Spring 2022

Deep Learning Based Generative Materials Design

Yong Zhao

Deep Learning based generative materials design

by

Yong Zhao

Bachelor of Electrical Engineering and Management
Qingdao University of Technology 2013

Master of Control Engineering
Qingdao University of Technology 2015

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

2022

Accepted by:

Jianjun Hu, Major Professor

John R. Rose, Committee Member

Song Wang, Committee Member

Yan Tong, Committee Member

Ming Hu, Committee Member

Tracey L. Weldon, Interim Dean and Vice Provost for Graduate Education

Acknowledgments

I gratefully acknowledge the mentoring of my advisor, Dr. Jianjun Hu, for his continual guidance and support throughout my PhD research path. Without him, the achievements I have realized could not have been accomplished. He enlightened me on my research topics in many ways. His passion for using Deep Learning/Machine Learning in material science always encourages me on my research path. His advice on both research and career development has been invaluable.

I want to express my sincere gratitude to my dissertation committee members, Dr. Yan Tong, Dr. John R. Rose, Dr. Song Wang and Dr. Ming Hu, for their time and advice on my work. It is such an honor to have them serving in my committee.

I would like to acknowledge my collaborators from the Department of Computer Science and Engineering for all the useful discussions and valuable comments. I want to specially thank Yuxin Cui and Dr. Zhonghao Liu for their effective and beneficial discussion of research, career development, and support living in South Carolina, as well as Zheng Xiong, Zhenyao Wu, Xinyi Wu, Steph-Yves Louis, Dr. Yanxia Liu, Dr. Fei Liu, Yuqi Song, Rui Xin for their contributions to my research.

Most importantly, I want to express my gratitude to my parents for their unconditional love and support. Their sacrifices and confidence in me have made this possible. Without them I would not have come this far. Furthermore, I would like to express my heartfelt gratitude to my girlfriend Shujie Chen for her love and support in this journey. Her optimism and reassurances were invaluable through the most difficult moments of this path.

Abstract

Discovery of novel functional materials is playing an increasingly important role in many key industries such as lithium batteries for electric vehicles and cell phones. However experimental tinkering of existing materials or Density Functional Theory (DFT) based screening of known crystal structures, two of the major current materials design approaches, are both severely constrained by the limited scale (around 250,000 in ICSD database) and diversity of existing materials and the lack of a sufficient number of materials with annotated properties. How to generate a large number of physically feasible, stable, and synthesizable crystal materials and build accurate property prediction models for screening are the two major unsolved challenges in modern materials science.

This dissertation is focused on addressing these two fundamental tasks in material science using deep learning/machine learning models. Deep learning and machine learning have already made tremendous progress in computer vision and natural language processing, as shown by autonomous driving cars and Google's translators, and have the potential to greatly transform the research of materials science. Compared to conventional tinkering based materials discovery methods, data-driven approaches have been increasingly used in material informatics due to their significantly faster screening speeds for new materials. In this dissertation, we design and develop novel deep learning-based algorithms to learn the hidden intricate chemical rules that assemble atoms into stable crystal structures from known crystals and to generate new crystal structures . We also explore and develop novel representation learning methods upon materials compositions and structures for high performance prediction

of materials structural characteristics and elastic properties.

In the first topic, we propose CubicGAN, a generative adversarial network (GAN) based deep neural network model for large-scale generative design of novel cubic materials. When trained on 375 749 ternary materials from the OQMD database, we show that the model can not only rediscover most of the currently known cubic materials but also generate hypothetical materials of new structure prototypes. A total of 506 such materials have been verified by DFT based phonon dispersion calculation. Our technique allows to generate tens of thousands of new materials given sufficient computing resources.

In the second topic, we propose a Physics Guided Crystal Generative Model (PGCGM) for new materials generation, which significantly expands the structural scope of CubicGAN by bringing the capability of generating crystals of 20 space groups. This is achieved by capturing and exploiting the pairwise atomic distance constraints among neighbor atoms, symmetric geometric constraints, and a novel data augmentation strategy using the base atom sites of materials. With atom clustering and merging on generated crystal structures, our method increases the generator's validity 8 times when compared to one of the baselines and by 143% compared to the previous CubicGAN, along with its superiority in properties distribution and diversity. We further validated our generated candidates by DFT calculations, which successfully optimized/relaxed 1869 materials out of 2000 generated ones, of which 39.6% had negative formation energy, indicating their stability.

In the third topic, we propose and evaluate machine-learning algorithms for determining the structure type of materials, given only their compositions. We couple random forest (RF) and multiple-layer perceptron (MLP) neural network models with three types of features: Magpie, atom vectors, and one-hot encoding (atom frequency) for the crystal system and space group prediction of materials. Four types of models for predicting crystal systems and space groups are proposed,

trained, and evaluated including one-versus-all binary classifiers, multiclass classifiers, polymorphism predictors, and multilabel classifiers. The synthetic minority over-sampling technique (SMOTE) is conducted to mitigate the effects of imbalanced data sets. Our results demonstrate that RF with Magpie features generally outperforms other algorithms for binary and multiclass prediction of crystal systems and space groups, while MLP with atom frequency features is the best method for structural polymorphism prediction.

Finally, we propose using electronic charge density (ECD) as a generic unified 3D descriptor for materials property prediction due to its advantage of possessing a close relation with the physical and chemical properties of materials. We develop an ECD-based 3D convolutional neural network (CNN) to predict the elastic properties of materials in which CNNs can learn effective hierarchical features with multiple convolving and pooling operations. Our experiments show that our method can achieve good performance for elasticity prediction over 2170 Fm-3m materials.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Discovering new materials and understanding their physical and chemical properties are key focuses in the material science community. Materials with excellent properties, such as very low lattice thermal conductivity [99], high-temperature superconductivity [84], ultra-hardness [147], and low electronic conductivity [137] are highly desired by various industries such as electric car companies, phone manufacturers, and mining companies. For a long time, material scientists discovered and explored materials and their properties primarily depending on experimental observations. Considering the huge number of possible combinations of material composition and crystal structures, it is infeasible to explore the whole compositional space as that is labor- and time-intensive.

Computational methodologies, such as Density Functional Theory (DFT) [70], provide a less expensive means to calculate materials' properties by simulating on the atomic level. Several large-scale databases of materials made by high-throughput computing such as the Materials Project [54], the Automatic Flow of Materials Discovery Library [25] and the Open Quantum Materials Database (OQMD) [132, 69] have been introduced in the material community. These databases contain DFT calculated properties of thousands of experimentally determined and hypothetical crystal structures. However, material discovery using DFT still is computationally demanding principally because materials design is a very complicated multi-optimization process.

Artificial intelligence has achieved tremendous successes in many fields such as Computer Vision [46], Natural Language Processing [93, 94] and Bioinformatics [60]. Large-scale open-source materials databases with emerging deep learning/machine learning models have led to the prosperity of Materials Informatics [97, 128]. Materials Informatics is a data-driven machine learning approach to employ statistical and machine learning methods to learn the relationship between materials and their physical and chemical representations. The aim of Materials Informatics is to screen

thousands of compounds for potential new industrial materials at a much faster speed.

Because of their time- and cost-efficiency, machine learning and deep learning models have been widely used in material discovery. Overall, data-driven material design can have two types: predicting properties and identifying special materials from existing databases [176, 56, 165, 165, 63, 137], and generating hypothetical crystal structures by learning compositional space of existing materials [63, 71, 85, 24, 105, 64, 101].

Despite the fruitful achievements in materials informatics, deep learning/machine Learning does not have the same success in material science as in other fields. The reasons are two-fold. Firstly, the scale and diversity of existing material databases are severely limited compared to images and text databases. For instance, the number of materials with certain properties, such as low thermal conductivity, is pretty small [19]. Secondly, due to the difficulty in encoding the crystal structures, using deep learning/machine learning to link crystal structures to their corresponding properties remains an outstanding challenge. Therefore, generating more stable crystal structures and engineering new structural representations for materials are extremely needed in the material community.

1.2 Scope of the Proposed Research

In this dissertation, we focus on four topics:

1. As described in the Introduction section, it is essential to generate materials with high diversity and beyond the scope of existing databases. In this topic, we propose the CubicGAN, a Wasserstein generative adversarial network (GAN) [44, 5, 43] based model for large scale generative design of novel cubic materials. In order to make the size of inputs to the GAN model invariant, we choose the base atom positions for ternary and quaternary materials. Considering the fact that it is not easy to generate atom coordinates as we generate image pixels, we choose

materials with the special atom coordinates (i.e., a multiplicative factor of 0.25). By generating 10 million hypothetical ternary and 10 million quaternary crystal structures, we perform three stage checks and re-discover most of the cubic materials in existing databases, and find dozens of new prototypes which do not exist in existing materials. By further stability verification, 506 new prototype materials have been generated and confirmed to be stable by phonon dispersion calculation.

2. Incorporating physical laws of materials is critical in generating inorganic crystals. In this topic, we propose a Physics Guided Crystal Generative Model (PGCGM) which adds two losses based on atom distance and symmetric constraints. Atom distance loss restricts the atom distance to a certain range, which avoids extremely small or large volume of crystals. On the other hand, symmetric constraints avoid the generation of atom coordinates crowding together. With the augmentation of base atom sites, the PGCGM can generate materials for 20 space groups. Then we propose a post-processing method on the generated materials to lower the number of atoms using clustering and merging. By DFT calculations/relaxations, 1869 materials out of 2000 ones are optimized successfully, which indicates the effectiveness of our physics guided model.

3. Structural information of compositions such as crystal structure and space group is of vast importance to determine the crystal structures and corresponding properties. However, existing methods, such as X-ray diffraction (XRD) or first-principle-based structure determination, are infeasible to determine the enormous composition space. Herein, we propose and evaluate machine-learning algorithms for determining the structure type of materials given only their compositions. Due to many-to-many relationships between compositions and crystal structures, we come up with four types of classification problems and

we compare the Random Forest and Multiple Layer Perceptron learning algorithms using different composition representations: one-hot encoding, atom embedding, and Magpie [165]. The four types of classification problems are binary classifiers, multi-label classifiers, multi-class classifiers, and polymorphism predictors. We train, validate, and evaluate these classifiers using data from Materials Project [54].

4. Materials representation plays a significant part in predicting materials' properties. However, it is not an easy task to find a generic representation that carries rich information for materials. Currently, both graph [176, 18] and three-dimensional (3D) voxel representations [62] are proposed based on the heterogeneous elements of the crystal structures. In this topic, we propose a new 3D representation called electronic charge density (ECD) [6] to process the relationship between materials and their physical and chemical properties. The ECD compacts electron distributions and local potentials that are critical to determine the materials' properties. With ECD representation, it is suitable for us to implement Convolutional Neural Networks (CNNs) [79]. We examine the 3D CNNs and the 2D CNNs embedded with Squeeze-and-Excitation (SE) block [52] by a regression test for two physical properties: the bulk modulus and the shear modulus. Extensive experiments show that the 2D CNNs infused with Magpie features [165] achieve the best regression results and significantly better extrapolation performance. As an extra validation, we evaluate the predictive performance of our models on 329 materials from (OQMD) [132, 69] of space group $Fm\bar{3}m$ by comparing to DFT calculated values, which shows a better prediction power of our model for bulk modulus than the shear modulus. The novel generic 3D representation for materials opens a new way to import excellent CNNs-based algorithms to predict other physical and chemical properties of materials.

1.3 Structure of the Dissertation

In chapter 2, we provide a brief review of deep learning and its application in Computer Vision, Bioinformatics, and common research trend in Material Informatics, along with some basic introduction of deep learning algorithms that we used in our research. In chapter 3, we present CubicGAN, a generative model to generate cubic materials in a high-throughput way. In chapter 4, we integrate atom distance and symmetric constraints based losses into the generative loss and with data augmentation and atom clustering and merging. The physics guided crystal generative model can generate high quality materials for 20 space groups. In chapter 5, we examine different machine learning methods and composition representation to predict structural information of materials. In chapter 6, we develop a new 3D voxel material representation and regress elasticity using CNNs. Furthermore, we apply our trained models to new materials and achieve good results. In chapter 7, we conclude our current works and introduce the future work that designs an unified framework to generate materials.

Chapter 2

Background

2.1 Deep Learning and Its Applications

Deep learning is a sub-field of Machine Learning, which uses multiple hidden layers of neural networks to hierarchically extract information from the vast amount of data. With its non-linear transformations and model abstractions, deep learning has gained notable attention from researchers and achieved great success in many fields, such as Computer Vision [46, 75], Natural Language Processing [158, 93, 94], Bioinformatics [60]. The reasons behind the booming of deep learning can be a lot, but the advancement of hardware (GPUs), the large volume of data, and the numerous innovated algorithms are the top 3 reasons. In this dissertation, we will elaborate on several representative algorithms in those fields that promote the development of deep learning in academia and industries.

AlexNet [75] consists of five convolutional layers, and some of them are followed by max-pooling layers and three fully connected layers. Nonlinearity transformation, ReLU, is engineered used to help speed up the training and the Dropout mechanism is introduced to reduce the over-fitting in fully connected layers. By spreading the network on two GPUs, AlexNet achieves state-of-the-art performance on imageNet classification competition. With the invention of the ReLU activation function and Dropout, AlexNet is believed to be one of the pioneering works that spark the emerging of deep learning. Residual neural network (ResNet) [46] is another work that reforms the Computer Vision field, which allows the depth of neural networks to be several hundred by skip-connections to jump over some layers. With skip-connections, the vanishing of gradients is mitigated so that the model can gain accuracy from significantly increased depth.

In Natural Language Processing field, lots of progress have been made. Word embeddings [93, 94, 116, 117, 118] are a type of distributed learned word representation that allows words with similar meanings to share close representations. The distributed

representation of text is a key breakthrough to achieving excellent performance on challenging Natural Language Processing tasks. Word2Vec [93, 94] is one of typical word embedding algorithms that use a statistical method to convert one-hot encoding of words to condensed vectors by learning from a text corpus. Word2Vec utilizes two different neural network architectures: the Skip-gram and continuous Bag-of-Words (CBOW). The skip-gram model learns by predicting the surrounding words given a current word and however CBOW model learns by predicting the current word given its context. Negative sampling and hierarchical softmax are used to accelerate the training. Word embeddings have achieved lots of improvements in downstream tasks, such as Sentiment Analysis [35], Spam Detection [89], and Sentence Classification [186].

Proteins play critical roles in all living organisms. Understanding their structures can help facilitate the understanding of mechanism of the living organisms. Traditional methods, such as experimentation, require a huge amount of time to determine the structure of proteins. Many efforts [191, 3, 115, 162, 138] have been employed to predict the 3D structure of proteins, but the atomic accuracy is not at a satisfactory level. The recent work of AlphaFold [60] provides the first computational method that can predict protein structures with atomic accuracy close to the experimentation results in a majority of cases. AlphaFold incorporates new equivariant attention architectures, graph inferences, and pair representations based on physical and geometric constraints into the design of deep learning and greatly outperforms other methods in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14) [96]. Deep learning-based AlphaFold will be the essential tool of the biology field in the future.

Next, we will briefly introduce several workhorses of deep learning used in this dissertation, including fully connected neural networks, convolutional neural networks, generative adversarial networks, and autoencoder.

2.1.1 Fully Connected Neural Networks (fcNNs)

Fully connected neural networks (fcNNs) are a series of fully connected layers in which all neurons on each layer are connected to the previous layer as shown in Figure 2.1. The size of input is two and there is one output value. The first hidden layer and second hidden layer have three and four neurons, respectively. The major advantage of fcNNs is that they are "structure agnostic". In other words, any structure of the data can be fed as the input to fcNNs [125].

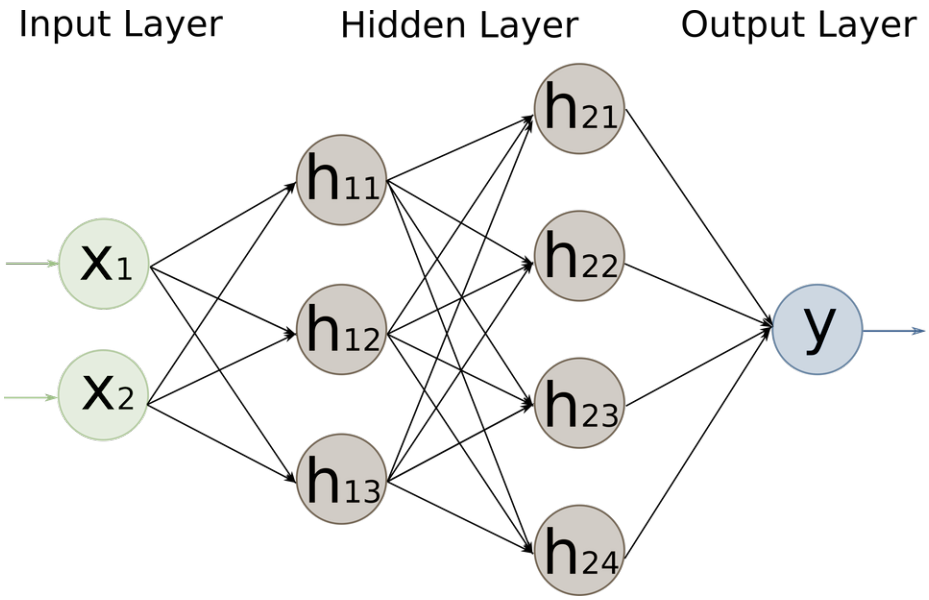


Figure 2.1 An example of a fully connected neural network model with two hidden layers.

The mathematical form of a fully connected layer can be formulated as below:

$$h^l = \sigma(W^l h^{l-1} + b^l) \tag{2.1}$$

where W^l is the weight and b^l is the bias parameter of the l-th layer, h^{l-1} and h^l are inputs and outputs of current layer, and σ is the activation function. The activation functions are differentiable and non-linear functions that help neural networks learn complex patterns from training data. The below equations show the commonly used activation functions in neural networks.

$$f(x) = \frac{1}{1+e^{-x}} \quad \text{sigmoid} \quad (2.2)$$

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{hyperbolic tangent} \quad (2.3)$$

$$f(x) = max(0, x) \quad \text{ReLU [75]} \quad (2.4)$$

$$f(x) = 1(x < 0)(\alpha x) + 1(x >= 0)(x) \quad \text{Leaky ReLU [87]} \quad (2.5)$$

2.1.2 Convolutional Neural Networks (CNNs)

Convolutional neural networks are first used in visual recognition tasks [79, 78]. It mainly consists of convolution, pooling, and fully connected layers stacked in some kind of order. Convolutional layers are the core building blocks of a convolutional neural network and it usually does most of the computational work. The convolutional layers convolve the local regions of the input to form feature maps by the kernels and pooling layers are used to reduce the size of feature maps. The parameter sharing mechanism and local connectivity of convolutional layers help reduce the number of parameters when dealing with high dimensional images since it is infeasible to connect neurons to all neurons in the previous layer as in fcNNs. Convolutional neural networks transform the original pixels from images into some class scores. Max-pooling and average-pooling are common pooling layers that extract the most important features by maximizing or averaging the local area in the input, respectively [12, 181].

When designing the convolutional neural networks, multiple convolutional layers can be stacked together followed by pooling layers [141], convolutional and pooling layers in parallel can formulate a block to increase the width of network [148, 150, 137], or we can use convolutional and pooling layers in series as blocks to build very deep neural networks with hundreds of layers [46, 53]. These innovative designs achieve state-of-the-art performance in image segmentation, object detection, and video tracking. Figure 2.2 shows a simple convolutional architecture.

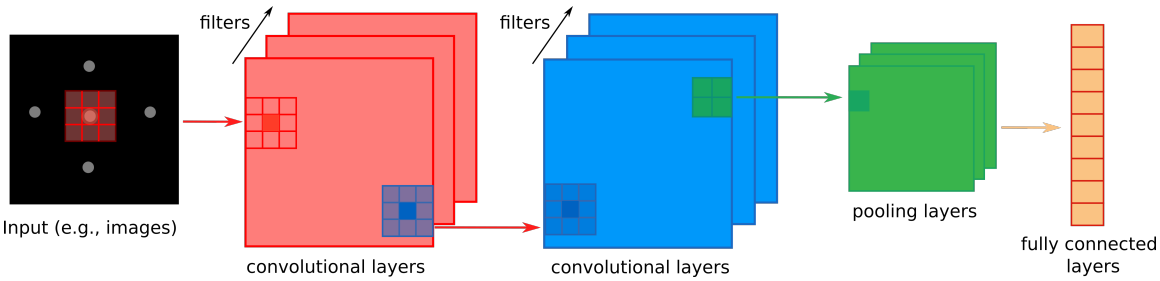


Figure 2.2 An example of a use of convolutional neural network with contiguous convolutional layers followed by pooling layers.

2.1.3 Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [43] is a type of generative models based on deep learning networks, such as CNNs. A GAN has two parts: generator and discriminator. The generator learns to generate plausible examples and the generated examples are the negative training data for the discriminator. The discriminator learns to tell real examples from the generator's fake examples and it penalizes the generator for creating fake examples. Two parts contest against each other to reach a balance status. The loss function is defined as below. The generator tries to minimize the following function while the discriminator tries to maximize it:

$$L_{GAN} = E_x[\log(D(x))] + E_z[\log(1 - D(G(x)))] \tag{2.6}$$

where $D(x)$ is the discriminator's estimation of the probability that a real sample is real, $G(x)$ is the fake data generated by the generator given random noise z, and $D(G(x))$ is the probability that a generated sample is fake estimated by the discriminator. We use the discriminator to optimize the parameters of the generator. The example use of an GAN is shown in Figure 2.3.

Many variants of GANs have been proposed. Wasserstein GAN [5, 44] is proposed to solve training instability of the standard GAN caused by gradients vanishing and mode collapse. In Cycle GAN [193], the authors present a system that maps an image from a source domain X to a target domain Y in the absence of paired samples. The

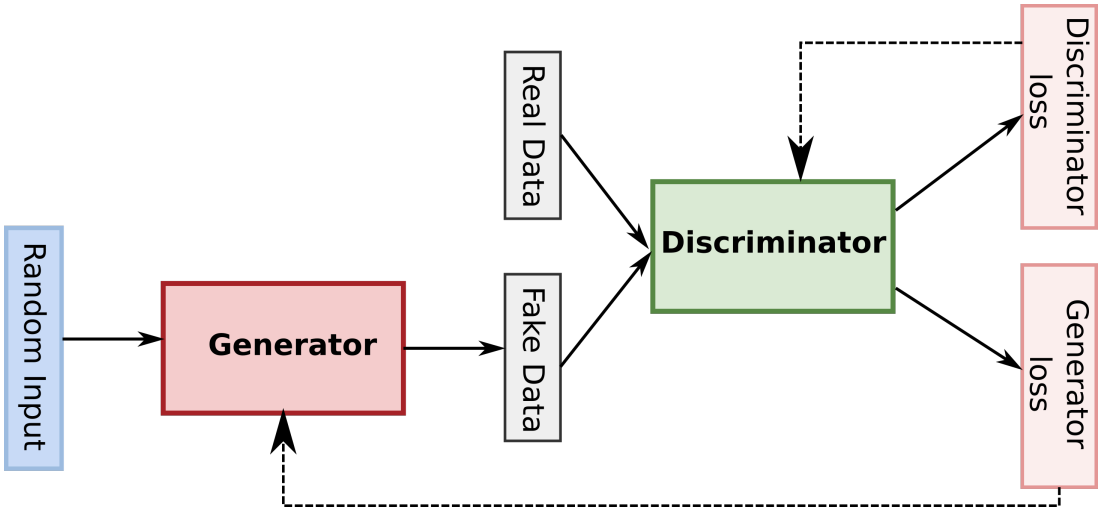


Figure 2.3 An example of the generative adversarial network architecture.

introduced consistency loss ensures to map the input images to the target domain. Conditional GAN [95] is proposed to add label information to the generator and discriminator and it generates corresponding samples.

2.1.4 Autoencoder (AE)

Autoencoder is an unsupervised learning method [42]. It consists of encoder and decoder architectures. An encoder maps the input into a latent vector and the latent vector can be reconstructed by the decoder. The learned latent vector can be used as the representation for a set of data by training the network to ignore noise data in the input. Variational autoencoder (VAE) [68] shares a similar structure to AE and belongs to the class of generative models. Instead of encoding the input as a latent vector, VAE encodes it as a distribution over the latent space. The distributions are decoded to reconstruct the original input.

2.2 Materials Informatics

In recent years, encouraged by the large volume of open-sourced material databases and by the algorithmic development and their success in various fields, materials

informatics has drawn more and more attention in material science [164, 124]. Materials informatics methods utilize data science methods to extract new patterns or predictive models from existing data, which can alleviate the burden of conventional methods such as DFT and trial-and-error experiments which have a much greater computational cost. Materials informatics approaches are based on three components: materials data, representations to quantitatively describe materials, and machine learning algorithms to map the data to its properties or to extract patterns from the data. In this dissertation, we will discuss key applications of materials informatics with a particular focus on inorganic materials. There has been a great amount of recent work using machine learning for materials, and we will not cover each of them. Instead, we only focus on some key achievements in the following subsections.

2.2.1 Properties Prediction

DFT based methods compute the properties of materials with minimal experimental input but at a large computational cost because individual energy evaluations are required [164]. It is not practical to perform calculations for thousands of materials. However, machine learning-based methods can predict the properties of materials in a high-throughput way at a little computational cost. The essential component of materials informatics is the representations for the materials. Many efforts have been made to represent materials numerically. There are broadly two types of representations: composition-based [165, 41] and crystal structure-based [176, 18, 62, 32, 135].

Composition-based representations take the stoichiometric attributes as input and use the descriptors to predict the properties of materials without crystal structures. Magpie [165] and Roost [41] are two excellent generic methods extracting knowledge from material formulas to predict materials properties such as formation energy and bandgap. Magpie [165] calculates the statistical information from chemical

formulas and the formed features serve as a general-purpose framework to discover new materials. Roost [41] treats the formulas as a dense weighted graph and learns improvable descriptors from data iteratively. Compared to Magpie, Roost is more flexible in learning patterns from a formula.

Crystal structure fundamentally determines the properties of materials. Designing features based on crystal structures becomes a prerequisite for a machine learning method with good performance. Researchers can not directly use the surface attributes (e.g., atom coordinates, types of atoms) in a crystal structure since they are neither invariant nor descriptive enough as a good input to machine learning models [164]. Therefore, how to design a structure-agnostic representation for materials remains an open question. In recent years, graph based methods have achieved notable performance in predicting properties of solid crystals and molecules [176, 18, 136, 59]. The Crystal Graph Convolutional Neural Networks (CGCNN) [176] encodes atomic information and bonding interaction between atoms and provides a universal and interpretable representation of crystalline materials. The CGCNN achieves a highly accurate prediction of 8 properties calculated by DFT. Figure 2.4 shows the crystal graph and the architecture of CGCNN. (a) Construction of the crystal graph. Nodes are atoms and edges are bonds between atoms. (b) The architecture of the graph convolutional neural network. Graph convolutional layer R convolves on the top of the crystal graph by iteratively updating atom features. The pooling layer then is used to produce a global vector for the crystal. Two hidden layers with the global vector as input are added to build the mapping between the crystal and its properties. Chen et al. [18] propose the MatErials Graph Network (MEGNet) that incorporates global states (e.g. temperature, enthalpy) into MEGNet blocks. MEGNet achieves better accuracy than DFT over a large dataset and outperforms prior graph models, such as SchNet [135] in 11 out of 13 properties of the QM9 molecule dataset. Additionally, the learned element embeddings by MEGNet can be used in transfer learning for small

datasets learning.

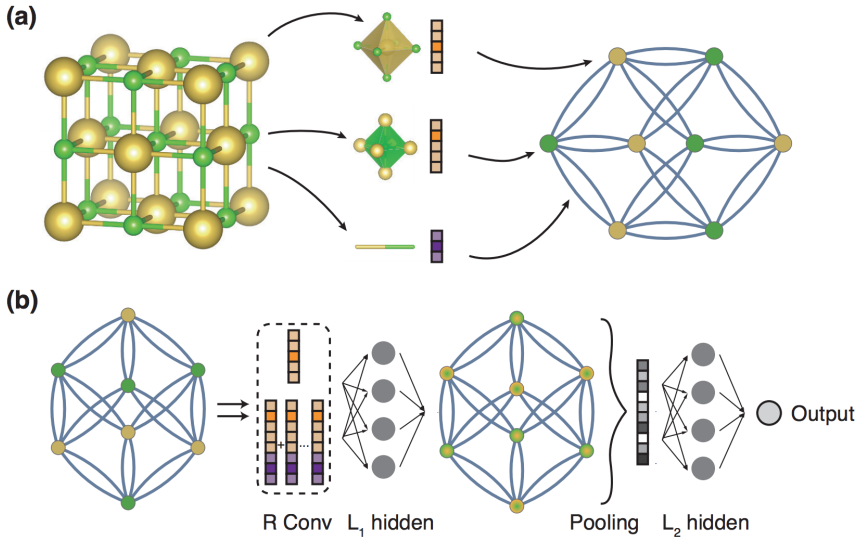


Figure 2.4 The illustration of CGCNN (image source: [176]).

2.2.2 Inverse Design

As mentioned in the last subsection, properties prediction is essential to learn materials-properties relationships and from them we discover materials with desired functional properties. Inverse design of inorganic materials goes in the opposite direction: given desire properties first, we want to find novel materials with those properties using mathematical algorithms and automations [102]. Conventionally, inverse design is based on humans' knowledge and candidate materials are examined computationally and experimentally, which is time-consuming and error-prone. Materials informatics mitigates the burden of traditional inverse design by replacing human knowledge with machine learning algorithms. High-throughput virtual screening (HTVS), global optimization (GO), and generative modeling (GM) are three primary approaches in the inverse design of inorganic crystals [133, 14, 101].

HTVS is defined as a process to search materials with desired properties in existing databases. However, one big disadvantage of HTVS is that the search is limited by the diversity and scope of the selected databases and human involvements are also

in the selection of databases. Thus some potential materials can be ignored since HTVS is implemented without a specific direction, which leads to inefficiency of the HTVS. One way to speed up HTVS is to perform global optimization. For instance, evolutionary algorithms can be used to find the most stable atom arrangement given a composition by mutations and crossovers [108].

Generative modeling is another way to expedite the inverse design. Generative modeling learns the distribution of the crystal functional space by adversarial training [64] and variational inference [101]. The difference between GO and GM is: global optimization learns the geometric landscapes of energy and properties of crystals by evolving iteratively, and generative modeling encodes the high-dimensional features as a latent vector and then it is mapped to novel materials with desired properties. We will introduce more methods of GO and GM in chapter 3.

2.3 New Generation of Deep Learning-based Generative Material Design Framework

Here, we propose a new framework as shown in Figure 2.5 which summarizes our work to explore the design space where the extraordinary materials exist. Past works use DFT calculation or experiments are used to discover exotic new materials and the discovered materials contribute to material databases [132, 69, 25, 54, 9], making for slow progress. Conversely, machine learning-based algorithms achieve progress at a much greater speed by establishing material-property connections, and the trained models are then used to quickly screen the existing databases to find materials with desired properties. DFT or experiments then can be utilized to only verify the needed materials. Current machine learning algorithms fail to find new materials with properties outside of the original boundary of the training data because of their limited extrapolative abilities. One way to solve this is to generate millions of new hypothetical materials using generative models [101]. We develop deep generative

models that incorporate explicit or implicit rules of crystals. Trained with materials from the existing databases, our generative models can generate materials that are beyond the space of training data. Material prediction models quickly screen materials in novel space and find potentially desired ones.

