Deep Learning Methods for the Design and Understanding of Solid Materials

by

Tian Xie

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Materials Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Department of Materials Science and Engineering
August 7, 2020

Certified by .....
Jeffrey Grossman
Professor of Materials Science and Engineering
Thesis Supervisor

Accepted by .....
Frances M. Ross
Chair, Departmental Committee on Graduate Studies

Deep Learning Methods for the Design and Understanding of Solid Materials

by

Tian Xie

Submitted to the Department of Materials Science and Engineering
on August 7, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Materials Science and Engineering

Abstract

The trend of open material data and automation in the past decade offers a unique opportunity for data-driven design of novel materials for various applications as well as fundamental scientific understanding, but it also poses a challenge for conventional machine learning approaches based on structure features. In this thesis, I develop a class of deep learning methods that solve various types of learning problems for solid materials, and demonstrate its application to both accelerate material design and understand scientific knowledge. First, I present a neural network architecture to learn the representations of an arbitrary solid material, which encodes several fundamental symmetries for solid materials as inductive biases. Then, I extend the approach to explore four different learning problems: 1) supervised learning to predict material properties from structures; 2) visualization to understand structure-property relations; 3) unsupervised learning to understand atomic scale dynamics from time series trajectories; 4) active learning to explore an unknown material space. In each learning problem, I demonstrate the performance of the approach compared with previous approaches, and apply it to solve several realistic materials design problems and extract scientific insights from data.

Thesis Supervisor: Jeffrey Grossman
Title: Professor of Materials Science and Engineering

Acknowledgments

I am grateful to many people around me and the wonderful academic environment at MIT. It is fortunate to work on an emerging field with many exciting opportunities to rethink how we do material science. The incredible journey of the last five years would not be possible without the help and support from them.

First and foremost, I would like to thank my advisor professor Jeffrey C. Grossman. Jeff is a great mentor who has guided and supported me throughout my PhD. He reminds me to think about fundamental breakthroughs rather than incremental improvements, and guides me to shape projects towards broader impact. I am grateful that he is always willing to make time for our discussions despite his busy schedule, and he offers me the freedom to explore new directions. He is also a charismatic leader for our group and creates a group culture that encourages sharing, collaboration, and fun. I am also grateful to him for supporting me to explore future career possibilities and offering advices for career development.

I am grateful to my collaborators who have expanded my knowledge to many different fields of materials science. In particular I would like to thank professor Yang Shao-Horn, professor Jeremiah A. Johnson, professor Adam P. Willard, and professor Rafael Gomez-Bombarelli from the MIT TRI polymer team, as well as current and former students and postdocs including Graham Leverich, Kaitlyn Duelle, Livia Giordano, Shuting Feng, Yivan Jang, Arthur France-Lanord, Yanming Wang, Jeffrey Lopez, Bo Qiao, Michael Stolberg, Megan Hill, Wujie Wang, Sheng Gong. This interdisciplinary team teaches me how to work with researchers from different fields and is extremely beneficial to my professional development. I also thank professor Venkat Viswanathan and Zeeshan Ahmad from Carnegie Mellon University for the collaboration on lithium metal batteries.

I am thankful to my thesis committee members professor Elsa A. Olivetti, professor Ju Li, and professor Rafael Gomez-Bombarelli. They provide many valuable advices and suggestions for my projects through committee meetings and individual discussions. I would like add some special thanks to Elsa, who first introduced me to

the field that combines materials science and machine learning in the first year of my PhD. Her passion of using natural language process to extract synthesis route from literature is an important reason why I chose to work on my current field.

It is also important for me to thank all the current and former members of the Grossman group. In particular, I am grateful to Huashan Li who taught me everything about computational material science when I joined the group. I also want to pay special tribute to Arthur France-Lanord and Yanming Wang, who I work closely in the polymer electrolyte project. I also thank Cuiying Jian and Zhengmao Lu for being wonderful officemates, as well as David S Bergsman, Anthony Straub, Brendan Smith, Xining Zang, Thomas Sannicolo, Beza Getachew, Taishan Zhu, Yun Liu, Eric Richard Fadel, Owen Morris, Adam Trebach, Xiang Zhang, Asmita Jana, Cédric Viry, David Chae, Emily Crabb, Ki-Jana Carter, Sheng Gong, Grace Han, and Nicola Ferralis for the time we spent togehter. Additional thanks should be given to Laura M. von Bosau for her passionate administrative support. I have the opportunity to advise several visiting students, William Xu, Pierre-Paul De Breuck, and Doosun Hong, and I thank them for being such awesome students. My graduate life will be much less rewarding and fun without the daily interactions with all the group members. The coffees and football games will always be a special part of my memory at MIT.

Another group of people I would like to thank are the friends who I am fortunate to meet in the last five years. They are Hongzhou Ye, Jiaming Luo, Xinhao Li, Ge Liu, Manxi Wu, Ruizhi Liao, Jiayue Wang, Zhiwei Ding, Yu Xia, Hongzi Mao, Hejin Huang, Yifei Zhang, Danhao Ma, Gufan Yin, Yiqi Ni, Chao Wu, Guo Zong, and many others. The dinners, movies, hiking and many other activities constitute an important part of my life outside campus and will forever be part of memory at MIT.

Finally, I would like to thank my parents for their support to my graduate study and career. I feel indebted as my time spent with them is significantly reduced after moving the US. They are always supportive of my decisions, and keep reminding me to relax more and eat healthy. I am grateful for their love and support in the past five years.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 Motivations for materials science

1.1.1 Materials discovery paradigms

The development of new materials plays a central role in the advancement of our civilization. From the stone age to modern society, fundamental innovations in materials have lead to exponential increases of productivity and prosperity. However, the creation and development of a novel material with desired properties is a notoriously difficult and slow process, mostly due to the lack of understanding of the structure-property relations. Consequently, a significant part of material innovations is still driven by trial-and-error, and many milestone materials are discovered by accident rather than careful design, like copper oxide superconductors in 1986 [1] and perovskite solar cells in 2013 [2].

The paradigms of materials discovery have been gradually shifting from empirical driven towards simulation/data driven over history. [1] Early innovations in materials are usually results from numerous trial-and-errors and empirical knowledge gathered by material scientists. One famous example of this first paradigm is the discovery of a carbonized cotton as the filament of the lightbulb by Thomas Edison in 1880, after

[1] There are multiple different ways to divide the paradigms in materials discovery. Here, we choose the narrative by Agrawal et al. in 2016 [3] which divides the entire history into four paradigms. Nevertheless, the overall trend stays the same despite the differences in ways of division.

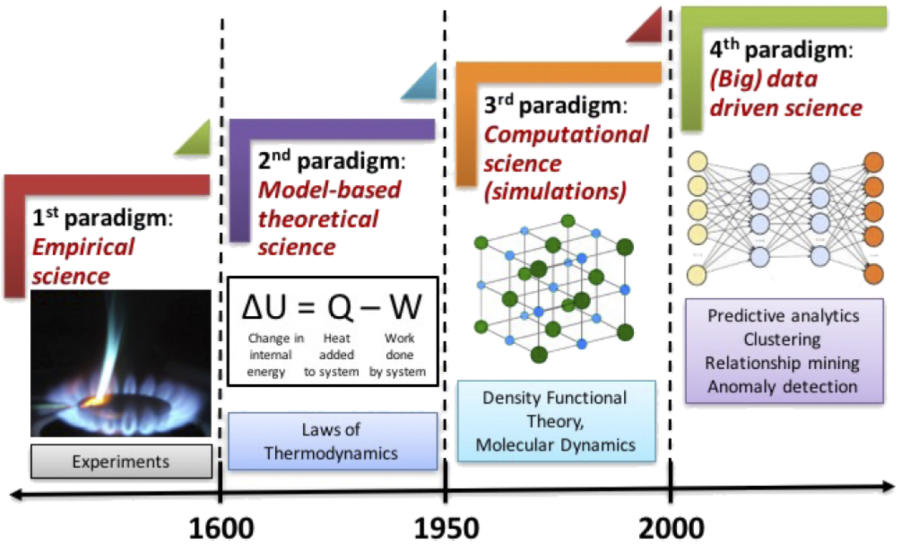


Figure 1-1: The four paradigms of science: empirical, theoretical, computational, and data-driven. [3]

thousands of failed experiments [4]. Since the early 20th century, the development of quantum mechanics and solid state physics shifted the paradigm to use physical laws and semi-empirical models to guide the design of new materials, as material scientists begin to understand that the atomic structure fundamentally determines material property. Up to today, many material science research are still motivated by exploring neighboring elements in the periodic table, which arranges elements according to periodic trends. In the late 20th century, increasingly powerful computers began to allow the direct computing of material properties by solving the Schrödinger equation, leading to the third paradigm of materials discovery. These simulation methods, called *ab-initio* simulations, are distinct from earlier physical models as they simulate material properties purely based on first principles like quantum mechanical theory, instead of relying on physical parameters that are fitted using experimental data. The success of *ab-initio* simulations, especially density functional theory [5], motivated the creation of the Materials Genome Initiative in 2011 [6] to explore the vast space of materials computationally and provide open materials data for the design of novel materials.

Many believe that machine learning and data driven approaches lead to the fourth

paradigm for materials discovery. [3, 7, 8] Since the introduction of Materials Genome Initiative, an increasing number of open databases have emerged that shares both simulation and experiment data of various classes of materials. For example, a recent review paper summarized 10 computational databases and 11 experimental databases that are publicly accessible [7]. There has also been a trend of increased automation in both material synthesis and characterization. [9, 10] The open databases and automation provide an opportunity to develop new data-driven methods that guide the design of novel materials and accelerate the discovery process, and they can also potentially lead to new theories about structure-property relations as patterns emerge from large amounts of data. However, the large size of open material data also indicates that traditional data analysis approaches are incapable of handling these large databases, requiring the development of new machine learning tools for materials. In Chapter 1.1.2, we will overview several classes machine learning tools for solid materials.

1.1.2 Application of machine learning in materials

In this chapter, we overview several types of goals that we aim to achieve by applying machine learning methods to materials discovery. We will focus on their impact to materials design and understanding rather than the methodology, but we list several review papers where various methods are discussed [7, 11, 12]. In chapter 1.4, we will discuss our deep learning approach to provide a unified framework that achieves these different goals.

Property prediction. The goal of property prediction models is to predict material properties based on their structure or chemical composition. Such models can usually run orders of magnitude faster than *ab-initio* simulations and experiment measurements when they are applied to new materials, thus significantly accelerate material discovery. They been successfully applied to accelerate the screening of organic light-emitting diodes [13], lithium ion batteries [14], etc. A special class of property prediction models are force field models [15], which aim to predict the forces on each atom given the structure of the material. These machine learning force fields

can run close to the speed of classical force fields but have the accuracy close to *ab-initio* force fields.

Interpretation and visualization. Material scientists are generally not only interested in discovering new materials, but also understanding how material structures affect their performances. The goal of interpreting machine learning models is to understand the key contributing factors to the property of interest, which can potentially lead to new theory for materials design. Visualization help to understand complex material spaces as they usually include tens of thousands of structures. For example, it can help navigate the complex space of ice structures from zeolite networks. [16].

Active learning. Active learning aims to explore a complex material space, like intermetallic alloys [17] and the configuration of atomic structures [18], in an iterative fashion that minimizes the number of simulations or experiments. In contrast to property prediction, active learning actively samples the material space and selects materials to evaluate in each iterative step, aiming to thoroughly explore the space with minimum amount of evaluations.

Inverse design. Inverse design aims to directly predict the material structure that has the optimum performance from existing data. Unlike property prediction models that are used to evaluate an existing material space, inverse design aims to predict material structures that are not in the dataset. [19]

1.2 Motivations for deep learning

1.2.1 Short introduction to deep learning

Deep learning is a class of machine learning algorithms that uses multiple layers of neural networks to progressively extract higher level features from the raw input directly from data [20]. In the past ten years, deep learning methods have dramatically improved the state-of-the-art in computer vision, speech recognition, and natural language processing tasks [21], which leads to Hinton, LeCun and Bengio winning

the Turing Award in 2018. Compared with other machine learning methods, deep learning do not rely on human designed features but aims to directly learn the task in an end-to-end fashion. As a result, it usually over-performs other methods when there are large amounts of data, which is a key factor of its success due to the increasing data sizes in multiple fields [22].

The simplest form of neural networks, feedforward networks, include multiple layers of linear transformations plus non-linear activation functions. In each layer, the input vector $\boldsymbol{x} \in \mathbb{R}^m$ is transformed by,

$$f(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}) = \boldsymbol{w}^\intercal \boldsymbol{x} + \boldsymbol{b}, \tag{1.1}$$

where both $\boldsymbol{w} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$ are learned weights. Often non-linearity is added via an activation function like ReLU [23] after f in each layer. Then, multiple layers of such transformations are applied in a chain, forming multi-layer feedforward networks $f^{(k)}(f^{(k-1)}(...f^{(1)}(\boldsymbol{x})...))$.

A more general neural network differs from this simple form in several aspects: 1) there might be more than one input vectors, 2) each layer might be more complicated than linear transformations, and 3) layers might be composed in different ways. However, the general idea is to build complex neural network architecture by combining simple components based on the type of data.

1.2.2 Inductive biases in neural networks

Inductive biases are the set of assumptions in a learning algorithm that are independent of the observed data. [24] In neural networks, one of the most important inductive biases is the symmetry, or the sharing of weights, within the network architecture. In the case of feedforward networks, there is no symmetry because the weights are not shared in Eq. 1.1. This means that each individual parameter in the matrix $\boldsymbol{w} \in \mathbb{R}^{m \times n}$ has to been learned independently. For a 640×480 pixel sized image, the input vector size m is $640 \times 480 = 307,200$, so the number of independent parameters in the first layer will be $307,200 \times 10 = 30,720,000$ if we choose an output

size n of 100, which is almost impossible to learn from data. The lack of sharing of weights in a feedforward neural network makes it unsuitable for complex tasks.

A key breakthrough in modern deep learning is to encode the symmetry of data into the neural network architecture in the form of weight sharing as inductive biases. [25] For example, image data has a translation symmetry for most tasks. To classify an image of a flower, the result should not change if the location of the flower shifts from left to right within the image. This symmetry is not encoded in a feedforward network, which means that a model learned from an image with a flower on the left does not generalize to an image with a flower on the right. The problem is solved by the introduction of convolutional neural network (CNN) [26], which shares the weights across the grid-like structure in images through a convolution operation, as illustrated in Fig. 1-2(a). The weight sharing significantly reduces the number of independent parameters in the networks by incorporating known symmetry in the image data. It has been tremendously successful in many practical applications, and nearly all best performing models in the ImageNet competition are variants of the CNN architecture (Fig. 1-2(b)).

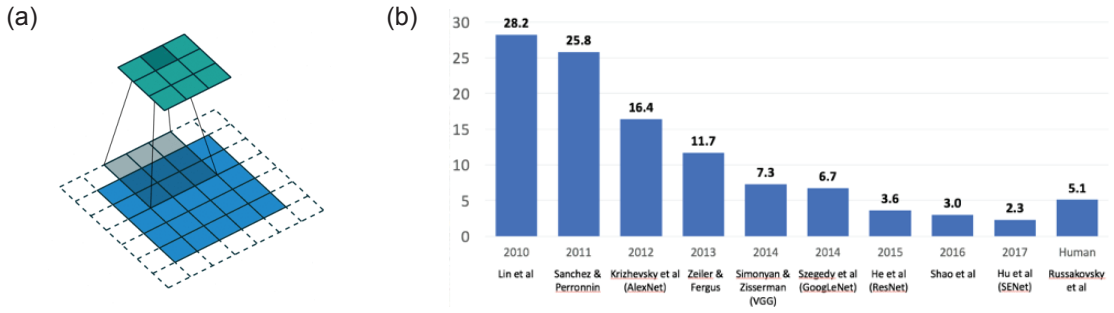


Figure 1-2: Weight sharing in convolution neural network (CNN) and its impact. (a) An illustrative diagram that demonstrates how weights are shared in CNN. [27] (b) The error rate of the best performing models in the ImageNet competition [28].

Different network architectures are developed to encode the symmetries of various data types into the neural networks. We list several examples below to demonstrate how this concept works in a broader context, and readers can refer to Ref. [25] which discussed the inductive biases in neural networks in depth.

Recurrent neural networks. Recurrent neural networks (RNNs) aim to capture

the time invariance in sequential data. [29] RNNs are typically used for language data where sentences and paragraphs are encoded as a sequence of words. Weights are shared across the sequence by applying the same weights to each work sequentially, which enables the generalization to different time in a sequence.

Graph neural networks. Graph neural networks (GNNs) aim to capture the node and edge invariances in graph structured data. [30] It has been used for learning social networks and molecular graphs, as well as many other problems. [31] Weights are shared across both nodes and edges in a graph, which enables the generalization across different graph sizes.

Point networks. Point networks (PointNets) aim to capture the permutation invariance of points in 3D point clouds. [32] Point clouds are a set of data points in 3D space, which are usually generated by 3D scanners. Weights are shared across the points which enable the generalization across 3D space.

1.3 Unified data representation for solid materials

1.3.1 Data representation format

In this section, we seek a general form to represent an arbitrary solid material. We hope this representation will cover various different types of solid materials in a unified way, but still capture the invariances shared by these materials.

We represent the structure of a solid materials as a collection of atoms under a 3D periodic boundary condition. Mathematically, a solid material made up with N atoms can be represented by a list of three vectors $\boldsymbol{m} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\}$, in which

- $\boldsymbol{x} \in \mathbb{R}^{N\times 3}$, denoting the coordinates of N atoms in 3D space;
- $\boldsymbol{z} \in \mathbb{N}^{N}$, denoting the atomic numbers of each atom in the structure;
- $\boldsymbol{l} \in \mathbb{R}^{3\times 3}$, denoting the periodic boundary condition of the system, which is represented by 3 lattice vectors.

This representation, which we call *periodic solid representation*, is widely used to represent solid materials in atomic scale simulations, supported by many existing standard file formats like the Crystallographic Information File (CIF) [33], Protein Data Bank (pdb) file [34], etc. Below, we list several examples of how different types of materials can be represented in this format.

Crystals. Crystal materials (Fig. 1-3(a)) can be naturally represented with periodic solid representation due to their periodicity. The structure of crystals can measured experimentally using X-ray diffraction (XRD) [35]. Large databases like the Inorganic Crystal Structure Database (ICSD) [36] collects hundreds of thousands of crystal structures.

Molecules. Molecular materials (Fig. 1-3(b)) are often represented by their molecular graphs which describes the connectivity between atoms. However, molecular graphs do not capture the 3D structure of molecules, which can be important for many molecular properties. Using the periodic solid representation, molecules can be represented by an isolated molecule with large vacuum space around it under a periodic boundary condition. Many molecules have structural polymorphism while forming crystals, which has important pharmaceutical significance. [37] These different structures can be easily capture with periodic solid representation.

Low dimensional materials. Low dimensional materials (Fig. 1-3(c)) are a large class of materials in which periodicity only exists in less than three dimensions. Typical 0D, 1D, and 2D materials includes quantum dots [38], carbon nanotube [39], and graphene [40]. They can be represented with periodic solid representation by introducing large vacuum spaces in dimensions that are not periodic.

Amorphous materials. Amorphous materials (Fig. 1-3(d)) are atomic systems that lack long range order in any direction. [41] Typical amorphous materials include glasses, polymers, gels, etc. Liquids are sometimes considered as amorphous materials as well. Despite the lack of periodicity, representative structures exist for most amorphous materials. Therefore, they can usually be represented by atomic structures in a very large periodic box where the boundary effects can be ignored.

Complex materials. Complex materials (Fig. 1-3(e)) are composed of more

than one types of above simpler materials. For example, metal-organic frameworks are a class of materials consisting of metal ions coordinated to organic compounds. [42] Solid polymer electrolytes are mixture of polymers and inorganic salts. [43] Depending on their periodicity, complex materials can also by represented using similar techniques as low dimensional materials and amorphous materials.

1.3.2 Quantum mechanical implications

The periodic solid representation of a material uniquely defines a quantum mechanical system. Given the list $\boldsymbol{m} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\}$, it defines an atomic system with N atoms and n electrons under a periodic boundary condition whose full Hamiltonian can be written as,

$$\hat{H} = -\sum_{I=1}^{N} \frac{\hbar^2}{2M_I} \nabla^2_{\boldsymbol{R}_I} + \frac{1}{2} \sum_{I \neq J} \frac{Z_I Z_J e^2}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} - \sum_{i=1}^{N} \frac{\hbar^2}{2m} \nabla^2_{\boldsymbol{r}_i} - \sum_{I=1}^{N} \sum_{i=1}^{n} \frac{Z_I e^2}{|\boldsymbol{R}_I - \boldsymbol{r}_j|} + \frac{1}{2} \sum_{i \neq j} \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \tag{1.2}$$

where M_I, Z_I, $\boldsymbol{R}_I$ are the mass, atomic number, and position of Ith nucleus, m, $\boldsymbol{r}_i$ are the mass, and position of ith electron, N and n are the total number of nuclei and electrons, $\hbar$ and e are Planck's constant and electron charge.

The many body wavefunction $\Psi(\boldsymbol{r}, \boldsymbol{R})$ of this system can be solved with the following eigenvalue problem,

$$\hat{H}\Psi(\boldsymbol{r}, \boldsymbol{R}) = E\Psi(\boldsymbol{r}, \boldsymbol{R}), \tag{1.3}$$

which includes both the nuclei and electrons. But in most cases, we can use the Born–Oppenheimer approximation to separate the motion of nuclei and electrons. Assuming that $\Psi(\boldsymbol{r}, \boldsymbol{R}) = \Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R})\Psi^{(\text{nuc})}(\boldsymbol{R})$, we could separate the full Hamiltonian into a nuclei and an electron part,

$$[\hat{H}_{\text{nuc}}(\boldsymbol{R}) + \hat{H}_{\text{elec}}(\boldsymbol{R}, \boldsymbol{r}))]\Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R})\Psi^{(\text{nuc})}(\boldsymbol{R}) = E\Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R})\Psi^{(\text{nuc})}(\boldsymbol{R}). \tag{1.4}$$

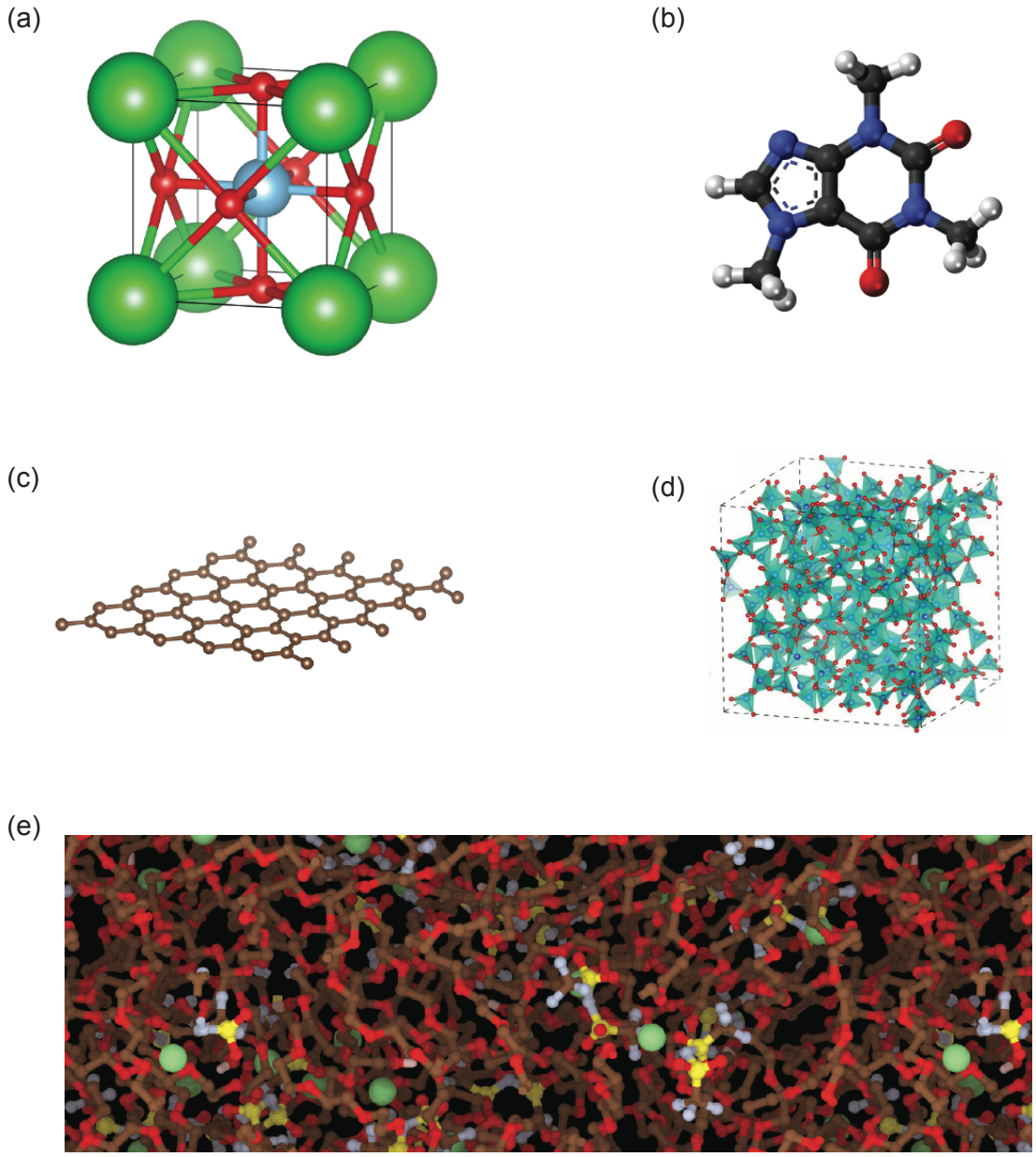


Figure 1-3: Representative types of different solid materials. (a) Crystals. Structure of $BaTiO_3$. (b) Molecules. Structure of caffeine. (c) Low dimensional materials. Structure of graphene. (d) Amorphous materials. Structure of silica glass. [44] (e) Complex materials. Structure of a mixture of polyethylene oxide (PEO) polymer and lithium bis-(trifluoromethanesulfonyl)-imide (LiTFSI) salts. [45]

Then, one could obtain an electronic only eigenvalue problem,

$$\hat{H}_{\text{elec}}(\boldsymbol{R}, \boldsymbol{r})\Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R}) = E(\boldsymbol{R})\Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R}). \tag{1.5}$$

In principle, after obtaining the electronic wavefunction $\Psi^{(\text{elec})}(\boldsymbol{r}, \boldsymbol{R})$, any material property can be computed with its corresponding quantum operators. Note that in this derivation, the periodicity of the system is ignored for simplicity. The derivation under periodic boundary condition is very similar.

Consequently, we know that the periodic solid representation uniquely defines a mapping,

$$f : \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\} \mapsto p, \tag{1.6}$$

where p represents an arbitrary material property. This mapping is valid under the Born–Oppenheimer approximation when the motion of the nuclei and electrons can be separated.

We need to pay attention to a special property, energy $E(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l})$, because the system tends to minimize its total energy by changing $\boldsymbol{x}$ and $\boldsymbol{l}$ and reach a ground state without an external force. It means that the majority of materials in the space of $\{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\} \in \{\mathbb{R}^{N\times 3}, \mathbb{N}^N, \mathbb{R}^{3\times 3}\}$ are not in the ground state and will not form a stable material. In most cases, we are interested in stable materials which form a manifold in the space of $\{\mathbb{R}^{N\times 3}, \mathbb{N}^N, \mathbb{R}^{3\times 3}\}$.

1.3.3 Thermodynamical implications

In section 1.3.2, our discussion did not include the temperature effect and is only valid at 0 K. At a finite temperature, the structure of a material will deviate from its ground state structure, forming an ensemble of structures whose distribution is determined by the temperature of the system. It means that macroscopic properties, material properties measured in real life, are determined by this ensemble of structures instead of a single structure. In statistical physics, the structures follow specific distributions according the thermodynamic ensembles. For example, in an isolated system at a fixed

temperature, structures follow a canonical ensemble with a probability distribution,

$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}) \propto e^{-E(\boldsymbol{x},\boldsymbol{z},\boldsymbol{l})/(k_B T)}, \tag{1.7}$$

where k_B is the Boltzmann's constant.

The temperature effect has different implications for different types of materials. For most solid materials, the atoms will only vibrate around its ground-state position at room temperature. It indicates that thermodynamical effect can be approximated with small perturbations to the ground state structure, and Eq. 1.6 still holds if we add the vibration terms into the properties. For liquids and amorphous materials, the energy landscape is more flat, making their macroscopic properties more difficult to define given a single ground state structure.

1.4 Problem statement and thesis overview

In this thesis, we aim to develop a set of unified deep learning methods that solve various learning problems for solid materials. In each chapter of the thesis, we focus on one different learning problem. Each chapter will start by introducing the motivations of the learning problem, then discuss the details of methodology, and end with applications to solve realistic material science problems. The thesis is roughly guided by the following four problems.

Problem 1: *how to create a neural network architecture that encodes material-specific inductive biases and whether such architecture outperforms existing methods?*

This problem is the foundation of this thesis. In section 1.3, we defined a unified data representation for the structure of solid materials. This data representation differs from existing data categories like graphs and point clouds, since it has both a discrete component and continuous component. This distinction makes it difficult to use existing neural network architectures for solid materials. In Chapter 2, we will study this problem by introducing a crystal graph convolutional neural networks (CGCNN) architecture [46] that encodes fundamental symmetries of solid materials,

and we will demonstrate the advantages of this architecture over traditional methods and its applications in the design of battery materials [47].

Problem 2: *how to extract intuitions that can be understood by human researchers from the learned representations?*

In materials science, it is often desirable to understand the relationship between material structures and properties, since it provides insights that may guide the design of new materials. But improving the interpretability of neural networks has been a challenge for deep learning. In chapter 3, we explore a variety of methods to extract insights from the learned representations, and demonstrate its application to several material systems. [46, 48]

Problem 3: *how to learn the representation of atoms in solid materials when no explicit property labels are available?*

In problem 1, we focused on the supervised learning setting that learns the structure-property relations in Eq. 1.6. But it is generally expensive to obtain property labels either from simulation or experiment. In chapter 4, we explore a different scenario to learn the representation of atoms from time-series data without the property labels. The key motivation of this problem is learn a low dimensional representation for atoms and small molecules to understand their complex dynamics behavior from molecular dynamics simulation data. We will also demonstrate the application of this method to study complex material systems like amorphous polymer electrolytes and liquid-solid interfaces. [45]

Problem 4: *how to search an unknown material space in a way that balances both exploration and exploitation?*

The previous problems focus on learning from existing material data. But no data is available when we try to explore an unknown material space. What is the most efficient way to search this space? The major challenge in the problem is the balance of exploration and exploitation. Exploration means searching unknown regions of the space, and exploitation means using existing data to find optimum materials. Designing a strategy to balance these two factors can lead to efficient search of a material space that does not miss potentially important materials. In chapter 5, we

use Bayesian optimization to tackle this problem and apply the approach to the search of polymer electrolytes. [49]

Chapter 2

Crystal graph convolutional neural networks for the representation learning of solid materials

2.1 Introduction

2.1.1 Chapter overview

In this chapter, we present a generalized crystal graph convolutional neural networks (CGCNN) framework for the representation learning of periodic solid materials that respects the fundamental invariances. We will first discuss the invariances of periodic solid materials that should be encoded in a neural network architecture. Then, we present how CGCNN encode these invariances into its architecture. Further, we demonstrate the performance of CGCNN in both regression and classification tasks to predict material properties. Finally, we explore the application of CGCNN to a real-world problem for materials design – designing solid electrolytes for lithium metal batteries.

2.1.2 Theoretical and practical motivations

From a theoretical perspective, we aim to solve the supervised learning problem to predict material properties from their structures,

$$y = f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}), \tag{2.1}$$

where f is a neural network, $\{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\}$ is a solid material, and y is a material property. The goal is to encode the known invariances of solid materials directly into the architecture of f as inductive biases, aiming to achieve better performance than other ML approaches and keep the methods general to various types of solid materials.

From a practical perspective, achieving good performance in the supervised learning problem enables the accelerated screening of new materials in many domains. It is extremely expensive to both measure material properties experimentally or compute them using simulation approaches. A machine learning model that predicts properties with good accuracy can usually run 10^4-10^6 orders of magnitude faster than quantum mechanical simulations.

2.1.3 Related prior research

The arbitrary size of crystal systems poses a challenge in representing solid materials, as they need to be represented as a fixed length vector in order to be compatible with most ML algorithms. This problem is previously resolved by manually constructing fixed-length feature vectors using simple material properties[50–54] or designing symmetry-invariant transformations of atom coordinates[55–57]. However, the former requires case-by-case design for predicting different properties and the latter makes it hard to interpret the models as a result of the complex transformations.

