Summer 2023

Predicting Material Structures and Properties Using Deep Learning And Machine Learning Algorithms

Yuqi Song

Predicting Material Structures and Properties Using Deep Learning and Machine Learning Algorithms

by

Yuqi Song

Bachelor of Software Engineering
Chongqing University 2016

Master of Software Engineering
Chongqing University 2019

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2023

Accepted by:

Jianjun Hu, Major Professor

Yan Tong, Committee Member

Forest Agostinelli, Committee Member

Qi Zhang, Committee Member

Ming Hu, Committee Member

Ann Vail, Dean of the Graduate School

Acknowledgments

I would like to express my deepest appreciation to my supervisor Dr. Jianjun Hu, who constructively guided and generously provided knowledge and expertise throughout my PhD program. His enthusiasm for research, passion for coding, and his invaluable insights deeply impact my research work and career planning. I could not have undertaken this interesting journey without him.

I would also like to extend my deepest gratitude to my dissertation committee members: Dr. Yan Tong, Dr. Forest Agostinelli, Dr. Qi Zhang and Dr. Ming Hu. I appreciate their time and useful suggestions on my work which are very important and valuable in the process of completing my dissertation.

I am also grateful to all my collaborators and classmates at the University of South Carolina for their help in various aspects of study and life, these are the intangible and valuable treasures of my last four years. Especially, I want to thank Dr. Yong Zhao, and Dr. Steph-Yves Louis for their practical suggestions and helpful contributions to my research and career development, as well as Dr. Edirisuriya M. Dilanga Siriwardane, Rongzhi Dong, Nihang Fu, Lai Wei, Qinyang Li, Sadman Sadeed Omee, and Rui Xin for their helpful discussion and advice to my research.

The completion of my dissertation would not have been possible without the support and nurturing of my parents and my husband Xin Zhang. Their belief in me has kept my spirits and motivation high during these years. I would also like to thank my puppy 'Captain', for all the entertainment and emotional support.

Abstract

Discovering new materials and understanding their crystal structures and chemical properties are critical tasks in the material sciences. Although computational methodologies such as Density Functional Theory (DFT), provide a convenient means for calculating certain properties of materials or predicting crystal structures when combined with search algorithms, DFT is computationally too demanding for structure prediction and property calculation for most material families, especially for those materials with a large number of atoms. This dissertation aims to address this limitation by developing novel deep learning and machine learning algorithms for effective prediction of material crystal structures and properties. Our data-driven machine learning modeling approaches allow to learn both explicit and implicit chemical and geometric knowledge in terms of patterns and constraints from known materials and then exploit them for efficient sampling in crystal structure prediction and feature extraction for material property prediction.

In the first topic, we present DeltaCrystal, a new deep learning based method for crystal structure prediction. This data-driven algorithm learns and exploits the abundant atom interaction distribution of known crystal material structures to achieve efficient structure search. It first learns to predict the atomic distance matrix for a given material composition based on a deep residual neural network and then employs this matrix to reconstruct its 3D crystal structure using a genetic algorithm. Through extensive experiments, we demonstrate that our model can learn the implicit interatomic relationships and its effectiveness and reliability in exploiting such information for crystal structure prediction. Compared to the global optimization based CSP method, our algorithm achieves better structure prediction performance for more complex crystals.

In the second topic, we shift our focus from individually predicting the positions of atoms in each material structure to the idea of crystal structure prediction based on structural polyhedron motifs based on the observation that these atom patterns appear frequently across different crystal materials with high geometric conservation, which has the potential to significantly reduce the search complexity. We extract a large set of structural motifs from a vast collection of material structures. Through the comprehensive analysis of motifs, we uncover common patterns and motifs that span across different materials. Our work represents a preliminary step in the exploration of material structures from the motif point of view and exploiting such motif for efficient crystal structure prediction.

In the third topic, we propose a machine learning based framework for discovering new hypothetical 2D materials. It first trains a deep learning generative model for material composition generation and trains a random forest-based 2D materials classifier to screen out potential 2D material compositions. Then, a template-based element substitution structure prediction approach is developed to predict the crystal structures for a subset of the newly predicted hypothetical 2D formulas, which allows us to confirm their structural stability using DFT calculations. So far, we have predicted 101 crystal structures and confirmed 92 2D/layered materials by DFT formation energy calculation.

In the last topic, we focus on machine learning models for predicting material properties, including piezoelectric coefficients and noncentrosymmetric of nonlinear optical materials, as they play important roles in many important applications, such as laser technology and X-ray shutters. We conduct a comprehensive study on developing advanced machine learning models and evaluating their performance for predicting piezoelectric modulus from materials' composition/structures. Next, we train several prediction models based on extensive feature engineering combined with machine learning models and automated feature learning based on deep graph neural networks. We use the best model to predict the piezoelectric coefficients for 12,680 materials and report the top 20 potential high-performance piezoelectric materials. Similarly, we develop machine learning models to screen potential noncentrosymmetric materials from 2,000,000 hypothetical materials generated by our material composition generative design model and report the top 80 candidate noncentrosymmetric nonlinear materials.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Materials play an important role in our daily life: clean water, fresh air, smart living and transportation, and so on. With the development of emerging science and technologies, as well as the increasing demands of various aspects, computational discovery of novel functional materials is a non-trivial task for a variety of industries, such as electric vehicles, cell phones, and quantum computing hardware [128]. The flourishing of these industries relies on many materials with special properties or/and structures. For example, piezoelectric materials serve as crucial units for energy-harvesting equipment or as active parts of sensors and motors, because of their special piezoelectric effect [151]. Two-dimensional(2D) materials have the potential to create new electronics and technologies such as spintronics, catalysis, and membranes owing to their exotic vibrational, electronic, optical, magnetic, and topological behaviors [98].

Traditionally, material researchers primarily discovered and analyzed materials and their properties based on experimental observations [152]. However, material discovery, optimization, and property prediction are time-consuming, labor-intensive, complex and expensive. Researchers should use more scientific and effective methods to conduct research. Recently, material researchers focus on high-throughput simulation methods such as Density Functional Theory (DFT) [37] which uses the quantum mechanical laws to find out the electrical properties of atoms, molecules and solids.

Furthermore, with the availability of faster communication technologies, vast amounts of materials data can be easily collected, there are several large-scale open-source materials databases that present opportunities for data-driven materials informatics [124], such as Material Project Database [67], the Open Quantum Materials Database (OQMD) [79], and Inorganic Crystal Structure Database (ICSD) [13].

In the footsteps of emerging computation technologies which include machine learning (ML) [72], deep learning (DL) [87] and high-performance computing[25], material science can benefit from those advanced approaches to efficiently and effectively discover new materials. Specifically, inspired by cutting-edge machine learning and deep learning methods which have achieved tremendous success in many fields such as computer vision [59] and natural language processing [101], there are many data-driven works are designed to determine material properties and structures that are hard to measure or compute using traditional methods. Currently, there are still many challenging tasks in the field of materials informatics. For example, how to predict materials' crystal structures only rely on their compound information, and how to discover new target materials more efficiently. In this work, we aim to explore three problems about predicting material structures and properties and make some contributions to material informatics [48].

1.2 Scope of the proposed research

In this dissertation, we focus on the following four topics:

1. Crystal structures are fundamental and important descriptors for inorganic compounds in the material science community [112]. Given only a material composition or formula, predicting its crystal structure is a promising and challenging task because most traditional *ab initio* [3] methods rely on global search with first principle free energy calculation which is time-consuming. Inspired by the recent success of deep learning approaches for protein structure prediction,

they use pairwise of amino acids to describe a single 3D structure [7]. Therefore, exploiting the abundant atom interaction distribution in existing known crystal structures, we present a new knowledge-based solution DeltaCrystal. It predicts the atomic distance matrix of a target crystal material and then employs this matrix to reconstruct its 3D crystal structures. Through a wealth of experiments, we demonstrate the effectiveness and reliability of plentiful inter-atomic relationships for structure prediction.

2. In material structures, a motif refers to a recurring and characteristic pattern or arrangement of atoms within the crystal lattice. Motif represents a higher-level atom pattern, it can be considered as building blocks or fundamental units that contribute to the overall structure and properties of the material. Analyzing and understanding motifs allows us to uncover common patterns, establish relationships between structure and properties, and facilitate the design of new materials with specific functionalities. Therefore, in this topic, we design a method for extracting motifs from existing material structures. We empower this method to extract a large number of motif structures, and subsequently conducted comprehensive statistical analysis on these structures. This approach allows us to gain insights into the prevalent motifs present in diverse materials and understand their significance in relation to the overall structures.

3. Two-dimensional (2D) materials have emerged as promising functional materials with many applications such as semiconductors and photovoltaics because of their unique optoelectronic properties [98]. Although several thousand 2D materials have been screened in existing materials databases, discovering new 2D materials remains to be challenging [164]. Herein, we propose a deep learning generative model [46] for composition generation combined with a random forest based 2D materials classifier to discover new hypothetical 2D materials. Fur-

thermore, a template-based element substitution structure prediction approach is developed to predict the crystal structures of a subset of the newly predicted hypothetical formulas, which allows us to confirm their structural stability using DFT calculations. Our results show that generative machine learning models provide an effective way to explore the vast chemical design space for new 2D materials discovery.

4. Piezoelectric materials are widely used in many industries and our daily life [156]. However, discovering high-performance piezoelectric materials is much more challenging than other material properties (such as formation energy and band gap) [113]. Here, we propose a comprehensive study on designing and evaluating advanced machine learning models for predicting piezoelectric modulus from materials' composition/structures. We train prediction models based on extensive feature engineering combined with machine learning models and automated feature learning based on deep graph neural networks [158]. We also use it to predict the piezoelectric coefficients for 12,680 materials and report the top 20 potential high-performance piezoelectric materials. Noncentrosymmetric materials also play a critical role in many important applications such as laser technology, communication systems, quantum computing, cybersecurity, etc [56]. However, the experimental discovery of new noncentrosymmetric materials is extremely difficult. Here we present a machine learning model that could predict whether the composition of a potential crystalline structure would be centrosymmetric or not. By evaluating a diverse set of composition features calculated using matminer feature package [154] coupled with different machine learning algorithms, we find that random forest classifiers give the best performance for noncentrosymmetric material prediction. We apply our ML model to screen potential noncentrosymmetric materials from 2,000,000 hypothetical materials generated by the inverse design engine and report the top 20 candidate

noncentrosymmetric materials.

1.3 Structure of this dissertation

In Chapter 2, we briefly introduce some related deep learning and machine learning methods as well as several prevalent research trends in Material Informatics. In Chapter 3, we propose a new knowledge-based method DeltaCrystal to predict crystal structures by exploiting the abundant atom interaction distribution. In Chapter 4, we extract and analysis of structural motifs from crystal materials. In Chapter 5, we design a model to discover new hypothetical 2D materials based on deep learning generative model for composition generation and the random forest method for predicting class. Chapter 6 focuses on predicting material properties, including piezoelectric coefficients and noncentrosymmetric. Finally, we conclude our research work in Chapter 7.

Chapter 2

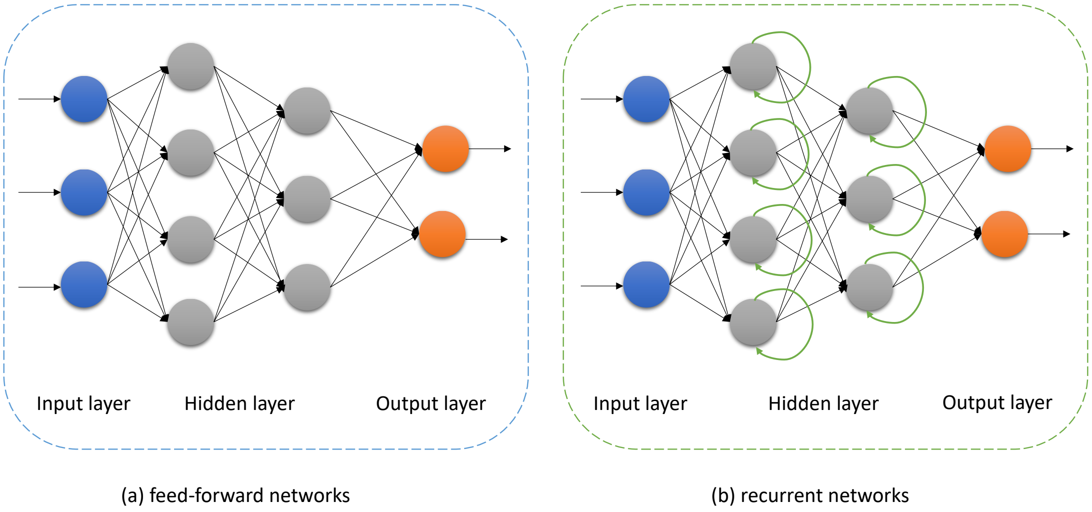
Background

Machine learning and deep learning have gained significant traction in the field of materials science, finding applications in various research areas. In this chapter, we provide an overview of relevant deep learning and machine learning techniques employed in material informatics. We also introduce the basic workflow of material informatics and delve into some main research tasks, such as crystal structure prediction and material properties prediction.

2.1 Deep Learning methods

Deep learning methods have driven rapid development in natural language processing, computer vision, and data mining [72]. Various deep learning methods have been used effectively in material informatics, these algorithms learn from existing material data, which include input information represented by descriptors and output responses that are usually material properties or/and performance of interest, in order to design and discover new materials with targeted properties or structures [123]. In this section, we present a few important and typical algorithms of related work.

Inspired by biological neural networks, in 1943, McCulloch and Pitts first proposed a binary threshold unit as a computational model for an artificial neuron [70], which promoted the development of neural network research. Artificial Neural Networks (ANNs) [66] are the most common neural networks which have been used to solve a wide variety of issues. ANNs can be regarded as weighted directed graphs in which neurons are nodes and directed edges (with weights) are connections between neuron outputs and neuron inputs. According to whether there are feedback connections in network structures, ANNs divide into feed-forward neural networks (FFNNs) and recurrent neural networks (RNNs), as shown in Figure 2.1. Each network usually consists of an input layer, hidden layers and an output layer.



(a) feed-forward networks

(b) recurrent networks

Figure 2.1 The comparison between FFNNs and RNNs, in (a) FFNNs there is only one direction for the data to move, (b) RNNs in which loops occur because of feedback connections.

Convolutional Neural Networks (CNNs) [51] are usually applied to image classification algorithms, which take images as input, assign learnable weights and biases to various aspects/objects in the image, and then this model can differentiate one from the other. CNNs are able to successfully capture the spatial and temporal dependencies in an image through the application of relevant filters. In a convolutional neural network, like Figure 2.2, the hidden layers include layers that perform convolutions which typically perform a dot product of the convolution kernel with the layer's input matrix.

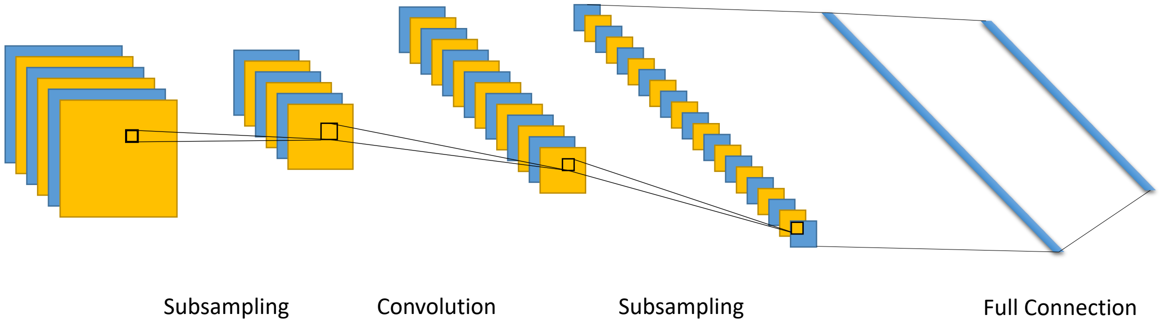


Figure 2.2 An example of convolutional neural networks

Residual Neural Networks (ResNets) [60] stack residual blocks on top of each other to form a network by skip connections, this design had a profound influence on how to build deep neural networks, as shown in Figure 2.3. The skip connections between layers add the outputs from previous layers to the outputs of stacked layers which results in the ability to train much deeper networks than what was previously possible.

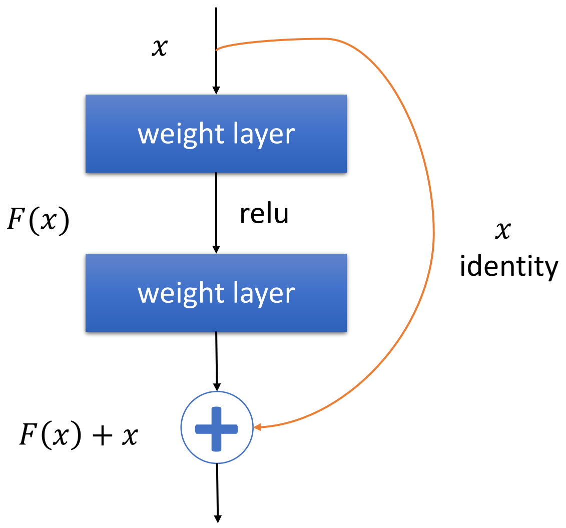


Figure 2.3 A building block in ResNets.

Graph neural networks (GNNs) [168] directly take graphs (composed of vertexes and edges) as the input. Graph neural networks have wide applications in various domains such as social networks, knowledge graphs, recommender systems, and life science. One of the significant advantages of graph neural networks is their capability to learn or model dependencies (interactions) between nodes in a graph which is highly suitable for modeling interactions between atoms in materials [159].

Generative Adversarial Networks (GANs) was proposed to solve the generative modeling problem by investigating a set of training examples and learning the probability distribution that generated them [47]. The framework of GANs is shown in Figure 2.4 which mainly includes the generative model (G) and the discriminative model (D). G aims to capture the training data distribution with random noise, and D focuses on distinguishing the sample from either the training set or generator with an estimated probability, while GANs try to find an equilibrium solution between G and D.

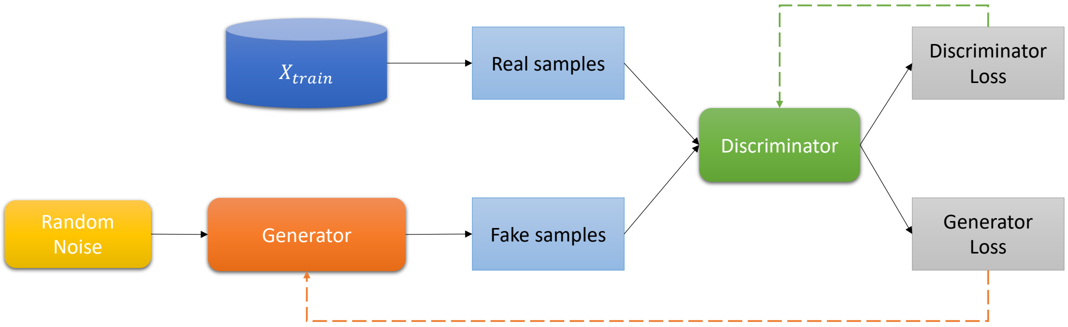


Figure 2.4 Schematic illustration of GANs

2.2 Machine Learning methods

Machine learning is developed along with the development of computer science, which enables algorithms to learn from data and anticipate outcomes based on data [102]. Currently, machine learning plays an essential role in data analyses and data prediction tasks [72]. As the foundation of the data-driven study of materials science, machine learning algorithms were introduced with various intentions. In this section, several related machine learning methods [102] are introduced.

Random Forest (RF) [90] is a kind of a supervised bagging ensemble learning algorithm. The idea behind random forests is to exploit the wisdom of the group. RF builds many decision trees in a random way with low correlation among them. After building the forest, when a new sample needs to be classified, each decision tree makes a judgment separately to vote which category the sample belongs to. Random forest improves the prediction accuracy without significantly increasing the amount of computation, and it is relatively robust to unbalanced data.

The Support Vector Machine (SVM) algorithm is another machine learning algorithm for classification, regression, and other tasks that works by creating a hyperplane (s) in a high- or infinite-dimensional space. Support vector machines are very effective in high-dimensional spaces, especially when the number of dimensions is greater than the number of samples. This method has been widely used in material science research [142, 1, 91].

The t-Distributed Stochastic Neighbor Embedding (t-SNE) method is a kind of machine learning approach that usually be used for data visualization by reducing high-dimensional data to two or three dimensions [144], it could address nonlinear dimensionality reduction tasks and also is a great tool to understand high-dimensional datasets. This data analysis tool has been applied in a variety of fields, such as computer security, cancer biology, and bioinformatics [9].

2.3 Material Informatics

With the rapid developments to improve the accuracy and efficiency of experimental and computational investigative approaches, there are huge volumes of collected data that led the field of materials science into data-driven scientific research, which promotes the advancement of computational techniques, and big-data analysis in material informatics [152]. Materials informatics usually utilizes statistical and machine learning methods to learn the relationship between materials and their physical and chemical representations, such as formulas, crystal structures and chemical properties.

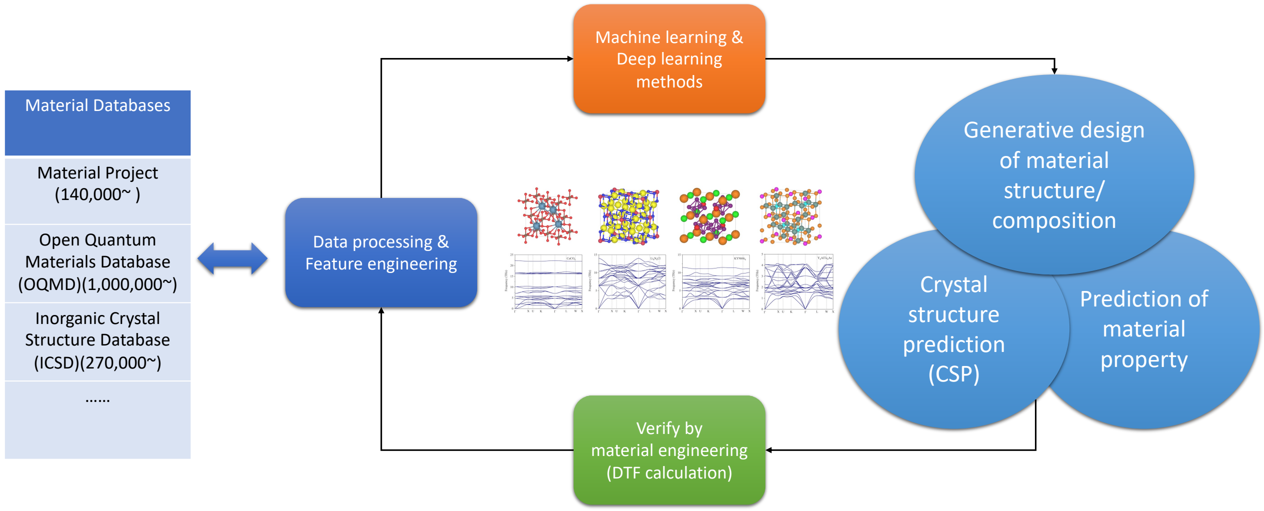

Material Databases
Material Project (140,000~)
Open Quantum Materials Database (OQMD)(1,000,000~)
Inorganic Crystal Structure Database (ICSD)(270,000~)
......
Data processing & Feature engineering
Machine learning & Deep learning methods
Generative design of material structure/ composition
Crystal structure prediction (CSP)
Prediction of material property
Verify by material engineering (DTF calculation)

Figure 2.5 Workflow of material informatics

The goal of materials informatics is to screen thousands of compounds at a much faster rate for potential new industrial materials [128].

2.3.1 The basic workflow of material informatics

Figure 2.5 depicts the fundamental workflow of material informatics. The process begins with researchers acquiring and preparing material data, which encompasses material formulas, properties, and structures. To facilitate data analysis, various representations of materials data are explored. These datasets are often sourced from publicly available material databases such as Material Project [67], OQMD (Open Quantum Materials Database) [79], and others. Subsequently, advanced data analysis techniques and deep learning/machine learning (DL/ML) models are employed to extract valuable insights from the vast material space. These methods enable predictions of material properties and structures, offering valuable guidance for further research. To validate the predicted outcomes, material engineering methods are applied, such as density functional theory (DFT) [**\cite {kohn1999nobel}**]. These experimental techniques serve to verify the accuracy and reliability of the predicted results, ensuring their practical applicability in real-world scenarios.

Currently, those data-driven material informatics can be divided into three categories: (1) constructing hypothetical crystal structures by learning the compositional space of known materials [108, 63, 77, 22]; (2) forecasting properties and recognizing special materials from existing databases [153, 159, 166, 43]; (3) generative design of material structure and composition. These three issues can be independent subjects, they also can be interconnected. For example, if we wanna design a novel superconductive material, we can generate potential novel compositions, and then predict their crystal structures, after that, predict the highly-related material properties, according to these predicted results, we pick up the final candidates to material scholars, they calculate and verify those formulas and structures, which indeed speed up the discovery of new materials.

In the following section, we introduce some remarkable achievements that are related to our work in crystal structure prediction and material properties prediction.

2.3.2 Crystal structure prediction

Crystal structures play an important role in materials because they fundamentally determine the properties of materials. Therefore, predicting structure is one of the most important and essential works in material science, such as predicting properties based on structural information, and exploring new materials with target structures and properties. While crystal structure prediction (CSP) is a notoriously challenging topic because material scientists have to discover crystal structures with the lowest free energy for a given chemical composition under specific pressure-temperature circumstances [97, 157, 95, 131]. However, most global free energy search-based algorithms have an obvious obstacle that limits their successes [112, 162] due to their dependence on the costly density functional theory (DFT) calculations of free energies for sampled structures. Hence, how to efficiently predict crystal structures becomes a key issue [111, 95]. To improve the sampling efficiency, a variety of strategies have been proposed such as exploiting symmetry[122], clustering, and machine-learning interatomic potentials with active learning [121]. However, the scalability of these approaches remains a challenging issue.

Recently, there are some emerging works using deep learning methods to predict material structures. Ryan et al. [131] reconstructed the crystal structure prediction problem as predicting the likelihoods of particular atomic sites in the structure. Their model is able to distinguish chemical components based on the topology crystallographic environment. They also use the model to analyze templates derived from known crystal structures in order to predict the likelihood of forming new compounds. Cheng et al. [22] lately proposed a framework that uses a graph network model to connect crystal structures and their formation enthalpies and then merged this model with an optimization algorithm for CSP. Although they predicted 29 crystal structures, the main disadvantage of this work is that it can only optimize some materials with simple composition, and for complex compounds, sometimes it is impossible to obtain good prediction results or even predict their structures.

2.3.3 Material properties prediction

Material properties give researchers and industries insight into the material's mechanical performance under specific conditions [104], such as hardness/softness, the density of the particles, fracture toughness, resistivity and thermal expansion, which are measured and determined by their structures [48]. Data-driven informatics methods are becoming useful to determine material properties that are hard to measure or compute using traditional methods, due to the cost, time or effort involved, but for which reliable data either already exists or can be generated for at least a subset of the critical cases. There are many approaches that typically employ the structural or composition features of known materials to predict target properties [32]. According to their representations, these methods can divide into two types: (1) composition-based properties prediction and (2) structure-based properties prediction.

With a sufficient amount of dataset, it has been shown that composition-based ML models can achieve highly accurate models for formation energy [153, 68] and band gap predictions [172]. However, some of those high-performance models are very likely due to the high redundancy of datasets that have many highly similar samples. Three recent solid benchmark evaluations have clearly shown that the structure-based prediction models most often outperform those composition models [32, 12, 35]. For example, Bartel et al. [12] showed that composition models failed to distinguish inorganic materials' relative stability. Instead, they found that including structure in the representation can lead to non-incremental improvement in stability predictions, which serves as a strong endorsement for structural models.

Chapter 3

Distance matrix based crystal structure prediction using deep residual neural networks

3.1 Introduction

Computational discovery of novel functional materials has enormous potential in transforming a variety of industries such as mobile communication, electric vehicles, quantum computing hardware, and catalysts [112]. Compared to traditional Edisonian or trial-and-error approaches which usually strongly depend on the expertise of the scientists, computational materials discovery has the advantage of efficient search in the vast chemical design space. Among these methods, inverse design[173, 76], generative machine learning models [28, 15, 76, 108, 125], and crystal structure predictions [44, 109, 86] are among the most promising approaches for new materials discovery.

Crystal structure prediction (CSP) is a notoriously hard problem [97, 157, 95] since scholars have to find a crystal structure with the lowest free energy for given chemical composition (or a chemical system such as Mg-Mn-O with variable composition) at given pressure-temperature conditions. With the crystal structure of a chemical substance, many physicochemical properties can be predicted reliably and routinely using first-principle calculation or machine learning models [159]. It is assumed that lower free energy corresponds to the more stable arrangement of atoms. The CSP approach for new materials discovery is especially appealing due to the efficient sampling algorithm that generates diverse chemically valid candidate compositions with low free energies[28]. CSP algorithms based on evolutionary algorithms [110] and particle swarm optimization [150] have led to a series of new materials discoveries [111, 112, 149]. However, these global free energy search-based algorithms have a major obstacle that limits their successes to relatively simple crystals [112, 162] (mostly binary materials with less than 20 atoms in the unit cell[112, 149]) due to their dependence on the costly density functional theory (DFT) calculations of free energies for sampled structures. With a limited DFT calculations budget, how to efficiently sample the atom configurations becomes a key issue [111, 95]. To improve the sampling efficiency, a variety of strategies have been proposed such as exploiting symmetry[122] and pseudosymmetry[95], smart variation operators, clustering, and machine-learning interatomic potentials with active learning [121]. However, the scalability of these approaches remains an unsolved issue.

With the mature development of deep learning techniques, a number of studies use those novel methods in the materials science area [4], especially in material property prediction [92], material discovery [139], and material design[78]. Recently, there are also several emerging works using deep learning methods to predict material structures. Ryan et al. [131] reconstructed the crystal structure prediction problem as predicting the likelihoods of particular atomic sites in the structure. The trained model successfully distinguishes chemical components based on the topology of their crystallographic environment. They use the model to analyze templates derived from the known crystal structures in order to predict the likelihood of forming new compounds. Cheng et al. [22] lately proposed a framework that uses a graph network model to connect crystal structures and their formation enthalpies and then merged this model with an optimization algorithm for CSP. Although they predicted 29 crystal structures, the main weakness of the study is the failure to predict structures of complicated chemical formulas.

Crystal structure prediction in material and protein structure prediction (PSP) in bio-informatics have some similarities [4]. (1) PSP aims to construct the three-dimensional (3D) protein structure from a protein only given its amino acid sequence [155], while CSP seeks to identify a crystal structure with the lowest free energy for certain chemical composition. (2) Features and functions of given protein and material can be explored more effectively if we uncover their structures. (3) These tasks both are time-consuming and expensive for lab experiments, calculation, and verification.

Accordingly, we investigated years of painstaking efforts on protein structure prediction [33, 85, 65] and were motivated by the recent breakthrough of deep learning in PSP with contact map [3] and DeepMind's AlphaFold [7, 167, 137, 73]. For a protein 3D structure, a contact map is a binary form of the distance matrix with a distance threshold, while the distance is calculated between every pair of residues [33]. Adhikari et al.[3] proposed DNCON2, an *ab initio* protein contact maps predictor based on two-level deep convolutional neural networks: at the first level, they used five convolution neural networks (CNNs) with multiple-distance thresholds to predict preliminary contact probabilities as the additional features in next step; then, the other CNN was used to predict the final contact probability map. As a distance map can reveal more structural information than a contact map, in [137], they presented the AlphaFold, a deep learning model which achieved high accuracy even though there are fewer homologous sequences in a sequence.

