3,** 252–262 (2013).

237. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **1,** 1–15 (2015).

238. Madsen, G. K., Carrete, J. & Verstraete, M. J. BoltzTraP2, a program for interpolating band structures and calculating semi-classical transport coefficients. *Comput. Phys. Commun.* **231,** 140 –145 (2018).

239. Zhang, Y., Tao, S., Chen, W. & Apley, D. W. A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors. *Technometrics,* 1–12 (Aug. 2019).

240. Zhang, Y., Apley, D. W. & Chen, W. Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. *Scientific Reports* **10,** 1–13 (2020).

241. Bautista, D. C. T. *A sequential design for approximating the pareto front using the expected pareto improvement function* PhD thesis (The Ohio State University, 2009).

242. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

Appendices

APPENDIX A
ADAPTIVE OPTIMIZATION ENGINE (AOE) IMPLEMENTATION DETAILS

A.1 Density Functional Calculation Details

We perform density functional theory (DFT) simulations as implemented in the Vienna *Ab initio* Simulation Package (VASP) [86, 87]. The projector augmented-wave (PAW) potentials [88] are used for all elements in our calculations with the following valence electron configurations: Al ($3s^23p^1$), Ga ($3d^{10}4s^24p^1$), In ($4d^{10}5s^25p^1$), V ($3s^23p^63d^44s^1$), Nb ($4s^24p^64d^45s^1$), Ta ($5p^65d^46s^1$), Cr ($3s^23p^63d^54s^1$), Mo ($4s^24p^64d^55s^1$), W($5s^25p^65d^56s^1$), S ($3s^23p^4$), Se ($4s^24p^4$), and Te($5s^25p^4$). We use exchange-correlation potentials as implemented by Perdew-Burke-Ernzerhof (PBE) [71]. The effect of on-site Coulomb interactions (PBE+U) is considered with a U value of 2.0 eV for all 6 transition metals. Previous studies have shown that such settings could nicely capture the complex electronic structures within the lacunar spinel family [51, 92]. Numerous spin configurations are evaluated to ensure the global ground state is achieved and that those states are consistent with available experimental magnetic data [233]. Spin-orbit interactions (SOI) are not considered in our calculations. Although it has been shown that SOI leads to interesting molecular j_{eff} states [92], this order does not strongly affect the size of the ground state electronic band gaps, even $5d$ transition metals lacunar spinels [51]. A Γ-centered $6 \times 6 \times 6$ k-point mesh with a 500 eV kinetic energy cutoff is used. We employ Gaussian smearing with a small 0.05 eV width. For density-of-state calculations, we use the tetrahedron method with Blöchl corrections [89]. Electric polarizations along the [111] direction are simulated using the Berry phase method [234].

The crystal structures of the existing lacunar spinels are obtained from our previous DFT studies [91], structures of new compositions are obtained by replacing the elements on the corresponding crystallographic sites from existing structures. We perform full lattice relaxations until the residual forces on each individual atom are less than 1.0 meVÅ^{-1}. The DFT-relaxed crystal structures of the Pareto front compositions are available at Ref. [235]. We initialize the relaxation with various magnetic moment configurations, the converged configuration with the lowest energy is reported as the DFT ground state. Zone center ($\mathbf{k} = \mathbf{0}$) phonon frequencies and eigendisplacements are obtained using the frozen-phonon method with pre- and post-processing performed with the Phonopy package [93]. The decomposition pathways are automatically generated using Grand Canonical Linear Programming [236] from the Open Quantum Materials Database [237].

Resistivity simulations are performed using electronic structures computed from VASP as previously described, but with an increased $24 \times 24 \times 24$ k-point mesh and the BoltzTrap2 package [238]. We also assume that all M^{a} sites have the same orientation within the crystal. In order to validate this model, we simulated a $2 \times 2 \times 2$ supercell of InNbMo$_3$Se$_8$ with one Nb atom oriented in a different direction from the other seven. We find that the ground state E_g as well as ΔH_d exhibit negligible changes from the homogeneous description. We also compared the change in properties with the anti-ferromagnetic spin configuration using a doubled simulation cell with the ferromagnetic ground state. As before, we find there are no significant changes in the aforementioned properties. These results are reasonable because the local structure of the TMC dictates the low-energy band structure near the Fermi level.