2.2 Invariances in periodic solid materials

In section 1.3, we proposed an unified representation for solid materials as a three member list $\boldsymbol{m} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\}$. For a specific solid material, the quantum mechanical

description of the system (Eq. 1.2) guarantees the following invariances.

- Permutation invariance. Exchanging two identical atoms (e.g. two hydrogen atoms) will not change the representation of the system. It means that exchanging $\boldsymbol{x}_i \in \mathbb{R}^3$ and $\boldsymbol{x}_j \in \mathbb{R}^3$ should not change the learned representation if $z_i = z_j$.

- Periodicity. Choosing different unit cells for the periodic system will not change the representations of the system. It means that $\boldsymbol{m}_1 = \{\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{l}_1\}$ and $\boldsymbol{m}_2 = \{\boldsymbol{x}_2, \boldsymbol{z}_2, \boldsymbol{l}_2\}$ should have the same learned representation if they are different unit cells of the same periodic structure.

2.3 Architecture of CGCNN

2.3.1 Graph representation of solid materials

The main idea in our approach is to represent the crystal structure by a crystal graph that encodes both atomic information and bonding interactions between atoms, and then build a convolutional neural network on top of the graph to automatically extract representations that are optimum for predicting target properties by training with data. As illustrated in Figure 2-1 (a), a crystal graph $\mathcal{G}$ is an undirected multigraph which is defined by nodes representing atoms and edges representing connections between atoms in a crystal. The crystal graph is unlike normal graphs since it allows multiple edges between the same pair of end nodes, a characteristic for crystal graphs due to their periodicity, in contrast to molecular graphs. Each node i is represented by a feature vector $\boldsymbol{v}_i$, encoding the property of the atom corresponding to node i. Similarly, each edge $(i,j)_k$ is represented by a feature vector $\boldsymbol{u}_{(i,j)_k}$ corresponding to the k-th bond connecting atom i and atom j.

For the same crystal structure $\boldsymbol{m} = \{\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{l}\}$, there can be multiple different graph representations depending on 1) how connectivity between nodes are determined; 2) how node feature vectors $\boldsymbol{v}_i^{(0)}$ are initialized; and 3) how edge feature vectors $\boldsymbol{u}_{(i,j)_k^{(0)}}$ are initialized.

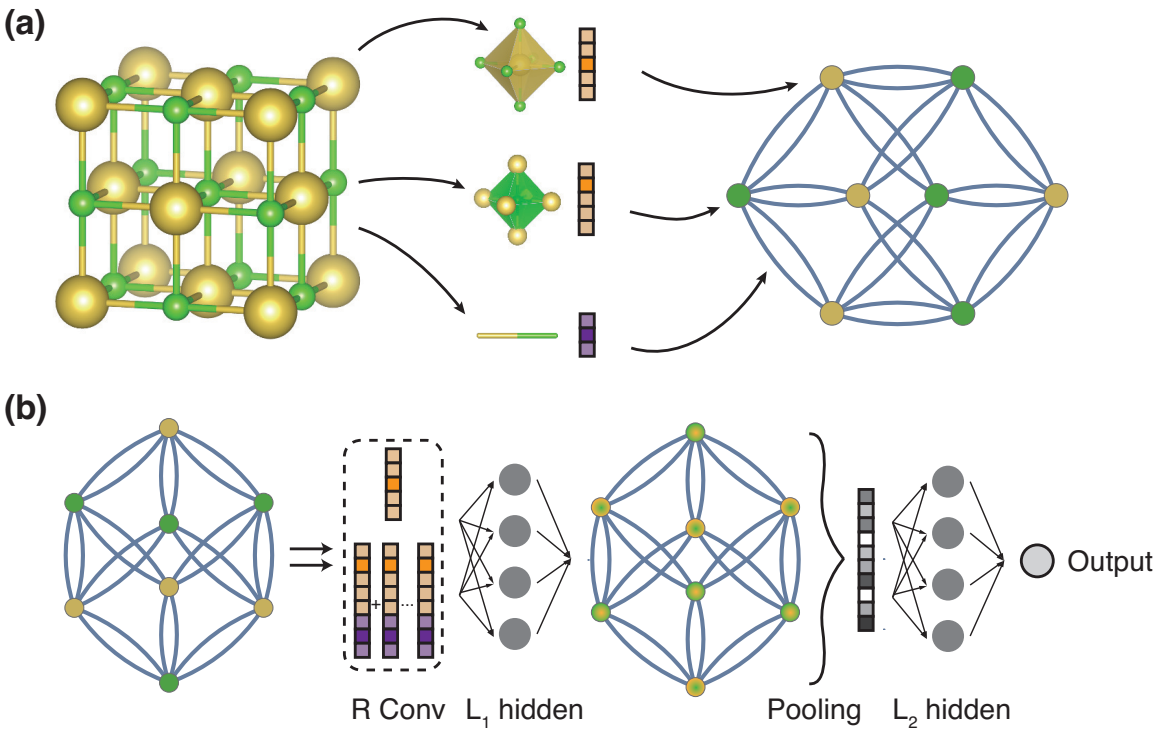


Figure 2-1: Illustration of the crystal graph convolutional neural network (CGCNN). (a) Construction of the crystal graph. Crystals are converted to graphs with nodes representing atoms in the unit cell and edges representing atom connections. Nodes and edges are characterized by vectors corresponding to the atoms and bonds in the crystal, respectively. (b) Structure of the convolutional neural network on top of the crystal graph. R convolutional layers and L_1 hidden layers are built on top of each node, resulting in a new graph with each node representing the local environment of each atom. After pooling, a vector representing the entire crystal is connected to L_2 hidden layers, followed by the output layer to provide the prediction.

In this work, we propose the following methods to construct the graph representations, but other methods are also possible under the framework and several such methods [58] have been proposed since the publication of our work.

Connectivity between nodes. We explore two different methods to determine the connectivity. The first is inspired by Ref. [50] which utilizes the Voronoi tessellation. Only strong bonding interactions are considered in this crystal graph construction. The second method is the nearest neighbor algorithm, where we connect 12 nearest neighbors in the initial graph construction. As we will discuss later, the introduction of a convolution function with gated structure (Eq. 2.5) allows the neural networks to automatically learn the importance of each edge. Practically, we find that both methods perform similarly using Eq. 2.5 as the convolution function, significantly outperforming Eq. 2.4.

Node features. The nodes are initialized by a unique mapping from the atomic number to a feature vector $f_{\text{node}} : z_i \mapsto \boldsymbol{v}_i^{(0)}$. We designed a feature vector that reflects the similarity between different elements using their element properties summarized in Table 2.1, which allows the neural networks to generalize to elements that appear less frequently in the dataset.

Table 2.1: Properties used in atom feature vector $\boldsymbol{v}_i^{(0)}$

Property	Unit	Range	# of categories
Group number	–	1,2, ..., 18	18
Period number	–	1,2, ..., 9[1]	9
Electronegativity[59, 60]	–	0.5–4.0	10
Covalent radius[61]	pm	25–250	10
Valence electrons	–	1, 2, ..., 12	12
First ionization energy[62](Log)	eV	1.3–3.3	10
Electron affinity[63]	eV	-3–3.7	10
Block	–	s, p, d, f	4
Atomic volume (Log)	cm^3/mol	1.5–4.3	10

Edge features. The edges are initialized by a unique mapping from the distance between the connected atoms $f_{\text{edge}} : d_{i,j} \mapsto \boldsymbol{u}_{(i,j)_k^{(0)}}$. In this work, we use the mapping function,

$$\boldsymbol{u}_{(i,j)}^{(0)}[t] = \exp(-(d_{(i,j)} - \mu_t)^2/\sigma^2), \tag{2.2}$$

where $\mu_t = t \cdot 0.2\,\text{Å}$ for $t = 0, 1, ..., K$, $\sigma = 0.2\,\text{Å}$.

2.3.2 Graph neural network architecture

The convolutional neural networks built on top of the crystal graph consists of two major components: convolutional layers and pooling layers. Similar architectures have been used for computer vision[64], natural language processing[65], molecular fingerprinting[66], and general graph-structured data[67, 68] but not for crystal property prediction to the best of our knowledge. The convolutional layers iteratively update the atom feature vector $\boldsymbol{v}_i$ by "convolution" with surrounding atoms and bonds with a non-linear graph convolution function.

$$\boldsymbol{v}_i^{(t+1)} = \text{Conv}\left(\boldsymbol{v}_i^{(t)}, \boldsymbol{v}_j^{(t)}, \boldsymbol{u}_{(i,j)_k}\right), (i,j)_k \in \mathcal{G} \tag{2.3}$$

The convolution function Eq. 2.3 needs to respect the permutation invariance of atoms, which is achieved by sharing weights between both nodes $\boldsymbol{v}_i^{(t)}$ and edges $\boldsymbol{u}_{(i,j)_k}$. In this work, we proposed two types of convolution functions. We start with a simple convolution function,

$$\boldsymbol{v}_i^{(t+1)} = g\left[\left(\sum_{j,k} \boldsymbol{v}_j^{(t)} \oplus \boldsymbol{u}_{(i,j)_k}\right)\boldsymbol{W}_c^{(t)} + \boldsymbol{v}_i^{(t)}\boldsymbol{W}_s^{(t)} + \boldsymbol{b}^{(t)}\right] \tag{2.4}$$

where $\oplus$ denotes concatenation of atom and bond feature vectors, $\boldsymbol{W}_c^{(t)}$, $\boldsymbol{W}_s^{(t)}$, $\boldsymbol{b}^{(t)}$ are the convolution weight matrix, self weight matrix, and bias of the t-th layer, respectively, and g is the activation function for introducing non-linear coupling between layers. By optimizing hyperparameters, the lowest mean absolute error (MAE) for the validation set is 0.108 eV/atom. One limitation of Eq. 2.4 is that it uses a shared weight matrix $\boldsymbol{W}_c^{(t)}$ for all neighbors of i, which neglects the differences of interaction strength between neighbors. To overcome this problem, we design a new convolution function that first concatenates neighbor vectors $\boldsymbol{z}_{(i,j)_k}^{(t)} = \boldsymbol{v}_i^{(t)} \oplus \boldsymbol{v}_j^{(t)} \oplus \boldsymbol{u}_{(i,j)_k}$, then

perform convolution by,

$$\boldsymbol{v}_i^{(t+1)} = \boldsymbol{v}_i^{(t)} + \sum_{j,k} \sigma(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_f^{(t)} + \boldsymbol{b}_f^{(t)}) \odot g(\boldsymbol{z}_{(i,j)_k}^{(t)} \boldsymbol{W}_s^{(t)} + \boldsymbol{b}_s^{(t)}) \tag{2.5}$$

where $\odot$ denotes element-wise multiplication and σ denotes a sigmoid function. In Eq. 2.5, the $\sigma(\cdot)$ functions as a learned weight matrix to differentiate interactions between neighbors, and adding $\boldsymbol{v}_i^{(t)}$ makes learning deeper networks easier[69]. We achieve MAE on the validation set of 0.039 eV/atom using the modified convolution function, a significant improvement compared to Eq. 2.4.

After R convolutions, the network automatically learns the feature vector $\boldsymbol{v}_i^{(R)}$ for each atom by iteratively including its surrounding environment. The pooling layer is then used for producing an overall feature vector $\boldsymbol{v}_c$ for the crystal, which can be represented by a pooling function,

$$\boldsymbol{v_c} = \text{Pool}\left(\boldsymbol{v}_0^{(0)}, \boldsymbol{v}_1^{(0)}, ..., \boldsymbol{v}_N^{(0)}, ..., \boldsymbol{v}_N^{(R)}\right) \tag{2.6}$$

that satisfies both permutational invariance with respect to atom indexing and periodicity with respect to unit cell choice. In this work, a normalized summation is used as the pooling function for simplicity but other functions can also be used. In addition to the convolutional and pooling layers, two fully-connected hidden layers with the depth of L_1 and L_2 are added to capture the complex mapping between crystal structure and property. Finally, an output layer is used to connect the L_2 hidden layer to predict the target property $\hat{y}$.

The training is performed by minimizing the difference between the predicted property $\hat{y}$ and the DFT calculated property y, defined by a cost function $J(y, \hat{y})$. The whole CGCNN can be considered as a function f parameterized by weights $\boldsymbol{W}$ that maps a crystal $\mathcal{C}$ to the target property $\hat{y}$. Using backpropagation and stochastic gradient descent (SGD), we can solve the following optimization problem by iteratively updating the weights with DFT calculated data,

$$\min_{\boldsymbol{W}} J(y, f(\mathcal{C}; \boldsymbol{W})) \tag{2.7}$$

the learned weights can then be used to predict material properties and provide chemical insights for future materials design.

2.4 Predictive performance

To demonstrate the performance of the CGCNN, we train the model using calculated properties from the Materials Project[70]. One key advantage of CGCNN over previous methods is its generality, since our representation cover various types of solid materials (section 1.3). We focus on two types of generality in this work: (1) The structure types and chemical compositions for which our model can be applied, and (2) the number of properties that our model can accurately predict.

The database we used includes a diverse set of inorganic crystals ranging from simple metals to complex minerals. After removing ill-converged crystals, the full database has 46744 materials covering 87 elements, 7 lattice systems and 216 space groups. As shown in Figure 2-2(a), the materials consist of as many as seven different elements, with 90% of them binary, ternary and quaternary compounds. The number of atoms in the primitive cell ranges from 1 to 200, and 90% of crystals have less than 60 atoms(Figure S2). Considering most of the crystals originate from the Inorganic Crystal Structure Database (ICSD)[72], this database is a good representation of known stoichiometric inorganic crystals.

Figure 2-2(b)(c) shows the performance of the two models on 9350 test crystals for predicting the formation energy per atom. We find a systematic decrease of the mean absolute error (MAE) of the predicted values compared with DFT calculated values for both convolution functions as the number of training data is increased. The best MAE's we achieved with Eq. 2.4 and Eq. 2.5 are 0.136 eV/atom and 0.039 eV/atom, and 90% of the crystals are predicted within 0.3 eV/atom and 0.08 eV/atom errors, respectively. In comparison, Kirklin *et al.* reports that the MAE of DFT calculation with respect to experimental measurements in the Open Quantum Materials Database (OQMD) is 0.081–0.136 eV/atom depending on whether the energies of the elemental reference states are fitted, although they also find a large MAE of 0.082 eV/atom

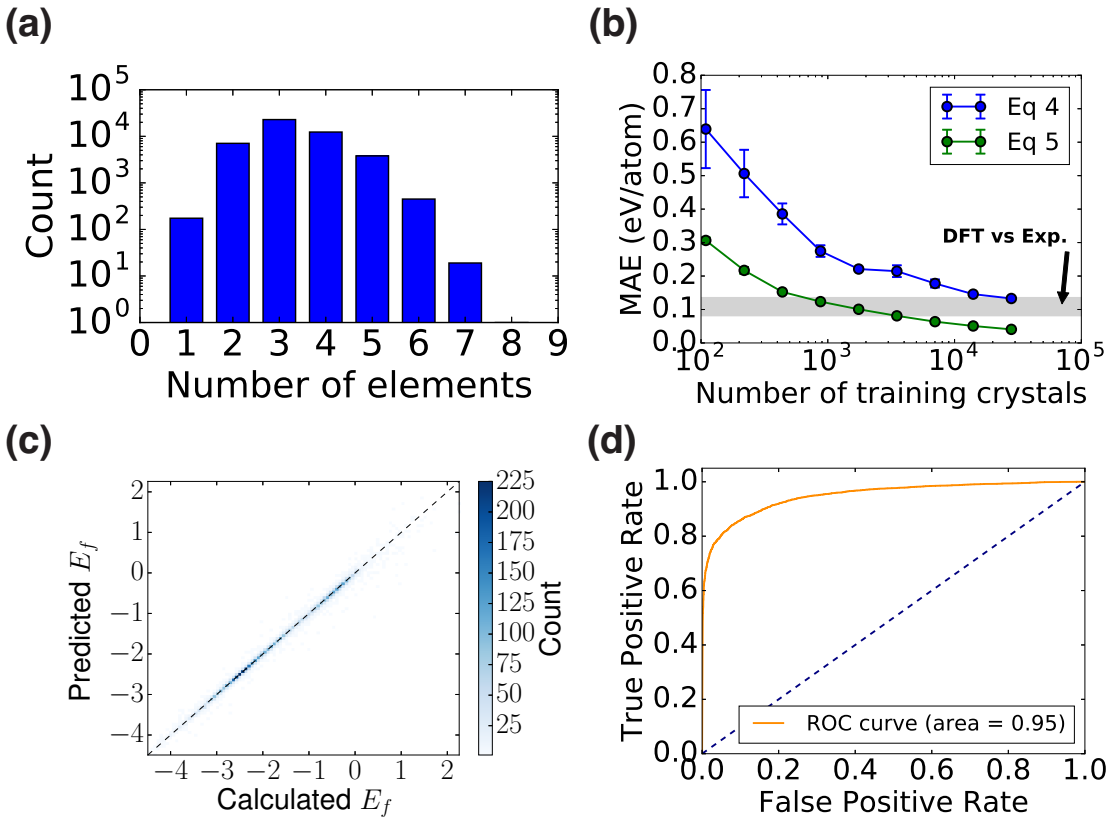


Figure 2-2: The performance of CGCNN on the Materials Project database[70]. (a) Histogram representing the distribution of the number of elements in each crystal. (b) Mean absolute error (MAE) as a function of training crystals for predicting formation energy per atom using different convolution functions. The shaded area denotes the MAE of DFT calculation compared with experiments[71]. (c) 2D histogram representing the predicted formation per atom against DFT calculated value. (d) Receiver operating characteristic (ROC) curve visualizing the result of metal-semiconductor classification. It plots the proportion of correctly identified metals (true positive rate) against the proportion of wrongly identified semiconductors (false positive rate) under different thresholds.

Table 2.2: Summary of the prediction performance of seven different properties on test sets.

Property	# of train data	Unit	$\text{MAE}_{\text{model}}$	MAE_{DFT}
Formation energy	28046	eV/atom	0.039	0.081–0.136[71]
Absolute energy	28046	eV/atom	0.072	–
Band gap	16458	eV	0.388	0.6[75]
Fermi energy	28046	eV	0.363	–
Bulk moduli	2041	log(GPa)	0.054	0.050[76]
Shear moduli	2041	log(GPa)	0.087	0.069[76]
Poisson ratio	2041	–	0.030	–

between different sources of experimental data. Given the comparison, our CGCNN approach provides a reliable estimation of DFT calculations and can potentially be applied to predict properties calculated by more accurate methods like GW[73] and quantum Monte Carlo[74].

After establishing the generality of CGCNN with respect to the diversity of crystals, we next explore its prediction performance for different material properties. We apply the same framework to predict the absolute energy, band gap, Fermi energy, bulk moduli, shear moduli, and Poisson ratio of crystals using DFT calculated data from the Materials Project[70]. The prediction performance of Eq. 2.5 is improved compared to Eq. 2.4 for all six properties (Table S4). We summarize the performance in Table 2.2 and the corresponding 2D histograms in Figure S4. As we can see, the MAE of our model are close to or higher than DFT accuracy relative to experiments for most properties when $\sim 10^4$ training data is used. For elastic properties, the errors are higher since less data is available, and the accuracy of DFT relative to experiments can be expected if $\sim 10^4$ training data is available.

In addition to predicting continuous properties, CGCNN can also predict discrete properties by changing the output layer. By using a softmax activation function for the output layer and a cross entropy cost function, we can predict the classification of metal and semiconductor with the same framework. In Figure 2-2(d), we show the receiver operating characteristic (ROC) curve of the prediction on 9350 test crystals. Excellent prediction performance is achieved with the area under the curve (AUC) at 0.95. By choosing a threshold of 0.5, we get metal prediction accuracy at 0.80,

semiconductor prediction accuracy at 0.95, and overall prediction accuracy at 0.90.

2.5 Application to the screening of solid electrolytes for batteries

2.5.1 Motivation

In this section, we aim to apply CGCNN to accelerate the discovery of an important type of materials – solid electrolytes for lithium metal batteries. Increased energy densities of Li-ion batteries are crucial for progress towards complete electrification of transportation [77–79]. Among the many possible routes, Li metal anodes have emerged as one of the most likely near-term commercialization options.[80] Coupled with a conventional intercalation cathode, batteries utilizing Li metal anodes could achieve specific energy of > 400 Wh/kg, much higher than the current state of the art $\sim$250 Wh/kg [81, 82]. Unstable and dendritic electrodeposition on Li metal anode coupled with capacity fade due to consumption of electrolyte has been one of the major hurdles in its commercialization [81, 83–88]. For large scale adoption, a stable, smooth and dendrite-free electrodeposition on Li metal is crucial.

Numerous approaches are being actively pursued for suppressing dendrite growth through the design of novel additives in liquid electrolytes [89–95], surface nanostructuring [96, 97], modified charging protocols[98, 99], artificial solid electrolyte interphase or protective coatings [100–102], polymers [103–105] or inorganic solid electrolytes [106–109]. Among these, dramatic improvements in the ionic conductivity of solid electrolytes [110, 111] have made them extremely attractive candidates for enabling Li metal anodes.

A comprehensive and precise criterion for dendrite suppression is still elusive. Interfacial effects [112, 113] and spatial inhomogeneities within the solid electrolyte like voids, grain boundaries and impurities [114] make the problem challenging. Monroe and Newman performed a dendrite initiation analysis and showed that solid polymer electrolytes with shear modulus roughly twice that of Li could achieve stable elec-

trodeposition [115]. In an earlier work, we extended this idea and showed that the criteria for the suppression of dendrite growth gets reversed for inorganic crystalline materials due to the difference in molar volume of Li^+. A softer solid electrolyte is required for stability in this case[116]. It is worth highlighting that this requirement applies only for dendrite initiation regime and other suppression approaches may be possible for the propagation regime. However, once initiated, dendrite growth is extremely hard to mitigate as pointed out by several studies[117–119]. Therefore, it is best to prevent dendrites from initiating to ensure smooth electrodeposition throughout cycling of the battery.

In recent years, high-throughput computational materials design has emerged as a major driver of discovery of novel materials for various applications [120, 121]. It typically involves a combination of first-principles quantum-mechanical approaches and database construction and mining techniques. Combined with machine learning methods that bypass the use of expensive quantum mechanical calculations through the use of structural descriptors [50, 122–125], one can accelerate the high-throughput screening by several orders of magnitude [126–128]. Previous high-throughput screening studies of solid electrolytes have used ionic conductivity, stability and electronic conductivity as screening criteria[126, 127]. However, dendrite suppression capability of solid electrolytes is an additional requirement that needs to be assessed.

Here, we carry out a large-scale data-driven search for solid electrolytes that might be promising candidates for suppressing dendrite growth during the initiation phase with a Li metal anode. We use machine learning techniques to train and predict the mechanical properties of inorganic solids which play a major role in stabilizing the interface. These properties are fed into the theoretical framework which uses the stability parameter [116, 129] to quantify the dendrite initiation with Li metal anode. At a mechanically isotropic interface, the screening results predict the crucial role of surface tension in stabilizing the interface since most solid electrolytes are not intrinsically stabilized by the stresses generated at the interface. Hence, surface nanostructuring may be essential to prevent initiation of dendrites for isotropic interfaces. We rank the materials based on the amount of nanostructuring (surface roughness

wavenumber) required for achieving a stable electrodeposition. We then performed a stability analysis of over 15,000 anisotropic interfaces between the Li metal and solid electrolyte using the Stroh formalism. This is essential to account for the highly anisotropic mechanical properties of Li[130] and texturing of electrodeposited Li at the interface[131]. A full anisotropic treatment of the interface reveals over twenty candidate interfaces that are predicted to suppress dendrite initiation. The materials obtained through screening are generally soft and with highly anisotropic mechanical properties. Since softer materials are generally faster ion conductors than stiffer materials due to availability of more volume per atom [132], the screened candidates present an opportunity to obtain both desirable mechanical properties and fast ion conduction.

2.5.2 Stability parameter

In solid electrolytes, the mechanical properties at the interface provide an additional degree of freedom for tuning the stability of electrodeposition. Previously, we developed a generalized stability diagram for assessing the stability of electrodeposition at a metal-solid electrolyte interface for isotropic mechanical response [116]. In these studies, we used the stability parameter first proposed by Monroe and Newman [119] to characterize the growth/decay of dendrites with time. The sign of the stability parameter, denoted hereafter as χ, determines whether the electrodeposition is stable or unstable. A positive χ implies higher current density at the peaks and lower current density at the valley leading to growth of dendrites while a negative χ leads to stabilization or suppression of dendrites. The stability parameter is related to the change in the electrochemical potential of the electron $\Delta\mu_{e^-}$ at a deformed interface $z = f(x)$ between the metal anode and the electrolyte (Fig. S1). It is convenient to compute properties of the interface in Fourier space with $f(x) = \int \mathrm{d}k[f_1(k)\cos(kx) + f_2(k)\sin(kx)]$ and then integrate over the surface roughness wavenumber k to obtain the overall behavior. The stability parameter can be calculated in a closed form at a given k. The change in the electrochemical potential at a given k is given by: $\Delta\mu_{e^-}(k) = \chi(k)[f_1(k)\cos(kx) + f_2(k)\sin(kx)]$ [129]. This

serves to define the stability parameter $\chi(k)$ at a given k as:

$$\chi(k) = \frac{\Delta\mu_{e^-}(k)}{f_1(k)\cos(kx) + f_2(k)\sin(kx)}; \tag{2.8}$$

$\Delta\mu_{e^-}$ can be obtained by including the effect of mechanical stresses and surface tension on the electrochemical potential of the species at a deformed interface as: [119]

$$\Delta\mu_{e^-} = -\frac{V_{\mathrm{M}}}{2\mathrm{z}}(1+v)\left(-\gamma\kappa - \mathbf{e_n}\cdot[(\boldsymbol{\tau_e} - \boldsymbol{\tau_s})\cdot\mathbf{e_n}]\right) + \frac{V_{\mathrm{M}}}{2\mathrm{z}}(1-v)\left(\Delta p_e + \Delta p_s\right). \tag{2.9}$$

From Eq. (2.9), $\Delta\mu_{e^-}$ depends on k through the surface tension γ, curvature κ, the hydrostatic stress Δp, and deviatoric stress $\boldsymbol{\tau}$ generated at the interface. e and s in the subscripts refer to the metal electrode (anode) and solid electrolyte respectively, V_{M} is the molar volume of the metal atom in the anode, v is the ratio of molar volume of the metal ion in electrolyte $V_{\mathrm{M}^{\mathrm{z}+}}$ to the metal atom in the anode V_{M}, z is the valence of the metal, and $\mathbf{e_n}$ is the normal to the interface pointing towards the electrolyte. The stability parameter consists of contributions from the surface tension and the stresses developed at the metal-electrolyte interface. For an isotropic metal anode with shear modulus G_e and Poisson's ratio ν_e in contact with an isotropic electrolyte with shear modulus G_s and Poisson's ratio ν_s, the stability parameter $\chi(k)$ can be computed exactly as[116]:

$$\begin{aligned}\chi = &\underbrace{-\frac{\gamma k^2 V_{\mathrm{M}}(1+v)}{2\mathrm{z}}}_{\text{surface tension}} \\ &+ \underbrace{\frac{2G_eG_skV_{\mathrm{M}}(1+v)(\nu_e(4\nu_s-3)-3\nu_s+2)}{\mathrm{z}(G_e(\nu_e-1)(4\nu_s-3)+G_s(4\nu_e-3)(\nu_s-1))}}_{\text{deviatoric stress}} \\ &+ \underbrace{\frac{kV_{\mathrm{M}}(1-v)\left(G_e^2(4\nu_s-3)+G_s^2(3-4\nu_e)\right)}{2\mathrm{z}(G_e(\nu_e-1)(4\nu_s-3)+G_s(4\nu_e-3)(\nu_s-1))}}_{\text{hydrostatic stress}}\end{aligned} \tag{2.10}$$

Using the shear modulus, Poisson's ratio and molar volume ratio of a solid electrolyte,

it is possible to calculate the stability parameter for its interface with Li metal anode and determine stability of electrodeposition. For a complete understanding of the interface growth and stability, it is necessary to determine the sign of stability parameter at all the Fourier components k. Fortunately, as we will see later, a negative stability parameter at a given k guarantees stability at all higher values.

The molar volume ratio $v = V_{M^{z+}}/V_M$ influences the range of shear moduli over which the electrodeposition is stable. $V_{M^{z+}}$ was calculated using the coordination number of Li in the crystal structure and mapping them to ionic radius using the values tabulated by Shannon [133]. The coordination number was calculated by generating polyhedra around a species through Voronoi analysis [134, 135] as implemented in pymatgen [136]. A linear interpolation was used for computing ionic radius corresponding to coordination numbers not in the Shannon's tabulated values. Predictions with $V_{M^{z+}} > V_M$ (true for just one candidate in the screening) were ignored since those correspond to very high Li coordination number where Shannon's tabulated values cannot be used. The partial molar volume of the metal in the electrolyte $V_{M^{z+}}$ can be measured directly in an experiment on the potential difference between a stressed and unstressed electrolyte as done by [137] and then using the relationship $V_{M^{z+}} = \partial\mu_{M^{z+}}/\partial p$ where $\mu_{M^{z+}}$ is the electrochemical potential of the metal ion.

2.5.3 Predicting the stability parameter with CGCNN ensembles

Since bulk and shear modulus are related to second derivatives of energy with respect to lattice constants at equilibrium, their calculation by first-principles requires fitting of the energy-strain relationship or the stress-strain relationship. Calculations on several deformed structures are required in order to get an accurate estimate of the fitting parameters. At each deformed state of the structure, the internal coordinates need to be relaxed to calculate the energy or the stress. The materials project database employed 24 relax calculations for a single material to compute the moduli. To perform a large scale screening over all Li-containing compounds (over

12,000) for use as solid electrolytes, it is necessary to choose a technique that can predict the properties reasonably accurately and without the high computational cost of multiple first-principles simulations. Hence, we used the crystal graph convolutional neural networks (CGCNN) framework [138] to predict the shear and bulk moduli of the crystalline solid electrolyte materials. At the core of the CGCNN is the multigraph representation of the crystal structure which encodes the atomic information and bonding interactions between atoms. The CGCNN builds a convolutional neural network directly on top of a multigraph that represents the crystal structure of the electrolytes, and predicts the elastic properties by extracting local structural features from the multigraph representation. Note that this method does not depend on any handcrafted geometric or topological features, and all the features are learned by the neural network automatically. This results in a model that is more general than the usual models replying on descriptors but also requires more data to train.

The training data for the mechanical properties required to compute χ through Eq. 2.10 was obtained from the materials project database [76, 139]. The calculated values in the database are typically within 15% of the experimental values which is sometimes the uncertainty in experimental data[76]. The moduli have been calculated using density functional theory (DFT) with the Perdew, Becke and Ernzerhof (PBE) Generalized Gradient Approximation (GGA) for the exchange-correlation functional [140]. GGA-level predictions for 104 systems were within 15% of the experimental value for all but 16 systems for the bulk modulus and 15 systems for the shear modulus [76]. Out of the outliers, many had a discrepancy of less than 10 GPa. Experimentally, the shear and bulk moduli can be calculated using the elastic tensor obtained through inelastic neutron scattering or pulse-echo measurements. The experimental measurements typically have a high degree of variability depending on the experimental technique and conditions. We used 2041 crystal structures with shear and bulk moduli, 60% of the entire dataset with elastic properties, to train our CGCNN model. We choose to minimize the mean squared errors between the log values of predicted and calculated elastic properties since we aim to minimize the relative prediction errors instead of absolute errors and avoid overweighing stiffer materials.

This also enabled us to always obtain positive values of the shear and bulk moduli. We then performed a hyperparameter optimization on 20% validation data via grid search to select the optimum learning rate, weight decay, and number of convolution layers. The best performing hyper-parameters are selected and the resulting model is evaluated on the rest 20% test data. The CGCNN was implemented in PyTorch [141] and the details of the architecture and optimized hyperparameters can be found in the Supporting Information and Ref. [138].

Considering the presence of uncertainty in training dataset, we developed a framework for obtaining uncertainty estimates on the results. The uncertainty in the model predictions was obtained by generating an ensemble of 100 CGCNNs using a random 60% of the training data for each model. Using each model, we obtained predictions of the shear and bulk modulus. The ensemble of moduli was then used together with Eq. (2.10) to obtain an ensemble of stability parameters for each material. The spread in the distribution of the stability parameter was used to quantify the uncertainty. Using the ensemble of stability parameters, we calculated the probability of stability P_s as the ratio of number of models that predict a negative stability parameter to the total number of models:

$$P_s = \frac{1}{N} \sum_{i=1}^{N} 1\{\chi_i < 0\} \tag{2.11}$$

Here N is the total number of models (100 in our calculations) and the indicator function $1\{X\}$ is equal to 1 if the condition X is true and 0 if it is false.

The performance of the CGCNN model was evaluated on 680 test data points. In Fig. 2-3a, we show the comparison between the shear modulus predicted by our model and the DFT-calculated value obtained from the materials project database and in Table 2.3, we show the root mean squared error (RMSE) for the shear and bulk moduli predicted by our model. The RMSE obtained using our model is comparable to previous work by [125]. However, it is worth noting that we evaluated our model on test data while de Jong et al. evaluated on the entire dataset, indicating that our model does not overfit and has better generalization capability.