Inspired by Alphafold's success in protein structure prediction while crystal structure prediction and protein structure prediction have many previously mentioned similarities, we aim to explore the relationship between each pair of atoms. Because currently in CSP work, there is no such concept as co-evolution or correlated mutation relationships among atoms and there is no existing method for crystal distance matrix prediction. Furthermore, there indeed is a large number of physical or chemical rules that determine whether or not two atoms can be bonded together, which leads to the predictability of the distance-matrix of crystal structures. Based on this foundation, in this study, we propose a novel distance matrix-based method for crystal structure prediction. The block diagram of which is illustrated in Fig 3.1. We show that it is possible to make accurate predictions about the structure for a given compound by training a deep neural network.

Our contributions can be summarized as follows:

- We design DeltaCrystal, which is a distance matrix-based method for predicting crystal structures from abundant atomic pair knowledge.

- We conduct extensive experiments to demonstrate DeltaCrystal's competitive effectiveness and reliability: it can achieve good performance not only for materials with simple compositions, but also those materials with complex compositions.

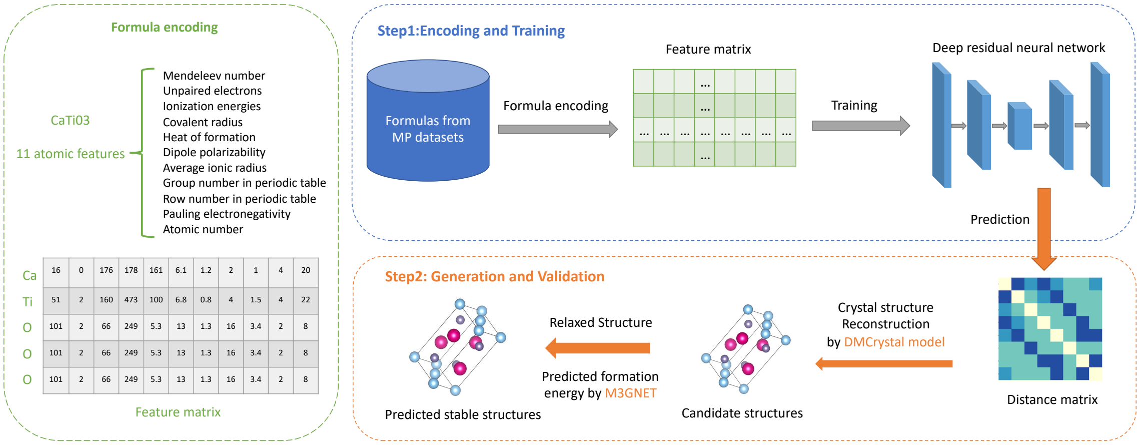
3.2 Methods

For crystal structure prediction tasks, mastering the relationships between atoms is an important factor. Therefore, we focus on exploring the pairwise atomic distance matrices and then reconstructing the crystal structures. We describe at a high-level our proposed method (as shown in Fig 3.1). First, with regard to each compound collected from Materials Project, we construct the feature matrix based on its formula and 11 chemical characteristics (Table 3.1, and then, we train the deep residual neural network. In the second stage, a new formula's distance matrix is predicted by the trained model, and its crystal structure is generated by the atomic coordinate reconstruction algorithm DMCCrystal [64]. Furthermore, we use M3GNET [18] to relax those predicted crystal structures and predicted their formation energies, so that we can pick the stable structures.

3.2.1 Feature matrix encoding

Representing material structure information in an appropriate format to learn atomic interaction is an essential and crucial preparation step, because atomic interactions are related to structural space. A few research demonstrated that many atomic characteristics are related to ionic bond. Therefore, we use 11 atomic features to encode the structural information of each element in each material as input for the deep neural network model.

Specifically, each element is represented by 11 chemical descriptors (see Table 3.1 for detailed) including Mendeleev Number, unpaired electrons, ionization energies,


Formula encoding
CaTi03
11 atomic features
Mendeleev number
Unpaired electrons
Ionization energies
Covalent radius
Heat of formation
Dipole polarizability
Average ionic radius
Group number in periodic table
Row number in periodic table
Pauling electronegativity
Atomic number
Feature matrix
Step1:Encoding and Training
Formulas from MP datasets
Formula encoding
Feature matrix
Training
Deep residual neural network
Prediction
Step2: Generation and Validation
Distance matrix
Crystal structure Reconstruction by DMCrystal model
Candidate structures
Relaxed Structure
Predicted formation energy by M3GNET
Predicted stable structures

Ca	16	0	176	178	161	6.1	1.2	2	1	4	20
Ti	51	2	160	473	100	6.8	0.8	4	1.5	4	22
O	101	2	66	249	5.3	13	1.3	16	3.4	2	8
O	101	2	66	249	5.3	13	1.3	16	3.4	2	8
O	101	2	66	249	5.3	13	1.3	16	3.4	2	8

Figure 3.1 The DeltaCrystal framework for distance matrix-based crystal structure prediction. In the first step, we encode the material feature matrix according to atom features and then train the deep neural network for predicting the distance matrix. In the second step, the crystal structure will be derived by DMCrystal and then relaxed by the M3GNET model, the stable structure will be picked according to predicted formation energies.

covalent radius, heat of formation, dipole polarizability, average ionic radius, group number and row number in the periodic table, pauling electronegativity, and atomic number. These 11 atomic features are represented in each row of the feature matrix.

Each column of the feature matrix represents an element symbol in its unit cell, where L is the number of atoms. We set the maximum number of atoms to be 12, and if the number of atoms is less than 12, the other empty position padding by 0. Hence, the dimension of each feature matrix is $12 * 11$. Since the atomic number of the formula most is less than 12, in order to maintain dimensional consistency, if the atomic number is set too large, it will cause too many 0 in the feature matrix and distance matrix.

Table 3.1 11 atomic features for encoding

Feature	Description
Mendeleev number (MN) [145]	Listing of chemical elements by column through the periodic system, which is effectively used to classify chemical systems.
Unpaired electrons	Electrons that occupies an orbital of an atom singly.
Ionization energies	The amount of energy required to remove an electron from an isolated atom or molecule.
Covalent radius	Half of the distance between two atoms bonded covalently.
Heat of formation	When one mole of a compound is produced from its basic elements, each substance is in its normal physical state, the quantity of heat received or released.
Dipole polarizability [135]	Describes the linear response of an electronic charge distribution with respect to an externally applied electric field.
Average ionic radius	Average of a monatomic ion's radius in an ionic crystal structure.
Group number in periodic table	The number of valence electrons of the elements in a certain group, a group is a vertical column of the periodic table.
Row number in periodic table	The number of rows of the element in the periodic table.
Pauling electronegativity	The power of an atom in a molecule to attract electrons to itself.
Atomic number	The charge number of an atomic nucleus.

3.2.2 Deep residual model for distance matrix prediction

For distance prediction, we train the deep residual neural network to grasp the intricate bonding interactions between atoms, this method takes advantage of the enormous atom interaction distributions that exist in the massive number of known crystal structures, therefore, it has the ability to predict pairwise distance with high accuracy.

Figure 3.2 depicts our deep neural network model, which is composed of three main parts. In the first part, a sequence of stacked 1-dimensional (1D) residual network layers to learn convoluted atom site features. In the second part, a 2-dimensional (2D) pairwise feature matrix is derived from the output of the 1D convolutional network by the outer product, and then we merge the convoluted sequential feature and pairwise feature as the input to the next module. The third part consists of a series of 2D residual network layers, which predicts the distances between two atoms, and finally gets the predicted distance matrix.

In our studies, the maximum number of atoms in a formula is set to L to accommodate the variable sizes of different crystal structure sites. In experiments, the L is set to 12 (the same as the feature matrix), when a formula has fewer atoms, the tensors are created by padding with zeros.

Residual Neural Network (ResNet) [60] is a kind of neural network that stacks residual blocks on top of each other to form a network by skip connections, this design had a profound influence on how to build deep neural networks. The skip connections between layers add the outputs from previous layers to the outputs of stacked layers which results in the ability to train much deeper networks than what was previously possible.

In the DeltaCrystal model, we design two residual network modules: one module aims to extract sequence features and the other module intends to derive pairwise features. For each residual network block, there are two convolutional layers, a batch normalization and two nonlinear transformations. We use 9 building blocks for each module in our main architecture. The number of filters is doubled per 3 blocks. The initial number of filters for the first and second modules are 32 and 256, respectively.

Since a distance map is a binary matrix, we use the cross-entropy loss as the loss function for neural network training. It is defined as:

$$LosscrossEntropy = \frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \tag{3.1}$$

Where N is the maximum length of the formula which is set to 12*5 and 12*10 in our experiments; y_i is the true distance matrix label at position i, and $\hat{y}_i$ is the

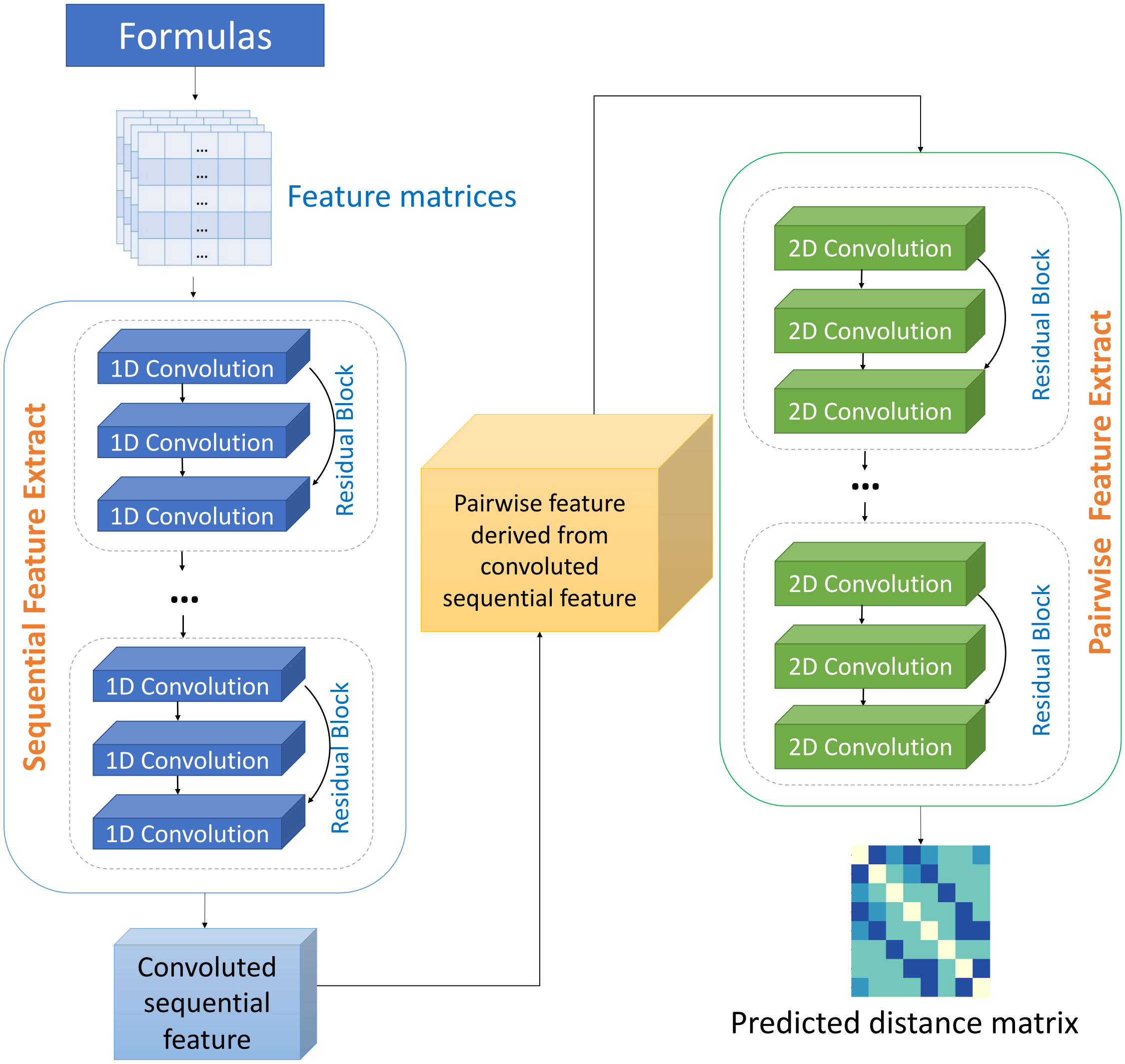


Figure 3.2 Deep neural network model for distance map prediction.

predicted label at position i.

3.2.3 3D crystal structure reconstruction algorithm

Our earlier proposed genetic algorithm DMCCrystal aimed to reconstruct atomic position based on the distance matrix. It has demonstrated that given the pairwise atomic distance matrix with space group, lattice parameters, and stoichiometry, this genetic algorithm can reconstruct the crystal structure which is close to the target crystal structures. For further improvement, these predicted structures can be used to seed the costly free-energy minimization-based ab initio CSP algorithms, as well as to acquire the correct crystal structure of certain components by DFT-based structural relaxation. Therefore, with all the predicted information, we then employ the DMCCrystal to predict the crystal structure.

3.2.4 Evaluation metrics

We use three metrics (MSE, RMSE, Overlap) to evaluate the performance of distance map based structure reconstruction. MSE and RMSE indicate the final structure similarity between the predicted structure and the true target structure. Overlap measures the accuracy of the predicted distance matrix with the true distance map. They are shown in the following equations:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

where n is the number of independent atoms in the target crystal structure, y_i and $\hat{y}_i$ are the corresponding atoms in the predicted crystal and the target crystal structure.

$$\text{Overlap} = \frac{|Target_{atom\ exist} \cap Predicted_{atom\ exist}|}{|Target\ atom|} \tag{3.4}$$

where $Target$ is the true distance map and $Prediction$ is the predicted distance matrix of a given composition, both only contain 1/0 entries. $Target\ atom$ means the total number of atoms in the target distance map. $Target\ atom_{exist} \cap Predicted\ atom_{exist}$ denotes the common exist atoms of $Target$ and $Prediction$. This function essentially measures the overlap of two matrix samples, with values ranging from 0 to 1 with 1 indicating perfect overlap. We also call this performance measure as distance map accuracy.

3.3 Experiments and Discussion

3.3.1 Dataset

We collect material ids, formulas and crystal structures (cif files) from Materialsproject.org by Pymatgen API. Since we set the maximum number of atoms in the formula to 12, after filtering, we get 18,800 compounds from the Materials Project and named this dataset as Mp_12. Furthermore, to compare our prediction results with the GNOA method [22], we removed 29 formulas that predicted structures with high performance from the training set, therefore, there are 18,776 compounds in Mp_12 datasets. Additionally, we extract materials whose crystal system is cubic to create Mp_12_cubic dataset; we also extract formulas with binary and ternary elements to form Mp_12_binary and Mp_12_ternary datasets, respectively. We train four distance matrix prediction models with each of these four datasets and evaluate their performance.

Table 3.2 Dataset

Dataset	Number of data	Describe
Mp_12	18,776	the original data collected from Materials Project is 18,800, we removed the 29 typical compounds in GNOA for prediction
Mp_12_cubic	3,298	extract the formula whose crystal system is cubic from Mp_12
Mp_12_binary	5,848	extract binary materials from Mp_12
Mp_12_ternary	11,615	extract ternary materials from Mp_12

Distance discretization We counted the atomic distances of 18,776 samples in Mp_12 dataset, the overall distribution is shown in Figure 3.3 where the smallest distance is 0.9488, the largest distance is 23.3361, and the majority of distances between 0 to 7 groups. In this work, we divide the continuous distance values with equal-width and store them in several groups by equation 3.5.

$$Interval\ width = (Maximum\_value - Minimum\_value)/N \tag{3.5}$$

where N represents the number of groups, we set it to 5, 10 or 20 in experiments. Furthermore, we use one-hot encoding to get the expression of each value. This method turns the regression problem into a classification problem by turning continuous distances into discrete values. Therefore, the cross-entropy loss can be used as the loss function for training neural networks, which makes the prediction more accurate.

Specifically, for each material, we encode 11 atomic features to 12*11 matrix (the maximum atomic number in unit cell is 12). After one-hot encoding, the dimension of distance matrix is 12*(12*N). In experiment, training samples are randomly selected with different sample sizes according to dataset size, such as 1000 or 10000 samples; after training, 100 or 1000 samples are selected for prediction and performance analysis.

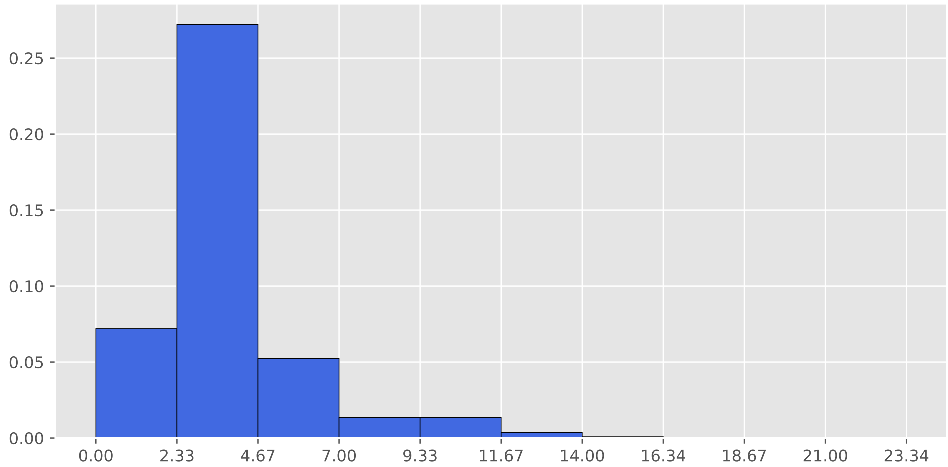


Figure 3.3 Overall atomic distance distribution of Mp_12 dataset. The x-axis represents the atomic distance and the y-axis represents the corresponding ratios. The most atomic distances are in the range of 2.33-4.67, accounting for more than 25% and there are just a few distances more than 14.

3.3.2 Prediction performance of distance matrix

To evaluate whether our DeltaCrystal models can learn the relationships between atom-pairs and whether they can predict the distance matrix with good performance, we did many experiments with different parameters, including the type of samples, the number of discrete groups, and the size of training and test datasets. We summarized the prediction performance in Table 3.3. In the training of these models, the number of epochs is set to 125, Adam optimizer is used to update model parameters and the learning rate to be 0.001.

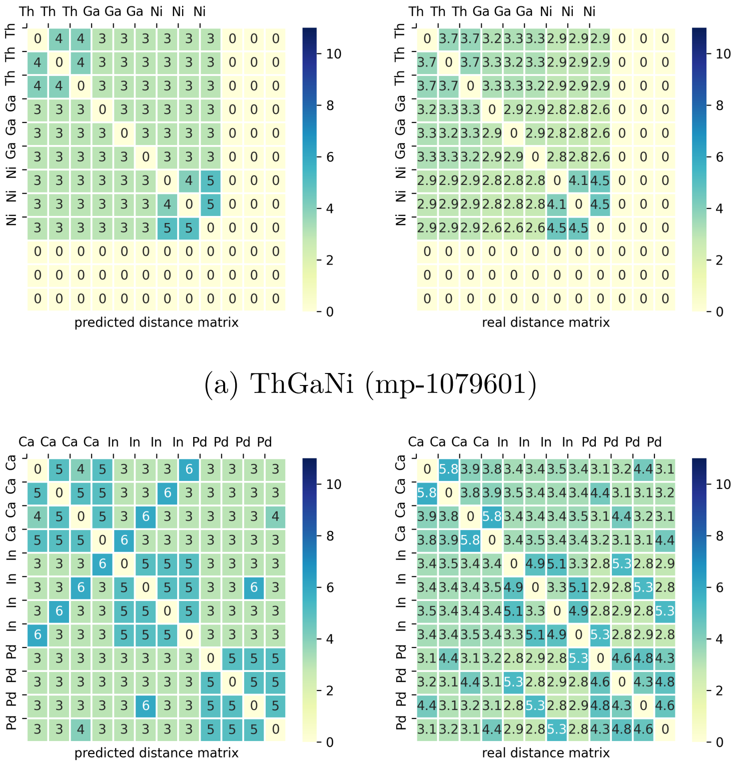
Table 3.3 Training performances of DeltaCrystal

Model	# Groups	# Training dataset	# Test dataset	MSE	RMSE	Overlap
DeltaCrystal	5	1000	100	4.9538	1.8841	0.9045
DeltaCrystal	5	10000	1000	2.3764	1.2269	0.9979
DeltaCrystal	10	1000	100	2.8586	1.3399	0.9981
DeltaCrystal	10	10000	1000	1.8932	1.0984	0.9989
DeltaCrystal	20	10000	1000	1.5591	0.9239	0.9992
DeltaCrystal_cubic	10	2638	200	1.0125	0.7394	0.9746
DeltaCrystal_binary	10	4678	400	2.7835	1.3109	0.9980
DeltaCrystal_ternary	10	9292	1000	2.0832	1.2779	0.9985

Table 3.3 summarizes the results of multiple comparison trials. Firstly, comparing different sizes of training datasets, there is a clear trend of decreasing MSE and RMSE with more training samples. For example, for the same DeltaCrystal model with a discretization group of 5, when the training samples are increased from 1000 to 10000, the RMSE is reduced by 53.57%. Secondly, we set different numbers of discrete groups for the basic DeltaCrystal model. Comparing discretization group 5, 10 and 20 with 10000 training samples, their RMSE are reduced by 11.69% and 18.88%, respectively. In order to explore whether only focusing on the same type of materials has a better effect, we specially trained three models: DeltaCrystal_cubic, DeltaCrystal_binary and DeltaCrystal_ternary. DeltaCrystal_cubic model is trained by materials whose

crystal system is cubic. DeltaCrystal_binary and DeltaCrystal_ternary are trained by binary and ternary materials.

Figure 3.4 shows two examples of the true distance matrix and the predicted distance matrix for ThGaNi and CaInPd, and their discretization group number is 10. Figure 3.5 illustrates the results for EuSi2 and ScTlAg2, their discretization group number is 20. These figures firstly demonstrate our methods can master the pair-wise relationship between atoms, secondly, as the number of discretized distance groups increases, the predicted distance labels are more accurate.



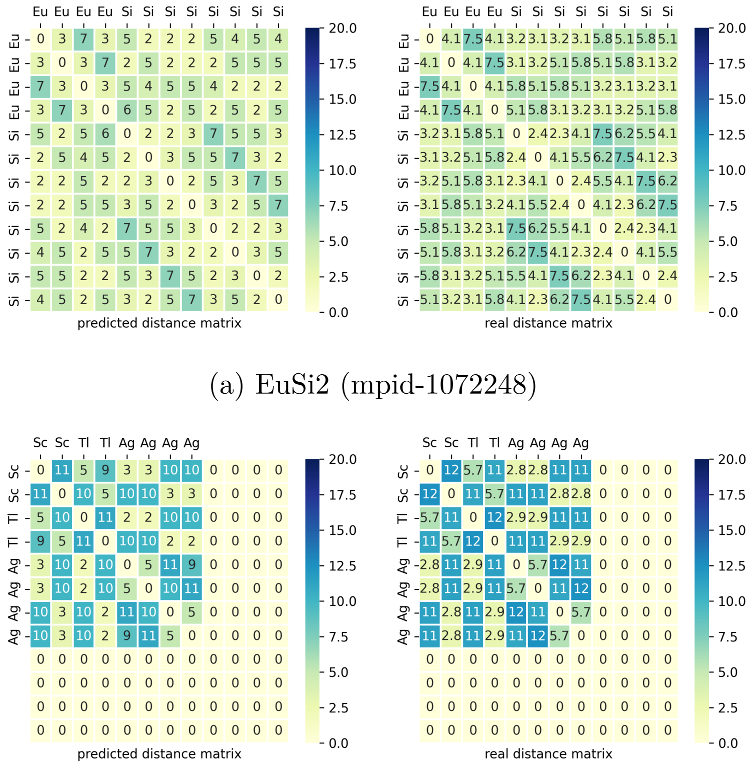
(a) ThGaNi (mp-1079601)

(b) CaInPd (mp-620571)

Figure 3.4 Compare the predicted and real distance matrix with distance group 10

3.3.3 Prediction of crystal structures

After predicting distance matrices, now we use the DMCrystal algorithm to predict their crystal structures and use M3GNET to relax these structures, as well as predict



(a) EuSi2 (mpid-1072248)

(b) ScTlAg2 (mp-1093619)

Figure 3.5 Compare the predicted and real distance matrix with distance group 20

their formation energies. Two samples of predicted crystal structures and their original (target) structures are shown in Figure 3.7.

3.3.4 Compare DeltaCrystal model with GNOA model

GNOA method [22] is based on a graph network and an optimization algorithm to predict crystal structure. Firstly, it used a graph network to construct a correlation model between the crystal structure and formation enthalpies, and then it utilized an optimization algorithm (OA) to search for the crystal structure that has the lowest formation enthalpy. Their experiments show 29 typical compounds with good predicted performance. To compare the predicted performance of GNOA and DeltaCrystal, we use these two models to predict 30 material structures separately, and the experimental results are summarized in Table 3.4 and Figure 3.6.

Table 3.4 shows RMSE performances of predicted distance matrices and predicted crystal structures by our DeltaCrystal model and RMSE of predicted crystal structures by the GNOA model. Although GNOA could achieve good performance for binary formulas, such as BePd and Ce3Pm, it is cannot handle complex formulas, especially for materials with the number of elements greater than or equal to 4, for example, GNOA cannot predict the structures of Rb2LiRhCl6, Y3Al3NiGe2 and LiFe2(ClO)2. While our method is able to predict the structure of complex materials and we can achieve lower RMSEs in many cases.

In addition to the aforementioned factors, formation energy is also an important criterion for assessing the stability of crystal structures. Hence, we have conducted a comparison of the lowest formation energies among the existing crystal structures, our predicted structures and GNOA predicted structures by M3GNET in Figure 3.6. In the figure, the red line represents the ground truth, the green line corresponds to our method, the yellow line indicates our relaxed results, and the blue line represents the GN-OA method. Notably, our method demonstrates a significantly closer alignment with the ground truth in terms of formation energies, outperforming the GN-OA predictions. This observation reinforces the efficacy and reliability of our approach in predicting crystal structures.

In figure 3.7, we present three examples showcasing the comparison between ground truth structures, GNOA predicted structures, and our predicted crystal structures. This illustration highlights the capability of our method to accurately predict crystal structures, particularly for complex chemical formulas. Our method demonstrates a remarkable ability to capture the intricate arrangements of atoms in these structures, yielding predictions that closely resemble the ground truth. This finding underscores the effectiveness of our approach in handling challenging crystal structure predictions.

