Improving supervised machine learning for materials science

by

Sheng Gong

Bachelor of Engineering, Peking University (2018)

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 2022

Author. . . . . .

Sheng Gong

Department of Materials Science and Engineering

June 21, 2022

Certified by. . . . . .

Jeffrey C. Grossman

Professor of Materials Science and Engineering

Thesis Supervisor

Accepted by. . . . . .

Frances M. Ross

Chair, Departmental Committee on Graduate Studies

Improving supervised machine learning for materials science

by

Sheng Gong

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

Abstract

Despite the widespread applications of machine learning models in materials science, in many cases the performance of machine learning models is not sufficiently accurate enough to meet the needs of materials design. In this thesis, we propose and apply a series of strategies to exam and improve upon the performance of machine learning models for specific materials problems. First, we exam whether current deep representation learning models for atomistic systems can capture human knowledge of crystal structures, and find that current graph neural networks can capture knowledge of local atomic environments but cannot capture periodicity of crystal structures. As an initial solution, we propose to hybridize human knowledge with deep representation learning models, and find that the hybridization can lead to large improvement for predicting vibrational properties of materials. Then, for situations where the datasets of target materials properties are small while there are large relevant materials datasets, we propose to use transfer learning and multi-fidelity learning to transfer information between the large and small datasets to facilitate the learning of target properties. We use experimentally measured formation enthalpy and lattice thermal conductivity as case studies to exam the usefulness of information transfer and understand where and why information transfer helps. For situations where expansion of datasets is necessary, we propose to use active learning/Bayesian Optimization to sample the materials space efficiently and mitigate bias, and as a case study, we apply Bayesian Optimization to find the optimal laser processing parameters for poly(acrylonitrile) sheet as porous carbon electrode. Finally, if generation of data is time-consuming, we propose to use machine learning to accelerate materials experiments and simulations. For this goal, we develop a framework to use graph neural networks to predict charge density distribution of materials. The machine learning models developed in this thesis not only deepen human understanding of where and how machine learning can be used to facilitate materials development, but also lead to the discovery of new materials systems, new processes, and new insights, such as new candidate thermoelectric materials, new processes for lasering poly(acrylonitrile), and new insights into the evaluation of the stability of materials.

Thesis Supervisor: Jeffrey C. Grossman

Title: Professor of Materials Science and Engineering

Head of the Department of Materials Science and Engineering, MIT

Acknowledgements

I am very fortunate to work and live in a very lovely environment here in MIT and in the Boston area, and I am very lucky to have wonderful colleagues, families and friends in my PhD.

First and foremost, I would like to thank my thesis supervisor, Professor Jeffrey C. Grossman. Jeff is a great scientist, mentor and leader. As a scientist, Jeff not only provides insightful and constructive suggestions to my machine learning projects, but also encourages me to collaborate with researchers with different backgrounds to apply my machine learning skills on different tasks of materials discovery. As a mentor, Jeff doesn't see his role as one of supervision, but instead of gentle guidance, inspiration, connection, and collaboration. Jeff is always supportive to me, gives me freedom to explore the directions I am interested in, and never pushes me to do anything. As a leader, Jeff leads the group and the department through the toughest pandemic period, and he always puts everyone's happiness at the first place. Jeff has been and will always be my model of knowledge, leadership and personality.

I would like to thank Professor Rafael Gomez-Bombarelli and Professor Ju Li as my thesis committee members, Professor Yang Shao-Horn from the MIT TRI team, and Dr. Mordechai Kornbluth from BOSCH as my intern supervisor. Feedbacks and suggestions from these distinguished scholars are invaluable to my researches.

I am grateful to have wonderful collaborators in my PhD. I would like to thank Dr. Tian Xie, Dr. Taishan Zhu, Dr. Shuo Wang, Jatin Patil, and all other collaborators. I always see Tian as my model of doing machine learning research, and he provides constructive suggestions to almost all projects in my PhD. Taishan is a great mentor, officemate, and friend of me, who influenced me deeply on both research and career. Shuo is my mentor and teacher in my undergraduate study. Although we are not in the same institute during my PhD, we chat with each other frequently about research, career, and life. Jatin guides me how to apply machine learning directly on experimental works, which is a special experience for me as a theorist.