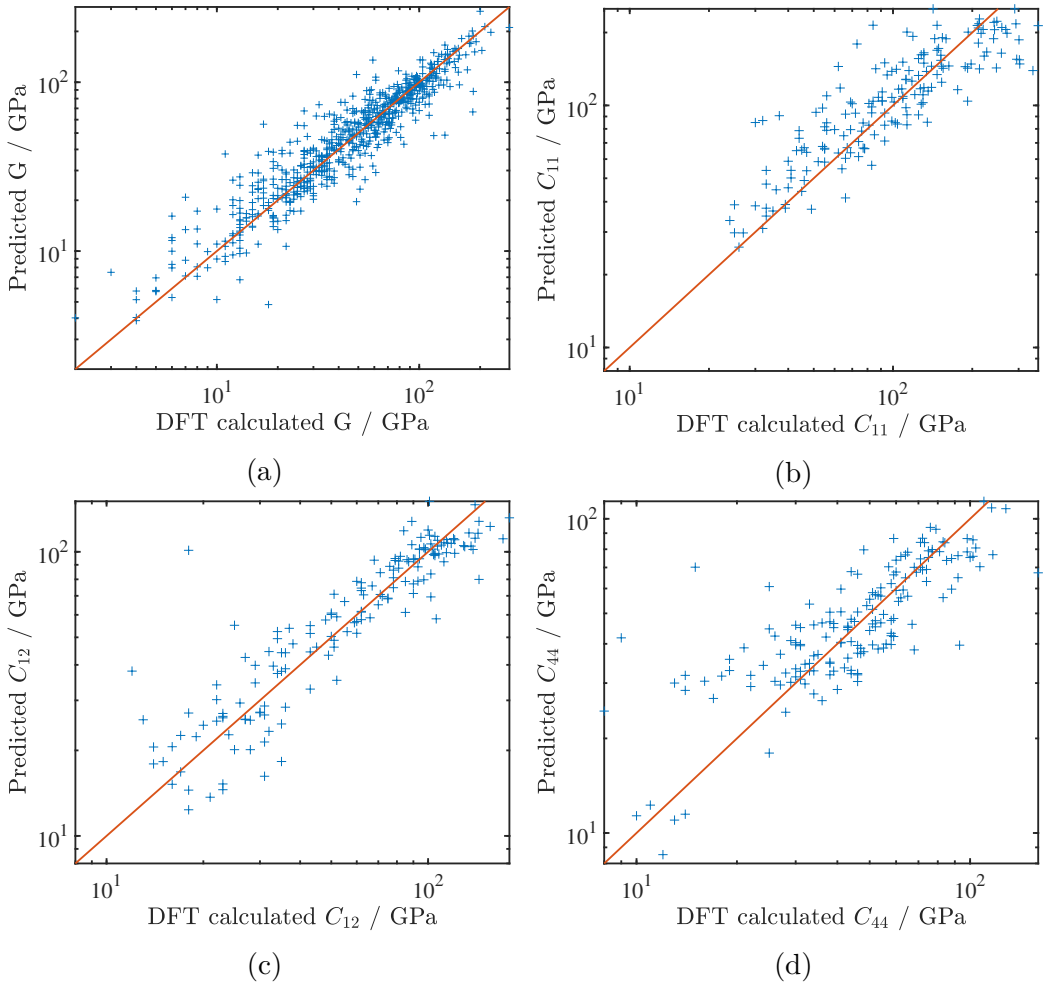


Figure 2-3: Parity plots comparing the elastic properties: (a) shear modulus G, and elastic constants (b) C_{11}, (c) C_{12} and (d) C_{44} predicted by the machine learning models to the DFT calculated values. The shear modulus is predicted using CGCNN and the elastic constants C_{11} and C_{44} are predicted using gradient boosting regression while C_{12} is predicted using Kernel ridge regression. The parity plot for shear modulus is on 680 test data points while that for the elastic constants contains all available data (170 points) where each prediction is a cross-validated value.

Table 2.3: Comparison of RMSE in log(GPa) for shear and bulk moduli

Method	log(G) RMSE	log(K) RMSE
This work	0.1268	0.1013
[125]	0.1378	0.0750

2.5.4 Screening of lithium containing compounds for interface stabilization

The shear modulus, Poisson's ratio and molar volume ratio $v = V_{M^{z+}}/V_M$ are the parameters determining the stability of electrodeposition at an interface where both materials are isotropic through Eq. 2.10. The role of surface tension in stabilizing electrodeposition is well established [96, 119, 142]. Since the contribution of the surface tension to the stability parameter increases as k^2 while that of stress increases linearly with k, the surface tension starts dominating the stability parameter as k is increased. This is elucidated in Fig. 2-4 through the contributions of the different terms to the total stability parameter for a material with $G = 3.4$ GPa and $v = 0.1$. The red line shows the fraction of contribution of surface tension to the overall stability parameter. All interfaces become stabilized as the value of k is increased beyond the critical surface roughness wavenumber. This motivates a distinction between two types of solid electrolytes – ones that are stabilized by the stress term alone and those that are stabilized by the surface tension beyond the critical value of k. For materials that are stabilized by stresses, the stability parameter remains negative for all values of surface roughness, and therefore, stability is guaranteed. However, for materials that have an overall destabilizing contribution due to stresses (hydrostatic + deviatoric), the stability parameter changes sign at an intermediate value of k since $\chi \to -\infty$ as $k \to \infty$. For such materials, the electrodeposition become stable at the critical surface wavenumber $k_{crit} = 2\pi/\lambda_{crit}$ (Fig. 2-4). If the surface roughness wavenumber so obtained is possible to achieve by nanostructuring the interface [96], the electrodeposition might be stabilized.

We calculated the stability parameter for 12,950 Li-containing compounds out of which the properties of ∼3400 were in training data and those of the remaining were predicted using CGCNN. In Fig. 2-5, we visualized the latent space representations of randomly selected 500 training and 500 predicted crystals in a two-dimensional plot using t-distributed stochastic neighbor embedding (t-SNE) algorithm. It is clear that training crystals cover the search space of predicted crystals, indicating the

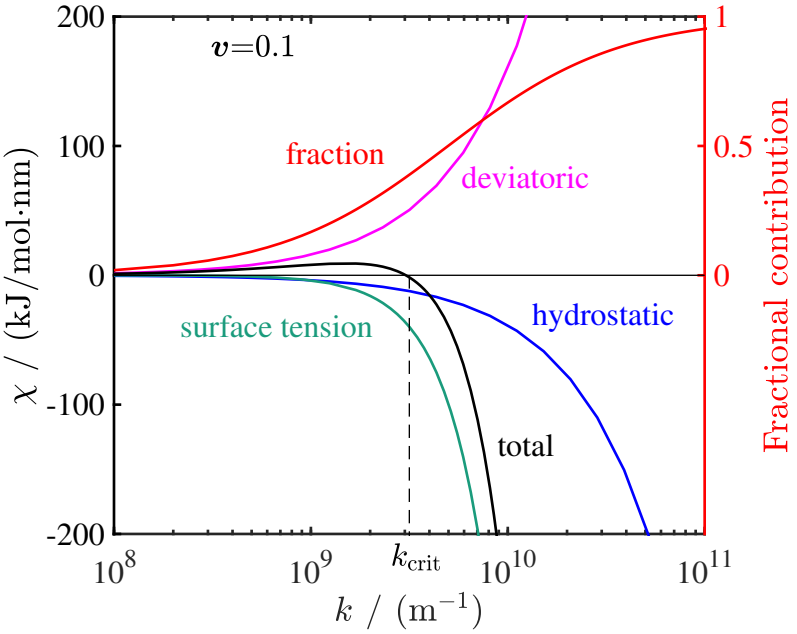


Figure 2-4: Contribution of hydrostatic stress, deviatoric stress and surface tension to the stability parameter as a function of surface roughness wavenumber. The surface tension term starts dominating at high k and ultimately stabilizes the interface after $k = k_{\text{crit}}$. The contributions are plotted for a material with shear modulus ratio $G/G_{\text{Li}} = 1$ and Poisson's ratio $\nu = 0.33$ which is not stable ($\chi > 0$) at $k = 10^8$ m^{-1}. The red line shows the fraction of surface tension contribution to the stability parameter obtained by dividing the absolute value of its contribution by the sum of absolute values of all components.

reliability of the prediction. The lower part of Fig. 2-5 includes compounds that do not contain Li atoms, which helps improving prediction even though they are not directly representative of the search space. The ensemble averaged stability parameter χ and the critical surface roughness λ_{crit} for all materials are shown as a histogram in Fig. 2-6. We found that none of the materials have a probability of stability over 5% at surface roughness wavenumber $k = 10^8$ m^{-1} [115].

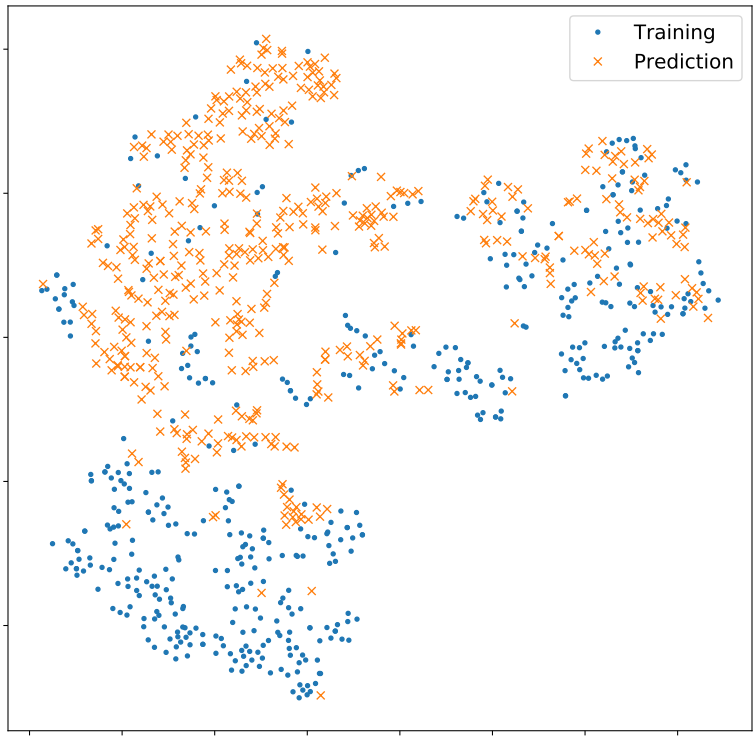


Figure 2-5: Visualization of the latent space representations of 500 random training and 500 random test crystals using t-distributed stochastic neighbor embedding algorithm for CGCNN.

The absence of any materials that can suppress dendrites without assistance from surface tension becomes clear from the isotropic stability diagram shown in Fig. 2-7. All materials have G/G_{Li} ratio higher than the critical value required to stabilize electrodeposition [116]. The highest number of materials are found in the region

where $G/G_{\mathrm{Li}} \sim 15$ and $v \sim 0.1$. The critical wavelength of surface roughening was used as the criteria for screening materials since a higher surface roughness is easier to achieve by nanostructuring.

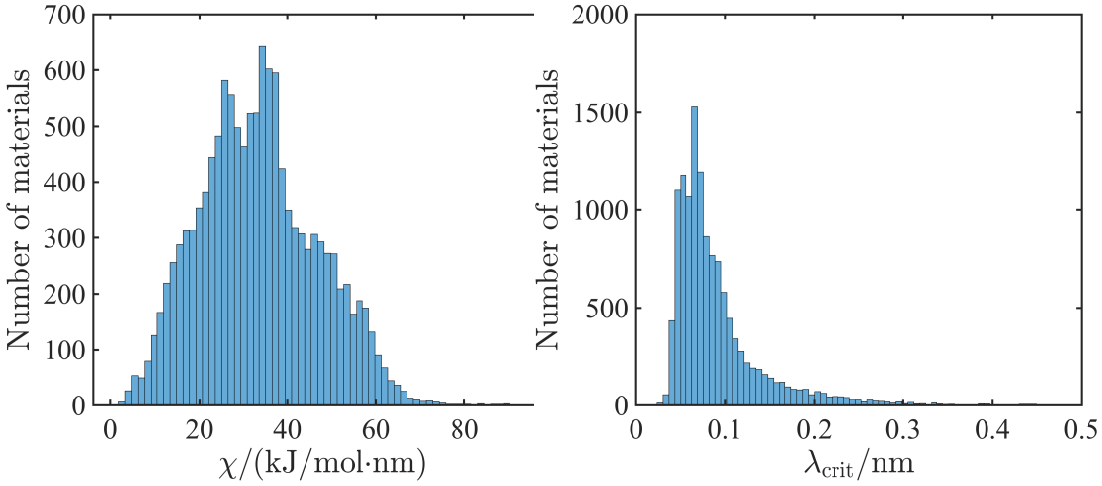


Figure 2-6: Results of isotropic screening for 12,950 Li- containing compounds. Distribution of ensemble averaged (a) stability parameter for isotropic Li-solid electrolyte interfaces at $k = 10^8$ m^{-1} and (b) critical wavelength of surface roughness required for stability. None of the materials in the database can be stabilized without the aid of surface tension. The required critical surface roughness wavenumber depends on the contribution of the stress term in the stability parameter.

The candidate materials with highest critical wavelength of roughening λ_{crit} are shown in Table 2.4 along with their stability parameter at different surface roughness wavenumbers and the corresponding probability of stability. While performing screening, we removed all materials which are electronically conducting i.e. those which have a zero band gap according to materials project database. However, we retained materials which were thermodynamically metastable (with energy above hull less than 0.1 eV) since many such solid electrolytes like $Li_{10}GeP_2S_{12}$ [110] and $Li_7P_3S_{11}$ [143] have been successfully synthesized. We find several candidate electrolytes with probability of stability P_s over 5% at surface roughness wavelength of 1 nm. It is worth noting that our screening identifies sulfide, borohydride and iodide-based solid electrolytes, classes to which many of the current solid electrolytes belong. The uncertainty in stability parameter is much higher at high surface roughness wavenumber.

Table 2.4: Solid electrolyte screening results for stable electrodeposition with Li metal anode together with their materials project id ranked by critical wavelength of surface roughening λ_{crit} required to stabilize electrodeposition. χ is the stability parameter in kJ/mol·nm which needs to be negative for stability, and $k = 2\pi/\lambda$ is the surface roughness wavenumber. Low k corresponds to $k = 10^8$ m^{-1} while high k corresponds to a wavelength $\lambda = 2\pi/k = 1$ nm. Only materials with probability of stability $P_s > 0.05$ at high k are shown. Uncertainty in χ and λ_{crit} (standard deviation of their distributions) and P_s are only shown for materials whose properties were predicted using CGCNN and not for those whose properties were available in training data.

Formula	Space Group	MP id	Low k		High k		λ_{crit}/nm
			χ	P_s	χ	P_s	
Li_2WS_4	P$\bar{4}$2m	mp-867695	0.62	-	-109.26	-	3.64
Li_2WS_4	I$\bar{4}$2m	mp-753195	1.75	-	-38.54	-	1.34
$LiBH_4$	P$\bar{1}$	mp-675926	1.98	-	-40.13	-	1.32
$LiAuI_4$	P2$_1$/c	mp-29520	2.7 ± 0.9	0	16.1 ± 55.2	0.43	1.02 ± 0.40
$LiGaI_4$	P2$_1$/c	mp-567967	3.2 ± 1.1	0	48.6 ± 67.0	0.28	0.85 ± 0.29
$LiWCl_6$	R3	mp-570512	3.2 ± 0.9	0	51.3 ± 56.6	0.17	0.82 ± 0.27
Cs_3LiI_4	P2$_1$/m	mp-569238	3.1 ± 0.7	0	46.9 ± 43.4	0.15	0.80 ± 0.17
$LiInI_4$	P2$_1$/c	mp-541001	3.5 ± 1.0	0	68.5 ± 62.8	0.12	0.74 ± 0.20
$Cs_2Li_3I_5$	C2	mp-608311	3.6 ± 0.9	0	77.2 ± 59.0	0.05	0.71 ± 0.17
$Ba_{19}Na_{29}Li_{35}$	F$\bar{4}$3m	mp-569025	4.2 ± 1.3	0	101.9 ± 81.3	0.08	0.68 ± 0.19
$Ba_{38}Na_{58}Li_{26}N$	F$\bar{4}$3m	mp-570185	4.2 ± 1.3	0	104.5 ± 82.3	0.08	0.67 ± 0.20
Li_2UI_6	P$\bar{3}$1c	mp-570813	4.2 ± 1.4	0	111.5 ± 86.8	0.11	0.66 ± 0.29

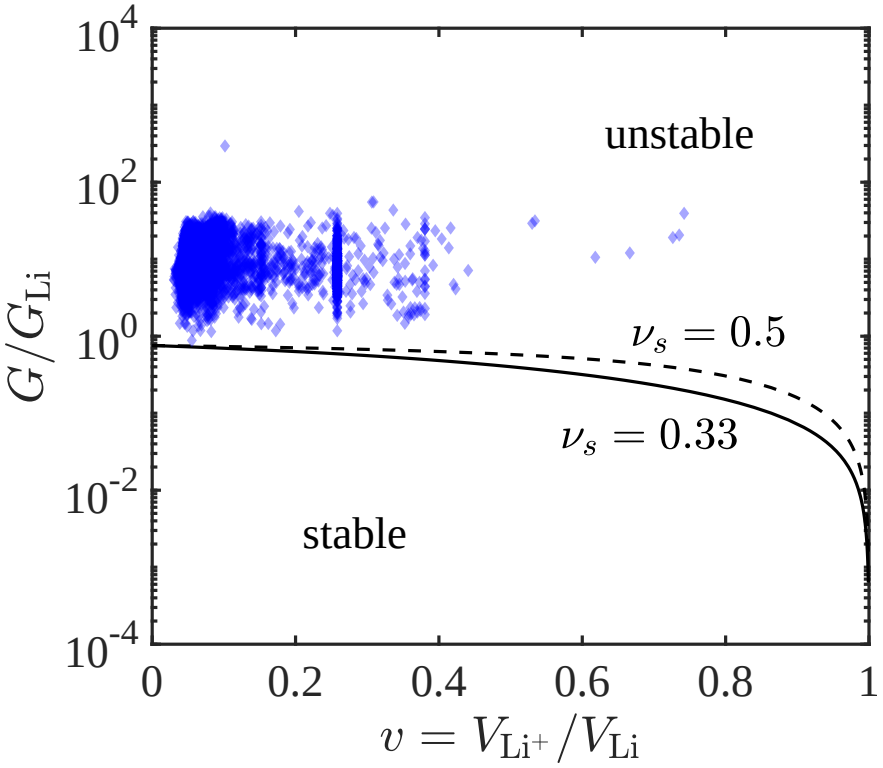


Figure 2-7: Isotropic stability diagram showing the position of all solid electrolytes involved in the screening. G_{Li} is the shear modulus of Li=3.4 GPa. The critical G/G_{Li} line separating the stable and unstable regions depends weakly on the Poisson's ratio, so the lines corresponding to $\nu_s = 0.33$ and 0.5 are good indicators for assessment of stability. The darker regions indicate more number of materials in the region.

Chapter 3

Visualization of crystal graph convolutional neural networks

3.1 Introduction

3.1.1 Chapter overview

In this chapter, we hope to understand chemical insights by visualizing the representations learned by the crystal graph convolutional neural networks (CGCNN) framework [46] we developed in chapter 2. Since the graph representation of solid materials has a clear correspondence between nodes/edges and atoms/bonds, visualizing the representations learned by the graph neural networks might help us understand the solid systems. We will first define a variant of CGCNN that has clear physical meanings in the first and last layers of its network. Then, we use the network to visualize three material spaces: perovskites, elemental boron, and general inorganic crystals, covering material spaces of different compositions, different structures, and both, respectively. We show that in all three cases pattern emerges automatically that might aid in the design of new materials, and we discover some empirical rules for understanding material stability that are consistent with our intuition.

3.1.2 Theoretical and practical motivations

From a theoretical perspective, we aim to explore the physical means of the representations learned by the graph neural networks in the context of solid materials. There are three types of representations learned by the graph neural networks: 1) node representations $\boldsymbol{v}_i^{(t)}$, 2) edge representations $\boldsymbol{u}_{(i,j)_k}$, 3) global representations $\boldsymbol{v}_c$. With the graph representation of solid materials, they correspond to 1) atoms, 2) bonds connecting atoms, 3) the entire material. Therefore, visualizing the similarities between these learned representations provide a way to understand the similarities between arbitrary solid materials at multiple scales.

From a practical perspective, we are interested in whether the representations learned by the model can result in knowledge that can be understood by human and help scientists to explore the vast space of solid materials. Efficient exploration of the materials space has been central to material discovery as a result of the limited experimental and computational resources compared with its vast size. Often compositional or structural patterns are sought from past experiences that might guide the design of new materials, improving the efficiency of material exploration [144–148]. Emerging high-throughput computation and machine learning techniques directly screen large amounts of candidate materials for specific applications [13, 149–155], which enables fast and direct exploration of the material space. However, the large quantities of material data generated makes the discovery of patterns challenging with traditional, human-centered approaches. Instead, an automated, data-centered method to visualize and understand a given materials design phase space is needed in order to improve the efficiency of exploration.

3.1.3 Related prior research

The key in visualizing material space is to map materials with different compositions and structures into a lower dimensional manifold where the similarity between materials can be measured by their Euclidean distances. One major challenge in finding such manifolds is to develop a unified representation for different materials. A

widely-used method is representing materials with feature vectors, where a set of descriptors are selected to represent each material [122, 156, 157]. There are also methods that automatically select descriptors that are best for predicting a desired target property [53]. Recent work has also developed atomic-scale representations to map complex atom configurations into low dimensional manifolds, such as atom centered symmetry functions [158], social permutation invariant (SPRINT) coordinates [159], global minimum of root-mean-square distance [160], smooth overlap of atomic positions (SOAP) [161], and many other methods [55, 162, 163]. These representations often have physically meaningful parameters that can highlight some structural or chemical features. Often material descriptors and atomic representations are used together to combine compositional and structural information [162, 164]. They have been used to visualize the material and molecular similarities[54, 165, 166], as well as explore the complex configurational space of biological systems [167–170] and water structures [16, 171]. In addition to Euclidean distances, similarity kernels are also used to compare material similarities [165, 166]. Combined with machine learning algorithms, these representations were also used to predict material properties [50, 51, 53, 55, 122, 154, 156, 157] and construct force fields [161, 172, 173].

In parallel to these efforts, the success of "deep learning" has inspired a group of representations purely based on neural networks. Instead of designing descriptors or atomic representations that are fixed or contain several physically meaningful parameters, they use relatively general neural network architectures with a large number of trainable weights to learn a representation directly. This field started with building neural networks on molecular graphs [66, 174–176], and was recently expanded to periodic material systems by us [46] and Schutt et al. [177]. The deep neural networks have shown many advantages over conventional machine learning methods given large amounts of data in computer vision and speech recognition [178], and they outperform conventional methods on 11/17 datasets for predicting molecular properties in a recent study [179]. However, the general neural network architecture may also limit performance when the data size is small since there is no material specific information built-in. It is worth noting that many machine learning force fields

combine atomic representations and neural networks [158, 172, 180], but they usually deal with different compositions separately and use a significantly smaller number of weights. It has been shown that the hidden layers of these neural networks can learn physically meaningful representations by proper design of the network architecture. For instance, several works have investigated the ideas of learning atom energies [46, 176, 181] and elemental similarities [182, 183]. In addition, recent work showed that element similarities can also be learned using a specially designed SOAP kernel [184].

3.2 Methods

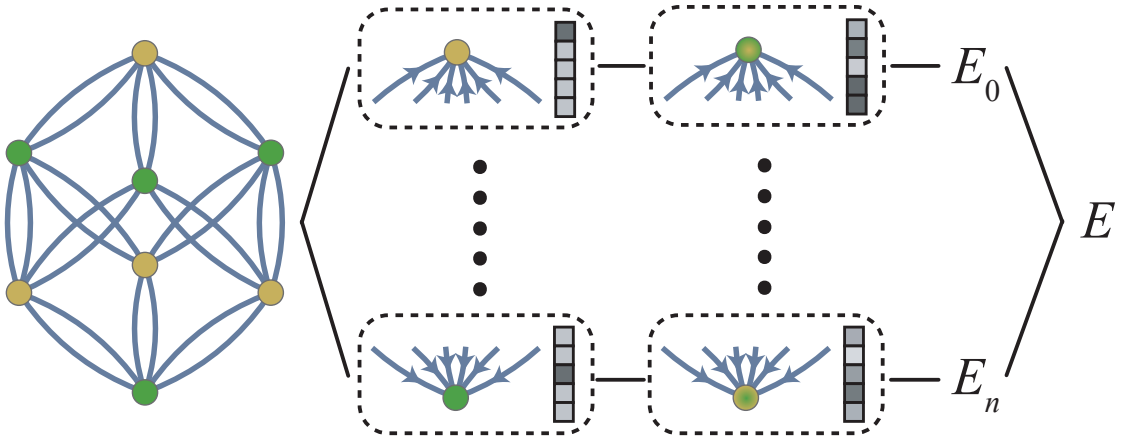


Figure 3-1: The structure of the crystal graph convolutional neural networks.

To visualize the crystal space at different scales, we design a variant of CGCNN [46] that has meaningful interpretation at different layers of the neural network. The learned CGCNN network provides a vector representation of the local environments in each crystal that only depends on its composition and structure without any human designed features, enabling us to explore the materials space hierarchically.

We first represent the crystal structure with a multigraph $\mathcal{G}$ that encodes the connectivity of atoms in the crystal. Each atom is represented by a node i in $\mathcal{G}$ which stores a vector $\boldsymbol{v}_i$ corresponding to the element type of the atom. To avoid introducing any human bias, we set $\boldsymbol{v}_i$ to be a random 64 dimensional vector for each element and allow it to evolve during the training process. Then, we search for the

12 nearest neighbors for each atom and introduce an edge $(i,j)_k$ between the center node i and neighbor j. The subscript k indicates that there can be multiple edges between the same end nodes as a result of the periodicity of the crystal. The edge $(i,j)_k$ stores a vector $\boldsymbol{u}_{(i,j)_k}$ whose tth element depends on the distance between i and j by,

$$\boldsymbol{u}_{(i,j)_k}[t] = \exp(-(d_{(i,j)_k} - \mu_t)^2/\sigma^2) \tag{3.1}$$

where $\mu_t = t \cdot 0.2\,\text{Å}$ for $t = 0, 1, ..., 40$ and $\sigma = 0.2\,\text{Å}$.

In graph $\mathcal{G}$, each atom i is initialized by a vector $\boldsymbol{v}_i$ whose value solely depends on the element type of atom i. We call this iteration 0 where

$$\boldsymbol{v}_i^{(0)} = \boldsymbol{v}_i \tag{3.2}$$

Then, we perform convolution operations on the multigraph $\mathcal{G}$ with the convolution function designed in Ref. [46] which allows atom i to interact with its neighbors iteratively. In iteration t, we first concatenate neighbor vectors $\boldsymbol{z}_{(i,j)_k}^{(t-1)} = \boldsymbol{v}_i^{(t-1)} \odot \boldsymbol{v}_j^{(t-1)} \odot \boldsymbol{u}_{(i,j)_k}$, and then perform the convolution by same as in Eq. 2.5,

$$\boldsymbol{v}_i^{(t)} = \boldsymbol{v}_i^{(t-1)} + \sum_{j,k} \sigma(\boldsymbol{z}_{(i,j)_k}^{(t-1)} \boldsymbol{W}_f^{(t-1)} + \boldsymbol{b}_f^{(t-1)}) \odot g(\boldsymbol{z}_{(i,j)_k}^{(t-1)} \boldsymbol{W}_s^{(t-1)} + \boldsymbol{b}_s^{(t-1)}) \tag{3.3}$$

where $\odot$ denotes element-wise multiplication, σ denotes a sigmoid function, and g denotes any non-linear activation function, and $\boldsymbol{W}$ and $\boldsymbol{b}$ denotes weights and biases in the neural network, respectively. During these convolution operations, $\boldsymbol{v}_i^{(t)}$ forms a series of representations of the local environments of atom i at different scales.

After K iterations, we perform a linear transformation to map $\boldsymbol{v}_i^{(K)}$ to a scalar E_i,

$$E_i = \boldsymbol{v}_i^{(K)} \boldsymbol{W}_l + b_l \tag{3.4}$$

and then use a normalized sum pooling to predict the averaged total energy per atom of the crystal,

$$E = \frac{1}{n} \sum_i E_i \tag{3.5}$$

where n is the number of atoms in the crystal. This introduces a physically meaningful term E_i to represent the energy of the local chemical environment.

The model is trained by minimizing the squared error between predicted properties relative to the DFT calculated properties using backpropagation and stochastic gradient descent.

In this CGCNN model, each vector represents the local environment of each atom at different scales. Here, we focus three vectors that has the most representative physical interpretations.

1. *Element representation* $\boldsymbol{v}_i^{(0)}$ that depends completely on the type of element that atom i is composed of, describing the similarities between elements.

2. *Local environment representation* $\boldsymbol{v}_i^{(K)}$ that depends on atom i and its Kth order neighbors, describing the similarities between local environments that combines the compositional and structural information.

3. *Local energy representation* E_i that describes the energy of atom i.

3.3 Visualization for different material spaces

3.3.1 Overview

To illustrate how this method can help visualize the compositional the structural aspects of the crystal space, we apply it to three datasets that representing different material spaces. 1) a group of perovskite crystals that share the same structure type but have different compositions; 2) different configurations of elemental boron that share the same composition but have different structures; and 3) inorganic crystals from the Materials Project [139] that have both different compositions and different structures.

For each material space, we train the CGCNN model with 60% of the data to predict the energy per atom of the materials. 20% of the data are used to select hyperparameters of the model and the last 20% are reserved for testing. In Fig. 3-

2, we show the learning curves for the three representative material spaces where a subset of training data is used to show how the number of training data affects the model prediction performance. As we will show below, the representations learned by predicting the energies automatically gain physical meanings and can be used to explore the materials spaces.

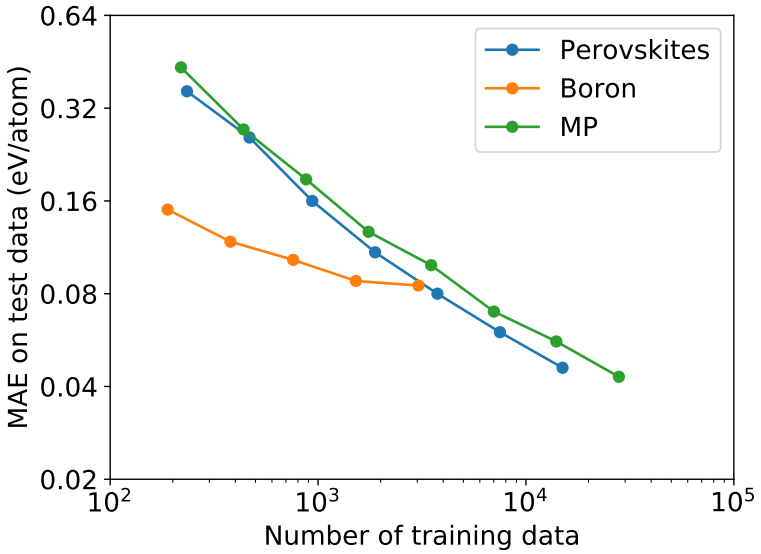


Figure 3-2: Learning curves for the three representative material spaces. The mean absolute errors (MAEs) on test data is shown as a function of the number of training data for the perovskites [185, 186], elemental boron [181], and materials project [139] datasets.

3.3.2 Perovskite: compositional space

First, we explore the compositional space of perovskites by visualizing the *element representations*. Perovskite is a crystal structure type with the form of ABC_3 as shown in Fig. 3-3(a). The dataset [185, 186] that we used includes 18,928 different perovskites where the elements A and B can be any nonradioactive metals and the element C can be one or several from O, N, S, and F. We trained our model to predict the energy above hull with 15,000 training data, and after hyperparameter optimization on 1,890 validation data, we achieve a prediction mean absolute error (MAE) of 0.042 eV/atom on 2,000 test data. The prediction performance is excellent

and lower than several recent ML models such as those of Schmidt et al. (0.121 eV/atom) [182] and Xie et al. (0.099 eV/atom) [46]. The learning curve in Fig. 3-2 shows a straight line in log-log scale, indicating a steady increase of prediction performance as the number of training data increases.