A.2 Multi-objective Bayesian Optimization Theory

Conventional Gaussian process (GP) modelling has been developed for only quantitative design variables and the associated correlation functions cannot handle categorical variables. To overcome this limitation, LVGP maps each categorical variable to a 2D Cartesian latent space [239, 240], establishing a numerical representation for different categories. With this mapping, the covariance model over categorical design variables can be any standard GP covariance model for quantitative variables, e.g., the Gaussian correlation function. In the AOE, two independent LVGP models with Gaussian correlation function are fit at each iteration to predict E_g and ΔH_d, respectively. In each LVGP model, categorical variables A, M^a, M^b and Q are represented by a 2D numerical latent variable vector to evaluate their correlation. Note that each categorical variable resides in its unique latent space. For the LVGP model predicting E_g, let $\boldsymbol{z}^A = [z_1^A, z_2^A]$ denote the latent variable for the A site. Similarly, $\boldsymbol{z}^{\mathrm{M}^a}$, $\boldsymbol{z}^{\mathrm{M}^b}$, and $\boldsymbol{z}^Q$ denote the latent variables for M^a, M^b and Q site, respectively. Then, the Gaussian correlation (ρ) between E_g of two compounds, e.g., GaMoV$_3$S$_8$ and AlNbW$_3$Se$_8$, is:

$$\begin{aligned}\rho\left(E_g^{\mathrm{GaMoV_3S_8}}, E_g^{\mathrm{AlNbW_3Se_8}}\right) = \exp\big(&-\|\boldsymbol{z}^{\mathrm{Ga}} - \boldsymbol{z}^{\mathrm{Al}}\|_2^2 - \|\boldsymbol{z}^{\mathrm{Mo}} - \boldsymbol{z}^{\mathrm{Nb}}\|_2^2 \\ &-\|\boldsymbol{z}^{\mathrm{V}} - \boldsymbol{z}^{\mathrm{W}}\|_2^2 - \|\boldsymbol{z}^{\mathrm{S}} - \boldsymbol{z}^{\mathrm{Se}}\|_2^2\big)\end{aligned} \tag{A.1}$$

where $\|.\|_2$ represents the Euclidean 2-norm. This procedure is used to compute the correlation matrix for properties of all evaluated compositions. The positioning of latent variables $\boldsymbol{z}^A$, $\boldsymbol{z}^{\mathrm{M}^a}$, $\boldsymbol{z}^{\mathrm{M}^b}$, and $\boldsymbol{z}^Q$ in their corresponding latent space are estimated via MLE as described in Ref. [239]. The LVGP model for ΔH_d also utilizes the 2D latent variable representation $\boldsymbol{\kappa}^A$, $\boldsymbol{\kappa}^{\mathrm{M}^a}$, $\boldsymbol{\kappa}^{\mathrm{M}^b}$, and $\boldsymbol{\kappa}^Q$ as previously defined to evaluate the correlation $\rho(\Delta H_d^{\mathrm{GaMoV_3S_8}}, \Delta H_d^{\mathrm{AlNbW_3Se_8}})$ in a similar manner.