Table 3.4 Compare the predicted performance (RMSE)of DeltaCrystal with GNOA

Type	Formula	Mp_id	Distance Matrix RMSE	GNOA RMSE	DeltaCrystal Top-10 RMSE	DeltaCrystal Top-20 RMSE
Binary	Tiln	mp-1216825	0.0011	0.3993	0.5000	0.4564
	ScAl	mp-331	0.0088	0.4194	**0**	0
	Tm3P	mp-971958	0.0525	0.4274	0.5590	0.5590
	BePd	mp-11274	0.0645	0	0.5000	0
	PaIn3	mp-861987	0.0928	0.500	0.5000	0.5000
	Ce3Pm	mp-1183767	0.1463	0	0	0
	GdCo5	mp-1077071	0.3185	0.3093	0.4044	0.3720
	LiC6	mp-1001581	0.3738	0.3762	**0.2795**	0.2795
	BaAu5	mp-30364	0.4725	0.3333	**0.2954**	0.2954
	Zr4Al3	mp-12752	0.4818	0.2473	0.4340	0.3293
Ternary	TmIn2Sn	mp-1216827	0.0831	0.3536	**0.3372**	**0.2041**
	CeAlO2	mp-1226604	0.1132	0.4508	**0.4077**	**0.3819**
	ThBeO3	mp-1187421	0.1743	0.3651	0.5627	0.5627
	Zn(CuN)2	mvc-15351	0.1830	0.4485	**0.2452**	0.2452
	LiVS2	mp-7543	0.2096	0.2734	**0.1922**	0.1922
	LiTiTe2	mp-10189	0.2454	0.2839	0.3612	0.3612
	Sr(AlGe)2	mp-1070483	0.2596	0.5112	**0.2435**	0.2435
	NdTiGe	mp-22331	0.2675	0.4748	**0.4540**	**0.4182**
	Mn4AsP3	mp-1221760	0.2916	0.4084	**0.2755**	0.2755
	Mn3FeP4	mp-1221749	0.3492	0.4543	**0.3943**	**0.3066**
≥ 4 elements	DyThCN	mp-1225528	0.1721	0.3933	0.4564	**0.2887**
	EuNbNO2	mp-1225127	0.2613	0.3535	**0.2887**	0.2887
	LaNiPO	mp-1079685	0.3550	0.3448	0.4086	0.4086
	SrFeMoO5	mp-690817	0.3802	0.4081	**0.3682**	**0.3154**
	FeCu2SnS4	mp-628568	0.3993	0.4364	**0.3446**	**0.3045**
	LiTb(CuP)2	mp-8220	0.4115	0.4270	**0.1782**	**0.1716**
	Rb2LiRhCl6	mp-1206187	0.5137	None	**0.2979**	0.2979
	Y3Al3NiGe2	mp-10209	0.5269	None	**0.4126**	0.4126
	LiFe2(ClO)2	mp-755254	0.5649	None	**0.3740**	0.3740
	Mg2VWO6	mp-1303315	0.5872	0.4347	**0.2951**	**0.2871**

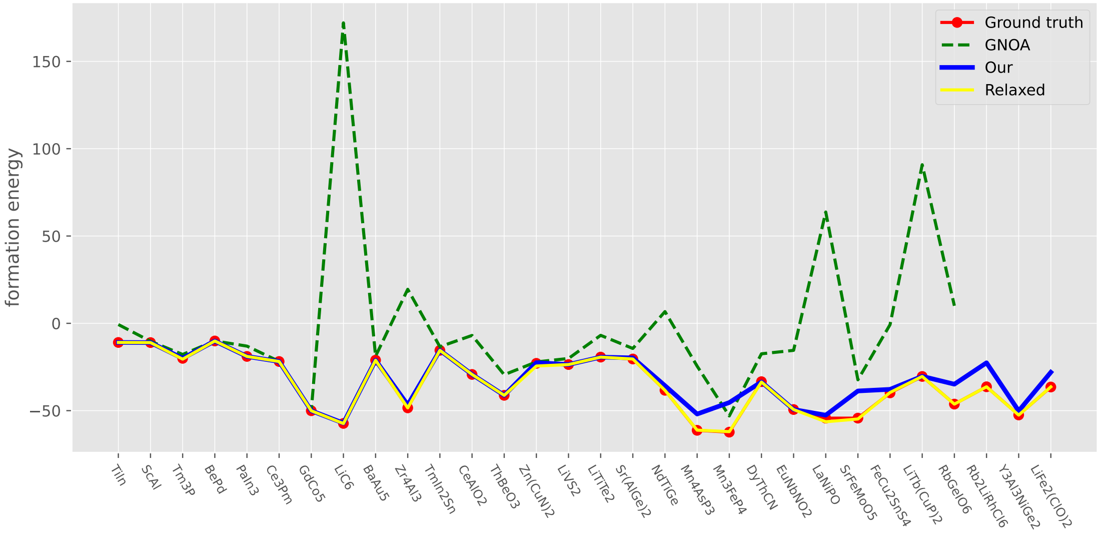
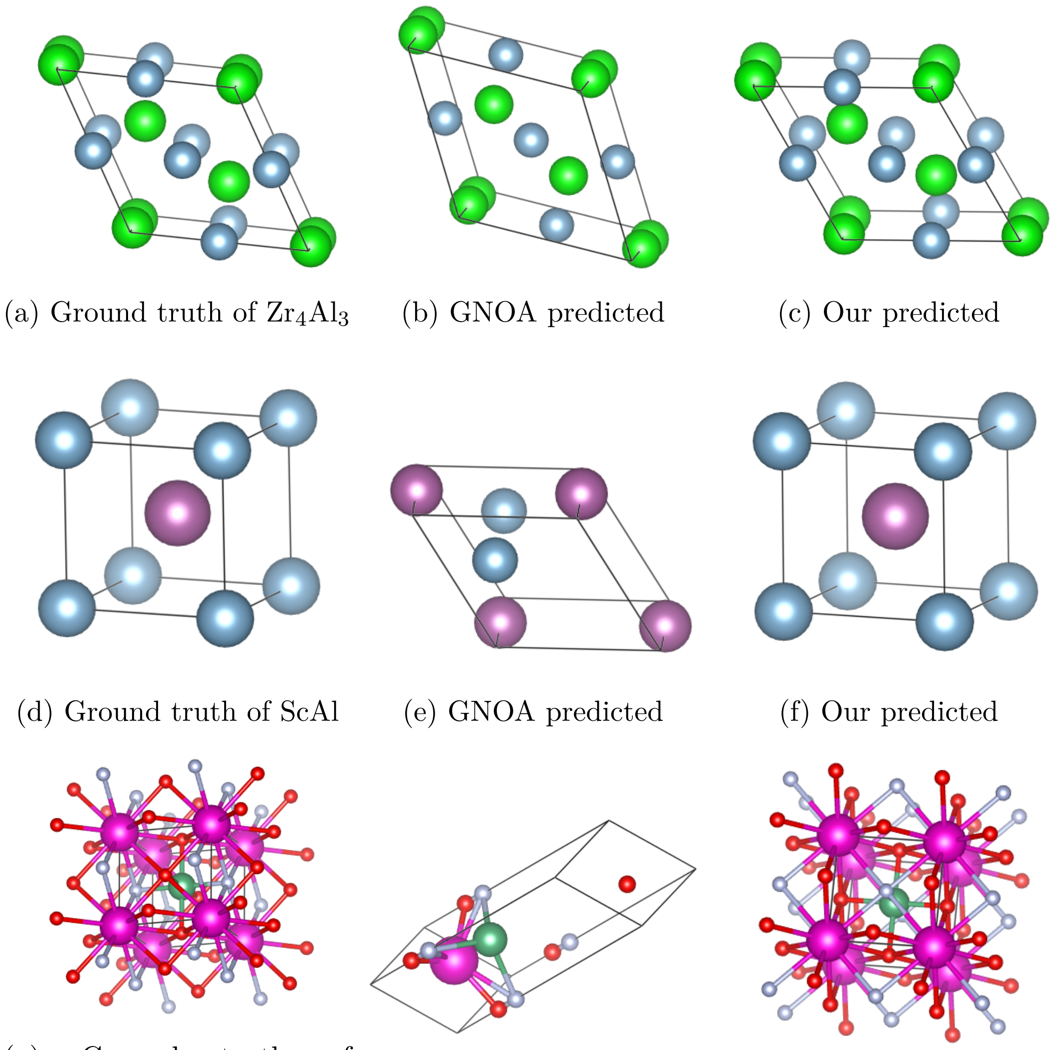


Figure 3.6 Comparison of formation energies

3.4 Conclusion

We propose DeltaCrystal, a deep residual neural network approach for crystal structure prediction by first predicting the distance matrix of atom pairs for given material composition, and then using it to predict its crystal structure using a genetic algorithm. Compared to the minimization of free energy during atomic configuration search in conventional ab initio CSP methods, our method takes advantage of the existing physical or geometric constraints (such as the symmetry of atom positions) of the existing crystal structures in the materials repositories. Our experiments show our DeltaCrystal algorithm is able to reconstruct the crystal structures for a large number of materials. Our predicted structures are so close to the ground truth crystal structures so that they can be used to seed the costly free energy minimization-based CSP algorithms for further structure refining. Our DeltaCrystal can be a strong new kind of machine learning or deep knowledge-guided CSP for large-scale prediction of crystal structures.



(a) Ground truth of Zr_4Al_3

(b) GNOA predicted

(c) Our predicted

(d) Ground truth of ScAl

(e) GNOA predicted

(f) Our predicted

(g) Ground truth of $EuNbNO_2$

(h) GNOA predicted

(i) Our predicted

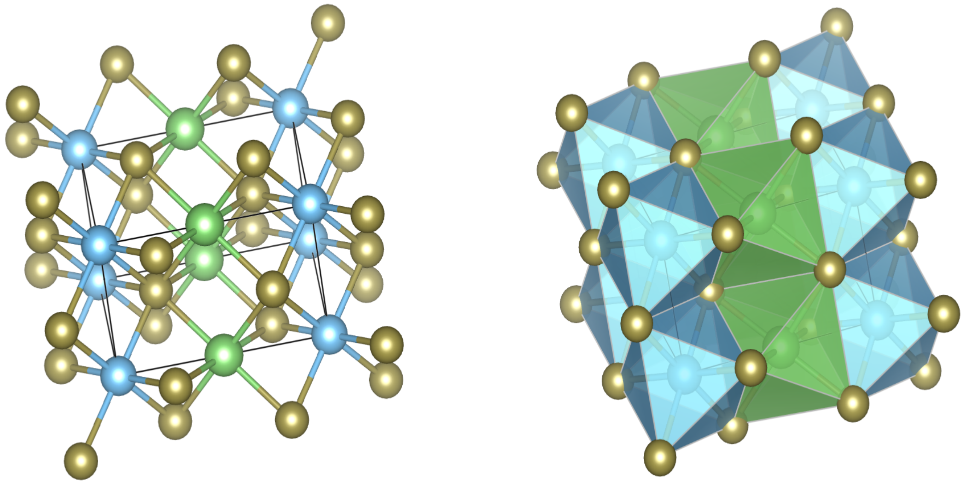
Figure 3.7 Examples of predicted crystal structures by DeltaCrystal

Chapter 4

Motif-driven crystal structure analysis and prediction

4.1 Introduction

Crystal structure prediction [97, 157, 95], generative design of crystal materials [165, 138], material structure searching [120, 71] are critical and challenging problems in exploring material crystal structure and they have been active areas of research for many decades. The main objective of these studies is to investigate the spatial arrangement of atoms and their interatomic relationships. Additionally, they can be used to design new structures or predict the functional properties of a material based on its structure. As crystal structure of a material has a significant impact on its properties and behavior, such as its mechanical strength, electrical conductivity, and optical properties. By exploring the crystal structure of a material, researchers can gain a deeper understanding of its properties and potential applications, as well as develop new materials with specific properties by designing their crystal structures. This exploration may involve theoretical calculations, experimental techniques, and computational methods. Overall, exploring material crystal structure is an important aspect of materials science that can lead to the discovery of new materials and technologies.



(a) Representation by atoms and bonds (b) Representation by polyhedral

Figure 4.1 Crystal structure of LiTiTe2.

Exploring material crystal structure involves studying the arrangement of atoms in a solid material at the atomic level. For example, in the CSP problem, most of these studies focus on predicting the positional information of each atom in the crystal structure, this type of prediction is commonly referred to as atomic-level crystal structure prediction. However, there are several drawbacks and limitations to these atomic-level studies. One major limitation is that predicting the crystal structure of a material is a complex problem that is computationally expensive and time-consuming, even with modern computing resources. It may require the use of advanced algorithms and techniques, and may still involve a significant degree of uncertainty and trial-and-error. Additionally, even if the crystal structure of a material can be accurately predicted or experimentally determined, understanding the behavior of the material under different conditions (such as high pressure or temperature) can be challenging, as these conditions can lead to changes in the crystal structure and properties. Another limitation is that atomic-level exploration of crystal structures may not fully capture the behavior of materials at larger scales. For example, defects or impurities in a crystal structure can have a significant impact on its properties, but may not be accurately captured by atomic-level simulations alone. Overall, while exploring material crystal structures at the atomic level can provide valuable insights and enable the discovery of new materials and properties, it is important to be aware of these limitations and to consider a range of approaches and scales in materials research.

In addition to directly predicting the position of each atom, there are some studies focusing on higher-level atomic relationships: motifs. A motif consists of a set of atoms arranged in a particular way within a unit cell, it also refers to a repeating unit or pattern of atoms or molecules in a crystal structure [69, 10]. These motifs can be considered the building blocks of a crystal structure, as they are repeated throughout the material to create a more giant crystal lattice. As shown in Figure

4.2, the left figure (a) shows the crystal structure of LiTiTe2 from the atoms and bonds relationships, it totally contains 44 atoms and 72 bonds, it is very difficult to predict the exact coordinates of each atom. Inspired by the polyhedral representation of crystal structure, as shown in Figure 4.2 (b), LiTiTe2 has 12 polyhedrals, we aim to study and extract the recurring patterns within a crystal structure, and contribute to the crystal structure prediction task from a higher level: we only need to predict positions of several motifs rather than a lot of atoms.

Motifs, these repeating patterns, have a significant impact on the crystal structure and the material properties [10]. For instance, the arrangement of atoms within a motif can affect the strength, electrical conductivity, and other properties of the material. By understanding the motifs present in a material, scholars can gain valuable insights into its structure and properties, and use this knowledge to develop new materials with desired properties. With this objective in mind, our work focuses on extracting motifs from existing crystal structures, analyzing their patterns, and utilizing these motifs to predict crystal structures.

By studying the motifs within crystal structures, we uncover recurring patterns that hold crucial information about the material's behavior. These motifs serve as building blocks that contribute to the overall structure and properties of the material. By comprehensively analyzing and understanding these motifs, we can make informed predictions about the crystal structures of new materials. Through this research, we advance our understanding of the relationship between motifs and crystal structures, ultimately paving the way for the design and synthesis of novel materials with tailored properties. By harnessing the power of motifs, we can accelerate the discovery and development of innovative materials for various applications.

Overall, our contributions can be summarized as follows:

- We highlight the significance of motifs in predicting crystal structures and underscores their potential for advancing materials informatics.

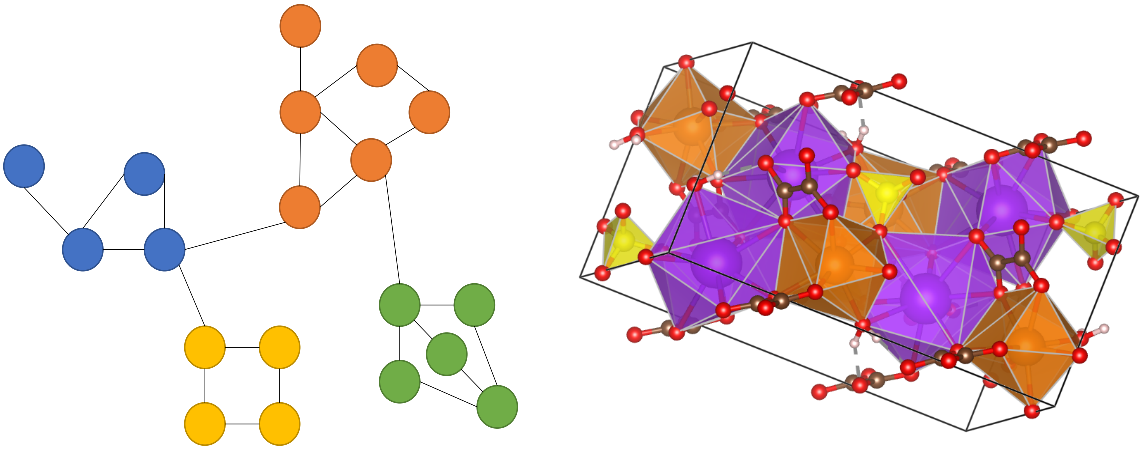
- We extract motifs from a diverse range of crystal structures and conduct in-depth analyses to uncover common patterns and motifs that span across different materials.

- We introduce a preliminary framework for the exploration of material structures based on motif knowledge.

4.2 Methods and Experiments

4.2.1 Motif detection method: the QG-CD model

Our method utilizes advanced algorithms and techniques to effectively identify recurring atomic patterns and motifs within the crystal structures. Through a systematic analysis of atom arrangements and their interatomic distances, we are able to extract meaningful and significant motifs within the material crystal structure and contribute to the understanding and exploration of material structures.



(a) Community with groups of node

(b) Material strictures with motifs

Figure 4.2 Community detection vs motif extraction

In order to detect and extract structural motifs in crystal structures, we design the Quotient Graph-Community Detection (QG-CD) model. Quotient graphs [146, 55] are

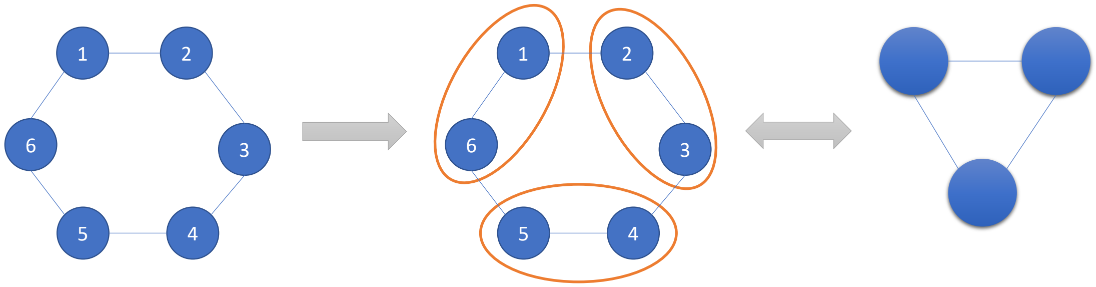


Figure 4.3 An example of a quotient graph

employed to transform a graph consisting of numerous clusters of nodes into a graph with several distinct "blocks" of nodes. By using community detection algorithms [42], such as the Girvan-Newman algorithm [30], it has been applied in social network analysis and biological network analysis. In our work, we aim to identify communities within complex systems by determining the edges that connect local clusters of nodes. Combining these two algorithms together, we can detect structural motifs with a smaller degree of errors.

The Girvan-Newman algorithm for the detection and analysis of community structure relies on the iterative elimination of edges that have the highest number of shortest paths between nodes passing through them. By removing edges from the graph one-by-one, the network breaks down into smaller pieces, so-called communities. The idea was to find which edges in a network occur most frequently between other pairs of nodes by finding edges betweenness centrality. The edges joining communities are then expected to have a high edge betweenness. The underlying community structure of the network will be much more fine-grained once the edges with the highest betweenness are eliminated which means that communities will be much easier to spot. In material structures, we can use this method to find motifs.

The Girvan-Newman algorithm can be divided into four main steps: (1)For every edge in a graph, calculate the edge betweenness centrality. (2) Remove the edge with the highest betweenness centrality. (3)Calculate the betweenness centrality for every remaining edge. (4) Repeat steps 2-4 until there are no more edges left.

As shown in Figure 4.4, in this example, it show how a typical graph looks like when edges are assigned weights based on the number of shortest paths passing through them. To keep things simple, we only calculated the number of undirected shortest paths that pass through an edge. The edge between nodes A and B has a strength of 1 because we don't count $A->B$ and $B->A$ as two different paths. It would remove the edge between nodes C and D because it is the one with the highest strength. This means that the edge is located between communities. After removing an edge, the betweenness centrality has to be recalculated for every remaining edge. In this example, we have come to the point where every edge has the same betweenness centrality. While the betweenness centrality measures the extent to which a vertex or edge lies on paths between vertices. Vertices and edges with high betweenness may have considerable influence within a network by virtue of their control over information passing between others.

The motif detection algorithm has been developed using libraries such as Networkx [54] and Pymatgen [116], these libraries provide essential functionalities for graph analysis, including the Girvan-Newman library, quotient graph library, and crystal structure library. Networkx is a Python library specifically designed for the creation, manipulation, and analysis of complex networks or graphs. It provides a comprehensive set of tools and functions for working with network data, including creating, modifying, and visualizing networks, as well as performing various network analysis tasks. Pymatgen (Python Materials Genomics) is a Python library specifically developed for the analysis and modeling of materials and molecules. It provides a wide range of tools and functionalities to work with various aspects of materials science, including crystal structures, electronic structures, and thermodynamics.

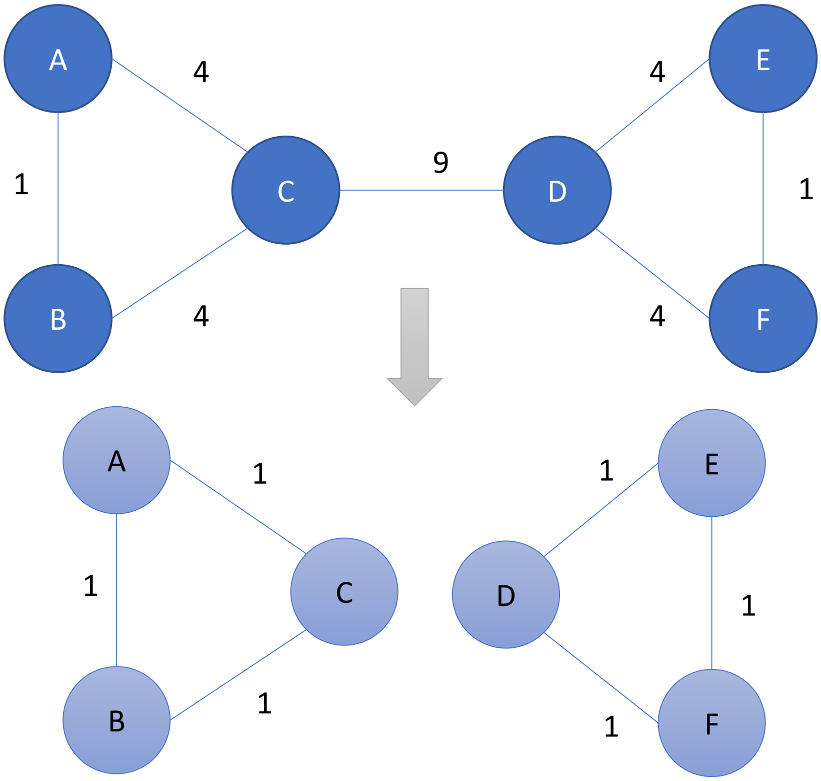


Figure 4.4 An example of Girvan-Newman algorithm

4.2.2 Motif extraction results

By extracting motifs from such a vast collection of material structures, we uncover common patterns and motifs that span across different materials. For instance, we show three material polyhedron structures and their motif structures in Figure 4.5, Figure 4.6 and Figure 4.7.

As illustrated in Figure 4.5, the left panel showcases the polyhedron structure of the LiTiTe2 material. On the right panel, we present two extracted motif structures: LiTe6 (represented by green-colored building blocks) and TiTe6 (represented by blue-colored building blocks).

In the Zn(CuN)2 material (in Figure 4.6), we have identified and extracted two distinct motifs: the tetrahedron motif ZnN4 and the linear motif CuN2. The ZnN4 motif represents the arrangement of four nitrogen (N) atoms surrounding a central

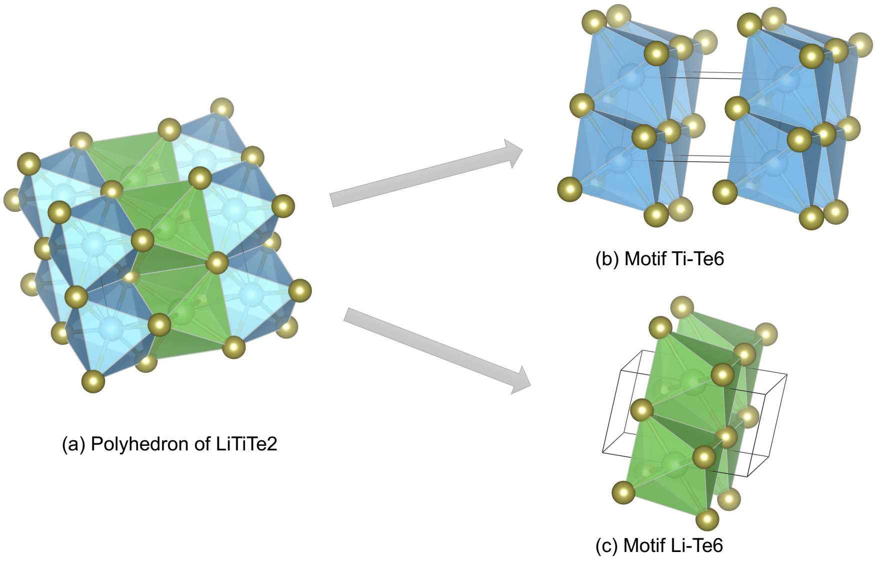


Figure 4.5 Extracted motifs of LiTiTe2

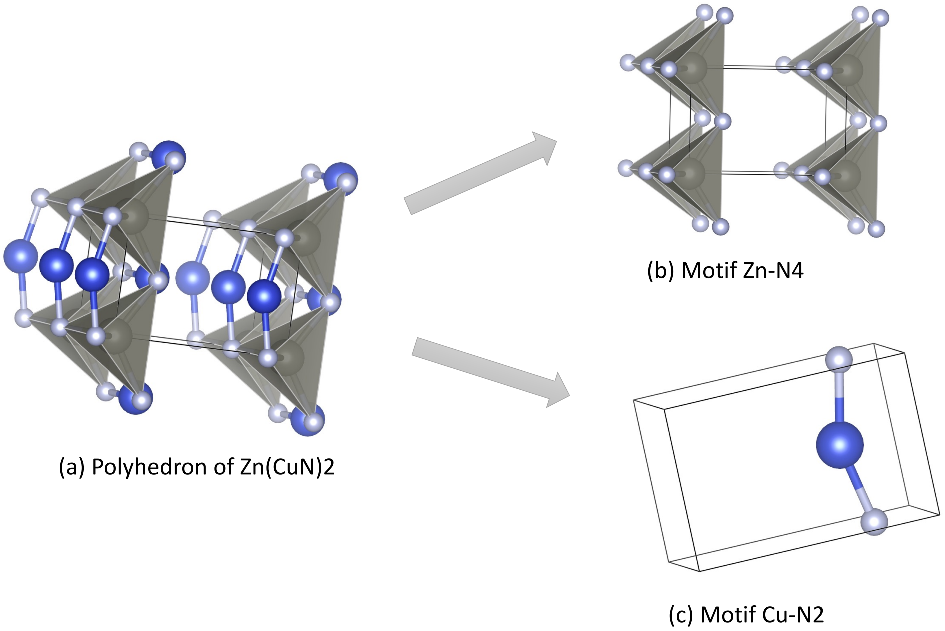


Figure 4.6 Motif of Zn(CuN)2

zinc (Zn) atom, forming a tetrahedral structure. On the other hand, the CuN2 motif consists of two nitrogen (N) atoms bonded to a central copper (Cu) atom in a linear configuration.

In the case of the KGdH2C2SO9 material (in Figure 4.7), we have identified and

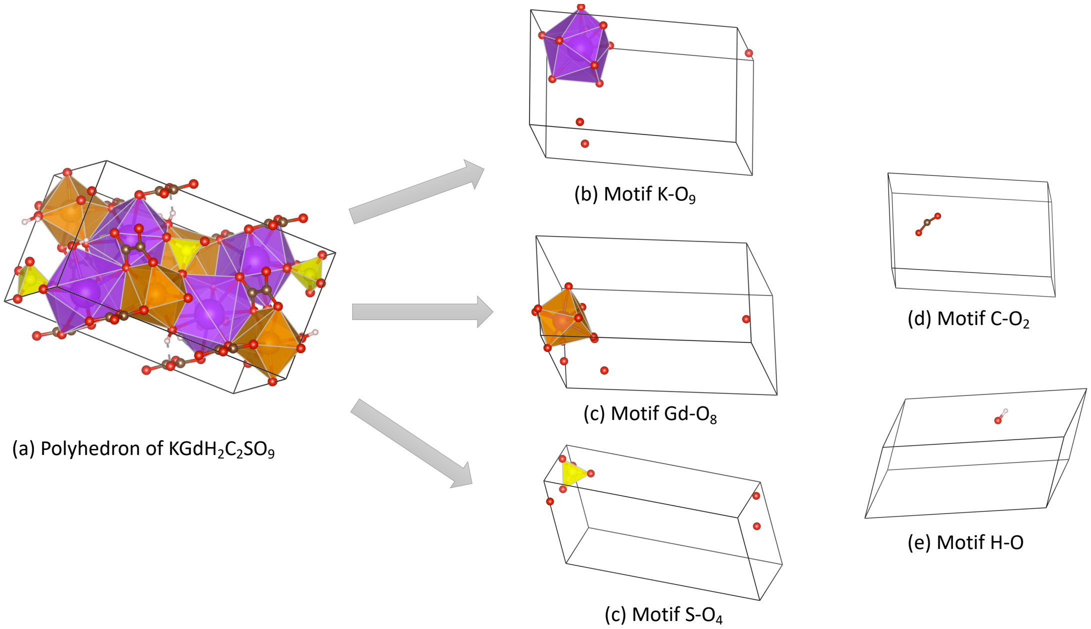
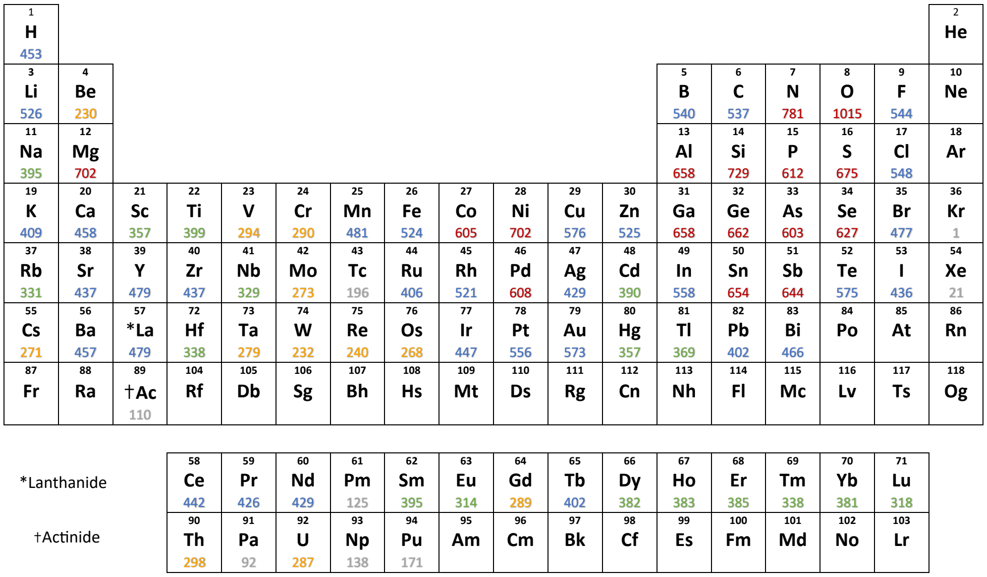


Figure 4.7 Motif of KGdH2C2SO9

extracted three polyhedral motifs: KO9 (depicted in purple), GdO8 (represented by orange), and SO4 (highlighted in yellow). Additionally, we have discovered two linear motifs: CO2 and HO. These motifs provide valuable insights into the atomic arrangements and bonding patterns within the KGdH2C2SO9 material, enabling a better understanding of its structural characteristics.

4.2.3 Motif analysis

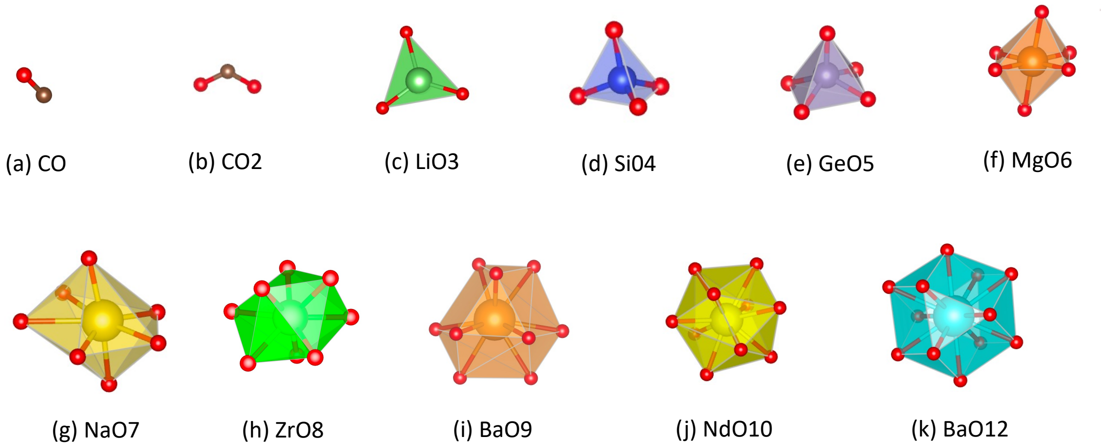
We have assembled a comprehensive dataset comprising 122,500 material structural cif files sourced from the Material Project database. Utilizing this extensive collection, we employ our method to extract motifs for each individual structure. As a result, we obtain a total of 18,534 motifs across 86 different types of elements. According to the periodic table, those elements and their corresponding motif numbers are summarized in Figure 4.8, the numbers with red color indicate their occurrence of more than 600, while the blue color indicates a range of 400 to 599, the green color represents a range of 300 to 399, the yellow color signifies 200 to 299 and the gray color indicates an occurrence of less than 200.



1 H 453																	2 He
3 Li 526	4 Be 230											5 B 540	6 C 537	7 N 781	8 O 1015	9 F 544	10 Ne
11 Na 395	12 Mg 702											13 Al 658	14 Si 729	15 P 612	16 S 675	17 Cl 548	18 Ar
19 K 409	20 Ca 458	21 Sc 357	22 Ti 399	23 V 294	24 Cr 290	25 Mn 481	26 Fe 524	27 Co 605	28 Ni 702	29 Cu 576	30 Zn 525	31 Ga 658	32 Ge 662	33 As 603	34 Se 627	35 Br 477	36 Kr 1
37 Rb 331	38 Sr 437	39 Y 479	40 Zr 437	41 Nb 329	42 Mo 273	43 Tc 196	44 Ru 406	45 Rh 521	46 Pd 608	47 Ag 429	48 Cd 390	49 In 558	50 Sn 654	51 Sb 644	52 Te 575	53 I 436	54 Xe 21
55 Cs 271	56 Ba 457	57 *La 479	72 Hf 338	73 Ta 279	74 W 232	75 Re 240	76 Os 268	77 Ir 447	78 Pt 556	79 Au 573	80 Hg 357	81 Tl 369	82 Pb 402	83 Bi 466	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 †Ac 110	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
*Lanthanide			58 Ce 442	59 Pr 426	60 Nd 429	61 Pm 125	62 Sm 395	63 Eu 314	64 Gd 289	65 Tb 402	66 Dy 382	67 Ho 383	68 Er 385	69 Tm 338	70 Yb 381	71 Lu 318	
†Actinide			90 Th 298	91 Pa 92	92 U 287	93 Np 138	94 Pu 171	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

Figure 4.8 Extracted motifs statistics. In this periodic table, the number below each element indicates how many motifs the element appears in.