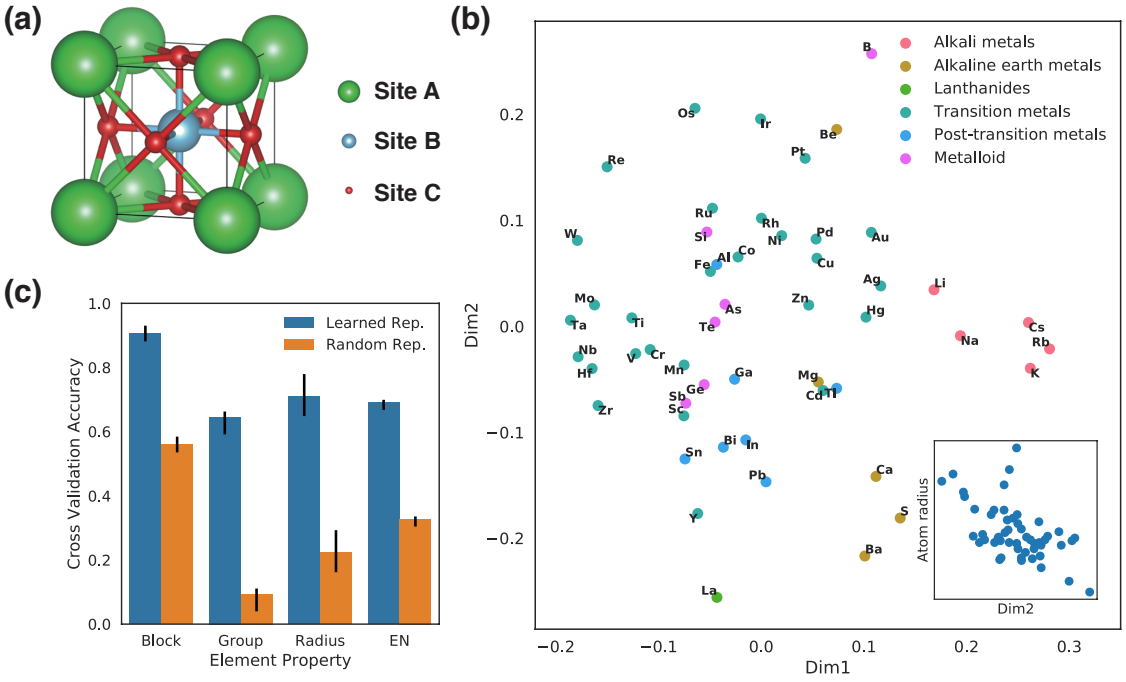


Figure 3-3: Visualization of the element representations learned from the perovskite dataset. (a) The perovskite structure type. (b) Visualization of the two principal dimensions with principal component analysis. (c) Prediction performance of several atom properties using a linear model on the element representations.

In Fig. 3-3(b)(c), the element representation $\boldsymbol{v}_i^{(0)}$, a 64 dimensional vector, is visualized for every nonradioactive metal element after training with the perovskite dataset. Fig. 3-3(b) shows the projection of these element representations on a 2D plane using principal component analysis, where elements are colored according to their elemental groups. We can clearly see that similar elements are grouped together based on their stability in perovskite structures. For instance, alkali metals are grouped on the right of the plot due to their similar properties. The large alkaline earth metals (Ba, Se, and Ca) are grouped on the bottom, distinct from Mg and Be, because their larger radius stabilizes them in the perovskite structure. On the left

side are elements such as W, Mo, and Ta that favor octahedral coordinations due to their configuration of d electrons, which might be related to their extra stability in the B site [46]. Interestingly, we can also observe a trend of decreasing atom radius from the bottom of the plot to the top as shown in the insert of Fig. 3-3(b), except for the alkali metals as outliers. This indicates that CGCNN learns the atom radius as an important feature for perovskite stability. Recently, Schutt et al. also discovered similar grouping of elements with data from the Materials Project [177]. In general, these visualizations can help discover similarities between elements for designing novel perovskite structures.

However, these 2D plots only account for part of the 64-dimensional element representation vectors. To fully understand how element properties are learned by CGCNN, we use linear logistic regression (LR) models to predict the block type, group number, radius, and electronegativity of each element from their learned representation vectors. In Fig. 3-3(c), we show the 3-fold cross validation accuracy of the LR models and compare them with LR models learned from random representations, which helps to rule out the possibility that the predictions are caused by coincidences. We discover a significantly higher prediction accuracy of the learned representations for all four properties, demonstrating that the element representations can reflect multiple aspects of element properties. For instance, the model predicts the block of the element with over 90% accuracy, and the same representation also predicts the group number, radius, and electronegativity with over 60% accuracy. This is surprising considering that there are 16 different elemental groups represented. It is worth noting that these representations are learned only from the perovskite structures and the total energy above hull, but they are in agreement with these empirical element properties reflecting decades of human chemical intuition.

Second, we visualize the *local energy representations* to understand how each site in the perovskite structure affects its stability. Figure 3-4(a, b) visualizes the mean of the predicted site energies when each element occupies the A and B site respectively. The most stable elements that occupy the A site are those with large radii due to the space needed for 12 coordinations. In contrast, elements with small radii like Be, B, Si

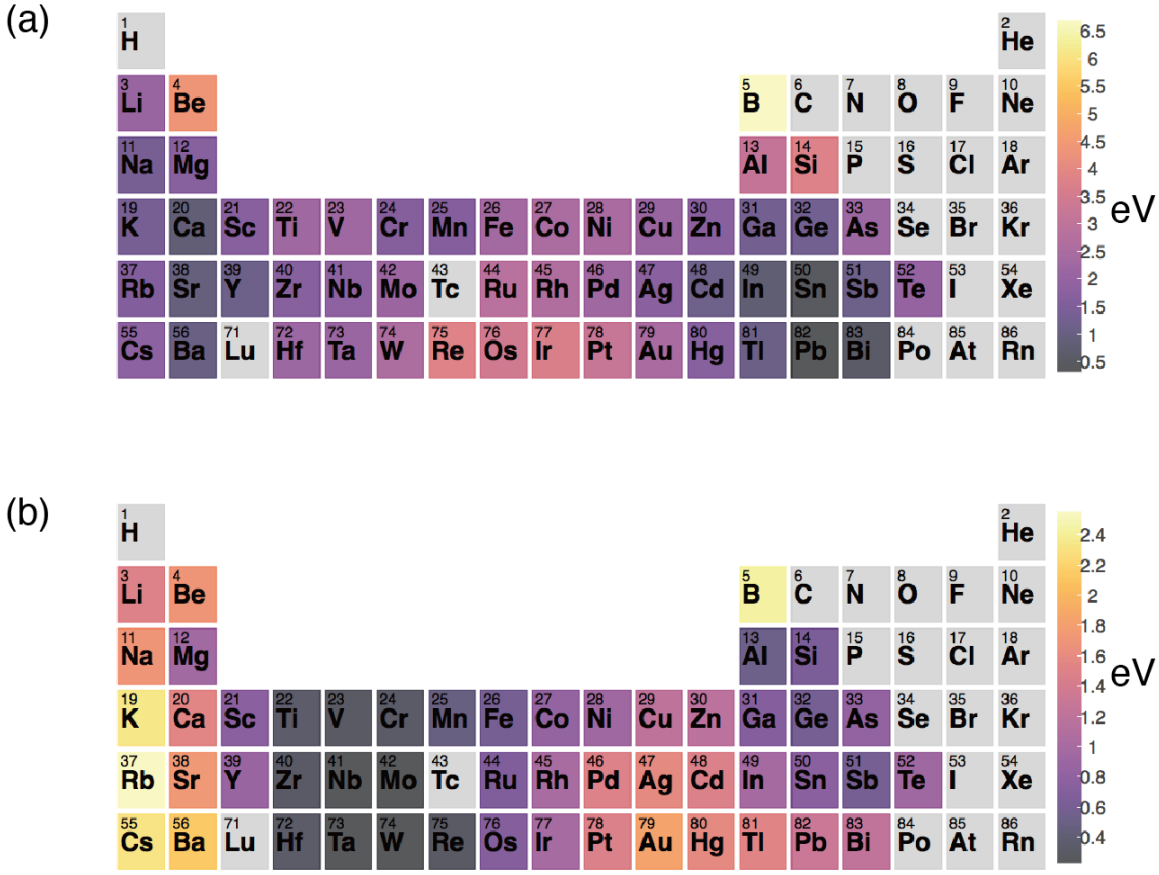


Figure 3-4: Extraction of site energy of perovskites from total energy above hull. (a, b) Periodic table with the color of each element representing the mean of the site energy when the element occupies A site (c) or B site (d).

are the most unstable for occupying the A site. For the B site, elements in groups 4, 5, and 6 are the most stable throughout the periodic table. This can be explained by crystal field theory, since the configuration of d electrons of these elements favors the octahedral coordination in the B site. Interestingly, the visualization shows that large atoms from groups 13-15 are stable in the A site, in addition to the well-known region of groups 1-3 elements. Inspired by this result, we applied a combinational search for stable perovskites using elements from group 13-15 as the A site and group 4-6 as the B site. Due to the theoretical inaccuracies of DFT calculations and the possibility of metastable phases that can be stabilized by temperature, defects, and substrates, many synthesizable inorganic crystals have positive calculated energies above hull at 0 K. Some metastable nitrides can even have energies up to 0.2 eV/atom above hull as a result of the strong bonding interactions[187]. In this work, since some of the perovskites are also nitrides, we choose to set the cutoff energy for potential synthesizability at 0.2 eV/atom. We discovered 33 perovskites that fall within this threshold out of 378 in the entire dataset, among which 8 are within the cutoff out of 58 in the test set (Table 3.1). Many of these compounds like $PbTiO_3$[188], $PbZrO_3$[188], $SnTaO_3$[189], and $PbMoO_3$[190] have been experimentally synthesized. Note that $PbMoO_3$ has calculated energy 0.18 eV/atom above hull, indicating that our choice of cutoff energy is reasonable. In general, chemical insights gained from CGCNN can significantly reduce the search space for high throughput screening. In comparison, there are only 228 potentially synthesizable perovskites out of 18928 in our database: the chemical insight increased the search efficiency by a factor of 7.

3.3.3 Elemental boron: structural space

As a second example, we explore the structural space of elemental boron by visualizing the *local environment representations* and the corresponding *local energies*. Elemental boron has a number of complex crystal structures due to its unique, electron-deficient bonding nature [181, 191]. We use a dataset that includes 5038 distinct elemental boron structures and their total energies calculated using density functional theory [181]. We train our CGCNN model with 3038 structures, and perform

Table 3.1: Perovskites with energy above hull lower than 0.2 eV/atom discovered using combinational search.

Formula	A site	B site	Formation energy per atom (eV/atom)
Training Set (60%)			
TlNbO3	Tl	Nb	0.0
SnTiO3	Sn	Ti	0.1
PbVO3	Pb	V	0.04
SnTaO3	Sn	Ta	0.0
TlWO3	Tl	W	0.12
PbMoO3	Pb	Mo	0.18
PbCrO3	Pb	Cr	0.14
SnNbO3	Sn	Nb	0.14
SnTaO2N	Sn	Ta	0.14
TlTaOFN	Tl	Ta	0.18
TlTaO2F	Tl	Ta	0.04
TlHfO2F	Tl	Hf	0.18
PbTiO3	Pb	Ti	0.06
InNbO3	In	Nb	0.06
InWO3	In	W	0.18
InTaO3	In	Ta	-0.16
InNbO2F	In	Nb	0.18
InTaO2S	In	Ta	0.18
Validation Set (20%)			
TlNbO2F	Tl	Nb	0.08
TlZrO2F	Tl	Zr	0.14
SnVO3	Sn	V	0.12
TlTiO2F	Tl	Ti	-0.02
PbNbO2N	Pb	Nb	0.18
PbZrO3	Pb	Zr	0.08
PbNbO3	Pb	Nb	0.04
Test Set (20%)			
BiCrO3	Bi	Cr	0.14
PbVO2F	Pb	V	0.14
SnNbO2N	Sn	Nb	0.18
PbHfO3	Pb	Hf	0.1
TlTaO3	Tl	Ta	0.1
PbTaO3	Pb	Ta	0.18
InTaO2F	In	Ta	0.08
InZrO2F	In	Zr	0.16

hyperparameter optimization with 1000 validation structures. The MAE of predicted energy relative to DFT results on the remaining 1000 test structures is 0.085 eV/atom. The learning curve in Fig. 3-2 shows a much smaller slope compared with the other material spaces. One explanation is that there exist many highly unstable boron structures in the dataset, whose energies might be hard to predict given the limited structures covered by the training data.

In Fig. 3-5, 1000 randomly sampled boron local environment representations are visualized in 2 dimensions using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [193]. We observe primarily four different regions of different boron local environments, and we discover a smooth transition of local energy, number of neighbor atoms, and the density between different regions. The disconnected region consists of boron atoms at the edge of boron clusters [Fig. 3-6(a-c)]. These atoms have very high local energies and lower number of neighbors, as to be expected, and their density varies depending on the distances between clusters. The amorphous region includes boron atoms in a relatively disordered local configuration, and their local energies are lower than the disconnected counterparts but higher than other other configurations [Fig. 3-6(d-f)]. We can see that the number of neighbors fluctuates drastically in these two regions due to the relatively disordered local structures. The layered region is composed of boron atoms in layered boron planes, where neighbors on one side are closely bonded and the neighbors on the other side are further away [Fig. 3-6(g-i)]. The B_{12} icosahedron region includes boron local environments with the lowest local energy, which have a characteristic icosahedron structure [Fig. 3-6(j-l)]. The local environments in each region share common characteristics but are slightly different in detail. For instance, most boron atoms in the B_{12} icosahedron region are in a slightly distorted icosahedron, and the local environments in Fig. 3-6(l) only have certain features of an icosahedron. Note that these representations are rather localized. The global structure of Fig. 3-6(c) is layered, but the representation of the highlighted atom at the edge is closer to the disconnected region locally. Some experimentally observed boron structures, like boron fullerenes, are not presented in the dataset. We calculate the local environment representations of every distinct

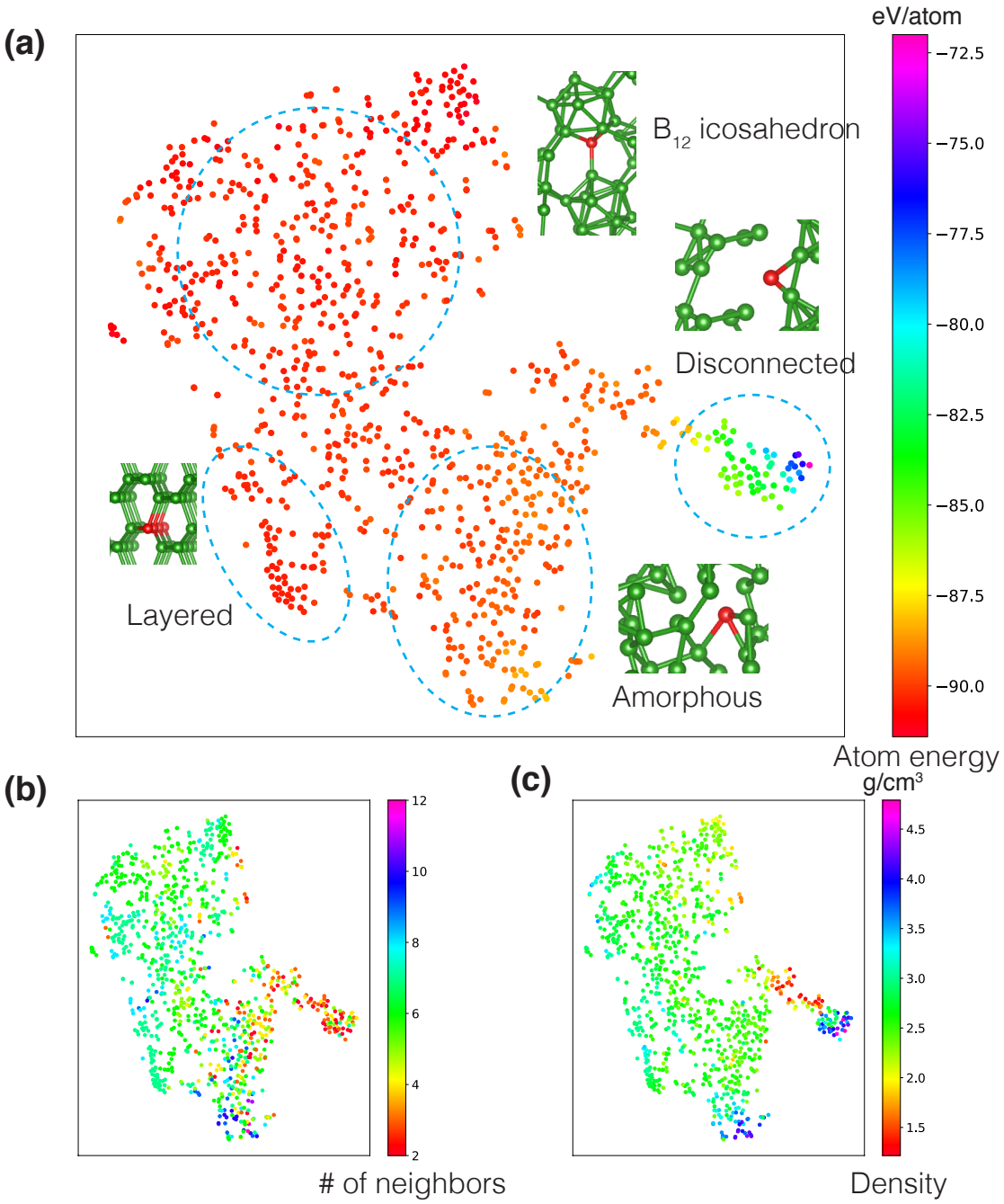


Figure 3-5: Visualization of the local environment representations learned from the elemental boron dataset. The original 64D vectors are reduced to 2D with the t-distributed stochastic neighbor embedding algorithm. The color of each plot is coded with learned local energy (a), number of neighbors calculated by Pymatgen package [192] (b), and density (c). Representative boron local environments are shown with the center atom colored in red.

boron atom of two boron fullerenes [194] using the trained CGCNN, and plot them into the original 2D visualization in Fig. 3-7. They form a small cluster close to the B_{12} icosahedron region. This can be explained by the fact that they share many common characteristics to the B_{12} icosahedron structure. In addition, the representations of the less symmetric $B_{40}(C_s)$ are more spread out than the more symmetric $B_{40}(D_{2d})$. Note that the pattern in Fig. 3-7 is slightly different from that in Fig. 3-5 due to the random nature of the t-SNE algorithm, but the overall structure of the patterns is preserved.

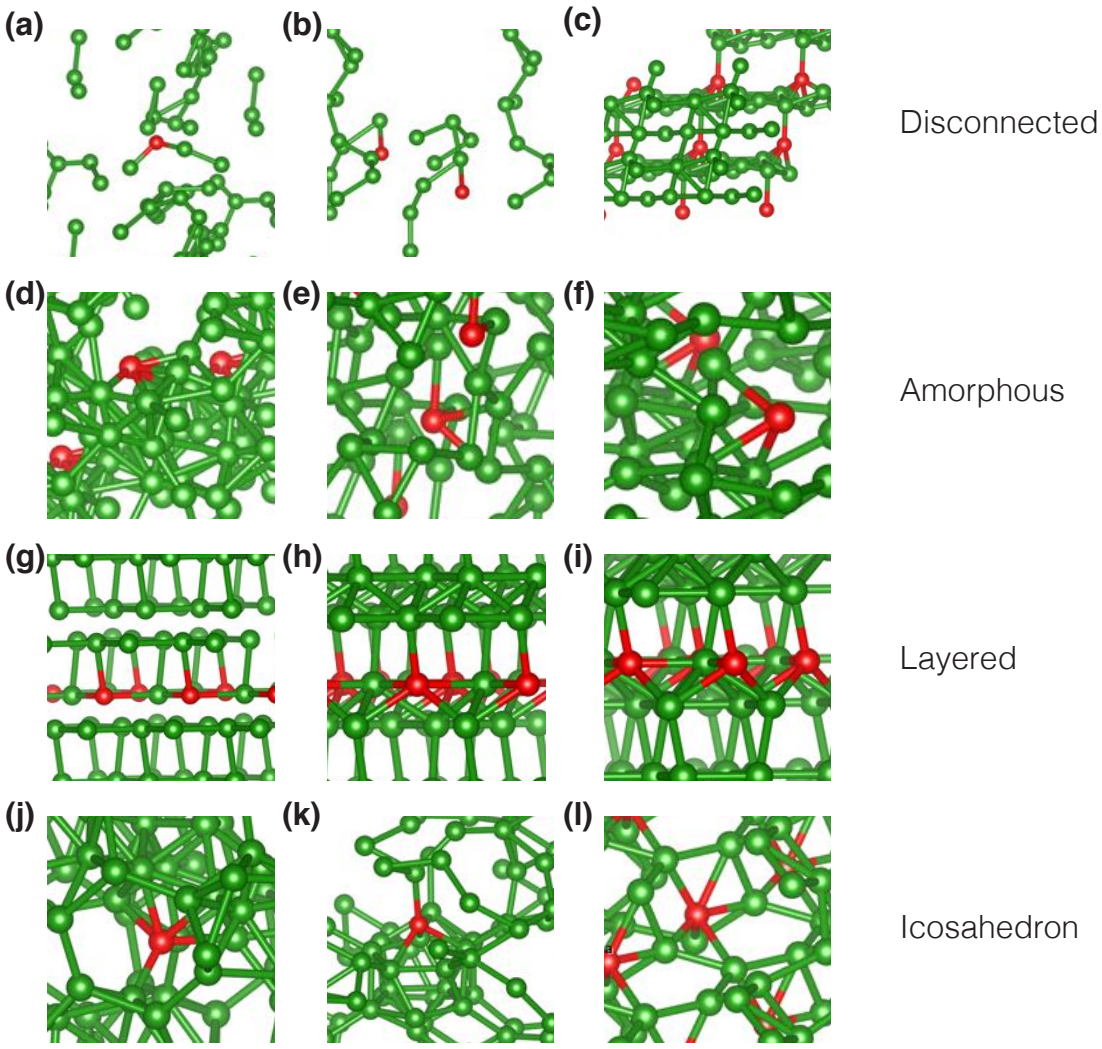


Figure 3-6: Example local environments of elemental boron in the four regions: (a-c) disconnected, (d-f) amorphous, (h-i) layered, and (j-l) icosahedron.

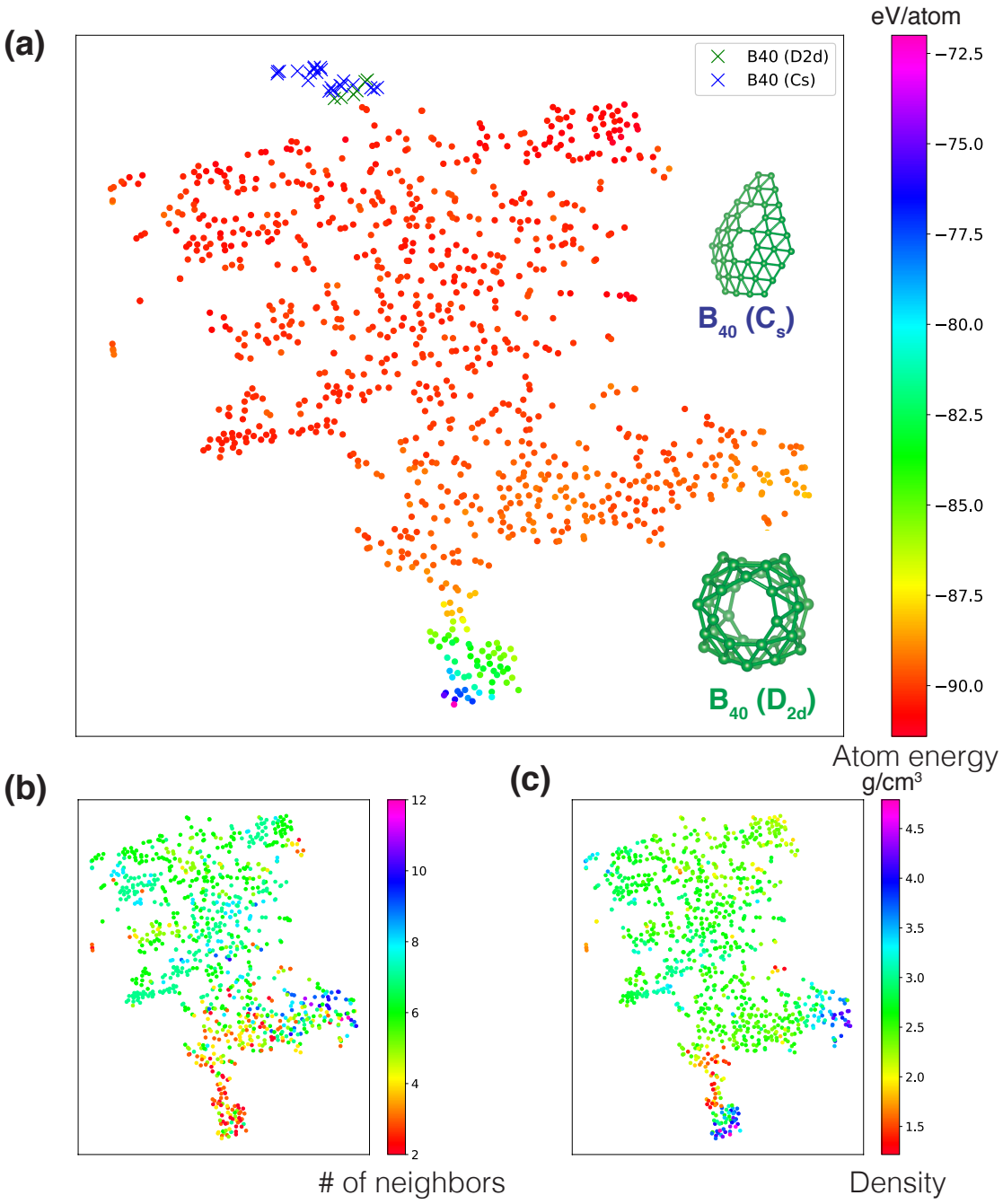


Figure 3-7: The boron fullerene local environments in the boron structural space. The representation of each distinct local environments in the two B_{40} structures are plotted in the original boron structural space in Fig. 4.

Taken together, such a visualization approach provides a convenient way to explore complex boron configurations, enabling the identification of characteristic structures and systematic exploration of structural space.

3.3.4 Materials Project: compositional and structural space

As a third example of applying this approach, we explore the material space of crystals in the Materials Project dataset [139], which includes both compositional and structural differences, by visualizing the *element representation*, *local environment representation*, and the *local energy representation*. The dataset includes 46744 materials that cover the majority of crystals from the Inorganic Crystal Structure Database [72], providing a good representation of known inorganic materials. After training with 28046 crystals and performing hyperparameter optimization with 9348 crystals, our model achieves MAE of predicted energy relative to DFT calculations on the 9348 test crystals of 0.042 eV/atom, slightly higher than the MAE of our previous work, 0.039 eV/atom, with a CGCNN structure focusing on prediction performance [46]. The learning curve in Fig. 3-2 is similar to that of the perovskites dataset, which might indicate a similar prediction performance to the datasets that are composed of stable inorganic compounds.

In Fig. 3-10, the element representation of 89 elements learned from the dataset is shown using the same method as that used to generate Fig. 3-3(b). We observe similar grouping of elements from the same elemental groups, but the overall pattern differs since it reflect the stability of each element in general inorganic crystals rather than perovskites. For instance, the non-metal and halogen elements stand out because their properties deviates from other metallic elements.

To illustrate how the compositional and structural spaces can be explored simultaneously, we visualize the oxygen and sulfur coordination environments in the Materials Project dataset using the local environment representation and local energy. 1000 oxygen and 803 sulfur coordination environments are randomly selected and visualized using the t-SNE algorithm. As shown in Fig. 3-9(a), the oxygen coordination environments are clustered into 4 major groups. The upper right group has the

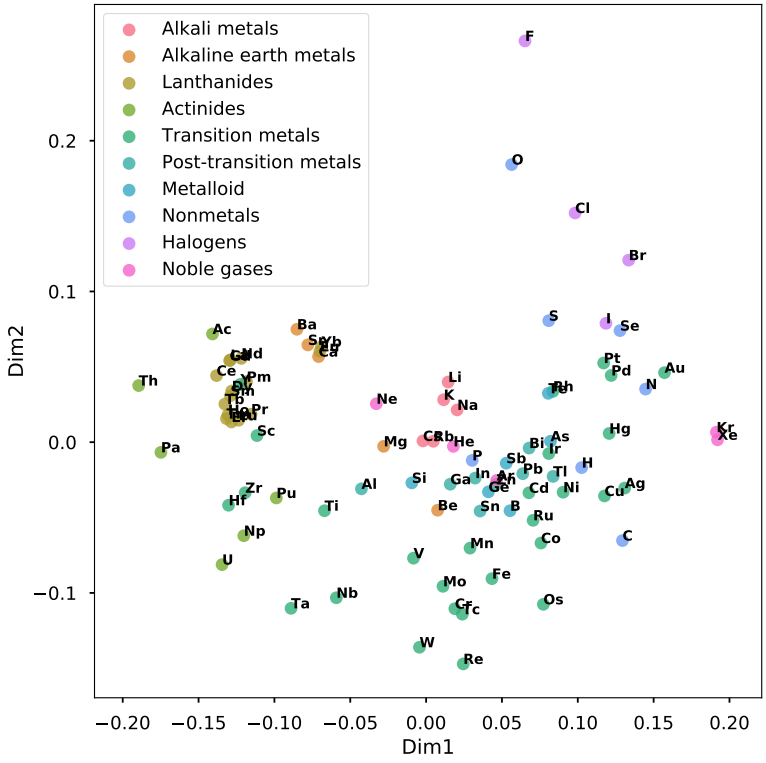


Figure 3-8: Visualization of the two principal dimensions of the element representations learned from the Materials Project dataset using principal component analysis.

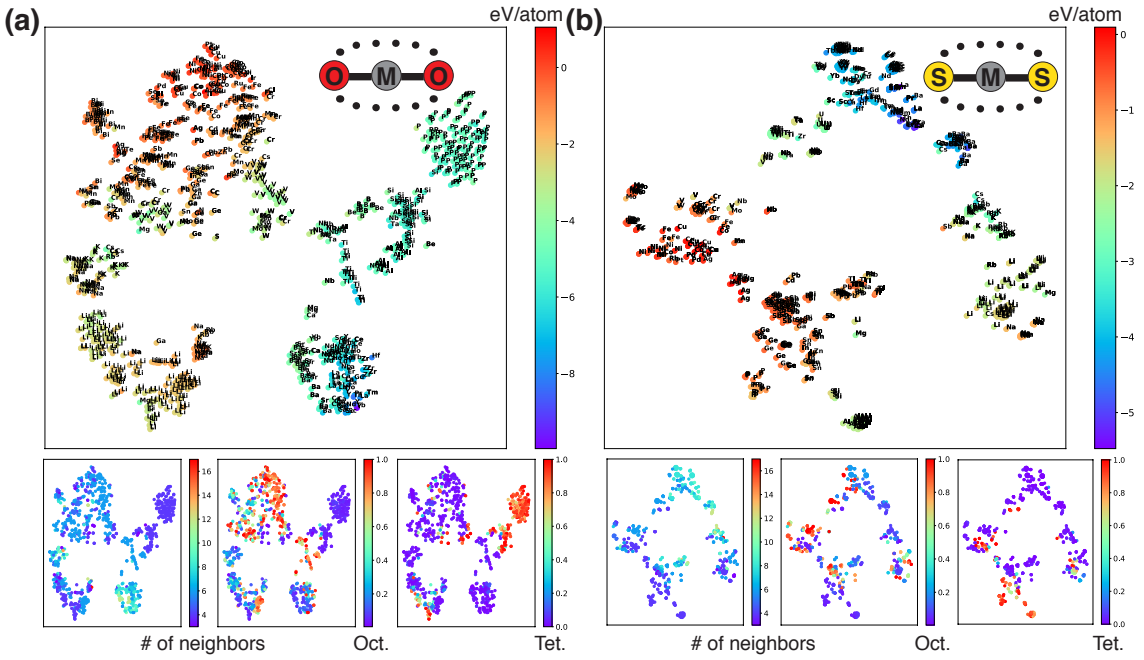


Figure 3-9: Visualization of the local oxygen (a) and sulfur (b) coordination environments. The points are labelled according to the type of the center atoms in the coordination environments. The colors of the upper parts are coded with learned local energies, and the color of the lower parts are coded with number of neighbors [192], octahedron order parameter, and tetrahedron order parameter [195].

center atom of non-metal elements like P, Al, Si, forming tetrahedron coordinations. The center atoms of the upper left environments are mostly transition metals, and they mostly form octahedron coordinations. The lower left group has center atoms of alkali metals, and the lower right group has those of alkaline earth metals and lanthanides which have larger radii and therefore higher coordination numbers. The sulfur coordination environment visualization [Fig. 3-9(b)] shares similar patterns due to the similarities between oxygen and sulfur, and a similar four-cluster structure can be observed. However, instead of non-metal elements, the lower center group has center atoms of metalloids like Ge, Sn, Sb, since these elements will be more stable in a sulfur vs. oxygen coordination environment.