The multi-objective Bayesian optimization starts from considering the lacunar spinel family

$A\mathrm{M}^a\mathrm{M}^b_3Q_8$ with $A \in \{$Al,Ga,In$\}$, $\mathrm{M}^a \in \{$V, Nb, Ta, Cr, Mo, W$\}$, $\mathrm{M}^b \in \{$V, Nb, Ta, Mo, W$\}$ and $Q \in \{$S,Se,Te$\}$. The design space ($\boldsymbol{C}$) comprises 270 compounds, each compound is represented by four design variables $A, \mathrm{M}^a, \mathrm{M}^b$ and Q with three, six, five, and three choices, respectively. Our objective is to maximize E_g and ΔH_d, which is represented in standard optimization formulation as:

$$\min_{\boldsymbol{c} \in \boldsymbol{C}} -E_g(\boldsymbol{c}), -\Delta H_d(\boldsymbol{c})\,. \tag{A.2}$$

Starting from the initial dataset, the AOE evaluates new candidate compounds by gauging their improvement in the design objectives. Here, we use the expected maximin improvement (EMI) metric [241] to guide the adaptive sampling framework. The Maximin Improvement (I_M) for compound $\boldsymbol{c}$ is:

$$I_M(\boldsymbol{c}) = \min_{\boldsymbol{c_i} \in \boldsymbol{C}_{PF}} \left\{ \max\left(\widetilde{E_g}(\boldsymbol{c}) - \widetilde{E_g}(\boldsymbol{c_i}), \widetilde{\Delta H_d}(\boldsymbol{c}) - \widetilde{\Delta H_d}(\boldsymbol{c_i}), 0 \right) \right\} \tag{A.3}$$

where $\boldsymbol{C}_{PF}$ is the current set of Pareto front compositions. To facilitate the comparison in Equation A.3, we scale the value of each design objective P using the scheme $\widetilde{P}(\cdot) = (P(\cdot) - P^{min})/(P^{max} - P^{min})$ where P^{max} and P^{min} are the maximum and minimum value of property observed so far. By scaling the properties, we ensure all design objectives are comparable and viewed equally. The EMI of compound $\boldsymbol{c}$ is defined as the expected value of I_M:

$$\mathrm{EMI}(\boldsymbol{c}) = \mathbb{E}[I_M(\boldsymbol{c})]\,. \tag{A.4}$$

We evaluate the EMI through Monte Carlo sampling with 500 trials. At each AOE iteration, the EMI is calculated for all compositions that are not yet present in the data repository. The composition with largest EMI will be sampled next in property evaluation and then added to the data repository.

APPENDIX B

DEEPKNET IMPLEMENTATION DETAILS

B.1 X-ray and Neutron Diffraction Simulations

We use a modified version of the X-ray diffraction (XRD) simulator as implemented in the open-source software `Pymatgen` to generate the input features for the `deepKNet` model [143]. Specifically, we consider all diffraction points within the limiting sphere of radius $4\pi/\lambda$, where $\lambda = 1.5406\,\text{Å}$ is the X-ray wavelength (Cu K_α in this case). The atomic form factors $f(s)$ are calculated using tabulated data to simulate the Fourier-transformed real-space atomic electron density function $\rho(\mathbf{r})$:

$$f(s) = Z - 41.78214 \cdot s^2 \cdot \sum_{i=1}^{n} a_i e^{-b_i s^2}\,, \tag{B.1}$$

where Z is the atomic number, $s = \frac{\sin\theta}{\lambda}$ and $\sin\theta = \frac{\lambda}{2d_{hkl}}$ (Bragg condition). The a_i and b_i coefficients are n fitting parameters for each element provided by `Pymatgen`. We then calculate

$$F_{hkl} = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{g}_{hkl}\cdot\mathbf{r_j}} \tag{B.2}$$

$$I_{hkl} = \frac{1}{V_{cell}^2} F_{hkl}^* F_{hkl} \tag{B.3}$$

where j runs over all atoms within the unit cell and V_{cell} is the conventional unit cell volume. Lorentz polarization and Debye-Waller factor are not considered in our simulation. Owing to the large values of the diffraction intensity, we take the natural logarithm of each I_{hkl}, i.e. $\tilde{I}_{hkl} = \ln(I_{hkl} + 1)$ to bring the intensity values within the range (0, 1].