From Figure 4.8, it is evident that the most frequently occurring element is oxygen, which appears in 998 motif structures. Additionally, in Figure 4.9, we provide 11 motifs for reference. Among these 86 elements, the top 10 elements that appear most frequently in motif structures are O (998), N (772), Si (727), Mg (702), Ni (702), S (673), Ge (662), Al (657) and Ga (658); while the 10 elements with the least number of occurrences are W (232), Be (230), Tc (196), Pu (171), Np (138), Pm (125), Ac (110), Pa (92), Xe (20), Kr (1).



(a) CO (b) CO2 (c) LiO3 (d) Si04 (e) GeO5 (f) MgO6

(g) NaO7 (h) ZrO8 (i) BaO9 (j) NdO10 (k) BaO12

Figure 4.9 Oxygen motifs

It is important to explore Lithium-ion materials as they play important roles [94]. For example, Lithium-ion batteries are widely used for energy storage in various applications, including electric vehicles, portable electronics, and renewable energy systems. Lithium-ion materials offer potential advancements in energy storage technologies, allowing for increased energy capacity, faster charging, and improved stability. Therefore, we focus on analyzing motifs with lithium as the central atom.

We conduct an analysis of motifs containing lithium elements, resulting in the extraction of 526 motifs. From this set, we filtered out structures where lithium atoms serve as the central atom. Based on the polyhedron type, we identified and categorized representative motifs, which are summarized in Figure 4.10.

Our analysis reveals six distinct types of motifs: linear motifs, flat motifs, tetrahedron motifs, pentahedron motifs, octahedron motifs, and polyhedron motifs. Linear motifs typically involve a lithium atom and 1 or 2 other atoms. For instance, the motif "Li-O" is formed by a lithium atom and an oxygen atom, while the motif "Li-N-O" consists of a lithium atom, a nitrogen atom, and an oxygen atom. Flat motifs encompass a central lithium atom surrounded by three other atoms that lie on the same plane. An example of such a motif is the "Li-N2-O", which comprises a lithium atom, two nitrogen atoms, and an oxygen atom. These motifs exhibit a distinct planar arrangement of atoms around the central lithium atom.

Tetrahedron motifs exhibit a central lithium atom surrounded by four other atoms, forming a tetrahedral arrangement. Examples of such motifs include "Li-F4," "Li-O3-F," and "Li-O2-F," where the central lithium atom is connected to four fluorine atoms or a combination of oxygen and fluorine atoms. Similarly, pentahedron motifs involve a central lithium atom surrounded by five other atoms, forming a pentahedral structure. Examples of pentahedron motifs include "Li-Cl5," "Li-O4-F" and "Li-O2-F3", where the central lithium atom is bonded to five chlorine atoms or a combination of oxygen and fluorine atoms. In addition, octahedron motifs encompass a central lithium atom surrounded by six other atoms, forming an octahedral structure, such as "Li-O6", "Li-O5-F" and "Li-O3-F3". These motifs provide insights into the spatial arrangement and coordination of lithium atoms in complex crystal structures.

Motifs that have more than seven surrounding atoms are classified as polyhedral structures. These motifs highlight the coordination of the central lithium atom with multiple atoms, forming intricate polyhedral arrangements. For instance, the motif "Li-O7" comprises a central lithium atom coordinated with seven oxygen atoms, while "Li-S8" indicates the coordination of the lithium atom with eight sulfur atoms. Similarly, "Li-O9" represents the coordination of the central lithium atom with nine oxygen atoms.

4.3 Motif-based crystal structure prediction: the framework

Traditional methods for crystal structure prediction primarily focus on predicting the positions of each atom individually, leading to computational intensity and time consumption issues. To address this challenge and enhance the efficiency of crystal structure prediction while facilitating the discovery of new materials, drawing inspiration from a motif-based structure searching method XMsearch [171], we aim to

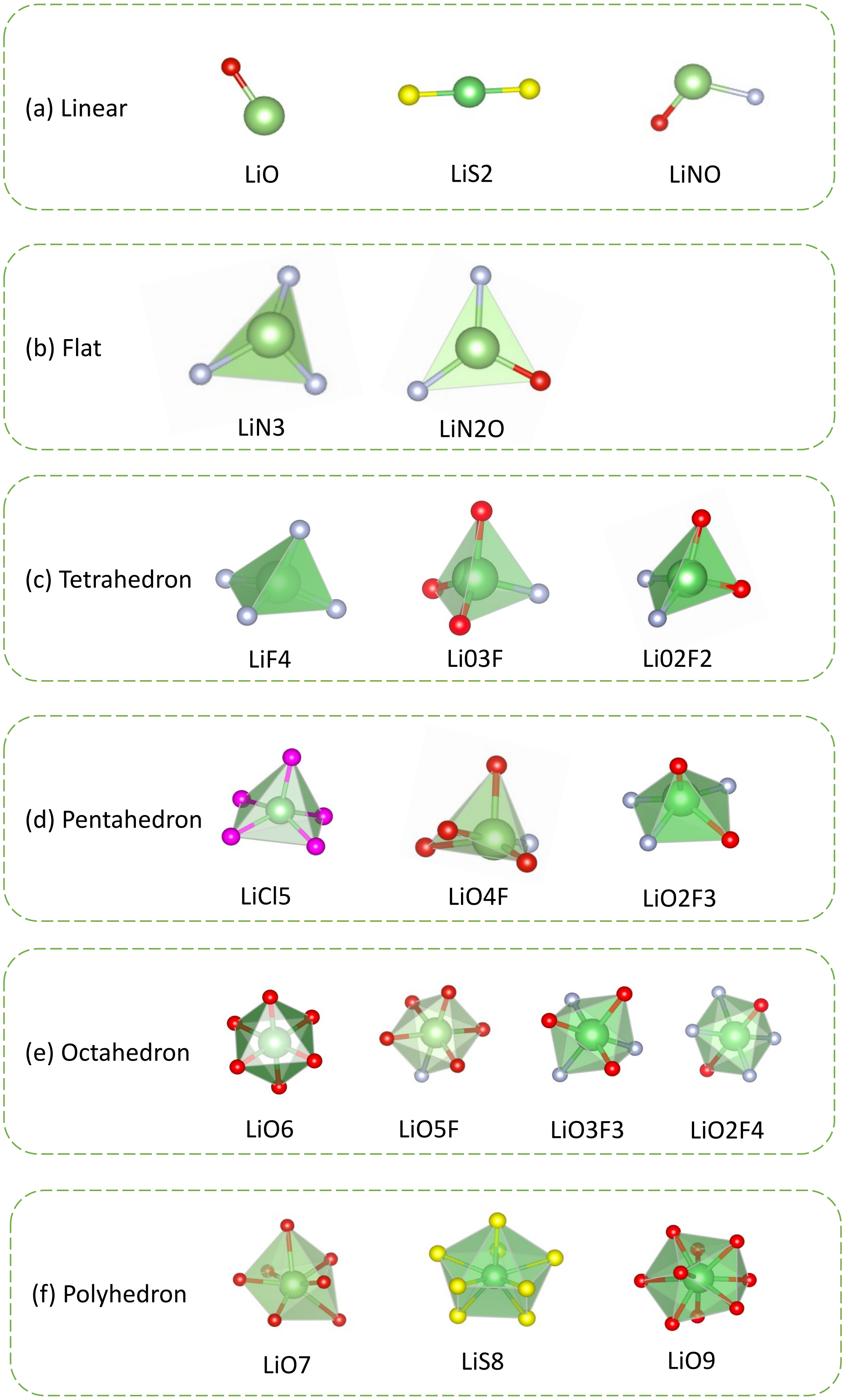


Figure 4.10 Motifs with lithium as the central atom

reduce the total degrees of freedom (DOF) by employing motifs to describe crystal structures. Therefore, we devise a preliminary framework based on structural motifs for the crystal structure prediction of materials.

The workflow of our preliminary motif-based crystal structure prediction is illustrated in Figure 4.11. This comprehensive approach consists of three primary steps, each contributing crucially to the successful prediction of crystal structures for materials.

In the first step, we extract motifs from existing material structures. In this initial phase, we gather a diverse set of existing material structures, which serve as the foundation for our prediction framework. And then employing our proposed QG-CD model, we meticulously extract key motifs present in these structures. This extraction process is fundamental to our motif-based framework, as it captures the recurring patterns and building blocks that underlie the crystal structures.

In the second step, we aim to generate and filtrate motif-based structures. Building on the knowledge gleaned from the extracted motifs, we now embark on the exciting task of generating novel material structures. By combining and assembling these motifs in various configurations, we can envision an extensive library of potential crystal structures. To ensure the feasibility and relevance of these candidates, a rigorous filtration process is employed. Only the most promising and physically plausible structures proceed to the next stage.

In the third step, we aim to relax generated structures and find stable structures. Having obtained a selection of potential crystal structures, our focus shifts to refining and validating their stability. Through a relaxation process carried out by the sophisticated M3GNet algorithm, the predicted structures undergo optimization to attain their most energetically favorable configurations. This step is critical in identifying the stable structures among the candidates. Subsequently, we employ Density Functional Theory (DFT) calculations to compute the formation energies

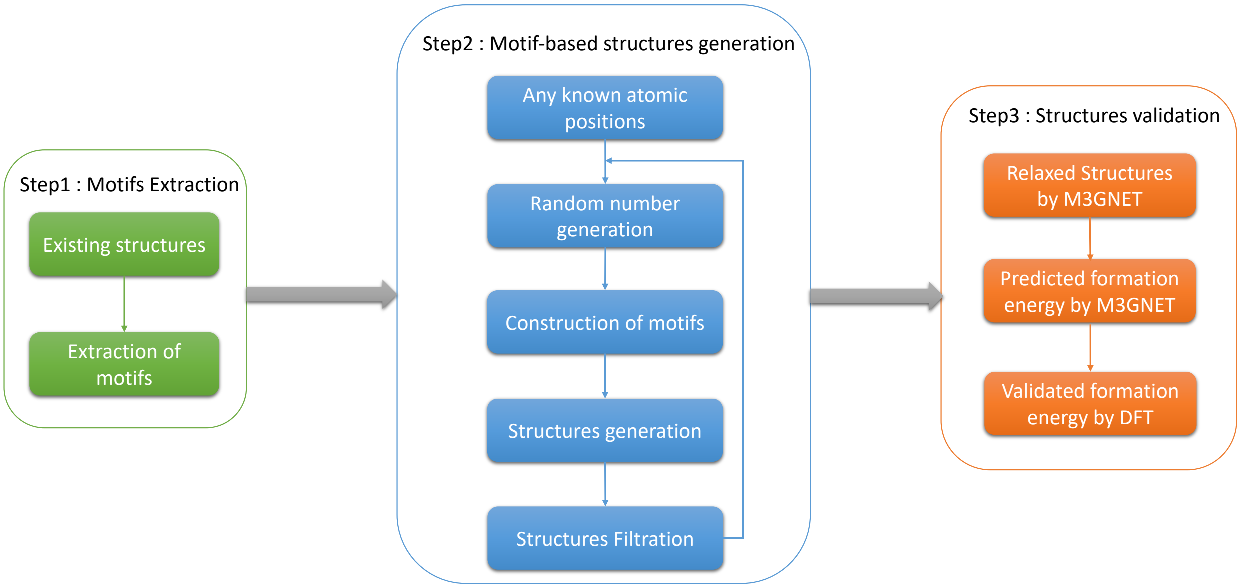

Step1 : Motifs Extraction
Existing structures
Extraction of motifs
Step2 : Motif-based structures generation
Any known atomic positions
Random number generation
Construction of motifs
Structures generation
Structures Filtration
Step3 : Structures validation
Relaxed Structures by M3GNET
Predicted formation energy by M3GNET
Validated formation energy by DFT

Figure 4.11 Framework of the motif-based structure prediction method

of these relaxed structures. The formation energy is a key indicator of a material's stability, helping us discern the most viable candidates for experimental synthesis and characterization.

By seamlessly integrating these three key steps, our motif-based framework offers a significant advancement in crystal structure prediction. Its capacity to efficiently explore the vast landscape of crystal structures holds immense promise in accelerating the discovery of novel materials, unlocking a treasure trove of materials with remarkable properties.

4.4 Conclusion

We explore and analyze material structures from a higher-level perspective: polyhedral motif structures. This approach identifies structural motifs based on exploring atom patterns that appear frequently across different crystal materials with high geometric conservation. Through our comprehensive extraction experiments, we extract 18,534 motifs and uncover common patterns and motifs that span across different materials.

For the crystal structure prediction task, unlike traditional methods that require the calculation of coordinates for each individual atom, the utilization of motifs offers a notable advantage by significantly reducing computational costs. In this work, we present a preliminary framework for the exploration of material structures from the motif point of view. By leveraging motif-based knowledge, we anticipate that it will serve as a guiding principle in the design and synthesis of novel materials in the near future, opening new avenues for advancements in material informatics.

Chapter 5

Computational Discovery of New 2D Materials Using Deep Learning Generative Models

5.1 Introduction

Two-dimensional materials such as graphene and hexagonal boron nitride have the potential to create new electronics and technologies such as spintronics, catalysis, and membranes owing to their exotic vibrational, electronic, optical[148], magnetic, and topological behaviors[40, 6, 5, 163]. Using density functional theory (DFT) based screening, Mounet et al. [105] have found 1,825 compounds with requisite geometric and bonding criteria that should make them relatively easy to exfoliate and so produce novel 2D materials with potentially interesting physical and electromagnetic properties. They discovered 56 ferromagnetic and antiferromagnetic systems, including half-metals and half-semiconductors. This greatly expands the list of predicted 2D materials and could fill the gaps in the characteristics and properties of the likes of graphene, phosphorene, and silicene. Zhou et al. [164] proposed a high-throughput computational materials design framework, which screened 5,000 compounds from the Materials Project Database (MP), and found 205 layered materials for water splitting photocatalysts and validated 36 kinds of 2D monolayers stability. One can also expand the list of 2D materials by chemical substitutions, alternative site decorations, crystal structure prediction and so on[117]. Several screening approaches have been proposed to find 2D materials from known layered bulk materials [24]. A simple criterion of comparing experimental lattice constants and lattice constants mainly obtained from Materials-Project DFT calculation repository is used to find potential 2D materials [24]: a relative difference between the two lattice constants for a specific material is greater than or equal to 5% is used to identify good candidates for 2D materials. Haastrup et al. [53] developed the Computational 2D Materials Database (C2DB), which contains a variety of structural, thermodynamic, elastic, electronic, magnetic, and optical properties of around 1500 2D materials distributed over more than 30 different crystal structures. More recently, Zhou et al. [169] developed 2DMatPedia, an open computational database of 6351 two-dimensional materials by screening all bulk materials in the database of Materials Project for layered structures by a topology-based algorithm and theoretically exfoliating them into monolayers. New 2D materials have also been generated by chemical substitution of elements in known 2D materials by others from the same group in the periodic table. These databases of experimental or hypothetical 2D materials have made it possible for discovering novel function materials [34, 161, 93, 74, 115].

Despite these efforts, the scale of experimental and hypothetical 2D materials is still limited because of long experimental period and high cost [20]. For example, computational generation of novel new materials have been proposed in the name of inverse materials design [173], in which new materials are to be searched to achieve a given specific function, most of these methods involve a global optimization or search/sampling procedure to explore the search space[21]. However, most of such inverse design research is based on screening known materials. Suleyman Er et al. [140] proposed an elemental substitution based approach and applied it to known 2D materials structural prototypes to generate a large number of hypothetical 2D materials, and then filtered those materials based on several criteria. They deposited their predicted 2D materials in their V2DB database.

To expand the scope of 2D materials, we propose to design a generative deep learning method to discover novel 2D materials in uncharted composition space. Our approach is based on a high-accuracy composition based 2D materials classifier, which is used to screen millions of hypothetical materials compositions generated using our MatGAN, a generative adversarial network (GAN) based model [27] that learns to generate chemically valid hypothetical materials. Based on 2.65 million generated samples, we have identified 267,489 hypothetical 2D materials. Furthermore, we use element substitute method to predicted the crystal structures, and then confirm their structure stability using DFT calculations.

Our contributions can be summarized as follows:

- We propose a composition-based 2D materials classifier model which achieves high prediction accuracy when trained with known 2D materials.
- We combine the 2D materials classifier and the composition-based generative machine learning to discover new 2D materials, which greatly expand the space of 2D materials.
- We apply a template-based element substitution-based structure prediction approach to get the structures of hypothetical 2D materials and verify them using DFT formation energy calculations, exfoliation energy calculation and phonon thermostability verification.

5.2 Methods

5.2.1 2D materials discovery framework

The schematic diagram of our 2D materials discovery framework includes the following four modules (Figure5.1): a GAN based hypothetical materials generator, a composition based 2D materials classifier, a template based structure predictor, and a DFT confirmation procedure. The hypothetical materials generator is trained with known inorganic materials in the Materials Project database to learn the composition rules of forming stable chemically valid materials compositions. Then, we use the generative module to breed a large number of hypothetical formulas (two million in our study). These formulas are then subjected to chemical validity tests including charge neutrality check and electronegativity check. After that, the remaining samples will be screened by the 2D materials classifier using composition alone. To verify the predicted 2D materials compositions, we apply template based element substitution to generate their hypothetical structures for a subset of 624 predicted 2D materials compositions. Using DFT calculations, the stability of these structures is calculated to verify the existence of these candidate 2D materials from which we identified twelve potentially stable materials.

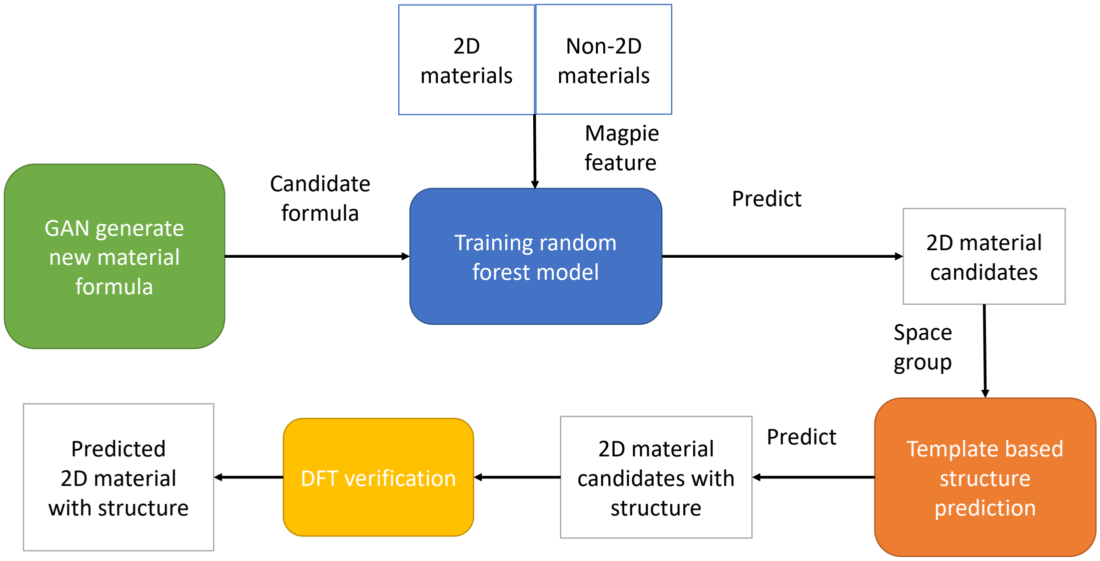

2D materials
Non-2D materials
Magpie feature
Candidate formula
GAN generate new material formula
Training random forest model
Predict
2D material candidates
Space group
Template based structure prediction
Predict
2D material candidates with structure
DFT verification
Predicted 2D material with structure

Figure 5.1 Framework for generation and prediction of 2D materials. It comprises four components. The green part: a GAN-based composition generation module for breeding chemically valid materials. The blue part: a composition-based random forest 2D materials classifier. The orange part: a template-based element substitution structure predictor, as well as the yellow part: DFT validation.

5.2.2 Generative deep learning for hypothetical inorganic materials

In the material design research area, one core task is to explore chemical space for searching new materials. In our previous work, a generative machine learning model (MatGAN) [27] is designed to efficiently generate new hypothetical inorganic materials composition based on generative adversarial network (GAN) [46]. There are two main tasks of MatGAN: one is how to suitably represent material composition; the other one is how to design the generative adversarial network for generating new materials.

When exploring the representation of inorganic materials, we found that there totally are 85 elements in ICSD dataset; and there are no more than 8 atoms per element in any specific compound. Therefore, each material could be represented as a

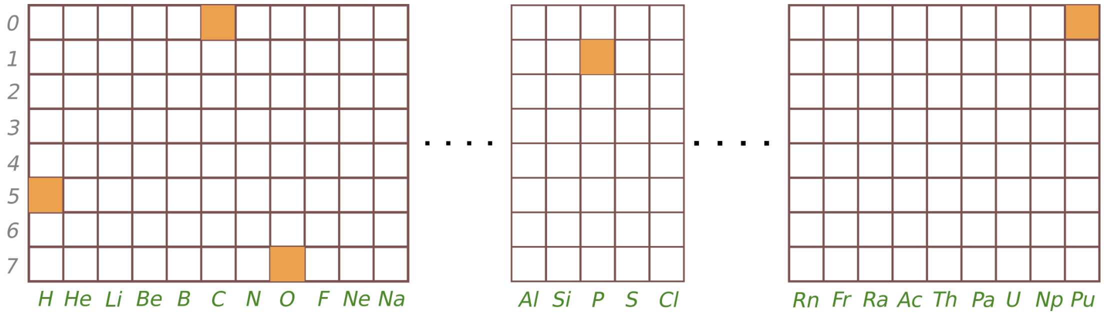


Figure 5.2 One hot representation of material composition $PuP_2H_6CO_8$. Brown color indicates the atom number of corresponding element in the specific material.

sparse matrix M of dimension 8 × 85 with 0/1 cell values, where $M_{i,j} = 1$ means the number of atoms of the element at column j is $i + 1$. Figure 5.2 shows the encoding matrix for $PuP_2H_6CO_8$.

The architecture of MatGAN is shown in Figure 5.3. In this generative adversarial network training, a generator is trained from existing real material representations to generate new samples. Meanwhile, the discriminator tries to differentiate real samples from generated samples; as the feedback, the discrimination loss is then used to guide the training of the generator and the discriminator's parameters to reduce this difference. These two training processes are repeated until good performances of both the generator and the discriminator are achieved. In order to avoid the gradient vanishing issue of standard GAN, we adopt the Wasserstein GAN [8], which replaces the JS divergence distance with the Wasserstein distance. The generator loss and discriminator loss are defined in the following equations:

$$Loss_G = -E_{x:P_g}[f_w(x)] \tag{5.1}$$

$$Loss_D = E_{x:P_g}[f_w(x)] - E_{x:P_r}[f_w(x)] \tag{5.2}$$

where, P_g and P_r are the distributions of generated materials and real materials; $f_w(x)$ is the discriminant network. Equation (5.1) and (5.2) are used to guide the training

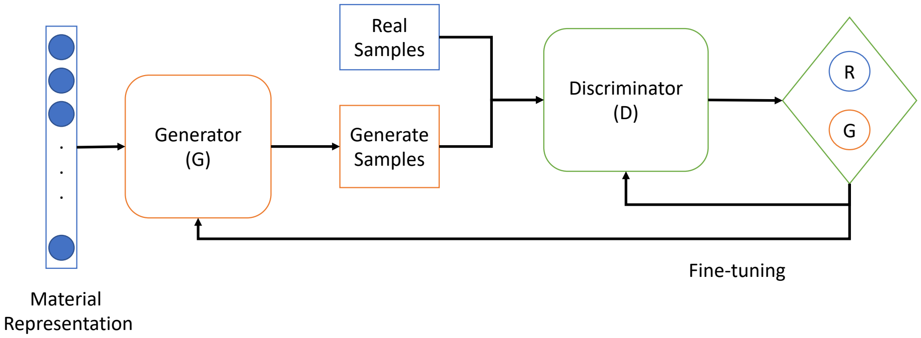


Figure 5.3 Architecture of MatGAN. Generator (G) learns from known material compositions to generate realistic samples while discriminator (D) learns to determine whether a sample is a real one or generated one. Alternative training of D and G will improve the performance of both G and D.

process. The smaller the $Loss_D$, the smaller the Wasserstein distance between the generated samples and the real samples and the better the GAN is trained.

We have generated 2,650,623 hypothetical materials compositions, and 1,940,209 of them satisfy both charge neutrality and electronegativity balance criteria via Semiconducting Materials from Analogy and Chemical Theory (SMACT) [29] tool. The charge neutrality check means that the total charge in a compound should be 0, namely $\sum_i Q_i n_i = 0$, where i are the elements in the compound and Q are the charges. Electronegativity is often used in high-throughput screening, Ginley [17] presented how the simple geometric mean of the electronegativities of a compound, SMACT tool has a built in function to calculate this property for a given composition.

5.2.3 Composition based classifiers for predicting 2D materials

Predicting whether a material is 2D structure can be regarded as a binary classification problem. Our goal is to screen unknown 2D materials from MatGAN-generated materials by training a random forest classifier with verified 2D and non-2D materials, and then predicting the probability of being a 2D material of each new material.

Specifically, we employ the Random Forest (RF) [90] as the surrogate model for predicting the 2D probability given a material' Magpie composition features. Magpie feature set[153]is a well-known descriptor set for composition based machine learning models, those features are calculated by the matminer library [154] which is a Python-based platform that facilitates data-driven methods for analyzing and predicting material properties by calculating a variety of descriptors from material compositions or crystal structures. Basically, magpie feature set calculate the mean, mean absolute deviation, range, minimum, maximum and mode for 22 different elemental properties for all the elements contained in a formula (132 features in total). This elemental property category includes attributes such as the maximum row on periodic table, average atomic number and the range of atomic radii between all elements present in the material. Those 132 features of each material will be calculated and used in our random forest model training.

RF is a supervised bagging ensemble learning algorithm. The idea behind random forests is to exploit the wisdom of the group. RF builds many decision trees in a random way with low correlation among them. After building the forest, when a new sample needs to be classified, each decision tree makes a judgment separately to vote which category the sample belongs to. Random forest improves the prediction accuracy without significantly increasing the amount of computation, and it is relatively robust to unbalanced data.

In the data preparation stage, we first collect known 2D and non-2D materials, as well as MatGAN generated new materials. Then, we calculate the Magpie features for all of them. In training the random forest model, known 2D materials and non-2D materials are treated as positive and negative samples to train the RF classifier with 10-fold cross-validation. The RF hyper-parameters are tuned to achieve good prediction performance with detailed settings explained in Section 3.2.2. Afterward, the trained RF model is utilized to predict the labels and probability scores of 2D for generated hypothetical new materials.

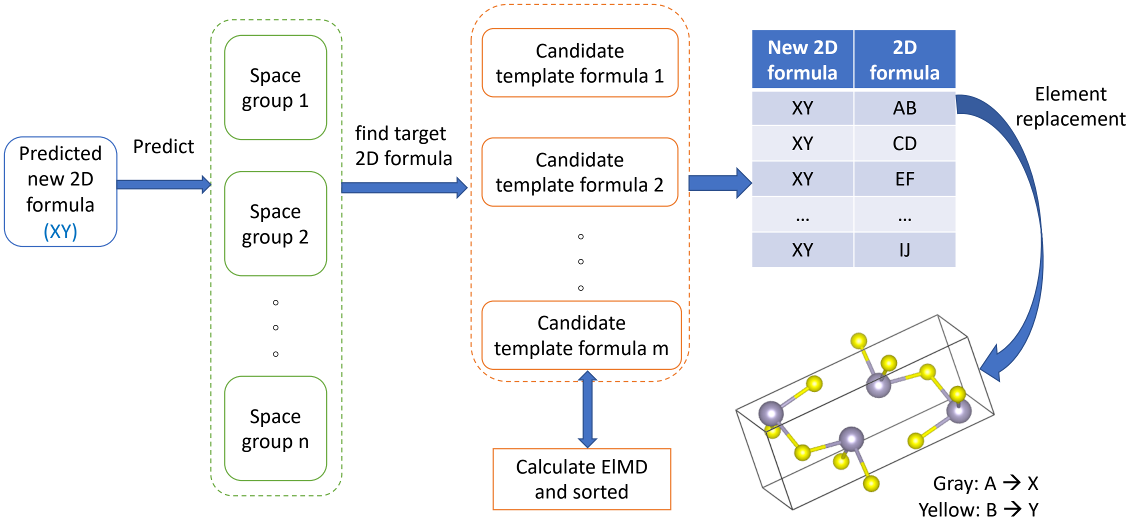
5.2.4 Template-based structure prediction

Although we have predicted the 2D probability scores of the hypothetical new materials generated by our MATGAN algorithm and the candidates with their probability scores greater than 95% are likely to be 2D materials, it is not enough to verify their existence by DFT calculation of their formation energy or phonon calculation based stability check. However, crystal structure prediction of complex compositions using current ab initio crystal structure prediction algorithms are not feasible [112]. To address this issue, we propose to use the template based or element substitution based structure prediction method, which is shown in Figure 5.4.

Firstly, for each predicted 2D formula, we use the Crystal Structure Prediction Network (CRYSPNet) [89] tool to predict its space group that the formula most likely belongs to. This method consists of many neural network models to predict the material's space group, Bravais lattice, and lattice constants. As CRYSPNet only needs chemical composition information as input, we use it to estimate the top 3 potential space groups for each new hypothetical 2D material.

Next, we try to find similar template materials from known 2D materials in the 2dMatpedia database. Specifically, For each new 2D formula with three potential space groups, we search the target 2D material that has the same number of elements and the same space group. However, one formula may lead to many potential target 2D material template. To identify the most similar template material, we use Element Movers Distance (ElMD) [57] machine learning model to calculate and sort similarities between the candidate materials and potential template materials. ElMD is a similarity measure for chemical compositions, which is measured Earth Mover's Distance (EMD) [130] between two compositions from the ratio of each of the elements and the absolute distance between the elements on the modified Pettifor scale.

Finally, we select the top 10 most similar known 2D materials as the structure templates according to the ElMD values. New 2D material's structures could be then predicted by one to one element substitution from those pairs. For example, as shown in Figure 5.4, the structure of XY is predicted from the structure of AB by using X to replace A (gray atom) and Y to replace B (yellow atom).


Predicted new 2D formula (XY)
Predict
Space group 1
Space group 2
Space group n
find target 2D formula
Candidate template formula 1
Candidate template formula 2
Candidate template formula m
Calculate ElMD and sorted
Element replacement
Gray: A → X
Yellow: B → Y

New 2D formula	2D formula
XY	AB
XY	CD
XY	EF
...	...
XY	IJ

Figure 5.4 The framework of template based structure prediction. The main parts are: predicting new formula's space group; finding candidate template formulas according to the same element number and the same space group; calculating and sorting ElMD between the new formula and template formulas; selecting top n templates to do element substitution to get the new formula's structure.