I would like thank all the current and past members of the Grossman group for the insightful scientific discussions and fun times spent together: Laura M. von Bosau, Nicola Ferralis, Zhengmao Lu, Dillon C. Yost, Tae Won Nam, Adam Trebach, Asmita Jana, Jatin Patil, Taishan Zhu, Xining Zang, Thomas Sannicolo, Yanming Wang, David Bergsman, Beza Getachew, Arthur France-Lanord, Emily Crabb, Ki-Jana Carter, David (Woo Hyun) Chae, Xiang Zhang, Tian Xie, Eric Fadel, Owen Morris, Yun Liu. I will always remember the group meetings, group lunches/dinners, group retreats, and time spent in the offices, hallways, and "caves" with them. I would also like to thank all my classmates, especially Runze Liu, Siying Huang, and Zhichu Ren. Studying, taking exams, and sharing experiences with them is another part my of memory at MIT beyond research.

On a more personal note, I would like to first thank my fiancée, Yu Wang. I could not survive the past nine years without her. We have sweet memories in Hunan (our hometown), Beijing, and Boston, and I believe we will have a happy life in the rest of our lives, regardless of where we will be. I would like to thank our cat, Erniu, who is the icon of my Zoom and social media profile and brings lots of accompany and love to us. I would like to thank my parents for their consistent care and support. Finally, I would like to thank all my friends I met in Boston: Celesta, Lun, Aria, Carrie, Eddy, Jason, Chen, James, K.D., Xiaoyu, Shufan, Wei, Shuo, Peter, Jiawen, Candy, Markus, Lanke Club, MZ Club, and many others. Chatting, drinking, skiing, traveling, playing poker, board gaming and doing other activities with you makes my life in Boston enjoyable and unforgettable.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

1. Introduction

1.1. Motivations for materials science

The development of materials plays a critical role in the human's civilizations. Historians define the periods of civilizations by the predominantly used material, such as Stone Age, Bronze Age, Iron Age, Steel Age, and Silicon Age[1]. Nowadays, it is urgent to accelerate materials development for various environmental and commercial purposes of human beings. For example, controlling the global warming demands the development of materials used in solar cells for more efficient generation of clean and renewable energy[2], and promoting electric vehicles requires the development of battery materials for larger energy storage, faster charging rate and higher safety[3].

In general, the development of novel materials is a notoriously difficult and slow process. An extreme example is the pitch drop experiment[4], where it takes 7 to 13 years to form a single drop. In Figure 1-1, we show the four paradigms of materials development over history[5]. Due to the lack of understanding of the structure-property relationships, in the first paradigm materials development was driven by trial-and-errors, and in the second paradigm, materials scientists started to gather empirical knowledge from the numerous experimental results, such as the law of thermodynamics. With the development of quantum mechanics, solid state physics and computation powers in the 20^{th} century, in the third paradigm, computational materials science started to compute materials properties based on the first principles, which motivates the creation of large computational materials datasets that explore the vast space of materials and provide open data for materials design, such as the Materials Project (MP)[6], Open Quantum Materials Database (OQMD)[7], the Automatic Flow of Materials Discovery

Library (AFLOW)[8], and the Joint Automated Repository for Various Integrated Simulations (JARVIS)[9]. More materials databases are summarized in Ref.[10-14]. With the access of tremendous amount of materials data, in the fourth paradigm, machine learning and data-driven approaches provide new opportunity for accelerating materials development. Ideally, machine learning models can extract the structure-property relations from the big materials datasets, and then apply the found relations to guide the design of materials, such as prediction of properties, and optimization of experimental conditions[15-18]. The main advantage of machine learning models over experiments and materials simulation is that, the decision process of machine learning models is usually much faster (seconds) for a given material than that of simulation (hours to days) and experiments (days to months), and the main advantage of machine learning over human summarization of physical rules is that, machine learning models can deal with very large datasets and extract very complex and non-linear relations between multiple inputs and outputs. In Chapter 1.2.2, we will overview the applications of machine learning in materials science.