For neutron scattering, Equation B.1 becomes a constant for each element (more specifically, for each isotope), which is independent of the momentum space position vector. The tabulated neutron scattering lengths are obtained from the `Pymatgen` package [143]. We then calculate the neutron diffraction (ND) patterns as before using Equation B.2 and Equation B.3, after obtaining the neutron scattering lengths.

B.2 Hyperparameter Optimization

The hyperparameters we considered for the model are tabulated in Table B.1. Note that the number of diffraction points n is considered as a variable instead of a hyperparameter of the model. We use a greedy approach to optimize each of these hyperparameters and take the average results from two randomly and independently generated training, validation, and test datasets. The reported data was generated using 1D convolutional layers with filter size 1×1 and filter channels $[4, 64, 128, 256, 512, 1024]$, the `max` pooling function, and multi-layer perceptions with hidden layer size $[1024, 512, 256, 256, k]$, where k is the number of output class. However, we find that the model performance on all classification tasks to be less dependent on the hyperparameters than the input perturbations. In other words, the same neural network architecture could capture most of the materials information, and increasing the number of model parameters does not significantly improve the model performance.

Table B.1: Hyperparameters explored in construction of `deepKNet`.

Hyperparameter	Values
number of convolution layers	3, 4, 5
convolution channels	64, 128, 256, 512, 1024
dimension of hidden crystal feature vector	256, 512, 1024
number of fully-connected layers	3, 4, 5
size of fully-connected layers	128, 256, 512, 1024
pooling	max, self-attention
number of self-attention layers	0, 2, 4
optimizer	SGD, Adam
initial learning rate	0.01, 0.001
dropout	0, 0.2, 0.4

B.3 Crystal System Classification

Before we learn the physical properties of materials using deepKNet, we first test whether the network architecture could learn to distinguish different crystal systems using the XRD patterns. This simple computer vision task is an important prerequisite before we analyze the structure-property relationship data. If the model is unable to recognize the crystal systems from structural input data, then we may not trust it in making property predictions.

We have a total of 48,524 crystal structures from seven crystal systems for this classification task (Table B.2). The dataset is available from the authors upon reasonable request. With only 3 reciprocal basis vectors visible to the model, we obtain an accuracy of 0.988 without random 3D rotations. As we check the misclassified crystals, we find most of the incorrect predictions are caused by very minor differences in lattice parameters (e.g., 89.9° and 90°, or $10.7\,\text{Å}$ and $10.8\,\text{Å}$), and the threshold of "equivalence" can make a difference in model predictions. However, the overall performance is very impressive, thus we can trust the model in further property learning tasks.

Table B.2: Distribution of structures by crystal system.

Crystal System	Number of Structures
cubic	6,899
tetragonal	6,983
orthorhombic	12,289
hexagonal	4,811
trigonal	3,898
monoclinic	11,120
triclinic	2,524

Once we turn on random 3D rotations with Euler angle between $[-\pi, \pi]$, the accuracy decreases to 0.83. We suspect that the significant performance loss originated from the model failing to identify the "principal axis" of the crystal, which is possibly a limitation of the PointNet-like network architecture. After we reduce the rotation angle to be $[-\frac{1}{4}\pi, \frac{1}{4}\pi]$, we regain an accuracy of 0.98. Interestingly, we later find the property classification tasks are completely immune to the range of random 3D rotations. It is reasonable since the crystal orientation does not matter to the scalar materials properties we examine.

In addition, we could also successfully distinguish trigonal cells from hexagonal cells using the `deepKNet`. We obtain an AUC value of 0.94 with 3 diffraction points, 0.97 with 27 diffraction points, and 0.98 with 125 diffraction points as input. This is an advantage of using 3D diffraction patterns compared with 2D projected versions.