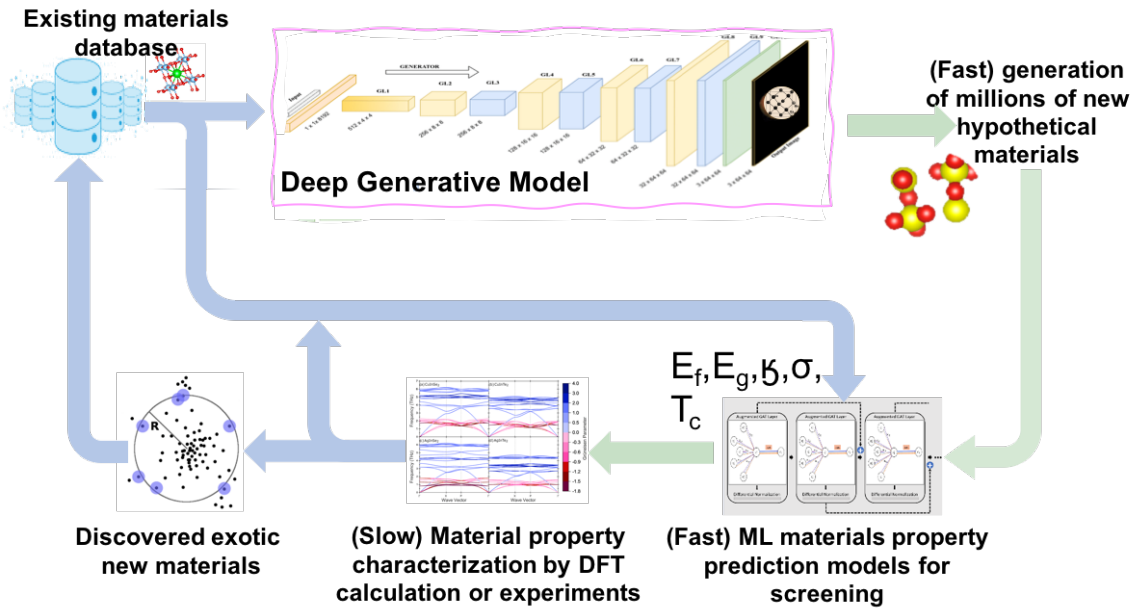


Figure 2.5 The overall framework of our proposed system.

Chapter 3

High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks

3.1 Introduction

Data-driven accelerated design of new materials is emerging as one of the most promising approaches for addressing the challenges in finding next-generation materials. Currently, one of the main strategies for materials discovery is screening existing materials databases [173, 137, 63, 143]. However, such approaches are severely limited by the scale and diversity of the existing structures in the repositories, such as ICSD and Materials Project (MP), which have about 165,000 and 125,000 materials, respectively, compared to the almost infinite chemical design space. For example, lithium compounds are widely used in electric vehicles and mobile phone batteries, but there are only 16,000 different lithium compounds in the MP database, which has been almost exhaustively screened for better lithium-ion battery[83, 103]. Large-scale generation of stable hypothetical crystal structures is strongly needed to significantly expand the current materials repositories in both the quantity and compositional and structural diversity to increase the success rate of high-throughput screening of novel functional materials.

The properties of materials are closely linked to their crystal structures. Traditionally, materials scientists discover new materials by either trial-and-error or heuristic random-guess approaches, both of which are notoriously labor-intensive. One example database is Inorganic Crystal Structure Database (ICSD) [9], which collects almost all discovered materials since 1913. To date, only around 165,000 experimental structures are reported in ICSD. Considering the number of elements in the periodic table and their possible combinations, the design space of materials would be infinite combinatorially. Hence, better approaches for new materials discovery is needed.

Several working directions are investigated for the generation of new materials [106, 109, 163, 36, 27, 85, 127, 101, 64]. There are mainly three different ways to generate or discover new crystal structures including doping/element substitution[45, 83, 139,

142], composition generation plus crystal structure prediction[109], and generative machine learning models[27, 102, 127, 61, 101, 64]. The element substitution approach is the most widely used strategy. But it is subject to the extremely limited known prototype structures in the database compared to the vast chemical design space. The second approach can exploit the recently developed generative models [27] to generate a large number of hypothetical materials compositions and then use crystal structure prediction codes to predict their structures. Many global optimization methods have been developed to search the appropriate compositions and structures, including simulated annealing [172], basin hopping [161], minima hopping [38], genetic and evolutionary algorithms [163, 36]. Those approaches generally guide the searches towards the local minima of free energy to identify the stable or meta-stable structures either by initial configuration space or chemical composition. However, these crystal structure prediction algorithms are usually too computationally expensive due to their reliance on DFT-based formation energy calculation and can thus only handle relatively simple structures. For complex structures, most of the time, these methods fail to find the ground truth structures corresponding to the global minimum formation energy.

One of the most promising approaches for new materials structure creation is deep generative machine learning models [27, 127, 101, 64, 105, 24, 85, 71]. Both variational autoencoder (VAE) [101, 127, 24, 71] and generative adversarial networks (GAN) [27, 85, 64, 105] have been adapted for inverse design of inorganic materials with different crystal structure representations. A VAE model contains two parts: an encoder and a decoder [49, 68, 31]. The encoder part encodes the crystal structure distribution into a latent space, and the decoder reconstructs the material structures from the latent space. After training, new material structures can be generated by sampling in the latent space. Conversely, a GAN generator model consists of two neural networks: a discriminator (critic) and a generator, both of which are trained simultaneously. The

discriminator is trained to differentiate real materials from fake ones generated by the generator, while the generator tries to generate fake materials as real as possible to fool the discriminator. The nash equilibrium achieved by the discriminator and the generator helps a GAN learn the distribution the materials implicitly. In the past few years, several inorganic materials generative models have been proposed. Those works are limited by their chemical family (e.g. special oxides) [101, 64, 85] or formulas generation [27] or hydrides [105]. Noh et al. [101] present a framework for learning a continuous vector for vanadium oxides using VAE, which is trained on a 3D image-like representation to attain the continuous materials space. Two sampling strategies are used in the latent space to generate only V_xO_y materials. Training the VAE model using 3D grid representation is computationally demanding and memory-hungry. In [64], Kim and Noh et al. trained a composition family-specific GAN model on the Mg-Mn-O system using the atom coordinates as the representation of materials. The crystal GAN model is composed of three modules: a generator, a critic (discriminator), and a classifier. The critic calculates the Wasserstein distance between real and fake materials [5]. The classifier module ensures that the generator generates desired composition and atom numbers in the unit cell. However, this model can only be used to generate structures of the Mg-MN-O system, and the model quality is limited by the small dataset since there are only limited known compounds of this chemical system. CrystalGAN [105], proposed by Nouira et al., consists of a cross-domain GAN model, which maps one hydride system into another using CycleGAN schema [193]. All these works focus on generating materials of a special material system. In a most recent work, Ren et al.[127] proposed a new VAE model that directly uses the atom coordinates and unit cell lattice parameters to encode the structures. To constrain the neural network model behavior, their invertible representation encodes the crystallographic information into the descriptors in both real space and reciprocal Fourier space crystal properties. Their model is trained

with 24,785 unique ternary materials and can generate interesting new structures. However, most of their new structures are generated by perturbing the latent vectors of known materials. Large-scale generation of stable crystal structures remains a challenging problem. Other than generating materials structures, Dan et.al. [27] proposed MatGAN to generate millions of novel materials formulas with chemical validity, which expands the candidates for inverse design of new solid materials.

In this chapter, we propose a novel deep generative model called CubicGAN to generate cubic materials structures on a large scale. Ternary materials selected from the OQMD [69] database are chosen as our training set because of its large size of materials and diverse compositions. In our model, material structures are represented by their lattice parameters, atom coordinates, element embedding, and the space group. The conditions of a specific space group and three elements are fed to the generator to generate desired crystal material structures. We trained ternary and a quarternary GAN models to generate novel cubic (ternary and quarternary) crystal structures of the space groups 216,255,221. Materials of these three space groups consist of 78.5% of all ternary and quarternary cubic materials in OQMD, covering a majority of known cubic materials space.

Our systematic experiments show that our CubicGAN model can recover not only many of the known cubic structures but also discover many new materials with new composition prototypes with different anonymous formulas (new prototypes). Additional large-scale DFT based validation has led to the discovery of 506 new cubic crystal materials of new prototypes. The detail of the CubicGAN model will be explained in the following sections. Compared to [101, 64, 105], our framework can generate a large variety of materials of different chemical systems. The only work that is similar to ours in terms of variety of materials is [127], in which Ren et al. use VAE rather than GAN as the generative model trained with train ternary materials in Materials Project [54] database. However, their model tends to generate new samples

by interpolation. The second major difference is a much simpler representation is used in our work without the momentum space representations.

Our contributions can be summarized as follows:

- We propose a novel GAN model to generate large-scale cubic materials conditioning on the elements and a specified spacegroup. In total, we generate 10 million hypothetical ternary and 10 million quaternary crystal structures for downstream analysis.
- We perform three stage checks on generated materials and extensively match the generated materials against existing databases. The results show that our method can rediscover a majority of cubic materials in the existing databases. In addition, most of the rediscovered materials from MP are confirmed as stable or meta-stable materials in terms of energy-above-hull.
- We perform DFT simulations on 108,897 hypothetical materials, of which 33.8% novel materials are successfully relaxed. By further analysis, we demonstrate that new crystal structure prototypes (with different anonymized formula types) can be found, such as ABC_6-216, ABC_6D_6-216, and AB_8C_{12}-221.
- By further stability verification, 506 new-prototype materials have been generated and confirmed to be stable by phonon dispersion calculation.

3.2 Related Work

VAE based Material Generation VAE [68] composes of two deep neural networks, an encoder and a decoder. The encoder is trained to encode the materials into latent vectors and the decoder reconstructs the materials from the latent vectors, then use different strategies to generate/reconstruct materials by sampling latent vectors. iMatGen [101] is believed to be the first work that uses VAE to realize the inverse design

of solid materials. iMatGen encodes unit cells into 3D grid based representations, and spherical linear interpolation and Gaussian random sampling are used to sample from the latent space to generate new materials. Hoffmann *et al.* [51] extend iMatGen by combining a UNet part to segment reconstructed 3D voxel images into atoms and assign atoms with a number. Base on iMatGen and Hoffmann *et al.*, ICSG3D [24] integrates formation energy per atom into 3D voxelized solid crystals and enables the VAE to encode materials and energy simultaneously, which makes it possible to generate materials subject to user-define formation condition. Another approach to represent 3D crystals is to encode 2D crystallographic representations as the combination of real space and reciprocal-space Fourier-transformed features [127]. In CDVAE [177], authors propose to generate materials in a diffusion process in the decoder. The diffusion process moves atoms into positions in the lower energy space to generate stable crystals.

GANs based Material Generation GANs [43] also consist of two deep neural networks, a generator and a discriminator (critic). The generator generates fake materials with inputs of random vectors sometimes conditioning on elements and space groups while the discriminator tries to tell real materials from generated ones. With learnt knowledge of forming crystals, the generator can directly create new materials. The first method to generate materials using GAN is CrystalGAN [104], which leverages a CycleGAN [193] to generate new ternary materials from existing binaries. However, it remains uncertain whether CrystalGAN can be extended to produce more complex crystals. Both GANCSP [64] and CubicGAN [189] use a "point cloud" (containing fractional coordinates, element properties, and lattice parameters) as inputs to build a model that generates crystals conditioned on composition or both composition and space group. The difference between them is GANCSP can only generate structures of the Mg-Mn-O system but CubicGAN can generate more

diverse systems under three space groups. In CCDCGAN [86], the authors use 3D voxelized crystals as inputs for an autoencoder. The autoencoder converts them to 2D crystal graphs, which is used as the inputs to the GAN model. A formation energy based constraint module is trained with the discriminator. This module automatically guides the searching for local minima in the latent space. Less related works include MatGAN [27] and CondGAN(x^{bp}) [134] developed for generating only chemical formulas.

3.3 Methods

In this work, we focus on training generators of ternary and quarternary cubic crystal structures of three space groups (216, 221, 225) to simplify our model design while ensuring coverage of a majority of the cubic design space. We find that in the OQMD dataset with 813,839 materials, 85.8% of them are ternary or quaternary materials. In addition, out of all the cubic crystals, 97.8% of them belong to these three space groups, again covering the majority of the known cubic materials space. These three space groups are selected because we find that for the materials of these three space groups, most of their nonequivalent atom fractional coordinates in the CIF files have a multiplicative factor of 0.25 or belong to this set [0, 0.25, 0.5, 0.75]. So, instead of generating cubic structures with arbitrary real-valued atom coordinates, we only aim to train a cubic material generator that only generates structures whose atom positions are sitting at positions with their fractional coordinate values to be from this set +/-0, 0.25, 0.5, 0.75. In this case, the special discrete fractional coordinates are much easier to generate accurately by our deep neural networks. This decision has dramatically simplified our generation model, and thus we choose the training data with these two criteria:ternary and quaternary cubic crystal structures of three space groups (216, 221, 225).

3.3.1 Dataset

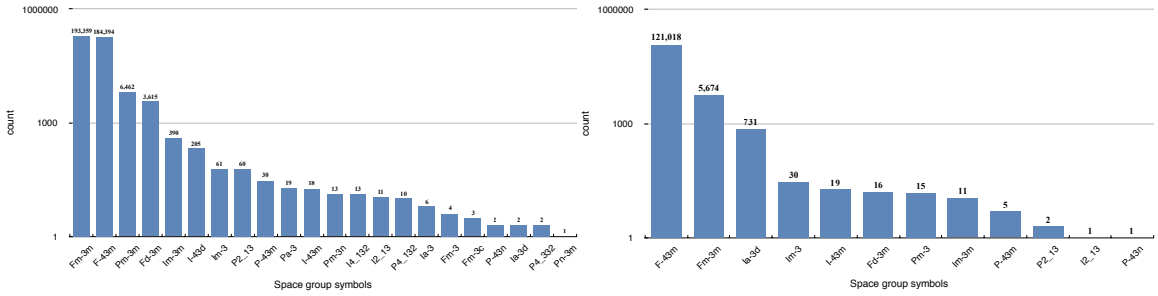
We collect the training data from OQMD [132, 69], which is an open-source database of experimental and DFT-calculated materials. Totally 813839 entries are retrieved from version 1.3 of OQMD. Entries calculated with local-density approximation (LDA) are also included. Among them, we successfully build 556,839 and 141,100 POSCAR files for ternary and quaternary materials in the OQMD, of which 505,456 and 127,659 structures belong to cubic crystal systems, respectively. After converting the POSCAR files to symmetrized CIF files, 411,646 ternary materials have three unique nonequivalent atom sites, of which 388,680 materials of cubic crystal systems are found; 129,514 quaternary materials have four unique nonequivalent atom sites, of which 127,523 materials belong to the cubic crystal systems. Table 3.1 shows the statistics of OQMD materials distributions. We can find that ternary materials of cubic crystal systems are the largest chuck (91%) out of all ternary materials. Similarly, it is observed that the ternary cubic structures with 3 nonequivalent sites are 94% out of all ternary materials with 3 nonequivalent sites. For quaternary materials, these two percentages are 90% and 98%, respectively. This means that our CubicGAN model can be used to generate hypothetical cubic materials that are the majority type of known material category.

Table 3.1 Statistics of OQMD ternary and quarternary materials (Total 813,839)

Type	Ternary	Ternary cubic
Count	556,839	505,456
Cubic Percentage	505456/556839=91%	
Type	Ternary with 3 nonequivalent sites	Ternary cubic with 3 nonequivalent sites
Count	411,646	388,680
Cubic Percentage	388,680/411,646=94%	

Type	Quaternary	Quaternary cubic
Count	141,100	127,659
Cubic Percentage	127659/141100=90%	
Type	Quaternary with 4 nonequivalent sites	Quaternary cubic with 4 nonequivalent sites
Count	129,514	127,523
Cubic Percentage	127,523/129,514=98%	

Another key criterion for selecting our training samples is that we only pick cubic structures with three nonequivalent atom positions (in CIFs) for training ternary GAN model (for quarternary GAN, the number is 4). Making this choice allows us to use a unified matrix of dimension (28×3) to represent all ternary cubic materials (for quarternary materials, the dimension is 27×3 where only one space group is used in this work). For a given material, once we have its nonequivalent positions and space group, the full atom positions within the unit cell can be converted to conventional atom positions by symmetry operations. We have identified 411,646 ternary materials with only three nonequivalent positions, of which 388,680 (94%) materials belong to cubic crystal systems as shown in Table 3.1. Out of these 388,680 materials, 22 space groups are found as shown in Figure 3.1a,. Among them, the space groups that have the most numbers of materials are $Fm\bar{3}m$ and $F\bar{4}3m$ (the total portion of these two space groups is 97.2%). $Pm\bar{3}m$ is the third one with only 6,462 samples or 1.7%. After removing the duplicate cubic materials within MP and ICSD, almost all the materials (375,749 out of 384,215) with the space groups of $Fm\bar{3}m$, $F\bar{4}3m$ and $Pm\bar{3}m$ follow this criterion.



(a) Ternary (Total is 388680) (b) Quaternary (Total is 127523)

Figure 3.1 Distribution of each space group of in Ternary and Quaternary Cubic systems. The bars' height is at logarithmic scale of real values.

Table 3.2 shows the overall statistics of our finalized training and validation datasets. In total, we have selected 375,749 ternary materials from three cubic system space groups from OQMD to form the OQMD-TC3 (T:Ternary, C-Cubic, 3-three space groups) training dataset: $Fm\bar{3}m$, $F\bar{4}3m$, and $Pm\bar{3}m$ each having 186,344, 184,162 and 5,243 materials respectively. These materials together correspond to 249,646 unique formulas. With this diversity of formulas, our CubicGAN model can efficiently learn valid combinations of ternary elements. The unique 84 elements in the datasets are utilized to generate random three-element combinations during GAN training. The same steps are applied to quaternary materials in OQMD. As shown in Figure 3.1b, materials with space group $F\bar{4}3m$ occupies 95% of the quaternary data. So for training the quaternary GAN model, we only choose materials of space group $F\bar{4}3m$.

We will use the ternary data from the Materials Project and ICSD databases as validation sets to check the rediscovery rates for our proposed method. We first process the ternary materials in Materials Project database [54] and ICSD [9] as we do for OQMD samples to create the MP-TC3 and ICSD-TC3 validation datasets. In total, 6,545 cubic materials are retrieved, of which the numbers of materials with $Fm\bar{3}m$, $F\bar{4}3m$ and $Pm\bar{3}m$ are 4576, 520, and 1449, respectively and there are 6,431 unique formulas existing in the whole retrieved data. From the ICSD database, 1,875 cubic materials are found to satisfy our seleciton criteria, of which the numbers of materials

are 804, 280, and 791 forf space groups $Fm\bar{3}m$, $F\bar{4}3m$ and $Pm\bar{3}m$.. For quaternary materials, the OQMD-QC1 training dataset has 121,008 samples. However, only 39 and 8 quaternary materials are found in MP and ICSD that satisfy our two selection criteria (See Table 3.2). Here the cubic materials (6,545+1,875) from MP/ICSD are used as validation set and are excluded from the training set. We have removed these samples from our training dataset selected from OQMD by removing the crystal structures with a minor difference of cube lengths from the samples in the validation sets).

Table 3.2 Statistics of our training data and validation datasets from OQMD, MP and ICSD.

Ternary Materials

Dataset	Total	$Fm\bar{3}m$	$F\bar{4}3m$	$Pm\bar{3}m$	Unique formula	Unique element
Training:OQMD-TC3	375,749	186,344	184,162	5,243	249,646	84
Validation:MP-TC3	6,545	4,576	520	1,449	6,431	84
Validation:ICSD-TC3	1,875	804	280	791	1,034	84

Quaternary Materials

Dataset	Total	Unique formula	Unique element
Training:OQMD-QC1	121,008	39,767	56
Validation:MP-QC1	39	39	39
Validation:ICSD-QC1	8	7	12

Table 3.3 Crystal prototypes existent in our training and validation sets (OQMD-TC3, MP-TC3 and ICSD-TC3) for ternary CubicGAN. There are only 8 different prototypes.

	ABC_2-225	ABC-216	ABC_3-221	AB_2C_6-225
OQMD-TC3	185170	184162	5237	1166
MP-TC3	4343	520	1410	196
ICSD-TC3	551	280	759	233
	ABC_6-225	AB_3C_3-221	AB_3C_8-221	AB_6C_6-225
OQMD-TC3	8	4	2	0
MP-TC3	36	16	23	1
ICSD-TC3	20	6	26	0

In terms of prototypes in the validation datasets MP-TC3 and ICSD-TC3, Table 3.3

shows details of the existing prototypes for materials that satisfy our selection criteria. We take the prototype "ABC2-225" as an example. Here ABC2 and 225 are the crystal prototype anonymous formula and the space group number used to denote a prototype, and we will use this format in the following content. Overall, the three databases have the same set of prototypes; other than that, MP has an extra one: AB6C6-225. However, only one material (mp-1147668) is found under AB6C6-225 and is unstable. For quaternary materials in OQMD, there are only two prototypes, including ABCD-216, with 121,006 materials and $ABCD_6$-216 with two materials. Moreover, we find that quaternary cubic materials distribution is highly biased with 121,018 belonging to space group 216, and only 5674 belonging to space group 225, and no samples found for space group 221. For simplicity, we train the quaternary CubicGAN using only the samples from space group 216 and it then can only generate samples of this space group.

3.3.2 CubicGAN Framework

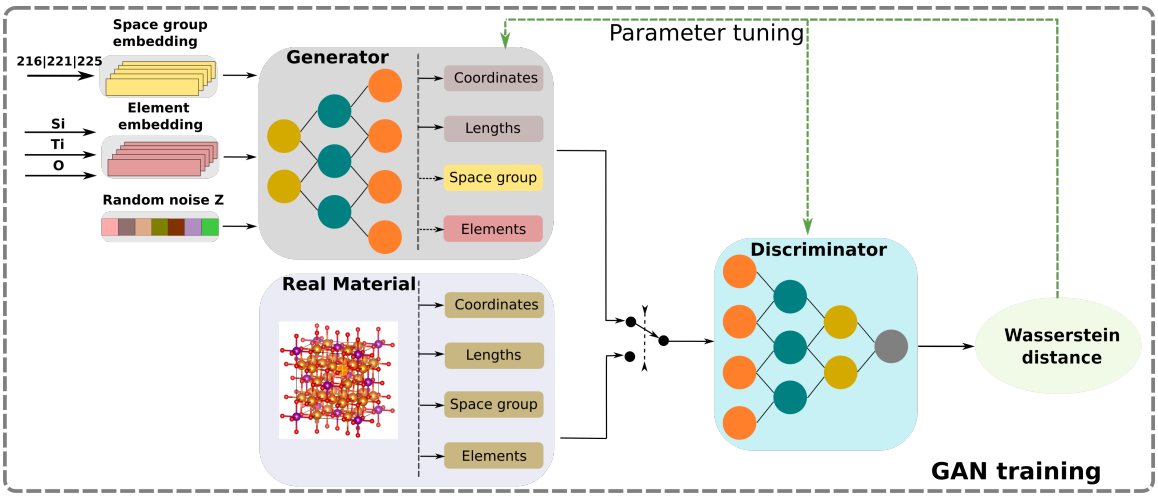


Figure 3.2 The workflow of our CubicGAN framework

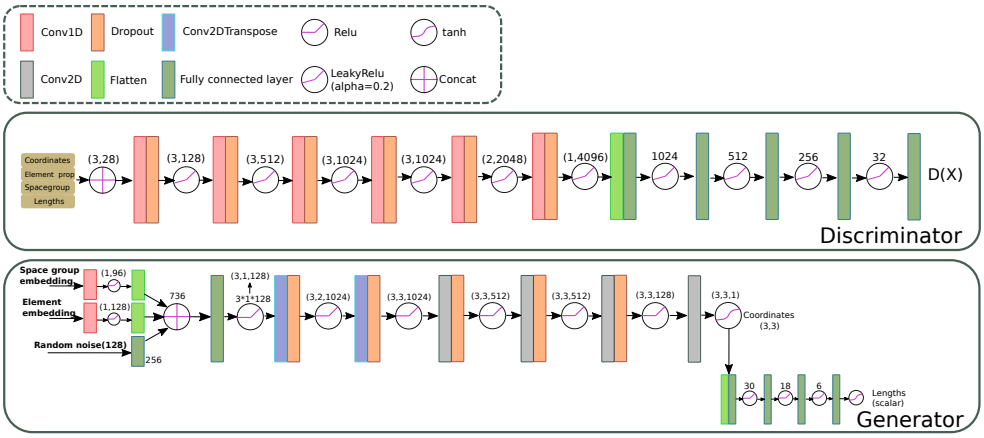


Figure 3.3 The detailed architecture of the generator and the discriminator of the CubicGAN framework.

Figure 3.2 illustrates the main framework of our method. The framework primarily contains two steps: GAN training and material generation. Our goal is to train a generator that learns the distribution from known materials data and then sample from it. To achieve this, the generator is trained to create fake material structures, conditioned on a given space group and a specification of three elements.The three elements are randomly chosen from 84 elements in the dataset. The 84 elements are one-hot encoded and are converted to a 3×23 element matrix by the embedding layer. There are several ways to encode elements such as learnt atom embedding [176, 18], atom2vec encoding [192], atom embedding learnt from research literature [155], and primitive element properties. In this work, we choose to use primitive elemental properties as the element embedding approach and use them to set the weights in the element embedding layer. The reason is that CubicGAN learns to build valid crystal structures from three elements and their properties, coordinates, and spacegroup. With raw elemental properties, it is beneficial for CubicGAN to find the most desirable combinations of elements, coordinates and space groups. The parameters of the embedding layer are initialized by 23 element properties as shown in Table 3.4. Taking a randomly selected space group (one-hot encoded), 3-element combinations (one-hot encoded), and random noise Z as inputs, the generator then generates material

structures with the specified space group and element constituents. Space groups and elements are mapped into dense vectors by their corresponding embedding layers. The number of atoms for each element does not need to be specified as it can be determined by the space group symmetry operations. The random selections of space groups are based on the portions of three cubic space groups considered in our model: $Fm\bar{3}m$, $F\bar{4}3m$ and $Pm\bar{3}m$. The detailed architecture is shown in Figure 3.3. The discriminator consists of 6 convolutional layers followed by Dropout layer and 5 fully connected layers. The inputs with shape (3×28) to discriminator is composed of four parts: atom coordinates and corresponding element properties, spacegroup encoding and lattice length. The inputs to Generator consist of spacegroup and element embedding and random noise vector. They are concatenated and de-convolutional layers map them to non-equivalent atom coordinates and lattice parameters.

Table 3.4 23 element properties used for element embedding in CubicGAN

Properties	**Properties**
Atomic number	Average ionic radius
Pauling electronegativity	Average cationic radius
Periodic table row	Average anionic radius
Periodic table group	Sum of all ionic radii
Atomic mass	Maximum oxidation state
Atomic radius	Minimum oxidation state
Mendeleev number	Average all common oxidation states
Molar volume	Average all known oxidation states
noble gas or not	transition metal or not
post transition metal or not	metalloid or not
alkali or not	alkaline or not
halogen or not	

An input to the discriminator has four parts: nonequivalent atomic coordinates, element properties, unit cell parameters, and space groups as shown in Figure 3.2. The coordinates part includes the fractional coordinates of three nonequivalent atoms. For three unique elements in each material, each element is represented by 23 properties as shown in the Table 3.4. Since the lattice lengths a, b, c are the same in cubic

crystals, we only need to use one value to represent it. Three cubic space groups are one-hot encoded. As shown in Figure 3.3, four parts are concatenated together to form a tensor with the dimension of 3×28. The input is then forwarded to four 1D convolutional layers, of which the kernel size is 1×1, which is used to capture the implicit relationships among the four parts. We use two CNN layers to reduce the dimension from three to one. Then, a few fully connected layers are used to map them to Wasserstein estimation [5]. The detailed network settings are shown in Figure 3.3. In standard conditional GAN, the input of the generator includes the random noise and a condition vector [95]. Here, we add a space group embedding layer and an element embedding layer as shown in Figure3.2 to map the randomly selected one-hot encoded space group (chosen from 216/221/225) and three randomly selected elements (one-hot encoded) into the latent vectors. The reasons for this design are as follows: 1) As only three dominant cubic space groups are used in this work, the combination of atom positions with corresponding elements, unit cell lengths, and one-hot encoded space group symmetry is sufficient to describe a material structure; 2) Using element properties as part of the representation makes the generator learn to generate chemically valid materials, e.g., structures that do not violate Pauling's rules. As our previous work [27] shows, the composition constraints can be learned from the compositions of existing materials. Here, our CubicGAN is also configured to learn both implicit compositional as well as structural constraints to help the generator generate only valid ternary or quaternary formulas as much as possible; 3) Our 2D representations of the cubic structures also matches well with the convolutional layers used in the discriminator, in which the convolutional operations can extract implicit relationships among four parts of information.

The generator and the discriminator of the CubicGAN model are trained with the loss function of Wasserstein distance [5] which measures the dissimilarity between distribution differences of real and fake materials. Compared to loss functions used in

traditional GAN [43], Wasserstein distance improves the model stability and prevents the mode collapse. We use the gradient penalty to clip weights in order to improve the stability of training as done by Gulrajani et al. [44]. The penalty of gradient norm with respect to the inputs works as a regularization term to stabilize the training process of the GAN. More formally, our cost function for GAN training is as follows:

$$L = \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\sim\mathbb{P}_g}[D(\tilde{\mathbf{x}})] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}_r}[D(\mathbf{x})] + \lambda \mathop{E}_{\hat{\mathbf{x}}\sim\mathbb{P}_{\hat{\mathbf{x}}}}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2] \tag{3.1}$$

where D is the discriminator, $\mathbb{P}_{\hat{\mathbf{x}}}$ is the distribution of interpolated samples between the distribution of real materials $\mathbb{P}_r$, and the distribution of generated materials $\mathbb{P}_g$. λ is the balancing parameter, which is set to 10 in this work.

After inspecting the generated structures by the GAN, we find that the generated lattice parameter a is often not good enough, leading to overlapping atom clusters. To address this issue, an additional post-processing step introduced to predict the lattice length a using a composition based machine learning model that we recently developed [80], which achieves a R^2 score of 0.979 for cubic lattice a prediction.

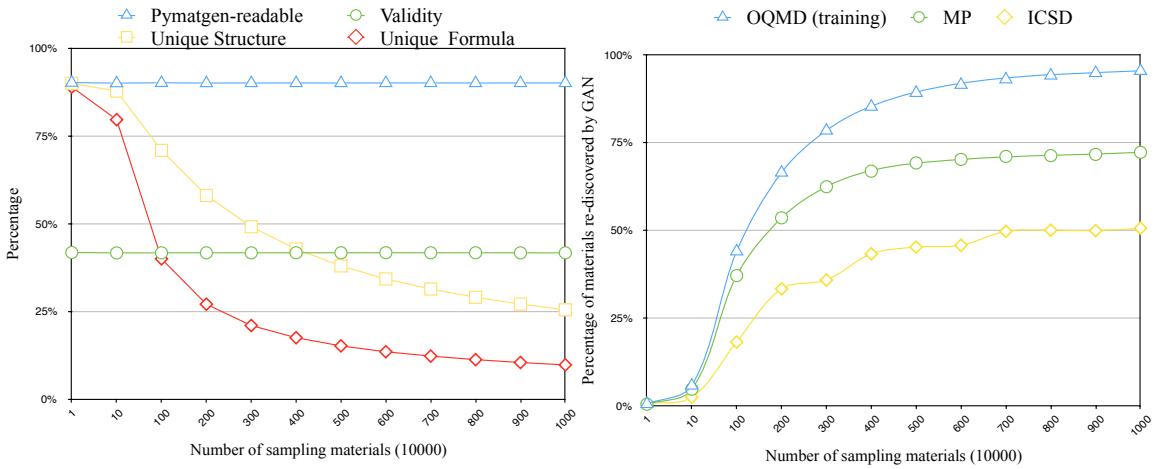
During training the GAN, real materials are randomly picked in batches. With the fused matrix of generated materials as shown in Figure 3.2, they are fed to the discriminator in a mixed manner. We set the number of iterations of the discriminator per generator iteration as 5. The GAN model is developed using the open-source libraries of TensorFlow [1] and Keras [22]. More details regarding model architecture and hyper-parameter setting can be found in Table 3.5.

Table 3.5 Hyper-parameters for training the CubicGAN Model

Hyper-parameter		Value
batch size		256
Adam optimizer	learning rate	0.00001
	β_1	0.5
	β_2	0.9
gradient penalty coefficient		10
the number of iterations of discriminator per generator iteration		5

3.4 Results and Discussion

3.4.1 Performance evaluation of CubicGAN: Validity, Uniqueness, and Rediscovery rate Analysis



(a) Validity, structural and compositional uniqueness of CubicGAN in terms of number of samplings.

(b) Rediscover rates of CubicGAN in term of the number of sampled materials.