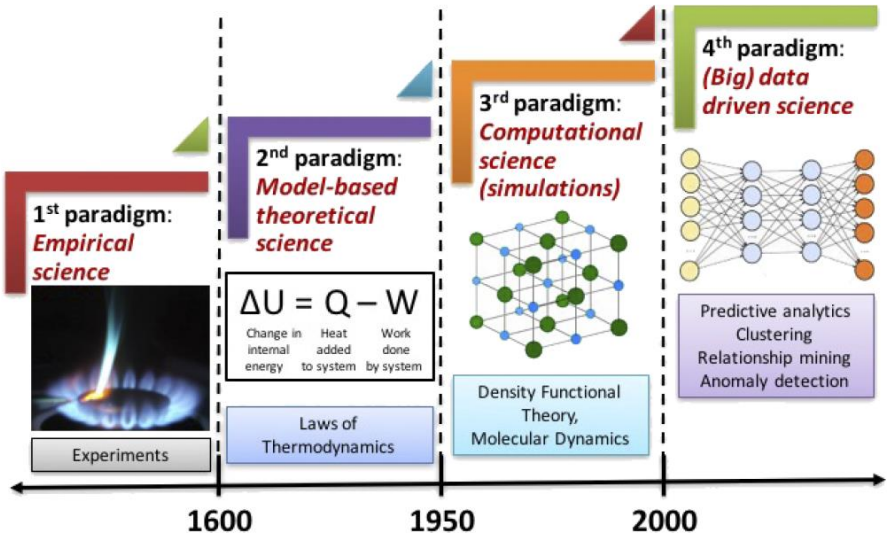

Figure 1-1. The four paradigms of materials science: empirical, theoretical, computational, and data-driven[5].

1.2. Motivations for machine learning

1.2.1. Overview of machine learning

"Machine learning" refers to the development of models that learn from experiences and make decisions based on experiences without explicitly being programmed for a given dataset, such as playing chess and social network recommendation, *etc*. Some of commonly used machine learning technologies are, linear regression, decision tree, random forest, and neural network.

Linear regression is one of the simplest machine learning models. The basic formulation of linear regression can be written as below:

$$y_{\text{pred}} = w^T x + b \text{ ..... (1-1)},$$

where $x \in \mathrm{R}^m$ is the given input, $y_{\text{pred}} \in \mathrm{R}^n$ is the predicted output, and $w \in \mathrm{R}^{m \times n}$ and b $\in \mathrm{R}^n$ are learned weights.

Decision tree is a flowchart-like model, where each node is a test on input, each branch is one of the outcomes of the test, and each leaf node represents one of the decisions (predicted values). The paths from the root to the leaf are the decision rules for the given dataset. Compared with linear regression, decision tree can learn highly nonlinear relation between input and output. More details about decision tree are provided in Ref.[19]. Random forest is an ensemble of decision trees, where multiple decision trees are used to learn the input-output relation for better performance[20].

Neural network is one of the most popular and powerful machine learning architectures for learning input-output relationships. The simplest form of neural network, 2-layer feedforward neural network, can be written as below:

$$y_{\text{pred}} = g\left(w^{(1)^T}x + b^{(1)}\right)w^{(2)^T} + b^{(2)} \text{ ..... (1-2),}$$

where g represents a non-linear activation function, and $(w^{(1)}, b^{(1)})$ and $(w^{(2)}, b^{(2)})$ are learned weights in the first and second layer, respectively. From equation (1-2), we can see that feedforward neural network is essentially multiple layers of linear transformations plus non-linear activation functions. More general neural networks can differ from the simple form in several aspects, such as larger number of layers, more complicated transformation in each layer, and constraints to the weights. According to the "universal approximation theorem"[21], theoretically neural network can approximate any function to arbitrary accuracy, although practically optimization of weights of neural network ("training") might not always be easy.