B.4 Baseline Performance for the Metal-Insulator Classification

We find that using only 3 reciprocal basis vectors as input, the metal-insulator classifier is able to achieve an AUC of 0.8. Typically, a binary classifier with AUC 0.8 can be considered as "effective." Our finding here shows that metallicity has a strong dependency on the cell shape and volume. In order to validate this discovery, we build a simple random forest classifier using `scikit-learn` [242] with only two features: the conventional standard cell volume (numerical variable) and the crystal system (categorical variable). The dataset we use is identical to the metal-insulator classification task using the `deepKNet`, where 20% of data is used as the test set, and others are used for training and validation purposes. We use grid search with 4-fold cross validation to select the hyperparameters of the random forest model, as shown in Table B.3. We take the average AUC value of 3 independent runs with different training-validation-test split, and finally obtain an AUC of 0.8, which is identical to the `deepKNet` model performance.

Our findings here establish the baseline performance of the metal-insulator classification. With only the primitive features of cell volume and crystal system (without any further feature engineering), we obtain an AUC of 0.8. Therefore, materials scientists should be comparing the classification performance with 0.8 instead of 0.5. We acknowledge the fact that determining the metallicity of a new material is not a trivial task, yet from existing materials databases, the baseline performance of a metal-insulator classifier is quite high. It remains challenging to interpret whether we really learned some new materials physics or rather learned the database statistics.

Table B.3: Hyperparameters explored in the random forest classifier using the `scikit-learn` package. The selected hyperparamter used to generate the results are bolded.

Hyperparameters	Values
number of estimators	**100**, 120, 150
criterion	gini, **entropy**
maximum depth	6, 8, **12**
maximum number of features	**None**, sqrt, log2
maximum number of leaf nodes	None, 30, **50**

B.5 Standard Deviation of Property Learning Tasks

Table B.4 shows the standard deviations for each property classification task from 3 randomly and independently generated training-validation-test datasets.

Table B.4: Standard deviations from 3 randomly and independently generated training-validation-test datasets for different property classification tasks. Features labeled XRD and ND correspond to X-ray diffraction and neutron diffraction, respectively.

Classification Task	Feature	φ_1	φ_2	φ_3
metal-insulator	XRD	0.001	0.002	0.009
	ND	0.001	0.001	0.005
bulk modulus	XRD	0.003	0.014	0.019
	ND	0.006	0.004	0.003
shear modulus	XRD	0.002	0.015	0.008
	ND	0.006	0.005	0.005
stability	XRD	0.005	0.004	0.009
	ND	0.003	0.024	0.005

B.6 XRD Critical Points of Selected Materials

To help us understand how the `deepKNet` differentiates metals from insulators, we use a small model to facilitate model interpretation. We use 125 diffraction points (Miller indices within range $\{\bar{2},\ \bar{1},\ 0,\ 1,\ 2\}$) for all materials and 32-dimensional crystal feature embedding for the metal-insulator classification task. The convolutional layer dimensions are [4, 8, 16, 32, 32, 32] and the fully-connected layer dimensions are [32, 32, 16, 16], respectively. The critical points of each material as shown in Figure 5.5 of the Chapter 5 are enumerated below. Since the same point may contribute to more than one hidden crystal feature, the number of unique critical points could be less than the embedding dimension.

Critical points of NaCl:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(\bar{1}\ \bar{1}\ 0)$ $(1\ \bar{1}\ 0)$ $(0\ 2\ 0)$

$(0\ 0\ 2)$ $(0\ \bar{2}\ 0)$ $(2\ 0\ 1)$ $(1\ 0\ 2)$ $(0\ 1\ 2)$

$(0\ \bar{2}\ 1)$ $(\bar{1}\ 0\ \bar{2})$ $(2\ 2\ 0)$ $(0\ 2\ 2)$ $(0\ 2\ \bar{2})$

$(0\ \bar{2}\ 2)$ $(\bar{2}\ 0\ 2)$ $(2\ 1\ \bar{2})$ $(1\ \bar{2}\ \bar{2})$ $(\bar{2}\ \bar{2}\ 2)$

Critical points of SiO_2:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(\bar{1}\ \bar{1}\ 1)$ $(1\ \bar{1}\ 1)$ $(\bar{1}\ 1\ \bar{1})$