Figure 3.4 Performance evaluation of CubicGAN

There are three major criteria for evaluating generative models, namely, validity, diversity, and uniqueness [133]. After training the ternary CubicGAN using the OQMD-TC3 dataset, we generate 10 million cubic structures of the specified three cubic space groups (225,216,221). The proportions of the samplings are set as identical to the training set, which is 49.6%-49.0%-1.4% respectively. To evaluate the generation performance, we first check how the percentage of the generated charge-neutral samples changes with respect to the total number of generated samples. The charge neutrality check is based on Pymatgen [111] using the common valence values of elements as defined in Pymatgen. As shown in Figure3.4a, the charge-neutral samples' percentage maintains around 41% over the whole process of generating 10 million samples, which means that when we generate 10 million samples, appropriately we can get 4.1 million charge-neutral samples for downstream screening. We then checked how the percentage

of the generated samples have pymatgen-readable CIFs (Crystallographic Information File), unique CIFs, and unique formulas, which reflect the diversity and uniqueness of the generator. In Figure 3.4a, the blue line demonstrates the percentage of cifs readable by pymatgen in terms of sampling size, and the sampling size is from ten thousand to ten million. In this work, pymatgen-readable means that CIFs can be recognized as the space group that is assigned to. We can find that the percentage of readable CIF files is stable no matter how we run the sampling. After removing the duplicate materials, we calculate the percentage of unique CIFs and unique formulas as denoted by the yellow and red lines in Figure 3.4a. Only those materials that have the same formula and the same corresponding atom positions are considered as duplicates here. It is found that the percentages of the unique CIFs and unique formulas are decreasing and growing flat. From these observations, we believe that our GAN model might have explored the majority of the cubic crystal structure space but have not exhausted it yet.

Another effective way to evaluate the CubicGAN's performance is to check how soon it can rediscover the known cubic crystals in leave-out datasets of existing databases. To do this, in our training dataset, we have removed all the materials of the three cubic space groups (216,225,221) existing in MP and ICSD databases, which are 6,545 and 1,875, respectively. It is interesting to see how many of those leave-out cubic materials can be rediscovered by our GAN model as the sampling size goes from ten thousand to ten million. Figure 3.4b shows how the rediscovery percentages of the cubic crystals of the three space groups (216,225,221) change as the sampling size increases.

Figure 3.4b shows the rediscovery rates over time of sampling. At first, we check how the percentage of the rediscovered cubic samples out of all training samples (blue line) changes while generating more samples. It is found that this training set rediscovery rate increases consistently over the sampling process. It soars quickly to

88% when the sampling size increases until 5 million samplings are reached. At the end of 10 million samplings, the rediscovery rate reaches 95.5%. Similar patterns can be observed for the rediscovery rate curve for the MP-TC3 validation dataset, as shown by the green line. With the increasing number of samplings, the rediscovery rate reaches 72.0%. This saturated percentage is much lower than that of the training set, which is due to MP-TC3 data has different proportions over the three space groups (225,216,221), which are 69.9%-7.9%-22.1% respectively compared to 49.6%-49.0%-1.4% of the training set. Since our generation process is based on the space group proportions of the training set (which focuses on generating candidates of space groups 225 and 216, the 72% rediscovery rate is close to the percentage of these two types of samples in MP-TC3 (69.9%+7.9%=77.8%). We also find that half of the rediscovered materials in MP-TC3 are stable based on the formation energy and e-above-hull criteria. Details of the stability analysis can be found in Figure 3.5. When sampling ten million materials, 4731 and 950 materials are re-discovered in MP and ICSD, respectively out of all ternary cubic materials of the space groups 216/225/221 which are 6545 and 1875 for MP and ICSD. For 4731 materials in MP, we downloaded the formation energy per atom and e-above-hull energy from Materials Project [54]. We find most of these 4731 materials have negative formation energy per atom and energy-above-hull equal to zero (2521) or close to 0 eV (4346 materials' e-above-hull is below 0.2 eV) which means that most materials we re-discovered from MP are stable or meta-stable. Materials in ICSD are mostly synthesizable and experimentally determined [9]. The re-discovered materials by our method demonstrate that our method could produce stable materials that are potentially synthesizable. The rediscovery rate pattern over ICSD-TC3 is similar to that of MP-TC3 except that the highest rediscovery rate is 50.7% at the sampling size of ten million, which is close to the percentage of total samples of space groups 225 and 216 (42.9%+14.9%=56.8%). These high rediscovery rates over the training set and the two validation sets demonstrate that our CubicGAN

has learned the implicit chemical rules of the cubic structures to generate in a much better way than random sampling. After the sampling size reaches 7 million, the number of materials rediscovered converges, indicating that ten million samplings could be a reasonable size to cover most of the cubic structures since they seem to have almost exhaustively explored the search space of materials that meet our criteria. Therefore, we use ten million samplings for further analysis.

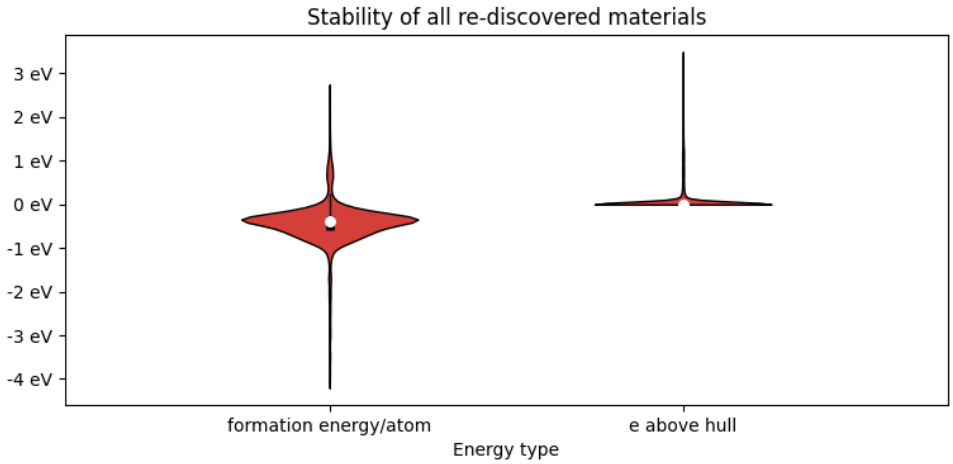


Figure 3.5 Distribution of the formation energy and e-above-hull of re-discovered materials in Materials Project.

To compare how our CubicGAN performs compared to random sampling or exhaustive enumeration, we calculate the enrichment score for our ternary CubicGAN. As we are searching candidates of three cubic space groups with three unique sites of three distinct elements and the only possible fraction coordinates are 0,0.25,0.5,0.75, the total possibility of configurations are $(4^3)^3 * 85 * 84 * 83 * 3 = 466,055,331,840$, which is much larger than the corresponding combinations of the ternary composition space [27]. Considering that with 10 million samplings, we have rediscovered 95.5% of the OQMD-TC3 dataset, the enrichment score is approximately 44,507, which is a significant boosting for generating chemically valid crystal structures compared to exhaustive enumeration.

3.4.2 Large scale generation of new cubic crystal materials structures

Table 3.6 Statistics of generated materials

	valid CIFs	unique formulas	crystal prototypes
Ternary	2,558,678	990,319	31 (24)
Quaternary	5,498,267	1,797,592	3 (1)
No Lanthanoid and Actinoid			
Ternary	1,064,650	403,337	31 (24)
Quaternary	4,382,130	1,431,500	3 (1)

While rediscovery rate analysis over the MP-TC3 and ICSD-TC3 validation sets have demonstrated the accelerated sampling in cubic structure space, there are only 6,545+1,875=8,420 validation samples plus the 358,840 rediscovered training samples. It is still desirable to check the chemical validity of the remaining 96.33% generated samples and filter out those promising new materials. With 10 million hypothetical cubic materials, it is impractical to perform DFT calculations for all of them to verify their chemical validity and stability. Here we adopt three stages of validation check to reduce the pool of samples for DFT validation. We use the CGCNN based graph neural network model for formation energy prediction, which was trained with samples from Materials Project database[176]. Then we scan the generated materials in the order of space group match, charge neutrality, and formation energy filtering. The nonequivalent coordinates are transformed by symmetry operations provided by relevant space groups used when generating samples. With the full coordinates set, elements, unit cell parameters (unit cell length a and angles, which are always 90 degrees in cubic systems), we could write a Crystallographic Information File. The space group check is performed by Pymatgen [111] in the first place (we refer to this check as a Pymatgen-recognizable check). If the generated sample cannot be recognized by Pymatgen or the space group analyzed by Pymatgen is not consistent with the space group given to the generated sample, this sample is considered as a failed generated case. As shown in Table 3.6, in total there are 2,558,678 and 5,498,267

valid ternary and quaternary CIFs have been found from 10 million generations, respectively. From them, candidate materials with charge neutrality and CGCNN-predicted negative formation energy are reserved for further DFT calculations based verification.

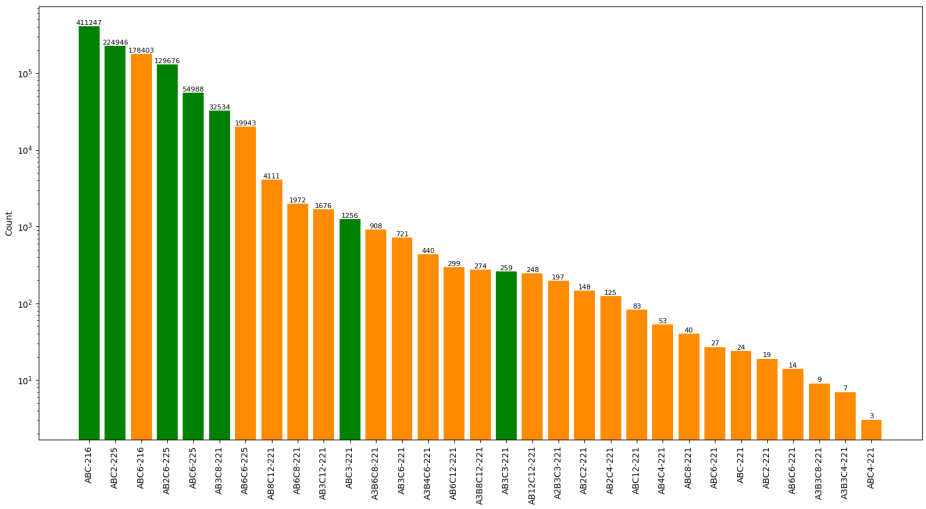


Figure 3.6 Distribution of the prototypes of generated materials after removing Lanthanoid and Actinoid.

A major evaluation of our CubicGAN model is to check whether it can generate new cubic materials with novel prototypes, which are represented by distinct anonymized formulas in Pymatgen. As shown in Table 3.6, we find that 24 and 1 novel prototypes for ternary and quaternary materials, respectively, have been found in our generated samples that are not existent in our training data. For relieving the burden of DFT calculations, we choose to remove the samples that contain Lanthanoid and Actinoid elements. In total, 1,064,650 ternary materials are left, of which 209,744 materials are of new crystal prototypes. The distribution of prototypes for 1,064,650 materials is shown in Figure 3.6. Green bars show known prototypes in the training data and orange bars show the number materials of new prototypes. The figure shows that our model has generated many new prototypes. The bars' height is at logarithmic

scale of real values. Similarly, 4,382,130 quaternary materials are left after removing Lanthanoid and Actinoid elements, of which 260,891 materials are of the new crystal prototype (the prototype ABC_6D_6-216). Since only two $ABCD_6$-216 materials exist in the quaternary training dataset OQMD-QC1, we also include $ABCD_6$-216 materials for the downstream DFT analysis considering the huge number of generated $ABCD_6$-216 samples (1,655,407). After searching thoroughly in databases of OQMD, MP, and ICSD, only a limited number of materials with $ABCD_6$-216 are found, as shown in Table 3.7. Then, we perform charge neutrality check by Pymatgen and CGCNN formation energy filtering on 209,744 ternary materials and 1,916,298 (260,891 + 1,655,407) quaternary materials. While each material might have different atom arrangements in the unit cell that maps to the same space group, in this work, we only choose one of them for DFT calculations. Finally, 17,303 ternary materials and 91,594 quaternary materials are left for DFT optimization. In total, 36847 candidate materials have been relaxed successfully with 14,433 ternary and 22,414 quaternary samples.

Table 3.7 Existing $ABCD_6$-216 materials in databases OQMD, MP, and ICSD

Database-ID	Formula
oqmd-24074	$BaNaH_6Ir$
oqmd-24073	$NaCaH_6Ir$
icsd-262196	$BaNaH_6Ir$
icsd-51205	$NbWNO_6$
icsd-262197	$BaNaH_6Ir$
icsd-96973	$NbWNO_6$
icsd-262195	$NaCaH_6Ir$
mp-1223322	$KRbMnF_6$
mp-1182061	$BaNaH_6Ir$
mvc-14934	$CaFeWO_6$
mp-1227207	$CaEuH_6Ru$
mp-1228944	$CsRbMnF_6$
mp-1180133	$NaCaH_6Ir$

3.4.3 Discovery of 506 new-prototype stable materials verified with DFT calculations

After filtering down materials with novel prototypes, we perform DFT optimization on materials with CGCNN-predicted negative formation energy, and we use Γ points and mechanic constants to further scale down the successfully relaxed structures. Phonon dispersion is the eventual criterion to determine the stability of structures.

Gamma points and mechanic constants filtering The vibrational frequencies at the Γ point together with the elastic constants of screened structures were obtained by calculating the Hessian matrix (matrix of the second derivatives of the energy with respect to the atomic positions) [77], which can be done by setting IBRION=6 (NFREE= 4) in VASP run. For cubic structures, the mechanical stability of lattice structures is verified as $C_{11} > 0, C_{44} > 0, C_{11} > |C_{12}|, C_{11} + 2C_{12} > 0$, where C_{ij} are components of elastic constant matrix [184]. After screening the materials with the mechanical criteria, we further narrow-down the materials by checking the vibrational frequencies at the Γ point. All materials with negative Γ point frequencies were discarded.

Phonon Dispersion calculation After the structures pass the mechanical stability criteria and all Γ point frequencies are positive, we further calculate the full phonon dispersions in the first Brinounion zone (BZ). All 2^{nd} interatomic force constants (IFCs) of the cubic structures were computed in a 2x2x2 supercell based on their corresponding primitive cell. Then, the phonon dispersions were calculated by using the PHONOPY package [154] with high symmetry paths $\Gamma \to X \to U \to K \to \Gamma \to L \to W \to X$ [50].

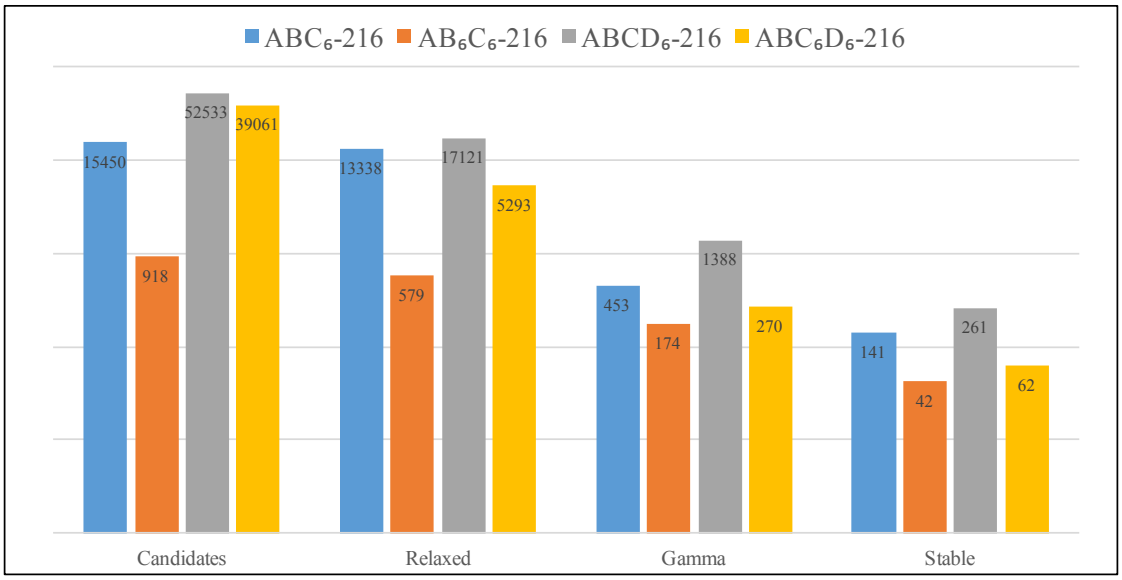
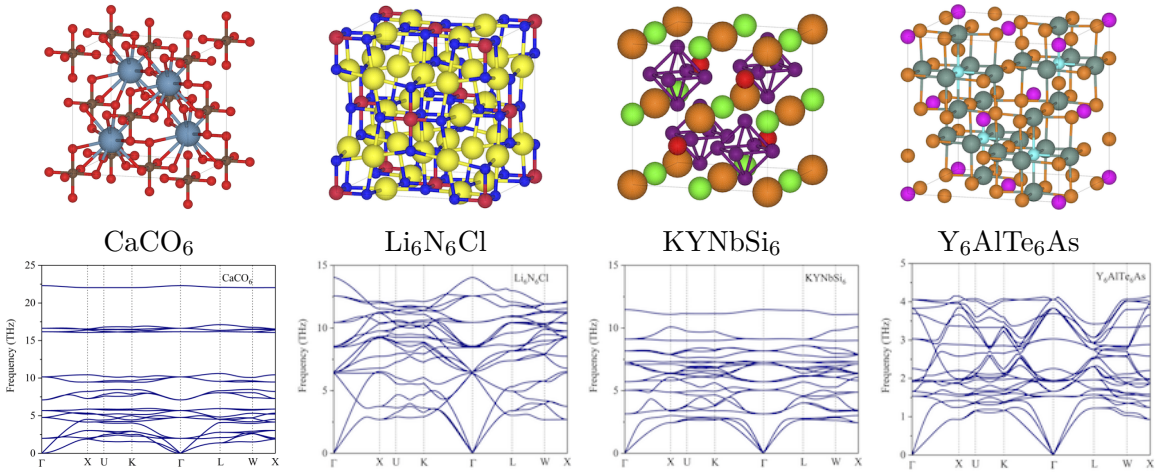


Figure 3.7 The number of materials after each filtering process for four new material prototypes (ABC_6-216,AB_6C_6-225,$ABCD_6$-216,ABC_6D_6-216) (The bars' height is at logarithmic scale of real values.).

In total, four prototypes with stable materials are discovered: ABC_6-216, AB_6C_6-225, $ABCD_6$-216, and ABC_6D_6-216. The details of the number of materials for each prototype are shown in Figure 3.7. We find that most of the generated structures can be successfully relaxed using DFT calculation. *Candidates* are generated structures that are charge neutral and have negative formation energy as predicted by CGCNN. *Relax* are candidate materials successfully optimized by DFT. *Gamma* are optimized structures that have positive vibrational frequencies at the *Gamma* point, indicating the structures are potentially stable. *stable* are the final stable structures with full positive phonon dispersions as verified by DFT. To our best of knowledge, ABC_6-216 and ABC_6D_6-216 are novel prototypes that are not in our training dataset, and the validation sets MP-TC3 and ICSD-TC3. Also, the AB_6C_6-225 prototype is not in the training dataset and only one unstable material can be found in MP. However, our method finds 42 stable ones. Two materials of $ABCD_6$-216 prototype are in the training dataset, and several others are in MP and ICSD. We expand the datasets by finding 62 stable materials of prototype $ABCD_6$-216. Overall, we find 183 stable

ternary materials and 323 stable quaternary materials. Figure 3.8 shows four newly discovered stable cubic materials with their phonon dispersion curves. The CIFs of the 506 new prototype materials can be found in Carolina Materials Database [188].



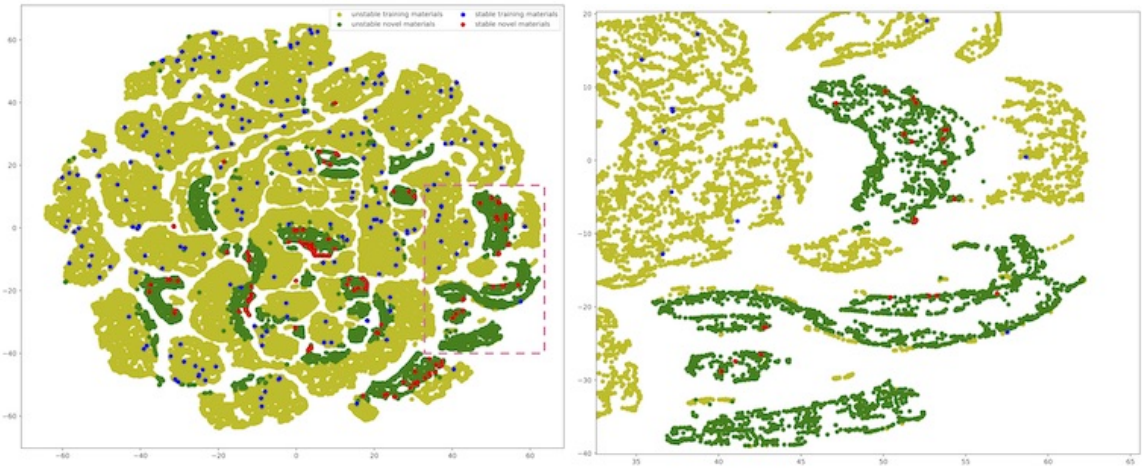
(a) $CaCO_6$ (b) Li_6N_6Cl (c) $KYNbSi_6$ (d) Y_6AlTe_6As

Figure 3.8 Examples of four new prototype materials. Top row is the crystal structures and the second row with corresponding phonon dispersion.

Some interesting features have been observed from the phonon dispersions of newly discovered materials. For instance, a couple of hundred cubic structures we have screened out possess significant but tunable phonon bandgaps (e.g., $CaCO_6$ as shown in Figure 3.8(a)). Such phonon bandgaps could lead to extraordinary hot carrier performance [23, 153, 185, 174], which is very promising for their potential application in photovoltaics, nonlinear optics (e.g, ultrashort pulsed lasers), multi-exciton generation devices, and even photocatalysis. Large phonon bandgaps at extremely high frequencies (such as H-containing materials not shown herein) deserve further investigation for their electron-phonon coupling properties [182, 179, 180], which could be beneficial for designing novel superconductors. Also, there are many cubic materials possessing very soft acoustic modes, e.g., the longitudinal acoustic (LA) phonon branch in $KYNbSi_6$ (Figure 3.8(c)), which indicate strong phonon anharmonicity and could be good candidates for waste-heat energy recovery (thermoelectrics). Last but not least, the phonon dispersion of Y_6AlTe_6 structure exhibits a very large gradient

in high-frequency optical phonon modes and thus their phonon group velocities will be very high, which could lead to a significant contribution to the overall thermal transport from these optical modes and thus unusual temperature-dependent lattice thermal conductivity [121].

3.4.4 Visualization of the relationship between new and existing prototypes within the same space group.



(a) Distribution of ABC_6 materials and the training samples of space group $F\bar{4}3m$

(b) Local zoomed region in (a)

Figure 3.9 Visualization of the structural distributions of the materials in training set and the generated new-prototype ABC_6 materials both belonging to $F\bar{4}3m$.

To qualitatively evaluate how the new-prototype materials are structurally different from existing materials, we represent both sets of materials using simulated (for generated samples) and real X-ray diffraction (XRD) spectrum of dimension 901, which is a way to analyze the structures of inorganic materials. The simulated XRDs are calculated using the Pymatgen package [111]. We then use the t-SNE embedding approach [88] to map the samples' XRD vectors into 2D space, which are then plotted together to visualize how existing and novel materials of the same space group 216. Figure 3.9 shows t-SNE embedding of existing materials in the training dataset and newly discovered ABC_6-216 materials. The new materials form

structurally distinct clusters. Sub-figure shows (a) Overall distributions of existing and new-prototype materials. Sub-figure shows (b) Zoomed region as marked in sub-figure (a). From Figure 3.9(a), we can find that new prototype materials (dark green dots) form distinct clusters, and there are apparent boundaries between known and unknown materials, which indicates that our model can generate materials beyond the scope of existing prototypes with significant structure deviations. Figure 3.9(b) shows a zoomed region of clusters of novel ABC_6-216 materials, which implies that even samples of the same prototype can form structurally different clusters. For the other three prototypes, the distribution of known structures and our new-prototype structures are shown in Figure 3.10- Figure 3.13. In Figure 3.10, we find that the new-prototype (AB_6C_6) materials are mostly located at the peripheral regions of know materials clusters, indicating their structural closeness to known structures. In contrast, Figure 3.12 shows that materials of two new-prototypes (ABC_6 and AB_6C_6) tend to form distinct clusters from known cubic materials in MP-TC3 and ICSD-TC3 validation sets, indicating their structural deviation from known materials. Additionally, for most of these new-prototype clusters, we have identified one or more DFT-verified stable materials. Figure 3.11 shows that materials of new-prototype $ABCD_6$ form multiple new clusters, each of which contains multiple DFT-verified stable materials. Sub-figure (a) shows the overall clustering. Sub-figure (b) is the zoomed region as marked in sub-figure (a). We find for this prototype, there are multiple new clusters, each of which contains multiple DFT-verified stable materials. Instead, materials of new-prototype ABC_6D_6 form much fewer cluster compared to the training set OQMD-TC3 as shown in 3.13.

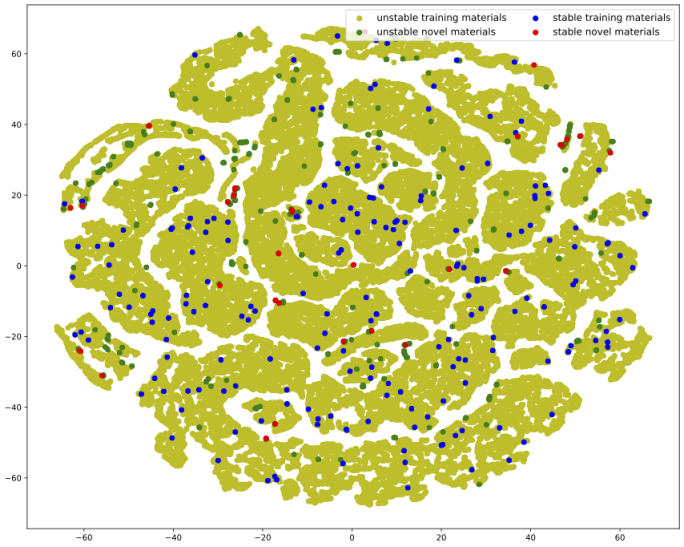
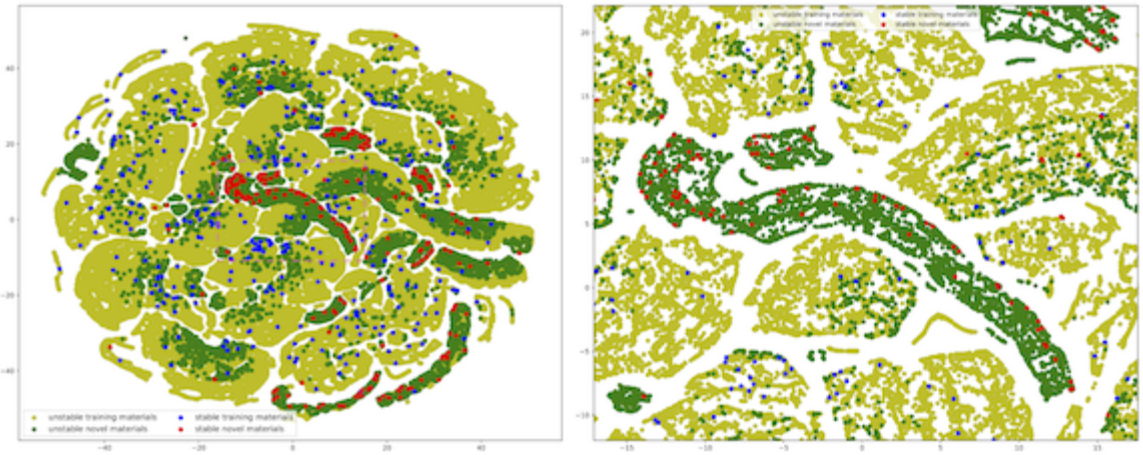


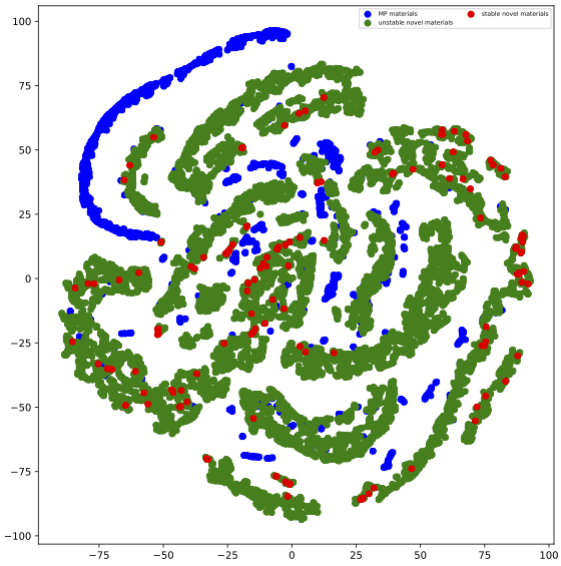
Figure 3.10 Visualization of the distributions of materials in the training set and the new prototype AB_6C_6 materials both with $Fm\bar{3}m$.



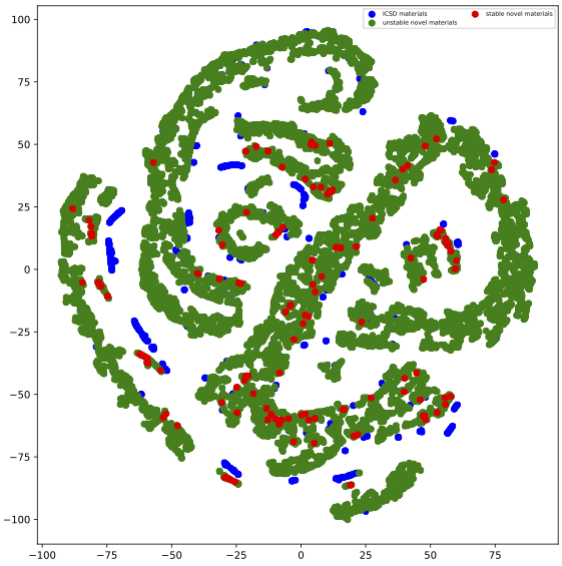
(a) Distribution of training samples and the new-prototype $ABCD_6$ materials with $F\bar{4}3m$.

(b) Local zoomed region.

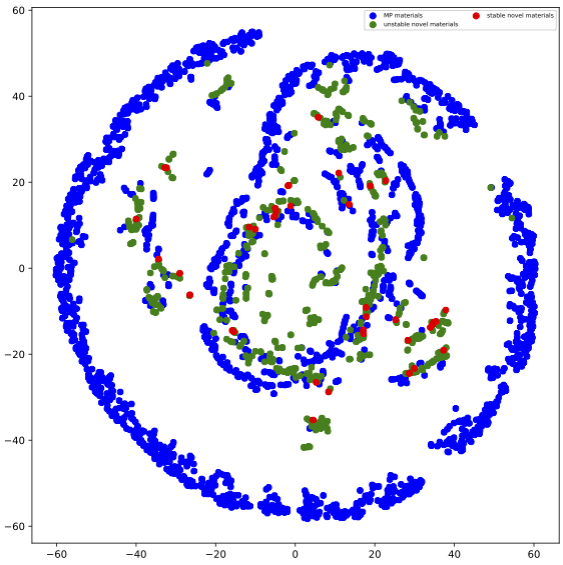
Figure 3.11 Visualization of the distributions of materials in the training data and the new-prototype $ABCD_6$ materials both with space group of $F\bar{4}3m$.



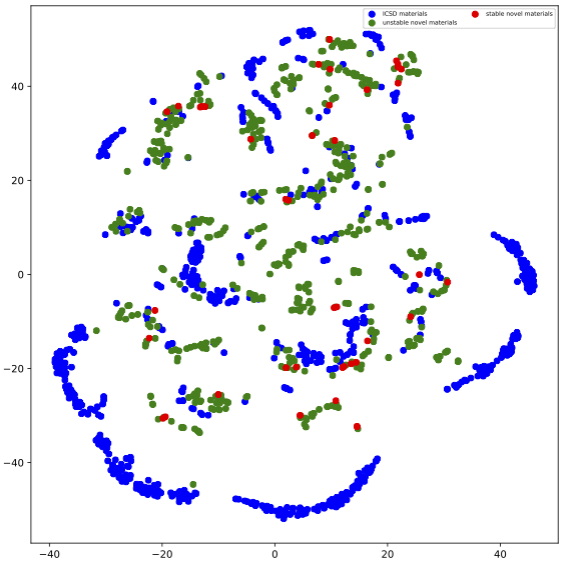
(a) Distribution of materials in MP-TC3 and new ABC_6 materials with space group $F\bar{4}3m$.