1.2.2. Applications of machine learning in materials science

In this chapter, we overview several purposes that researchers have applied machine learning models to achieve. We will focus on the formats of input and output data and impact of prediction of the output data on materials development, and detailed architectures of machine learning models to realize these purposes are discussed in Ref.[10-12, 22] and later chapters.

In Figure 1-2, we summarize the commonly seen input and output data formats used in machine learning models for materials science. Atomistic structure of materials is one of the most frequently used input data formats, because fundamentally all materials properties are determined by structure. Conversion of atomistic structure into machine-readable numerical representation has become a central task for machine learning applications in materials science, and we will overview several methods for this challenge in Chapter 1.3.1.

In cases where structural information of materials is missing, compositions can also be used as input data for various purposes, such as scalar property prediction[23-25], and prediction of possible stable structures for the given compositions[26-28]. Recently, Schmidt *et al.* have used

composition plus incomplete structural information (such as structural prototype) to predict scalar property[29].

Spectrums of materials are also used as input for machine learning models. For example, spectrums from materials structure characterizations, such as X-ray diffraction (XRD) and X-ray absorption (XAS), are used as input of machine learning models to predict the structural information with higher speed than manual analysis[10]. Another type of materials spectrums, density of states, can also be used to predict materials properties. For example, electronic density of states are used in DOSnet[30] to predict adsorption energy of adsorbates on surfaces.

With the rapid development of computer vision, images of materials, such as those from optical microscopy (OM), scanning electron microscopy (SEM), scanning tunneling microscopy (STM), atomic force microscopy (AFM), and transmission electron microscopy (TEM), are also used as input to machine learning models. Machine learning models can extract information from images with higher speed and better consistency of measurements than manual analysis. For example, SEM images can be used to classify different materials systems[31] and locate defects inside a material[32].

In addition to the intrinsic description of materials as mentioned above, scientific literature of materials can also be used as input to machine learning models to learn and make inferences from the text information, such as prediction of synthesis conditions[33-35] and properties[36-38].

Finally, synthesis conditions[16-18, 39-42] and computation settings[43-45] can be used as input of machine learning models to predict the materials properties from the corresponding experiments and simulations.

As for the output data formats, scalar materials properties, such as band gap, formation energy, and bulk modulus, are frequently seen outputs of machine learning models for materials

science[25, 46-75]. Recently, machine learning models for predicting spectral and tensorial properties have been developed, such as Mat2Spec[76] and modified E3NN[77] for prediction of density of states, and ETGNN[78] for prediction of force, dielectric and piezoelectric tensors.

In addition to direct prediction of materials properties, in cases where direct predictions are not accurate because of limitations of models and/or scarcity of data as discussed in Chapter 1.3, machine learning models can be used to predict the intermediate physical quantities, such as interatomic energy and forces, and charge density distributions and wavefunctions. Prediction of interatomic energy and forces enables machine learning accelerated molecular dynamics[79-83], and prediction of charge density distributions and wavefunctions can accelerate density function theory calculations (DFT)[73, 84-90]. Beyond acceleration of routine simulations, such predictions of intermediate physical quantities pushes the boundary of materials simulations. For example, machine learning prediction of charge density can push the limit of density functional theory to thousands of atoms[86], and machine learning can solve the fractional electron problem in DFT[90].

Meanwhile, atomistic structures of materials can also be predicted by machine learning models for three purposes, searching for the most stable structures under constraints[26-28], enlarging the space of stable materials structures[12, 26, 28, 91], and inverse design of atomistic structures with target properties[91-94]. Experimental parameters that result in the optimal materials properties can also be suggested by machine learning models, especially based on the scientific literature as the input[10, 34, 35, 95].

Finally, although often not as the direct output, insights of materials systems can be extracted from machine learning models. The most commonly seen insight is how characteristics of materials affect materials properties revealed by the feature importance or similar metrics such as impact on model output[37, 60, 68, 96-99]. Other types of knowledge from machine learning

models include physical formulas from symbolic regression[17, 80, 100-102], physical concepts[103], knowledge graph[104], visualization of similarity between materials[105], and capturing dynamical information from trajectories of molecular dynamics simulations[106].