$(\bar{1}\ 0\ \bar{2})$ $(1\ \bar{2}\ 0)$ $(0\ \bar{2}\ 1)$ $(\bar{1}\ \bar{2}\ 0)$ $(1\ 1\ 2)$

$(1\ \bar{1}\ 2)$ $(0\ 2\ \bar{2})$ $(\bar{2}\ 0\ 2)$ $(2\ \bar{2}\ 0)$ $(2\ 1\ 2)$

$(\bar{2}\ \bar{2}\ \bar{1})$ $(2\ 1\ \bar{2})$ $(1\ \bar{2}\ \bar{2})$ $(\bar{2}\ \bar{1}\ 2)$ $(2\ 2\ 2)$

$(\bar{2}\ \bar{2}\ 2)$

Critical points of Al_2O_3:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(0\ 0\ \bar{2})$ $(2\ \bar{1}\ \bar{1})$ $(\bar{2}\ \bar{2}\ 2)$

$(1\ 0\ 1)$ $(1\ \bar{1}\ 1)$ $(0\ 1\ \bar{2})$ $(\bar{2}\ \bar{1}\ 2)$ $(2\ 2\ 0)$

$(2\ 0\ \bar{2})$ $(1\ 0\ 2)$ $(2\ 2\ 2)$ $(0\ 1\ 2)$ $(\bar{2}\ 2\ \bar{2})$

$(1\ \bar{1}\ 2)$

Critical points of Cu:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(0\ \bar{1}\ 0)$ $(1\ 0\ 1)$ $(0\ 1\ 1)$

$(1\ 1\ 1)$ $(\bar{1}\ \bar{1}\ 1)$ $(\bar{1}\ \bar{1}\ \bar{1})$ $(1\ \bar{1}\ 1)$ $(2\ 0\ 0)$

$(1\ 0\ \bar{2})$ $(2\ 2\ 0)$ $(0\ 2\ 2)$ $(0\ 2\ \bar{2})$ $(0\ \bar{2}\ 2)$

$(\bar{2}\ 0\ 2)$ $(2\ 1\ 2)$ $(2\ \bar{2}\ \bar{1})$ $(1\ \bar{2}\ \bar{2})$ $(\bar{2}\ \bar{2}\ 2)$

$(\bar{2}\ \bar{2}\ \bar{2})$

Critical points of Ag:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(1\ 0\ 1)$ $(\bar{1}\ \bar{1}\ 0)$ $(1\ 1\ 1)$

$(\bar{1}\ \bar{1}\ \bar{1})$ $(1\ \bar{1}\ 1)$ $(\bar{1}\ \bar{1}\ 1)$ $(2\ 0\ 0)$ $(1\ 0\ \bar{2})$

$(2\ 2\ 0)$ $(0\ 2\ 2)$ $(2\ \bar{2}\ 0)$ $(0\ 2\ \bar{2})$ $(\bar{2}\ 0\ 2)$

$(2\ 1\ 2)$ $(1\ \bar{2}\ \bar{2})$ $(2\ \bar{2}\ \bar{1})$ $(\bar{2}\ 1\ 2)$ $(2\ 2\ 2)$

$(\bar{2}\ \bar{2}\ \bar{2})$ $(\bar{2}\ 2\ 2)$ $(\bar{2}\ \bar{2}\ 2)$

Critical points of Au:

$(0\ 0\ 0)$ $(0\ 0\ \bar{1})$ $(1\ 0\ 1)$ $(0\ 1\ 1)$ $(\bar{1}\ \bar{1}\ 0)$

$(1\ 1\ 1)$ $(\bar{1}\ \bar{1}\ \bar{1})$ $(1\ \bar{1}\ 1)$ $(\bar{1}\ \bar{1}\ 1)$ $(2\ 0\ 0)$