The local energy of oxygen and sulfur coordination environments are determined by their relative stability to the pure elemental states since the model is trained using the formation energy data, which treats the pure elemental states as the reference energy states. In Fig. 3-10, we show the change of local energy of oxygen and sulfur local energies as a function of atomic number. We can clearly see that it follows a similar trend as the electronegativity of the elements: elements with lower electronegativity tend to have lower local energy and vice versa. This is because elements with lower electronegativity tends to give the oxygen and sulfur more electrons and thus form stronger bonds. The local energies of alkali metals are slightly higher since they form weaker ionic bonds due to lower charges. Interestingly, the strong covalent bonds between oxygen and Al, Si, P, S forms a V-shaped curve in the figure, with Si-O environments having the lowest energy, contrasting the trend of electronegativity and sulfur coordination environments, whose local energies are dominated by the strength of ionic bonds. We also observe a larger span of local energies in oxygen coordination environments than their sulfur counterparts due to the stronger ionic interactions.

Inspired by these results, we visualize the averaged local energy of 734,077 distinct coordination environments in the Materials Project by combining different center and neighbor atoms in Fig. 3-11. This figure illustrates the stability of the local coordination environment while combining the corresponding center and neighbor elements. The diagonal line represents coordination environments made up with

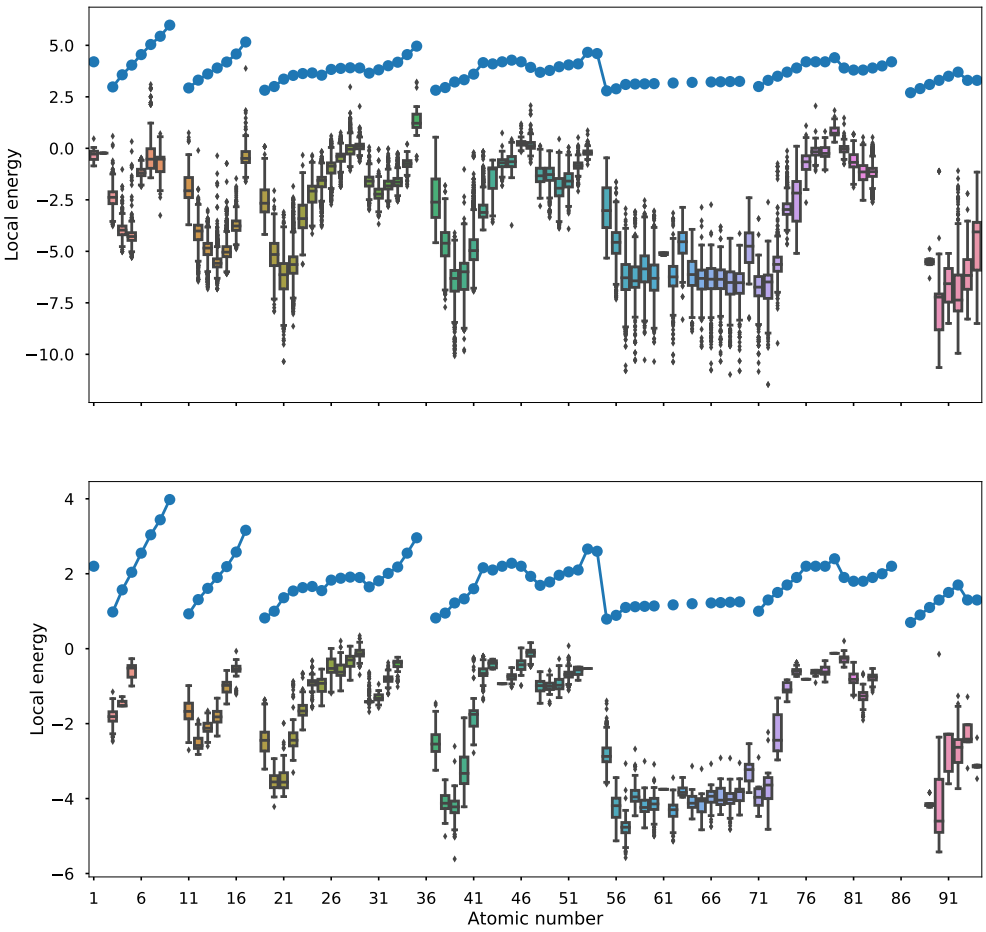


Figure 3-10: The local energy of oxygen (upper) and sulfur (lower) coordination environments as a function of atomic number. The blue dotted line denotes the electronegativity of each element.

the same elements with local energy close to zero, which corresponds to elemental substances with zero formation energy. The coordination environments with lowest local energy consist of high valence metals and high electronegativity non-metals, which can be explained by the large cohesive energies due to strong ionic bonds. One abnormality is the stable Al-O, Si-O, P-O, S-O coordination environments, although this can be attributed to their strong covalent bonds. We can also see that Tm-H coordination stands out as a stable hydrogen solid solution [196]. It is worth noting that each local energy in Fig. 3-11 is the average of many coordination environments with different shape and outer layer chemistry, and we can obtain more information by using additional visualizations similar to Fig. 3-9.

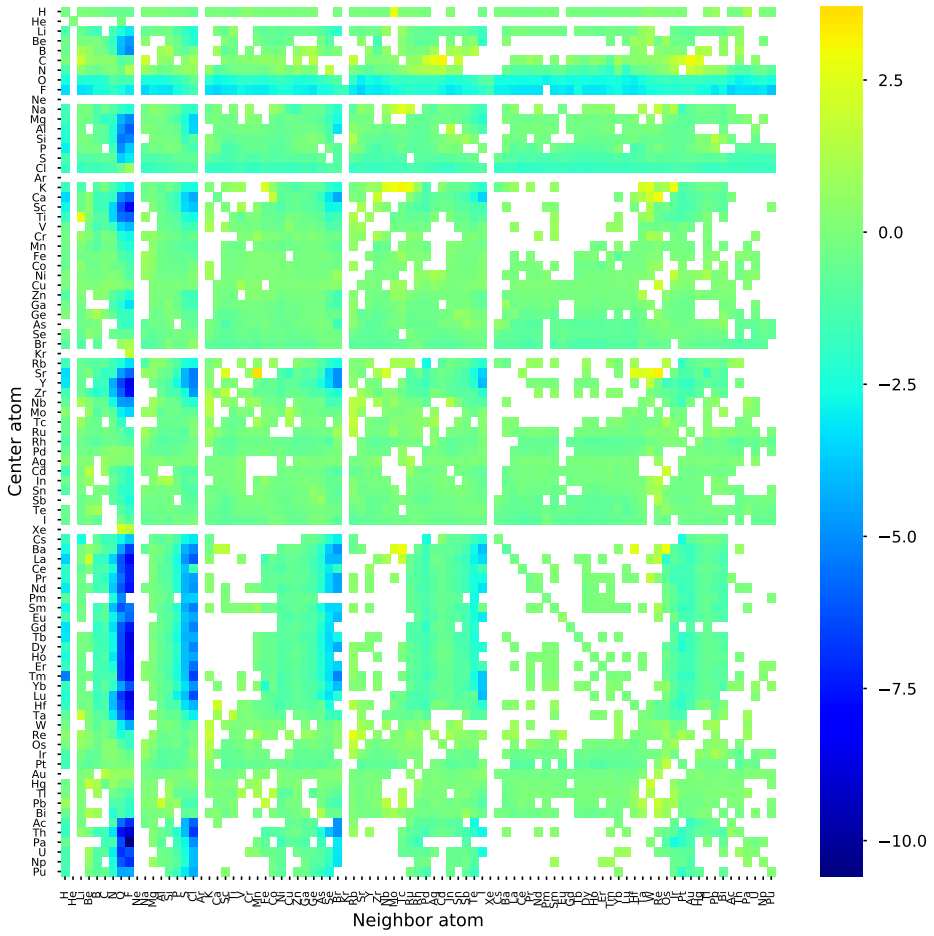


Figure 3-11: The averaged local energy of 734,077 distinct coordination environments in the Materials Project dataset. The color is coded with the average of learned local energies while having the corresponding elements as the center atom and the first neighbor atom. White is used when no such coordination environment exists in the dataset.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials

4.1 Introduction

4.1.1 Section overview

In this chapter, we aim to explore learning representations for solid materials when there is no property available, i.e. unsupervised learning. Obtaining the property of a solid material, either by computation or experiment, is expensive and in many cases, we only have the structure of a material. One of such cases is the time series data from molecular dynamics (MD) simulations, which describes how the structure of a material evolves as a function time. In this chapter, we develop a deep learning architecture, Graph Dynamical Networks (GDyNets), that combines Koopman analysis and graph convolutional neural networks to learn the dynamics of individual atoms in material systems from molecular dynamics trajectory data. We first describe the architecture of the GDyNets and how it extends the CGCNN to learn the dynamics of atoms in materials. Then, we will show through a toy system that the invariances built in graph neural networks allow the sharing of information between similar lo-

cal environments, which significantly improves the sampling of complex dynamics in materials. Finally, we apply our method to two realistic material systems – silicon dynamics at solid-liquid interfaces and lithium ion transport in amorphous polymer electrolytes – to demonstrate the new dynamical information one can extract for such complex materials and environments. Given the enormous amount of MD data generated in materials research, we believe the broad applicability of this method could help uncover important new physical insights from atomic scale dynamics that may have otherwise been overlooked.

4.1.2 Theoretical and practical motivations

From a theoretical perspective, we aim to extend the CGCNN framework to unsupervised learning by learning from time series data. The problem can be framed as a clustering problem, where we hope to assign each atom in a solid material a class by learning from time series data.

$$c_i = f(i, \boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{l}(t)), \tag{4.1}$$

where $c_i \in \{C_1, ..., C_K\}$ is the class label for atom i in the material, $\{\boldsymbol{x}(t), \boldsymbol{z}(t), \boldsymbol{l}(t)\}$ is the structure of the entire solid material at time step t. Note that we do not have the class labels c_i from the training data, and the model need to discover these classes by itself. This clustering problem provides a method to understand complex material systems, in which atoms experience various different local chemical environments that are hard to study with conventional approaches. The key question is what criteria we use to determine these class labels c_i, and we will solve this problem using a variational loss function that optimizes for the slowest transition in the system.

From a practical perspective, understanding the atomic scale dynamics in condensed phase is essential for the design of functional materials to tackle the global energy and environmental challenges. [197–199] The performance of many materials, like electrolytes and membranes, depend on the dynamics of individual atoms or small molecules in complex local environments. Despite the rapid advances in experimental

techniques [200–202], molecular dynamics (MD) simulations remain one of the few tools for probing these dynamical processes with both atomic scale time and spatial resolutions. However, due to the large amounts of data generated in each MD simulation, it is often challenging to extract statistically relevant dynamics for each atom especially in multi-component, amorphous material systems. At present, atomic scale dynamics are usually learned by designing system-specific description of coordination environments or computing the average behavior of atoms. [132, 203–205] A general approach for understanding the dynamics in different types of condensed phases, including solid, liquid, and amorphous, is still lacking.

4.1.3 Related prior research

The advances in applying deep learning to scientific research open new opportunities for utilizing the full trajectory data from MD simulations in an automated fashion. Ideally, we want to trace every atom or small molecule of interest in the MD trajectories, and summarize their dynamics into a linear, low dimensional model that describes how their local environments evolve over time. Recent studies show that combining Koopman analysis and deep neural networks provides a powerful tool to understand complex biological processes and fluid dynamics from data. [206–208] In particular, VAMPnets [208] develop a variational approach for Markov processes to learn an optimal latent space representation that encodes the long-time dynamics, which enables the end-to-end learning of a linear dynamical model directly from MD data. However, in order to learn the atomic dynamics in complex, multi-component material systems, sharing knowledge learned for similar local chemical environments is essential to reduce the amount of data needed. Recent development of graph convolutional neural networks (GCN) has led to a series of new representations of molecules [66, 174–176] and materials [46, 177] that are invariant to permutation and rotation operations. These representations provide a general approach to encode the chemical structures in neural networks which shares parameters between different local environments, and they have been used for predicting properties of molecules and materials [46, 66, 174–177], generating force fields [177, 209], and visualizing

structural similarities [48, 183].

4.2 Architecture of graph dynamical networks

4.2.1 Koopman analysis of atomic scale dynamics.

In materials design, the dynamics of target atoms, like the lithium ion in electrolytes and the water molecule in membranes, provide key information to material performance. We describe the dynamics of the target atoms and their surrounding atoms as a discrete process in MD simulations,

$$\boldsymbol{x}_{t+\tau} = \boldsymbol{F}(\boldsymbol{x}_t), \tag{4.2}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+\tau}$ denote the local configuration of the target atoms and their surrounding atoms at time steps t and $t+\tau$, respectively. Note that Eq. (4.2) implies that the dynamics of $\boldsymbol{x}$ is Markovian, i.e. $\boldsymbol{x}_{t+\tau}$ only depends on $\boldsymbol{x}_t$ not the configurations before it. This is exact when $\boldsymbol{x}$ includes all atoms in the system, but an approximation if only neighbor atoms are included. We also assume that each set of target atoms follow the same dynamics $\boldsymbol{F}$. These are valid assumptions since 1) most interactions in materials are short-range, 2) most materials are either periodic or have similar local structures, and we could test them by validating the dynamical models using new MD data which we will discuss later.

The Koopman theory [210] states that there exists a function $\boldsymbol{\chi}(\boldsymbol{x})$ that maps the local configuration of target atoms $\boldsymbol{x}$ into a lower dimensional feature space, such that the non-linear dynamics $\boldsymbol{F}$ can be approximated by a linear transition matrix $\boldsymbol{K}$,

$$\boldsymbol{\chi}(\boldsymbol{x}_{t+\tau}) \approx \boldsymbol{K}^T \boldsymbol{\chi}(\boldsymbol{x}_t). \tag{4.3}$$

The approximation becomes exact when the feature space has infinite dimensions. However, for most dynamics in material systems, it is possible to approximate it with a low dimensional feature space with a sufficiently large τ due to the existence of

characteristic slow processes. The goal is to identify such slow processes by finding the feature map function $\boldsymbol{\chi}(\boldsymbol{x})$.

4.2.2 Learning feature map function with graph dynamical networks.

In this work, we use graph convolutional neural networks (GCN) to learn the feature map function $\boldsymbol{\chi}(\boldsymbol{x})$. GCN provides a general framework to encode the structure of materials that is invariant to permutation, rotation, and reflection [46, 177]. As shown in Fig. 4-1, for each time step in the MD trajectory, a graph $\mathcal{G}$ is constructed based on its current configuration with each node $\boldsymbol{v}_i$ representing an atom and each edge $\boldsymbol{u}_{i,j}$ representing a bond connecting nearby atoms. We connect M nearest neighbors considering periodic boundary condition while constructing the graph, and a gated architecture [46] is used in GCN to reweigh the strength of each connection (see section 2.3 for details). Note that the graphs are constructed separately for each step, so the topology of each graph may be different. Also, the 3-dimensional information is preserved in the graphs since the bond length is encoded in $\boldsymbol{u}_{i,j}$. Then, each graph is input to the same GCN to learn an embedding for each atom through graph convolution (or neural message passing [175]) that incorporates the information of its surrounding environments.

$$\boldsymbol{v}_i' = \mathrm{Conv}(\boldsymbol{v}_i, \boldsymbol{v}_j, \boldsymbol{u}_{(i,j)}), \quad (i,j) \in \mathcal{G}. \tag{4.4}$$

After K convolution operations, information from the Kth neighbors will be propagated to each atom, resulting in an embedding $\boldsymbol{v}_i^{(K)}$ that encodes its local environment.

To learn a feature map function for the target atoms whose dynamics we want to model, we focus on the embeddings learned for these atoms. Assume that there are n sets of target atoms each made up with k atoms in the material system. For instance, in a system of 10 water molecules, $n = 10$ and $k = 3$. We use the label $\boldsymbol{v}_{[l,m]}$ to denote the mth atom in the lth set of target atoms. With a pooling function [46], we can get an overall embedding $\boldsymbol{v}_{[l]}$ for each set of target atoms to represent its local

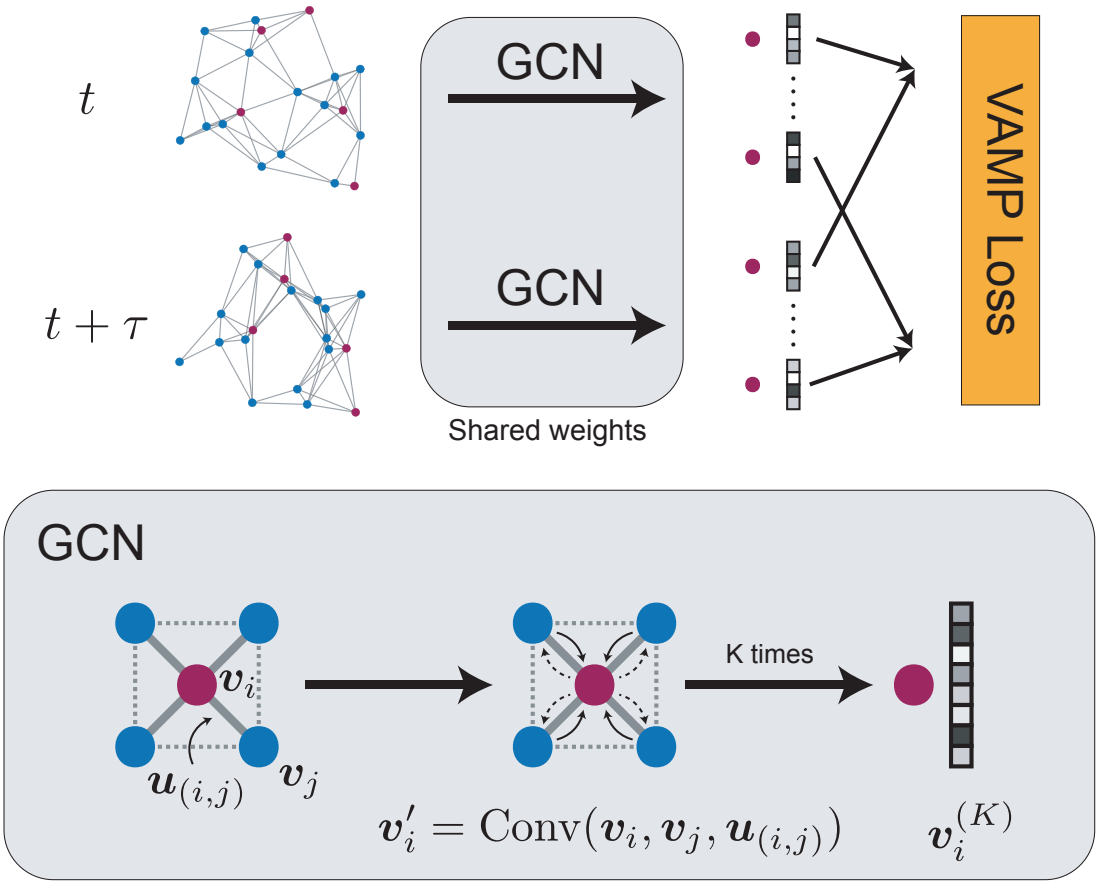


Figure 4-1: Illustration of the graph dynamical networks architecture. The MD trajectories are represented by a series of graphs dynamically constructed at each time step. The red nodes denote the target atoms whose dynamics we are interested in, and the blue nodes denote the rest of the atoms. The graphs are input to the same graph convolutional neural network to learn an embedding $\boldsymbol{v}_i^{(K)}$ for each atom that represents its local configuration. The embeddings of the target atoms at t and $t+\tau$ are merged to compute a VAMP loss that minimizes the errors in Eq. (4.3) [208, 211].

configuration,

$$\boldsymbol{v}_{[l]} = \mathrm{Pool}(\boldsymbol{v}_{[l,0]}, \boldsymbol{v}_{[l,1]}, \ldots, \boldsymbol{v}_{[l,k]}). \tag{4.5}$$

Finally, we build a shared output layer with a Softmax activation function to map the embeddings $\boldsymbol{v}_{[l]}$ to a feature space $\tilde{\boldsymbol{v}}_{[l]}$ with a pre-determined dimension. This is the feature space described in Eq. (4.3), and we can select an appropriate dimension to capture the important dynamics in the material system. The Softmax function used here allows us to interpret the feature space as a probability over several states [208]. Below, we will use the term "number of states" and "dimension of feature space" interchangeably.

To minimize the errors of the approximation in Eq. (4.3), we compute the loss of the system using a VAMP-2 score [208, 211] that measures the consistency between the feature vectors learned at timesteps t and $t + \tau$,

$$\mathrm{Loss} = -\mathrm{VAMP}(\tilde{\boldsymbol{v}}_{[l],t}, \tilde{\boldsymbol{v}}_{[l],t+\tau}), \quad t \in [0, T-\tau], l \in [0, n]. \tag{4.6}$$

This means that a single VAMP-2 score is computed over the whole trajectory and all sets of target atoms. The entire network is trained by minimizing the VAMP loss, i.e. maximizing the VAMP-2 score, with the trajectories from the MD simulations.

4.2.3 Hyperparameter optimization and model validation.

There are several hyperparameters in the GDyNets that need to be optimized, including the architecture of GCN, the dimension of the feature space, and lag time τ. We divide the MD trajectory into training, validation, and testing sets. The models are trained with trajectories from the training set, and a VAMP-2 score is computed with trajectories from the validation set. The GCN architecture is optimized according to the VAMP-2 score similar to ref. [46].

The accuracy of Eq. (4.3) can be evaluated with a Chapman-Kolmogorov (CK) equation,

$$\boldsymbol{K}(n\tau) = \boldsymbol{K}^n(\tau), \quad n = 1, 2, \ldots. \tag{4.7}$$

This equation holds if the dynamic model learned is Markovian, and it can predict the long-term dynamics of the system. In general, increasing the dimension of feature space makes the dynamic model more accurate, but it may result in overfitting when the dimension is very large. Since a higher feature space dimension and a larger τ make the model harder to understand and contain less dynamical details, we select the smallest feature space dimension and τ that fulfills the CK equation within statistical uncertainty. Therefore, the resulting model is interpretable and contains more dynamical details. More about the effects of feature space dimension and τ can be found in refs. [208, 211].

4.3 Advantage of learning local dynamics

The key advantage of GDyNets over the VAMPnet is that graph neural networks allow for the sharing of knowledge learned for similar local environments across the system, so it focuses on the modeling of local atomic dynamics instead of global dynamics. This significantly improves the sampling of the atomic dynamical processes because a typical material system includes a large number of atoms or small molecules moving in structurally similar but distinct local environments.

To demonstrate the advantage of learning local dynamics in material systems, we compare the dynamics learned by the GDyNet with VAMP loss and a standard VAMPnet with fully connected neural networks that learns global dynamics for a simple model system using the same input data. As shown in Fig. 4-2(a), we generated a 200 ns MD trajectory of lithium atom moving in a face-centered cubic (FCC) lattice of sulfur atoms at a constant temperature, which describes an important lithium ion transport mechanism in solid-state electrolytes [132]. There are two different sites for the lithium atom to occupy in a FCC lattice, tetrahedral sites and octahedral sites, and the hopping between the two sites should be the only dynamics in this system. As shown in Fig. 4-2(b-d), after training and validation with the first 100 ns trajectory, the GDyNet correctly identified the transition between the two sites with a relaxation timescale of 42.3 ps while testing on the second 100 ns trajectory, and it

performs well in the CK test. In contrast, the standard VAMPnet, which inputs the same data as the GDyNet, learns a global transition with a much longer relaxation timescale at 236 ps, and it performs much worse in the CK test. This is because the model views the four octahedral sites as different sites due to their different spatial location. As a result, the transition between these identical sites are learned as the slowest global dynamics.

It is theoretically possible to identify the faster local dynamics from a global dynamical model when we increase the dimension of feature space (Fig. 4-7). However, when the size of system increases, the number of slower global transitions will increase exponentially, making it practically impossible to discover important atomic scale dynamics within a reasonable simulation time. In addition, it is possible in this simple system to design a symmetrically invariant coordinate to include the equivalence of the octahedral and tetrahedral sites. But in a more complicated multi-component or amorphous material system, it is difficult to design such coordinates that take into account the complex atomic local environments. Finally, it is also possible to reconstruct global dynamics from the local dynamics. Since we know how the 4 octahedral and 8 tetrahedral sites are connected in a FCC lattice, we can construct the 12 dimensional global transition matrix from the 2 dimensional local transition matrix (see section 4.6.1 for details). We obtain the slowest global relaxation timescale to be 531 ps, which is close to the observed slowest timescale of 528 ps from the global dynamical model in Fig. 4-7. Note that the timescale from the two-state global model in Fig. 4-2 is less accurate since it fails to learn the correct transition. In sum, the built-in invariances in GCN provides a good tool to reduce the complexity of learning atomic dynamics in material systems.

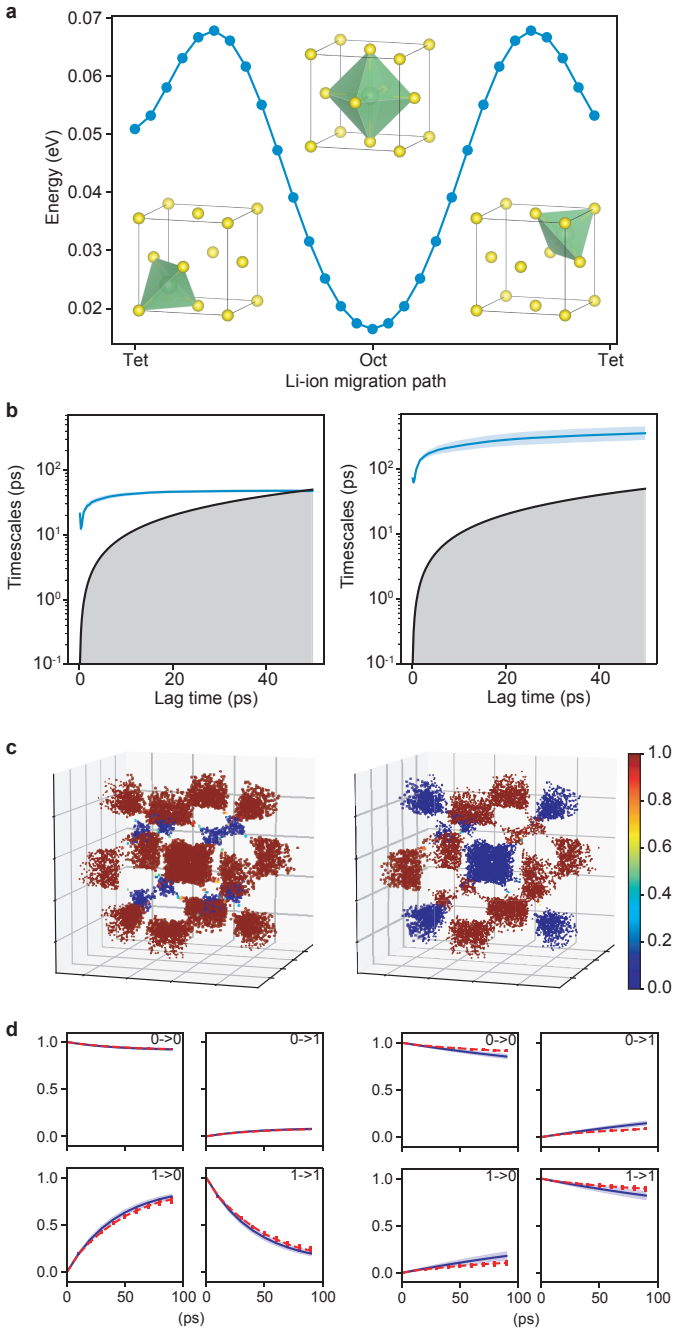


Figure 4-2: A two-state dynamic model learned for lithium ion in the face-centered cubic lattice. (a) Structure of the FCC lattice and the relative energies of the tetrahedral and octahedral sites. (b-d) Comparison between the local dynamics (left) learned with GDyNet and the global dynamics (right) learned with a standard VAMPnet. (b) Relaxation timescales computed from the Koopman models. (c) Assignment of the two states in the FCC lattice. The color denotes the probability of being in state 0. (d) CK test comparing the long-term dynamics predicted by Koopman models at $\tau = 10\,\text{ps}$ (blue) and actual dynamics (red). The shaded areas and error bars in (b, d) report the 95% confidence interval from five independent trajectories by dividing the test data equally into chunks.

4.4 Application to the understanding of complex dynamics

4.4.1 Silicon dynamics in solid-liquid interface

To evaluate performance of the GDyNets with VAMP loss for a more complicated system, we study the dynamics of silicon atoms at a binary solid-liquid interface. Understanding the dynamics at interfaces is notoriously difficult due to the complex local structures formed during phase transitions. [212, 213] As shown in Fig. 4-3(a), an equilibrium system made of two crystalline Si {110} surfaces and a liquid Si-Au solution is constructed at the eutectic point (629 K, 23.4% Si [214]) and simulated for 25 ns using MD. We train and validate a four-state model using the first 12.5 ns trajectory, and use it to identify the dynamics of Si atoms in the last 12.5 ns trajectory. Note that we only use the Si atoms in the liquid phase and the first two layers of {110} surfaces as the target atoms (Fig. 4-3(b)). This is because the Koopman models are optimized for finding the slowest transition in the system, and including additional solid Si atoms will result in a model that learns the slower Si hopping in the solid phase which is not our focus.

In Fig. 4-3(b, c), the model identified four states that are crucial for the Si dynamics at the solid-liquid interface – liquid Si at the interface (state 0), solid Si (state 1), solid Si at the interface (state 2), and liquid Si (state 3). These states provide a more detailed description of the solid-liquid interface structure than conventional methods. In Fig. 4-4, we compare our results with the distribution of the q_3 order parameter of the Si atoms in the system, which measures how much a site deviates from a diamond-like structure and is often used for studying Si interfaces [215]. We learn from the comparison that 1) our method successfully identifies the bulk liquid and solid states, and learns additional interface states that cannot be obtained from q_3; 2) the states learned by our method are more robust due to access to dynamical information, while q_3 can be affected by the accidental order structures in the liquid phase; 3) q_3 is system specific and only works for diamond-like structures, but the

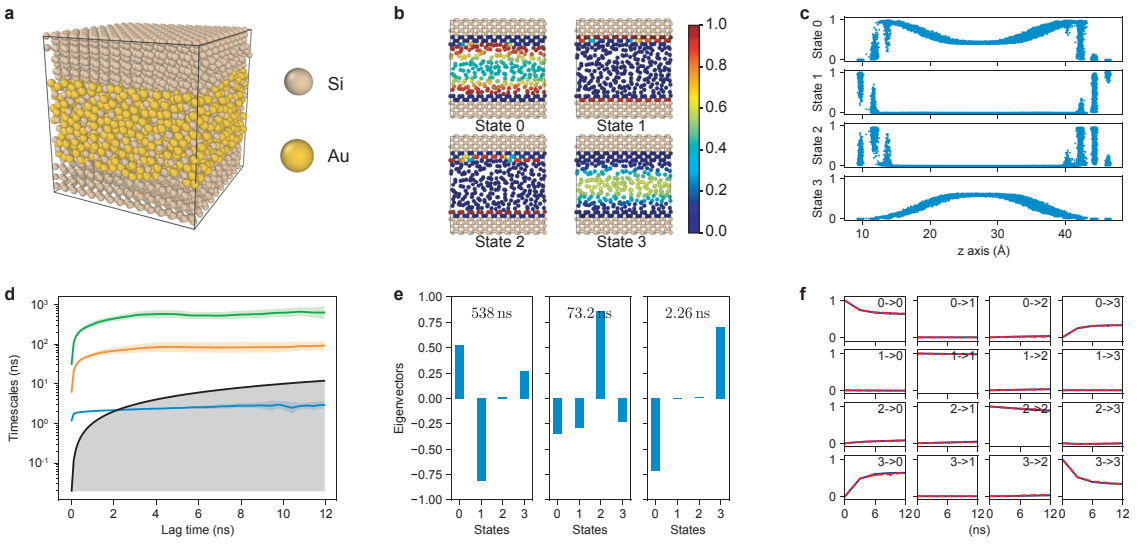


Figure 4-3: A four-state dynamical model learned for silicon atoms at solid-liquid interface. (a) Structure of the silicon-gold two-phase system. (b) Cross section of the system, where only silicon atoms are shown and color-coded with the probability of being in each state. (c) The distribution of silicon atoms in each state as a function of z-axis coordinate. (d) Relaxation timescales computed from the Koopman models. (e) Eigenvectors projected to each state for the three relaxations of Koopman models at $\tau = 3\,\text{ns}$. (f) CK test comparing the long-term dynamics predicted by Koopman models at $\tau = 3\,\text{ns}$ (blue) and actual dynamics (red). The shaded areas and error bars in (d, f) report the 95% confidence interval from five sets of Si atoms by randomly dividing the target atoms in the test data.

GDyNets can potentially be applied to any material given the MD data.