(b) Distribution of materials in ICSD-TC3 and new ABC_6 materials with space group $F\bar{4}3m$.

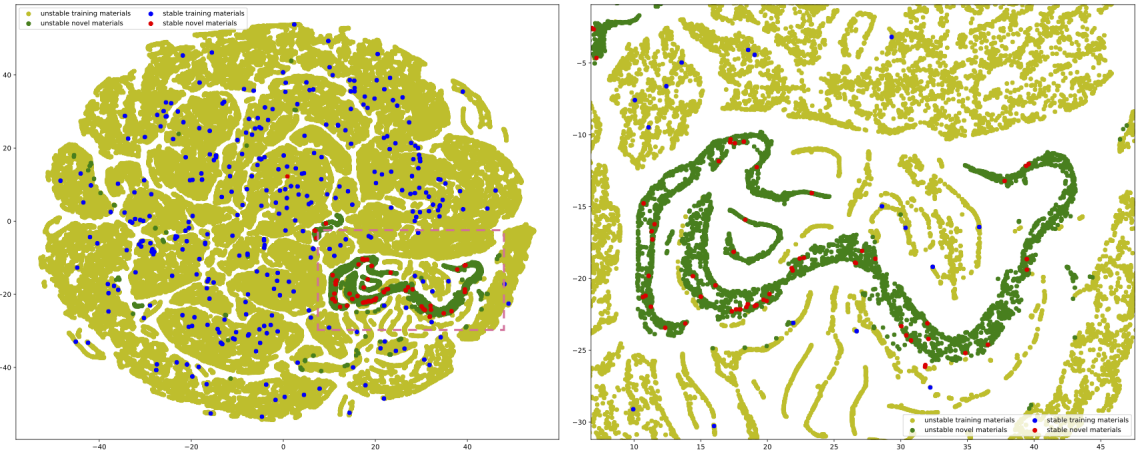


(c) Distribution of materials in MP-TC3 and new AB_6C_6 materials with space group $Fm\bar{3}m$.



(d) Distribution of materials in ICSD-TC3 and new AB_6C_6 materials with space group $Fm\bar{3}m$.

Figure 3.12 Visualization of the distributions of materials in the MP-TC3/ICSD-TC3 validation sets and the new prototype (ABC_6 and AB_6C_6) materials both with space group of $F\bar{4}3m$ and $Fm\bar{3}m$.



(a) Distribution of training set samples and the new-prototype ABC6D6 materials with space group of F$\bar{4}$3m.

(b) Local zoomed region.

Figure 3.13 Visualization of the distributions of materials in the training data and the new-prototype ABC_6D_6 materials both with space group of F$\bar{4}$3m.

3.5 Chapter Summary

Large scale generation of new materials with distinct structures and functions are highly desirable for widely used high-throughput screening based materials discovery. Faced with astronomically large structural design space (compared to the space of the chemical compositions), the generator models have to exploit the implicit sophisticated physicochemical and geometric rules and constraints embedded in the existing crystal materials. Here we propose a novel GAN-based deep generative model for large-scale generation of three major types (space groups:216, 225, 221) of cubic materials structures. Trained with 375,749 ternary cubic crystal structures from OQMD, our CubicGAN model can rediscover most of the known cubic structures as curated over more than 100 years of history within 10 million samplings. Especially, further analysis shows that our GAN model can generate not only new materials of existing prototypes but also new-prototype materials with distinct structural novelty. In total, we have identified 24 new prototypes of cubic materials. With rigorous DFT-based relaxation and phonon dispersion calculation, we have identified and verified 506 new-

prototype cubic materials, which are shared via Carolina Materials Database(http://www.carolinamatdb.org/). From them, we have already identified several crystal structures with exceptional properties to be exploited in future. Together, our CubicGAN has demonstrated a promising path to large-scale generation and discovery of new materials.

Chapter 4

Physics Guided Generative Adversarial Networks for Generations of Crystal Materials with Symmetry Constraints

4.1 Introduction

In chapter 3, we introduced the CubicGAN, a WGAN based crystal generative model which can generate crystals for three cubic space groups. Although the CubicGAN has shown a promising path to generate stable cubic materials at a high-throughput means, it depends on using special fractional coordinates and the number of space groups generated is limited.

While those works discussed in the last chapter open the door to generative design of new materials, several unique challenges still remain that prevents effective generative design: (1) how to learn the physical atomic constraints of stable materials to enable efficient sampling; (2) how to achieve precise generation of atom fractional coordinates and lattice parameters; (3) how to handle the extreme bias of the distribution of materials in 230 space groups. In this work. we introduce a new physics guided GAN architecture to exploit the physical rules for addressing aforementioned challenges. Our contributions are summarized as follows:

1. We present a new physics guided deep generative model for crystal generation that combines the space group affine transformation and an efficient self-augmentation method.

2. We propose two physics-oriented losses based on atomic pairwise distance constraints and symmetry to fuse the physical laws into deep learning model training.

3. We evaluate our model against two baselines to show its superiority and perform DFT calculations to validate our generated structures with high success rate (93.5% can be optimized successfully).

4.2 Problem Statement and Notations

The structure of an inorganic material can be represented by a unit cell in material science. The unit cell is the smallest unit that completely reflects the arrangement of atoms in the 3D space. Additionally, the unit cell describes the periodic structure of an inorganic material and it can be repeated infinite times along three directions to form a super cell. A material $\mathcal{M}$ can be denoted as following:

$$\mathcal{M} = (\boldsymbol{E}, \boldsymbol{B}, \boldsymbol{P}, \boldsymbol{O}), \tag{4.1}$$

where: (a) $\boldsymbol{E} = (e_0, e_1, e_2) \in \mathbb{E}$ denotes elements in materials, where $\mathbb{E}$ is the element set in periodic table. In this work, we only deal with ternary materials so that there are only three unique elements in the unit cell;

(b) $\boldsymbol{B} = (\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^{3\times3}$ denotes the symmetry equivalent positions termed as base atom sites. $\boldsymbol{b}_i$ is fractional coordinates of an atom denoted by $[u, v, w]^T$. We choose materials that one element only has one base atom site so that three atom sites can be used to represent the atom positions. Moreover, any one atom site of each element can be considered as the base atom site for that element;

(c) $\boldsymbol{P} = (a, b, c, \alpha, \beta, \gamma) \in \mathbb{R}$ are six lattice parameters that define three lengths and three angles of the unit cell;

(d) $\boldsymbol{O} = (\boldsymbol{t}_0, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_n) \in \mathbb{R}^{4\times4}$ denotes affine matrix that represents the symmetry operations defined by space groups *sgp*. $\boldsymbol{t}_j$ is one affine operator containing the rotation and translation matrices. n is determined by space groups. Generally the higher symmetry of a space group, the larger n. n can be as small as 1 or as large as 192.

Figure 4.1 shows an example of inorganic material. The left figure shows the unit cell of the periodic structure. The right figure shows a bigger cell that repeats the unit cell three times along with three directions. In this material, $\boldsymbol{E}$ is (Ca, Ti, O), $\boldsymbol{B}$ can

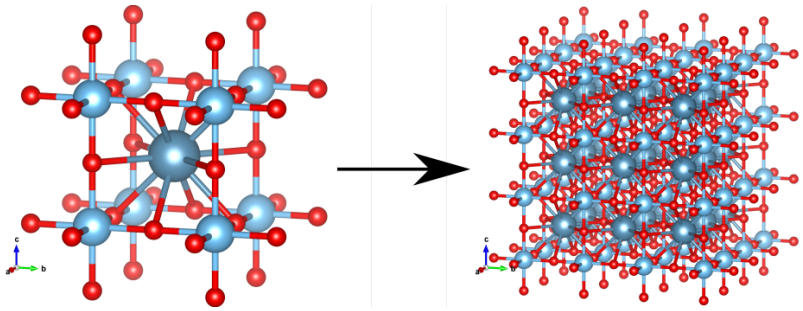


Figure 4.1 The periodic structure of calcium titanium oxide ($CaTiO_3$).

be $(\begin{bmatrix}0.5\\0.5\\0.5\end{bmatrix}, \begin{bmatrix}0.0\\0.0\\0.0\end{bmatrix}, \begin{bmatrix}0.0\\0.0\\0.5\end{bmatrix})$ in which each set of coordinates is corresponding to each element in $\boldsymbol{E}$, $\boldsymbol{P}$ is $(3.89, 3.89, 3.89, 90.0, 90.0, 90.0)$, and $\boldsymbol{O}$ has 48 symmetry operations. We can find that the lengths are same and all angles are 90° since $CaTiO_3$ is a cubic structure.

In order to acquire all atom positions in the unit cell, each base atom site can be converted by affine matrix $\boldsymbol{O}$. The conversion procedure is summarized in Algorithm 1. Different materials vary from the number of atoms and the number of elements. In order to make a fixed size of inputs, we only use ternary materials in this research. After conversion shown in Algorithm 1, the number of atom (sites) also differs from materials. That is the reason why base atom sites (one element one base site) are used to represent atom positions. In addition, it should be noted that the calculation of the uniqueness at line 10 of Algorithm 1 is not differentiable and time-consuming.

Fractional coordinates can be converted to Cartesian coordinates $[x, y, z]^T$ using [33]:

$$\begin{bmatrix}x\\y\\z\end{bmatrix} = \boldsymbol{A} \cdot \begin{bmatrix}u\\v\\w\end{bmatrix}, \tag{4.2}$$

where $\boldsymbol{A}$ is a lattice matrix calculated by lattice parameters $\boldsymbol{P}$ using:

Algorithm 1 Generate unique coordinates using base sites and affine matrix

Require: The space group *sgp*, the base atom sites $\boldsymbol{B}$

$\boldsymbol{O} \leftarrow Lib(sgp)$ ▷ Lib saves affine matrix
$n \leftarrow len(\boldsymbol{O})$
$\boldsymbol{coords} \leftarrow$ *an empty list*
for $i \leftarrow 1$ to 3 **do**
    add 0 to $\boldsymbol{b}_i$
    $\boldsymbol{uniq} \leftarrow$ *an empty list*
    for $j \leftarrow 1$ to n **do**
        $\boldsymbol{c} \leftarrow \boldsymbol{b}_i \cdot \boldsymbol{t}_j - \lfloor \boldsymbol{b}_i \cdot \boldsymbol{t}_j \rfloor$
        pop last element from $\boldsymbol{c}$
        if $\boldsymbol{c}$ *not in* $\boldsymbol{uniq}$ **then**
            add $\boldsymbol{c}$ *to* $\boldsymbol{uniq}$
        end if
    end for
    add $\boldsymbol{uniq}$ *to* $\boldsymbol{coords}$
end for
return $\boldsymbol{coords}$

$$\boldsymbol{A} = \begin{bmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & c\frac{\cos\alpha - \cos\beta\cos\gamma}{\sin\gamma} \\ 0 & 0 & \frac{V}{ab\sin\gamma} \end{bmatrix}, \tag{4.3}$$

where $V = abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$ is the volume of the unit cell.

Now we can model the generation of materials as follows:

$$(\boldsymbol{B}, \boldsymbol{P}) = f_\theta(\boldsymbol{Z}, \boldsymbol{E}, sgp), \tag{4.4}$$

where f_θ is the generative model that learns the knowledge of forming crystal structures given inputs of random noise $\boldsymbol{Z}$, element set $\boldsymbol{E}$, and space group *sgp*.

4.3 Proposed Method

We describe our material generation model and its three major components: (1) discriminator, (2) generator, and (3) atom distance matrix calculation module. These components are designed to be differentiable so that the whole pipeline can be trained

end-to-end. In addition, the atom distance- and coordinates- based losses and self-augmentation of materials are introduced. Figure 4.2 illustrates our proposed Physics Guided Crystal Generative Model (**PGCGM**).

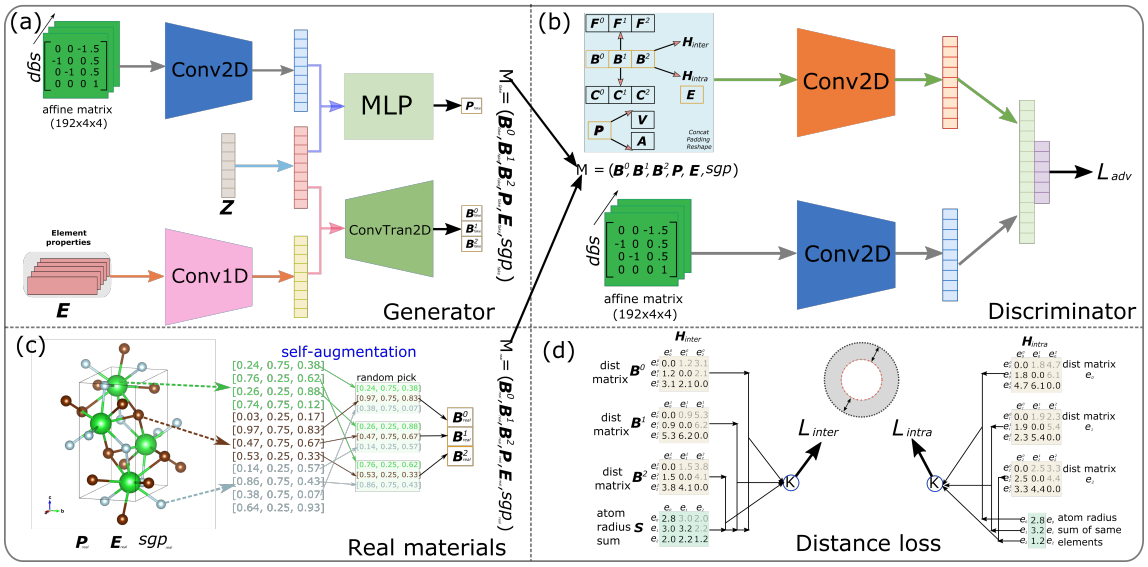


Figure 4.2 The main framework of our proposed method.

4.3.1 Self-augmentation

The generation of atom coordinates that meet the symmetry constraints is one of the most challenging tasks in crystal generation. In order to make the fixed size of representation for crystals, we use base atom sites. As shown in sub-figure (c) of Figure 4.2, we can use any atom site of each element to form a set of base atom sites. Instead of randomly picking up them, we choose three atoms for three elements individually using steps as shown in below:

Step 1. Shuffle three elements in $\boldsymbol{E}$;

Step 2. Randomly pick up the first element e_0 and one atom position $\boldsymbol{b}_0$ for it;

Step 3. Randomly pick up the second element e_1 from the rest two elements and find the closest atom $\boldsymbol{b}_1$ of the second element to the atom in the first step;

Step 4. Calculate the atom distance from the atoms of the last element e_2 to the two atoms selected in the first and second steps respectively, then sum the atom distance element-wise and the atom of the last element with the smallest sum is considered as the closest atom $\boldsymbol{b}_2$ to the selected atoms in the first and second steps;

Step 5. Repeat Steps 2, 3, and 4 three times to obtain three sets of base atom sites $(\boldsymbol{B}^0_{real}, \boldsymbol{B}^1_{real}, \boldsymbol{B}^2_{real})$;

Step 6. Repeat last five steps 32 times.

4.3.2 Materials Representation

In previous section we use $\mathcal{M} = (\boldsymbol{E}, \boldsymbol{B}, \boldsymbol{P}, \boldsymbol{O})$ to completely describe a crystal material. As shown in sub-figure (b) of Figure 4.2, however, we use three sets of base atom sites $(\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2)$. Thus here we re-formulate a material as $\mathcal{M}^* = (\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2, \boldsymbol{P}, \boldsymbol{E}, sgp)$. The space group sgp is used to link to the affine matrix O. We can use $(\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2)$ in $\mathcal{M}^*$ to calculate physical properties as inputs to the discriminator and to design physics-based losses. Three sets of base atom sites are useful for two reasons: (1) we want to add more crystal information for the discriminator and let the discriminator have enough information to tell real materials from fake ones; (2) With more base atom sites, we can calculate more atom distances as the physical constraints in the generator and the inputs to the discriminator.

All Fractional Coordinates We use affine matrix O to acquire the whole atom sites in the unit cell as shown in Algorithm 1. Since the number of affine operators in O varies in space groups, we zero-pad the affine matrices as large as $192 \times 4 \times 4$. We then transform each base atom site by the affine matrix and get a coordinates matrix $\boldsymbol{F}_{all}$ with shape of $192 \times 3 \times 3$. Affine transformation leads to duplicate fractional coordinates. In material science, practitioners usually remove the duplicates. However,

uniqueness calculation is not differentiable and it requires lots of time to do it. We choose to average along with the first dimension of $\boldsymbol{F}_{all}$ to get three sets of averaged full fractional coordinates $(\boldsymbol{F}^0, \boldsymbol{F}^1, \boldsymbol{F}^2)$, each of which is with shape of 3×3.

For a real material, base atom sites $(\boldsymbol{B}^0_{real}, \boldsymbol{B}^1_{real}, \boldsymbol{B}^2_{real})$ can be transformed into the same average full fractional coordinates, which means $\boldsymbol{F}^0_{real} = \boldsymbol{F}^1_{real} = \boldsymbol{F}^2_{real}$. When generating a fake material, base atom sites $(\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake})$ are supposed to belong to the same fake material, which hopefully results in $\boldsymbol{F}^0_{fake} = \boldsymbol{F}^1_{fake} = \boldsymbol{F}^2_{fake}$. However, the transformation of $(\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake})$ might slightly deviate from the goal. Thus using $(\boldsymbol{F}^0, \boldsymbol{F}^1, \boldsymbol{F}^2)$ in real and fake materials implicitly adds physical constraints, which somehow pushes the generator to generate different sets of base atom sites for a same material, which increases chances to generate good materials in return.

Base Cartesian Coordinates Three sets of Cartesian coordinates can be calculated for each set of base atom sites by Eq. (4.2) and we denote them by $(\boldsymbol{C}^0, \boldsymbol{C}^1, \boldsymbol{C}^2)$.

Atom Distance Matrices Given three sets of base atom sites $(\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2)$, we calculate the atom distance matrices $\boldsymbol{H}_{inter}$ and $\boldsymbol{H}_{intra}$ as shown in sub-figure (d) of Figure 4.2. We firstly calculate pair-wise different atom distance matrix for each base atom site $\boldsymbol{B}^j,\ j = 0, 1, 2$ and return only values in upper triangle of corresponding distance matrix termed by $\boldsymbol{H}_{inter}$. Then we select three atoms belonging to the same element to form a set of three atom sites for three elements and calculate pair-wise same atom distance matrix and again return only values in upper triangle of corresponding distance matrix termed by $\boldsymbol{H}_{intra}$. The final shape of $\boldsymbol{H}_{inter}$ and $\boldsymbol{H}_{intra}$ both is 3×3.

Lattice Parameters The volume of the unit cell can be calculated by lattice parameters $\boldsymbol{P}$. We repeat the scalar volume three times to get the volume vector $\boldsymbol{V}$.

We also use the lattice matrix $\boldsymbol{A}$ in Eq. (4.3) as part of the inputs to the discriminator.

Element Properties We use 23 properties as in **CubicGAN** to formalize element matrix $\boldsymbol{E}$.

Now we list all parts of inputs to the discriminator in Table 4.1. $\boldsymbol{P}^*$ only contains the lengths because the angles are either $(90°, 90°, 90°)$ or $(90°, 90°, 120°)$ in the training materials. Thus instead of generating three angles in $\boldsymbol{P}$ for fake materials, we build a map between angles and the space group *sgp*. Then we concatenate all parts and a zero matrix of shape 3×3 into a matrix of shape of 3×64. The matrix is finally reshaped into $3 \times 8 \times 8$ as the inputs to the discriminator.

Table 4.1 Symbols and their shape used in inputs to the discriminator.

symbol	shape	symbol	shape
$(\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2)$	3×9	$\boldsymbol{P}^*$	3×1
$(\boldsymbol{F}^0, \boldsymbol{F}^1, \boldsymbol{F}^2)$	3×9	$\boldsymbol{V}$	3×1
$(\boldsymbol{C}^0, \boldsymbol{C}^1, \boldsymbol{C}^2)$	3×9	$(\boldsymbol{H}_{inter}, \boldsymbol{H}_{intra})$	3×6
$\boldsymbol{E}$	3×23	$\boldsymbol{A}$	3×3

4.3.3 Discriminator

There are two input branches for crystal representation and affine matrix in Discriminator as in sub-figure (b) of Figure 4.2. Each branch is forwarded to a 2D convolutional block and the learnt features are concatenated together. The concatenated vector is sent to a couple of fully connected layers to get the discriminative score. We have three different sets of base atom sites in our inputs and with the affine matrix branch, it helps to implicitly learn the knowledge of how affine matrix transforms base atom sites into full atom sites. The detailed architectures of two convolutional blocks can be found in Table 4.2. **Mat** is the input material representations with shape of $3 \times 8 \times 8$. **SymOp** is the zero-padded symmetric operation matrix for space groups of materials. The 2D convolutional layer parameters are denoted as "C2D-<number of channels>-<receptive field size>". The fully connected layer parameters are denotes

as "FC-<number of neurons>". The concatenation is denoted as "CAT-<number of neurons>". We use *LeakyReLU* as the activation function after each layer except for the last layer. The negative slope for it is 0.2.

Table 4.2 Discriminator configuration.

Discriminator Configuration	
Mat-$3 \times 8 \times 8$	
C2D-16-2	
C2D-32-2	
C2D-64-2	
C2D-96-2	**SymOp**-$192 \times 4 \times 4$
C2D-128-2	C2D-64-2
C2D-192-2	C2D-128-2
C2D-256-2	C2D-256-2
CAT-512	
FC-265	
FC-1	

4.3.4 Generator

The architecture of generator is shown in sub-figure (a) of Figure 4.2. Three branches are found. Conditioning on element constituents and space group, the generator outputs three sets of base atom sites ($\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake}$) and unit cell length $\boldsymbol{P}^*$. Then we re-formalize Eq. (4.4) as follow:

$$(\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake}, \boldsymbol{P}^*_{fake}) = f^*_\theta(\boldsymbol{Z}, \boldsymbol{E}, sgp). \tag{4.5}$$

Taking random noise $\boldsymbol{Z}$, space group sgp, and element properties matrix $\boldsymbol{E}$ as inputs, the generator can generate a material with the same lattice parameters and space group but different representations of the base atom sites when merely sampling one material. Our goal here is that the generated three sets of base atom sites belong to the same material. Random noise $\boldsymbol{Z}$ is mapped to a dense vector a fully connected layer. The space group branch is the same as in discriminator. Element matrix $\boldsymbol{E}$ is forwarded to a 1D convolutional layer (Conv1D). The outputs of random noise and

space group branches are combined together as the inputs to a multi-layer perceptron (MLP) block to generate unit cell length $\boldsymbol{P}^*_{fake}$. The outputs of random noise and element branches are combined together as the inputs to 2D deconvolutional layers (ConvTran2D) to generate three sets of base atom sites ($\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake}$). The detailed descriptions for MLP, Conv1D, and ConvTran2D can be found in Table 4.3. **SymOp** is the zero-padded symmetric operation matrix for space groups of materials. **Z** is the random noise with shape of 128 and it shared by two branches for generating unit cell length $\boldsymbol{P}^*$ and three set of base atom sites ($\boldsymbol{B}^0_{fake}, \boldsymbol{B}^1_{fake}, \boldsymbol{B}^2_{fake}$). The 2D convolutional layer parameters are denoted as "C2D-<number of channels>-<receptive field size>". The 2D deconvolutional layer parameters are denoted as "TC2D-<number of channels>-<receptive field size>". The fully connected layer parameters are denotes as "FC-<number of neurons>". The concatenation is denoted as "CAT-<number of neurons>". We use batch normalization and *ReLU* after each layer except for the last layers of two branches. They are followed by a *Tanh* activation to generate lengths and atom coordinates.

Table 4.3 Generator configuration.

<table>
<tr><td colspan="3">Generator Configuration</td></tr>
<tr><td colspan="2"></td><td>ElemProp-23 × 3</td></tr>
<tr><td>SymOp-192 × 4 × 4</td><td rowspan="3"></td><td>C1D-64-2</td></tr>
<tr><td>C2D-64-2</td><td>C1D-128-2</td></tr>
<tr><td>C2D-128-2</td><td>Z-128</td><td>flatten</td></tr>
<tr><td>C2D-256-2</td><td>FC-256</td><td>FC-256</td></tr>
<tr><td colspan="2">CAT-512</td><td>CAT-512</td></tr>
<tr><td colspan="2">FC-128</td><td>TC2D-1024-2</td></tr>
<tr><td colspan="2">FC-64</td><td>TC2D-512-2</td></tr>
<tr><td colspan="2">FC-32</td><td>TC2D-256-1</td></tr>
<tr><td colspan="2">FC-16</td><td>TC2D-128-1</td></tr>
<tr><td colspan="2">FC-3</td><td>TC2D-64-1</td></tr>
<tr><td colspan="2" rowspan="2">output: $\boldsymbol{P}^*$-3</td><td>TC2D-3-1</td></tr>
<tr><td>output: $\boldsymbol{B}$ - 3 × 3 × 3</td></tr>
</table>

4.3.5 Loss Function

The original GAN [43] is notoriously hard to train because of saturation and mode collapse in discriminator. We take advantage of WGAN-GP [44] with gradient penalty to enhance the training stability in our network. WGAN-GP changes the Sigmoid function of the discriminator to a 1-Lipschitz function while introducing a gradients penalty term to enforce the norm of gradients to be close to 1. The loss function is described in Eq. (4.6):

$$\begin{aligned}
\hat{\mathcal{M}}^* &= \epsilon \mathcal{M}^*_{real} + (1-\epsilon)\mathcal{M}^*_{fake}, \quad \epsilon \sim U(0,1), \\
\mathcal{L}_{dis} &= D(\mathcal{M}^*_{fake}) - D(\mathcal{M}^*_{real}) + \lambda_d(\left\| \nabla_{\hat{\mathcal{M}}^*} D(\hat{\mathcal{M}}^*) \right\|_2 - 1)^2, \\
\mathcal{L}_{adv} &= -D(\mathcal{M}^*_{fake}),
\end{aligned} \tag{4.6}$$

where $\hat{\mathcal{M}}^*$ is linearly interpolated between real and fake materials and ϵ is uniformly sampled from 0 and 1. $\mathcal{L}_{dis}$ and $\mathcal{L}_{adv}$ represent the loss function of the discriminator and adversarial loss for generator respectively. The third term in $\mathcal{L}_{dis}$ is the gradient penalty and λ_d is set 10. $D(.)$ means the score result from the discriminator.

Atom Distance Losses To ensure that the atoms in generated crystal structures are not crowded or not too far apart from each other, we introduce the inter- and intra-atom distance based losses as following:

$$\begin{aligned}
\mathcal{L}_{inter} = \frac{1}{N*9}\sum_{i=1}^{N}\{&[\max(\boldsymbol{H}_{inter}, \phi^{upper}_{inter}\boldsymbol{S}_{inter}) - \phi^{upper}_{inter}\boldsymbol{S}_{inter}]^2 \\
&+ [\min(\boldsymbol{H}_{inter}, \phi^{lower}_{inter}\boldsymbol{S}_{inter}) - \phi^{lower}_{inter}\boldsymbol{S}_{inter}]^2\}, \\
\mathcal{L}_{intra} = \frac{1}{N*9}\sum_{i=1}^{N}\{&[\max(\boldsymbol{H}_{intra}, \phi^{upper}_{intra}\boldsymbol{S}_{intra}) - \phi^{upper}_{intra}\boldsymbol{S}_{intra}]^2 \\
&+ [\min(\boldsymbol{H}_{intra}, \phi^{lower}_{intra}\boldsymbol{S}_{intra}) - \phi^{lower}_{intra}\boldsymbol{S}_{intra}]^2\},
\end{aligned} \tag{4.7}$$

where $\mathcal{L}_{inter}$ constrains the distance in $\boldsymbol{H}_{inter}$. $[\max(\boldsymbol{H}_{inter}, \phi^{upper}_{inter}\boldsymbol{S}_{inter}) - \phi^{upper}_{inter}\boldsymbol{S}_{inter}]^2$ enforces the atom distance to be smaller than $\phi^{upper}_{inter}\boldsymbol{S}_{inter}$ and $[\min(\boldsymbol{H}_{inter}, \phi^{lower}_{inter}\boldsymbol{S}_{inter}) -$

$\phi_{inter}^{lower}\boldsymbol{S}_{inter}]^2$ enforces the atom distance to be bigger than $\phi_{inter}^{lower}\boldsymbol{S}_{inter}$. $\boldsymbol{S}_{inter}$ are atom radius sum corresponding to each pair of atoms in $\boldsymbol{H}_{inter}$ and ϕ_{inter}^{upper} and ϕ_{inter}^{lower} are control weights for upper and lower bound of inter-atom distance, respectively. In this way, the distance of two atoms is constrained to be in the grey area indicated by two circles in sub-figure (d) of Figure 4.2. Similarly, $\mathcal{L}_{intra}$ constrains the distance in a range in $\boldsymbol{H}_{intra}$. N is batch size and 9 is the number of distance value in $\boldsymbol{H}_{inter}$ and $\boldsymbol{H}_{intra}$.

Base and Average Full Coordinates Losses The generator generates three sets of base atom sites $(\boldsymbol{B}_{fake}^0, \boldsymbol{B}_{fake}^1, \boldsymbol{B}_{fake}^2)$ which are supposed to be different but a part of full coordinates. The averaged transformation to $(\boldsymbol{F}_{fake}^0, \boldsymbol{F}_{fake}^1, \boldsymbol{F}_{fake}^2)$ from base atom sites should be exactly same. With these implicit rules, we design two losses to explicitly enforce them in the generator as expressed below:

$$\begin{aligned}
\mathcal{L}_{full} = \frac{1}{N*9}\sum_{i=1}^{N}\{&\max(0, cos(\frac{\boldsymbol{F}_{fake}^0}{\left\|\boldsymbol{F}_{fake}^0\right\|_2}, \frac{\boldsymbol{F}_{fake}^1}{\left\|\boldsymbol{F}_{fake}^1\right\|_2})) \\
&+ \max(0, cos(\frac{\boldsymbol{F}_{fake}^1}{\left\|\boldsymbol{F}_{fake}^1\right\|_2}, \frac{\boldsymbol{F}_{fake}^2}{\left\|\boldsymbol{F}_{fake}^2\right\|_2})) \\
&+ \max(0, cos(\frac{\boldsymbol{F}_{fake}^0}{\left\|\boldsymbol{F}_{fake}^0\right\|_2}, \frac{\boldsymbol{F}_{fake}^2}{\left\|\boldsymbol{F}_{fake}^2\right\|_2}))\}, \\
\mathcal{L}_{base} = \frac{1}{N*9}\sum_{i=1}^{N}\{&(1 - cos(\frac{\boldsymbol{B}_{fake}^0}{\left\|\boldsymbol{B}_{fake}^0\right\|_2}, \frac{\boldsymbol{B}_{fake}^1}{\left\|\boldsymbol{B}_{fake}^1\right\|_2})) \\
&+ (1 - cos(\frac{\boldsymbol{B}_{fake}^1}{\left\|\boldsymbol{B}_{fake}^1\right\|_2}, \frac{\boldsymbol{B}_{fake}^2}{\left\|\boldsymbol{B}_{fake}^2\right\|_2})) \\
&+ (1 - cos(\frac{\boldsymbol{B}_{fake}^0}{\left\|\boldsymbol{B}_{fake}^0\right\|_2}, \frac{\boldsymbol{B}_{fake}^2}{\left\|\boldsymbol{B}_{fake}^2\right\|_2}))\},
\end{aligned} \tag{4.8}$$

where *cos* is cosine similarity function. We normalize each coordinate value across the mini-batch of size N. 9 is the number of coordinates.

Full Generator Loss By combining above losses, we can achieve our full loss for the generator:

$$\mathcal{L}_{gen} = \mathcal{L}_{adv} + \lambda_1\mathcal{L}_{inter} + \lambda_2\mathcal{L}_{intra} + \lambda_3\mathcal{L}_{full} + \lambda_4\mathcal{L}_{base}. \tag{4.9}$$

4.4 Experiments

We carry out experiments on materials from three databases and evaluate the performance of our **PGCGM** on newly released materials from OQMD [69]. In addition,

we compare our method with two generative models that can generate crystals of multiple chemical systems to validate the effectiveness of our **PGCGM** model.