$(0\ 0\ 2)$ $(1\ 0\ \bar{2})$ $(0\ 2\ \bar{2})$ $(\bar{2}\ 0\ 2)$ $(2\ 1\ 2)$

$(2\ \bar{2}\ \bar{1})$ $(1\ \bar{2}\ \bar{2})$ $(\bar{2}\ 1\ 2)$ $(2\ 2\ 2)$ $(\bar{2}\ \bar{2}\ \bar{2})$

$(2\ 2\ \bar{2})$ $(\bar{2}\ 2\ 2)$ $(\bar{2}\ \bar{2}\ 2)$

B.7 Critical Point Distribution from Neutron Diffraction Data

We also train a small network using neutron diffraction data to interpret how the model distinguishes metals from insulators. We use ND data with $n = 125$ diffraction points and `deepKNet` with a 32-dimensional hidden crystal vector, i.e., the same as that described in the manuscript. The model performance is quite poor with an AUC of 0.84. It misclassifies NaCl as metal. Interestingly, Figure B.1 shows the critical point distributions are quite different from the XRD model appearing in Figure 5.5.

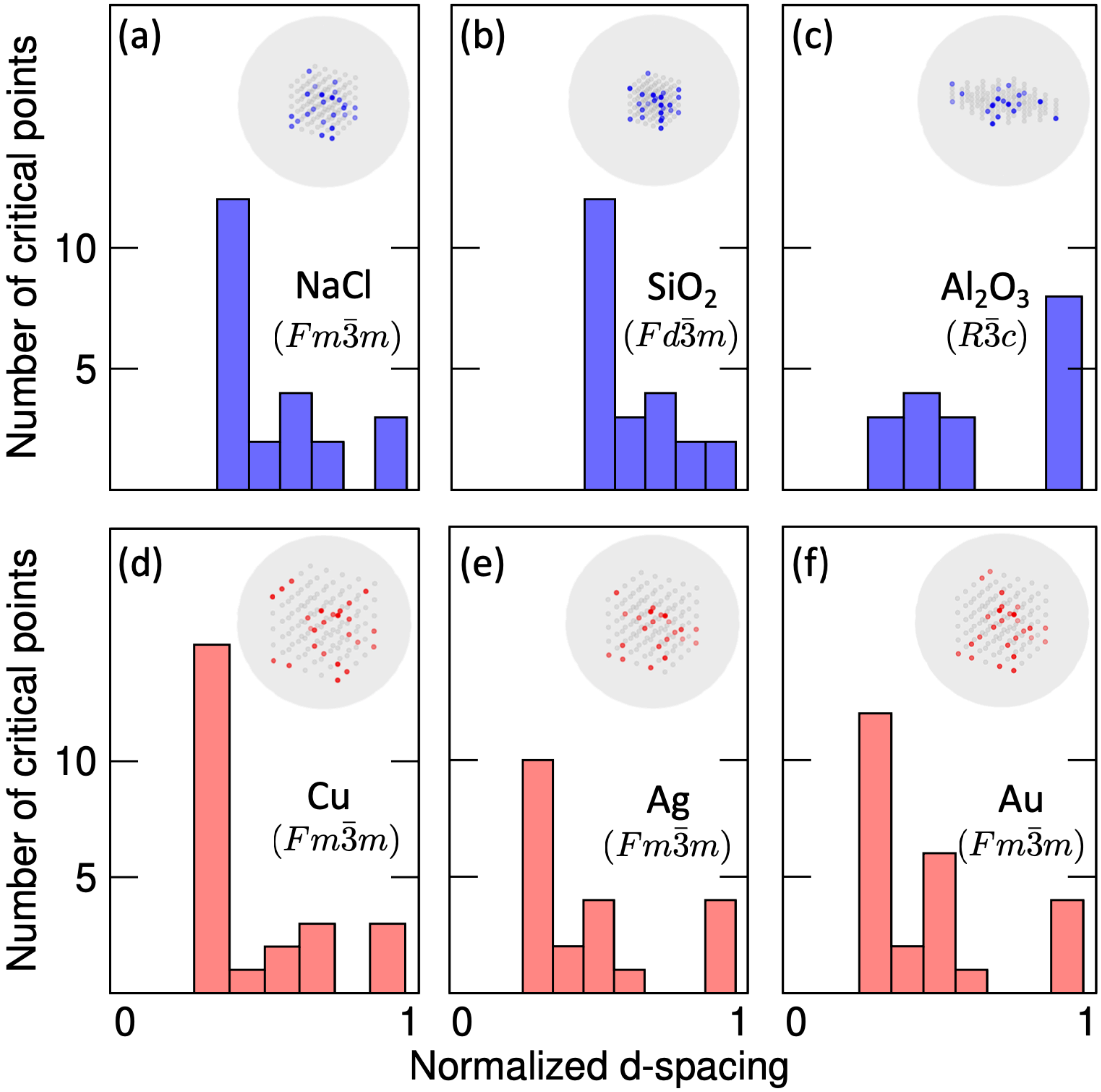


Figure B.1: Distribution of neutron diffraction critical diffraction points with normalized interplanar d_{hkl} spacings of a few common insulators (NaCl, SiO$_2$, Al$_2$O$_3$) and metals (Cu, Ag, Au). The critical points in the limiting (gray) sphere are those that contribute to the final crystal feature vector after `max` pooling, and are marked with blue for insulators, and red for metals, respectively. Non-critical points are represented with light gray points.

VITA

Yiqun (Raymond) Wang

raywang.sci@gmail.com | +1 (224) 714-8582 | raymond931118@GitHub

EDUCATION

Northwestern University, Evanston, IL, USA 09/2016 - 06/2021 (*Expected*)
Ph.D. in Chemistry

Fudan University, Shanghai, China 09/2012 - 06/2016
B.S. in Chemistry

PROFESSIONAL EXPERIENCE

Northwestern Research Computing Services, Evanston, IL, USA 02/2019 - Present
High Performance Computing Consultant