5.2.5 DFT calculation for verification

The density functional theory (DFT) calculations were performed based on the Vienna *ab initio* simulation package (VASP) [82, 83, 36, 81]. The electron-ion interactions were considered by using the projected augmented wave (PAW) method [14, 84]. The energy cutoff value was set as 500 eV. The generalized gradient approximation (GGA) based on the Perdew-Burke-Ernzerhof (PBE) pseudopotentials [118, 119] represented the exchange-correlation potentials. The energy convergence criterion was set as 10^{-7} eV, while the force convergence criterion of the ionic steps was considered as 10^{-2} eV/Å. The Γ-centered Monkhorst-Pack k-meshes were considered to perform the Brillouin zone integration for the unit cells. The van der Waals interactions were considered using DFT+D3 with Becke-Jonson damping[50, 49]. Formation energy per atom (E_{form}) of a material were calculated based on Eq.5.3. Here, E[Layered] is the total energy per unit formula of the corresponding material, $E[\text{A}_i]$ is the energy of i^{th} element of the material, x_i represents the number of atoms of i^{th} element in a unit formula, and N indicates the total number of atoms in a unit formula of the material (e.i., N=$\sum_i x_i$). Exfoliation energies were determined using Eq. 5.4. In this eqaution, E[Monolayer] is the total energy per unit formula of the monolayer exfoliated from a 2D layered material. S is the area of the layered material's surface, which is perpendicular to out-of-plane direction of the monolayer.

$$E_{\text{form}} = \frac{1}{N}(E[\text{Layered}] - x_i \sum_i E[\text{A}_i]) \tag{5.3}$$

$$E_{\text{exf}} = \frac{-1}{2S}(E[\text{Layered}] - E[\text{Monolayer}]) \tag{5.4}$$

5.3 Experiments and Discussion

5.3.1 2D materials Dataset

All the 2D materials are collected from 2DMatPedia[169], an open computational database of two-dimensional materials, which is constructed by a topology based screening algorithm and element substitutions. There are 6,351 2D materials in total, they are regarded as positive training samples in our work. We also collect all existing materials from the materials project database with 126,356 materials in total. After removing known 2D materials, there are 115,498 negative samples. We use the MatGAN model to generate 2,650,264 new materials as candidates for 2D materials prediction.

Table 5.1 Datasets

Dataset	Amount	Role
2dMaterials	6,351	positive training sample
MaterialProject	126,356	negative training sample (exclude 2D materials)
ICSD_2M	2,650,624	potential new material
V2DB	294,077	comparative dataset

5.3.2 Results

Generation of candidate inorganic materials

Trained with 291,840 inorganic materials compositions in Material Project database, a generative deep learning model (MATGAN) is used to generate 2,650,624 new compositions, and then charge neutrality and balanced electronegativity are used to screen out 1,947,792 formulas, among which 1,940,209 are not in the training set.

The number of generated 2-element materials, 3-element, 4-element, ≥5-element are 1217, 18946, 36827, 78639, respectively. There are two reasons to explain why binary materials account for the least proportion: one is the diversity of the combination of two elements is much less than that of five elements. The other is many binary materials have already been discovered. The generated 2-element materials play an important role in subsequent research because most known 2D materials are binary materials.

In order to better display the generated materials distribution information, we draw a line chart (as shown in Figure 5.5) to show the frequencies of 112 elements ordered by atomic number in three datasets: generated ICSD-2M candidate materials by MatGAN, the published 2D material dataset and the predicted 2D materials. From these three curves, we can see that the top 5 crests positions basically overlap, the number ranges are 7-9, 15-17, 32-35, 50-53, and 80-83. It proves that the space of candidate materials generated by MatGAN is consistent with the real 2D materials. Furthermore, it provides a solid candidate range for the following 2D new material prediction.

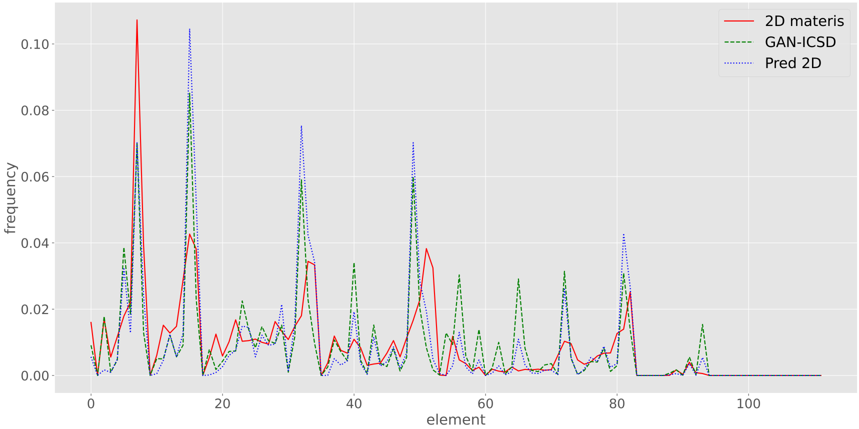


Figure 5.5 Element frequency distribution. The abscissa represents 112 chemical elements arranged according to the atomic number from hydrogen (H) to Copernicium (Cn), vertical axis indicates the frequency of each element. The red curve denotes the distribution information of 2dMaterials dataset. The green curve shows our generated candidate materials. The blue line represents our predicted 2D materials.

Performance of the 2D materials classifier

The hyper-parameter configuration for training 2D random forest classifier is set as follows: we set the maximum tree depth (max_depth) to be 20 and the number of decision trees ($n\_estimators$) as 250. There are 6351 2D material samples and 15,959 non-2D samples. In order to mitigate the imbalanced positive and negative samples, we randomly select 1.5 times the number of positive samples as negative samples. Besides, the class weight parameter is set to be balanced. With these settings tuned per feature iteration, we train the RF 2D materials prediction models and evaluate their performance. Our algorithm is implemented using the Scikit-Learn library in Python 3.6.

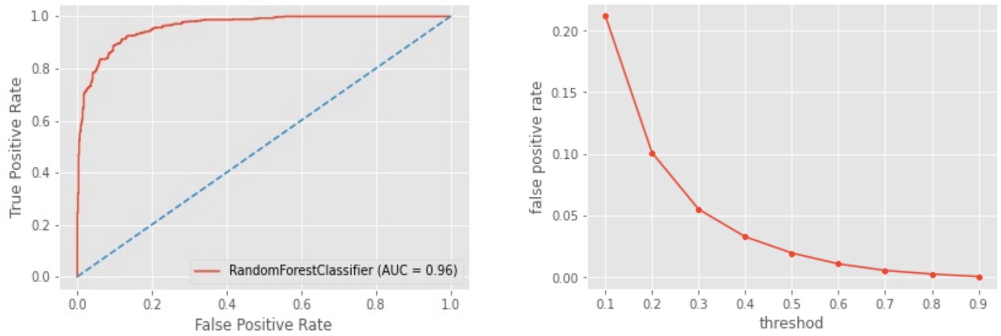
To evaluate the prediction performance of our model, precision, recall, accuracy, F1 score, and receiver operating characteristic area under the curve ROC are used as performance metrics. ROC is plotted with TPR and FPR as the vertical and horizontal axes under different threshold settings.

$$TPR = \frac{TP}{TP + FN} \tag{5.5}$$

$$FPR = \frac{FP}{FP + TN} \tag{5.6}$$

The accuracy, precision, recall, and F1-measure of the RF classifier with 10 fold cross-validation is 88.97%, 88.98%, 88.96%, and 88.96%. Figure 5.6(a) shows the ROC curve of the classifier with an AUC score reaching 96%.

We also use a series of thresholds to differentiate 2D and non-2D materials to evaluate the performance of the RF classifier, as shown in Figure 5.6(b). The abscissa represents the predicted probability threshold to declare a 2D materials when its probability is higher than this threshold value; the y coordinate indicates the corresponding false-positive rate. As the threshold increases, the higher the probability score is required to be judged as a 2D material leading to lower false-positive rate.



(a)

(b)

Figure 5.6 Performance of our RF random forest classifier. (a) ROC curve and AUC score; (b) false-positive rates with different thresholds.

Finding new 2D materials using our framework

To identify interesting hypothetical new 2D materials, we applied our RF-based 2D materials classifiation model to screen the 2.6 million hypothetical materials generated by our Generative Adversarial Network (GAN) based on new materials composition generator [27]. After predicting the probability of each candidate belonging to 2D materials, we sort them by the probability scores. The statistics of the predicted 2D materials with different probability thresholds are shown in Table5.2. With a stringent probability threshold of 0.95, our algorithm has identified 1,485 hypothetical 2D material formulas with 266 binary, 361 ternary, 327 quartenary candidates. When the threshold is lowered to 0.9, the number of candidate 2D formulas increases to 5,034 or to 18,451 with threshold of 0.8.

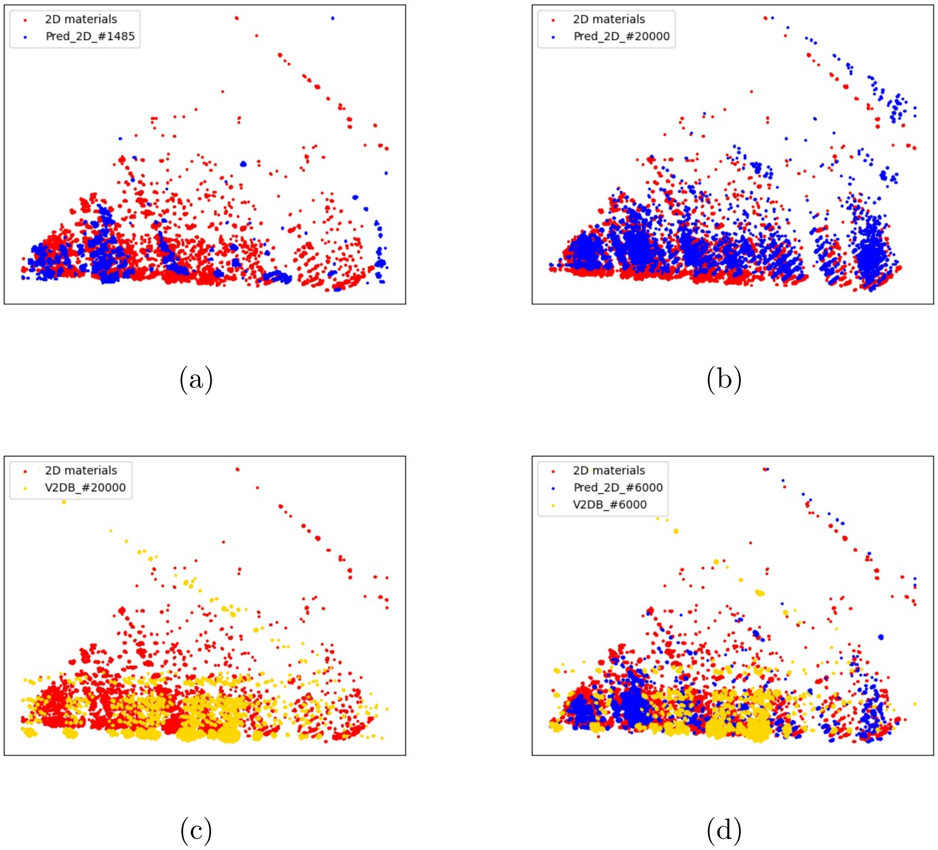
To demonstrate how the newly predicted 2D materials are distributed in the composition space, we apply t-sne dimension reduction tool [96] to map the normalized Magpie features of the 6351 2D materials in the 2Dmatpedia database and the predicted 2D materials, and then plot their distribution in Figure 5.7. In Figure 5.7(a), We apply the same dimension reduction transformation to both the training set and the newly predicted 2D materials and visualize their distribution where red points are training samples, and blue points are the 1485 predicted 2D materials with the highest probability scores. Similarly, top 20,0000 predicted 2D materials are drawn as blue points in Figure 5.7(b) together with known 2D materials used for training. We found that in Figure 5.7(a) the majority of blue points are located in the dense red point areas in the bottom left corner (which can be seen also from Figure 5.7(c)), indicating that our predicted 2D materials have similar composition distribution with regard to known 2D materials. Figure 5.7(b) further confirms this composition distribution match, in which we find that the blue points in general only appear in areas with red points. The areas with sparse red points also contain few blue points. Figure 5.7(c) shows the distribution of 20,000 predicted 2D materials in the V2DB dataset against the known 2D materials. It is found these candidate 2D materials have different composition distribution as regard to the known 2D materials: many yellow points appear in areas without red points. Quite many yellow points reach out of the boundary defined by the red points. For better comparison, top 6000 predicted 2D materials by our method and 6000 predicted 2D materials of V2DB are drawn in Figure 5.7(d). It can be seen that the majority of the overlapped blue and yellow points (117 as shown in Table 5.3) are located in the lower-left corner, which means there are new 2D materials that are jointly predicted by both methods.

To screen out top candidate materials, we use the Roost algorithm [45] for formation energy prediction, which is a graph network based machine learning model for materials property prediction using only composition information. After predicting the formation energies of all candidates, we draw a histogram of formation energy distribution, as shown in Figure 5.5. Furthermore, we filter out those with probability scores greater than 0.95 and then sort them by the formation energy in ascending order and pick the top 40 candidates with 2, 3, and 4 elements respectively. The results are in Table 5.2.

Furthermore, we analyze the 2DMatPedia dataset, 2-element materials occupy 65%, 3-element materials, and 4-element materials account for 25% and 9%, therefore, from the perspective of the probability distribution, our prediction is meaningful. We also find that the predicted 2D probabilities of 2-element materials are in general higher than those of 3-element materials and 4-element materials, corresponding to the fact that the majority of known 2D materials (65%) are binary materials. We also count the number of our predicted new 2D materials with 2D probability greater than 0.5 that overlap with those in V2DB and 117 hypothetical 2D materials are found to be predicted by both methods. Table 5.3 shows the overlapped candidate 2D materials in five parts according to their 2D probability scores. It is found the overlapped materials with 2D probability greater than 0.8 account for nearly 50% of all overlapped candidates.

Table 5.2 Statistics of predicted 2D materials

2D Prob	# of Predicted 2D formula	# 2 element	# 3 element	# 4 element	# ≥ 5 element
0.95	1,485	266	861	327	31
0.9	5,034	439	2,617	1,695	283
0.8	18,451	729	8,123	7,316	2,283
0.7	48,592	942	16,827	21,430	9,393
0.6	119,489	1,146	28,172	51,998	38,173
0.5	267,489	1,340	40,382	99,943	125,824



(a)

(b)

(c)

(d)

Figure 5.7 Distribution of the new and existing 2D materials in 2D space generated by t-SNE embedding from magpie features. Red ones are known 2D materials, blue ones are our predicted 2D materials while yellow ones are predicted 2D materials from V2DB. (a) Distribution of 1485 predicted 2D materials with probability score >0.95. (b) Distribution of 20,000 predicted 2D materials randomly selected from 260,000 candidates with probability score >0.5. (c) Distribution of 20,000 predicted 2D materials randomly selected from 290,000 candidates of the V2DB database. (d) Distribution of 6000 predicted 2D materials by our model and 6000 from V2DB randomly selected from 260,000 and 290,000 candidates respectively.

Table 5.3 Predicted hypothetical 2D materials that overlap with V2DB

Formula	Prob	Formula	Prob	Formula	Prob	Formula	Prob	Formula	Prob
	>0.9		>0.8		>0.7		>0.6		>0.5
AgS	0.9999	ZrClS	0.8997	TiSeCl	0.7913	VISe	0.6879	NiF	0.5998
AlS	0.9982	AlIS	0.8926	ZrSeF	0.7900	NiSeCl	0.6830	ZrTeN	0.5993
ScI	0.9840	VClF	0.8917	SbSeO	0.7856	CoBrF	0.6797	MnSeS	0.5872
InTeS	0.9753	FeClS	0.8909	ZrTeSe	0.7802	NiSF	0.6794	SrSF	0.5840
SnTeSe	0.9748	PbTeS	0.8884	MnBr	0.7793	RhSeS	0.6759	YSeS	0.5709
SnTeS	0.9737	GeSF	0.8852	MnIS	0.7782	CoTeS	0.6740	NbSeO	0.5669
PbTeSe	0.9465	BiSF	0.8830	CuTeO	0.7647	CrSeO	0.6732	CoF	0.5638
SbSeF	0.9421	TiClF	0.8800	CoCl	0.7602	CoSF	0.6714	NbTeO	0.5636
SbClS	0.9418	ZnBrS	0.8777	GeSO	0.7588	RhTeS	0.6710	TaSO	0.5604
GeClS	0.9409	MnClS	0.8757	ZnSF	0.7579	MnBrN	0.6697	TaTeS	0.5506
AsClS	0.9374	SnSeO	0.8727	ZnSO	0.7579	AlSeO	0.6609	NiSO	0.5474
SbSF	0.9224	ZnClS	0.8725	AgSO	0.7557	PbTeO	0.6540	NiSeO	0.5419
NbSe	0.9221	AgIS	0.8670	CuSeF	0.7546	MnSeO	0.6473	YTeS	0.5249
AlSeS	0.9197	PbSeO	0.8535	MnSF	0.7518	VSeS	0.6451	TiSeN	0.5186
SnSO	0.9181	AsSF	0.8532	MnSeF	0.7466	NbSeF	0.6367	SrClS	0.5120
BiTeS	0.9102	AgTeS	0.8524	AsSO	0.7422	RuTeS	0.6310		
AlTeS	0.9060	PbSO	0.8494	AgSeO	0.7406	TaSeO	0.6217		
		YClS	0.8471	ZrSeS	0.7392	FeSeF	0.6190		
		BiTeF	0.8462	CoClF	0.7362	TaTeSe	0.6077		
		NiClS	0.8457	PdFO	0.7319	CrTeO	0.6066		
		YTeF	0.8457	ZrTeO	0.7317	ZrSeN	0.6014		
		CoClS	0.8433	MnI	0.7312				
		VClS	0.8397	CuSF	0.7245				
		VSF	0.8395	AlSO	0.7239				
		AlTeCl	0.8285	TiTeS	0.7230				
		NiBr	0.8254	NiClO	0.7077				
		SbSeS	0.8220	NiTeCl	0.7056				
		AgSeS	0.8217	NbSO	0.7005				
		BiSO	0.8201						
		ZnClF	0.8176						
		TiSeF	0.8155						
		MnCl	0.8077						
		FeI	0.8074						
		AlSeF	0.8017						
		AlTeF	0.8015						
		CoI	0.8002						

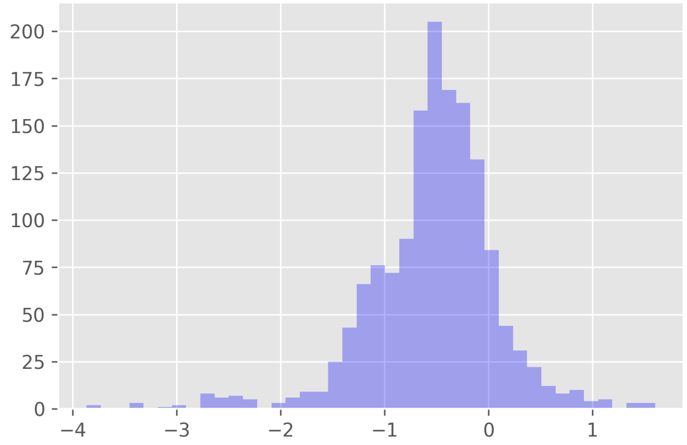


Figure 5.8 Formation energy distribution of predicted 2D materials.

Structure prediction and verification

To verify the predicted 2D materials, we pick 1485 predicted material formulas with the highest probability scores (≥ 0.95) and use the template based structure prediction method to find their structures. With the result that we find 101 materials' templates of known layered materials or 2D materials in the 2DMatPedia database. We then predict the space group of our predicted 2D materials and choose the templates with the same or similar space groups. Next, element substitution method has been used to get their crystal structures. Some of the predicted 2D materials structures are shown in Figure 5.11. In total, 101 predicted materials with structures have been obtained.

To further verify whether these hypothetical materials are thermodynamically stable, we applied DFT first principle calculation to compute the formation energies per atom for the 101 2D-layered materials which have template structures. In total we found 92 hypothetical materials with negative formation energies (see Supplementary Table S1), there are 79 binary 2D materials, 10 ternary 2D materials and 3 quaternary 2D materials. The materials with high formation energies like TaF_4 (-2.9431 eV/atom) and SiF3(-2.1204 eV/atom) imply that the proposed method in this research is able

Table 5.4 Hypothetical 2D materials sorted by predicted formation energy(only top 120 are listed here)

2 Elements			3 Elements			4 Elements		
Formula	Prob	E_{form}[Layered-ML] (eV/atom)	Formula	Prob	E_{form}[Layered-ML] (eV/atom)	Formula	Prob	E_{form}[Layered-ML] (eV/atom)
ZrF3	0.9876	-3.8661	ScYF3	0.9760	-3.3429	NbMoCl5S2	0.9757	-1.5188
YF2	0.9880	-3.7549	NbMoF6	0.9720	-3.0069	CrNbMoCl8	0.9638	-1.4547
TaF4	0.9990	-3.4457	ZnTaF5	0.9755	-2.9848	CrNbRuCl8	0.9598	-1.3912
SiF3	0.9607	-3.3363	NbRuF7	0.9665	-2.7521	NbRuCl5S2	0.9753	-1.3857
ZrF2	0.9997	-3.1413	InSnF5	0.9519	-2.7437	NbRuCl4S2	0.9712	-1.3826
ScF	0.9760	-2.7647	TaIrF7	0.9794	-2.7379	NbRuCl6S	0.9673	-1.3719
YF	0.9640	-2.747	NbRuF6	0.9708	-2.7165	NbRuCl6S2	0.9633	-1.3688
GeF3	0.9705	-2.6997	TaWO5	0.9732	-2.6845	CrMoCl5S	0.9517	-1.3163
NbF2	0.9671	-2.5717	ScYCl6	0.9677	-2.6222	NbMoRuCl6	0.9676	-1.2988
YCl2	0.9919	-2.4657	NbMoO5	0.9816	-2.6101	Ga3AsCl6S2	0.9514	-1.2548
ZrF	0.9720	-2.4356	TaWF5	0.9760	-2.5844	InSn3SeCl6	0.9513	-1.2127
WF3	0.9999	-2.4182	TaIrF5	0.9511	-2.5832	SnSbAsCl8	0.9593	-1.205
TaF2	0.9880	-2.4126	ScYCl4	0.9679	-2.527	MoRuCl5S2	0.9633	-1.171
AlF	0.9999	-2.4107	YZrCl6	0.9598	-2.4481	Sn2AsCl6S	0.9550	-1.1705
ScCl2	0.9799	-2.338	ScZrCl6	0.9518	-2.3783	InSn3Cl4S2	0.9907	-1.1704
GaF	0.9995	-2.0765	MoRuF6	0.9708	-2.352	MoRuCl6S2	0.9633	-1.1499
FB	0.9880	-2.0746	TaOsF5	0.9640	-2.3308	Sb2BrAsCl8	0.9671	-1.1406
F3C	0.9800	-2.0634	ScTiCl4	0.9600	-2.2386	InSnCl2S2	0.9745	-1.1381
InF	0.9880	-1.9397	VTc2F7	0.9661	-2.2366	InSnCl4S2	0.9709	-1.1229
Tc2O5	0.9650	-1.9029	VZrCl6	0.9560	-1.9388	SnSbAsCl4	0.9514	-1.1043
ErBr3	0.9674	-1.8561	YLaBr6	0.9547	-1.9262	Sn2AsCl4S	0.9628	-1.0983
OsF3	0.9837	-1.7279	TbDyBr6	0.9729	-1.8741	Sn2AsCl8S2	0.9627	-1.0955
NbCl3	0.9959	-1.7273	TiVCl6	0.9519	-1.8074	Sn3SeCl6S2	0.9509	-1.0928
S2O5	0.9859	-1.7237	Nb3Cl8S	0.9919	-1.7429	Sn2AsCl7S2	0.9587	-1.0901
NbCl2	0.9799	-1.7137	NbCl2S	0.9599	-1.6969	SnPb2Br2Cl2	0.9573	-1.0892
OB	0.9600	-1.7046	AlClS	0.9596	-1.6887	Sn2AsCl6S2	0.959	-1.0838
TaCl2	0.9950	-1.6658	TaRe2F6	0.9640	-1.6724	InSn2Cl2S2	0.9755	-1.0789
RuF2	0.9864	-1.5312	CrNbCl7	0.9677	-1.6054	Sn2As2Cl6S	0.9511	-1.0705
SF	0.9798	-1.4954	CrNbCl6	0.9758	-1.602	InSnClS2	0.9820	-1.0414
MnCl3	0.9997	-1.4641	CrNbCl8	0.9637	-1.5821	Sn2AsCl4S2	0.9628	-1.0387
CrCl4	0.9838	-1.4529	TiYBr6	0.9538	-1.5738	CuRuAgCl6	0.9629	-1.0133
SeF	0.9795	-1.436	SnPbCl6	0.9577	-1.5723	Sn2AsCl6S3	0.9587	-1.011
O3C2	0.992	-1.3673	VMnCl4	0.96	-1.5719	Sn2SeCl6S2	0.9509	-1.0109
GeCl3	0.9988	-1.3663	CrNbCl5	0.9598	-1.5622	Sn2SeCl5S2	0.9549	-1.0048
YS3	0.9618	-1.3661	Te2SO5	0.9508	-1.5303	AsGeCl6S	0.9590	-1.0045
V2S3	0.9661	-1.3311	InSnCl6	0.9795	-1.5166	AsGe2Cl6S2	0.9636	-0.9982
TiCl	0.9599	-1.3269	SnTlCl5	0.9507	-1.5121	Sn2As2Cl6S3	0.9547	-0.9832
HoS3	0.9755	-1.3082	TaWCl6	0.9794	-1.5067	Sn2SeCl4S2	0.9549	-0.9821
SmS3	0.9555	-1.3041	NbMoCl6	0.9839	-1.5054	Sn2AsCl6S4	0.9590	-0.977
YI2	0.9999	-1.2846	Cl6SSi2	0.9877	-1.5012	SnPb2Se2Cl2	0.9538	-0.9457

to discover 2D layered materials which are highly thermodynamically stable against the parent compounds of their elements.

Furthermore, we studied the exfoliation possibility of 2D-layered materials using exfoliation energy based on Eq. 5.4. Modeling monolayers from around 100 2D-layered materials is computationally expensive. Thus, here we modeled only 31 monolayers to show that the proposed computational technique can find 2D materials with very low exfoliation energies (see Supporting Information Table S2). The 12 materials with lowest exfoliation energies are mentioned in Table 5.5. The formation energies of those monolayers are also negative indicating they are stable relative to the compounds of respective elements.

In order to demonstrate that our predicted compositions and structures can be used to discover stable layered materials, we computed the elastic constants using density functional perturbation theory (DFPT) [11] and phonon bands using Phonopy code [143] for V_2S_3. Our DFT calculations show that V_2S_3 is a stable 2D-layered material. The structure of this material is shown in Figure 5.9. V_2S_3 has Triclinic crystal symmetry with P-1 (2) space group symmetry. The lattice parameters were found as $a = 3.072$ Å, $b = 7.130$ Å, $c = 9.216$ Å, $\alpha = 87.06$, $\beta = 80.40$, and $\gamma = 77.59$. The formation energy calculated based on Eq. 5.3 is around -0.618 eV/atom. The enthalpy difference between competitive phases determined based on the expression $\Delta H = E[\text{Material}] - E[\text{competitive phases}]$ using the total energy E of the material and its competitive phases. The competitive phases were found from the Material Project database. The computed maximum enthalpy difference for V_2S_3 is 0.046 eV/atom against the stable phases V_5S_8 and V_3S_4.

Based on the Hill approach, we found the Bulk modulus, Shear modulus, and Young's modulus as 26.29, 18.88, 45.69 GPa, respectively [61]. The total energy of a material can be denoted by $E = E_0 + \frac{1}{2}V_0 \sum_{i,j=1}^{6} C_{i,j}\epsilon_i\epsilon_j + O(\epsilon^3)$, when an infinitesimal strain (ϵ) is applied. Here, C is the matrix of second-order elastic constants. The Born elastic stability criteria require that the matrix be definite positive, and all eigenvalues of C and all the principal components should be positive. The relationships between the elastic constants of Born criteria of P-1 (2) crystal symmetry are very complicated since the triclinic systems have 21 independent elastic constants [103]. Therefore, VASPKIT [147] code was employed to calculate the elastic properties. It confirms that V_2S_3 is mechanically stable. The all phonon frequencies in Figure 5.10 are positive, implying the material is dynamically stable at 0 K temperature.

Moreover, we performed phonon calculations for V_2S_3 monolayer as shown in the Supporting Information Figure S1. It is clear that this nanosheet also dynamically stable at 0 K temperature. The exfoliation energy and formation energy of the monolayer are negative suggesting that V_2S_3 nanosheet can be exfoliated from the parent layered-material (see Table 5.5).