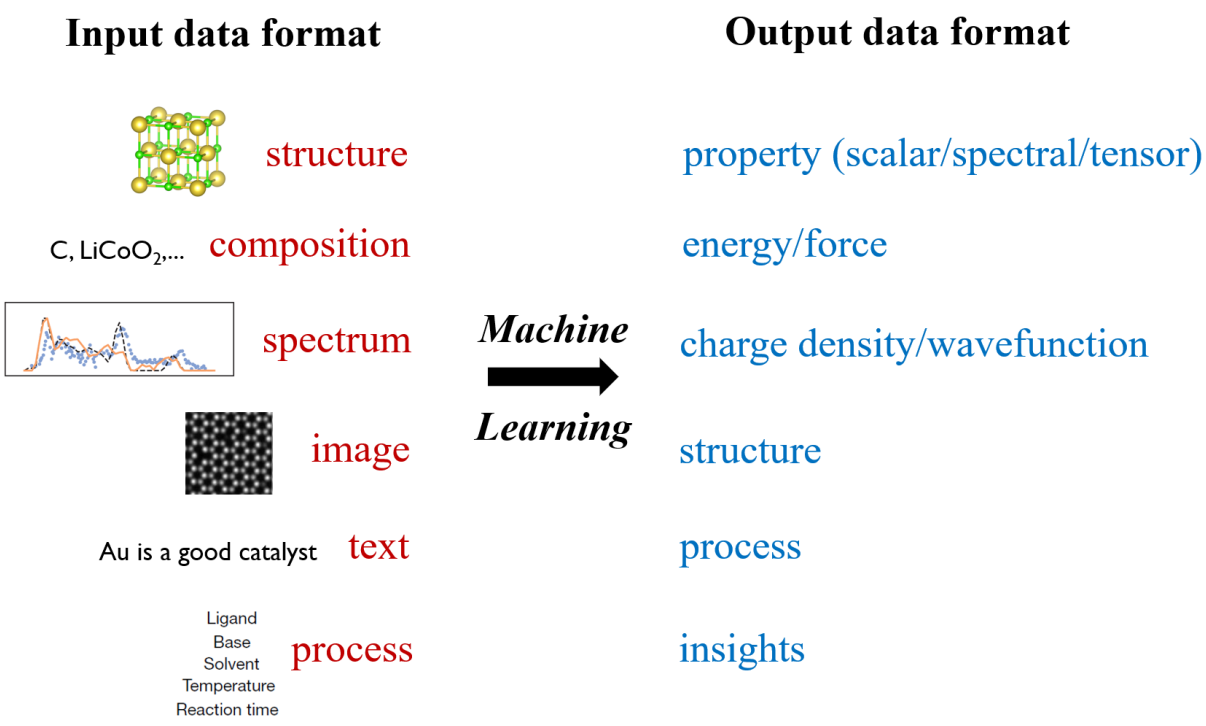

Figure 1-2. Illustration of input and output data formats used in machine learning models for materials science.

1.3.Current limitations of machine learning for materials science

Although machine learning models have been widely applied in materials science, in many cases the performance of machine learning models is still limited and not sufficiently accurate enough to meet the needs of materials design. Such limitations are shown in two aspects: on the one hand, the prediction accuracy of machine learning models is not satisfying for many materials properties. For example, in the study of ALIGNN[69], one of the state-of-the-art machine learning models taking atomistic structures as input, the authors report that ALIGNN

can only achieve satisfying predictive performance[25] for 6 tasks out of 29 tasks; on the other hand, sometimes even if the errors of machine learning models are small, the prediction results are still not very useful. For example, Bartel *et al*.[67] report that even if some machine learning models can learn the formation enthalpy of materials with errors similar with that of DFT compared with experiment, the machine learning predicted formation enthalpy cannot estimate the stability of compounds as correct as DFT, because errors from DFT are beneficially systematic whereas errors from machine learning are not. In the following, we will overview the current limitations of machine learning models on materials data from two aspects: representations of materials, and datasets of materials.