- Developed and open-sourced a library of distributed machine learning regression and classification model templates on multiple code platforms including PyTorch, scikit-learn, and Spark MLlib for in-house supercomputer users
- Hosted quarterly data science workshops on Linux, Apache Spark, and parallel programming
- Realized distributed in-memory processing of large Linux system files by deploying a standalone Spark cluster

University of California, Los Angeles, Los Angeles, CA, USA 07/2015 - 09/2015
Summer Research Intern

- Performed geometric optimization and ground state wave function calculation on a novel Boron catalyst
- Simulated the photoexcitation process using time-dependent density functional theory simulations, results validated through experimental characterizations
- Implemented Markov chain Monte Carlo simulations in Matlab to study the atomic coalescence process on Pt surface

PUBLICATIONS

1. **Y. Wang**, X. Zhang, F. Xia, E. A. Olivetti, R. Seshadri, and J. M. Rondinelli, "Learning the Crystal Structure Genome for Property Classification", *arXiv:2101.01773* (2021)
2. **Y. Wang**, A. Iyer, W. Chen, and J. M. Rondinelli, "Featureless adaptive optimization accelerates electronic materials design", *Appl. Phys. Rev.* 7, 041403 (2020)
3. **Y. Wang**, D. Puggioni, and J.M.Rondinelli, "Assessing exchange-correlation functional performance in the chalcogenide lacunar spinels GaM_4Q_8 (M = Mo, V, Nb, Ta; Q = S, Se)", *Phys. Rev. B* 100, 115149 (2019)
4. **Y. Wang**, N. Wagner, and J.M.Rondinelli, "Symbolic Regression in Materials Science", *MRS Commun.*, 9(3), 793 (2019)
5. M. S. Messina, J. C. Axtell, **Y. Wang**, *et al.*, "Visible-Light-Induced Olefin Activation Using 3D Aromatic Boron-Rich Cluster Photooxidants", *J. Am. Chem. Soc.* 138, 22, 6952 (2016)