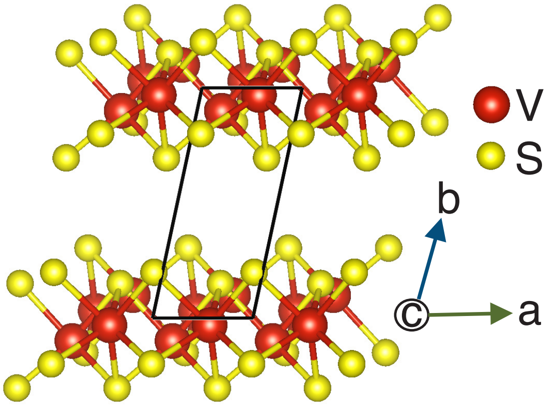


Figure 5.9 Structure of V_2S_3 2D-layered material

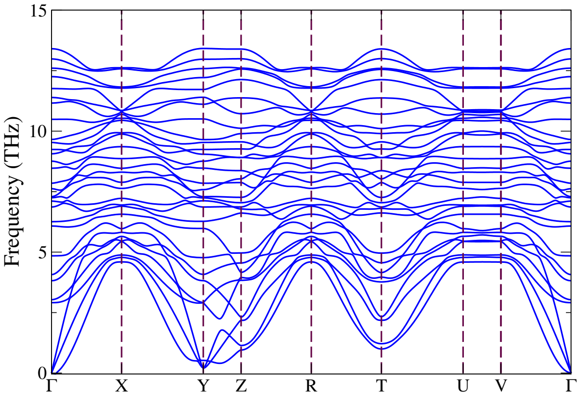


Figure 5.10 Phonon bands of V_2S_3 2D-layered material

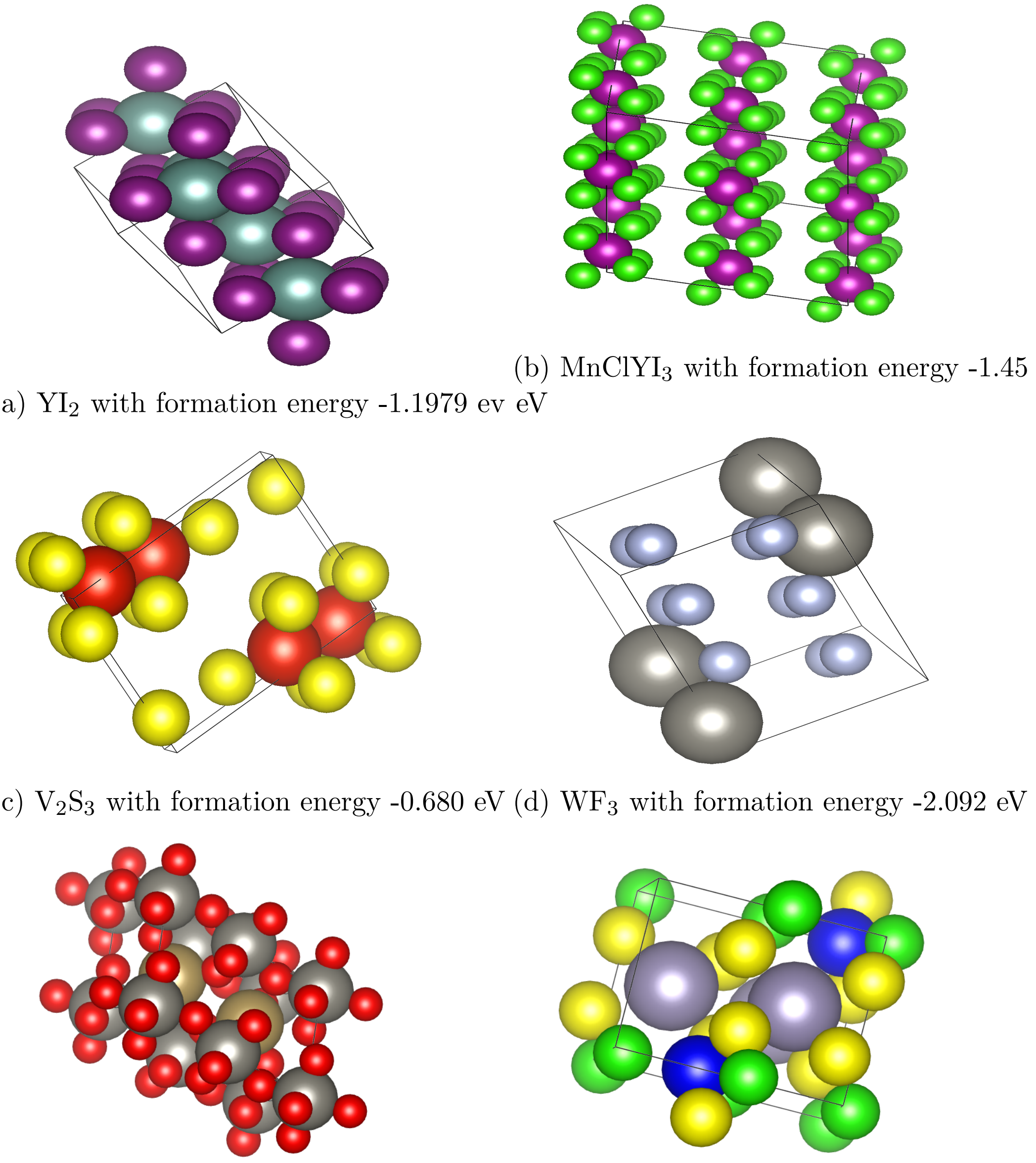
5.4 Conclusion

We propose a generative inverse design approach for finding hypothetical new 2D materials. It includes a GAN based composition generation model for generating chemically valid materials formulas, a composition based random forest 2D materials classifier, a template based element substitution structure predictor, and DFT

Table 5.5 The 12 compositions with lowest exfoliation energies found using DFT are mentioned in the table. The formation energies of layered materials using DFT (E_{form}[Layered-DFT]) and using the machine learning model (E_{form}[Layered-ML]), and the formation energies of monolayers using DFT (E_{form}[Monolayer]) are also stated.

Formula	E_{form}[Layered-DFT] (eV/atom)	E_{form}[Layered-ML] (eV/atom)	E_{exf} (meV)	E_{form}[Monolayer] (eV/atom)
S2O5	-0.6417	-1.7237	-87.8131	-1.4659
V2S3	-0.6830	-1.3311	-11.7012	-0.8130
CoCl3	-0.2434	-1.0299	-2.4151	-1.8472
YI2	-1.1979	-1.2846	-1.1606	-1.2315
OB	-2.0188	-1.7046	-0.6184	-2.0268
TaF4	-2.9431	-3.4457	-0.1287	-2.8880
YCl2	-1.9251	-2.4657	0.5745	-3.3604
ScS3	-0.9223	-1.1463	1.6923	-0.8984
WS3	-0.2765	-0.4336	2.3043	-0.1630
SiF3	-2.1204	-3.336	2.5126	-2.0726
GaCl	-0.6627	-1.0418	2.5508	-1.6550
PBr4	-0.2054	-0.3209	2.6124	-0.0463

verification. Using this pipeline, we have generated 1485 hypothetical 2D material compositions with probability scores greater 95%. We computationally verified that 92 materials have negative formation energies using DFT. We also modeled 31 monolayers from the proposed structures and found that the all 31 materials provide exfoliation energies less than 200 meV showing high possibility of exfoliating nanosheets from their layered materials. These new hypothetical materials can be used to guide the screening of 2D materials for special functions using materials property prediction models. The experiments demonstrate the effectiveness of the proposed approach for discovering new 2D materials and can be used as a complement of the prototype based element substitution based generation approach. Currently, our method is constrained by the limited capability of the crystal structure prediction step, which is an unsolved problem. More powerful crystal structure prediction methods are needed to identify 2D materials of novel structural prototypes, which cannot be identified using template-based structure modeling approach as we use here.



(a) YI_2 with formation energy -1.1979 ev eV

(b) $MnClYI_3$ with formation energy -1.45

(c) V_2S_3 with formation energy -0.680 eV

(d) WF_3 with formation energy -2.092 eV

(e) $TaWO_5$ with formation energy -2.747 eV

(f) $SnAsClS_2$ with formation energy -0.360eV

Figure 5.11 Selected structures of the discovered new 2D materials with DFT validation. (a) structure of predicted YI_2 with DFT calculated formation energy -1.1979 eV. The predicted crystal structures and DFT calculated formation energies for $MnClYI_3$ V_2S_3, WF_3, $TaWO_5$, $SnAsClS_2$ are shown in (b),(c),(d),(e),(f), respectively.

Chapter 6

Machine learning based Prediction of piezoelectric and noncentrosymmetric materials

6.1 Introduction

Piezoelectric materials can generate charge from applied stress (Figure 6.1). They are also able to exhibit the inverse piezoelectric effect (as shown in Figure 6.2), which is the generation of mechanical strain reacting to an applied electrical field [107, 31]. These two unique properties have enabled a variety of applications such as ultrasonic detectors, microphones, sonar devices, and ignition systems, which are all based on the piezoelectric effect [156]. The piezoelectric materials themselves are used in many daily appliances such as electric cigarette lighters, gas grills and burners, and cold cathode fluorescent lamps, electric guitars, electronic drum pads, and medical acceleromyography. Piezoelectric materials also can be used as actuators for accurately positioning objects, which is helpful in loudspeakers, piezoelectric motors, laser electronics, inkjet printers, diesel engines, and x-ray shutters.

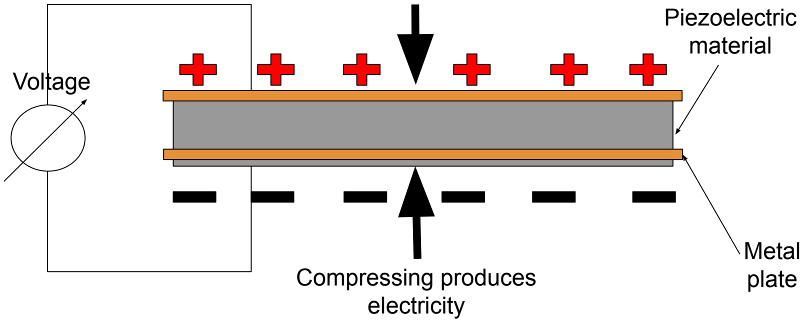


Figure 6.1 Piezoelectric effect.

In general, there are three types of piezoelectric materials: naturally occurring, man-made, and ceramic materials. Naturally occurring crystals include quartz, sucrose, Rochelle salt, topaz, tourmaline, and berlinite ($AlPO_4$). Man-made piezoelectric crystals include gallium orthophosphate ($GaPO_4$) and langasite ($LA_3Ga_5SiO_14$). Ceramic piezoelectric materials include $BaTiO_3$, $PbTiO_3$, PZT, $KNbO_3$, $LiNBO_3$, $LiTaO_3$, Na_2WO_4. PZT is currently the most commonly used piezoelectric ceramic; as it is not environmentally friendly, several alternative materials were discovered include sodium

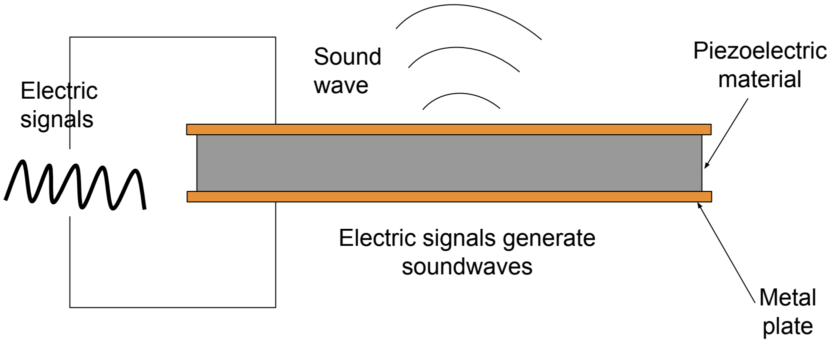


Figure 6.2 Inverse piezoelectric effect.

potassium niobate (NaKNb), bismuth ferrite ($BiFeO_3$), sodium niobate (NaNbO3). New applications such as energy harvesting call for developing new types of lead-free piezoelectric materials. Despite the importance of piezoelectric materials, there is no solid understanding of how the crystal structure determines the piezoelectric modulus and how to design materials with higher piezoelectric coefficients.

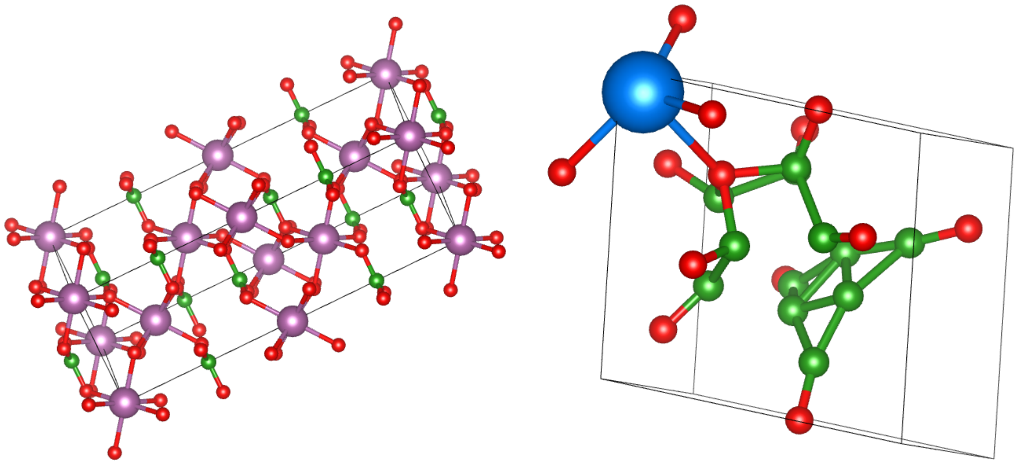
In this work, we aim to develop a machine learning model for the accurate prediction of piezoelectric modulus to be used for screening novel environmentally friendly piezoelectric materials. The piezoelectric coefficient or piezoelectric modulus, usually written d33, measures the volume change when a piezoelectric material is subject to an electric field or the polarization on the application of a stress. There have been a variety of machine learning models developed for materials property predictions such as formation energy, band gaps [106, 43], fermi energy [159], hardness [99], Poisson's ratios, elastic (shear/bulk moduli) [159, 166, 127], superconductor transition temperature [100, 141, 88, 26, 126], ion conductivity [32, 52, 136, 58], flexoelectricity [38, 39, 31] and etc. These ML models can be categorized into three main categories in terms of the input information: composition descriptors based models, structure information based models, and hybrids. See [32] for a comprehensive set of descriptors and related ML models. With a sufficient amount of dataset, it has been shown that composition based ML models alone can achieve highly accurate models for formation energy [153, 68] and band gap predictions [172]. Actually, some of those high-performance reports of composition based ML models are very likely due to the high redundancy of the test sets as regards to the training set when random splitting or cross-validation evaluations are used for large datasets such as Materials Project. Due to the tinkering discovery and study process of materials over history, these datasets tend to have many highly similar samples. Three recent solid benchmark evaluations have clearly shown that the structure based prediction models most often outperform those composition models [32, 12, 35]. For example, Bartel et al. [12] showed that composition models failed to distinguish inorganic materials' relative stability. Instead, they found that including structure in the representation can lead to non-incremental improvement in stability predictions (with CGCNN graph neural network), which serves as a strong endorsement for structural models. Even with this range of ML models for diverse materials properties, there is currently no study on ML prediction of piezoelectric coefficients in the literature.

Appropriately representing materials' characteristics is an important and necessary task in machine learning. In general, material features [133] can be divided into composition features, structure features, and other more complex representations calculated by specific models. Composition features mainly include chemical element stoichiometric information such as atomic element types, the number of elements, atomic weight; Magpie [153] is a commonly used composition feature set. Structural features focus on crystal system information and present the chemical species and atomic coordinates instance. Typical structural features include Coulomb matrix, Ewald sum matrix, sine matrix, Many-body Tensor Representation (MBTR), Atom-centered Symmetry Function (ACSF), and Smooth Overlap of Atomic Positions (SOAP) as implemented in the Dscribe library [62]. There are more complex representations from molecule-oriented features to descriptors for extended materials systems and tensorial properties, such as symmetrized gradient-domain machine learning (sGDML) [23].

However, a recent benchmark study shows that representation learning enabled by graph neural networks such as crystal graph convolutional neural networks (CGCNN) [159] tends to greatly outperform traditional heuristic features such as SOAP features. Other deep representation learning models have also been applied to the grid or voxel-like representations for materials property prediction and generations [75, 132, 166].

Currently, there are 1,705 materials in the Materials Project database with measured piezoelectric modulus. Based on those known piezoelectric materials, this paper aims to train a machine learning model that can predict piezoelectric coefficients with good performance. Here, we explore 6 types of features to train random forest models and support vector machine models. Besides, 5 graph neural networks combined with composition and structural features are also trained. Their performances are evaluated by k-fold cross-validation in experiments. Finally, we apply the trained SVM model to predict 12,680 materials' piezoelectric coefficients and report the top 20 potential piezoelectric materials.

Nonlinear optical materials (NLO), in which light waves interact with each other, are one of the key enablers for next generation of new lasers, fast telecommunication, quantum computing, quantum encryption, dynamic or optical storage data, and many other applications [114, 56, 80, 2]. NLO materials are most broadly defined as those compounds capable of altering the frequency of light. Depending on the chemical and physical construct of the materials they can combine multiple photons to generate shorter wavelength photons or split one photon into several new photons of longer wavelengths. These new photons can be employed to perform all of the above applications as well as many others. The classes of NLO materials range broadly from inorganic oxides such as $KTiOPO_4$ and $LiNbO_3$ to semiconductors like to periodically poled GaAs, to organic polymers to metal organic framework (MOFs), and to simple small organic molecules like stilbene. This broad range of materials has many different properties and characteristics but all are united by one common factor, i.e. their lattice structure must not contain a center of symmetry and must be acentric [114, 56]. This is a rigorous requirement that can only be met in well-ordered lattice structures, meaning ordered crystals. It is generally difficult to design and grow acentric single crystals and less than 15% of all known structures are acentric. This demands exceptional determination on the part of the synthetic and crystal growth experimentalists. The process is made even more difficult by the fact that the NLO processes that enable frequency modification are inherently inefficient. Moreover, the ability to prepare new NLO materials and study their properties is not trivial and requires patient and detailed investigations. The payoff is enormous however, as the materials enable the development of devices used in next generation laser surgery, imaging, optical communication, advanced spectroscopy, optical data storage and a vast array of applications dependent on the interaction of light with matter. In Figure 6.3, We show the crystal structures of a centrosymmetric material and a noncentrosymmetric material, namely $ScBO_3$ and SrB_12O_7.



(a) Centrosymmetric: ScBO3 (R$\bar{3}$c) (b) Noncentrosymmetric: SrB12O7 (R3)

Figure 6.3 Crystal Structures of centrosymmetric and noncentrosymmetric materials. (a) The crystal structures of ScBO3 of space group R$\bar{3}$c, where the purple nodes represent Sc atoms, the green nodes represent B atoms and red nodes are O atoms. (b) The crystal structure of SrB12O7 of space group R3, where the blue node represents Sr atom, the green nodes represent B atoms, and the red nodes are O atoms.

we also propose and evaluate two machine learning models including RF and multilayer perceptron (MLP) neural network models for noncentrosymmetric classification given only material composition. The Magpie composition descriptors are used in our study. Cross-validation and hold-out experiments show that RF with Magpie features achieved the best results. A further application of our RF noncentrosymmetric prediction model to screening two million hypothetical materials generated by our generative ML model [28] allows us to identify and predict dozens of potential novel noncentrosymmetric materials with high confidence scores.

6.2 Methods

We downloaded the piezoelectric coefficient dataset from the Materials Project (MP) database, containing 1,705 inorganic materials, 65 samples' piezoelectric coefficient (PC) is 0, and there are only 8 materials with a coefficient greater than 20

C/m^2 including: $AgBiO_3$(24.84), Sm_2CdSe_4(32.504), Li_4CO_4(33.35), $MnCO_3$(39.83), Na_4CO_4(50.45), $Ba(Si_3N_4)2$(67.67), $CdGeO_3$(75.01), Pr_3NF_6(86.09).

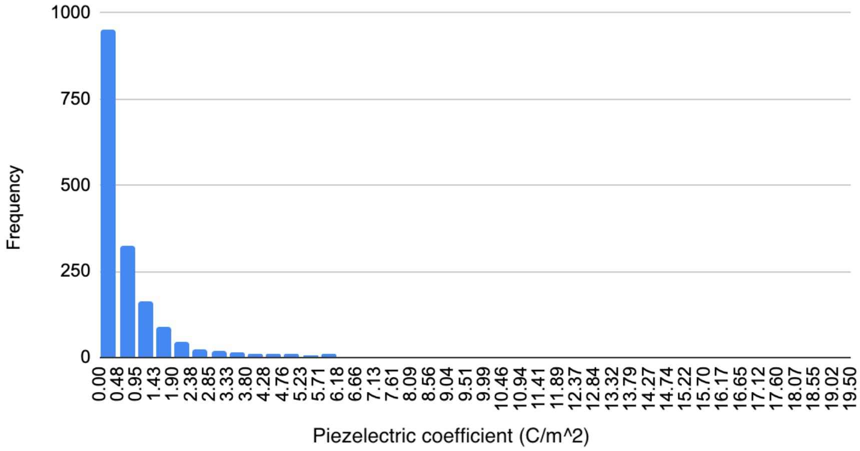


Figure 6.4 Piezelectric coefficient distribution of the MP dataset. For better visualization, we have excluded the 8 samples with piezoelectric coefficent >20 C/m^2.

Based on the graph shown in Figure 6.4, the piezoelectric coefficient values are highly unbalanced with very few samples with high piezoelectric coefficients, which makes their prediction very challenging.

6.2.1 Features

Selecting a set of appropriate material features is critical to train a successful machine learning model for predicting piezoelectric properties, which are unique and challenging to predict, as we found during our research process. To solve this problem, we try a wide range of different features, including the following:

- Composition based magpie features [153]: these are elemental based property features which are composed of 6 statistics (mean, mean absolute deviation, range, minimum, maximum, and mode) of a set of 22 elemental properties, including atom number, Mendeleev number, atomic weight, melting temperature, periodic table column, periodic table row, covalent radius, electronegativity, Ns valence(number of s orbital valence), Np valence, Nd valence, Nf valence, N valence, Ns unfilled, Np unfilled, Nd unfilled, Nf unfilled, N unfilled, GS volume pa, GS band gap, GS magnetism moments, space group number.

- Oxidation states features: these are statistics (maximum, range, standard deviation) about the oxidation states for each species.

- Structural features: we found that the symmetry degree of the crystal structure has substantial effects on the piezoelectric effects, so we defined a set of crystal structural features based on the crystal systems of the materials. These features include (1) number of perpendicular face pairs of the unit cell, (2) the number of equal edges of the unit cell, (3) one hot encoding of crystal systems (cubic, hexagonal, monoclinic, orthorhombic, tetragonal, triclinic, trigonal). We also included the following features: (1) maximum atom radius, (2) minimum atom radius, (3) average atom radius, (4) geometric mean of atom radius, (5) standard deviation of atom radius, (6) maximum Poisson's ratio, (7) minimum Poisson ratio, (8) average Poisson's ratio, (9) geometric mean of Poisson's ratio, (10) standard deviation of Poisson's ratio.

- Energy and magnetism features: (1) e_above_hull, (2) band gap, (3) total magnetization of magnetism, and (4) total magnetization. We downloaded these features from the Materials Project database.

- Elastic modulus features: (1) shear modulus, which indicates an object's tendency to shear when acted upon by opposing forces, (2) bulk modulus, which describes volumetric elasticity.

- Raw crystal structures: graph neural network based feature learning.

6.2.2 Machine learning algorithms

Random Forest Random forest (also known as random decision forest) is a machine learning method specifically made for classification and regression that runs on the construction of decision trees during the training process. Advantages of random forest include its ability to reduce over fitting in decision trees which improves accuracy and compatibility with different values. Another useful feature of the random forest model is that it can effectively rank the importance of variables in regression tasks. In the training part, we set the hyper-parameter number of the tree as 200 and the number of estimators as 50.

Support Vector Machines The support vector machine algorithm is another machine learning algorithm for classification, regression, and other tasks that works by creating a hyper plane(s) in a high- or infinite-dimensional space. Support vector machines are very effective in high dimensional spaces, especially when the number of dimensions is greater than the number of samples. This method has been widely used in material science research [142, 1, 91]. In our model, we select rbf kernel, and we set the regularization parameter c = 1.0 and epsilon = 0.2.

Graph Neural Networks Graph neural network [168] is a machine learning model that directly takes graphs (composed of vertexes and edges) as an input. Graph neural networks have wide applications in various domains such as social networks, knowledge graphs, recommender systems, and life science. One of the significant advantages of graph neural networks is their capability to learn or model dependencies (interactions) between nodes in a graph which is highly suitable for modeling interactions between atoms in materials [159].

In this paper, we use five graph neural networks including: SchNet [134], CGCNN [159], MPNN [41], MEGNET [19], and graph attention neural network(GATGNN) [92] for piezoelectric coefficient predictions. These models have recently been evaluated in a benchmark study with comparable performances [35]. In our experiments, we use those models' default parameters.

SchNet SchNet is a deep neural network used to predict molecular energies and atomic properties. This network observes physical laws and achieves rotation and translation invariance. Three interaction blocks describe interactions between atoms. This model adopts continuous-filter convolutional layers as the main building block for neural network architecture, which allows it to model local correlations without requiring the data to lie on the grid.

CGCNN CGCNN (Crystal Graph Convolutional Neural Network) is a type of deep graph neural network used to learn material properties from the relationships of atoms within a crystal which gives in depth representations of crystalline materials. CGCNN is flexible in predicting properties and accurately extracting information of different materials. It has been successfully applied to predict a variety of materials properties such as formation energy, absolute energy, Fermi energy, band gaps, bulk/shear moduli, and Poisson's ratio. Among them, the bulk/shear moduli and Posisson ratio prediction models are trained with only 2041 samples, which is close to the dataset size in our study.

MPNN MPNN (Message Passing Neural Networks) is a neural network that accurately predicts important molecular properties of materials. MPNN is favorable in predicting molecular properties with relatively high accuracy. This model is invariant to graph isomorphism, composed of message functions, vertex update functions, and readout functions, all of which are all differential functions. MPNN generally works on directed graphs with separate channels for incoming and outgoing edges.

MEGNET MEGNET (MatErials Graph Network) is a new kind of graph neural network algorithm for material property prediction that uses two new strategies to address the data scarcity problem. One strategy is to build a single free energy model by incorporating the temperature, press, and entropy as global state inputs. MEGNet models are found to outperform previous ML models in predicting properties and achieve higher accuracy dealing with larger datasets. MEGNET is high-performing in targeting various properties for both molecules and crystals. MEGNet models use learned element embedding that encodes periodic chemical trends, which can be learned from other property models with larger datasets.

GATGNN This graph neural network model comprises multiple graph-attention layers (GAT) and global attention layers. These GAT layers enable this model to efficiently learn complex bonds shared among atoms in each atom's local environments. We aim to train machine learning models to learn the structural features for piezoelectric coefficient predictions using this model.

The graph neural network models used here are adapted from the benchmark study in [35] except that we have added the differentiable group normalization operator [170] into the above five neural network models and skip-connection which is first proposed in ResNet and recently in graph neural networks. Such modifications have allowed us to train graph neural network models to perform better. We use the default parameters for training these models with 250 epochs.

6.2.3 Performance evaluation criteria

We use Mean Absolute Error (MAE) and R-squared (R^2) metrics to evaluate those models' performance. MAE is defined as the average of the absolute difference between the target values and the values predicted by the model.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{6.1}$$

R^2 is defined as the part of the variance of the dependent variable based on the independent variables of our model.

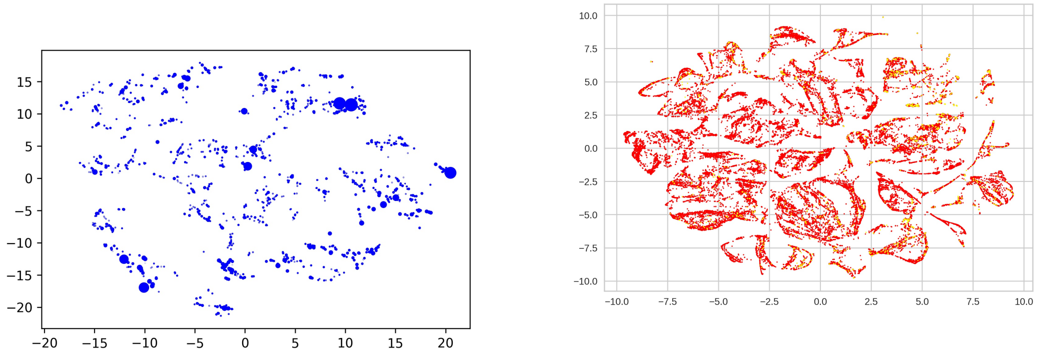
$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} \tag{6.2}$$

In our work, k-Fold Cross-Validation is used to evaluate the performance of trained machine learning models on the limited dataset. Compared to random training test splitting, k-Fold Cross-Validation allows us to obtain less biased, less optimistic, and more stable estimates of model performance. The basic procedure is that the dataset is shuffled randomly and split into k number of groups, take each group in turn as the test dataset, and take the remaining groups as the training dataset. Fit the model on the training set and evaluate it on the test set. The performance of each of the trained models is averaged and reported as the overall cross validation performance. In our evaluations, we used 10-Fold Cross-Validation to ensure fair and stable results.

6.3 Experiments and Discussion about piezoelectric materials

6.3.1 Global distribution of piezoelectric materials

To explore how the piezoelectric materials and their piezoelectric modulus are distributed, we visualize our piezoelectric dataset with 1,705 samples, all using the t-sne algorithm [144]. This algorithm can map high dimension data points into 2D space by preserving the neighborhood relationships. Each sample is first featurized by calculating its magpie features, which are then fed to the t-sne algorithm that maps it into 2D space. Two distribution maps are generated using this procedure. We only include the 1,705 materials samples with piezoelectric modulus values. The map is shown in Figure 6.5a with blue colors and their size indicates a high-piezoelectic effect. We find that only a small number of materials have higher piezoelectric properties.



(a) T-sne map of piezoelectric moduli.

(b) Distribution of 1,705 known piezoelectric materials vs other ternary materials.

Figure 6.5 Distribution of piezoelectric moduli and materials in composition space. (a) the sizes of the points are proportionate to their piezoelectric moduli. There are a few exceptionally high values located in 4 clusters. (b) Ternary materials (red points) are distributed in composition clusters. Known piezoelectric materials (yellow points) are similarly grouped in clusters with two groups with a significant number of known materials.

To further show how the know piezoelectric materials are distributed among all ternary crystal materials, we include all ternary materials samples (with the 1,705 piezoelectric materials included). The distribution map is shown in Figure 6.5b. Red dots are the 59,211 ternary materials samples and blue dots are the 1,705 piezoelectric materials. It can be observed that the overlap area between red and blue dots is tiny indicating that piezoelectric materials are different from general materials distribution.

6.3.2 Performance of machine learning models

The performances of our Random forest models and SVM models with different types of features (magpie features, oxidation states features, structural features, feature transformations, energy and magnetism features, and elastic modulus) are reported in Table 6.1.

Table 6.1 Machine learning performances on piezoelectric constant prediction (10-fold cross-validation).

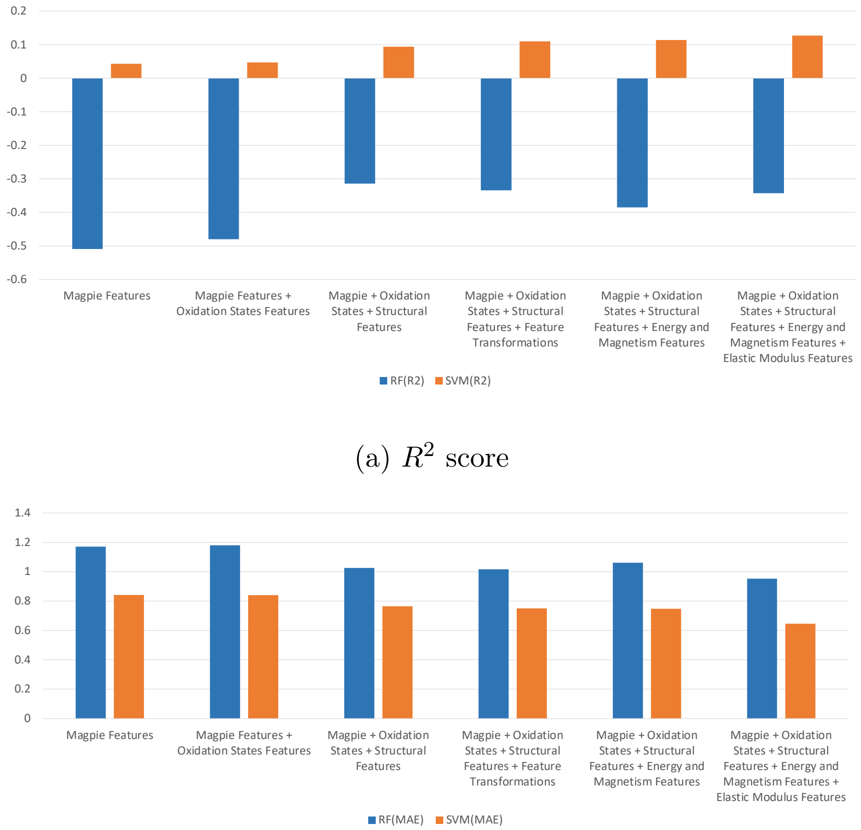
	RF (R^2)	**RF (MAE)**	**SVM (R^2)**	**SVM (MAE)**
Magpie Features	-0.509	1.17	0.043	0.841
Magpie Features + Oxidation States Features	-0.480	1.18	0.047	0.840
Magpie + Oxidation States + Structural Features	-0.314	1.026	0.094	0.764
Magpie + Oxidation States + Structural Features+ Feature Transformations	-0.334	1.017	0.110	0.750
Magpie + Oxidation States + Structural Features + Energy and Magnetism Features	-0.385	1.061	0.114	0.748
Magpie + Oxidation States + Structural Features + Energy and Magnetism Features+ Elastic Modulus Features	-0.343	0.953	0.127	0.646

Initially, the random forest model is only based on the magpie features, with its MAE and R^2 scores being 1.17 C/m^2 and -0.509. After adding the oxidation state features, structural features, energy and magnetism features, and elastic features, the MAE decreases 18.5% to 0.953, and the R^2 increases 32.6% to -0.343.