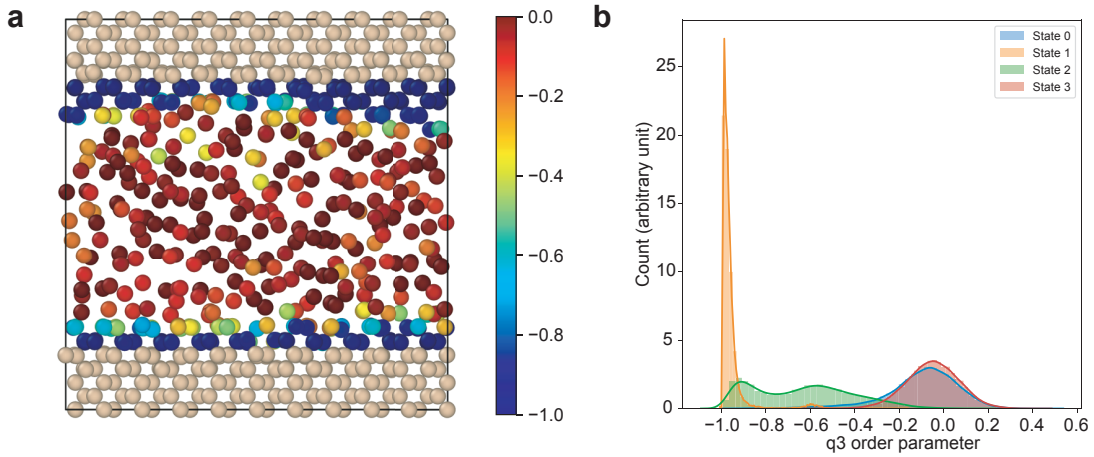


Figure 4-4: Comparison between the learned states and q_3 order parameters for silicon atoms at the solid-liquid interface. (a) Cross section of the system, where the silicon atoms are color-coded with their q_3 order parameters. (b) Distribution of the q_3 order parameter for the silicon atoms of each state.

In addition, important dynamical processes at the solid-liquid interface can be learned with the model. Remarkably, the model identified the relaxation process of the solid-liquid transition with a timescale of 538 ns (Fig. 4-3(d, e)), which is one order of magnitude longer than the simulation time of 12.5 ns. This is because the large number of Si atoms in the material system provide an ensemble of independent trajectories that enable the identification of rare events [216–218]. The other two relaxation processes corresponds to the transitions of solid Si atoms at the interface (73.2 ns) and liquid Si atoms at interface (2.26 ns), respectively. These processes are difficult to obtain with conventional methods due to the complex structures at solid-liquid interfaces, and the results are consistent with our understanding that the former solid relaxation is significantly slower than the latter liquid relaxation. Finally, the model performs excellently in the CK test on predicting the long-term dynamics.

4.4.2 Lithium ion dynamics in polymer electrolytes

Finally, we apply GDyNets with VAMP loss to study the dynamics of lithium ions (Li-ions) in solid polymer electrolytes (SPEs), an amorphous material system composed of

multiple chemical species. SPEs are candidates for next generation battery technology due to their safety, stability, and low manufacturing cost, but they suffer from low Li-ion conductivity compared with liquid electrolytes. [219, 220] Understanding the key dynamics that affect the transport of Li-ions is important to the improvement of Li-ion conductivity in SPEs.

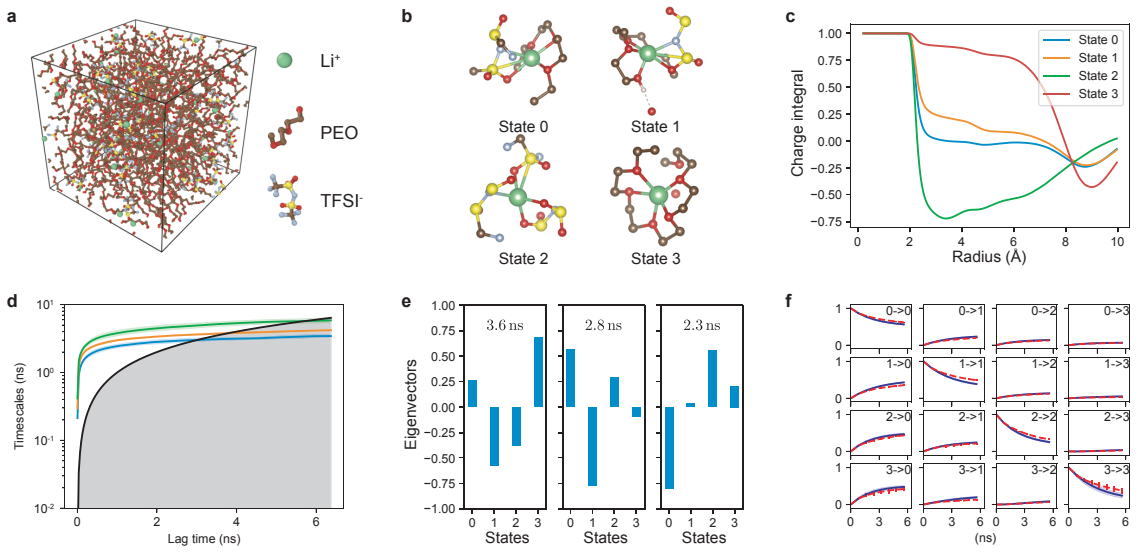


Figure 4-5: A four-state dynamical model learned for lithium ion in a PEO/LiTFSI polymer electrolyte. (a) Structure of the PEO/LiTFSI polymer electrolyte. (b) Representative configurations of the four Li-ion states learned by the dynamical model. (c) Charge integral of each state around a Li-ion as a function of radius. (d) Relaxation timescales computed from the Koopman models. (e) Eigenvectors projected to each state for the three relaxations of Koopman models at $\tau = 0.8\,\text{ns}$. (f) CK test comparing the long-term dynamics predicted by Koopman models at $\tau = 0.8\,\text{ns}$ (blue) and actual dynamics (red). The shaded areas and error bars in (d, f) report the 95% confidence interval from four independent trajectories in the test data.

We focus on the state-of-the-art [220] SPE system – a mixture of poly(ethylene oxide) (PEO) and lithium bis-trifluoromethyl sulfonimide (LiTFSI) with Li/EO = 0.05 as shown in Fig. 4-5(a). Five independent 80 ns trajectories are generated to model the Li-ion transport at 363 K. We train a four-state GDyNet with one of the trajectories, and use the model to identify the dynamics of Li-ions in the remaining four trajectories. The model identified four different solvation environments, i.e. states, for the Li-ions in the SPE. In Fig. 4-5(b), the state 0 Li-ion has a population of $50.6 \pm 0.8\%$, and it is coordinated by a PEO chain on one side and a TFSI anion

on the other side. The state 1 has a similar structure as state 0 with a population of $27.3 \pm 0.4\%$, but the Li-ion is coordinated by a hydroxyl group on the PEO side rather than an oxygen. In state 2, the Li-ion is completely coordinated by TFSI anion ions, which has a population of $15.1 \pm 0.4\%$. And the state 3 Li-ion is coordinated by PEO chains with a population of $7.0 \pm 0.9\%$. Note that the structures in Fig. 4-5(b) only show a representative configuration for each state. We compute the element-wise radial distribution function (RDF) for each state to demonstrate the average configurations, which is consistent with above description. We also analyze the total charge carried by the Li-ions in each state considering their solvation environments in Fig. 4-5(c) (Table 4.1). Interestingly, both state 0 and state 1 carry almost zero total charge in their first solvation shell due to the one TFSI anion in their solvation environments.

Table 4.1: The charge carried by each state in PEO/LiTFSI.

State	0	1	2	3
Charge	+0.040	+0.262	-0.637	+0.889

We further study the transition between the four Li-ion states. Three relaxation processes are identified in the dynamical model as shown in Fig. 4-5(d, e). By analyzing the eigenvectors, we learn that the slowest relaxation is a process involving the transport of a Li-ion into and out of a PEO coordinated environment. The second slowest relaxation happens mainly between state 0 and state 1, corresponding to a movement of the hydroxyl group. The transitions from state 0 to states 2 and 3 constitute the last relaxation process, as state 0 can be thought of an intermediate state between state 2 and state 3. The model performs well in CK tests (Fig. 4-5(f)). Relaxation processes in the PEO/LiTFSI systems have been extensively studied experimentally [221, 222], but it is difficult to pinpoint the exact atomic scale dynamics related to these relaxations. The dynamical model learned by GDyNet provides additional insights into the understanding of Li-ion transport in polymer electrolytes.

4.4.3 Implications to lithium ion conduction

The state configurations and dynamical model allow us to further quantify the transitions that are responsible for the Li-ion conduction. In Fig. 4-6, we compute the contribution from each state transition to the Li-ion conduction using the Koopman model at $\tau = 0.8\,\text{ns}$. First, we learn that the majority of conduction results from transitions within the same states ($i \to i$). This is because the transport of Li-ions in PEO is strongly coupled with segmental motion of the polymer chains [203, 223], in contrast to the hopping mechanism in inorganic solid electrolytes [224]. In addition, due to the low charge carried by state 0 and state 1, the majority of charge conduction results from the diffusion of states 2 and 3, despite their relatively low populations. Interestingly, the diffusion of state 2, a negatively charged species, accounts for $\sim 40\%$ of the Li-ion conduction. This provides an atomic scale explanation to the recently observed negative transference number at high salt concentration PEO/LiTFSI system [225].

4.5 Discussion

We have developed a general approach, Graph Dynamical Networks (GDyNets), to understand the atomic scale dynamics in material systems. Despite being widely used in biophysics [218], fluid dynamics [226], and kinetic modeling of chemical reactions [227–229], Koopman models, (or Markov state models [218], master equation methods [230, 231]) have not been used in learning atomic scale dynamics in materials from MD simulations except for a few examples in understanding solvent dynamics [232–234]. Our approach also differs from several other unsupervised learning methods [235–237] by directly learning a linear Koopman model from MD data. Many crucial processes that affect the performance of materials involve the local dynamics of atoms or small molecules, like the dynamics of lithium ions in battery electrolytes [238, 239], the transport of water and salt ions in water desalination membranes [240, 241], the adsorption of gas molecules in metal organic frameworks [242, 243], among many other examples. With the improvement of computational power

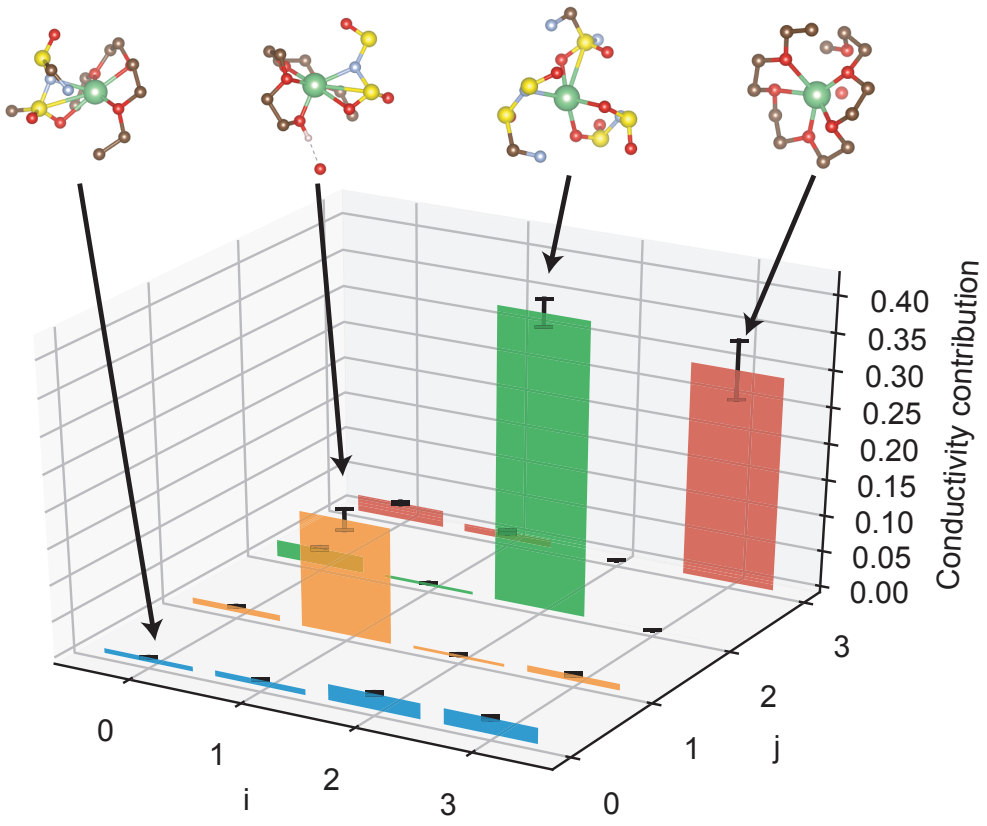


Figure 4-6: Contribution from each transition to lithium ion conduction. Each bar denotes the percentage that the transition from state i to state j contributes to the overall lithium ion conduction. The error bars report the 95% confidence interval from four independent trajectories in test data.

and continued increase in the use of molecular dynamics to study materials, this work could have broad applicability as a general framework for understanding the atomic scale dynamics from MD trajectory data.

Compared with the Koopman models previously used in biophysics and fluid dynamics, the introduction of graph convolutional neural networks enables parameter sharing between the atoms and an encoding of local environments that is invariant to permutation, rotation, and reflection. This symmetry facilitates the identification of similar local environments throughout the materials, which allows the learning of local dynamics instead of exponentially more complicated global dynamics. In addition, it is easy to extend this method to learn global dynamics with a global pooling function [46]. However, a hierarchical pooling function is potentially needed to directly learn the global dynamics of large biological systems including thousands of atoms. It is also possible to represent the local environments using other symmetry functions like smooth overlap of atomic positions (SOAP) [161], social permutation invariant (SPRINT) coordinates [159], etc. By adding a few layers of neural networks, a similar architecture can be designed to learn the local dynamics of atoms. However, these built-in invariances may also cause the Koopman model to ignore dynamics between symmetrically equivalent structures which might be important to the material performance. One simple example is the flip of ammonia molecule – the two states are mirror symmetric to each other so the GCN will not be able to differentiate them by design. This can potentially be resolved by partially break the symmetry of GCN based on the symmetry of the material systems.

The graph dynamical networks can be further improved by incorporating ideas from both the fields of Koopman models and graph neural networks. For instance, the auto-encoder architecture [207, 244, 245] and deep generative models [246] start to enable the direct generation of future structures in the configuration space. Our method currently lacks a generative component, but this can potentially be achieved with a proper graph decoder [247, 248]. Furthermore, transfer learning on graph embeddings may reduce the number of MD trajectories needed for learning the dynamics [249, 250].

In summary, graph dynamical networks present a general approach for understanding the atomic scale dynamics in materials. With a toy system of lithium ion transporting in a face-centered cubic lattice, we demonstrate that learning local dynamics of atoms can be exponentially easier than global dynamics in material systems with representative local structures. The dynamics learned from two more complicated systems, solid-liquid interfaces and solid polymer electrolytes, indicate the potential of applying the method to a wide range of material systems and understanding atomic dynamics that are crucial to their performances.

4.6 Supplementary notes

4.6.1 Computation of global dynamics from local dynamics in the toy system

To compute the global dynamics from local dynamics, we first assume that the transition matrix of the local Koopman model has the form,

$$\boldsymbol{K}_{\text{local}} = \begin{bmatrix} p_o & 1-p_o \\ 1-p_t & p_t \end{bmatrix}, \tag{4.8}$$

where p_o and p_t denotes the probability of the lithium atom staying in the octahedral and tetrahedral sites, respectively. Since there are 4 octahedral sites and 8 tetrahedral sites that are connected to each other in the FCC lattice, we can write the transition

matrix of the global Koopman model as,

$$\boldsymbol{K}_{\text{global}} = \begin{bmatrix} p_o & 0 & 0 & 0 & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} \\ 0 & p_o & 0 & 0 & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} \\ 0 & 0 & p_o & 0 & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} \\ 0 & 0 & 0 & p_o & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} & \frac{1-p_o}{8} \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & p_t & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & p_t & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & p_t & 0 & 0 & 0 & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & 0 & p_t & 0 & 0 & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & 0 & 0 & p_t & 0 & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & 0 & 0 & 0 & p_t & 0 & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & 0 & 0 & 0 & 0 & p_t & 0 \\ \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & \frac{1-p_t}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_t \end{bmatrix}. \tag{4.9}$$

By computing the eigenvalues of $\boldsymbol{K}_{\text{global}}$, we could obtain the relaxation timescales and understand the global dynamics of the toy system. There are two major reasons for the discrepancy between the computed and observed global Koopman model (Fig. 4-7): 1) the amount of MD data is not large enough to capture of full global dynamics of the lithium atom by directly learning a global dynamical model; 2) the probability of lithium atom transporting to nearby sites of the same type is not strictly zero at a given τ, so the p_o and p_t in $\boldsymbol{K}_{\text{local}}$ and the zero terms in $\boldsymbol{K}_{\text{global}}$ are approximate.

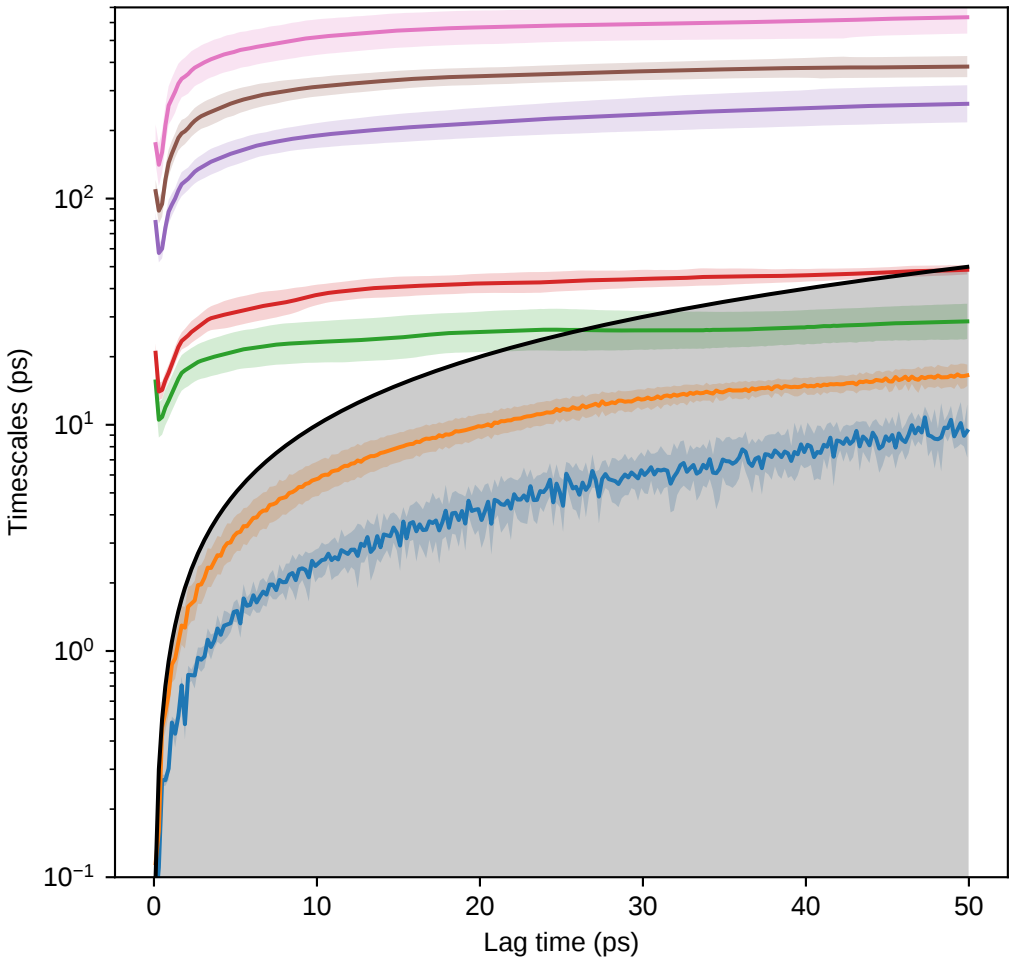


Figure 4-7: Global relaxation timescales computed for lithium ion hopping in face-centered cubic (FCC) lattice with a 8 dimensional feature space.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Autonomous exploration of the space of polymer electrolytes with Bayesian optimization and coarse-grained molecular dynamics

5.1 Introduction

5.1.1 Chapter overview

In this chapter, we explore active learning to autonomously explore a material space combining Bayesian optimization and coarse-grained molecular dynamics simulations. This is a different learning problem compared with both supervised and unsupervised learning. We aim not only to predict the property of a given material, but also propose material to simulate in a next iteration step in a way the minimizes the total number of simulation needed. We will start by introducing the material design problem – the discovery of solid polymer electrolyte materials for lithium ion batteries. Then, we will define the target material space and develop the Bayesian optimization approach to efficient explore this space. Further, we will demonstrate the performance of our approach and discuss the material design knowledge learned from the exploration.

Finally, we discuss the implications of our models for materials design.

5.1.2 Motivations

Lithium-ion batteries haven been applied in a wide range of applications, from personal devices to grid scale energy storage [251]. In pursuit of safer and more durable lithium-ion batteries at lower cost, solid polymer electrolytes (SPEs) are promising building blocks, due to their unique advantages such as absence of flammable solvents, compatibility with roll-to-roll processes, and intrinsic flexibility and stretchability [252, 253]. However, the low ionic conductivity of current SPEs prevents their further incorporation into real-world applications [254]. This challenge motivates tremendous research efforts towards the design of highly conductive SPE materials [204, 252, 255, 256], via investigating the ionic transport mechanisms [257, 258] and exploring candidate SPE materials [259, 260] (where simulations and modeling have already made valuable contributions [203, 261]). With the recent rapid development of artificial intelligence (AI) [7], machine learning (ML) techniques have started to play roles in improving and reforming the design loop of SPE materials [47, 259, 260]. AI and ML developments provide immense opportunities to examine molecular moieties in polymer electrolytes and correlate their dynamics and energetics with ion transport properties. In order to fully understand what governs the ion mobility and achieve global optimization of the SPE system, the identification of universally applicable descriptors is of great importance, but this is very difficult in a typical design space constructed by chemical species. Besides, in the case of SPEs, the complexity of the system, which mixes long chain polymers and lithium compounds, places it beyond the capacity of conventional fully atomistic (FA) simulations as a means to generate datasets with the sizes suitable for many popular ML algorithms (e.g. a training dataset containing 10^4-10^5 samples is usually preferred [46]).

5.2 Coarse Grained Molecular Dynamics-Bayesian Optimization framework

In this work, we propose a new framework for design of SPE materials that combines coarse grained molecular dynamics (CGMD) with Bayesian Optimization (BO). In addition to a great reduction in the computational cost, the CG simulation also preserves molecular level information, converting the discrete chemical species space to a continuous space constructed by the CGMD parameters. The adoption of the BO algorithm enables efficient exploration of this high dimensional design space. From this, we can predict the relationships between the associated molecular level material properties (e.g. molecule sizes and intermolecular interactions) and the electrolyte performance, to gain useful mechanistic insights and optimize SPE functions.

To train the CGMD-BO model, we first construct a design space using a set of CGMD parameters, including the molecule sizes and intermolecular interaction strengths, that completely define the properties of our designated "improvable components" in a SPE system; specifically the anions, backbone chains and possible secondary sites (e.g. by introducing chemical variations in PEO chains[262]). In this CG space, we set the starting point of our exploration at the parameters that represent the lithium bis(trifluoromethanesulfonyl)imide-poly(ethlyene oxide) (PEO-LiTFSI) system, considering that the PEO-LiTFSI exhibits the highest conductivity among the SPE candidates that have been extensively studied [256, 263]. Once the CG space is constructed, we then aim to optimize the lithium ionic conductivity σ_{Li^+}, by an iterative parallel BO training process. The learned CGMD-BO model provides a detailed description of the relationships between the σ_{Li^+} and the CGMD parameters, from which we propose the directions and principles for changing TFSI^-, introducing secondary sites and replacing PEO backbone chains.

Figure 5-1 demonstrates the concept of the SPE materials design pathway via BO guided CGMD simulations. Conventionally, computation guided materials design starts with the proposal of a set of chemical species. These serve as the inputs to fully atomistic (FA) simulations to obtain a detailed and accurate description of

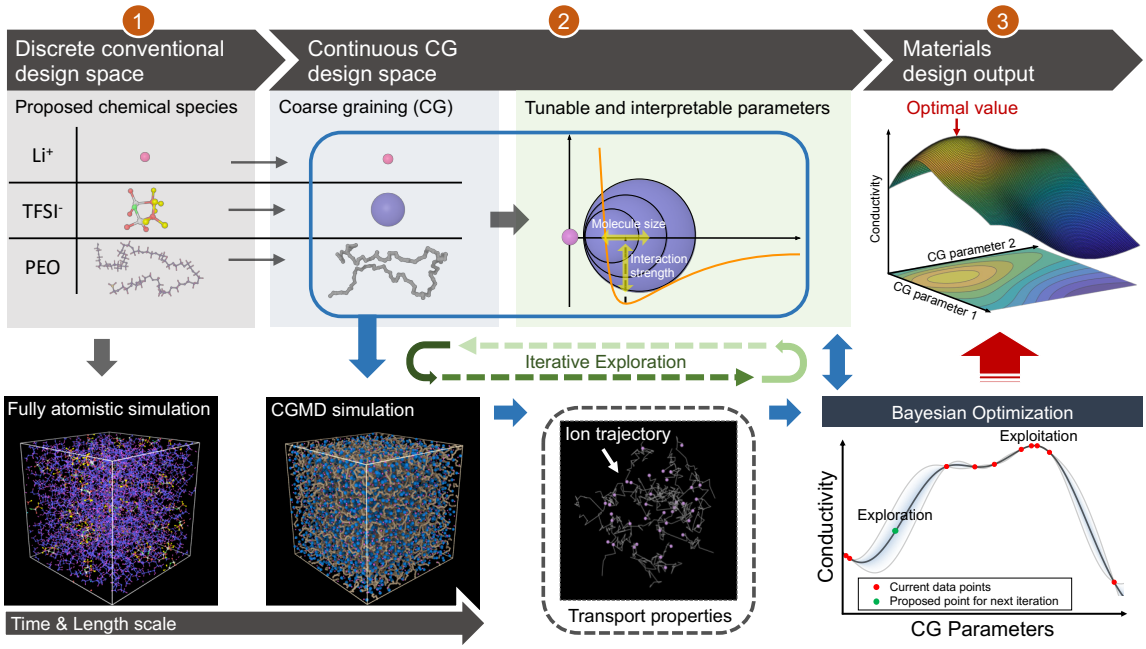


Figure 5-1: Illustration of the Coarse Grained Molecular Dynamics-Bayesian Optimization framework. Schematics of the polymer electrolyte materials design pathway by Bayesian Optimization (BO) guided coarse grained molecular dynamics (CGMD) simulation. Materials design starts with the coarse graining process, to transform the conventional chemical species space to a continuous space composed of CG parameters (①→②). This space is then explored by BO guided CGMD simulations in iterations, to predict the relationships between the transport properties and the associated CG parameters (②→③).

the system. However, considering the time and length scale limitations of the FA models (for example, a typical 10 ns classical MD run for only one PEO/LiTFSI system with 20,000 atoms requires 1,500 CPU hours on an Intel Xeon Gold core), we propose an alternative coarse graining (CG) process that abstracts the polymer chains and the anion molecules with a bead-spring representation [264]. Compared to the FA model, the CG configuration maintains most of the capability to capture the polymer conformation, while using fewer particles in the simulation cell to reduce the computational cost. Through the CG process, molecular level information, such as molecular size and intermolecular interaction strength, become CGMD parameters. Calibration of the CG system by the FA model provides the values of these parameters for a desired electrolyte system (e.g. PEO-LiTFSI), and also accomplishes the transformation from a discrete conventional design space to a continuous CG design space.

The CGMD simulation defines a function f that maps from the continuous CG design space to the performance of the SPE material. Given a set of input parameters, the CGMD simulation generates the trajectory of the particles, from which the transport properties such as the ionic conductivity and the transference number can be extracted to compute performance metrics for the corresponding SPE material. We can thus reformulate our goal as to find the set of input parameters, i.e. the SPE material, that maximizes the performance metrics.

As illustrated in Figure 5-1, to efficiently explore the continuous CG design space and maximize the performance of SPE materials, we design a Bayesian Optimization (BO) approach that utilizes the information from past simulations iteratively. In each iteration, we compute the a posterior estimation of the target function $p(f|\mathcal{D}_i)$ using the current simulation data $\mathcal{D}_i$, and propose the next points in the CG design space as inputs to CGMD by balancing exploration and exploitation. By the end of this process, the model outputs a posterior estimation of the objective function $p(f|\mathcal{D})$ from which the optimal lithium conductivity and its dependences on all the input CGMD parameters can be extracted. Compared with existing works that use BO in materials design [51, 265, 266], our approach uses several key characteristics of the

system to improve the efficiency further: 1) adopting the local penalization algorithm [267], multiple points are proposed in each iteration to better utilize the parallel computation of modern super computers; 2) the continuity of f and the intrinsic noise of CGMD simulations are built into the BO model to provide a more reliable estimation of f.

5.3 Exploration of the polymer electrolyte space

5.3.1 Defining three search spaces

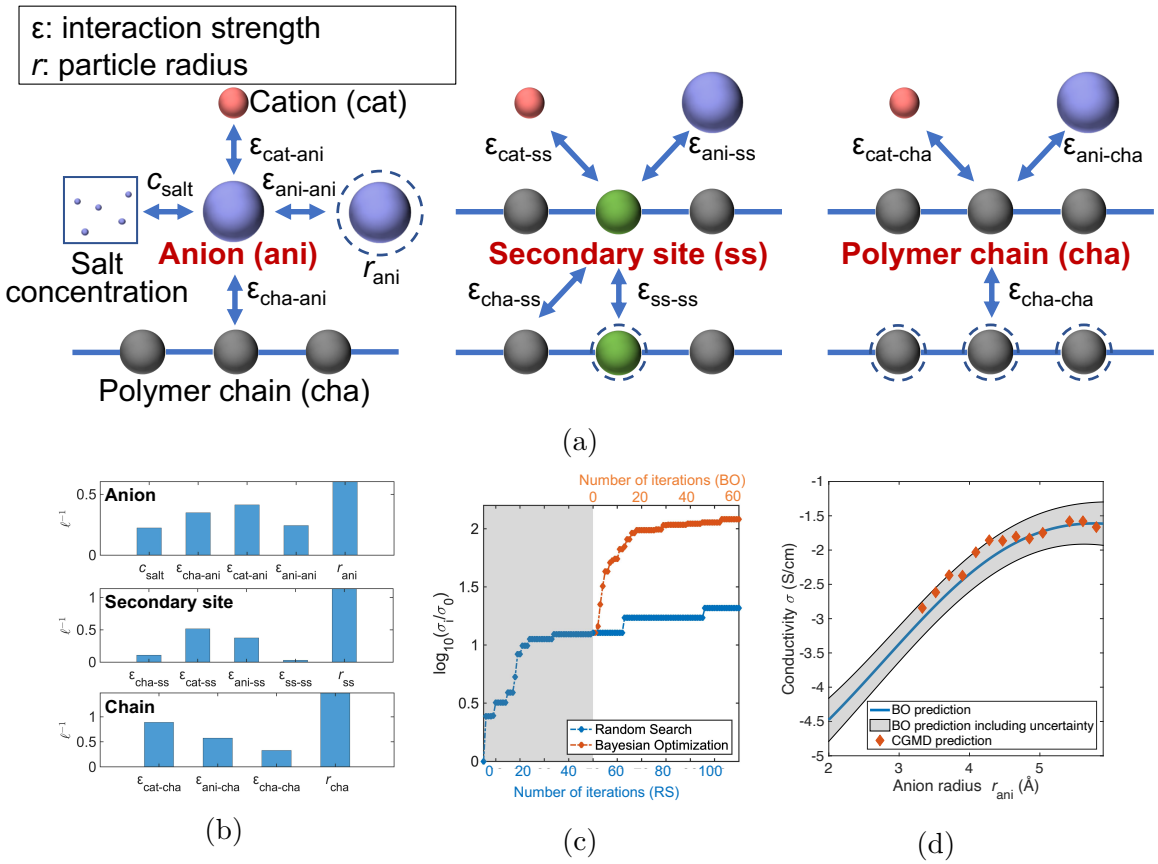


Figure 5-2: Evaluation of the Bayesian Optimization training process. (a) Illustration of the CGMD parameters, which are divided into three groups for describing the properties associated with the anions, secondary sites and backbone chains respectively (from left to right), (b) the inverse of characteristic length scale for each CGMD parameter in the BO training process, (c) the design space exploration efficiency of BO in comparison with random search, and (d) the BO predicted conductivities in comparison with the CGMD test data.