4.4.1 Dataset

We collect our material data from MP [54], ICSD [8] and OQMD (v1.4) [69]. In total, 42072 ternary materials with 20 space groups are curated when we start this project. We use a 80-20 random training/validation split for all of our experiments. We term the dataset with 42072 materials as **MIO**. When conducting this project, the newest version of OQMD is just yet released. There are 9441 ternary materials that are filtered by the same criteria and are brand new materials in the new OQMD (v1.5). We use these 9441 ternary materials as our test dataset **TST** to compare our method with two baselines.

We select the material data from three databases: MP [54], ICSD [8], OQMD [69]. The selection criteria are described following:

1. Ternary materials with only three base atom sites (a.k.a. one element is allowed to have only one base atom site);
2. Only keep materials that do not contain elements in lanthanoid and actinoid;
3. Ternary materials whose space group has more than 400 materials in three databases;
4. Ternary materials in OQMD whose fractional coordinates does not all belong to the set $[0.0, 0.25, 0.5, 0.75]$ since materials with fractional coordinates all falling in that set dominate the database [189].

In total, 42072 materials are selected and 20 space groups are found in those materials following above criteria. The statistics of materials in each space group is shown in Table 4.4.

Table 4.4 The distribution of 20 space groups in dataset **MIO**.

SG	#	SG	#
$P4/mmm$	1180	$Immm$	4679
$Fm\bar{3}m$	3716	$Cmcm$	1004
$I4_1/amd$	588	$I\bar{4}2d$	749
$Pm\bar{3}m$	1462	$R\bar{3}$	1969
$F\bar{4}3m$	898	$I4/mmm$	6162
$P6_3/mmc$	5599	$Fd\bar{3}m$	3292
$P\bar{3}m1$	1191	$Pnma$	2527
$P6/mmm$	2214	$R\bar{3}m$	1479
$I4/mcm$	433	$P6_3mc$	692
$R\bar{3}c$	1246	$P4/nmm$	992

Table 4.5 The distribution of 20 space groups in dataset **TST**.

SG	#	SG	#
$P4/mmm$	317	$Immm$	59
$Fm\bar{3}m$	675	$Cmcm$	507
$I4_1/amd$	168	$I\bar{4}2d$	482
$P4/nmm$	719	$R\bar{3}$	374
$F\bar{4}3m$	60	$I4/mmm$	768
$P6_3/mmc$	1713	$Fd\bar{3}m$	239
$P\bar{3}m1$	674	$Pnma$	1386
$P6/mmm$	281	$R\bar{3}m$	576
$I4/mcm$	81	$P6_3mc$	151
$R\bar{3}c$	211	$Pm\bar{3}m$	0

We use first, second, and four criteria above to select materials in new released OQMD and the distribution of materials in 20 space groups is shown in Table 4.5. 9441 materials are chosen and space group $Pm\bar{3}m$ does not have new released materials.

4.4.2 Baselines

We compare **PGCGM** with two latest algorithms that can generate crystals with multiple chemical systems instead of only a special group of materials, such as VxOy and Mg-Mn-O systems [101, 64]. **FTCP** [127] combines real space properties (e.g., atom coordinates) and momentum-space properties to represent crystal structures. Then a CNN based VAE is trained for materials generation. **CubicGAN** [189] trains a WGAN-GP [44] to generate cubic structures in three space groups and here we expand the original method to 20 space groups.

4.4.3 Material generation

Evaluation Metrics Past studies in crystal generation used different evaluation metrics, making it hard to compare different methods. Here, we create a set of metrics to evaluate our method and two baselines. 1) *Validity.* Following [24], we consider a crystal structure as valid when the shortest distance between any two atoms is

bigger than 0.5Å. Following **CubicGAN**, we calculate the overall charge of a crystal structure and if it is neutral, then it is valid. Also, we count the number of structures after post-processing in our method and we apply the same post-processing onto the **CubicGAN**. 2) *Property distribution.* We calculate wasserstein distance (WD) between the property distribution of generated materials and materials in test dataset **TST**. The properties we used are minimum atom distance, maximum atom distance, and density. 3) *Diversity.* We calculate the diversity of compositions, which means the ratio of unique number of compositions in generated structures.

Atom Clustering and Merging For crystals with high symmetry, the number of atoms in the unit cell tends to be very large after conversion by Algorithm 1. We propose a post-processing method to reduce the number of atoms by clustering and merging. Firstly, we cluster the nearby atoms of the same elements using hierarchical clustering. The maximum atom distance allowed in our research is 1.2 times the atom radius. Secondly, we merge the atoms in the same clusters considering periodic attributes of crystal structures.

Table 4.6 Material Generation Performance.

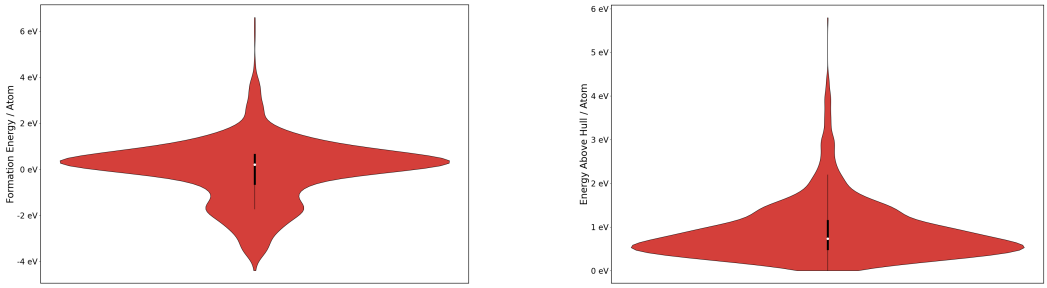
Method	Validity (%)			Prop. Dist.			(%)
	CIFs	Distance	Charge	minD	maxD	density	Diversity
FTCP	0.88	63.28	49.89	1.685	0.754	2.895	89.9
CubicGAN	4.97	99.0	59.47	0.626	3.476	3.871	98.0
PGCGM	1.98	**99.54**	57.36	**0.224**	3.664	2.675	**98.4**
PGCGM +dist	**7.14**	99.47	61.82	0.405	0.520	**0.765**	96.3
PGCGM +dist +coor	6.07	99.43	**63.34**	0.357	**0.490**	0.791	97.0

Results The performance is shown in Table 4.6. For each method, we sample 500,000 structures and the percentage of Crystallographic Information Files (CIFs) that are readable are shown in the *CIFs* column. For **PGCGM** and **CubicGAN**, we perform atom clustering and merging. We can found that **PGCGM+dist** has the

largest percentage of materials left and **PGCGM+dist+coor** comes next. It shows that distance and coordinates losses play a big part in generating valid materials. For next percentage related metrics, we use the number of CIFs left of each method as denominator. Our model outperforms **FTCP** and **CubicGAN** in terms of distance and charge validity. Since validity is relatively weak, property distribution provides a stronger metric to evaluate whether the generated materials are realistic. Our model significantly outperforms both two baselines, indicating that our generated crystals have much higher potential to be realistic materials. **PGCGM** also achieves the best diversity score.

Table 4.8 Lattice parameters P^* generation performance comparison.

	$P(a)$	$P(b)$	$P(c)$	216&225	CubicGAN
R^2	0.173	0.205	0.238	-0.245	-5.545
RMSE	2.132	2.259	3.060	1.721	2.454

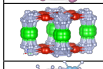
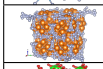


(a) Formation energy. (b) Energy above hull.

Figure 4.3 The distribution of formation energy 1579 materials and energy above hull for 1863 materials.

DFT Verification We randomly select 100 materials with less than and equal to 32 atoms in the unit cell for each space group. Out of 2000 generated crystals, 93.5% (1869) are successfully optimized, which is significantly better than 33.8% of CubicGAN as reported in [189]. Among 1,869 materials, 39.6% have negative formation energy and 106 ones have e above hull less than 0.25 eV/atom, which indicates they may be potentially stable and synthesizable as in Figure 4.3a and 4.3b. Table 4.7 shows 20 structures of 20 space groups. Before any post-processing, both

Table 4.7 20 example optimized crystals with lowest energy for 20 space group. GEN: generated; #: no.of atoms; MER: merged; OPT: optimized; FE: formation energy; SG: space group.

GEN	#	MER	OPT	#	FE(eV)	SG
	36			14	-4.400	164
	432			32	-1.711	227
	36			22	-3.659	139
	36			16	-3.680	186
	32			20	-3.532	129
	60			30	-3.536	194
	40			10	-3.314	123
	120			28	-2.579	221
	32			22	-3.690	71
	60			16	-3.666	191
	24			20	-3.819	62
	216			32	-2.267	216
	108			27	-3.487	166
	48			24	-3.721	63
	56			28	-3.672	141
	48			28	-3.681	122
	312			32	-3.178	225
	80			28	-2.667	140
	54			24	-3.945	148
	108			30	-4.340	167

materials has a large number of atoms. We found that the atoms of the same elements are crowded together in column of *GEN*. After clustering and merging the atoms of the same elements, the number of atoms drop rapidly from 36 to 14 for space group 164 and 432 to 32 for space group 227 as shown in column of *MER*. Column *OPT* shows the crystal structures after DFT optimization. We also compare the generated and optimized lattice parameters P^* of our model with **CubicGAN**. As shown in

Table 4.8, our model has the overall better performance than **CubicGAN** except for the RMSE of lattice parameter c. For direct comparison, we merely calculate R^2 and RMSE for space groups of 216 & 225 as in **CubicGAN**.

4.5 Chapter Summary

We propose a physics guided deep generative model for crystal generation to improve the generation performance of realistic materials. With atom distance and fractional coordinate losses, our generator learns to generate crystals that better satisfy the physical constraints. Extensive experiments demonstrate that our method outperforms the two baselines in terms of major evaluation metrics. Rigorous DFT calculations for candidate structure relaxation and optimization shows that our model can generate more valid crystals with higher efficiency and success rate.

Chapter 5

Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions

5.1 Introduction

Computational materials screening based on high-speed machine learning algorithms have become a reality as shown by a growing number of related works [56, 63, 137, 110, 113]. There are two types of screenings: one for screening known materials with desired properties [187, 47, 137]; one for screening hypothetical materials that have not been discovered or synthesized and usually have only composition information available [56, 63, 110]. Usually some kind of enumeration procedures [29] or generative machine learning models [28] can be used to generate many (millions) of hypothetical materials compositions as the combinations of selected set of elements, which requires fast algorithms to evaluate their stability [57], to predict their crystal structures [109] or physical properties of interest [40].

The crystal structure plays a critical role in determining the properties of materials. Knowing how the atoms of a material are arranged in the space helps understand its properties [166]. The structural information such as atomic coordinates or space groups can then be incorporated into the advancement of material design. Takahashi et al. [152] use the Gaussian mixture model to reveal two data clusters and then Random Forest is used to classify the crystal structures by eight descriptors. Further, first-principles calculations are performed to confirm the stability of predicted materials. However, predicting the atomic coordinates of a crystalline only from its composition using crystal structure prediction algorithms such as USPEX [107] (Universal Structure Predictor: Evolutionary Xtallography) is challenging and time-consuming as expensive density functional theory (DFT) simulations are needed [167]. In this case, prediction of the space groups or other structural information (such as atomic bonding angles and relative distances) of materials that have no crystal structure information can be useful to understand their physicochemical properties. For example, Ward et al. [167] use composition based features of elemental properties and the Voronoi tessellation of

the materials' crystal structure as inputs to ML to predict formation energies in their work [167].

Conventionally, the crystal structures of materials can be determined experimentally by the X-ray diffraction (XRD) technique, in which X-ray beams are used to hit nano particles and the scattered intensity of the beams are observed and measured. Novel materials can be unveiled by mapping XRD patterns to the measured or simulated XRD patterns of known materials. This method has led to the determination of a huge number of crystal structures as deposited in databases such as Material Projects [54] and ICSD [10, 8]. A large number of methods have been developed to analyze the XRD data such as programs for indexing and space-group determination. ITO [160], TREOR [170], DICVOL [11], McMaille [76], EXPO [4], and X-CELL [100] are part of cutting-edge software packages for indexing and space group determination. Space groups of materials can also be determined using machine learning methods from their XRD data. Recently, Park et al. [114] showed that deep learning techniques can outperform rule-based programs without human involvement for space group determination from XRD data. The successful prediction of the crystal system of two novel inorganic compounds further confirms the potential of their method [114] in crystal structure determination. Another deep neural network algorithm [113] is proposed by F. Oviedo et al. to predict the space group and crystal dimensionality of materials through limited number of experimental thin-film XRD data. This method augments small datasets based on physics knowledge and their deep neural networks achieved high accuracy among other machine learning algorithms.

Despite the success of XRD based methods for materials structure determination, this is not a feasible solution for high throughput material screening, in which millions of possible elemental compositions need to be evaluated which makes experimental method to be expensive, time-consuming or just infeasible [56]. Next, the success of X-ray diffraction method is heavily reliant on the quality of X-ray diffraction results,

which is not always easy to achieve [163]. It also takes hours to acquire and analyze XRD data to recognize the crystal structure for each material [113].

Theoretically, given the chemical composition of a material, computational prediction of its crystal structure is possible. A couple of works utilize evolutionary algorithms or particle swarm optimization and DFT to determine crystal structures [107, 163]. USPEX [107] leverages evolutionary algorithm to find the most stable crystal structures, of which local optimization, spatial heredity, and lattice mutation are three key components to minimize the free energy. [163] searches the free-energy space by the Particle Swarm Optimization algorithm (PSO) within the evolutionary scheme together with ab initio structural optimization, symmetry constraint, and the geometrical structure parameter technique. GAtor [26] uses various setting of first principles calculation and genetic algorithms to increase the chance of locating the numerous low-energy minima. Despite their powerful prediction abilities, these first-principles based approaches are computationally demanding, which makes it impossible to perform high-speed screening for novel materials discovery. For example, it is shown that it takes tens of thousands of CPU hours to calculate 45 DFT calculations of formation energy [195].

In this paper, we propose a machine learning based method for predicting the space group and the crystal system for an inorganic material given only its composition information. Such models allow to conduct fast screening of millions of potential chemicals as done in [56]. We evaluate three types of features/descriptors: Magpie [165], atom vector [192], and one hot encoding (atom frequency) as the inputs of our machine learning algorithms. Neither XRD data nor DFT calculation is involved in feature calculations. Due to the fact that one composition may correspond to multiple crystal structures, four classifiers are developed to predict material structures in terms of the crystal system and space group: one-versus-all classifiers, multi-class classifiers, polymorphism classifiers, and multi-label classifiers. We leverage Multi

Layer Perceptron (MLP) and Random Forest (RF) to analyze how those feature sets can help determine the crystal structure using 10 fold cross-validation. By evaluating with different combinations of feature sets and machine learning techniques, we find that RF with Magpie features are the best in one-versus-all classification of space groups; one hot encoding is better than other two when classifying the multi-structure polymorphism and multiple space group labeling. Moreover, because most of the materials have a single crystal system or space group, we apply RF and MLP to assign these two labels to such materials. Our results indicate that RF with Magpie performs the best in determining the single crystal system or space group.

5.2 Methods

5.2.1 Datasets

We describe how we create the datasets for training and evaluating our prediction models. Our materials samples are extracted from the Materials Project [54], which is an extensive database that deposits the properties (e.g. crystallographic parameters, formation energy, band gap) of all known inorganic materials [54]. It is continuously growing and when we started this work, it contained 86,106 compounds in total. Table 5.1 summarizes the distribution of compounds as regard to the number of elements existent in the compounds. We find that the number of composition elements ranges from 2 to 8 and materials with 2, 3, 4, and 5 elements occupy 98.9% of the database (We exclude those materials of a single element).

We eliminate duplicate entries with identical formulas and space group information by keeping one sample for each such group. Besides, we remove a material (HeSiO2) that has None values in its Magpie features [165]. After this preprocessing, the total number of samples in our dataset is 60,636. These materials can be classified into 7 crystal systems and 223 space groups which we aim to predict. For each material, we generate three types of features merely based on its composition including Magpie,

Table 5.1 Distribution of materials with respect to the no. of elements

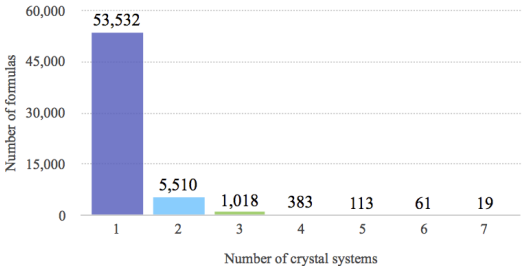
No. of Elements	No. of compounds
2	14,026
3	41,751
4	22,798
5	6,585
6	874
7	67
8	5

atom vector [192] and one-hot encoding (atom frequency) , which are detailed in next section.

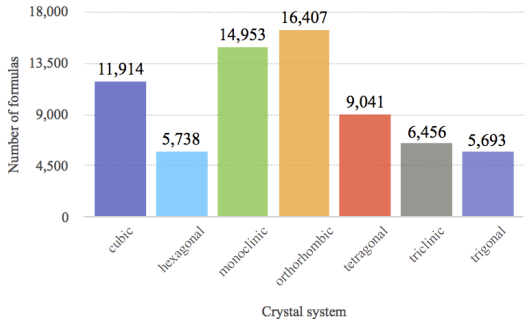
The goal of this paper is to develop classification algorithms to predict the crystal systems and space groups from materials compositions. Since each inorganic compound formula might correspond to materials with multiple different crystal systems or space groups, the crystal system/space group assignment problem can be mapped as a multi-label classification problem. To understand the distribution of samples in these crystal systems and space groups, Figures 5.1 and 5.2 show the distributions of samples in these categories. Figure 5.1a shows that most of the formulas in the dataset (88.3% or 53,532 formulas) have unique crystal systems. Similar observation applies to space group (85.7% or 51,988 formulas) as shown in Figure 5.2a. Among those formulas having multiple crystal systems, the number of 2-element formula is the largest, 3-element formula is the second, and few formulas have more than 3 elements. Figure 5.1b shows the distribution of materials in each crystal systems. We can find that the number of formulas is above 10,000 in orthorhombic, monoclinic and cubic systems. Besides, the number is close to 10,000 for tetragonal system. For the other three remaining systems, they have around 5,000 formulas.

There are 223 unique space groups in our dataset. Some formulas may correspond to materials with multiple space groups. Figure 5.2a shows the distribution of formula with different numbers of space groups. It shows that a majority of compositions

(51,993) only exist with one space group, and 5,977 formulas have 2 space groups. In Figure 5.2b, we can find that most of space groups have number of formulas less than 1,000. In our space groups classification problems, we only consider those space groups that have more than 1,000 formulas and the total number of space groups are 18. The space group symbols are Fm-3m, P2_1/c, Pnma, P-1, P1, C2/c, C2/m, Immm, Pm-3m, I4/mmm, P6_3/mmc, Ccmm, P4/mmm, R-3m, Cm2m, P2_1/m, Cm, and F-43m. From the space group and crystal system classification system, we find that 8 out of these top 10 space groups belong to the top 4 crystal systems [145]: 2,647 Immm and 3,891 Pnma belong to the orthorhombic crystal system, 5,220 P2_1/c, 2,707 C2/c, and 2,647 C2/m belong to the monoclinic crystal system, 6,171 Fm-3m and 2,142 Pm-3m belong to the cubic crystal system, and 2,124 I4/mmm belong to the tetragonal crystal system.

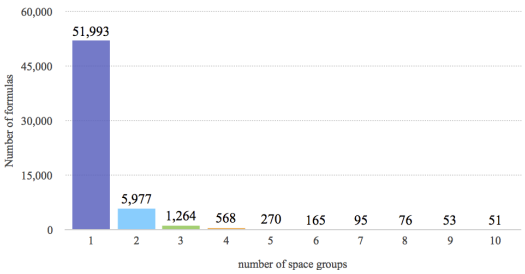


(a) The distribution of the number of crystal systems

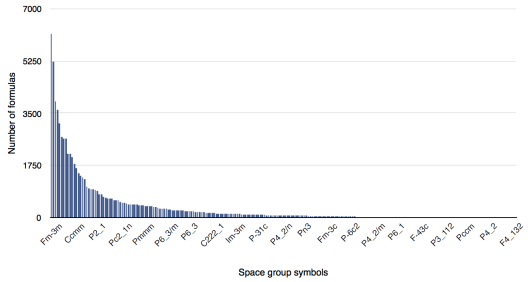


(b) The number of formulas in each crystal system

Figure 5.1 Distribution of crystal systems in Dataset



(a) Top 10 space group polymorphism



(b) The number of formulas in each space-group

Figure 5.2 Distribution of space groups in Dataset

We develop four types of classifiers to predict the crystal systems and space groups of the materials:

- one-versus-all classifier, which predicts whether a given composition/formula can form compounds of a specific crystal system or space group. We need to train one classifier for each crystal system or space group.
- multi-class classifier, which determines the single label for the materials with unique crystal system or space group. We only need to train one classifier for crystal system prediction and another classifier for space group prediction.
- polymorphism classifier, which predicts whether a composition can form compounds of multiple (>= 2) crystal systems or space groups.
- multi-label classifier, which predicts with what crystal systems or space groups a composition can form compounds.

5.2.2 Descriptors

The machine learning classifiers that we aim to develop are based on combinations of different machine learning algorithms and feature encodings. In this paper, we explore three kinds of features for predicting the crystal system and space group from materials compositions: Magpie [165], atom vector [192] and one-hot encoding (atom frequency). These features depend only on materials compositions or the formula themselves. In other words, we will not use any other structure information or physical properties calculated from first principle.

- Magpie features

  Magpie (Materials-Agnostic Platform for Informatics and Exploration) [165] is an extensive set of features related to the constituent elements in materials. The set covers a broad range of physical and chemical properties that fall into

four different categories: stoichiometric features, elemental property statistics, electronic structure features, and ionic compound features. Stoichiometric features only contain the number of elements in the compound and their several L^p norms. Elemental property statistics are calculated by computing several statistics (i.e., average, minimum, maximum, range, mean absolute deviation and mode) of 22 different elemental properties. Electronic structure features are the average fraction of s, p, d and f valence electrons [92]. Ionic compound features are the possibility of forming an ionic compound when we assume all elements present in a single oxidation state and two adaptions for calculating the fractions of a compound based on electronegativity [15]. Compared to the atom vector and one-hot encoding, Magpie is a general-purpose feature type that can be used to predict the properties of materials based on their formulas, e.g. it can describe the difference of heavy atoms and light atoms in a compound and link it to e.g. thermal conductivity prediction. Matminer [168] is used to retrieve the features and we remove the features with respects to crystal space group.

- Atom2vec

  Atom2vec [192] is a representation scheme for elements, which is calculated based on learning the relationship of elements among known materials. These learned properties are represented in terms of high-dimensional vectors for all elements. Atom-environment pairs are invented. The model maps the collection of all atoms in the environment to a feature vector for the composition. Suppose a $n \times d$ matrix $V = [v_1, v_2, \ldots, v_n]$ is given, where n is the number of atoms and d is the dimension of atom vector. Assume that the environment contains k atoms, then the environment vector can be represented as following:

$$E = C(v_1, v_2, \ldots, v_k)$$

where C is the summation over all atoms in our work. A score function (i.e., normalization score) is define as $S(v_i, E)$ which evaluates the likelihood of the target atom v_i appears with the environment E. Atom vectors are trained by maximizing the normalization score over the whole dataset. Compared to other representations, Atom2vec represents atoms in terms of high-dimensional vectors which capture how atoms relate to each other in high dimensional space. Based on the atom/element vectors calculated by atom2vec for all elements, for a given formula, we sum up all the atom vectors for the elements in the formula as the representation vector for the material.

- One-hot encoding (atom frequency)

  This encoding approach represents each compound by a vector of atom numbers of each element. We first count the frequency of atoms for each element in the given inorganic compound. Then a vector with 87 values is used to represent a formula since there are 87 unique elements in our dataset. Each component of the vector stores the frequency of a given element that exists in the formula or set to zero if a specific element is not available. Despite its simplicity, [56] shows that with large dataset and powerful models such as deep neural networks, even one-hot encoding can achieve highly predictive models.

5.2.3 Machine Learning methods

Two widely used machine learning algorithms including MLP and RF and three multi-label learning algorithms are evaluated in this study:

- We design two MLP structures. The first structure is for one-versus-all, multi-class and polymorphism classifiers. It has 11 layers and the numbers of nodes on hidden layers are 1024, 1024, 1024, 512, 512, 512, 256, 128, 64, 32, respectively. The second structure is only for multi-label classifier. It has 13 layers and

the numbers of nodes on hidden layers are 4096, 2048, 2048, 2048, 1024, 1024, 512, 512, 256, 256, 128, 32, respectively. The number of neurons on last layers of both structures is decided by the specific classifier. For instance, the last layer of first structure has 7 neurons in predicting multi-label crystal structure. ReLU [98] is used to activate neurons except for the last output layers. The activation function on last layers depend on the classification problem. We use Sigmoid for one-versus-all, multi-label and polymorphism classifiers and softmax normalization for multi-class classifier. It should be noted that two basic deep fully connected MLP architectures are used here due to their demonstrated performance in materials property prediction [56]. While more advanced deep neural networks such as convolutional or recurrent neural networks may be used and explored for each predicting task, however, tuning of model hyper-parameters and architectures is left for future work.

- RF [13] is a popular machine learning algorithm widely used in material informatics due to its robustness and capability to train with large datasets [98, 63, 146]. As an ensemble algorithm, RF aggregates the results of different decision trees (in our work, the number of decision tree is set as 50) to make more accurate models. Each decision tree is trained with a randomly selected subset of samples and features. The output of the final model either votes or averages the output of each decision tree depending on the specific task of regression or classification.

- Binary relevance (BR) [156, 37, 183] is considered as the most intuitive solution for multi-label classification. It transforms a multi-label problem into multiple independent binary learning problems. The number of independent binary classifiers is reliant on the number of unique labels in the dataset. Each binary classifier corresponds to one label in the label space. All binary classifiers are

trained on decomposed dataset. For example, we have 7 crystal systems and the dataset is decomposed into 7 datasets, of which labels in each dataset belong to one crystal system or not.

- ClassifierChain (CC) [126] is also a binary relevance method. However, CC differs from BR in that the feature space is augmented by the predictions of all previous binary classifiers in the chain. The added label information allows CC to take into account correlations among labels. If strong correlations exist in the label space, CC gives each base binary classifier relatively more predictive power.

- LabelPowerset [157] transforms the multi-label problem into a multi-class problem with one multi-class is trained on all unique label combinations formed in the dataset. In other words, it considers each combination in the power set as a single label in the dataset. This technique needs worst case of 2^L classifiers, where L is the number of labels in the label space. When L increases, the distinct label combinations can grow exponentially, which leads to memory and computing time explosion easily.

In addition, because of the imbalanced datasets, we investigate whether over-sampling method (i.e., Synthetic Minority Over-sampling Technique (SMOTE) [17]) improves the performance in predicting crystal systems and space groups using one-versus-all, multi-class and polymorphism classifiers. To illustrate how SMOTE works, we take Magpie features as an example. For minority class (e.g. cubic), we take a sample from the dataset and consider its k nearest neighbors in Magpie feature space. To synthesize a new sample, we take one sample from current sample and its k nearest neighbors. Then we multiply Magpie feature with a random real number between 0 and 1.

5.2.4 Evaluation metrics

Since our datasets are imbalanced, we use F1-score and Matthews correlation coefficient (MCC) to evaluate the performances of one-vs-all classifiers and polymorphism classifiers. F1-score is the harmonic mean of precision and recall with the maximum value of 1 and minimum value of 0 as the worst. MCC is also used to measure the quality of binary and multi-class classifiers, which takes into account of the balance ratios of true positive, true negative, false positive and false negative of the predictions. A MCC of 1 means prefect prediction, 0 is an average random guess, and -1 is inverse prediction.

In multi-label classification problems, a sample can be labeled with one or more categories. The predicted labels for each sample can thus be fully correct, partially correct, or fully incorrect. Traditional evaluation metrics such as precision or recall no longer apply to multi-label classifiers for performance evaluation. Thus, we re-defined the accuracy, precision, recall and F1-score to evaluate the performance according to [144]. In addition, we add Exact Match Ratio [144] as one additional performance measure. Assuming n is the number of samples and T_i and P_i are real and predicted labels that the sample i has, then the precision, recall, F1-score, and ExactMatchRatio can be calculated as follows:

$$
\begin{aligned}
Accuracy &= \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|T_i \cup P_i|} \\
Precision &= \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|P_i|} \\
Recall &= \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|T_i|} \\
F1-score &= \frac{1}{n}\sum_{i=1}^{n}\frac{2*|T_i \cap P_i|}{|T_i|+|P_i|} \\
ExactmatchRatio &= \frac{1}{n}\sum_{i=1}^{n} I(T_i = P_i)
\end{aligned}
$$

where accuracy is the intersection over union between real and predicted labels.

Precision is the average of the ratio of predicted correct labels over the total number of real labels. Recall is the average of the ratio of predicted correct labels over the total number of predicted labels. F1-score is the harmonic mean of precision and recall above. Exact Math Ratio is the proportion of entirely correct predictions over the total number of samples, where I is the indicator function.

We use 10 fold cross-validation to evaluate the performance of all classifiers composed of different machine learning algorithms and features sets. This evaluation strategy randomly splits the whole dataset into 10 equal partitions. Then for each fold, train a classification model over 9 of the 10 partitions and test the model over the remaining partition. The process is repeated until all 10 partitions are used as test sets once for each. The final performance is aggregated as the average performance over the whole dataset.

5.3 Results

5.3.1 Crystal system prediction

Materials crystal system prediction from composition using one-versus-all binary classifiers

For each of the 7 crystal systems, we train an one-versus-all binary classifier with the formulas of the selected crystal systems set as positive samples and all other samples as negative ones. The sample distribution for all crystal systems is shown in Figure 5.1b. Table 5.2 and Table 5.3 show the F1-scores and MCC for predicting crystal systems using RF and MLP, respectively. First, we find that RF achieves the highest performance with F1-scores ranging from 0.723 to 0.844 for all crystal systems except triclinic for which the RF+atom frequency encoding achieves the best performance with F1-score of 0.704 and MCC of 0.434. In comparison, the atom vector encoding works the worst among all three encoding methods with RF.

When we compare the performance of MLP with three encoding methods, it is found that the atom frequency encoding achieves the best performance for all the 7 crystal systems. Comparing the best combination of RF with Magpie with the best combination of MLP with atom frequency, the F1-score of RF with Magpie is slightly better than the MLP with atom frequency in predicting some crystal systems (e.g. cubic, hexagonal, monoclinic and tetragonal). For predicting orthorhombic, triclinic and trigonal, MLP with atom frequency outperforms RF with Magpie slightly. However, RF with Magpie is better than MLP with atom frequency overall in terms of Matthews correlation coefficient (MCC). Indeed, we find that atom vectors and atom frequency using MLP outperform their counterparts using RF and that RF and MLP using Magpie have close performance among all seven crystal systems. The possible reason is that atom vectors and atom frequency encode the internal connections inside a formula. Non-linear operations by MLP help to discriminate objects well. Plus, MLP can efficiently learns the mappings between the inputs and their labels. Since a F1-score of 0.844 is a relatively high score, this shows that the machine learning algorithms have done a good job in materials crystal system prediction from the compositions.

The results by over-sampling are shown in Tables 5.4 and 5.5. It can be found that with over-sampling, the best performance of RF has not been improved by a large margin and instead scores of MLP is decreased. The possible reason is that the ratios of between positive and negative labels of each dataset are between $\frac{1}{11}$ and $\frac{1}{4}$, which is acceptable to machine learning algorithms. On the contrary, one interesting finding is that the performance of RF with atom vectors is improved by SMOTE significantly. On average, both F1-score and MCC are increased by 0.05.