The SVM's R^2 performance score starts at 0.043 with magpie features and then increases 117% (more than doubled) with oxidation states and structural features. It shows the importance of structural information for piezoelectric modulus prediction. Unlike the random forest model, the SVM model performs better with energy and magnetism features (increased 22%) showing that SVM is more stable than the random forest models. The MAEs for both random forest and SVM show moderate change with the addition of the diverse features showing that MAE has more potential to improve (e.g. undiscovered features). As elastic moduli are important quantities that measure the resistance to being deformed elastically, we also add the bulk and shear moduli to train the random forest and the SVM models. The MAE and R^2 of the random forest model improve 10.17% and 10.9%, respectively; while the SVM model's MAE and R^2 improve 13.63% and 11.4% accordingly.

For better comparison, we draw two bar charts in Figure 6.6 to show the performance changes with different combinations of features. Figure 6.6a shows the R^2 performance where the blue bars represent the R^2 scores of the random forest models and the red bars represent those of the SVM models with different feature sets. From the figure, we find that the SVM's R^2 scores remain positive throughout the six different feature sets. In contrast, the random forest's R^2 scores remain negative, indicating that SVM models perform better than random forests in terms of the R^2 performance measure. As more features are added throughout the feature sets, SVM consistently increases its accuracy. Random forest, instead, shows different patterns. Starting with the magpie features, adding oxidation states and structural features increases the accuracy whereas adding feature transformations and energy and magnetism features lowers the accuracy score. In Figure 6.6b, the blue bars represent the MAEs of the random forest model and the red bars represent the MAEs of the SVM models for piezoelectric prediction. The figure shows that the MAE for the random forest model (blue bars) is significantly greater than the MAE for SVM models (red bars) indicating random forest performs worse than SVM. Another pattern we find is that the error for both random forest and SVM decreases as more features are added.

To compare the predicted piezoelectric coefficients with the actual piezoelectric coefficients, we draw two scatter plots in Figure 6.7, the x-axis indicates the true value, and the y-axis shows the predicted value. In addition, the red color line is the regression line. The results of the random forest model with five types of features are shown in Figure 6.7a and results of the SVM model are shown in Figure 6.7b. The more blue dots close to the red line, the better performance the model achieves. For the RF model, more blue dots are far away from the red regression line than SVM,

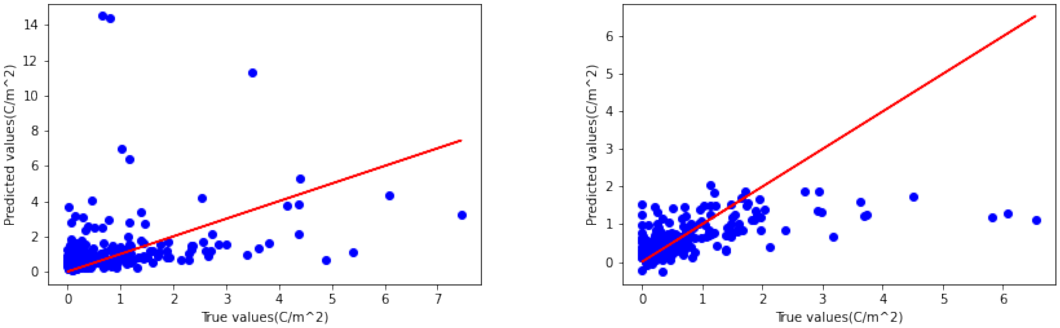


(a) R^2 score

(b) MAE error

Figure 6.6 Performance comparison of random forest and SVM in terms of R^2 scores and MAE errors.

which indicates that the performance of SVM is better. Besides, several dots have higher predicted piezoelectric coefficients (around 14) while their true values are small (around 1). For the SVM model, the overall fitting within 2 C/m^2 is good, consistent with the fact that the range of most piezoelectric coefficients is between 0 to 2 C/m^2 (refer to Figure 6.6).



(a) RF prediction with five types of features (b) SVM prediction with five types of features

Figure 6.7 Scatter plots of predicted piezoelectric coefficients by random forest model and SVM model

6.3.3 Performance of graph neural networks

Table 6.2 shows the five-fold cross validation results of piezoelectric modulus prediction by graph neural networks. The best graph neural network model is CGCNN with a MAE of 0.97439 C/m^2 and the worst model is SchNet with a MAE of 1.34294 C/m^2. Compared to random forest and SVM in Table 6.1, every graph neural network underperformed the SVM model. SchNet, CGCNN, MPNN, and MEGNET all performed slightly better than random forest while GATGNN performed somewhat worse.

Table 6.2 Performance of graph neural network for piezoelectric coefficient prediction. The MAE errors are much larger than those of SVM models.

Model	SchNet	CGCNN	MPNN	MEGNET	GATGNN
MAE (C/m^2)	1.34294	0.97439	1.04362	0.98129	1.09634

6.3.4 Feature analysis

To obtain physical insights from our machine learning models and the dataset, we analyze the distribution of the piezoelectric materials in terms of the crystal systems and conducted the feature importance analysis to identify physical factors that affect materials' piezoelectric moduli.

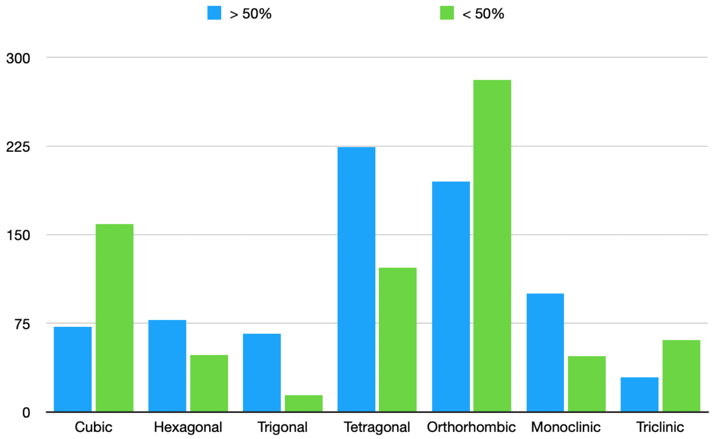


Figure 6.8 Distribution of the number of materials in different crystal systems with high (>50% percentile) and low(<50% percentile) piezoelectric modulus. It is found that materials with high piezoelectric modulus tend to belong to crystal systems with intermediate symmetry (tetragonal, orthorhombic, monoclinic).

From Figure 6.8, we can see that the distribution of known piezoelectric materials is highly unbalanced among the seven different crystal systems. The orthorhombic crystal system has the highest number of known piezoelectric materials, followed by tetragonal and cubic systems. It seems that the structure property has a profound influence on the piezoelectric effect. We suspect that the reason that there are a smaller number of monoclinic and triclinic piezoelectric materials exist is the low symmetry of the crystal structure. We also observed that the high symmetry crystal systems also have a smaller number of known piezoelectric materials. We find that the cubic crystal system has 231 piezoelectric materials despite its high symmetry, which may be since cubic materials are from the most well-studied material family. Based on this analysis, we believe structure features are essential and significant in building high-performance models for piezoelectric modulus.

To further understand how different features contribute to the final prediction performance, we trained the random forest model which provides a built-in ranking of the feature importance. The top 30 features are shown in Figure 6.9.

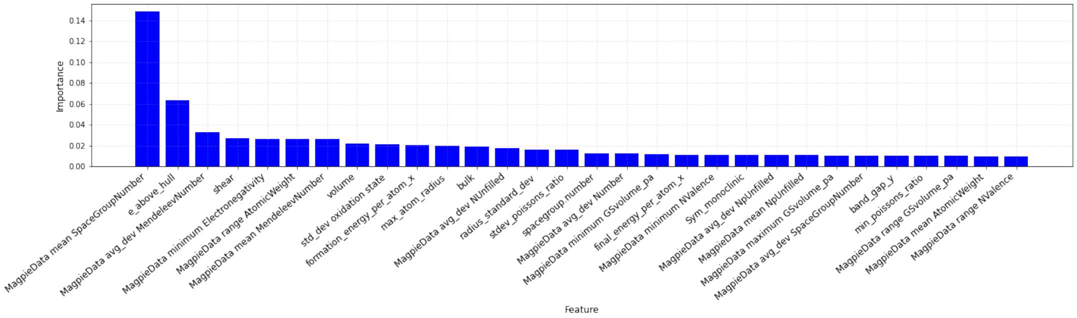


Figure 6.9 Ranking of top 30 features in the RF model for piezoelectric modulus prediction. It brings physical insights by showing the key physical factors that affect the piezoelectric effect.

The most important feature is the mean SpaceGroupNumber followed by energy above hull, average MendeleevNumber, shear and minimum Electronegativity. Since element space group number represents the symmetry trends of crystal materials, it can explain why this feature is important. In terms of shear and bulk, both of them are elastic modulus, shear indicates an object's tendency to shear when acted upon by opposing forces and bulk describes volumetric elasticity. This is reasonable that they are important to the piezoelectric coefficient.

We also found that Poissons' ratio is ranked 15th, reflecting the degree of volume expansion perpendicular to the applied force closely related to the piezoelectric effect. It's interesting that our random forest model can identify Poisson's ratio as one of the top features of the piezoelectric modulus prediction model, which is also confirmed in an experimental study [113].

Next, the space group number of the crystal material ranked 16th, consistent with our analysis in Figure 6.8 that crystal structure plays a crucial role in the distribution of piezoelectric modulus. Another critical feature is Nvalence, which relates to how easily the atoms lose electrons. It can explain why valence related features are ranked high in this prediction model.

Lastly, we found that Nd unfilled features show several times. Those features represent the capability of the atom to acquire electrons so that they may affect the piezoelectric effect.

6.3.5 Predicted potential piezoelectric materials

To discover new potential piezoelectric materials, we applied our trained SVM model with six features to predict the piezoelectric properties of 12,680 materials from the Materials Project database with elastic modulus values (bulk and shear). We sort the predicted piezoelectric coefficients and find 12,650 materials' predicted values are positive, with 1,498 of them are larger than 1 C/m^2 and 12 materials with predicted piezoelectric coefficients greater than 2 C/m^2. We listed the top 20 predicted piezoelectric materials in Table 6.3. We also summarize a total of 150 materials with high predicted piezoelectric coefficients in supplementary file Table S1, Table S2 and Table S3, which are classified according to the number of elements in the materials.

Table 6.3 Top 20 predicted potential piezoelectric materials (unit: C/m^2)

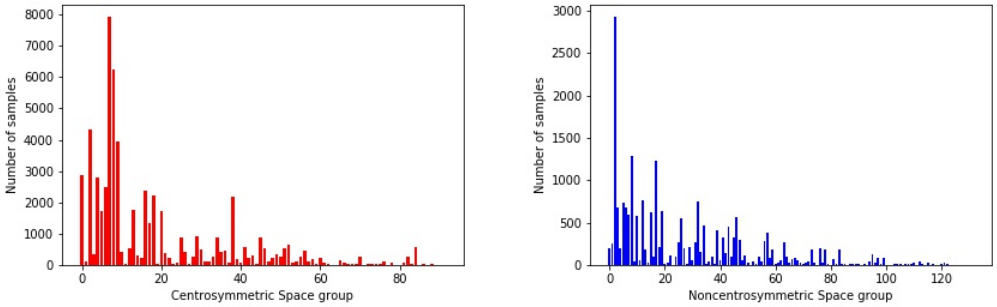
Material ID	Formula	Predicted Value	Material ID	Formula	Predicted Value
mp-578601	NaNbO2	2.448	mp-756683	HfBiO4	2.015
mp-10426	Nb2O5	2.341	mp-7017	NaNbN2	2.007
mp-760401	Nb3O7F	2.184	mp-549490	KNb4O5F	1.978
mp-754698	NbO2	2.153	mp-552588	LiNbO3	1.966
mp-753459	Nb3O7F	2.140	mp-7240	NaRuO2	1.958
mp-1595	Nb2O5	2.114	mp-754375	NaTi2O3	1.949
mp-557680	NbAgO3	2.085	mp-505517	BaNb4O6	1.938
mp-755690	NbO2	2.059	mp-644497	BaTiO3	1.935
mp-753380	La(BiO2)2	2.058	mp-28254	LiRuO2	1.932
mp-2533	NbO2	2.024	mp-1029267	CaZrN2	1.928

6.4 Experiments and Discussion about noncentrosymmetric materials

Herein, we describe the datasets, the evaluation criteria, and the experimental results. We analyze and compare the prediction performance of RF and DNN models. Besides, we discuss the application of our model to screening new hypothetical noncentrosymmetric materials. Our experiments on classifying noncentrosymmetry from composition include three parts: cross-validation experiments, holdout experiments on Borates, and screening a two million hypothetical materials.

6.4.1 Datasets

Crystal structures with different space groups have different centrosymmetric tendencies. It is known that there are 138 noncentrosymmetric space groups and 92 centrosymmetric space groups, the detailed space group IDs and names and their centrosymmetric property are summarized in Table 6.4.



(a) Centrosymmetric space groups (b) Non-centrosymmetric space groups

Figure 6.10 Sample distribution of noncentrosymmetric and centrosymmetric space groups in MPF dataset

We first downloaded the composition formulas of 97,217 crystal materials from the Materials Project database. We then remove those compositions belonging to multiple space groups with conflicting centrosymmetric tendencies. In total, we collecte 82,506 material compositions and assign the noncentrosymmetric property labels according to their corresponding space group. The dataset is called **MPF**, which have 60,587 positive (noncentrosymmetric) samples and 21,919 negative (centrosymmetric) samples, as shown in Table 6.5. The distribution of noncentrosymmetric and centrosymmetric space groups in MPF dataset are shown in Figure 6.10. We find that the distribution of samples over different space groups are not well balanced.

In order to evaluate the extrapolation prediction performance of our machine learning prediction model of noncentrosymmetry, we select all the 315 borate compounds

Table 6.4 Space groups with noncentrosymmetric and centrosymmetric structures

	group IDs	**group names**
centrosymmetric	2, 10-15, 47-74, 83-88, 123-142, 147-148, 162-167, 175-176, 191-194, 200-206, 221-230	$P\bar{1}$, P2/m, $P2_1/m$, C2/m, P2/c, $P2_1/c$, C2/c, Pmmm, Pnnn, Pccm, Pban, Pmma, Pnna, Pmna, Pcca, Pbam, Pccn, Pbcm, Pnnm, Pmmn, Pbcn, Pbca, Pnma, Cmcm, Cmca, Cmmm, Cccm, Cmma, Ccca, Fmmm, Fddd, Immm, Ibam, Ibca, Imma, P4/m, $P4_2/m$, P4/n, $P4_2/n$, I4/m, $I4_1/a$, P4/mmm, P4/mcc, P4/nbm, P4/nnc, P4/mbm, P4/mnc, P4/nmm, P4/ncc, $P4_2/mmc$, $P4_2/mcm$, $P4_2/nbc$, $P4_2/nnm$, $P4_2/mbc$, $P4_2/mnm$, $P4_2/nmc$, $P4_2/ncm$, I4/mmm, I4/mcm, $I4_1/amd$, $I4_1/acd$, $P\bar{3}$, $R\bar{3}$, $P\bar{3}1m$, $P\bar{3}1c$, $P\bar{3}m1$, $P\bar{3}c1$, $R\bar{3}m$, $R\bar{3}c$, P6/m, $P6_3/m$, P6/mmm, P6/mcc, $P6_3/mcm$, $P6_3/mmc$, $Pm\bar{3}$, $Pn\bar{3}$, $Fm\bar{3}$, $Fd\bar{3}$, $Im\bar{3}$, $Pa\bar{3}$, $Ia\bar{3}$, $Pm\bar{3}m$, $Pn\bar{3}n$, $Pm\bar{3}n$, $Pn\bar{3}m$, $Fm\bar{3}m$, $Fm\bar{3}c$, $Fd\bar{3}m$, $Fd\bar{3}c$, $Im\bar{3}m$, $Ia\bar{3}d$
noncentrosymmetric	1, 3-9, 16-46, 75-82, 89-122, 143-146, 149-161, 168-174, 177-190, 195-199, 207-220	P1, P2, $P2_1$, C2, Pm, Pc, Cm, Cc, P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, $C222_1$, C222, F222, I222, $I2_12_12_1$, Pmm2, $Pmc2_1$, Pcc2, Pma2, $Pca2_1$, Pnc2, $Pmn2_1$, Pba2, $Pna2_1$, Pnn2, Cmm2, $Cmc2_1$, Ccc2, Amm2, Aem2, Ama2, Aea2, Fmm2, Fdd2, Imm2, Iba2, Ima2, P4, $P4_1$, $P4_2$, $P4_3$, I4, $I4_1$, $P\bar{4}$, $I\bar{4}$, P422, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_322$, $P4_32_12$, I422, $I4_122$, P4mm, P4bm, $P4_2cm$, $P4_2nm$, P4cc, P4nc, $P4_2mc$, $P4_2bc$, I4mm, I4cm, $I4_1md$, $I4_1cd$, $P\bar{4}2m$, $P\bar{4}2c$, $P\bar{4}2_1m$, $P\bar{4}2_1c$, $P\bar{4}m2$, $P\bar{4}c2$, $P\bar{4}b2$, $P\bar{4}n2$, $I\bar{4}m2$, $I\bar{4}c2$, $I\bar{4}2m$, $I\bar{4}2d$, P3, $P3_1$, $P3_2$, R3, P312, P321, $P3_112$, $P3_121$, $P3_212$, $P3_221$, R32, P3m1, P31m, P3c1, P31c, R3m, R3c, P6, $P6_1$, $P6_5$, $P6_2$, $P6_4$, $P6_3$, $P\bar{6}$, P622, $P6_122$, $P6_522$, $P6_222$, $P6_422$, $P6_322$, P6mm, P6cc, $P6_3cm$, $P6_3mc$, $P\bar{6}m2$, $P\bar{6}c2$, $P\bar{6}2m$, $P\bar{6}2c$, P23, F23, I23, $P2_13$, $I2_13$, P432, $P4_232$, F432, $F4_132$, I432, $P4_332$, $P4_132$, $I4_132$, $P\bar{4}3m$, $F\bar{4}3m$, $I\bar{4}3m$, $P\bar{4}3n$, $F\bar{4}3c$, $I\bar{4}3d$

from MPF dataset and assign them as the hold-out test dataset Borates315. Borates contain boron (B) element and oxygen (O) element, which are a ubiquitous family of flame retardants found as boric acid and as a variety of salts. Previous research found that compared to other material family, borates tend to have a higher percentage of nonlinear proprieties, which makes it a good hold-out test set [16].

We further find that most borate materials include 3 elements. It is interesting to see if ML models trained with 3-element training samples can achieve better prediction performance. We select all 3-element materials from the MPF dataset and assigned them to the **MP3** dataset, which includes 30,762 centrosymmetric materials and 8,964 noncentrosymmetric materials as shown in Table 6.5. The motivation is to check if our classification models trained with MP3 dataset can achieve better performance when testing on the hold-out borates dataset.

Table 6.5 Dataset

	#symmetry	**#non symmetry**	**#total**
MPF	63,376	19,130	82,506
MP3	30,762	8,964	39,726
Borates315	250	65	315

6.4.2 Prediction performance

To evaluate how our machine learning models can predict whether a crystal material's structure is noncentrosymmetry or not, we used two evaluation approaches: one is cross-validation over the MPF dataset and the other is the hold-out evaluation trained with non-borates datasets MPF and MP3 and tested on the Borates315 dataset. This hold-out test is especially important as the cross-validation performance can usually be over-estimated due to the redundancy of the training samples in most of the large-scale datasets such as the Materials Projects and the OQMD [160].

10-fold cross-validation performance

We set the maximum tree depth to be 20 and the number of decision trees as 200. This was later expanded to include the minimum number of samples per leaf node, the minimum number of samples required to split a node, and the maximum number of leaf nodes. With these 5 settings tuned per featurizer iteration, we then train the final prediciton RF models and make prediction, and caculate the performance scores. To further verify the performance of our RF-based models, we compare it with those of the DNN-based models. Table 6.6 shows the performances we achieved on two datasets using four evaluation criteria.

Table 6.6 Ten-fold cross-validation performance of ML models for noncentrosymmetry prediction

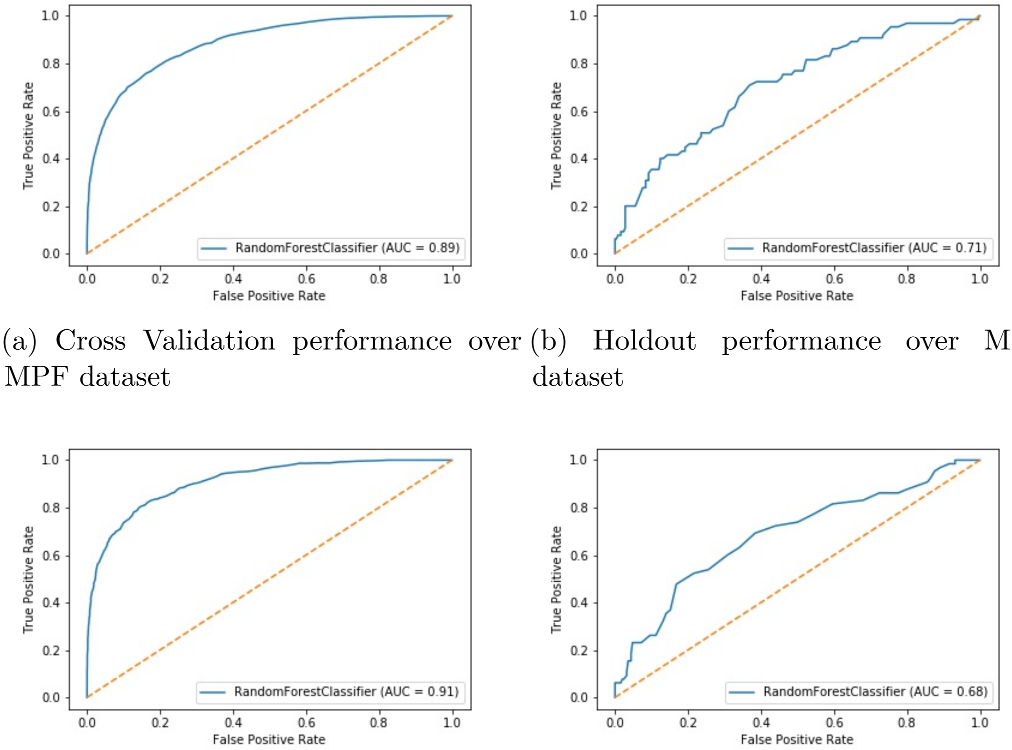
Model	Dataset	Precision	Recall	Accuracy	F1 score
RF-based	MPF	0.834	0.754	0.848	0.781
RF-based	MP3	**0.845**	0.755	**0.869**	**0.786**
DNN-based	MPF	0.773	0.769	0.785	0.771
DNN-based	MP3	0.784	**0.780**	0.792	0.782

Firstly, we found that the precision and accuracy of the RF model are significantly better in comparison with DNN models: the 10 fold cross-validation accuracy of RF model on the MPF dataset is 0.848 compared to 0.785, which indicates 7.89% improvement. The F1 score of RF model is 0.781 compared to 0.771 of DNN. Although DNN achieves better Recall score, the F1-score of RF is higher than DNN's. This validates the effectiveness of our RF-based model for predicting the noncentrosymmetric property for a given material. This is consistent with a recent evaluation of different ML methods for materials property prediction [129].

Secondly, comparing the results of the same RF and DNN model on the MPF dataset and the MP3 dataset, we found that each model achieved better prediction performance for the MP3 dataset. Particularly, the precision, accuracy and F1 score of the RF classifier increase to 0.845, 0.869 and 0.786, respectively.

Hold out experiment results

To explore the effectiveness of our model for extrapolative prediction of noncentrosymmetry where the test samples may not have the same distribution with the training set, we conducted a hold-out test over the Borates315 dataset.The training dataset is generated by filtering out all the samples of the Borates315 dataset from the MPF dataset and keeping the remaining ones, which includes 82,191 samples. Similarly, we also conduct a hold-out test for the MP3 dataset for which the training set is generated by removing all borates in the MP3 dataset. The number of samples of the no-borates 3-element training set is 39411. Their ROC curves and AUC scores are shown in Figure 6.11.



(a) Cross Validation performance over MPF dataset

(b) Holdout performance over MPF dataset

(c) Cross Validation performance over MP3 dataset

(d) Holdout performance over MP3 dataset

Figure 6.11 ROC curves for cross-validation and hold-out experiments for the RF prediction models trained with the whole dataset and the 3-element dataset.

In Figure 6.11, each dotted yellow line corresponds to the ROC curve of a random predictor with AUC value of 0.5. Each blue curve represents the ROC curve of the classifier. As is well known the higher value of AUC, the better performance of the classifier. Among the four sub-figures, figure (c) shows the best result, with AUC reaching 0.91. Furthermore, comparing (a) (c) with (b) (d), we can find AUC scores of cross-validation experiments are higher than those of hold-out experiments over the same two datasets, which suggests the over-estimation of model performance due to dataset sample redundancy. Meanwhile, although the performance of hold out experiments is not as good as cross validation experiments, it only uses the non-borate materials as the training data for predicting the 315 borate materials, which interprets the 0.71 and 0.68 AUC are acceptable since this is extrapolation prediction performance.

6.4.3 Predicting new noncentrosymmetric materials

To identify interesting hypothetical new NLO noncentrosymmetric materials, we applied our RF-based noncentrosymmetric materials prediction model to screen the two million hypothetical materials generated by our Generative Adversarial Network (GAN) based new materials composition generator [28]. After predicting the probability of each candidate belonging to noncentrosymmetric materials, we sort them by the probability scores and report top 20 hypothetical noncentrosymmetric materials with 2, 3 and 4 elements here in Table 6.7. Furthermore, as we mentioned above that most borate materials are NLO materials. So we also reported top 20 borate materials with highest proability here. Please note that materials containing lanthanide and actinide elements have been filtered in these results because they are very rare.

Table 6.7 Predicted hypothetical noncentrosymmetric materials with 2, 3, and 4 elements and predicted noncentrosymmetric borates (only top 20 are listed here)

2 element	score	3 element	score	4 element	score	Borate	Score
Li4Ge	0.935	AlCuSe3	0.960	LaCeNdS4	0.975	CB2O6	0.840
Cu2S3	0.875	Cu2AsS3	0.955	LaCeNdSe4	0.965	N2B4O7	0.715
NO5	0.835	Cu3As2S4	0.945	CeNdEuS4	0.960	CB4O6	0.700
Li4Pb	0.830	Y2CeO5	0.945	CuZnInS3	0.955	S3B2O8	0.670
Li4Sn	0.800	CeTb2S4	0.935	AlCuZnTe4	0.925	CB2O4	0.665
Cl3S	0.745	DyErC3	0.930	MnNiAgSn	0.925	NCB4O6	0.665
SbC	0.740	MnDy2S4	0.925	AlCuInSe2	0.915	CoIB4O6	0.660
Pd2S	0.735	LaSm2S4	0.920	MnCoRuSn	0.915	EuB4O6	0.655
AsC	0.720	ZnGaSe2	0.920	LaNdUTe4	0.915	ZnSnO6B4	0.650
SeO6	0.715	AlCu2Te3	0.915	Cu2ZnInS6	0.900	As2B2O7	0.635
Ni3Ge2	0.715	AlCu2S4	0.910	NiCuSnSe3	0.895	PB2O6	0.630
Cl5S	0.710	CoCd2S3	0.905	MnCoAgSn	0.895	ZnB2O4	0.625
Zr2S3	0.695	NbSnIr	0.900	TiCoRhSn	0.880	SB2O6	0.620
S2O5	0.690	NbWTe4	0.900	MnFeSbO6	0.875	MnZnLaEuO6B2	0.610
LiOs	0.690	VSnAu	0.900	MnCu2AgS4	0.875	Zn3SB2O6	0.600
NH2	0.690	CrCu2S3	0.895	FeLaPbO6	0.875	Sr2TaB2O6	0.600
CrI	0.685	SnTaOs	0.890	V2Ni2RuSn2	0.875	PbB4O6	0.595
F3N	0.680	NdDySi3	0.885	MnFeBi2O6	0.875	AlB2O4	0.585
Cl6S	0.675	Dy2GeS4	0.885	TiCoBi2O6	0.870	NbRuCl2B4O6	0.585
S2O3	0.660	Mg6MnSn	0.885	SrLaNdS4	0.865	C3B4O6	0.580

As shown in Table 6.7, the probability score range of the top 20 2-element materials, 3-element materials, 4-element materials and borate materials are 0.935 to 0.660, 0.960 to 0.885, 0.975 to 0.865 and 0.885 to 0.670, respectively. It is clear that the predicted noncentrosymmetic probabilities of 3 element materials are higher than those of 2-element materials and 4-element materials. As those material are generated and hypothetical, we can only give the predicted noncentrosymmetry scores, which may guide experimental work to verify them in future research, which may further validate the effectiveness and the predictive capability of our models.

6.5 Conclusion

In this study, we developed and applied two machine learning algorithms (Random forest and support vector machines) and five graph neural network models for predicting the piezoelectric modulus using various composition and structural features. Extensive evaluations have been done over the dataset composed of 1,705 samples downloaded from the Materials Project data repository. Our experiment results show that the composition only descriptors/features alone do not help build a good piezoelectric modulus prediction model, which proves that the piezoelectric effect is strongly affected by their crystal lattice structures as well as other electronic and magnetic properties, and elastic moduli. By adding the structural features, magnetic features and elastic features, we have been able to increase the regression R^2 score of the SVM model from 0.043 to 0.127. Our study also shows that the Random forest models in general perform much worse than the SVM models. We also explored five popular graph neural network models for the piezoelectric modulus prediction. Our results show that it is much more challenging and all these models have much lower performance compared to those of other property prediction [35], which may be due to the limited dataset size and the sophisticated relationships between crystal structures and the piezoelectric effect. However, our study does show that the graph neural network models can achieve performance compared to RF and SVM in terms of the MAE errors. Compared to other properties such as formation energy, shear/bulk moduli, and band gaps, we find that piezoelectric modulus is notoriously much more challenging to predict accurately. Much more detailed feature engineering or deep neural network based representation learning are needed to further improve the prediction performance of piezoelectric modulus to enable large-scale screening and design of novel piezoelectric materials and uncover the complex relationships of the crystal structure and the piezoelectric effect. More importantly, we utilize the trained SVM model to predict the piezoelectric coefficients for 12,680 materials from the materials project database, and find 1,498 materials whose predicted piezoelectric coefficients are larger than 1 C/m^2. We report the top 20 materials with all related information (material ids in the material project dataset, formulas and the predicted piezoelectric coefficient values), inspiring experimental material scientists to verify some of these new piezoelectric material candidates.

Computational prediction of noncentrosymmetry of a given composition can be used for fast screening new nonlinear optical materials. Here we developed and evaluated two machine learning models including a Random Forest Classifier and a neural network model for computational prediction of materials noncentrosymmetry given only their composition information. By using the Magpie composition features, our best prediction model based on Random forest can achieve an accuracy of 84.8% when evaluated using 10-fold cross-validation over the Material Projects database. Further experiments showed that when the prediction model is trained only on 3-element samples, it can achieve even higher performance for the test set, which is made of mostly 3-element materials. A feature importance calculation shows the top six contribution factors for predicting noncentrosymmetry, many of which are related to the distribution of valence electrons. which is consistent with current physicochemical principles. Our developed model can be applied to discovering novel nonlinear materials as we conduct large-scale screening over two million hypothetical materials.