In the practice of the above design method, as demonstrated in Figure 5-2a, we divide all the designated independent CGMD parameters into three categories: five related to the properties of the anions (anion size, salt concentration and anion involved vdW interaction strengths), five related to the properties of the secondary sites introduced in the polymer chain (molecule size and secondary site involved non-bonding interaction strengths), and four to determine the properties of the polymer chain itself (monomer size and polymer involved non-bonding interaction strengths). This division naturally sets up three exploration directions in the CG design space, corresponding to modifications to the anions, the secondary sites, and the polymer backbone chains respectively. The search range of these parameters is determined based on their values of the reference PEO-LiTFSI system, with the lower and upper bounds capped by their physical interpretations, e.g. typically the vdW interaction strength $\epsilon_{ij} \in (0.4, 6)$ Kcal/mol and the particle radius $r_i \in (1.5, 5)$ Å. Specifically, we set the target property of the materials optimization to the lithium conductivity σ_{Li^+}, which is defined as the product of the overall conductivity σ and the lithium transference number t_{Li^+}. This setup enables our exploration to not only maximize σ, but also put emphasis on t_{Li^+}, considering that increasing t_{Li^+} could effectively reduce the polarization and improve the stability of the electrolyte system in the charge-discharge cycles [225, 258, 268].

5.3.2 Performance of the exploration

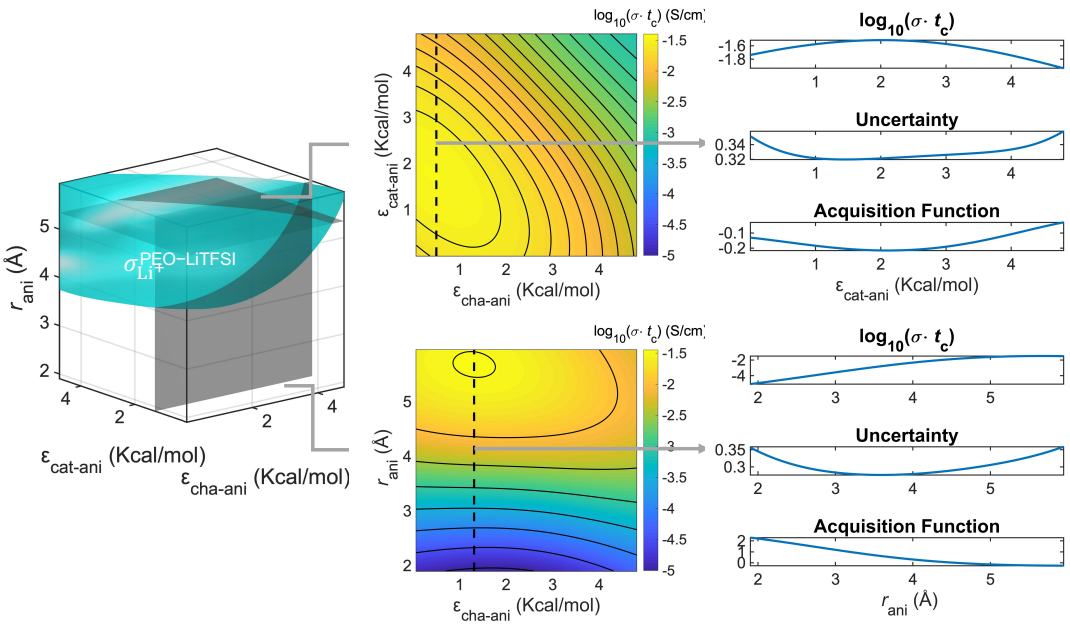
Figure 5-2b shows the inverse of characteristic length scale ℓ^{-1} (a hyper-parameter in the BO model) of each CG parameter, which can be considered as a measure of the parameter importance in the BO training process. A larger ℓ^{-1} indicates that the change of this parameter will be likely to make more impact on σ_{Li^+}. (ℓ^{-1} only reflects the average effect of each CG parameter on σ_{Li^+} for the entire design space we explore. Thus, it is possible that the parameter with the highest ℓ^{-1} may not be the most influential factor for search in some subspaces.) In all three explorations, the size of the particle was found to be the most influential factor, while the interaction strength associated with the cation ranked 2nd. In contrast, the role of the inter-

chain interaction was not crucial, possibly because its strength was small compared to other interactions involving charged particles.

To compare the searching efficiency of the BO method with a random search (RS), Figure 5-2c plots the normalized best-so-far (BSF) conductivity as a function of the iteration number, where for RS the BSF value plateaus after around 50 random explorations. In contrast, given a RS-generated initial data set with the size of 50, the BO improves the BSF values more efficiently, and converges to a much higher conductivity within only around 60 iterations, i.e. to reach the same BSF value, the RS method will take a much longer time. This result indicates that the CGMD-BO method is an efficient approach for SPE materials optimization. To validate the model, a test data set was built, containing the conductivities calculated from a series of CGMD simulations performed at different anion sizes (with the other parameters kept the same as the reference PEO-LiTFSI system). As shown in Figure 5-2d, both the values and the trend presented by these test data were well reproduced by the trained BO model. It should be noted that the agreement was achieved under the condition that the BO model used only around 100 sampling points to search this high dimensional CG space, and none of these training data were close to the test data in the design space.

5.3.3 Understanding the effects of structural modification

The CGMD-BO model was adopt to investigate the consequences of possible modifications to the anions on the lithium conductivity σ_{Li^+}. The obtained relation between σ_{Li^+} and the most influential three anion-related parameters ($\varepsilon_{\text{cat-ani}}$, $\varepsilon_{\text{cha-ani}}$ and r_{ani}, referring to Figure 5-2b) is described in Figure 5-3a, in the form of an isosurface plot at the σ_{Li^+} value of the reference PEO-LiTFSI system ($\sim 10^{-3}$ S/cm [269]). In Figure 5-3b, at a fixed anion radius, a 2D landscape is drawn, to describe σ_{Li^+} as a function of $\varepsilon_{\text{cat-ani}}$ and $\varepsilon_{\text{cha-ani}}$. In general, while a moderate value of $\varepsilon_{\text{cha-ani}}$ was necessary for dissolving the anions, if the $\varepsilon_{\text{cha-ani}}$ was too large, it would lead to the reduction of σ_{Li^+}, which could result from the increased population of polymer cross linking through the anions. Also, the optimal value of $\varepsilon_{\text{cha-ani}}$ was found



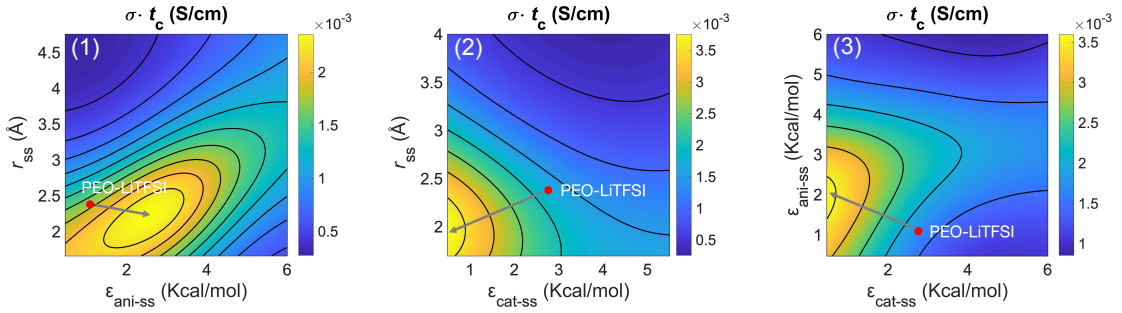
(a) (b) (c)

Figure 5-3: Anion effects on lithium conductivity. (a) 3D isosurface plot at the lithium conductivity value of PEO-LiTFSI, (b) 2D σ_{Li^+} landscape projected in $\varepsilon_{cat\text{-}ani}$-$\varepsilon_{cha\text{-}ani}$ and r_{ani}-$\varepsilon_{cha\text{-}ani}$ planes, and (c) 1D cross sectional plots showing the dependence of σ_{Li^+} on $\varepsilon_{cat\text{-}ani}$ and r_{ani} respectively, with the uncertainty evaluations and the acquisition function values.

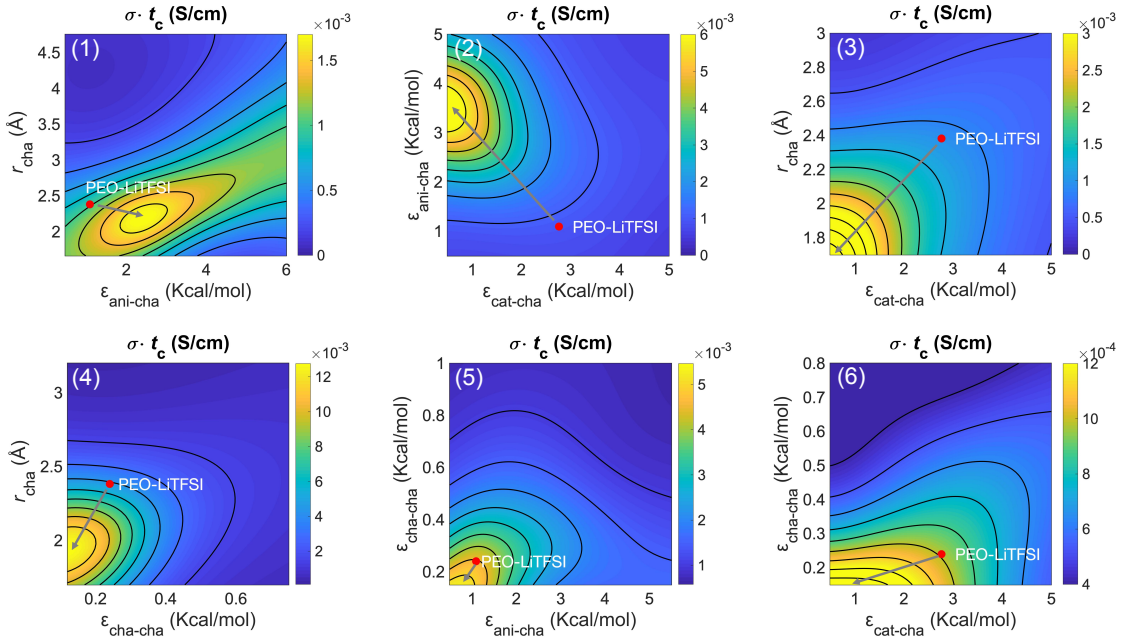
to be positively correlated with $\varepsilon_{\text{cat-ani}}$, namely, decreasing $\varepsilon_{\text{cha-ani}}$ and $\varepsilon_{\text{cat-ani}}$ together would be beneficial to the lithium conductivity, where an optimum value of $\varepsilon_{\text{cat-ani}}$ was needed to maximize σ_{Li^+}. At this $\varepsilon_{\text{cat-ani}}$, the repulsive contribution between the cation and the anion may be balanced with their attractive Coulombic interaction, effectively lowering the ion dissociation energy in polymer (Figure S1). When we project σ_{Li^+} in the r_{ani}-$\varepsilon_{\text{cha-ani}}$ plane, a significant increase of σ_{Li^+} can be observed by increasing r_{ani} in most of the BO searching area, mainly due to the decaying Coulombic interaction at larger charge separation distance. This is consistent with the fact that in many cases a larger anion could enhance the delocalization of the negative charge [270, 271]. (In our current CG model, the size increase of the anion inherently implies an effective higher degree of charge delocalization, however, it should be noted that this may not always be true in reality.) In opposition to the above positive relation between σ_{Li^+} and r_{ani}, larger anion volume also introduces a more severe obstacle that suppresses the system diffusion. Therefore, on the σ_{Li^+} landscape there exists an optimal value of r_{ani}, which slightly decreases with stronger $\varepsilon_{\text{cha-ani}}$.

We also examined the σ_{Li^+} dependence on a single factor by cross sectioning the 2D landscape. For example as shown in Figure 5-3c, we provide the curves representing the σ_{Li^+}-$\varepsilon_{\text{cat-ani}}$ and σ_{Li^+}-r_{ani} relations respectively, which further supports the above analysis. Figure 5-3c also gives the estimates of uncertainty (in the form of standard deviation), in conjunction with the BO acquisition function, whose value is proportional to the probability of the region that will be evaluated in the next iteration. Overall, the uncertainty fluctuates within a small range, thus, the minimum of the acquisition function (AF_{min}) occurs near the CGMD parameter value that tends to yield high σ_{Li^+}, leading to further exploitation of this region. On the other hand, at the beginning of the BO training, the AF_{min} may appear at the position where the uncertainty is relatively high, to enforce the exploration of the entire parameter space.

The CGMD-BO method was also adopted to investigate the effects of introducing secondary sites (SS) to PEO chains on the lithium conductivity. For simplicity, setting



(a)



(b)

Figure 5-4: Effects of secondary sites and polymer backbone chains on lithium conductivity. A series of 2D σ_{Li^+} landscape plots for the materials exploration of (a) secondary sites, and (b) polymer backbone chains. Each subfigure shows the dependence of σ_{Li^+} on a pair of CGMD parameters, with the other parameters fixed at the values of the reference PEO-LiTFSI system. The red dots on the graphs denote the reference PEO-LiTFSI system, with the arrows pointing out the directions to maximize σ_{Li^+}.

the reference system to PEO-LiTFSI, we tuned the properties of SS, with keeping the SS to EO ratio at 1/5. The model predictions were collected, as a set of 2D σ_{Li^+} landscapes in Figure 5-4a. In each of the landscape plot, the lithium conductivities were contoured on a plane determined by a pair of CGMD parameters, with the red dot marking the position of the unmodified PEO-LiTFSI system. From Figure 5-4a, one can conclude that a promising SS is expected to have a size smaller but close to the EO monomer, and a selective and modestly strong interaction with the anion. These properties enhance the cation diffusion and immobilize the anions. This is in opposition to the properties of PEO, but in accordance with the rationale for the proposal of single lithium-ion conducting polymers in the literature. [270, 271]

We noted that the improvement of σ_{Li^+} due to the introduction of SS, in comparison with unmodified PEO-LiTFSI, was rather limited. This inspired us to probe the possibility of designing non-PEO based SPE materials, that is actually an active research direction where several novel non-PEO based polymer architectures have been proposed and investigated experimentally. [272–276] Figure 5-4b presents the change of σ_{Li^+} induced by varying any two CGMD parameters away from the PEO reference. Based on Figure 5-4b, favorable polymer candidates tend to shrink their sizes and weaken their inter-chain interaction (e.g. Figure 5-4b(4)), presumably to achieve high diffusivity and flexibility. In addition, as here we aim to maximize the conductivity contributed by the lithium-ion, our CGMD-BO model consistently suggests the design direction of increasing $\varepsilon_{\text{ani-cha}}$ together with decreasing $\varepsilon_{\text{cat-cha}}$, to create more free Li^+ in the SPE system [277]. Even so, one may consider the negative effects of increasing $\varepsilon_{\text{ani-cha}}$ on chain diffusivity, that explains why there exists an optimal $\varepsilon_{\text{ani-cha}}$ (as shown in Figure 5-4b(1) and (2)).

5.4 Discussion

The trained CGMD-BO model can be utilized as a rich SPE materials database. To obtain the transport properties of one specific SPE system, the model only requires a CG parameterization of the molecule species, through a set of straightforward energy

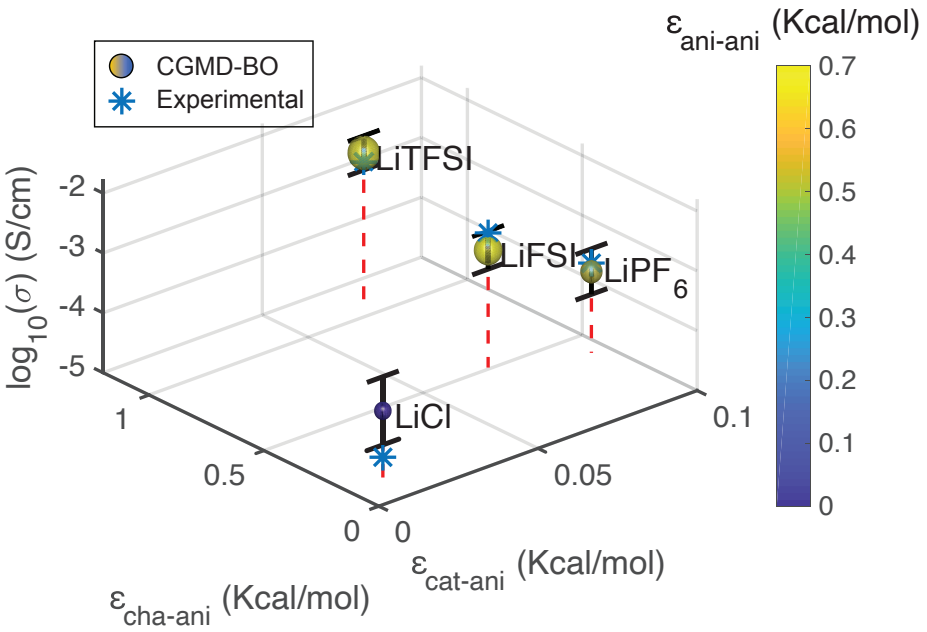


Figure 5-5: CGMD-BO predictions on conductivity for several common electrolyte systems. The trained BO model predicts conductivities for the PEO-LiTFSI, PEO-LiFSI, PEO-$LiPF_6$ and PEO-LiCl systems, which are plotted with the uncertainty information (shown as error bars) and their corresponding CG parameters, in comparison with experimental measurements (represented by asterisks)[263, 269, 278, 279].

evaluations by either all-atom force fields or DFT. In comparison with classical fully atomistic approaches that require days or even months to obtain conductivity values, the CGMD-BO model reduces the time of the process down to minutes. For instance, without performing any additional simulations, as shown in Figure 5-5, the trained BO model predicts and compares the conductivities of four common electrolyte systems (the input parameters to the CGMD-BO model are shown on the plot and listed in Table S1) to reasonable agreement with the experimental measurements[263, 269, 278, 279]. This suggests the potential of the CGMD-BO model as a convenient tool for rapid screening of the candidate materials prior to synthesis. Besides, Figure 5-5 shows that rather than being determined by a single factor, the change of conductivity is more likely to be a joint effect of all the molecular properties on this plot (e.g. anion size and the anion associated intermolecular interactions). This again implies the complexity of the SPE materials design space, which is almost impossible to be explored without the CGMD-BO approach.

In this work, using the molecular-level material properties as descriptors, the CGMD-BO framework has shown its unique advantages of efficiency and flexibility, in the optimization of σ_{Li^+}. Actually, with minor modifications, the model can be extended to adopt additional descriptors from microstructural features (e.g. the Li-ion solvation-site connectivity [275]) to macroscopic material properties (e.g. polymer stiffness and glass transition temperature), to discover the correlations among the descriptors at different scales and their joint effects on the lithium conductivity. Broadly speaking, we expect the CGMD-BO framework to be a promising approach to understanding the collective effects of the molecular descriptors on a wide range of properties of a given system (not limited to polymer-salt mixture), for design of complex multi-component material systems. So far, all the CG parameters are independently adjustable, enabling us to reach every corner of the design space. In reality, correlations usually exist among the parameters, confining the exploration to one or several subspaces of our current CG design space. In principle, these constraints could be better understood with data from the CG parameterization of more SPE materials. Besides, the process of the CG parameterization and the CG model itself could be

more refined, to further improve the prediction accuracy of the current CGMD-BO model. (For example, the accuracy of the CG model could be improved by calibrating its parameters to FA simulations with a polarizable force field [280]. Taking the information of cation solvation-site structures and distribution from FA trajectories, a dynamic bond percolation model could be adopted, to further accelerate the CG simulations [281].) Last, we anticipate that the CGMD-BO model can go beyond its current capability to make further contributions to new materials design. By taking advantage of machine learning to understand the structure-property relationships [46], it progressively becomes achievable to recognize and decode the similarities between the micro-structural features of coarse grained and fully atomistic models[124]. We believe that the joint efforts from more advanced ML algorithms, more accurate CG models and sufficient training data should ultimately achieve the recovery of the atomistic details from molecular level information, that will grant the CGMD-BO method the ability to suggest chemical species directly.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Conclusion and outlook

6.1 Summary of the thesis

In summary, the main methodological contribution of this thesis is the development of a unified neural network framework to solve multiple material design and understanding problems for solid materials. The key distinction of this approach and previous approaches is its generality: the same framework can be easily adapted to learn different types of materials, like inorganic materials in section 2.5 and polymers in section 4.4.2, as well as solve different types of problems, like property prediction in section 2.3 and learning dynamical processes in section 4.2. This generality is also a feature in many other types of deep learning models used for computer vision and natural language processing, which has led to large models like BERT [282] that achieves state-of-the-art performances in a wide range of tasks. As the size of open material data continue to increase, we hope this thesis would open the possibility of utilizing material data from multiple sources and create general models to help us design and understand materials better in various fields.

Another key feature of this framework is the inductive biases, or symmetries, that are built into the neural network. The clear correspondences between nodes/edges and atoms/bonds combined with the symmetries built into network through parameter sharing improves both the performance of model in various tasks and its interpretability. In chapter 2, the performance of CGCNN for supervised learning is

significantly improved by including the permutation invariance and periodicity in the models. In chapter 3, the visualization of the similarities between solid materials at different scales is achieved because each layer of the neural network has clear physical meanings, a result of the inductive biases in the model. In chapter 4, the learning of atomic scale dynamics is not possible without the sharing the knowledge between similar local environments in the materials, which is achieved through the parameter sharing in the neural networks. There are clearly other types of inductive biases that can potentially be included in the neural networks, like the long-range Coulomb interactions and electronic structures. We believe including the correct amount of inductive biases in the model while making the rest of the network flexible is a key towards machine learning models with both performance and interpretability.

From an application perspective, we demonstrate several examples on how this framework can be used to accelerate material design. In section 2.5, we applied CGCNN to predict the interface stability of solid electrolytes for lithium metal battery. In this example, we train our model on 3,400 open materials data and then apply it to screen 12,950 solid materials. This greatly accelerates the screening of new materials since we save the time to compute the properties of these materials with *ab-initio* simulations. Considering the wide applicability of CGCNN, it has the potential to be used to discover new materials in other types of materials and properties. In chapter 5, we applied Bayesian optimization to accelerate discovery of solid polymer electrolytes for lithium ion batteries. This is an example of active learning in which no initial training data is available. The framework automatically explores the material space and we achieved the optimum material with around 100 simulations. This is useful when the computation is more expensive and less data is available from open databases. The major limitation is the lack of decoding methods to generate the structure of solid materials. Therefore, we can only decode to predict the optimum interaction strengths between different components of the system instead of their chemical structures.

We also show that this framework can help us to understand complex material systems that are difficult to study with conventional methods. In section 3.3, we visualize

the similarities between materials for several large material spaces with 10^3-10^4 materials. Useful knowledge is obtained like identifying contribution to material stability by putting different elements into local chemical environments and the clustering of characteristic material structures. In section 4.4, we applied GDyNets to understand the key dynamical processes in solid-liquid interfaces and complex amorphous polymer systems. Scientific insights are learned that provides atomic scale explanations to some of recent experimental discoveries. In these examples, our framework provides a unified method to extract a simple, low dimensional representation for complex material system. This is especially valuable for amorphous, multi-component materials, which are difficult to study with conventional methods due to their intrinsic complexity.

6.2 Future directions

Despite the progresses made in this thesis, there are still many open questions that require further development of deep learning methods for materials. We come back to the four question proposed in section 1.4 and discuss the remaining challenges.

Problem 1: *how to create a neural network architecture that encodes material-specific inductive biases and whether such architecture outperforms existing methods?*

In chapter 2, we have largely solved this problem by developing a generalized neural network architecture (CGCNN) that encodes several material-specific inductive biases like periodicity, permutation, and rotation invariances. The method has shown to outperform other methods when more than 10^3-10^4 data points are available across various materials and properties. [46, 283] Since its publication, CGCNN has been broadly applied to solve many materials design problems in multiple fields. [47, 284, 285] One key challenge is to improve the performance of the method when less data are available. In other deep learning fields, techniques like transfer learning and pre-training have shown promising results to improve performance when less data is available, e.g. medical images [286]. Another direction is to develop neural networks that incorporate different types of inductive biases based on the applications. For ex-

ample, rotational invariance is incorporated in CGCNN so the model can generalize to materials after rotation. However, it also prevents the model from differentiating left-handed and right-handed structures. Developing a series of neural networks including different types of inductive biases would be beneficial to solve problems with different constrains.

Problem 2: *how to extract intuitions that can be understood by human researchers from the learned representations?*

In chapter 3, we have explored one aspect of the problem by extracting material insights from different layers of the neural networks, leveraging the built-in invariances. We only showcase the application of this approach to understand structural and compositional contributions to formation energy, but it can potentially be applied to other types of properties as well. In chapter 4, we explore another aspect with GDyNets by providing a way to understand complex dynamical systems by learning a low dimensional representation of local structures. These methods have enabled the rediscovery of some empirical rules for materials design, like the stability rule for perovskites (section 3.3.2), as well as the discovery new insights for complex materials, like the solvation structures and dynamics of lithium ion in polymer electrolytes (section 4.4.2). The challenge for the future will be to discover quantitative rules, instead of qualitative rules, for complex material systems, which requires the development of symbolic learning methods.

Problem 3: *how to learn the representation of atoms in solid materials when no explicit property labels are available?*

In chapter 4, we have achieved the goal of unsupervised learning of material representation by leveraging the dynamical information in time-series molecular dynamics data, which is applied to understanding the atomic dynamics in complex material systems like the solid-liquid interface (section 4.4.1) and the polymer electrolytes (section 4.4.2). This approach focuses on the dynamical aspect of material structures, but other unsupervised learning methods can be developed that focuses on different aspects, like the structural similarity with regularized entropy match [165]. Unsupervised learning approaches that focus on different aspects will provide pow-

erful tools to navigate the complex space of materials and extract useful knowledge from unstructured material data.

Problem 4: *how to search an unknown material space in a way that balances both exploration and exploitation?*

In chapter 5, we have moved forward a small step towards the autonomous exploration of material spaces. We focus on the coarse-grained space of interactions between various components of polymer electrolytes, and develop a Bayesian optimization framework to explore the space autonomously. In this coarse-grained space, we have achieved the optimum electrolyte with just over 100 simulations and extracted insights for the importance of each interaction parameter. The key challenge ahead is the autonomous exploration of the unconstrained, discrete material space space directly, bypassing the simplified coarse-grained space. The unconstrained, discrete nature of material space makes the autonomous exploration difficult since it is hard to construct a new material from its continuous representation. Solving this challenge could potentially enable unbiased exploration of material spaces and lead to the discovery of materials completely different from the ones we know.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

1. Bednorz, J. G. & Müller, K. A. Possible highT c superconductivity in the Ba-La- Cu- O system. *Zeitschrift für Physik B Condensed Matter* **64,** 189–193 (1986).
2. Liu, M., Johnston, M. B. & Snaith, H. J. Efficient planar heterojunction perovskite solar cells by vapour deposition. *Nature* **501,** 395–398 (2013).
3. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials* **4,** 053208 (2016).
4. *Edison's Lightbulb* `https://www.fi.edu/history-resources/edisons-lightbulb`. Accessed: 2020-06-16.
5. Parr, R. G. in *Horizons of Quantum Chemistry* 5–15 (Springer, 1980).
6. White, A. The materials genome initiative: One year on. *MRS Bulletin* **37,** 715–716 (2012).
7. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559,** 547–555 (2018).
8. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3,** 1–13 (2017).
9. Correa-Baena, J.-P. *et al.* Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2,** 1410–1420 (2018).
10. Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials* **3,** 5–20 (2018).
11. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **5,** 1–17 (2019).
12. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5,** 1–36 (2019).
13. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials* **15,** 1120–1127 (2016).

14. Fujimura, K. *et al.* Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms. *Advanced Energy Materials* **3,** 980–985 (2013).

15. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C* **121,** 511–522 (2017).

16. Engel, E. A., Anelli, A., Ceriotti, M., Pickard, C. J. & Needs, R. J. Mapping uncharted territory in ice from zeolite networks to ice structures. *Nature communications* **9,** 1–7 (2018).

17. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO 2 reduction and H 2 evolution. *Nature Catalysis* **1,** 696–703 (2018).

18. Vandermause, J., Torrisi, S. B., Batzner, S., Kolpak, A. M. & Kozinsky, B. On-the-fly Bayesian active learning of interpretable force-fields for atomistic rare events. *arXiv preprint arXiv:1904.02042* (2019).

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

20. Deng, L., Yu, D., *et al.* Deep learning: methods and applications. *Foundations and Trends in Signal Processing* **7,** 197–387 (2014).

21. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).

22. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).

23. Nair, V. & Hinton, G. E. *Rectified linear units improve restricted boltzmann machines* in *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), 807–814.

24. Mitchell, T. M. *The need for biases in learning generalizations* (Department of Computer Science, Laboratory for Computer Science Research, 1980).

25. Battaglia, P. W. *et al.* Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).

26. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1,** 541–551 (1989).

27. *A Comprehensive Guide to Convolutional Neural Networks* `https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53`. Accessed: 2020-06-18.

28. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115,** 211–252 (2015).

29. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *nature* **323,** 533–536 (1986).

30. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Transactions on Neural Networks* **20,** 61–80 (2008).

31. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).

32. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. *Pointnet: Deep learning on point sets for 3d classification and segmentation* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 652–660.

33. Hall, S. R., Allen, F. H. & Brown, I. D. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations of Crystallography* **47,** 655–685 (1991).

34. Bernstein, F. C. *et al.* The Protein Data Bank: A computer-based archival file for macromolecular structures. *European journal of biochemistry* **80,** 319–324 (1977).

35. Warren, B. E. *X-ray Diffraction* (Courier Corporation, 1990).

36. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science* **58,** 364–369 (2002).

37. Singhal, D. & Curatolo, W. Drug polymorphism and dosage form design: a practical perspective. *Advanced drug delivery reviews* **56,** 335–347 (2004).

38. Jacak, L., Hawrylak, P. & Wojs, A. *Quantum dots* (Springer Science & Business Media, 2013).

39. Dresselhaus, G., Riichiro, S., *et al. Physical properties of carbon nanotubes* (World scientific, 1998).

40. Geim, A. K. Graphene: status and prospects. *science* **324,** 1530–1534 (2009).

41. Elliott, S. R. Physics of amorphous materials. *Longman Group, Longman House, Burnt Mill, Harlow, Essex CM 20 2 JE, England, 1983.* (1983).

42. Lee, J. *et al.* Metal–organic framework materials as catalysts. *Chemical Society Reviews* **38,** 1450–1459 (2009).

43. Gray, F. M. & Gray, F. M. *Solid polymer electrolytes: fundamentals and technological applications* (VCH New York, 1991).

44. Niu, H., Piaggi, P. M., Invernizzi, M. & Parrinello, M. Molecular dynamics simulations of liquid silica crystallization. *Proceedings of the National Academy of Sciences* **115,** 5348–5352 (2018).

45. Xie, T., France-Lanord, A., Wang, Y., Shao-Horn, Y. & Grossman, J. C. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature communications* **10,** 1–9 (2019).

46. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120,** 145301 (2018).

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

48. Xie, T. & Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *The Journal of chemical physics* **149,** 174111 (2018).

49. Wang, Y. *et al.* Toward Designing Highly Conductive Polymer Electrolytes by Machine Learning Assisted Coarse-Grained Molecular Dynamics. *Chemistry of Materials* **32,** 4144–4151 (2020).

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

51. Seko, A. *et al.* Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. *Physical review letters* **115,** 205901 (2015).

52. Xue, D. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nature communications* **7** (2016).

53. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: Critical role of the descriptor. *Physical review letters* **114,** 105503 (2015).

54. Isayev, O. *et al.* Materials cartography: Representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials* **27,** 735–743 (2015).

55. Schütt, K. *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **89,** 205118 (2014).

56. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **115,** 1094–1101 (2015).

57. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **95,** 144110 (2017).

58. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials* **4,** 063801 (2020).

59. Sanderson, R. An interpretation of bond lengths and a classification of bonds. *Science* **114,** 670–672 (1951).

60. Sanderson, R. An explanation of chemical variations within periodic major groups. *Journal of the American Chemical Society* **74,** 4792–4794 (1952).

61. Cordero, B. *et al.* Covalent radii revisited. *Dalton Transactions,* 2832–2838 (2008).

62. Kramida, A., Ralchenko, Y., Reader, J., *et al.* NIST atomic spectra database (ver. 5.2). *National Institute of Standards and Technology, Gaithersburg, MD* (2013).

63. Haynes, W. M. *CRC handbook of chemistry and physics* (CRC press, 2014).

64. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *Imagenet classification with deep convolutional neural networks* in *Advances in neural information processing systems* (2012), 1097–1105.

65. Collobert, R. & Weston, J. *A unified architecture for natural language processing: Deep neural networks with multitask learning* in *Proceedings of the 25th international conference on Machine learning* (2008), 160–167.

66. Duvenaud, D. K. *et al. Convolutional networks on graphs for learning molecular fingerprints* in *Advances in neural information processing systems* (2015), 2224–2232.