Table 5.2 Performance of RF for predicting crystal systems

Crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
cubic	**0.844/0.698**	0.753/0.538	0.775/0.457
hexagonal	**0.794/0.618**	0.647/0.374	0.704/0.433
monoclinic	**0.736/0.482**	0.670/0.360	0.730/0.467
orthorhombic	**0.729/0.485**	0.611/0.297	0.705/0.425
tetragonal	**0.797/0.623**	0.654/0.388	0.723/0.477
triclinic	0.686/0.412	0.644/0.337	**0.704/0.434**
trigonal	**0.723/0.498**	0.616/0.320	0.703/0.436

Table 5.3 Performance of MLP for predicting crystal systems

Crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
cubic	0.815/0.632	0.805/0.612	**0.830/0.660**
hexagonal	0.774/0.553	0.741/0.486	**0.781/0.566**
monoclinic	0.699/0.399	0.698/0.396	**0.732/0.465**
orthorhombic	0.692/0.385	0.689/0.380	**0.731/0.463**
tetragonal	0.767/0.536	0.743/0.488	**0.773/0.548**
triclinic	0.663/0.331	0.676/0.353	**0.709/0.421**
trigonal	0.701/0.409	0.705/0.412	**0.743/0.489**

Table 5.4 Performance of RF for predicting crystal systems by over-sampling

Crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
cubic	0.846/0.693	0.779/0.557	0.777/0.556
hexagonal	0.808/0.622	0.714/0.428	0.674/0.361
monoclinic	0.750/0.500	0.707/0.418	0.725/0.450
orthorhombic	0.739/0.485	0.667/0.336	0.698/0.405
tetragonal	0.803/0.613	0.720/0.441	0.695/0.409
triclinic	0.714/0.429	0.690/0.383	0.707/0.419
trigonal	0.742/0.494	0.680/0.360	0.693/0.402

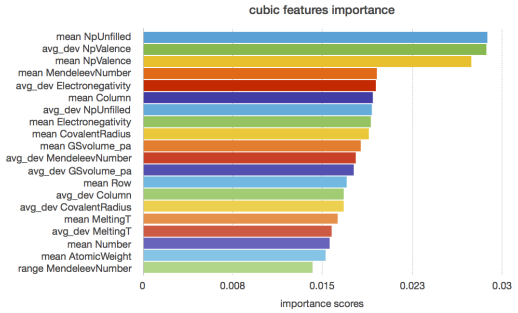
Table 5.5 Performance of MLP for predicting crystal systems by over-sampling

Crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
cubic	0.806/0.613	0.788/0.575	0.820/0.640
hexagonal	0.752/0.507	0.717/0.435	0.759/0.518
monoclinic	0.701/0.410	0.696/0.393	0.731/0.463
orthorhombic	0.682/0.365	0.678/0.358	0.727/0.454
tetragonal	0.749/0.501	0.725/0.450	0.757/0.517
triclinic	0.677/0.368	0.682/0.366	0.705/0.410
trigonal	0.683/0.372	0.686/0.372	0.722/0.445

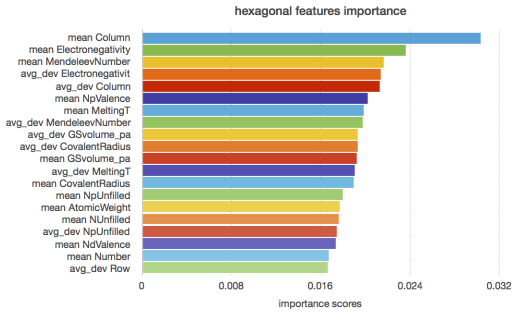
To show what features contribute most to the prediction of crystal systems, we calculate and rank the top 20 features by their feature importance scores for each crystal system when the RF with Magpie (the best classifier) is applied for classification. The results are shown in Figure 5.3 (i.e. from subfigure 5.3a to subfigure 5.3g). We find that shared important features include: mean and average deviation of melting temperature, mean and average deviation of Mendeleev number, mean and average deviation of covalent radius, mean and average deviation of GSvolume per pa, mean and average deviation of electronegativity, mean atom number, mean atomic weight, and mean Np valence. These features describe physical properties which are known to be involved in crystal system formation.

Crystal system prediction using multi-class prediction models

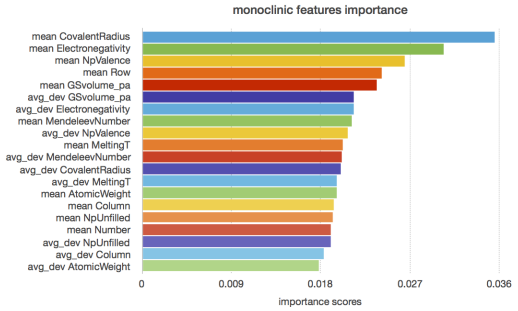
As shown in Figure 5.1a, 88.3% formulas (53,532 in total) have a unique crystal system. It is reasonable to develop a single classifier to assign the crystal system for a given composition, which is much more efficient than predicting its crystal system by running through 7 binary classifiers. Here we train one single RF classifier and MLP classifier for materials crystal system prediction for each encoding approach. We only choose the formulas with a single crystal system. A stratified 10-fold cross validation is used here for evaluating the classifiers. The 10-fold cross-validation results are shown in Table 5.6. Again, we find that RF with Magpie achieves the strongest



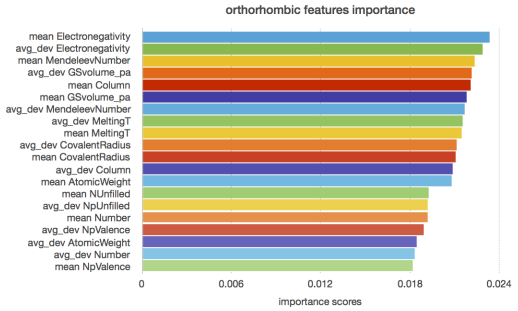
(a) cubic



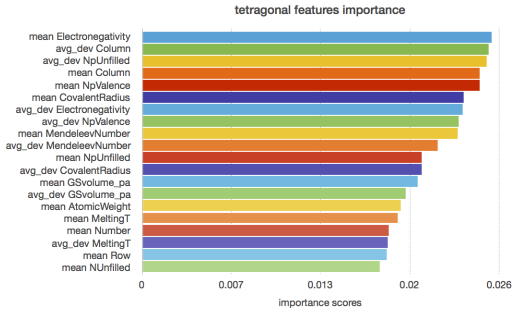
(b) hexagonal



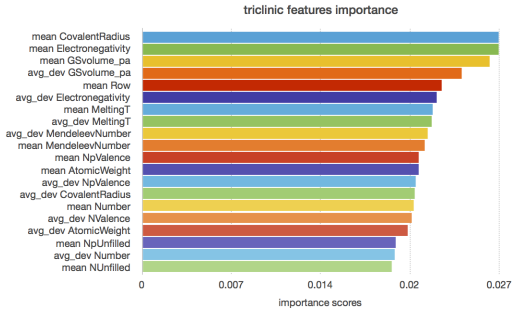
(c) monoclinic



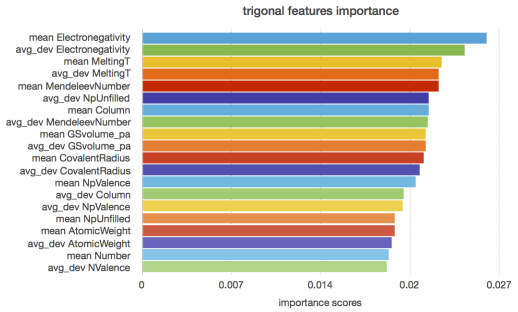
(d) orthorhombic



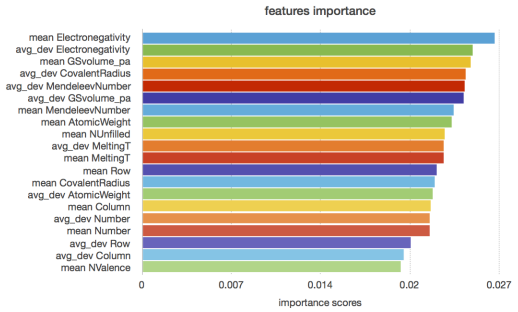
(e) tetragonal



(f) triclinic



(g) trigonal



(h) polymorphism

Figure 5.3 Ranking of Magpie Features for crystal system prediction

performance with F1-score and MCC of 0.650 and 0.591 against other combinations of models and feature sets. Compared with RF, we find that MLP is inferior for all three feature types. It is interesting that again, for MLP, the best encoding is atom frequency rather than Magpie which achieves the best performance with RF. It should be noted that while we have spent sufficient effort for tuning the MLP model parameters to maximize its performance, We find it is not easy to further significantly improve the MLP performances here by simple parameter tuning or structure modification. New descriptors and machine learning methods may be needed to improve the predictive performance. Besides, SMOTE shows inferior results across all combinations of learning methods and feature sets except for RF with atom vectors. The improvement for RF with atom vectors is only marginal.

Table 5.6 Performance for multi-class prediction of crystal system

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	**0.650/0.591**	0.511/0.445	0.575/0.511
MLP	0.559/0.486	0.559/0.489	0.615/0.551
RF-oversample	0.644/0.585	0.524/0.448	0.562/0.494
MLP-oversample	0.509/0.424	0.541/0.469	0.598/0.533

Crystal system polymorphism prediction using binary classifiers

Knowing whether or not a formula/composition can form compounds of multiple crystal systems is interesting to the materials community. Here, we select all formula with multiple crystal systems as positive samples (7,104 samples in total) while the remaining samples as negative ones (53,532 samples in total). Then, we train two binary classifiers using RF and MLP respectively to predict whether a given material composition can form multiple crystal systems or not. The 10-fold cross-validation results are shown in Table 5.7.

First, we find that MLP with atom frequency encoding achieves the best performance with F1-score of 0.704 and MCC of 0.409. The RF with atom frequency is the

second with F1-score of 0.668 and MCC of 0.354. In comparison, the MLP and RF with Magpie and MLP with atom vector achieve similar performance and are both much lower than those of RF and MLP+atom frequency. Over-sampling increases the F1-score of RF with all feature sets slightly but decreases the MCC of them. However, over-sampling decreases both F1-score and MCC of MLP with all feature sets.

Figure 5.3h shows the top 20 important features for crystal system polymorphism prediction. The shared features of mean and avg_devMendeleevNumber, mean and avg_dev GSvolume_pa, mean and avg_dev Electronegativity, mean and avg_dev MeltingT, mean and avg_dev CovalentRadius, mean Number, and mean Atomic Weight with one-versus-all case are keys for predicting crystal system.

Table 5.7 Performance for crystal system polymorphism prediction

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.652/0.350	0.610/0.293	0.668/0.354
MLP	0.646/0.308	0.642/0.289	**0.704/0.409**
RF-oversample	0.670/0.343	0.636/0.272	0.672/0.348
MLP-oversample	0.646/0.304	0.633/0.267	0.699/0.399

Crystal system prediction using multi-label Classifiers

It is known that many materials with different crystal systems can share the same formula or composition. So the crystal system prediction problem can be formulated as a multi-label classification problem. Here, we apply multi-label classifiers to explore how machine learning algorithms perform as regard to crystal system prediction from composition. We evaluate four multi-label prediction algorithms each with three encoding. The algorithms include MLP and three transformation algorithms (BinaryRelevance [157], ClassifierChain [126], and LabelPowerset [157]) for multi-label classification, all using the RF as the base classifier. 10 fold cross-validation is applied for performance evaluation. Table 5.8 shows the best results for each evaluated algorithm.

We find that LabelPowerset with Magpie and MLP with atom frequency achieve close performance and they are much better than other two transformation methods. LabelPowerset has the best performance with Exact MR of 0.598, Accuracy of 0.638, Precision of 0.673, Recall of 0.649, and F1-score of 0.652, of which Recall is 0.010 lower than MLP with atom frequency. BinaryRelevance has the worst results, which is reasonable. Because MLP, ClassifierChain and LabelPowerset take the internal label relationships into account in the label space. Instead, BinaryRelevance assumes an independent classifier for each label.

Table 5.8 Performance for multi-label crystal system prediction

	AF+MLP	Magpie+BR	Magpie+CC	Magpie+LP
Exact MR	0.579	0.469	0.534	**0.598**
Accuracy	0.631	0.504	0.568	**0.638**
Precision	0.660	0.531	0.601	**0.673**
Recall	**0.659**	0.516	0.574	0.649
F1-score	0.650	0.516	0.580	**0.652**

AF = atom frequency, BR = BinaryRelevance
CC = ClassifierChain, LP = LabelPowerset

5.3.2 Space group prediction

Determining the space group for a given material composition tells a lot of information about its physical properties. However, compared to 7 crystal systems, there are 223 space groups in total in the Material Project dataset, which makes it much more challenging to build the prediction models. Here, We select top 18 space groups, each having more than 1,000 compositions for exploring four classifiers for space group prediction from composition. We show the results of machine learning models for space group prediction as evaluated via 10-fold cross validation.

For each of the 18 space groups and with selected machine learning algorithm (RF or MLP) and selected encoding methods, we train 10 binary classifiers for 10 fold cross-validation. Together, we have trained 180 space group classifier. So instead of reporting the classifier performances for each of the space groups, we merely calculate the average F1-score and MCC for the 10 fold cross-validation performances of each space group over all space group categories and the results are shown in Table 5.9.

Table 5.9 shows the average F1-score and MCC over 18 space groups using RF and MLP with different materials encoding. Without over-sampling, we can find that RF with Magpie and MLP with atom frequency are the best combinations for predicting the space groups of inorganic materials using composition. The MCC of RF with Magpie is slightly better than that of MCC of MLP with atom frequency. However, F1-score of RF with Magpie is slightly worse than MCC of MLP with atom frequency. These scores are considerably lower compared with the performance scores for predicting crystal systems since there are much more categories of space groups than crystal systems. Both F1-score and MCC of RF are improved by over-sampling. The best combination becomes RF with Magpie after over-sampling. The scores for MLP are decreased slightly by over-sampling for all feature sets. For instance, F1-score and MCC of MLP with atom frequency are decreased to 0.753 and 0.508. The biggest improvement achieved by SMOTE is RF with atom vectors, whose F1-score and MCC are increased by 0.077 and 0.089, respectively.

Table 5.9 Average performance for predicting space group using RF and MLP

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.765/**0.566**	0.649/0.365	0.722/0.470
MLP	0.751/0.507	0.729/0.461	**0.768**/0.540
RF-oversample	0.787/0.579	0.726/0.454	0.725/0.459
MLP-oversample	0.743/0.493	0.718/0.437	0.753/0.508

Space group prediction using multi-class prediction models

Instead of building 18 binary classifiers for space group prediction, here we build a multi-class predictor for determining the space group given a material composition. We focus on the materials with a single space group. The stratified 10 fold cross-validation results are shown in Table 5.10.

Similar as in multi-class prediction for crystal systems, the combination of RF and Magpie features has the best performance with F1-score and MCC of 0.652 and 0.627 as shown in Table 5.10, respectively. In this case, the performance of each case is worse than the counterparts in the multi-class prediction of crystal systems. The possible reason is that the number of space groups is larger than the number of crystal systems so that the samples in each group are more sparse compared to crystal systems. Again, over-sampling slightly decreases the performance of all combinations except for RF with atom vector.

Table 5.10 Performance for multi-class prediction of space groups

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	**0.652/0.627**	0.519/0.501	0.576/0.556
MLP	0.571/0.540	0.540/0.517	0.616/0.591
RF-oversample	0.643/0.619	0.531/0.505	0.566/0.543
MLP-oversample	0.557/0.528	0.525/0.502	0.597/0.573

Space group polymorphism prediction using binary classifiers

Here we develop algorithms for predicting whether a material composition can form materials of multiple space groups. We set the compositions with multiple space groups in the dataset as positive samples, and the remaining as negative ones. Then RF or MLP based predictors combined with one of three encoding methods are evaluated in terms of their prediction power. The 10 fold cross-validation results are shown in Table 5.11.

It is found that MLP with atom frequency achieves the best result for predicting space group polymorphism with an F1-score and MCC of 0.670 and 0.342, respectively. RF with Magpie features by over-sampling achieves comparable but slightly lower performance (F1-score 0.651) as MLP with atom frequency. SMOTE improves the performance of all cases other than the combination MLP with frequency. MLP with Magpie and RF with atom vector have the largest improvement by SMOTE. Both scores are improved by 0.05 on average. Figure 5.4 shows the top 20 important features for space group polymorphism prediction. It is interesting but as expected that the top features here overlap a lot with those top 20 important features for predicting crystal systems. It means these features such as electronegativity, GSVolume, Mendeleev Number, play a critical role in predicting crystal symmetry.

Table 5.11 Performance for space group polymorphism prediction

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.610/0.273	0.540/0.147	0.614/0.253
MLP	0.582/0.205	0.591/0.190	**0.670/0.342**
RF-oversample	0.651/0.305	0.607/0.218	0.635/0.275
MLP-oversample	0.626/0.267	0.597/0.198	0.663/0.326

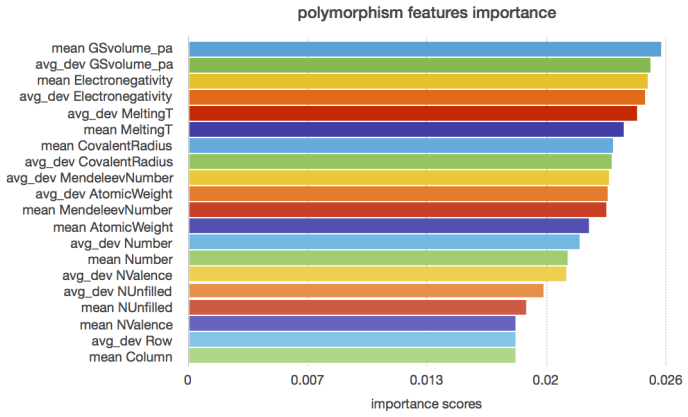


Figure 5.4 Magpie feature importance ranking for space group polymorphism prediction

Since each elemental composition may form materials of multiple different space groups, here we evaluate how current machine learning algorithms can predict all the multiple space groups for a given composition. Similar to multi-label predictions for crystal systems, we use BinaryRelevance [157], ClassifierChain [126] and LabelPowerset [157] plus MLP as multi-label classifiers, each evaluated with 3 features sets. 10 fold cross-validation results for the best combinations of algorithms and features sets are shown in Table 5.12. We can find that the performance of multi-label predictors for space group prediction is slightly inferior to their counterparts in the multi-label classification of crystal systems, which is expected due to the large number of space groups compared to the number of samples. Similar observations apply to space group predictions. LabelPowerset with Magpie has the best learning power and MLP with atom frequency comes next with close performance. BinaryRelevance is still the worst one due to the assumed independence of binary classifiers.

Table 5.12 Performance for multi-label space group prediction using MLP

	AF+MLP	Magpie+BR	Magpie+CC	Magpie+LP
Exact MR	0.569	0.446	0.472	**0.597**
Accuracy	0.612	0.467	0.491	**0.626**
Precision	0.633	0.485	0.510	**0.651**
Recall	0.634	0.472	0.495	**0.636**
F1-score	0.626	0.474	0.498	**0.637**

AF = atom frequency, BR = BinaryRelevance
CC = ClassifierChain, LP = LabelPowerset

5.4 Chapter Summary

We propose and evaluate machine learning algorithms for predicting the crystal systems and space groups of materials merely from their compositions. Two popular machine learning algorithms including random forests and multi-layered Perceptron neural networks combined with three material representations are evaluated for four types of

structure classification problems for both crystal system prediction and space group prediction: one-vs-all binary classification, multi-class classification, polymorphism prediction, and multi-label classification. Our extensive experiments show that the Random Forest with Magpie features achieves the highest performance for one-vs-all binary classification, multi-label prediction, and multi-class classification of both crystal systems and space groups. In contrast, Random Forest with atom frequency obtains the best results for polymorphism prediction of both crystal systems and space groups. However, the modest MCC scores of 0.591 and 0.627 for multi-class crystal system and space group prediction shows current machine learning algorithms and descriptors are far from achieving satisfactory performance, which calls for development of more advanced algorithms. One possible reason is that some artificial compounds have very high energy above hull, which might lead to unreasonable and misleading prediction over the crystal system and space group. In the future work, we may try to set a formation energy threshold to filter out those materials in the Material Project dataset. In addition, our feature importance analysis shows that electronegativity, covalent radius, Mendeleev number, melting temperature, GAS volume pascal, and mean atomic weight are crucial factors for predicting the crystal system and space group for a given material composition. Moreover, compared to the determination of crystal structures for materials, the performance for predicting space groups pales. That is because the data is distributed more unevenly over 18 space groups in our study, which may call for more advanced techniques to address this issue.

Our prediction models for crystal systems and space groups pave a way for performing large scale fast structural screening of materials when only compositions are available. This is especially true when compared to XRD data and DFT based approaches for space group determination, which is too expensive or slow for large-scale screening.

Chapter 6

Predicting Elastic Properties of Materials from Electronic Charge Density Using 3D Deep Convolutional Neural Networks

6.1 Introduction

As discussed in the Background chapter, designing a feature representation for materials are essential in Materials Informatics. In the past decade, a large number of descriptors have been proposed to encode materials [131, 16, 165, 159, 32, 130, 62, 7, 151, 30, 34, 92, 119],which is one of the most critical factors in machine learning applications in materials property prediction as shown in the review by Liu et al. [82]. In general, those descriptors are based on materials composition, their electronic or geometric structures as shown by the integrated feature calculation routines as implemented in the Matminer package [168]. A widely used set of material composition features is the Magpie features, which are based on the statistics of elemental properties in a material [165]. Mendeleev numbers (MN) has also been used by P. Villars et al. [159] to classify chemical systems by using the minimum and maximum MN versus the ratio between the minimum and the maximum MV. Ghiringhelli et al. [34] developed 23 primary features, based on atomic properties, to explore the energy difference of zinc blende, wurtzite, and rocksalt semiconductors. Han et al. [169] leveraged three key factors as the descriptors for classic machine learning methods to predict thermal conductivity effectively. Logan Ward et al. [165] presented a comprehensive set of features for a wide variety of material compositions. This set contains four unique categories: stoichiometric attributes, elemental property statistics, electronic structure attributes, and ionic compound attributes. Elemental descriptors have achieved great success in predicting band gaps [194], formation energies[58], crystal system[190], and etc. But these descriptors have their severe limitations: elemental descriptors are merely based on material compositions while most materials properties are strongly dependent on their atomic structures. There are also materials that share the same composition but exist in completely different structures. It is a common understanding that the most important information for analyzing a material's property is its structure.

How atoms coordinate and interact with each other conveys rich information on the properties of the materials. Therefore, structural features play a key role in developing prediction models of materials. Currently, there are several successful applications that use structural features to predict materials properties [16, 32, 130, 62, 7, 151, 30, 119]. Rupp and colleagues applied the Coulomb matrix (CM) features for predicting the atomization energies of small isolated organic molecules [131, 32, 130]. CM formulates the electrostatic interaction between nuclei into a matrix representation. Pham et al. [119] developed the orbital-field matrix (OFM) descriptor, based on the distribution of valence shell electrons, to predict formation energies and atomization energies with high accuracy. Bartók et al. [7] proposed the Smooth overlap of atomic positions (SOAP), which describes the similarity between two atomic environments to define a metric in the structural cell. The local similarity can be combined further to form a global measure of similarity for the evaluation of molecular properties [30]. More recently, voxel grid representation with atom features has been proposed to predict Hartree energies[62]. Atom density and related continuous representations have also been proposed for materials representation and are used for crystal structure generation[101, 171]. Graph neural networks have also been introduced to learn structural representation from material structures for predicting formation energy, band gaps, bulk modulus and etc with great success [176, 18]. On the other hand, deep learning has been utilized to extract three dimensional (3D) spatial features for material property prediction. In [16, 62], 3D CNNs have been applied to extract 3D geometric features from material microstructures represented as 3D matrices [16]. In this chapter, a dataset with 5,900 microstructures was created, where a microstructure is the quantification of the material structure. Each microstructure is represented by a feature matrix of dimension $51 \times 51 \times 51$, where each feature corresponds to a vector. Kajita et al. [62] developed a descriptor called Reciprocal 3D Voxel Space Descriptor (R3DVS) from the distributions of the valence electron density for 680

oxides. The authors enlarge the dataset by rotating R3DVS for testifying invariance of R3DVS to rotation and translation. R3DVS compacts the density of electrons in the bond generation.

In this chapter, we propose to leverage convolutional neural networks (CNNs) to learn physically meaningful features from the three dimensional electronic charge density (ECD) of materials for elastic property prediction. Since physical and chemical properties of materials are related to the transferability between atoms (nuclei) and the presence of electronic charges or electronic multipoles on atoms or molecules [48, 112, 122, 123], extraction of informative features from materials ECD can help predict materials properties. The ECD of a material can be calculated as a 3D matrix that describes the amount of electronic charge per volume. It represents the charge of electrons in the effective material space. By explicitly encoding the geometry of materials, ECD is supposed to have high transferability with respect to different compositions and structures [39]. As ECD captures both geometrical and electronic structural features, 3D distribution of electronic charge density would have the advantage over classical 1D and 2D descriptors as well as heterogeneous 3D structural descriptors in terms of the correlation with the electrochemical properties of materials. Indeed, ECD and its related electronic properties such as the electrostatic potential, electron localization function and non-covalent interaction index have been used to analyze many materials characteristics, including bonding, defects, stability,reactivity, and electron, ion and thermal transport [39]. For example, ECD was used to predict 8 materials properties by using the Fourier coefficients of the planar averaged Kohn-Sham charge density fingerprint features [120]. Abraham et al. [2] calculated 2D charge density to predict the chemical bonding and charge transfer in magnetic compounds. However both approaches failed to take advantage of the flexibility of the 3D representation [21]. Compared to conventional ML models, 3D CNNs can better link 3D descriptors to the properties efficiently as shown by [16]

(linkage between microstructure and homogenized property) and [62] (linkage between R3DVS and Hartree energies, testify the invariance to rotation for R3DVS). We believe that the unified representation of ECD makes it easier to learn unified continuous representation for facilitating downstream prediction tasks by deep convolutional neural networks [101]. In [129], the authors used the particle packing and the quartet structure generation set (QSGS) methods to generate microstructures of 3D composites. Instead of directly applying 3D CNNs to the 3D matrix, 2D CNNs are used to predict thermal conductivity of composites in the work by obtaining a series of cross-section slices from the 3D structure, which are stacked in order as the channel direction. This approach, however, may lead to too much global information loss in our ECD based elastic property prediction.

We explored two types of convolutional neural network models for ECD based elastic property prediction. One is the standard 3D CNNs with two convolutional layers. The other one is a projected 2D CNN models, in which the ECD matrix is converted to three different image-like representations which are then fed to three 2D CNN networks whose outputs are then fused together. The difference of 2D CNNs used in this chapter and in [129] is we compressed 3D structures from 3 directions and this strategy can preserve the global structural information to a large extent, compared to selecting some slices from the 3D matrix. This allows us to exploit the powerful hierarchical representation learning capability of 2D CNNs as demonstrated by their success in computer vision. [20, 141, 90, 149, 75]. We then conducted extensive benchmark experiments based on a dataset consisting of 2170 Fcc structured materials and 11 non-redundant datasets generated by leaving one-element-out at a time.

Our contributions can be summarized as follows:

- We propose to exploit the ECD descriptor as unified 3D materials representation and combine it with two types of 3D CNNs for materials elastic property

prediction. We also developed a fusion CNN model based on CNN+Magpie and CNN+ECD models.

- We developed a standard benchmark ECD dataset, named "FCC2170" calculated from 2710 Fcc Structured materials from ICSD. This database is characterized by its highly redundant samples with very similar compositions. We also developed 11 non-redundant datasets for evaluating the extrapolation capability of ECD+CNN models.

- We performed extensive prediction experiments over the aforementioned datasets using 5-fold cross validation. Our results show that our ECD+CNN can be complementary to elemental Magpie feature based models while can significantly outperform them over non-redundant datasets, thus demonstrating superior performance on some extrapolation experiments.

- We analyzed the situations when our ECD+CNN models perform better by visually inspecting the distribution of test samples and training samples in the 2D space mapped from the learned features.

- We validated the prediction performance of our models by comparison with DFT-calculated bulk and shear modulus for a set of 329 materials of the space group $Fm\bar{3}m$ collected from the Open Quantum Materials Database (OQMD) database.

6.2 Methods

6.2.1 Datasets

Here we discuss how we create the benchmark datasets for training and validating our proposed method. Due to the high computational cost to calculate electronic charge density for all the materials in the Materials Project database, we decide to focus only

on materials of one specific space group. First we retrieved 2170 material structures of $Fm\bar{3}m$ space group (excluding Lanthanides and Actinides) from the Materials Project (MP) database (https://materialsproject.org). We chose the $Fm\bar{3}m$ structure because its structure is simple and it takes less time to calculate the related elastic properties using DFT. Most materials of the $Fm\bar{3}m$ space group do not have the charge density or elastic properties available in the MP database. Hence, we calculated both the electronic charge density [6] and the elastic property [175] using VASP [74, 72, 73] for the 2,170 samples to form the "FCC2170" dataset . Table 6.1 lists top 11 elements that are contained in at least 200 materials of our FCC2070 dataset. Among them, most are from Group 1 (Lithium, Sodium, Potassium, Rubidium, and Caesium), Group 13 (Indium and Thallium) and Group 17 (Fluorine, Chlorine, and Bromine). The rest includes Scandium from Group 3. With this dataset, we then use the commonly used cross-validation method to evaluate our model's interpolation performance as done in most machine learning based property prediction studies [18].

To validate our model's extrapolation capability, we define a set of leave-one-element-out datasets, which are better for evaluating the extrapolation capability of ML models [178, 65, 91]. For all samples in FCC2170, we first select those samples containing one specific element E as the test set, and then designate the remaining samples as the training set. These datasets are called FCC-E-N datasets, where E is the element of interest and N is the number of training samples without element E. Statistics of all these non-redundant datasets generated from FCC2170 for elements contained in more than 200 materials are shown in Table 6.1.

Table 6.1 Statistics of non-redundant datasets

Element	F	K	Rb	Cs	Na	Cl
dataset	FCC-F-1775	FCC-K-1800	FCC-Rb-1802	FCC-Cs-1814	FCC-Na-1877	FCC-Cl-1880
train set size	1755	1800	1802	1814	1877	1880
test set size	415	370	368	356	293	290
Element	In	Br	Li	Sc	Tl	-
dataset	FCC-In-1937	FCC-Br-1938	FCC-Li-2148	FCC-Sc-1952	FCC-Tl-1966	-
train set size	1937	1938	2148	1952	1966	-
test set size	233	232	222	218	204	-

6.2.2 Representations of Materials

We studied and compared two material representations for elastic property prediction including Magpie [165] and electronic charge density(ECD) [140].

- Magpie features

  Magpie (Materials-Agnostic Platform for Informatics and Exploration) is an extensive set of features related to the constituent elements in materials. The set covers a broad range of physical and chemical properties that fall into four different categories: stoichiometric features, elemental property statistics, electronic structure features, and ionic compound features [165]. Stoichiometric features only contain the number of elements in the compound and their several L^p norms [165]. Elemental property statistics are calculated by computing several statistics (e.g., average, minimum, maximum, range and mode) of 22 different elemental properties [165]. These properties include row and column on the periodic table, average atomic number, the range of atomic radii between all elements presenting in compositions, Mendeleev number, atomic weight, covalent radius, electro-negativity. Electronic structure features are the average fraction of s, p, d and f valence electrons [92]. Ionic compound features include the capability of forming an ionic compound (when we assume all elements present in a single oxidation state) and two adaptions for calculating the fractions of a compound based on electronegativity [15].