Chapter 7

Conclusion

In this dissertation, we present our research in the prediction of crystal structures, and encompasses the design of a motif extraction method and the analysis of these motifs. Additionally, we explore the discovery of new 2D materials, as well as investigate the prediction of piezoelectric coefficients and noncentrosymmetric properties. Through this work, we explore several main tasks in material informatics and pave the way for crystal structure prediction and material property prediction.

In the first topic, we propose the DeltaCrystal model for crystal structure prediction which only relies on material composition. In the first step, we predict the distance matrix of atom pairs for the given material composition based on the feature matrix and trained deep residual neural network. In the second step, we generate crystal structure by the atomic coordinate reconstruction algorithm DMCrystal, and then use M3GNET to relax those structures. Our experiments show DeltaCrystal can reconstruct the crystal structures for a large number of materials, even for complex compounds. DeltaCrystal can be a strong new kind of deep knowledge-guided CSP for large-scale prediction of crystal structures.

In the second work, we present a motif extraction method, we use this method to extract a total of 18,534 motifs from 122,500 material structures. The detection and extraction of motifs within any crystal structure could have a major impact on the materials discovery field. Currently, CSP models either base their prediction on a few small molecules or on individual atoms due to the computing power needed for deep learning. If CSP models based their prediction on pre-determined motifs instead, it would take less computing power to create larger crystal structures, leading to even more materials being predicted.

In the third topic, we design a model for finding hypothetical new 2D materials, which mainly includes four parts: (1) generate chemically valid materials formulas based on the deep learning generative model, (2) train a composition-based random forest 2D materials classifier to predict potential 2D materials, (3) develop a template based element substitution method to predict those predicted materials' structures and (4) finally, we use DFT verification to confirm their structural stability. Currently, we generate 1485 hypothetical 2D material compositions with probability scores greater 95% and verify that 92 materials have stable formation energies by DFT calculation. These novel hypothetical materials can be utilized to guide the screening of 2D materials for material scientists.

In the last topic, we explore several different algorithms to predict piezoelectric coefficients, including Random forest support vector machines and five graph neural network models using various composition and structural features. Due to the limited data, machine learning methods, especially the random forest model, achieve better predictive results than deep learning methods, which demonstrates the advantages of machine learning in dealing with small datasets. Similarly, we also use the random forest model to predict materials' noncentrosymmetric and achieve good performance.

Bibliography

[1] Abbas M Abd and Suhad M Abd. “Modelling the strength of lightweight foamed concrete using support vector machine (SVM)”. In: *Case studies in construction materials* 6 (2017), pp. 8–15.

[2] Hossin A. Abdeldayem and Donald O. Frazier, eds. *Nonlinear Optics and Applications*. Research Signpost, 2007.

[3] Badri Adhikari, Jie Hou, and Jianlin Cheng. “DNCON2: improved protein contact prediction using two-level deep convolutional neural networks”. In: *Bioinformatics* 34.9 (2018), pp. 1466–1472.

[4] Ankit Agrawal and Alok Choudhary. “Deep materials informatics: Applications of deep learning in materials science”. In: *MRS Communications* 9.3 (2019), pp. 779–792.

[5] Ethan C Ahn. “2D Materials for Spintronic Devices”. In: *npj 2D Materials and Applications* 4.1 (2020), pp. 1–14.

[6] Deji Akinwande et al. “Graphene and Two-dimensional Materials for Silicon Technology”. In: *Nature* 573.7775 (2019), pp. 507–518.

[7] Mohammed AlQuraishi. “AlphaFold at CASP13”. In: *Bioinformatics* 35.22 (2019), pp. 4862–4865.

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[9] Sanjeev Arora, Wei Hu, and Pravesh K Kothari. “An analysis of the t-sne algorithm for data visualization”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 1455–1462.

[10] Huta R Banjade et al. “Structure motif–centric learning framework for inorganic crystalline systems”. In: *Science advances* 7.17 (2021), eabf1754.

[11] Stefano Baroni et al. “Phonons and Related Crystal Properties from Density-functional Perturbation Theory”. In: *Rev. Mod. Phys.* 73 (2 July 2001), pp. 515–562.

[12] Christopher J Bartel et al. “A critical examination of compound stability predictions from machine-learned formation energies”. In: *npj Computational Materials* 6.1 (2020), pp. 1–11.

[13] Alec Belsky et al. “New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design”. In: *Acta Crystallographica Section B: Structural Science* 58.3 (2002), pp. 364–369.

[14] P. E. Blöchl. “Projector Augmented-Wave Method”. In: *Phys. Rev. B* 50 (24 Dec. 1994), pp. 17953–17979. DOI: 10.1103/PhysRevB.50.17953. URL: https://link.aps.org/doi/10.1103/PhysRevB.50.17953.

[15] John Bradshaw et al. “A model to search for synthesizable molecules”. In: *Advances in Neural Information Processing Systems.* 2019, pp. 7937–7949.

[16] Rimma Bubnova et al. “Borates—crystal structures of prospective nonlinear optical materials: High anisotropy of the thermal expansion caused by anharmonic atomic vibrations”. In: *Crystals* 7.3 (2017), p. 93.

[17] MA Butler and DS Ginley. “Prediction of Flatband Potentials at Semiconductor-electrolyte Interfaces from Atomic Electronegativities”. In: *Journal of the Electrochemical Society* 125.2 (1978), p. 228.

[18] Chi Chen and Shyue Ping Ong. “A universal graph deep learning interatomic potential for the periodic table”. In: *arXiv preprint arXiv:2202.02450* (2022).

[19] Chi Chen et al. “Graph networks as a universal machine learning framework for molecules and crystals”. In: *Chemistry of Materials* 31.9 (2019), pp. 3564–3572.

[20] Guang Chen et al. “Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges”. In: *Polymers* 12.1 (2020), p. 163.

[21] Yingshi Chen et al. “Smart Inverse Design of Graphene-based Photonic Metamaterials by an Adaptive Artificial Neural Network”. In: *Nanoscale* 11.19 (2019), pp. 9749–9755.

[22] Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. “Crystal structure prediction by combining graph network and optimization algorithm”. In: *Nature communications* 13.1 (2022), pp. 1–8.

[23] Stefan Chmiela et al. "Towards exact molecular dynamics simulations with machine-learned force fields". In: *Nature communications* 9.1 (2018), pp. 1–10.

[24] Kamal Choudhary et al. "High-throughput Identification and Characterization of Two-dimensional Materials Using Density Functional Theory". In: *Scientific Reports* 7.1 (2017), pp. 1–16.

[25] Juan-Pablo Correa-Baena et al. "Accelerating materials development via automation, machine learning, and high-performance computing". In: *Joule* 2.8 (2018), pp. 1410–1420.

[26] Yabo Dan et al. "Computational Prediction of Critical Temperatures of Superconductors Based on Convolutional Gradient Boosting Decision Trees". In: *IEEE Access* 8 (2020), pp. 57868–57878.

[27] Yabo Dan et al. "Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials". In: *npj Computational Materials* 6.1 (2020), pp. 1–7.

[28] Yabo Dan et al. "Generative adversarial networks (GAN) based efficient sampling of chemical space for inverse design of inorganic materials". In: *arXiv preprint arXiv:1911.05020* (2019).

[29] Daniel W Davies et al. "SMACT: Semiconducting Materials by Analogy and Chemical Theory". In: *Journal of Open Source Software* 4.38 (2019), p. 1361.

[30] Ljiljana Despalatović, Tanja Vojković, and Damir Vukicević. "Community structure in networks: Girvan-Newman algorithm improvement". In: *2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2014, pp. 997–1002.

[31] Hien V Do et al. "An isogeometric analysis to identify the full flexoelectric complex material properties based on electrical impedance curve". In: *Computers & Structures* 214 (2019), pp. 1–14.

[32] Alexander Dunn et al. "Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm". In: *npj Computational Materials* 6.1 (2020), pp. 1–10.

[33] Isaac Arnold Emerson and Arumugam Amala. "Protein contact maps: A binary depiction of protein 3D structures". In: *Physica A: Statistical Mechanics and its Applications* 465 (2017), pp. 782–791.

[34] Marco Fronzi et al. “Impressive Computational Acceleration by Using Machine Learning for 2-dimensional Super-lubricant Materials Discovery”. In: *arXiv preprint arXiv:1911.11559* (2019).

[35] Victor Fung et al. “Benchmarking graph neural networks for materials chemistry”. In: *npj Computational Materials* 7.1 (2021), pp. 1–8.

[36] J. Furthmüller G. Kresse. “Efficiency of ab initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set”. In: *Comput. Mater. Sci.* 6 (July 1996), pp. 15–50.

[37] Kevin F Garrity et al. “Pseudopotentials for high-throughput DFT calculations”. In: *Computational Materials Science* 81 (2014), pp. 446–452.

[38] Hamid Ghasemi, Harold S Park, and Timon Rabczuk. “A level-set based IGA formulation for topology optimization of flexoelectric materials”. In: *Computer Methods in Applied Mechanics and Engineering* 313 (2017), pp. 239–258.

[39] Hamid Ghasemi, Harold S Park, and Timon Rabczuk. “A multi-material level set-based topology optimization of flexoelectric composites”. In: *Computer Methods in Applied Mechanics and Engineering* 332 (2018), pp. 47–62.

[40] M Gibertini et al. “Magnetic 2D Materials and Heterostructures”. In: *Nature nanotechnology* 14.5 (2019), pp. 408–419.

[41] Justin Gilmer et al. “Neural message passing for quantum chemistry”. In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272.

[42] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks”. In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[43] Vladislav Gladkikh et al. “Machine learning for predicting the band gaps of ABX3 perovskites from elemental properties”. In: *The Journal of Physical Chemistry C* 124.16 (2020), pp. 8905–8918.

[44] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. “USPEX—Evolutionary crystal structure prediction”. In: *Computer physics communications* 175.11-12 (2006), pp. 713–720.

[45] Rhys EA Goodall and Alpha A Lee. “Predicting materials properties without crystal structure: Deep representation learning from stoichiometry”. In: *arXiv preprint arXiv:1910.00617* (2019).

[46] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[47] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[48] Lipichanda Goswami, Manoj Deka, and Mohendra Roy. “Artificial Intelligence in Material Engineering: A review on applications of AI in Material Engineering”. In: *arXiv preprint arXiv:2209.11234* (2022).

[49] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. “Effect of the Damping Function in Dispersion Corrected Density Functional Theory”. In: *Journal of Computational Chemistry* 32.7 (), pp. 1456–1465. DOI: 10.1002/jcc.21759. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21759. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21759.

[50] Stefan Grimme et al. “A Consistent and Accurate ab initio Parametrization of Density Functional Dispersion correction (DFT-D) for the 94 elements H-Pu”. In: *The Journal of Chemical Physics* 132.15 (2010), p. 154104. DOI: 10.1063/1.3382344. eprint: https://doi.org/10.1063/1.3382344. URL: https://doi.org/10.1063/1.3382344.

[51] Jiuxiang Gu et al. “Recent advances in convolutional neural networks”. In: *Pattern recognition* 77 (2018), pp. 354–377.

[52] Shu-Hui Guan, Cheng Shang, and Zhi-Pan Liu. “Resolving the temperature and composition dependence of Ion conductivity for yttria-stabilized zirconia from machine learning simulation”. In: *The Journal of Physical Chemistry C* 124.28 (2020), pp. 15085–15093.

[53] Sten Haastrup et al. “The Computational 2D Materials Database: High-throughput Modeling and Discovery of Atomically Thin Crystals”. In: *2D Materials* 5.4 (2018), p. 042002.

[54] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[55] Hossein Hajiabolhassan et al. “FunQG: Molecular Representation Learning Via Quotient Graphs”. In: *arXiv preprint arXiv:2207.08597* (2022).

[56] P Shiv Halasyamani and Kenneth R Poeppelmeier. “Noncentrosymmetric oxides”. In: *Chemistry of Materials* 10.10 (1998), pp. 2753–2769.

[57] Cameron J Hargreaves et al. “The Earth Mover’s Distance as a Metric for the Space of Inorganic Compositions”. In: *Chemistry of Materials* (2020).

[58] Kan Hatakeyama-Sato et al. “Synthesis of lithium-ion conducting polymers designed by machine learning-based prediction and screening”. In: *Chemistry Letters* 48.2 (2019), pp. 130–132.

[59] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778.

[60] Kaiming He et al. “Identity mappings in deep residual networks”. In: *European conference on computer vision.* Springer. 2016, pp. 630–645.

[61] R Hill. “The Elastic Behaviour of a Crystalline Aggregate”. In: *Proceedings of the Physical Society. Section A* 65.5 (May 1952), pp. 349–354. DOI: 10.1088/0370-1298/65/5/307. URL: https://doi.org/10.1088/0370-1298/65/5/307.

[62] Lauri Himanen et al. “DScribe: Library of descriptors for machine learning in materials science”. In: *Computer Physics Communications* 247 (2020), p. 106949.

[63] Jordan Hoffmann et al. “Data-driven approach to encoding and decoding 3-d crystal structures”. In: *arXiv preprint arXiv:1909.00949* (2019).

[64] Jianjun Hu, Wenhui Yang, and Edirisuriya M Dilanga Siriwardane. “Distance matrix-based crystal structure prediction using evolutionary algorithms”. In: *The Journal of Physical Chemistry A* 124.51 (2020), pp. 10909–10919.

[65] John Ingraham et al. “Generative models for graph-based protein design”. In: *Advances in Neural Information Processing Systems.* 2019, pp. 15820–15831.

[66] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. “Artificial neural networks: A tutorial”. In: *Computer* 29.3 (1996), pp. 31–44.

[67] Anubhav Jain et al. “The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL materials* 1.1 (2013), p. 011002.

[68] Dipendra Jha et al. “Elemnet: Deep learning the chemistry of materials from only elemental composition”. In: *Scientific reports* 8.1 (2018), pp. 1–13.

[69] Xiao-Ming Jiang et al. “Material research from the viewpoint of functional motifs”. In: *National Science Review* (2022).

[70] McCulloch JL. “A logical calculus of ideas immanent in nervous activity”. In: *Bull. of Math. Biophysics* 5 (1943), pp. 115–133.

[71] Gisli Holmar Johannesson et al. “Combined electronic structure and evolutionary search approach to materials design”. In: *Physical Review Letters* 88.25 (2002), p. 255506.

[72] Michael I Jordan and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.

[73] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[74] Arnab Kabiraj, Mayank Kumar, and Santanu Mahapatra. “High-throughput Discovery of High Curie Point Two-dimensional Ferromagnetic Materials”. In: *npj Computational Materials* 6.1 (2020), pp. 1–9.

[75] Seiji Kajita et al. “A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks”. In: *Scientific reports* 7.1 (2017), pp. 1–9.

[76] Baekjun Kim, Sangwon Lee, and Jihan Kim. “Inverse design of porous materials using artificial neural networks”. In: *Science advances* 6.1 (2020), eaax9324.

[77] Sungwon Kim et al. “Generative adversarial networks for crystal structure prediction”. In: *arXiv preprint arXiv:2004.01396* (2020).

[78] Yongtae Kim et al. “Deep learning framework for material design space exploration using active transfer learning and data augmentation”. In: *npj Computational Materials* 7.1 (2021), pp. 1–7.

[79] Scott Kirklin et al. “The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies”. In: *npj Computational Materials* 1.1 (2015), pp. 1–15.

[80] Walter Kohn. “Nobel Lecture: Electronic structure of matter—wave functions and density functionals”. In: *Reviews of Modern Physics* 71.5 (1999), p. 1253.

[81] G. Kresse and J. Furthmüller. “Efficient Iterative Schemes for ab initio Total-Energy Calculations Using a Plane-Wave Basis Set”. In: *Phys. Rev. B* 54 (16 Oct. 1996), pp. 11169–11186. DOI: 10.1103/PhysRevB.54.11169. URL: https://link.aps.org/doi/10.1103/PhysRevB.54.11169.

[82] G. Kresse and J. Hafner. “ab initio”. In: *Phys. Rev. B* 47 (1 Jan. 1993), pp. 558–561. DOI: 10.1103/PhysRevB.47.558. URL: https://link.aps.org/doi/10.1103/PhysRevB.47.558.

[83] G. Kresse and J. Hafner. “ab initio”. In: *Phys. Rev. B* 49 (20 May 1994), pp. 14251–14269. DOI: 10.1103/PhysRevB.49.14251. URL: https://link.aps.org/doi/10.1103/PhysRevB.49.14251.

[84] G. Kresse and D. Joubert. “From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method”. In: *Phys. Rev. B* 59 (3 Jan. 1999), pp. 1758–1775. DOI: 10.1103/PhysRevB.59.1758. URL: https://link.aps.org/doi/10.1103/PhysRevB.59.1758.

[85] Brian Kuhlman and Philip Bradley. “Advances in protein structure prediction and design”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 681–697.

[86] Alexander G Kvashnin, Zahed Allahyari, and Artem R Oganov. “Computational discovery of hard and superhard materials”. In: *Journal of Applied Physics* 126.4 (2019), p. 040901.

[87] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[88] Shaobo Li et al. “Critical temperature prediction of superconductors based on atomic vectors and deep learning”. In: *Symmetry* 12.2 (2020), p. 262.

[89] Haotong Liang et al. “CRYSPNet: Crystal Structure Predictions via Neural Network”. In: *arXiv preprint arXiv:2003.14328* (2020).

[90] Andy Liaw, Matthew Wiener, et al. “Classification and regression by randomForest”. In: *R news* 2.3 (2002), pp. 18–22.

[91] Zhixin Liu et al. “Modelling and parameter optimization for filament deformation in 3D cementitious material printing using support vector machine”. In: *Composites Part B: Engineering* 193 (2020), p. 108018.

[92] Steph-Yves Louis et al. “Graph convolutional neural networks with global attention for improved materials property prediction”. In: *Physical Chemistry Chemical Physics* 22.32 (2020), pp. 18141–18148.

[93] Shuaihua Lu et al. “Coupling a Crystal Graph Multilayer Descriptor to Active Learning for Rapid Discovery of 2D Ferromagnetic Semiconductors/Half-Metals/Metals”. In: *Advanced Materials* (2020), p. 2002658.

[94] Chade Lv et al. “Machine learning: an advanced platform for materials development and state prediction in lithium-ion batteries”. In: *Advanced Materials* 34.25 (2022), p. 2101474.

[95] Andriy O Lyakhov et al. “New developments in evolutionary structure prediction algorithm USPEX”. In: *Computer Physics Communications* 184.4 (2013), pp. 1172–1182.

[96] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using T-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[97] John Maddox. “Crystals from first principles”. In: *Nature* 335.6187 (1988), pp. 201–201.

[98] Ruben Mas-Balleste et al. “2D materials: to graphene and beyond”. In: *Nanoscale* 3.1 (2011), pp. 20–30.

[99] Efim Mazhnik and Artem R Oganov. “Application of machine learning methods for predicting new superhard materials”. In: *Journal of Applied Physics* 128.7 (2020), p. 075102.

[100] Bryce Meredig et al. “Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery”. In: *Molecular Systems Design & Engineering* 3.5 (2018), pp. 819–825.

[101] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[102] Tom M Mitchell and Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[103] Félix Mouhat and François-Xavier Coudert. “Necessary and Sufficient Elastic Stability Conditions in Various Crystal Systems”. In: *Physical Review B* 90 (Sept. 2014). DOI: 10.1103/PhysRevB.90.224104.

[104] Anthony J Moulson and John M Herbert. *Electroceramics: materials, properties, applications*. John Wiley & Sons, 2003.

[105] Nicolas Mounet et al. “Two-dimensional Materials from High-throughput Computational Exfoliation of Experimentally Known Compounds”. In: *Nature nanotechnology* 13.3 (2018), pp. 246–252.

[106] Gyoung S Na et al. “Tuplewise Material Representation Based Machine Learning for Accurate Band Gap Prediction”. In: *The Journal of Physical Chemistry A* 124.50 (2020), pp. 10616–10623.

[107] SS Nanthakumar et al. “Detection of material interfaces using a regularized level set method in piezoelectric structures”. In: *Inverse Problems in Science and Engineering* 24.1 (2016), pp. 153–176.

[108] Juhwan Noh et al. “Inverse design of solid-state materials via a continuous representation”. In: *Matter* 1.5 (2019), pp. 1370–1384.

[109] Artem R Oganov. *Modern methods of crystal structure prediction*. John Wiley & Sons, 2011.

[110] Artem R Oganov and Colin W Glass. “Crystal structure prediction using ab initio evolutionary techniques: Principles and applications”. In: *The Journal of chemical physics* 124.24 (2006), p. 244704.

[111] Artem R Oganov, Andriy O Lyakhov, and Mario Valle. “How Evolutionary Crystal Structure Prediction Works and Why”. In: *Accounts of chemical research* 44.3 (2011), pp. 227–237.

[112] Artem R Oganov et al. “Structure prediction drives materials discovery”. In: *Nature Reviews Materials* 4.5 (2019), pp. 331–348.

[113] Hirotsugu Ogi et al. “Elastic constants, internal friction, and piezoelectric coefficient of α- TeO 2”. In: *Physical Review B* 69.2 (2004), p. 024104.

[114] Kang Min Ok, Eun Ok Chi, and P Shiv Halasyamani. “Bulk characterization methods for non-centrosymmetric materials: second-harmonic generation, piezoelectricity, pyroelectricity, and ferroelectricity”. In: *Chemical Society Reviews* 35.8 (2006), pp. 710–717.

[115] Thomas Olsen et al. “Discovering Two-dimensional Topological Insulators from High-throughput Computations”. In: *Physical Review Materials* 3.2 (2019), p. 024005.

[116] Shyue Ping Ong et al. “Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis”. In: *Computational Materials Science* 68 (2013), pp. 314–319.

[117] JT Paul et al. “Computational Methods for 2D Materials: Discovery, Property Characterization, and Application Design”. In: *Journal of Physics: Condensed Matter* 29.47 (2017), p. 473001.

[118] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized Gradient Approximation Made Simple". In: *Phys. Rev. Lett.* 77 (18 Oct. 1996), pp. 3865–3868. DOI: 10.1103/PhysRevLett.77.3865. URL: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[119] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]". In: *Phys. Rev. Lett.* 78 (7 Feb. 1997), pp. 1396–1396. DOI: 10.1103/PhysRevLett.78.1396. URL: https://link.aps.org/doi/10.1103/PhysRevLett.78.1396.

[120] Chris J Pickard and RJ Needs. "Ab initio random structure searching". In: *Journal of Physics: Condensed Matter* 23.5 (2011), p. 053201.

[121] Evgeny V Podryabinkin et al. "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning". In: *Physical Review B* 99.6 (2019), p. 064114.

[122] Evan Pretti et al. "Symmetry-Based Crystal Structure Enumeration in Two Dimensions". In: *The Journal of Physical Chemistry A* 124.16 (2020), pp. 3276–3285.

[123] Seeram Ramakrishna et al. "Materials informatics". In: *Journal of Intelligent Manufacturing* 30.6 (2019), pp. 2307–2326.

[124] Rampi Ramprasad et al. "Machine learning in materials informatics: recent applications and prospects". In: *npj Computational Materials* 3.1 (2017), pp. 1–13.

[125] Zekun Ren et al. "Inverse design of crystals using generalized invertible crystallographic representation". In: *arXiv preprint arXiv:2005.07609* (2020).

[126] G Revathy, V Rajendran, and P Sathish Kumar. "Prediction study on critical temperature (C) of different atomic numbers superconductors (both gaseous/solid elements) using machine learning techniques". In: *Materials Today: Proceedings* 44 (2021), pp. 3627–3632.

[127] Vivek Revi et al. "Machine learning elastic constants of multi-component alloys". In: *Computational Materials Science* 198 (2021), p. 110671.

[128] Jeffrey M Rickman, Turab Lookman, and Sergei V Kalinin. "Materials informatics: From the atomic-level to the continuum". In: *Acta Materialia* 168 (2019), pp. 473–510.

[129] Matthew C Robinson, Robert C Glen, and Alpha A Lee. "Validating the Validation: Reanalyzing a large-scale comparison of Deep Learning and Machine

Learning models for bioactivity prediction". In: *arXiv preprint arXiv:1905.11681* (2019).

[130] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "The Earth Mover's Distance as a Metric for Image Retrieval". In: *International journal of computer vision* 40.2 (2000), pp. 99–121.

[131] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. "Crystal structure prediction via deep learning". In: *Journal of the American Chemical Society* 140.32 (2018), pp. 10158–10168.

[132] Esteban Samaniego et al. "An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications". In: *Computer Methods in Applied Mechanics and Engineering* 362 (2020), p. 112790.

[133] Gabriel R Schleder et al. "From DFT to machine learning: recent approaches to materials science–a review". In: *Journal of Physics: Materials* 2.3 (2019), p. 032001.

[134] Kristof T Schütt et al. "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions". In: *arXiv preprint arXiv:1706.08566* (2017).

[135] Peter Schwerdtfeger and Jeffrey K Nagle. "2018 Table of static dipole polarizabilities of the neutral elements in the periodic table". In: *Molecular Physics* 117.9-12 (2019), pp. 1200–1225.

[136] Austin D Sendek et al. "Machine learning-assisted discovery of solid Li-ion conducting materials". In: *Chemistry of Materials* 31.2 (2018), pp. 342–352.

[137] Andrew W Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710.

[138] Edirisuriya M Dilanga Siriwardane et al. "Generative design of stable semiconductor materials using deep learning and density functional theory". In: *npj Computational Materials* 8.1 (2022), p. 164.

[139] Yuqi Song et al. "Computational discovery of new 2D materials using deep learning generative models". In: *ACS Applied Materials & Interfaces* 13.45 (2021), pp. 53303–53313.

[140] Murat Cihan Sorkun et al. "An Artificial Intelligence-aided Virtual Screening Recipe for Two-dimensional Materials Discovery". In: *npj Computational Materials* 6.1 (2020), pp. 1–10.

[141] Valentin Stanev et al. “Machine learning modeling of superconducting critical temperature”. In: *npj Computational Materials* 4.1 (2018), pp. 1–14.

[142] Jia-li Tang, Qiu-ru Cai, and Yi-jun Liu. “Prediction of material mechanical properties with Support Vector Machine”. In: *2010 International Conference on Machine Vision and Human-machine Interface*. IEEE. 2010, pp. 592–595.

[143] A Togo and I Tanaka. “First Principles Phonon Calculations in Materials Science”. In: *Scr. Mater.* 108 (Nov. 2015), pp. 1–5.

[144] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[145] P Villars et al. “Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB”. In: *Journal of alloys and compounds* 367.1-2 (2004), pp. 167–175.

[146] Junjie Wang et al. “MAGUS: machine learning and graph theory assisted universal structure searcher”. In: *National Science Review* (2023), nwad128.

[147] Vei Wang et al. “VASPKIT: a User-friendly Interface Facilitating High-throughput Computing and Analysis Using VASP Code”. In: *arXiv preprint arXiv:1908.08269* (2019).

[148] Xiaoting Wang et al. “Recent Advances in the Functional 2D Photonic and Optoelectronic Devices”. In: *Advanced Optical Materials* 7.3 (2019), p. 1801274.

[149] Yanchao Wang et al. “CALYPSO method for structure prediction and its applications to materials discovery”. In: *Handbook of Materials Modeling: Applications: Current and Emerging Materials* (2020), pp. 2729–2756.

[150] Yanchao Wang et al. “Materials discovery via CALYPSO methodology”. In: *Journal of Physics: Condensed Matter* 27.20 (2015), p. 203203.

[151] Zhong Lin Wang. “Piezotronic and piezophototronic effects”. In: *The Journal of Physical Chemistry Letters* 1.9 (2010), pp. 1388–1393.

[152] Zhuo Wang et al. “Data-Driven Materials Innovation and Applications”. In: *Advanced Materials* 34.36 (2022), p. 2104113.

[153] Logan Ward et al. “A general-purpose machine learning framework for predicting properties of inorganic materials”. In: *npj Computational Materials* 2.1 (2016), pp. 1–7.

[154] Logan Ward et al. “Matminer: An Open Source Toolkit for Materials Data Mining”. In: *Computational Materials Science* 152 (2018), pp. 60–69.

[155] Guo-Wei Wei. “Protein structure prediction beyond AlphaFold”. In: *Nature Machine Intelligence* 1.8 (2019), pp. 336–337.

[156] *What Are Piezoelectric Materials? Kernel Description.* https://sciencing.com/piezoelectric-materials-8251088.html. Accessed: 2010-09-30.

[157] Scott M Woodley and Richard Catlow. “Crystal structure prediction from first principles”. In: *Nature materials* 7.12 (2008), pp. 937–946.

[158] Zonghan Wu et al. “A comprehensive survey on graph neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[159] Tian Xie and Jeffrey C Grossman. “Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties”. In: *Physical review letters* 120.14 (2018), p. 145301.

[160] Zheng Xiong et al. “Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation”. In: *Computational Materials Science* 171 (2020), p. 109203.

[161] Tong Yang et al. “High-Throughput Identifications of Exfoliable Two-Dimensional Materials with Active Basal Planes for Hydrogen Evolution”. In: *ACS Energy Letters* (2020).

[162] Lijun Zhang et al. “Materials discovery at high pressures”. In: *Nature Reviews Materials* 2.4 (2017), pp. 1–16.

[163] Xu Zhang, An Chen, and Zhen Zhou. “High-throughput Computational Screening of Layered and Two-dimensional Materials”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 9.1 (2019), e1385.

[164] Xu Zhang et al. “Computational Screening of 2D Materials and Rational Design of Heterojunctions for Water Splitting Photocatalysts”. In: *Small Methods* 2.5 (2018), p. 1700359.

[165] Yong Zhao et al. “Physics guided deep learning for generative design of crystal materials with symmetry constraints”. In: *npj Computational Materials* 9.1 (2023), p. 38.

[166] Yong Zhao et al. “Predicting elastic properties of materials from electronic charge density using 3d deep convolutional neural networks”. In: *The Journal of Physical Chemistry C* 124.31 (2020), pp. 17262–17273.

[167] Wei Zheng et al. “Deep-learning contact-map guided protein structure prediction in CASP13”. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1149–1164.

[168] Jie Zhou et al. “Graph neural networks: A review of methods and applications”. In: *AI Open* 1 (2020), pp. 57–81.

[169] Jun Zhou et al. “2DMatPedia, an Open Computational Database of Two-dimensional Materials from Top-down and Bottom-up Approaches”. In: *Scientific data* 6.1 (2019), pp. 1–10.

[170] Kaixiong Zhou et al. “Towards Deeper Graph Neural Networks with Differentiable Group Normalization”. In: *arXiv preprint arXiv:2006.06972* (2020).

[171] Zizhong Zhu et al. “An Efficient Scheme for Crystal Structure Prediction Based on Structural Motifs”. In: *The Journal of Physical Chemistry C* 121.21 (2017), pp. 11891–11896.

[172] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. “Predicting the band gaps of inorganic solids by machine learning”. In: *The journal of physical chemistry letters* 9.7 (2018), pp. 1668–1673.

[173] Alex Zunger. “Inverse design in search of materials with target functionalities”. In: *Nature Reviews Chemistry* 2.4 (2018), pp. 1–16.