67. Henaff, M., Bruna, J. & LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163* (2015).

68. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. *Neural Message Passing for Quantum Chemistry* in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), 1263–1272. <`http://proceedings.mlr.press/v70/gilmer17a.html`>.

69. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

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

71. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **1,** 15010 (2015).

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

73. Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Physical Review B* **34,** 5390 (1986).

74. Foulkes, W., Mitas, L., Needs, R. & Rajagopal, G. Quantum Monte Carlo simulations of solids. *Reviews of Modern Physics* **73,** 33 (2001).

75. Jain, A. *et al.* A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science* **50,** 2295–2310 (2011).

76. De Jong, M. *et al.* Charting the complete elastic properties of inorganic crystalline compounds. *Scientific data* **2,** 150009 (2015).

77. Sripad, S. & Viswanathan, V. Evaluation of Current, Future, and Beyond Li-Ion Batteries for the Electrification of Light Commercial Vehicles: Challenges and Opportunities. *J. Electrochem. Soc.* **164,** E3635–E3646 (2017).

78. Sripad, S. & Viswanathan, V. Performance Metrics Required of Next-Generation Batteries to Make a Practical Electric Semi Truck. *ACS Energy Lett.* **2,** 1669–1673 (2017).

79. Moore, M. D. in (American Institute of Aeronautics and Astronautics, Jan. 10, 2014). doi:10.2514/6.2014-0535. <https://doi.org/10.2514/6.2014-0535>.

80. Cheng, X.-B., Zhang, R., Zhao, C.-Z. & Zhang, Q. Toward Safe Lithium Metal Anode in Rechargeable Batteries: A Review. *Chem. Rev.* **117,** 10403–10473 (2017).

81. Lin, D., Liu, Y. & Cui, Y. Reviving the lithium metal anode for high-energy batteries. *Nat. Nanotechnol.* **12,** 194–206. ISSN: 1748-3387 (Mar. 2017).

82. Christensen, J. *et al.* A Critical Review of Li/Air Batteries. *J. Electrochem. Soc.* **159,** R1–R30 (2011).

83. Xu, W. *et al.* Lithium metal anodes for rechargeable batteries. *Energy Environ. Sci.* **7,** 513–537 (2 2014).

84. Tikekar, M. D., Choudhury, S., Tu, Z. & Archer, L. A. Design principles for electrolytes and interfaces for stable lithium-metal batteries. *Nat. Energy* **1,** 16114 (Sept. 8, 2016).

85. Aurbach, D., Zinigrad, E., Teller, H. & Dan, P. Factors Which Limit the Cycle Life of Rechargeable Lithium (Metal) Batteries. *J. Electrochem. Soc.* **147,** 1274–1279 (2000).

86. Aurbach, D., Zinigrad, E., Cohen, Y. & Teller, H. A short review of failure mechanisms of lithium metal and lithiated graphite anodes in liquid electrolyte solutions. *Solid State Ionics* **148,** 405–416 (2002).

87. Steiger, J., Kramer, D. & Monig, R. Mechanisms of dendritic growth investigated by in situ light microscopy during electrodeposition and dissolution of lithium. *J. Power Sources* **261,** 112–119. ISSN: 0378-7753 (2014).

88. Albertus, P., Babinec, S., Litzelman, S. & Newman, A. Status and challenges in enabling the lithium metal electrode for high-energy and low-cost rechargeable batteries. *Nat. Energy* **3,** 16–21. ISSN: 2058-7546 (2018).

89. Aurbach, D., Markovsky, B., Shechter, A., Ein-Eli, Y. & Cohen, H. A Comparative Study of Synthetic Graphite and Li Electrodes in Electrolyte Solutions Based on Ethylene Carbonate-Dimethyl Carbonate Mixtures. *J. Electrochem. Soc.* **143,** 3809–3820 (1996).

90. Hirai, T., Yoshimatsu, I. & Yamaki, J. Effect of Additives on Lithium Cycling Efficiency. *J. Electrochem. Soc.* **141,** 2300–2305 (1994).

91. Ding, F. *et al.* Dendrite-Free Lithium Deposition via Self-Healing Electrostatic Shield Mechanism. *J. Am. Chem. Soc.* **135,** 4450–4456 (2013).

92. Qian, J. *et al.* High rate and stable cycling of lithium metal anode. *Nat. Commun.* **6,** 6362 (Feb. 20, 2015).

93. Suo, L., Hu, Y.-S., Li, H., Armand, M. & Chen, L. A new class of Solvent-in-Salt electrolyte for high-energy rechargeable metallic lithium batteries. *Nat. Commun.* **4,** 1481 (Feb. 12, 2013).

94. Lu, Y., Tu, Z. & Archer, L. A. Stable lithium electrodeposition in liquid and nanoporous solid electrolytes. *Nat. Mater.* **13,** 961 (Aug. 10, 2014).

95. Zhang, X., Cheng, X., Chen, X., Yan, C. & Zhang, Q. Fluoroethylene Carbonate Additives to Render Uniform Li Deposits in Lithium Metal Batteries. *Adv. Funct. Mater.* **27,** 1605989 (2017).

96. Wang, D. *et al.* Towards High-Safe Lithium Metal Anodes: Suppressing Lithium Dendrites via Tuning Surface Energy. *Adv. Sci.* **4,** 1600168. ISSN: 2198-3844 (2017).

97. Zhang, Y. *et al.* Dendrite-Free Lithium Deposition with Self-Aligned Nanorod Structure. *Nano Lett.* **14,** 6889–6896 (2014).

98. Mayers, M. Z., Kaminski, J. W. & Miller, T. F. Suppression of Dendrite Formation via Pulse Charging in Rechargeable Lithium Metal Batteries. *J. Phys. Chem. C* **116,** 26214–26221 (2012).

99. Aryanfar, A. *et al.* Dynamics of Lithium Dendrite Growth and Inhibition: Pulse Charging Experiments and Monte Carlo Calculations. *J. Phys. Chem. Lett.* **5,** 1721–1726 (2014).

100. Liu, Q. *et al.* Artificial Protection Film on Lithium Metal Anode toward Long-Cycle-Life Lithium-Oxygen Batteries. *Adv. Mater.* **27,** 5241–5247 (2015).

101. Yan, K. *et al.* Ultrathin Two-Dimensional Atomic Crystals as Stable Interfacial Layer for Improvement of Lithium Metal Anode. *Nano Lett.* **14,** 6016–6022 (2014).

102. Liu, Y. *et al.* An Artificial Solid Electrolyte Interphase with High Li-Ion Conductivity, Mechanical Strength, and Flexibility for Stable Lithium Metal Anodes. *Adv. Mater.* **29,** 1605531 (2017).

103. Khurana, R., Schaefer, J. L., Archer, L. A. & Coates, G. W. Suppression of Lithium Dendrite Growth Using Cross-Linked Polyethylene/Poly(ethylene oxide) Electrolytes: A New Approach for Practical Lithium-Metal Polymer Batteries. *J. Am. Chem. Soc.* **136,** 7395–7402 (2014).

104. Stone, G. M. *et al.* Resolution of the Modulus versus Adhesion Dilemma in Solid Polymer Electrolytes for Rechargeable Lithium Metal Batteries. *J. Electrochem. Soc.* **159,** A222–A227 (2012).

105. Yue, L. *et al.* All solid-state polymer electrolytes for high-performance lithium ion batteries. *Energy Storage Mater.* **5,** 139–164. ISSN: 2405-8297 (2016).

106. Janek, J. & Zeier, W. G. A solid future for battery development. *Nat. Energy* **1,** 16141 (Sept. 8, 2016).

107. Li, J., Ma, C., Chi, M., Liang, C. & Dudney, N. J. Solid Electrolyte: the Key for High-Voltage Lithium Batteries. *Adv. Energy Mater.* **5,** 1401408 (2015).

108. Suzuki, Y. *et al.* Transparent cubic garnet-type solid electrolyte of Al2O3-doped Li7La3Zr2O12. *Solid State Ionics* **278,** 172–176. ISSN: 0167-2738 (2015).

109. Manthiram, A., Yu, X. & Wang, S. Lithium battery chemistries enabled by solid-state electrolytes. *Nat. Rev. Mater.* **2,** 16103 (Feb. 14, 2017).

110. Kamaya, N. *et al.* A lithium superionic conductor. *Nat. Mater.* **10,** 682–686. ISSN: 1476-1122 (Sept. 2011).

111. Kato, Y. *et al.* High-power all-solid-state batteries using sulfide superionic conductors. *Nat. Energy* **1,** 16030 (Mar. 21, 2016).

112. Kerman, K., Luntz, A., Viswanathan, V., Chiang, Y.-M. & Chen, Z. Review—Practical Challenges Hindering the Development of Solid State Li Ion Batteries. *Journal of The Electrochemical Society* **164,** A1731–A1744 (2017).

113. Sharafi, A. *et al.* Impact of air exposure and surface chemistry on Li-Li7La3Zr2O12 interfacial resistance. *J. Mater. Chem. A* **5,** 13475–13487 (26 2017).

114. Sharafi, A., Haslam, C. G., Kerns, R. D., Wolfenstine, J. & Sakamoto, J. Controlling and correlating the effect of grain size with the mechanical and electrochemical properties of Li7La3Zr2O12 solid-state electrolyte. *J. Mater. Chem. A* **5,** 21491–21504 (40 2017).

115. Monroe, C. & Newman, J. The Impact of Elastic Deformation on Deposition Kinetics at Lithium/Polymer Interfaces. *J. Electrochem. Soc.* **152,** A396–A404 (2005).

116. Ahmad, Z. & Viswanathan, V. Stability of Electrodeposition at Solid-Solid Interfaces and Implications for Metal Anodes. *Phys. Rev. Lett.* **119,** 056003 (5 Aug. 2017).

117. Diggle, J. W., Despic, A. R. & Bockris, J. O. The Mechanism of the Dendritic Electrocrystallization of Zinc. *J. Electrochem. Soc.* **116,** 1503–1514 (1969).

118. Monroe, C. & Newman, J. Dendrite Growth in Lithium/Polymer Systems: A Propagation Model for Liquid Electrolytes under Galvanostatic Conditions. *J. Electrochem. Soc.* **150,** A1377–A1384 (2003).

119. Monroe, C. & Newman, J. The Effect of Interfacial Deformation on Electrodeposition Kinetics. *J. Electrochem. Soc.* **151,** A880–A886 (2004).

120. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nat. Mater.* **12,** 191 (Feb. 20, 2013).

121. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65,** 1501–1509. ISSN: 1543-1851 (Nov. 1, 2013).

122. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3,** 2810 (Sept. 30, 2013).

123. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **3,** 159–177. ISSN: 2352-8478 (2017).

124. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4,** 268–276 (2018).

125. De Jong, M. *et al.* A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Sci. Rep.* **6,** 34256 (Oct. 3, 2016).

126. Fujimura, K. *et al.* Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms. *Adv. Energy Mater.* **3,** 980–985 (2013).

127. Sendek, A. D. *et al.* Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10,** 306–320 (2017).

128. Evans, J. D. & Coudert, F.-X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **29,** 7833–7839 (2017).

129. Ahmad, Z. & Viswanathan, V. Role of anisotropy in determining stability of electrodeposition at solid-solid interfaces. *Phys. Rev. Materials* **1,** 055403 (5 Oct. 2017).

130. Xu, C., Ahmad, Z., Aryanfar, A., Viswanathan, V. & Greer, J. R. Enhanced strength and temperature dependence of mechanical properties of Li at small scales and its implications for Li metal anodes. *Proc. Natl. Acad. Sci. USA* **114,** 57–61 (2017).

131. Shi, F. *et al.* Strong texturing of lithium metal in batteries. *Proc. Natl. Acad. Sci. USA* **114,** 12138–12143 (2017).

132. Wang, Y. *et al.* Design principles for solid-state lithium superionic conductors. *Nat. Mater.* **14,** 1026–1031. ISSN: 1476-1122 (Oct. 2015).

133. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr., Sect. A* **32,** 751–767 (Sept. 1976).

134. Gotoh, K. & Finney, J. L. Statistical geometrical approach to random packing density of equal spheres. *Nature* **252,** 202 (Nov. 15, 1974).

135. Stepanyuk, V. *et al.* Microstructure and its relaxation in FeB amorphous system simulated by moleculular dynamics. *J. Non-Cryst. Solids* **159,** 80–87. ISSN: 0022-3093 (1993).

136. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68,** 314–319. ISSN: 0927-0256 (2013).

137. Pannikkat, A. & Raj, R. Measurement of an electrical potential induced by normal stress applied to the interface of an ionic material at elevated temperatures. *Acta Mater.* **47,** 3423–3431. ISSN: 1359-6454 (1999).

138. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120,** 145301 (14 Apr. 2018).

139. Jain, A. *et al.* The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1,** 011002. ISSN: 2166532X (2013).

140. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77,** 3865 (1996).

141. Paszke, A. *et al. Automatic differentiation in PyTorch* in *NIPS-W* (2017).

142. Tikekar, M. D., Archer, L. A. & Koch, D. L. Stabilizing electrodeposition in elastic solid electrolytes containing immobilized anions. *Sci. Adv.* **2,** 1600320 (2016).

143. Seino, Y., Ota, T., Takada, K., Hayashi, A. & Tatsumisago, M. A sulphide lithium super ion conductor is superior to liquid ion conductors for use in rechargeable batteries. *Energy Environ. Sci.* **7,** 627–631 (2 2014).

144. Niu, G., Guo, X. & Wang, L. Review of recent progress in chemical stability of perovskite solar cells. *Journal of Materials Chemistry A* **3,** 8970–8980 (2015).

145. Snaith, H. J. Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells. *The Journal of Physical Chemistry Letters* **4,** 3623–3630 (2013).

146. Xu, M., Liang, T., Shi, M. & Chen, H. Graphene-like two-dimensional materials. *Chemical reviews* **113,** 3766–3798 (2013).

147. Butler, S. Z. *et al.* Progress, challenges, and opportunities in two-dimensional materials beyond graphene. *ACS nano* **7,** 2898–2926 (2013).

148. Madelung, O. *Physics of III-V compounds* (J. Wiley, 1964).

149. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature materials* **5,** 909 (2006).

150. Senkan, S. M. High-throughput screening of solid-state catalyst libraries. *Nature* **394,** 350 (1998).

151. Potyrailo, R. *et al.* Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS combinatorial science* **13,** 579–633 (2011).

152. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature materials* **12,** 191 (2013).

153. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chemistry of Materials* **22,** 3762–3767 (2010).

154. Faber, F. A., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Machine Learning Energies of 2 Million Elpasolite (A B C 2 D 6) Crystals. *Physical Review Letters* **117,** 135502 (2016).

155. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **108,** 058301 (2012).

156. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* **89,** 094104 (2014).

157. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2,** 16028 (2016).

158. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics* **134,** 074106 (2011).

159. Pietrucci, F. & Andreoni, W. Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. *Physical review letters* **107,** 085504 (2011).

160. Sadeghi, A. *et al.* Metrics for measuring distances in configuration spaces. *The Journal of chemical physics* **139,** 184118 (2013).

161. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B* **87,** 184115 (2013).

162. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics* **148,** 241717 (2018).

163. Glielmo, A., Zeni, C. & De Vita, A. Efficient nonparametric n-body force fields from machine learning. *Physical Review B* **97,** 184307 (2018).

164. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **96,** 024104 (2017).

165. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* **18,** 13754–13769 (2016).

166. Musil, F. *et al.* Machine learning for the structure–energy–property landscapes of molecular crystals. *Chemical science* **9,** 1289–1300 (2018).

167. Das, P., Moll, M., Stamati, H., Kavraki, L. E. & Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences* **103,** 9885–9890 (2006).

168. Ceriotti, M., Tribello, G. A. & Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **108,** 13023–13028 (2011).

169. Spiwok, V. & Králová, B. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *The Journal of chemical physics* **135,** 224504 (2011).

170. Rohrdanz, M. A., Zheng, W. & Clementi, C. Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annual review of physical chemistry* **64,** 295–316 (2013).

171. Pietrucci, F. & Martoňák, R. Systematic comparison of crystalline and amorphous phases: Charting the landscape of water structures and transformations. *The Journal of chemical physics* **142,** 104704 (2015).

172. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **98,** 146401 (2007).

173. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C* **121,** 511–522 (2016).

174. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* **30,** 595–608 (2016).

175. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212* (2017).

176. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature communications* **8,** 13890 (2017).

177. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148,** 241722 (2018).

178. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* `http://www.deeplearningbook.org` (MIT Press, 2016).

179. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **9,** 513–530 (2018).

180. Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. Neural network models of potential energy surfaces. *The Journal of chemical physics* **103,** 4129–4137 (1995).

181. Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Physical review letters* **120,** 156001 (2018).

182. Schmidt, J. *et al.* Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chemistry of Materials* **29,** 5090–5103 (2017).

183. Zhou, Q. *et al.* Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences,* 201801181 (2018).

184. Willatt, M. J., Musil, F. & Ceriotti, M. A Data-Driven Construction of the Periodic Table of the Elements. *arXiv preprint arXiv:1807.00236* (2018).

185. Castelli, I. E. *et al.* New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy & Environmental Science* **5,** 9034–9043 (2012).

186. Castelli, I. E. *et al.* Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science* **5,** 5814–5819 (2012).

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

188. Shirane, G., Suzuki, K. & Takeda, A. Phase transitions in solid solutions of PbZrO3 and PbTiO3 (II) X-ray study. *Journal of the Physical Society of Japan* **7,** 12–18 (1952).

189. Lang, J., Li, C., Wang, X., *et al.* Improved Photocatalytic Performance under Solar Light Irradiation by Integrating Wide-band-gap Semiconductors, SnO2, SnTaO3 and Sn2Ta2O7. *Materials Today: Proceedings* **3,** 424–428 (2016).

190. Takatsu, H. *et al.* Cubic lead perovskite PbMoO 3 with anomalous metallic behavior. *Physical Review B* **95,** 155105 (2017).

191. Ogitsu, T., Schwegler, E. & Galli, G. β-Rhombohedral boron: at the crossroads of the chemistry of boron and the physics of frustration. *Chemical reviews* **113,** 3425–3449 (2013).

192. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319. ISSN: 0927-0256 (2013).

193. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

194. Zhai, H.-J. *et al.* Observation of an all-boron fullerene. *Nature chemistry* **6,** 727 (2014).

195. Zimmermann, N. E. R., Horton, M. K., Jain, A. & Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Frontiers in Materials* **4,** 34. ISSN: 2296-8016 (2017).

196. Bonnet, J. & Daou, J. Study of the hydrogen solid solution in thulium. *Journal of Physics and Chemistry of Solids* **40,** 421–425 (1979).

197. Etacheri, V., Marom, R., Elazari, R., Salitra, G. & Aurbach, D. Challenges in the development of advanced Li-ion batteries: a review. *Energy & Environmental Science* **4,** 3243–3262 (2011).

198. Imbrogno, J. & Belfort, G. Membrane desalination: Where are we, and what can we learn from fundamentals? *Annual review of chemical and biomolecular engineering* **7,** 29–64 (2016).

199. Peighambardoust, S. J., Rowshanzamir, S. & Amjadi, M. Review of the proton exchange membranes for fuel cell applications. *International journal of hydrogen energy* **35,** 9349–9384 (2010).

200. Zheng, A., Li, S., Liu, S.-B. & Deng, F. Acidic properties and structure–activity correlations of solid acid catalysts revealed by solid-state nmr spectroscopy. *Accounts of chemical research* **49,** 655–663 (2016).

201. Yu, C. *et al.* Unravelling Li-ion transport from picoseconds to seconds: bulk versus interfaces in an argyrodite Li6PS5Cl–Li2S all-solid-state Li-Ion battery. *Journal of the American Chemical Society* **138,** 11192–11201 (2016).

202. Perakis, F. *et al.* Vibrational spectroscopy and dynamics of water. *Chemical reviews* **116,** 7590–7607 (2016).

203. Borodin, O. & Smith, G. D. Mechanism of ion transport in amorphous poly (ethylene oxide)/LiTFSI from molecular dynamics simulations. *Macromolecules* **39,** 1620–1629 (2006).

204. Miller III, T. F., Wang, Z.-G., Coates, G. W. & Balsara, N. P. Designing polymer electrolytes for safe and high capacity rechargeable lithium batteries. *Accounts of chemical research* **50,** 590–593 (2017).

205. Getman, R. B., Bae, Y.-S., Wilmer, C. E. & Snurr, R. Q. Review and analysis of molecular simulations of methane, hydrogen, and acetylene storage in metal–organic frameworks. *Chemical reviews* **112,** 703–723 (2011).

206. Li, Q., Dietrich, F., Bollt, E. M. & Kevrekidis, I. G. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **27,** 103111 (2017).

207. Lusch, B., Kutz, J. N. & Brunton, S. L. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications* **9,** 4950 (2018).

208. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **9,** 5 (2018).

209. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters* **120,** 143001 (2018).

210. Koopman, B. O. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences* **17,** 315–318 (1931).

211. Wu, H. & Noé, F. Variational approach for learning Markov processes from time series data. *arXiv preprint arXiv:1707.04659* (2017).

212. Sastry, S. & Angell, C. A. Liquid–liquid phase transition in supercooled silicon. *Nature materials* **2,** 739 (2003).

213. Angell, C. A. Insights into phases of liquid water from study of its unusual glass-forming properties. *Science* **319,** 582–587 (2008).

214. Ryu, S. & Cai, W. A gold–silicon potential fitted to the binary phase diagram. *Journal of Physics: Condensed Matter* **22,** 055401 (2010).

215. Wang, Y., Santana, A. & Cai, W. Atomistic mechanisms of orientation and temperature dependence in gold-catalyzed silicon growth. *Journal of Applied Physics* **122,** 085106 (2017).

216. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).

217. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology* **25,** 135–144 (2014).

218. Husic, B. E. & Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **140,** 2386–2396 (2018).

219. Meyer, W. H. Polymer electrolytes for lithium-ion batteries. *Advanced materials* **10,** 439–448 (1998).

220. Hallinan Jr, D. T. & Balsara, N. P. Polymer electrolytes. *Annual review of materials research* **43,** 503–525 (2013).

221. Mao, G., Perea, R. F., Howells, W. S., Price, D. L. & Saboungi, M.-L. Relaxation in polymer electrolytes on the nanosecond timescale. *Nature* **405,** 163 (2000).

222. Do, C. *et al.* Li+ transport in poly (ethylene oxide) based electrolytes: neutron scattering, dielectric spectroscopy, and molecular dynamics simulations. *Physical review letters* **111,** 018301 (2013).

223. Diddens, D., Heuer, A. & Borodin, O. Understanding the lithium transport within a rouse-based model for a PEO/LiTFSI polymer electrolyte. *Macromolecules* **43,** 2028–2036 (2010).

224. Bachman, J. C. *et al.* Inorganic solid-state electrolytes for lithium batteries: mechanisms and properties governing ion conduction. *Chemical reviews* **116,** 140–162 (2015).

225. Pesko, D. M. *et al.* Negative transference numbers in poly (ethylene oxide)-based electrolytes. *Journal of The Electrochemical Society* **164,** E3569–E3575 (2017).

226. Mezić, I. Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics* **45,** 357–378 (2013).

227. Georgiev, G., Georgieva, V. & Plieth, W. Markov chain model of electrochemical alloy deposition. *Electrochimica acta* **51,** 870–876 (2005).

228. Valor, A., Caleyo, F., Alfonso, L., Velázquez, J. & Hallen, J. Markov chain models for the stochastic modeling of pitting corrosion. *Mathematical Problems in Engineering* **2013** (2013).

229. Miller, J. A. & Klippenstein, S. J. Master equation methods in gas phase chemical kinetics. *The Journal of Physical Chemistry A* **110,** 10528–10544 (2006).

230. Buchete, N.-V. & Hummer, G. Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B* **112,** 6057–6069 (2008).

231. Sriraman, S., Kevrekidis, I. G. & Hummer, G. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *The Journal of Physical Chemistry B* **109,** 6479–6484 (2005).

232. Gu, C. *et al. Building Markov state models with solvent dynamics* in *BMC bioinformatics* **14** (2013), S8.

233. Hamm, P. Markov state model of the two-state behaviour of water. *The Journal of chemical physics* **145,** 134501 (2016).

234. Schulz, R. *et al.* Collective hydrogen-bond rearrangement dynamics in liquid water. *The Journal of chemical physics* **149,** 244504 (2018).

235. Cubuk, E. D., Schoenholz, S. S., Kaxiras, E. & Liu, A. J. Structural properties of defects in glassy liquids. *The Journal of Physical Chemistry B* **120,** 6139–6146 (2016).

236. Nussinov, Z. *et al.* in *Information Science for Materials Discovery and Design* 115–138 (Springer, 2016).

237. Kahle, L., Musaelian, A., Marzari, N. & Kozinsky, B. Unsupervised landmark analysis for jump detection in molecular dynamics simulations. *arXiv preprint arXiv:1902.02107* (2019).

238. Funke, K. Jump relaxation in solid electrolytes. *Progress in Solid State Chemistry* **22,** 111–195 (1993).

239. Xu, K. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chemical reviews* **104,** 4303–4418 (2004).

240. Corry, B. Designing carbon nanotube membranes for efficient water desalination. *The Journal of Physical Chemistry B* **112,** 1427–1434 (2008).

241. Cohen-Tanugi, D. & Grossman, J. C. Water desalination across nanoporous graphene. *Nano letters* **12,** 3602–3608 (2012).

242. Rowsell, J. L., Spencer, E. C., Eckert, J., Howard, J. A. & Yaghi, O. M. Gas adsorption sites in a large-pore metal-organic framework. *Science* **309,** 1350–1354 (2005).

243. Li, J.-R., Kuppler, R. J. & Zhou, H.-C. Selective gas adsorption and separation in metal–organic frameworks. *Chemical Society Reviews* **38,** 1477–1504 (2009).

244. Wehmeyer, C. & Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics* **148,** 241703 (2018).

245. Ribeiro, J. M. L., Bravo, P., Wang, Y. & Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *The Journal of Chemical Physics* **149,** 072301 (2018).

246. Wu, H., Mardt, A., Pasquali, L. & Noe, F. *Deep Generative Markov State Models* in *Advances in Neural Information Processing Systems* (2018), 3979–3988.

247. Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv preprint arXiv:1802.04364* (2018).

248. Simonovsky, M. & Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv preprint arXiv:1802.03480* (2018).

249. Sultan, M. M. & Pande, V. S. Transfer Learning from Markov models leads to efficient sampling of related systems. *The Journal of Physical Chemistry B* (2017).

250. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS central science* **3,** 283–293 (2017).

251. Dunn, B., Kamath, H. & Tarascon, J.-M. Electrical energy storage for the grid: a battery of choices. *Science* **334,** 928–935 (2011).

252. Agrawal, R. & Pandey, G. Solid polymer electrolytes: materials designing and all-solid-state battery applications: an overview. *Journal of Physics D: Applied Physics* **41,** 223001 (2008).

253. Li, L. *et al.* Recent advances in flexible/stretchable supercapacitors for wearable electronics. *Small* **14,** 1702829 (2018).

254. Blomgren, G. E. The development and future of lithium ion batteries. *Journal of The Electrochemical Society* **164,** A5019–A5025 (2017).

255. Tikekar, M. D., Choudhury, S., Tu, Z. & Archer, L. A. Design principles for electrolytes and interfaces for stable lithium-metal batteries. *Nature Energy* **1,** 16114 (2016).

256. Yue, L. *et al.* All solid-state polymer electrolytes for high-performance lithium ion batteries. *Energy Storage Materials* **5,** 139–164 (2016).

257. Maitra, A. & Heuer, A. Cation transport in polymer electrolytes: A microscopic approach. *Physical review letters* **98,** 227802 (2007).

258. Chintapalli, M. *et al.* Relationship between conductivity, ion diffusion, and transference number in perfluoropolyether electrolytes. *Macromolecules* **49,** 3508–3515 (2016).

259. Mannodi-Kanakkithodi, A. *et al.* Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Materials Today* **21,** 785–796 (2018).

260. Hatakeyama-Sato, K., Tezuka, T., Nishikitani, Y., Nishide, H. & Oyaizu, K. Synthesis of Lithium-ion Conducting Polymers Designed by Machine Learning-based Prediction and Screening. *Chemistry Letters* **48** (2018).

261. Mogurampelly, S., Borodin, O. & Ganesan, V. Computer simulations of ion transport in polymer electrolyte membranes. *Annual review of chemical and biomolecular engineering* **7,** 349–371 (2016).

262. France-Lanord, A. *et al.* The effect of chemical variations in the structure of poly (ethylene oxide)-based polymers on lithium transport in concentrated electrolytes. *Chemistry of Materials* (2019).

263. Gorecki, W., Jeannin, M., Belorizky, E., Roux, C. & Armand, M. Physical properties of solid polymer electrolyte PEO (LiTFSI) complexes. *Journal of Physics: Condensed Matter* **7,** 6823 (1995).

264. Underhill, P. T. & Doyle, P. S. On the coarse-graining of polymers into bead-spring chains. *Journal of non-newtonian fluid mechanics* **122,** 3–31 (2004).

265. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J. & Lookman, T. Adaptive strategies for materials design using uncertainties. *Scientific reports* **6,** 19660 (2016).

266. Ju, S. *et al.* Designing nanostructures for phonon transport via Bayesian optimization. *Physical Review X* **7,** 021024 (2017).

267. González, J., Dai, Z., Hennig, P. & Lawrence, N. *Batch bayesian optimization via local penalization* in *Artificial Intelligence and Statistics* (2016), 648–657.

268. Shah, D. B. *et al.* Effect of anion size on conductivity and transference number of perfluoroether electrolytes with lithium salts. *Journal of The Electrochemical Society* **164,** A3511–A3517 (2017).

269. Zhang, H. *et al.* Lithium bis (fluorosulfonyl) imide/poly (ethylene oxide) polymer electrolyte. *Electrochimica Acta* **133,** 529–538 (2014).

270. Ma, Q. *et al.* Single lithium-ion conducting polymer electrolytes based on a super-delocalized polyanion. *Angewandte Chemie International Edition* **55,** 2521–2525 (2016).

271. Zhang, H. *et al.* Single lithium-ion conducting solid polymer electrolytes: advances and perspectives. *Chemical Society Reviews* **46,** 797–815 (2017).

272. Lee, Y.-C., Ratner, M. A. & Shriver, D. F. Ionic conductivity in the poly (ethylene malonate)/lithium triflate system. *Solid State Ionics* **138,** 273–276 (2001).

273. Zhang, Z. *et al.* Ion conductive characteristics of cross-linked network polysiloxane-based solid polymer electrolytes. *Solid state ionics* **170,** 233–238 (2004).

274. Tominaga, Y., Shimomura, T. & Nakamura, M. Alternating copolymers of carbon dioxide with glycidyl ethers for novel ion-conductive polymer electrolytes. *Polymer* **51,** 4295–4298 (2010).

275. Webb, M. A. *et al.* Systematic computational and experimental investigation of lithium-ion transport mechanisms in polyester-based polymer electrolytes. *ACS central science* **1,** 198–205 (2015).

276. Zhao, Q., Liu, X., Stalin, S., Khan, K. & Archer, L. A. Solid-state polymer electrolytes with in-built fast interfacial transport for secondary lithium batteries. *Nature Energy* **4,** 365–373 (2019).

277. Savoie, B. M., Webb, M. A. & Miller III, T. F. Enhancing cation diffusion and suppressing anion diffusion via Lewis-acidic polymer electrolytes. *The journal of physical chemistry letters* **8,** 641–646 (2017).

278. Yang, X. *et al. Ion pair dissociation effects of aza-based anion receptors on lithium salts in polymer electrolytes* (The Office of Scientific and Technical Information, Oak Ridge, TN, 1996).

279. Ibrahim, S., Yasin, S. M. M., Nee, N. M., Ahmad, R. & Johan, M. R. Conductivity, thermal and infrared studies on plasticized polymer electrolytes with carbon nanotubes as filler. *Journal of Non-Crystalline Solids* **358,** 210–216 (2012).

280. Borodin, O. Polarizable force field development and molecular dynamics simulations of ionic liquids. *The Journal of Physical Chemistry B* **113,** 11463–11478 (2009).

281. Webb, M. A., Savoie, B. M., Wang, Z.-G. & Miller III, T. F. Chemically specific dynamic bond percolation model for ion transport in polymer electrolytes. *Macromolecules* **48,** 7346–7358 (2015).

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

283. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *arXiv preprint arXiv:2005.00707* (2020).

284. Lym, J., Gu, G. H., Jung, Y. & Vlachos, D. G. Lattice Convolutional Neural Network Modeling of Adsorbate Coverage Effects. *The Journal of Physical Chemistry C* **123,** 18951–18959 (2019).

285. Kim, M. *et al.* Artificial Intelligence to Accelerate the Discovery of N2 Electroreduction Catalysts. *Chemistry of Materials* **32,** 709–720 (2019).

286. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annual review of biomedical engineering* **19,** 221–248 (2017).