- Electronic charge density

  ECD in the form of 3D structural matrix represents the spatial distribution of electronic charge density in crystalline materials. It can be calculated by local quantum-mechanical functions related to the Pauli exclusion principle [140]. The *ab initio* method is used to calculate Hartree-Fock wavefunction and the electron

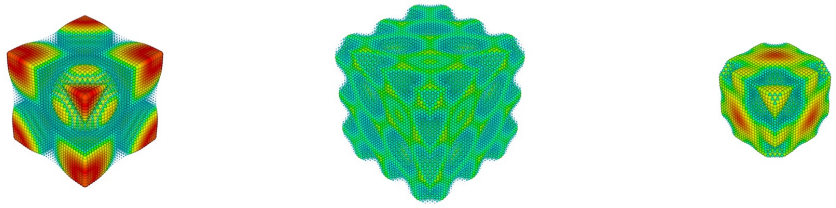
localization function (ELF) [6]. A single determinant wave function is calculated on a grid in the 3D space by hartree Fock or Kohn Sham orbitals φ_i as following:

$$ELF = \frac{1}{1 + (\frac{D}{D_h})^2} \tag{6.1}$$

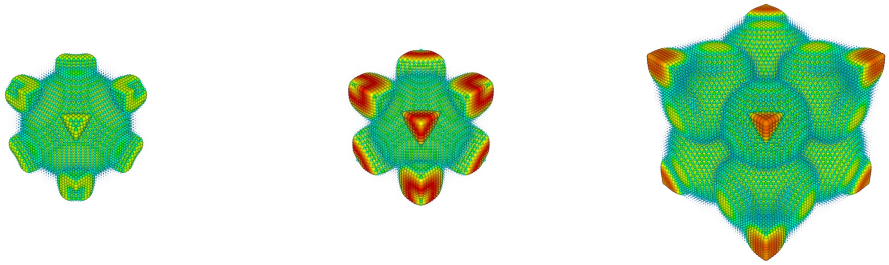
where

$$\begin{aligned} D &= \frac{1}{2}\sum_i |\Delta\varphi_i|^2 - \frac{1}{8}\frac{|\Delta\rho|^2}{\rho} \\ D_h &= \frac{3}{10}(3\pi^2)^{5/3}\rho^{5/3} \end{aligned} \tag{6.2}$$

where ELF has values between 0 and 1, where 1 means the perfect localization. Figure 6.1 shows the visualizations for the ECDs of six representative materials, namely $SrCaIn_2$, $Mn_{23}C_6$, $VSiOs_2$, RbI, CsBr, and Rb_2TeBr_6, where $SRCaIn_2$, $Mn_{23}C_6$, and $VSiOs_2$ possess high bulk modulus. These visualizations consist of points that correspond to the values in a material's ECD matrix. The color and area of each point represents the size of each value and together show the distribution of a material's electron clouds. When the value of these points are plotted, we found that points appear in both thick and thin clouds, within the cubes, as shown in subfigures 6.1a, 6.1b, and 6.1c. Subfigures 6.1d, 6.1e, and 6.1f show a clear difference from the top-row figures. In these figures, there are some empty spaces in the cubes and some dense clusters present in the remaining area. These observations correspond to the physical reality that materials with high bulk modulus usually have active electrons orbiting across the whole space strongly when compared to materials with lower bulk modulus. Among all six materials, we find that although the ECD visualizations share many similar characteristics, there are a few distinct differences between them. These minor variations make it possible for us to employ 3D CNNs to learn the structural and physical patterns that may characterize the material's elastic properties.



(a) $SrCaIn_2$ ($32 \times 32 \times 32$) (b) $Mn_{23}C_6$ ($40 \times 40 \times 40$) (c) $VSiOs_2$ ($24 \times 24 \times 24$)



(d) RbI ($30 \times 30 \times 30$) (e) CsBr ($30 \times 30 \times 30$) (f) $VSiOs_2$ ($48 \times 48 \times 48$)

Figure 6.1 Visualization of ECDs for six materials showing clearly contrasting structural features (top and bottom rows). $l \times w \times h$ is the actual length, width and height of each ECD matrix.

6.2.3 Machine Learning Methods

In this chapter, we use Random Forest and Convolutional Neural Networks (CNNs) with Magpie features as the baseline methods. We propose that CNNs with ECD can capture certain characteristic relationships between material structures and their elastic properties.

Random Forest (RF) [13] is a widely used machine learning model in material informatics because of its high accuracy and robustness [152, 113, 167]. As an ensemble learning algorithm, a RF aggregates the results from different decision trees (50 in this chapter). The decision trees are randomly trained based on subsets of training samples and features. Within a decision tree, a set of decision rules (e.g. Melting temperature > 200.0) is learned by minimizing the variance of the decision tree. For

predicting elastic properties, RF calculates the final results by averaging outputs of all decision trees.

Convolutional Neural Networks are a type of feed-forward neural network interleaved with convolutional, pooling, and fully connected layers. It has achieved state-of-the-art(SOTA) performance when applied to computer vision and natural language processing [141, 75, 66]. The convolutional unit is the core building block of CNNs, which is inspired by the multi-layered organization of the visual cortex. The unit consists of multiple learnable filters with a given receptive field and weight parameters. In our case, the filters are convolved across the full depth of the input volume of the ECD [79]. The filters are learned hierarchically, where low-level features generate more condensed representations. The computational unit can be constructed by a transformation $U = F_{tr}(X), X \in \mathbb{R}^{L' \times W' \times H' \times C'}, U \in \mathbb{R}^{L \times W \times H \times C}$. F_{tr} denotes the convolutional operation. Let $V = [v_1, v_2, \ldots, v_C]$ be the learnable convolutional filters. Then the outputs of F_{tr} can be written as $U = [u_1, u_2, \ldots, u_C]$, where

$$u_c = v_c * X = \sum_{i=1}^{C'} v_c^i * x^i \tag{6.3}$$

Here $*$ denotes the dot product, $v_c = [v_c^1, v_c^2, \ldots, v_c^{C'}]$, $X = [x^1, x^2, \ldots, x^{C'}]$. We removed the bias terms for simplicity. v_c^i is a 3D spatial filter convolving on a single channel of X. Stacked outputs of filters produce a 4D tensor activation map [79]. A pooling layer is used to do non-linear downsampling. It partitions the 3D input into a set of rectangular boxes. In max-pooling, the pooling layer outputs the maximum value of each sub-region. Then a 3D tensor is activated through a rectified linear unit (ReLu) [75]. The ReLu operation can be denoted by $max(0, P)$, where P is the tensor generated by the max-pooling operation. The same procedure can be applied repeatedly to the whole activation map. Finally, the output of the convolutional layers is fed to one or more fully connected layers to accomplish the regression step. Similar procedures are applied in the CNN block in Figure 6.3.

We implemented two types of convolutional neural networks for learning ECD based features for elastic property prediction. Figure 6.2 depicts the 3D CNN architecture in our work. The "Scalar" stands for the bulk or shear modulus. The numbers above each convolutional layers are its parameter settings. For instance, $200@(5 \times 5 \times 5)$ means 200 filters with size of $5 \times 5 \times 5$. Unless it is specified, the stride is always the same with the filter size. Two consecutive convolutional layers are followed by a max-pooling with pooling size and strides both of $2 \times 2 \times 2$. The number below are outputs of each layer. For fully connected layers, the numbers above them are the number of neurons. This model has two consecutive convolutional layers followed by a max pooling layer, and then seven fully connected layers. For simplicity, we did not show the ReLu [75] activation for all neural layers in Figure 6.2. The filter size of 2 convolution layers are $5 \times 5 \times 5$ and $4 \times 4 \times 4$, respectively and the stride has the same size as that of the convolution filters. For all max pooling layers, the sizes of filters and strides are $2 \times 2 \times 2$. The ECD matrices are fed to the 3D convolutional and pooling layers, and then the output matrix is flattened and passed to succeeding fully connected layers to calculate final predictions.

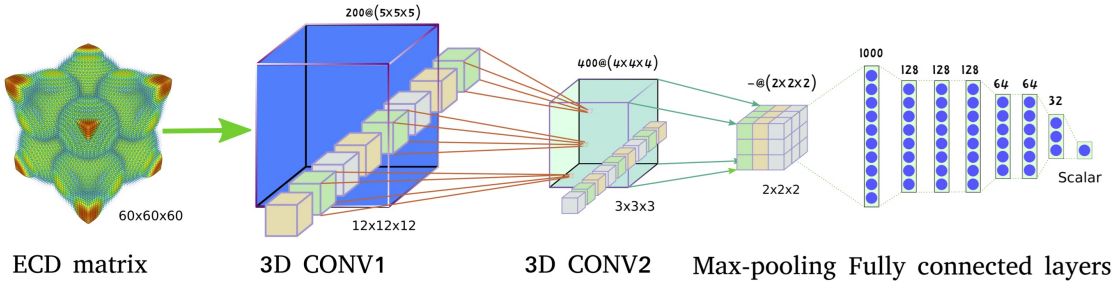


Figure 6.2 The architecture of 3D CNN with ECD representation.

Figure 6.3 shows the architecture of our 2D CNNs for elastic property prediction. The "Scalar" stands for the bulk or shear modulus. The model includes three parts: mainframe, SE block and CNN block. In the mainframe, we have three branches whose outputs are concatenated and fed into six fully connected layers. The numbers above each component/layer are the number of neurons of that layer. In SE block,

the labels of **R** and **S** are reshape and channel-wise multiplication operations. For simplicity, we ignore the max-pooling layers following every convolutional layer in the CNN block. Numbers below each component are the output dimension of that layer. The ECD matrix does not have the concept of channel as images. Thus, we rotated the ECD matrix so that we can face the cube from x,y,z axes as shown in different colors of cubes in Figure 6.3. Then the direction facing to us is considered as the channel direction. To model the inter-dependencies between channels, we used the Squeeze-and-Excitation (SE) network [52], which can exploit this inter-dependency by feature recalibration. This model selectively highlights the informative features and suppresses less useful ones. A SE block is shown in the left corner of Figure 6.3. In this module, 24 filters of size of 1×1 are used to down-sample the ECD matrix, which was first proposed in [81]. A nonlinearity operation is performed on each pixel across the channels. After the nonlinear projection, the ECD matrix X of size $60 \times 60 \times 60$ is reduced to the feature map U of size $60 \times 60 \times 24$. A global average pooling is then used to shrink the feature map into a vector of size 24 along with the dimensions of width and height. Then we use a small set of fully connected layers to transform this vector into higher level features. The number of neurons on each layers are 4 and 24, respectively. The output s of the last fully connected layer is reshaped into size of $1 \times 1 \times 24$. The last step is nonlinear excitation and the final output U' of block is achieved by rescaling the U with the activated s:

$$U' = U \odot \sigma(s) \tag{6.4}$$

where σ is the Sigmoid activation function that implements the nonlinear transformation. And $\odot$ denotes the channel-wise multiplication between the scalar s and the feature map U'.

The SE block in our 2D CNN architecture is followed by CNN blocks. A CNN block has two regular convolutional layers followed by a max-pooling layer. The first convolution neural has the same filter size and strides of 6×6 and there are 64 filters

in this layer. The second CNN layer has a filter size of 5×5 and stride of 2×2 and there are 128 filters in total. All max-pooling layers have the same pooling size and strides of 2×2, respectively.

For each of the projection map of x, y, and z, there is a SE and CNN block for feature extraction. The outputs of them are concatenated into a vector of size 384. Six fully connected layers are then used to map this learned features into elastic property values. The number of neurons on these fully connected layer are 4096, 4096, 128, 128, 128 and 32 respectively.

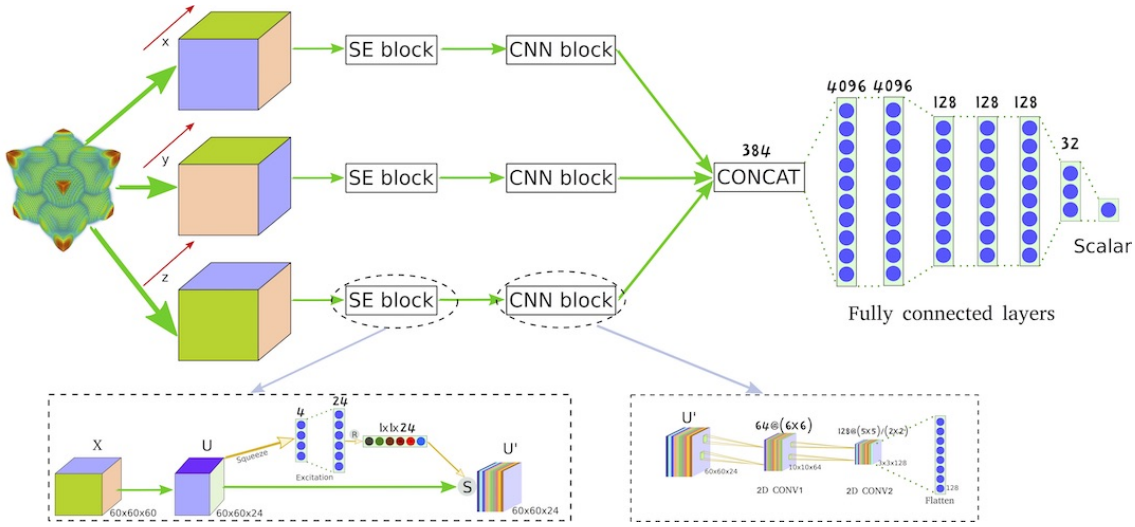


Figure 6.3 The architecture of the 2D CNN with ECD representation.

For the baseline algorithm, we also train a 2D CNN model with the Magpie features. To do that, we append 12 zeros to the Magpie features to get a vector of 1x144, which is then reshaped into a 2D matrix of size 12×12. The CNN model for Magpie features has two consecutive convolutional layers followed by an average pooling layer. Then an additional convolutional layer is added followed by two fully connected layers. The model parameters are set as follows: the kernel size and strides of the first convolutional layer are 3x3 and 1x1 and the number of filters is 32; the kernel size and strides of the second convolutional layer are 3x3 and 1x1 and the number of filters is 48; the pooling size and strides of the average pooling layer are

both set as 2x2; the kernel size and strides of the third convolutional layer are 3x3 and 1x1 and the number of filters is 64; the number of neurons of the two fully connected layers are 48 and 32, respectively.

6.2.4 Training and Implementation

Figure 6.2 shows the detailed architecture of our 3D CNN and its parameters. The models are implemented using the open-source libraries of TensorFlow (https://www.tensorflow.org) and Keras (https://keras.io). The performance of the models are evaluated using 5-fold cross validation. The input ECD has a shape of $60 \times 60 \times 60$ by interpolation for smaller matrices. The CNN for Magpie is also trained using the Adam optimizer [67] with a batch size of 32 and learning rate of 0.001. The 3D CNN model parameters are learned using the Adam optimizer [67] with a initial learning rate of 0.0005. For the 2D CNNs with ECD, we use the SGD optimizer to learn the model parameters. The initial learning rate is 0.001 and it drops by $0.5^{\lfloor \frac{epoch}{10} \rfloor}$, where *epoch* is the current epoch. The mean absolute error (MAE) is used as the loss function for all three CNN models. The open source matminer (https://hackingmaterials.lbl.gov/matminer/) is utilized to calculate the Magpie features.

6.3 Results and Discussions

In this section, we discuss the experiments demonstrating the potential of ECD for material representation and elastic property prediction. The experiments are separated into two parts in terms of the evaluation approaches: experiments with 5-fold cross validation and experiments focusing on extrapolation performance evaluation. All experiments of CNN models are repeated 5 times and the result presented herein is the average of their outputs.

6.3.1 5-fold cross validation experiments with redundant dataset

Table 6.2 shows the results from 5-fold cross validation on the whole dataset with 2170 samples. We find that the baseline models using Magpie features are better than CNNs with ECD across all evaluation metrics for predicting bulk and shear moduli. Overall, RF with Magpie performs slightly better than CNNs with Magpie. Although R^2 of RF with Magpie is 0.001 lower than that of CNNs with Magpie in predicting bulk modulus, RF with Magpie achieves much better results in predicting shear modulus (R^2 is 0.049 higher). Similar observations apply to performance evaluated in terms of Root Mean Square Error (RMSE). This better performance of Magpie based RF models are not unexpected. First, all samples in this FCC2170 dataset belong to the $Fm\bar{3}m$ space group. By sharing similar structures, the Magpie features are able to capture most of the elastic property variation due to composition difference. The high structural similarity of the dataset helps the baseline methods based on composition Magpie features predict the elastic properties well. Another reason is that the FCC2170 contains many similar samples in terms of compositions. The high redundant samples also makes the baseline models with Magpie features to make precise predictions by exploiting redundant neighbor samples in the training set when evaluated on the test set during cross-validation. However, the machine learning models trained with a redundant training set can lead to low extrapolation performance as shown in our previous study [178]. In terms of dimension size of the Magpie and ECD descriptors, ECD has a much larger dimension of $60 \times 60 \times 60$ compared to 132 of the Magpie features. Since higher input data dimensions usually lead to machine learning models with more parameters, and more training samples are needed to achieve good prediction performance. From this perspective, the limited dataset size in our problem actually favors the baseline models with Magpie features.

Here we show that ECD can be used as a complementary materials descriptor for elastic property prediction together with the Magpie features. To verify this, We

pre-trained a CNN model with Magpie features and a 2D CNN model with ECD. Then we fused these two models by concatenating the outputs of the penultimate layers of these two models to generate a output latent feature vector of dimension 64, which is then fed to three fully connected layers with 128, 64, and 32 neurons respectively. The Adam optimizer [67] is used for training with a learning rate of 0.001. This fusion neural network model with mixed Magpie and ECD descriptor yielded the best R^2 and RMSE of 0.955 (0.804) and 16.530 (15.780) in predicting bulk (shear) modulus respectively as shown in Table 6.2 . This confirms that ECD and Magpie can work together to achieve better performance for elastic property prediction. In addition, our experiments also showed that the projected 2D CNN achieved significantly better performance than the basic 3D CNN models. The R^2 and RMSE of 2D-CNN with ECD are 0.912 and 23.401 in predicting bulk modulus compared to 0.884 and 26.819 of 3D-CNNs with ECD. The R^2 and RMSE of 2D CNN with ECD are 0.768 and 17.192 in predicting shear modulus compared to 0.745 and 17.944 of 3D-CNNs with ECD.

Table 6.2 Performance Comparisons of models with Magpie and ECD descriptors using 5-fold cross validation

Type	RF+Magpie		CNN+Magpie		3D-CNN+ECD		2D CNN+ECD		Fusion	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
bulk	0.943	18.721	0.944	18.423	0.884	26.819	0.912	23.401	**0.955**	**16.530**
shear	0.794	16.142	0.745	17.959	0.745	17.944	0.768	17.192	**0.804**	**15.780**

Table 6.3 Extrapolation prediction performance comparison on non-redundant leave-one-element-out datasets

Elem	Type	RF+Magpie		CNN+Magpie		3D-CNN+ECD		2D-CNN+ECD		CGCNN	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
F	bulk	-0.529	26.797	-0.809	29.102	**-0.051**	**22.212**	-0.448	26.080	-2.217	35.554
	shear	-3.350	18.117	-6.912	24.315	-1.202	12.878	-1.293	13.151	**-0.548**	**10.657**
K	bulk	**0.776**	**6.067**	0.646	7.573	0.510	8.969	0.570	8.397	0.474	9.055
	shear	**0.810**	**2.641**	0.548	4.014	0.389	4.733	0.367	4.817	0.146	5.523
Rb	bulk	0.867	4.603	**0.869**	**4.579**	0.753	6.287	0.777	5.966	0.275	10.290
	shear	**0.778**	**2.767**	0.727	3.064	0.608	3.657	0.719	3.111	0.268	4.944
Cs	bulk	-0.128	11.232	**0.760**	**5.166**	0.448	7.818	0.067	10.158	-0.144	10.934
	shear	-4.327	11.199	**0.492**	**3.446**	0.014	4.743	-1.137	7.083	0.344	3.881
Na	bulk	0.630	16.398	**0.833**	**11.013**	0.660	15.708	0.616	16.689	0.605	16.223
	shear	0.545	8.366	0.386	9.716	**0.548**	**8.340**	0.451	9.196	0.351	9.863
Cl	bulk	0.410	15.935	0.529	14.151	0.591	13.009	**0.716**	**11.05**	0.534	13.119
	shear	-0.477	10.715	0.213	7.765	**0.339**	**7.160**	0.093	8.394	-0.197	9.366
In	bulk	**0.829**	**20.780**	0.780	23.550	0.725	26.326	0.773	23.908	0.761	24.460
	shear	0.791	8.250	0.771	8.618	0.683	10.136	**0.793**	**8.207**	0.655	10.416
Br	bulk	0.921	4.464	0.923	4.585	0.912	4.700	**0.923**	**4.411**	0.631	9.245
	shear	0.630	2.290	-0.078	3.857	0.755	1.861	**0.824**	**1.579**	-2.661	6.975
Li	bulk	0.418	29.869	**0.867**	**14.253**	0.519	27.142	0.454	28.937	0.732	20.121
	shear	-0.232	17.799	0.416	12.239	0.428	12.126	**0.451**	**11.881**	0.388	12.488
Sc	bulk	0.855	23.276	**0.908**	**18.538**	0.756	30.195	0.850	23.688	0.818	25.983
	shear	**0.781**	**12.996**	0.707	15.024	0.682	15.650	0.635	16.786	0.667	16.007
Tl	bulk	-0.370	24.574	0.421	15.973	0.219	18.529	0.501	14.833	**0.550**	**14.040**
	shear	0.456	6.745	0.437	6.815	**0.559**	**6.068**	0.557	6.084	0.427	6.818
# of the best		**5**		**6**		**4**		**5**		2	

6.3.2 Extrapolation Experiments with non-redundant datasets

ML models with elemental descriptors such as Magpie can achieve good cross-validation performance for datasets consisting of redundant (computationally very similar samples) such as FCC2170. However, the better performance of the fusion model with CNN with Magpie and 2D-CNN with ECD implies that for the ECD descriptor can help to make better predictions over a certain subset of test samples. In this section, we aim to construct non-redundant dataset and show that our CNN models with the ECD descriptor can achieve better performance on non-redundant datasets or for test samples with few highly similar neighbor samples.

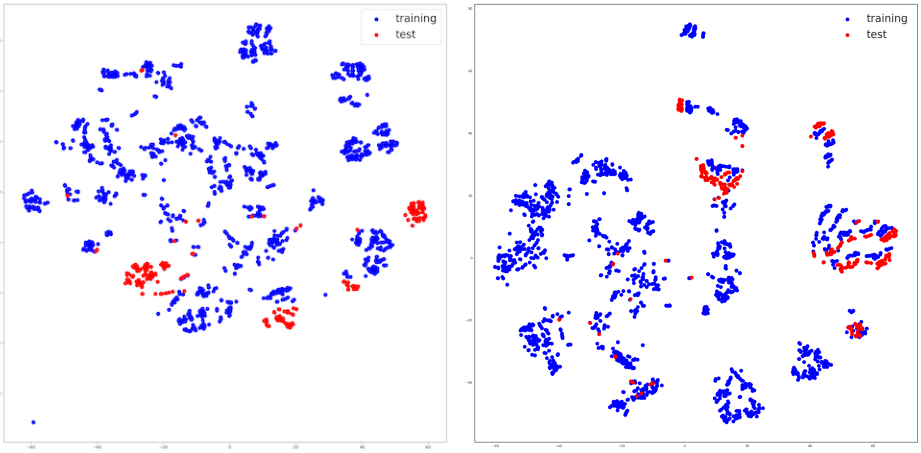
For these extrapolation experiments, we trained and tested the prediction models over the FCC-E-N datasets as described in Section 6.2.1. The performance comparison results of the extrapolation experiments for bulk and shear modulus prediction are shown in Table 6.3. There are 22 sets of experiments with 11 of them for predicting bulk modulus and the other 11 for predicting shear modulus by five different algorithms including RF+Magpie, CNN+Magpie, 3D-CNN+ECD, 2D-CNN+ECD, and the latest crystal graph convolutional neural network (CGCNN) [176], which also uses structural information. We highlighted the best performance scores for each experiments and count how many experiments each algorithm achieved the best scores. As shown in Table 6.3, the RF with Magpie and CNN with Magpie worked the best for 5 and 6 experiments respectively. However, impressively, for these non-redundant training/testing experiments, our ECD descriptor based 3D-CNN-ECD and 2D-CNN-ECD outperformed the others for 4 and 5 experiments respectively, which reflecting the importance of the structure based ECD descriptor for elastic property prediction. In contrast, the popular CGCNN only achieved the best performance out of 2 experiments, which demonstrated the advantage of our ECD based atomic structure representation.

6.3.3 Visualization Study of when ECD descriptor works better

To understand on why our ECD based CNN models worked better than Magpie features on some datasets but not others, we conducted a visualization study for all the extrapolation experiments. For magpie features, we directly apply the t-distributed Stochastic Neighbor Embedding (t-SNE) [88] to the dataset. For the ECD based features, directly applying t-SNE is not feasible due to the memory limit. So we first applied max-pooling to the 3D ECD matrices with strides of $(6, 6, 6)$ and pooling size of $(6, 6, 6)$ before feeding them into t-SNE. Hence the final size of the ECD matrices is $(10, 10, 10)$, which are then flattened to a 1D vector of 1,000 elements. Then we applied t-SNE to this 1D vector to reduce the dimension to 2.

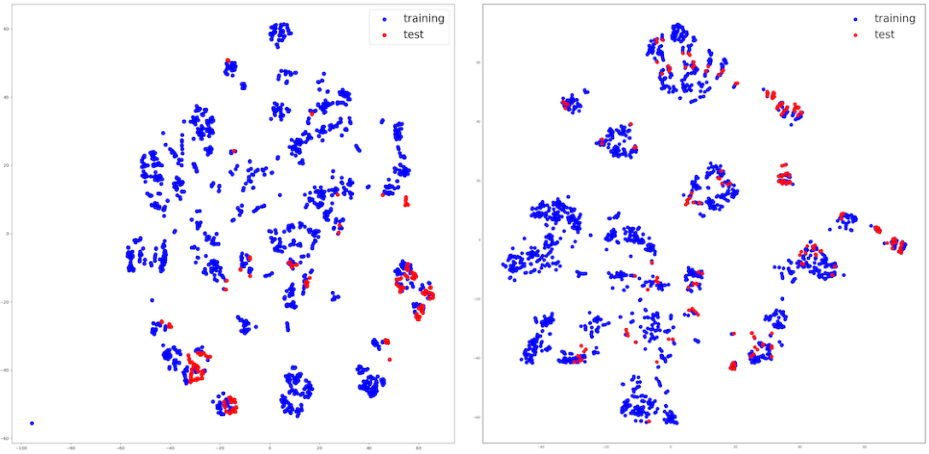
Figure 6.4 shows 2D visualization of the high-dimension Magpie and ECD features for two datasets: FCC-Chlorine-1880 and FCC-Thallium-1966 over which the ECD based models outperform Magpie feature based models. The training samples are labelled as blue points while the test samples are red points. First, Figure 6.4 (a) and (b) show the distribution of training and test samples with Magpie features and with ECD features respectively for the FCC-Chlorine-1880 dataset. In subfigure 6.4a, we found that there exist three large clusters of test samples (red points) that have few similar training samples around. This corresponds to the low prediction performance for Magpie based models. The best performance for both bulk and shear modulus prediction is achieved by CNN+Magpie with R2 of 0.529 and 0.213 respectively. In contrast, subfigure 6.4b shows the 2D distribution of the samples represented with ECD features. It can be found that the test samples are mostly mixed with training samples, leading to much better prediction performance: the best performance for bulk modulus prediction is achieved by 2D-CNN+ECD with R2 of 0.716, which is significantly better (35%)than 0.529, the best prediction performance achieved by Magpie based models. The best performance for shear prediction is achieved by 3D-CNN+ECD with R2 of 0.339, which is also 59% better than 0.213, the best R2

score of Magpie based models.



(a) 2D map of Magpie features for FCC-Cl-1880 dataset

(b) 2D map of ECD features for FCC-Cl-1880 dataset



(c) 2D map of Magpie features for FCC-Tl-1966 dataset

(d) 2D map of ECD features for FCC-Tl-1966 dataset

Figure 6.4 Visualization of high-dimensional features for elements Chlorine and Thallium by t-SNE. Blue dots are training data and red dots are test data.

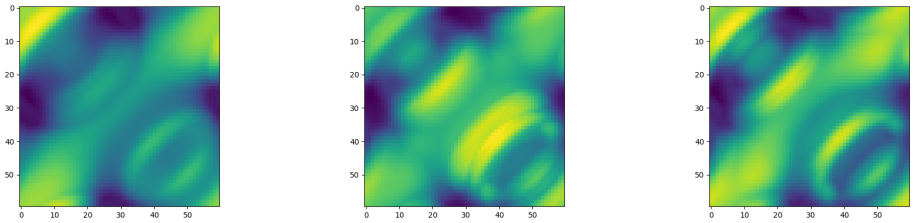
Figure 6.4 (c) and (d) show the distribution of training and test samples with Magpie features and with ECD features respectively for the FCC-Thallium-1966 dataset. In subfigure 6.4c, we found that clusters of test samples (red points) are closer to training samples compared to subfigure 6.4a. There is no large clusters of isolated test samples. The best performance for bulk modulus is achieved by CNN+Magpie with R2 of 0.421. The best performance for shear modulus prediction

is achieved by RF+Magpie with R2 of 0.456. In contrast, subfigure 6.4d shows the 2D distribution of the samples represented with ECD features. It can be found that the test samples are better mixed with training samples than subfigure 6.4a, leading to better prediction performance. The best performance for both bulk modulus prediction is achieved by 2D-CNN+ECD with R2 of 0.501 and the best shear modulus prediction performance is achieved by 3D-CNN+ECD with R2 of 0.559. In this dataset, the best ECD based model is (0.559-0.421)/0.421 = 19% better than the best Magpie based model for bulk modulus prediction. The performance gap is much smaller compared to that (35%) on the FCC-Chlorine-1880 dataset. The best ECD based model is also (0.559-0.456)/0.456 = 24.9% better than the best Magpie based model for shear modulus prediction, which is however much smaller than the performance gap over the FCC-Chlorine-1880 dataset, which is 59%. These findings can partially explain why ECD based models are superior to Magpie based models in predicting elastic properties for these two datasets. It shows the structure based ECD descriptor can be a complementary descriptor to elemental Magpie features for elastic property prediction due to their better neighborhood structure of the samples. This analysis is consistent to those observation that neighbor sample distribution significantly affects the performance of neural network based prediction models [55].

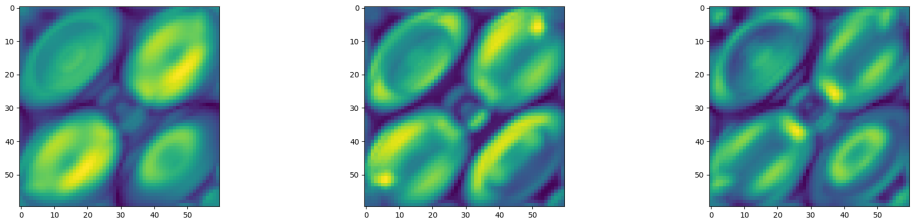
6.3.4 Visualization of averaged SE block outputs

Figure 6.5 shows the visualization of the average output of the 24 channels of the SE block as shown in Figure 6.3 from x, y, and z directions. On the top row, bright and dark areas/patterns overlap together in the $SrCaIn_2$ with high bulk modulus with fuzzy boundaries. However, clear boundaries (four clear ovals in each direction) between dark and bright regions can be found on the bottom row, which are the patterns for the material with low bulk modulus. These findings are consistent with the patterns as we have discussed in the sub-section 6.2.2. This distinct patterns

extracted by 2D CNNs help to differentiate materials and effectively predict their elastic properties. Moreover, although visualizations for three directions share many similar patterns for the same materials, there are variations among them. For example, the darkness of subfigures 6.5d, 6.5e, and 6.5f is different. Among them, overall Figure 6.5e has the darkest area and 6.5d has the brightest ones. We believe that the slight variations detected by the 2D CNNs might be one of the reasons that 2D CNNs outperform 3D CNNs in predicting elastic properties.



(a) x direction (b) y direction (c) z direction



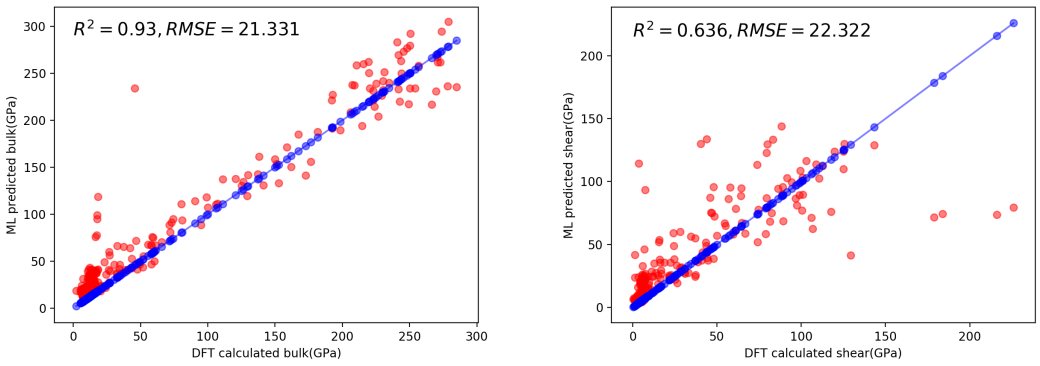
(d) x direction (e) y direction (f) z direction

Figure 6.5 Visualization of average output of 24 channels of the SE block for three directions for $SrCaIn_2$ and K_3YI_6.

6.3.5 DFT validation

To further validate our neural network models, we predict the bulk and shear modulus of a set of external materials from the OQMD [132] database and compare them to DFT calculated ones. We first collect all the materials of the space group $Fm\bar{3}m$ from OQMD and then remove the duplicates existing in the Material Project database that we used as the training set. We also filter out the materials having more than 40 atoms in the unit cell. We finally obtain 329 materials as our test set. Then

we apply the trained fusion model (Magpie + ECD features) trained with Material Project samples to predict the bulk and shear modulus of the 329 samples in the test set and compared them with DFT-calculated ones as shown in Figure 6.6. We find that our fusion model successfully predicted the bulk modulus for the 329 materials with good alignment with DFT calculated values. The R^2 and RMSE in predicting bulk are 0.93 and 21.331 as shown in Figure 6.6a. However, we also find that the ML-predicted the predicted shear modulus values deviate much more from the DFT calculated ones compared to the bulk modulus, which reflects the fact that it is more difficult to predict shear modulus than bulk modulus. We also observe that most of the deviations of the predicted values compared with DFT calculated ones are from the regions with low bulk or shear modulus and the predicted values usually are larger than the DFT calculated ones.



(a) (b)

Figure 6.6 Panels (a) and (b) show ML-predicted versus DFT-calculated bulk and shear modulus respectively.

6.4 Chapter Summary

We propose to combine deep convolutional neural networks and electronic charge density (ECD) for materials elasticity prediction. We demonstrate that the ECD descriptor can be used to predict bulk and shear modulus with CNNs model. We created a benchmark dataset named "FCC2170" with 2,170 materials of $Fm\bar{3}m$ space

group from Materials Project database and derived 11 non-redundant leave-one-element-out datasets for benchmarking the proposed ML models with ECD and elemental Magpie features. Our computational experiments showed that due to the structural similarity among the samples of the FCC2170 dataset, the elemental Magpie feature with CNN models achieved the best results, which however, can be enhanced by the fusion models with both Magpie and ECD features. In addition, our benchmark studies on the non-redundant datasets showed that the structure-based ECD feature with CNNs can achieve better extrapolation prediction performances over half prediction tasks out of the total 22 experiments for prediction bulk and shear modulus.

To further understand the power of the ECD descriptor, we visualized the distribution of training and test datasets of two descriptor types using t-SNE. It shows that when the training set and testing set of the non-redundant datasets have higher level of mixing, the Magpie-based CNN models work better. When they have lower level of mixing, the ECD descriptor based models significantly outperform the Magpie based CNN models. The results demonstrate the importance of structure based features for achieving higher extrapolation and generalization prediction capability. It is expected that our ECD descriptor with CNN models can also be applied to a variety of problems in material science, especially with the development of algorithms for predicting ECD [39]. Currently, we are generating more ECD dataset with more space groups to extend this method to more materials with diverse structures.

Chapter 7

Conclusions

7.1 Conclusion

In this dissertation, we introduce our work in crystal structure generation, crystal system prediction, and elastic properties prediction. For the first work, we propose CubicGAN, a novel WGAN-based model to generate cubic crystals on a large scale. We represent crystals with lattice parameters, base atom coordinates, element properties, and space group encoding. Elements and space groups work as the input of the generator to control the type of generated crystals. We train our models for ternary and quaternary materials with space groups of $Fm\bar{3}m$, $F\bar{4}3m$, and $Pm\bar{3}m$ from OQMD. We generate 10 million samples and perform CIFs readability, charge neutrality, and negative formation energy checks. It is found that our method not only can re-discover most cubic materials in existing databases, but can also create 24 and 1 prototype(s) for ternary and quaternary materials, respectively. After performing DFT simulations on 108,897 hypothetical materials, four new prototypes with stable materials are confirmed by dispersion calculation.

On the second topic, we integrate physical insights (atom distance) and symmetric constraints into the generation of materials. Extensive experiments show that our model outperforms two baseline models in terms of several widely used evaluation metrics. With the atom clustering and merging process, the physics guided crystal generative model can generate high quality crystal structures. After further rigorous DFT calculations, 1869 out of 2000 randomly selected materials are relaxed/optimized successfully, of which 39.6% are with negative formation energies.

On the third topic, we use machine learning to build the relationship between materials composition and the crystal system/space group. We compare Random Forest with multi-layer perceptron using three different feature representations: atom embedding, Magpie, and atom one-hot encoding. Due to the many-to-many relationship between composition and crystal system/space group of materials, four types

of classification (one-vs-all binary, multi-class, polymorphism and multi-label) are explored. Our extensive experiments show that Random Forest with Magpie achieves the best performance in one-vs-all binary classification, multi-label and multi-class classification for crystal systems and space groups. On the contrary, Random Forest with atom one-hot encoding is the best for polymorphism determination of crystal systems and space groups. In addition, we analyze the feature importance of Magpie and the results show that electronegativity, covalent radius, Mendeleev number, melting temperature, GAS volume pascal, and mean atomic weight play a critical part in determining the space groups and crystal systems of compositions. Our method paves a new way to quickly screen structures of materials when compositions are only available.

In the last work, we introduce a new 3D representation termed electronic charge density (ECD) for materials property prediction. We use 3D and 2D convolutional neural networks to predict elastic properties, in which convolutional operations can extract hierarchical features from the ECD. The results show that our method can achieve good prediction performance for bulk and shear properties. In particular, the CNN model based on the fusion of Magpie and ECD achieves the best results. We then validate our best model by comparing the predicted elastic properties with DFT calculated ones and find that our model can successfully predict the properties with respective R^2 and RMSE of 0.93 and 21.331 for bulk and 0.636 and 22.322 for shear.

7.2 Future work

7.2.1 An unified framework for generative design of materials

No current crystal generative models are able to generate any types of inorganic materials. They either generate a special family of materials (e.g., Mb-Mn-O system) or materials with a part of specific space groups. In reality, the number of elements in existing materials is from 1 to 9. The distribution of the number of materials in

230 space groups is severely biased. Some space groups have thousands of materials but other space groups only have several materials in the existing materials databases. Moreover, the number of atoms in a unit cell can be one-digit or even three-digit. Different methods, such as padding, are used to represent the various materials as vectors of fixed sizes. But padding is inefficient in generating materials because it introduces noise in the representations of materials.

Therefore it is needed for us to design an unified framework to generate materials in the full spectrum of space groups and elements. Right now, we have three options in hand: (i) include the number of atoms and the number of elements in the generative process; (ii) use electronic charge density as the representation for materials; and (iii) apply score based generative modeling.

7.2.2 Incorporating Additional Physical Properties Based Losses

As we mentioned before, finding stable materials is essential in material science. In PGCGM, we integrate atom distances and space group symmetry into the generative losses. Although the two losses have increased the ratio of successfully relaxed/optimized materials, stability related properties are needed in generating process. We have optimized thousands of ternary materials. In next move, we firstly will train a CGCNN model to build a relationship between the generated materials and formation energy/energy above hull. Then we will embed the pre-trained CGCNN model to predict formation energy/energy above hull into the generator. With a energy based loss, it will push the generator to generate materials with higher stability.

Bibliography

[1] Martıén Abadi et al. "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 2016, pp. 265–283.

[2] Jisha Annie Abraham, Gitanjali Pagare, and Sankar P Sanyal. "Electronic structure, electronic charge density, and optical properties analysis of GdX3 (X= In, Sn, Tl, and Pb) compounds: dft calculations". In: *Indian Journal of Materials Science* 2015 (2015).

[3] Luciano A Abriata, Giorgio E Tamò, and Matteo Dal Peraro. "A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments". In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019), pp. 1100–1112.

[4] Angela Altomare et al. "EXPO2009: structure solution by powder data in direct and reciprocal space". In: *J. Appl. Crystallogr.* 42.6 (2009), pp. 1197–1202. DOI: 10.1107/S0021889809042915.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[6] Kurt Artmann. "Berechnung der Seitenversetzung des totalreflektierten Strahles". In: *Annalen der Physik* 437.1-2 (1948), pp. 87–102.

[7] Albert P Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Physical Review B* 87.18 (2013), p. 184115.

[8] Alec Belsky et al. "New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design". In: *Acta Crystallogr., Sect. B: Struct. Sci.* 58.3 (2002), pp. 364–369. DOI: 10.1107/S0108768102006948.

[9] G Bergerhoff, ID Brown, F Allen, et al. "Crystallographic databases". In: *International Union of Crystallography, Chester* 360 (1987), pp. 77–95.

[10] Guenter Bergerhoff et al. “The inorganic crystal structure data base”. In: *J Chem Inf Comput Sci* 23.2 (1983), pp. 66–69. DOI: 10.1021/ci00038a003.

[11] Ali Boultif and Daniel Louër. “Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method”. In: *J. Appl. Crystallogr.* 24.6 (1991), pp. 987–993. DOI: 10.1107/S0021889891006441.

[12] Y-Lan Boureau, Jean Ponce, and Yann LeCun. “A theoretical analysis of feature pooling in visual recognition”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 111–118.

[13] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[14] Keith T Butler et al. “Computational materials design of crystalline solids”. In: *Chemical Society Reviews* 45.22 (2016), pp. 6138–6146.

[15] William D Callister and David G Rethwisch. *Materials science and engineering: an introduction*. Vol. 7. John Wiley & Sons New York, 2007.

[16] Ahmet Cecen et al. “Material structure-property linkages using three-dimensional convolutional neural networks”. In: *Acta Materialia* 146 (2018), pp. 76–84.

[17] Nitesh V Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *JAIR* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[18] 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.

[19] Lihua Chen et al. “Machine learning models for the lattice thermal conductivity prediction of inorganic materials”. In: *Computational Materials Science* 170 (2019), p. 109155.

[20] Xiaozhi Chen et al. “Multi-view 3d object detection network for autonomous driving”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1907–1915.

[21] Hwanho Choi et al. “Predicting the Electrochemical Properties of Lithium-Ion Battery Electrode Materials with the Quantum Neural Network Algorithm”. In: *The Journal of Physical Chemistry C* 123.8 (2019), pp. 4682–4690.

[22] François Chollet et al. *Keras*. https://keras.io. 2015.

[23] S Chung et al. “Evidence for a large phononic band gap leading to slow hot carrier thermalisation”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 68. 1. IOP Publishing. 2014, p. 012002.

[24] Callum Court et al. “3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning”. In: *Journal of Chemical Information and Modeling* (2020).

[25] Stefano Curtarolo et al. “AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations”. In: *Computational Materials Science* 58 (2012), pp. 227–235.

[26] Farren Curtis et al. “GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction”. In: *J. Chem. Theory Comput.* 14.4 (2018), pp. 2246–2264. DOI: 10.1021/acs.jctc.7b01152.

[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. “Computational screening of all stoichiometric inorganic materials”. In: *Chem* 1.4 (2016), pp. 617–627. DOI: 10.1016/j.chempr.2016.09.010.

[30] Sandip De et al. “Comparing molecules and solids across structural and alchemical space”. In: *Physical Chemistry Chemical Physics* 18.20 (2016), pp. 13754–13769.

[31] Carl Doersch. “Tutorial on variational autoencoders”. In: *arXiv preprint arXiv:1606.05908* (2016).

[32] Felix Faber et al. “Crystal structure representations for machine learning models of formation energies”. In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1094–1101.

[33] Fractional coordinates. *Fractional coordinates — Wikipedia, The Free Encyclopedia*. [Online; accessed 11-November-2021]. 2021. URL: https://en.wikipedia.org/wiki/Fractional_coordinates#cite_note-3.

[34] Luca M Ghiringhelli et al. “Big data of materials science: critical role of the descriptor”. In: *Physical review letters* 114.10 (2015), p. 105503.

[35] Maria Giatsoglou et al. “Sentiment analysis leveraging emotions and word embeddings”. In: *Expert Systems with Applications* 69 (2017), pp. 214–224.

[36] 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.

[37] Shantanu Godbole and Sunita Sarawagi. “Discriminative methods for multi-labeled classification”. In: *PAKDD*. Springer. 2004, pp. 22–30. DOI: 10.1007/978-3-540-24775-3_5.

[38] Stefan Goedecker. “Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems”. In: *The Journal of chemical physics* 120.21 (2004), pp. 9911–9917.

[39] Sheng Gong et al. “Predicting charge density distribution of materials using a local-environment-based graph convolutional network”. In: *Physical Review B* 100.18 (2019), p. 184103.

[40] 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).

[41] Rhys EA Goodall and Alpha A Lee. “Predicting materials properties without crystal structure: Deep representation learning from stoichiometry”. In: *Nature Communications* 11.1 (2020), pp. 1–9.

[42] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[43] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[44] Ishaan Gulrajani et al. “Improved training of wasserstein gans”. In: *Advances in neural information processing systems*. 2017, pp. 5767–5777.

[45] Geoffroy Hautier et al. “Data mined ionic substitutions for the discovery of new compounds”. In: *Inorganic chemistry* 50.2 (2011), pp. 656–663.

[46] 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.

[47] Xingfeng He et al. “Crystal Structural Framework of Lithium Super-Ionic Conductors”. In: *Adv. Energy Mater.* (2019), p. 1902078. DOI: 10.1002/aenm.201902078.

[48] Graeme Henkelman, Andri Arnaldsson, and Hannes Jónsson. “A fast and robust algorithm for Bader decomposition of charge density”. In: *Computational Materials Science* 36.3 (2006), pp. 354–360.

[49] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.

[50] Yoyo Hinuma et al. “Band structure diagram paths based on crystallography”. In: *Computational Materials Science* 128 (2017), pp. 140–184. ISSN: 0927-0256. DOI: https://doi.org/10.1016/j.commatsci.2016.10.015. URL: http://www.sciencedirect.com/science/article/pii/S0927025616305110.

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

[52] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[53] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

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

[55] Jon Paul Janet et al. “A quantitative uncertainty metric controls error in neural network-driven chemical discovery”. In: *Chemical Science* (2019).

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

[57] Dipendra Jha et al. “Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning”. In: *Nat. Commun.* 10 (2019), p. 5316. DOI: 10.1038/s41467-019-13297-w.

[58] Dipendra Jha et al. “Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning”. In: *Nature communications* 10.1 (2019), pp. 1–12.

[59] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N Schmidt. "Neural message passing with edge updates for predicting properties of molecules and materials". In: *arXiv preprint arXiv:1806.03146* (2018).

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

[61] Yousung Jung. "Machine learning approaches for materials discovery: Predictive and generative models". In: *Telluride Workshop Machine Learning and Informatics for Chemistry and Materials*. Telluride Science Research Center. 2018.

[62] 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), p. 16991.

[63] Kyoungdoc Kim et al. "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds". In: *Phys. Rev. Mater.* 2.12 (2018), p. 123801. DOI: 10.1103/PhysRevMaterials.2.123801.

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

[65] Yoolhee Kim et al. "Machine-learned metrics for predicting thelikelihood of success in materials discovery". In: *arXiv preprint arXiv:1911.11201* (2019).

[66] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[67] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[68] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[69] Scott Kirklin et al. "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies". In: *npj Comput. Mater.* 1 (2015), p. 15010. DOI: 10.1038/npjcompumats.2015.10.

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

[71] Vadim Korolev et al. "Machine-learning-assisted search for functional materials over extended chemical space". In: *Materials Horizons* 7.10 (2020), pp. 2710–2718.

[72] Georg Kresse and Jürgen Furthmüller. “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set”. In: *Computational materials science* 6.1 (1996), pp. 15–50.

[73] Georg Kresse and Jürgen Furthmüller. “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set”. In: *Physical review B* 54.16 (1996), p. 11169.

[74] Georg Kresse and Jürgen Hafner. “Ab initio molecular dynamics for liquid metals”. In: *Physical Review B* 47.1 (1993), p. 558.

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[76] Armel Le Bail. “Monte carlo indexing with mcmaille”. In: *Powder Diffr.* 19.3 (2004), pp. 249–254. DOI: 10.1154/1.1763152.

[77] Yvon Le Page and Paul Saxe. “Symmetry-general least-squares extraction of elastic data for strained materials from ab initio calculations of stress”. In: *Phys. Rev. B* 65 (10 Feb. 2002), p. 104104. DOI: 10.1103/PhysRevB.65.104104. URL: https://link.aps.org/doi/10.1103/PhysRevB.65.104104.

[78] Yann LeCun, Yoshua Bengio, et al. “Convolutional networks for images, speech, and time series”. In: ().

[79] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[80] Yuxin Li et al. “MLatticeABC: generic lattice constant prediction of crystal materials using machine learning”. In: *arXiv preprint arXiv:2010.16099* (2020).

[81] Min Lin, Qiang Chen, and Shuicheng Yan. “Network in network”. In: *arXiv preprint arXiv:1312.4400* (2013).

[82] Yue Liu et al. “Materials discovery and design using machine learning”. In: *Journal of Materiomics* 3.3 (2017), pp. 159–177.

[83] Yunsheng Liu et al. “Tailoring the Cation Lattice for Chloride Lithium-Ion Conductors”. In: *Advanced Energy Materials* 10.40 (2020), p. 2002356.

[84] J Lloberas et al. “A review of high temperature superconductors for offshore wind power synchronous generators”. In: *Renewable and Sustainable Energy Reviews* 38 (2014), pp. 404–414.

[85] Teng Long et al. “CCDCGAN: Inverse design of crystal structures”. In: *arXiv preprint arXiv:2007.11228* (2020).

[86] Teng Long et al. “Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures”. In: *npj Computational Materials* 7.1 (2021), pp. 1–7.

[87] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. “Rectifier nonlinearities improve neural network acoustic models”. In: *Proc. icml*. Vol. 30. 1. Citeseer. 2013, p. 3.

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

[89] Sreekanth Madisetty and Maunendra Sankar Desarkar. “A neural network-based ensemble approach for spam detection in Twitter”. In: *IEEE Transactions on Computational Social Systems* 5.4 (2018), pp. 973–984.

[90] Daniel Maturana and Sebastian Scherer. “Voxnet: A 3d convolutional neural network for real-time object recognition”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 922–928.

[91] 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.

[92] Bryce Meredig et al. “Combinatorial screening for new materials in unconstrained composition space with machine learning”. In: *Phys. Rev. B* 89.9 (2014), p. 094104. DOI: 10.1103/PhysRevB.89.094104.

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

[94] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[95] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014).

[96] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. and Topf, M. “Critical assessment of techniques for protein structure prediction, fourteenth round”. In: *CASP 14 Abstract Book* (2020). URL: https://www.predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf.

[97] Gregory J Mulholland and Sean P Paradiso. "Perspective: Materials informatics across the product lifecycle: Selection, manufacturing, and certification". In: *Apl Materials* 4.5 (2016), p. 053207.

[98] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *ICML-10*. 2010, pp. 807–814.

[99] Pinku Nath et al. "High throughput combinatorial method for fast and robust prediction of lattice thermal conductivity". In: *Scripta Materialia* 129 (2017), pp. 88–93.

[100] Marcus A Neumann. "X-Cell: a novel indexing algorithm for routine tasks and difficult cases". In: *J. Appl. Crystallogr.* 36.2 (2003), pp. 356–365. DOI: 10.1107/S0021889802023348.

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

[102] Juhwan Noh et al. "Machine-enabled inverse design of inorganic solid materials: promises and challenges". In: *Chemical Science* 11.19 (2020), pp. 4871–4881.

[103] Adelaide M Nolan et al. "Computation-accelerated design of materials and interfaces for all-solid-state lithium-ion batteries". In: *Joule* 2.10 (2018), pp. 2016–2046.

[104] Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. "CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks". In: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2019.

[105] Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. "Crystalgan: learning to discover crystallographic structures with generative adversarial networks". In: *arXiv preprint arXiv:1810.11203* (2018).

[106] AR Oganov et al. "Crystal structure prediction using the USPEX code". In: *CECAM-Workshop Lausanne*. 2012, pp. 22–26.

[107] Artem R Oganov and Colin W Glass. "Crystal structure prediction using ab initio evolutionary techniques: Principles and applications". In: *J. Chem. Phys.* 124.24 (2006), p. 244704. DOI: 10.1063/1.2210932.

[108] 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.

[109] Artem R Oganov et al. “Structure prediction drives materials discovery”. In: *Nat. Rev. Mater.* 4.5 (2019), pp. 331–348. DOI: 10.1038/s41578-019-0101-8.

[110] Bart Olsthoorn et al. “Band gap prediction for large organic crystal structures with machine learning”. In: *arXiv preprint arXiv:1810.12814* (2018). DOI: 10.1002/qute.201900023.

[111] 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.

[112] Tao Ouyang and Ming Hu. “Competing mechanism driving diverse pressure dependence of thermal conductivity of X Te (X= Hg, Cd, and Zn)”. In: *Physical Review B* 92.23 (2015), p. 235204.

[113] Felipe Oviedo et al. “Fast classification of small X-ray diffraction datasets using data augmentation and deep neural networks”. In: *arXiv preprint arXiv:1811.08425* (2018). DOI: 10.1038/s41524-019-0196-x.

[114] Woon Bae Park et al. “Classification of crystal structure using a convolutional neural network”. In: *IUCrJ* 4.4 (2017), pp. 486–494. DOI: 10.1107/S205225251700714X.

[115] Robin Pearce and Yang Zhang. “Deep learning techniques have significantly impacted protein structure prediction and protein design”. In: *Current Opinion in Structural Biology* 68 (2021), pp. 194–207.

[116] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[117] Matthew E Peters et al. “Deep contextualized word representations”. In: *arXiv preprint arXiv:1802.05365* (2018).

[118] Matthew E. Peters et al. *Semi-supervised sequence tagging with bidirectional language models*. 2017. arXiv: 1705.00108 [cs.CL].

[119] Tien Lam Pham et al. “Machine learning reveals orbital interaction in materials”. In: *Science and technology of advanced materials* 18.1 (2017), p. 756.

[120] Ghanshyam Pilania et al. “Accelerating materials property predictions using machine learning”. In: *Scientific reports* 3 (2013), p. 2810.

[121] Guangzhao Qin et al. “Anomalously temperature-dependent thermal conductivity of monolayer GaN with large deviations from the traditional 1/T law”. In: *Physical Review B* 95.19 (2017), p. 195416.

[122] Guangzhao Qin et al. “External electric field driving the ultra-low thermal conductivity of silicene”. In: *Nanoscale* 9.21 (2017), pp. 7227–7234.

[123] Guangzhao Qin et al. “Lone-pair electrons induced anomalous enhancement of thermal transport in strained planar two-dimensional materials”. In: *Nano Energy* 50 (2018), pp. 425–430.

[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] Bharath Ramsundar and Reza Bosagh Zadeh. *TensorFlow for deep learning: from linear regression to reinforcement learning*. " O’Reilly Media, Inc.", 2018.

[126] Jesse Read et al. “Classifier chains for multi-label classification”. In: *ECML PKDD*. Springer. 2009, pp. 254–269. DOI: 10.1007/978-3-642-04174-7_17.

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

[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] Qingyuan Rong et al. “Predicting the effective thermal conductivity of composites from cross sections images using deep learning methods”. In: *Composites Science and Technology* 184 (2019), p. 107861.

[130] Matthias Rupp. “Machine learning for quantum mechanics in a nutshell”. In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1058–1073.

[131] Matthias Rupp et al. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical review letters* 108.5 (2012), p. 058301.

[132] James E Saal et al. “Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)”. In: *Jom* 65.11 (2013), pp. 1501–1509.

[133] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. "Inverse molecular design using machine learning: Generative models for matter engineering". In: *Science* 361.6400 (2018), pp. 360–365.

[134] Yoshihide Sawada, Koji Morikawa, and Mikiya Fujii. "Study of deep generative models for inorganic chemical compositions". In: *arXiv preprint arXiv:1910.11499* (2019).

[135] Kristof T Schütt et al. "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties". In: *Physical Review B* 89.20 (2014), p. 205118.

[136] Kristof T Schütt et al. "SchNet–A deep learning architecture for molecules and materials". In: *The Journal of Chemical Physics* 148.24 (2018), p. 241722.

[137] Austin D Sendek et al. "Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials". In: *Energy & Environmental Science* 10.1 (2017), pp. 306–320.

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

[139] Jimmy-Xuan Shen, Matthew Horton, and Kristin A Persson. "A charge-density-based general cation insertion algorithm for generating new Li-ion cathode materials". In: *npj Computational Materials* 6.1 (2020), pp. 1–7.

[140] Bernard Silvi and Andreas Savin. "Classification of chemical bonds based on topological analysis of electron localization functions". In: *Nature* 371.6499 (1994), p. 683.

[141] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[142] Yuqi Song et al. "Computational discovery of new 2D materials using deep learning generative models". In: *arXiv preprint arXiv:2012.09314* (2020).

[143] 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.

[144] Mohammad S Sorower. "A literature survey on algorithms for multi-label learning". In: *Oregon State University, Corvallis* 18 (2010).

[145] Space group. *Space group — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-January-2019]. 2019. URL: https://en.wikipedia.org/wiki/Space_group.

[146] Valentin Stanev et al. "Machine learning modeling of superconducting critical temperature". In: *npj Comput. Mater.* 4.1 (2018), p. 29. DOI: 10.1038/s41524-018-0085-8.

[147] Hitoshi Sumiya. "Novel Development of High-Pressure Synthetic Diamonds Ultra-Hard Nano-Polycrystalline Diamonds". In: *SEI Technical Review* 74 (2012), pp. 15–23.

[148] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[149] Christian Szegedy et al. "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[150] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[151] Wojciech J Szlachta, Albert P Bartók, and Gábor Csányi. "Accuracy and transferability of Gaussian approximation potential models for tungsten". In: *Physical Review B* 90.10 (2014), p. 104108.

[152] Keisuke Takahashi and Lauren Takahashi. "Creating Machine Learning-Driven Material Recipes Based on Crystal Structure". In: *J. Phys. Chem. Lett.* 10.2 (2019), pp. 283–288. DOI: 10.1021/acs.jpclett.8b03527.

[153] B. Thapa et al. "Hot carrier dynamics in nitrogen – rich hafnium nitride thin film". In: *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*. 2020, pp. 0793–0797. DOI: 10.1109/PVSC45281.2020.9300430.

[154] A Togo and I Tanaka. "First principles phonon calculations in materials science". In: *Scr. Mater.* 108 (Nov. 2015), pp. 1–5.

[155] Vahe Tshitoyan et al. "Unsupervised word embeddings capture latent knowledge from materials science literature". In: *Nature* 571.7763 (2019), pp. 95–98.

[156] Grigorios Tsoumakas and Ioannis Katakis. "Multi-label classification: An overview". In: *IJDWM* 3.3 (2007), pp. 1–13.

[157] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. "Mining multi-label data". In: *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685. DOI: 10.1007/978-0-387-09823-4_34.

[158] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[159] 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.

[160] JW Visser. "A fully automatic program for finding the unit cell from powder data". In: *J. Appl. Crystallogr.* 2.3 (1969), pp. 89–95. DOI: 10.1107/S0021889869006649.

[161] David J Wales and Jonathan PK Doye. "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms". In: *The Journal of Physical Chemistry A* 101.28 (1997), pp. 5111–5116.

[162] Sheng Wang et al. "Accurate de novo prediction of protein contact map by ultra-deep learning model". In: *PLoS computational biology* 13.1 (2017), e1005324.

[163] Yanchao Wang et al. "Crystal structure prediction via particle-swarm optimization". In: *Phys. Rev. B* 82.9 (2010), p. 094116. DOI: 10.1103/PhysRevB.82.094116.

[164] Logan Ward and Chris Wolverton. "Atomistic calculations and materials informatics: A review". In: *Current Opinion in Solid State and Materials Science* 21.3 (2017), pp. 167–176.

[165] 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.

[166] Logan Ward et al. "A machine learning approach for engineering bulk metallic glass alloys". In: *Acta Mater.* 159 (2018), pp. 102–111. DOI: 10.1016/j.actamat.2018.08.002.

[167] Logan Ward et al. "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations". In: *Phys. Rev. B* 96.2 (2017), p. 024104. DOI: 10.1103/PhysRevB.96.024104.

[168] Logan Ward et al. "Matminer: An open source toolkit for materials data mining". In: *Computational Materials Science* 152 (2018), pp. 60–69.

[169] Han Wei et al. “Predicting the effective thermal conductivities of composite materials and porous media by machine learning methods”. In: *International Journal of Heat and Mass Transfer* 127 (2018), pp. 908–916.

[170] P-E Werner, L Eriksson, and M Westdahl. “TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries”. In: *J. Appl. Crystallogr.* 18.5 (1985), pp. 367–370. DOI: 10.1107/S0021889885010512.

[171] Michael J Willatt, Félix Musil, and Michele Ceriotti. “Atom-density representations for machine learning”. In: *The Journal of chemical physics* 150.15 (2019), p. 154110.

[172] LT Wille. “Searching potential energy surfaces by simulated annealing”. In: *Nature* 324.6092 (1986), pp. 46–48.

[173] Philipp Wollmann et al. “High-throughput screening: speeding up porous materials discovery”. In: *Chemical Communications* 47.18 (2011), pp. 5151–5153.

[174] Adam D Wright et al. “Electron-phonon coupling in hybrid lead halide perovskites”. In: *Nature communications* 7.1 (2016), pp. 1–9.

[175] Zhi-jian Wu et al. “Crystal structures and elastic properties of superhard Ir N 2 and Ir N 3 from first principles”. In: *Physical Review B* 76.5 (2007), p. 054115.

[176] 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.

[177] Tian Xie et al. “Crystal Diffusion Variational Autoencoder for Periodic Material Generation”. In: *arXiv preprint arXiv:2110.06197* (2021).

[178] 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.

[179] Jia-Yue Yang and Ming Hu. “Strong electron–phonon interaction retarding phonon transport in superconducting hydrogen sulfide at high pressures”. In: *Physical Chemistry Chemical Physics* 20.37 (2018), pp. 24222–24226.

[180] Jia-Yue Yang et al. “Strong electron-phonon coupling induced anomalous phonon transport in ultrahigh temperature ceramics ZrB2 and TiB2”. In: *International Journal of Heat and Mass Transfer* 152 (2020), p. 119481.

[181] Dingjun Yu et al. “Mixed pooling for convolutional neural networks”. In: *International conference on rough sets and knowledge technology*. Springer. 2014, pp. 364–375.

[182] Sheng-Ying Yue et al. “Electron–phonon interaction and superconductivity in the high-pressure cI 16 phase of lithium from first principles”. In: *Physical Chemistry Chemical Physics* 20.42 (2018), pp. 27125–27130.

[183] Min-Ling Zhang and Zhi-Hua Zhou. “A k-nearest neighbor based algorithm for multi-label classification.” In: *GrC* 5 (2005), pp. 718–721. DOI: 10.1109/GRC.2005.1547385.

[184] S.H. Zhang and R.F. Zhang. “AELAS: Automatic ELAStic property derivations via high-throughput first-principles computation”. In: *Computer Physics Communications* 220 (2017), pp. 403–416. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2017.07.020. URL: http://www.sciencedirect.com/science/article/pii/S0010465517302400.

[185] Xiu Zhang et al. “Ultrafast hot carrier dynamics of $ZrTe_5$ from time-resolved optical reflectivity”. In: *Phys. Rev. B* 99 (12 Mar. 2019), p. 125141. DOI: 10.1103/PhysRevB.99.125141. URL: https://link.aps.org/doi/10.1103/PhysRevB.99.125141.

[186] Ye Zhang, Stephen Roller, and Byron Wallace. “MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification”. In: *arXiv preprint arXiv:1603.00968* (2016).

[187] Ying Zhang et al. “Unsupervised discovery of solid-state lithium ion conductors”. In: *Nat. Commun.* 10.1 (2019), pp. 1–7. DOI: 10.1038/s41467-019-13214-1.

[188] Yong Zhao. *Carolina Materials Database*. http://www.carolinamatdb.org/. [Online; accessed 01-August-2021]. 2021.

[189] Yong Zhao et al. “High-throughput discovery of novel cubic crystal materials using deep generative neural networks”. In: *arXiv preprint arXiv:2102.01880* (2021).

[190] Yong Zhao et al. “Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions”. In: *ACS omega* 5.7 (2020), pp. 3596–3606.

[191] 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.

[192] Quan Zhou et al. “Learning atoms for materials discovery”. In: *Proceedings of the National Academy of Sciences* 115.28 (2018), E6411–E6417.

[193] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[194] 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.

[195] Angelo Ziletti et al. “Insightful classification of crystal structures using deep learning”. In: *Nat. Commun.* 9.1 (2018), p. 2775. DOI: 10.1038/s41467-018-05169-6.