1.3.1. Representations of materials structures

Conversion of atomistic structures into machine-readable numerical representation, or designing the x based on atomistic structures of materials for machine learning models, is one of the most critical tasks for applications of machine learning in materials science[11]. In general, there are two approaches to convert materials structures into numbers: human-designed description and deep representation learning. In this chapter, we will overview the two types of representations, and discuss their limits. More detailed discussions about representations of materials are provided in Ref.[10, 11].

Human-designed descriptors. Human-designed descriptors are based on the collection of human understanding of compositions and structures of materials. Generally, people can easily understand the meaning of such descriptors. Mean electronegativity and difference of atomic radius of elements in materials are examples of compositional descriptors, and mean bond length and difference of coordination number of atoms in materials are examples of structural descriptors of materials. Beyond such simple descriptors, recently, researchers have proposed a series of descriptors for materials, such as Magpie[25] compositional descriptors, classical

force-field inspired descriptors (CFID)[72], Coulomb matrix[107], fragment descriptors[62], voxel descriptors[108], and partial radial distribution function[74]. In addition to the general descriptors, system-specific descriptors are also proposed, such as those for MOF[109], zeolite[110], and surface of materials[111].

Although ML models based on the human-designed descriptors have achieved some successes in revealing the trend between human-understandable characteristics of materials and properties[55, 60, 70, 97, 101, 112, 113], these descriptors contain only information known to human-beings, and employing only descriptors to learn and predict materials properties might miss key structure-property relation unknown to human.

Deep representation learning. Deep representation learning, by definition, refers to the ML models that learn the numerical representation of materials automatically during the training of machine learning. Although the learned representations are generally less understandable by human compared with human-designed descriptors, deep representation learning can uncover the pattern of structure-property relation unknown to human. Since materials can be intuitively represented as graphs, where atoms forming the nodes and bonds forming the edges, graph neural networks (GNN) have become the state-of-the-art deep representation learning method for materials science. SchNet[57] and CGCNN[71] are two classic GNN architectures designed for materials. They update the representations of each atom by neighboring atoms and bond length between atoms, and pool all the updated atom representations into an overall representation of each material. In later variants of GNN such as iCGCNN[75], MEGNet[114] and GATGNN[115], bond representations are also updated during the convolution. Through multiple layers of graph convolutions, these models can implicitly encode many-body interactions. To explicitly encode many-body interactions, Gasteiger *et al.* proposed DimeNet[116] and GemNet[117] for molecules, and Choudhary *et al.* proposed ALIGNN for

materials[69], where atom representations (one-body), bond representations (two-body) and bond angle representations (three-body) are all updated during the convolution via the construction of line graph (the nodes of the line graph are edges in the original graph, and the edges of the line graph are angles between edges in the original graph). Together with other studies using higher-order information for improving expressiveness of GNN[118, 119], ALIGNN-*d*[120], a recent variant of ALIGNN, updates the dihedral angle representation (four-body) by constructing line graph of line graph. Very recently, Batatia *et al.* proposed a general formalism to encode the local atomic environments by graph neural networks and atomic cluster expansion with arbitrary body-order[121]. Other efforts have also been made to improve GNN for crystal structures, such as inclusion of state attributes in MEGNet[114], attention mechanism in GATGNN[115], representations equivariant to rotations in E3NN[77], use of structure motifs in AMDNet[63], and exploitation of correlations in spectral properties in Mat2Spec[76].

Although these variants of GNN have achieved some success in learning materials' properties, for capturing the atomic structure of crystalline materials, the improvements are mainly based on human intuition of local bonding environment, such as explicitly encoding bond angle (three-body) and dihedral angle (four-body) information, structure motif, and representations equivariant to rotations. Currently, prediction of materials properties is in general still challenging[69], and there is still no systematic approach and quantitative metric to analyze and understand the limitations of GNNs for capturing crystal structures. Moreover, the two methods of converting crystal structures into numbers, human-designed descriptors and deep representation learning, are now developed separately, not synergically.

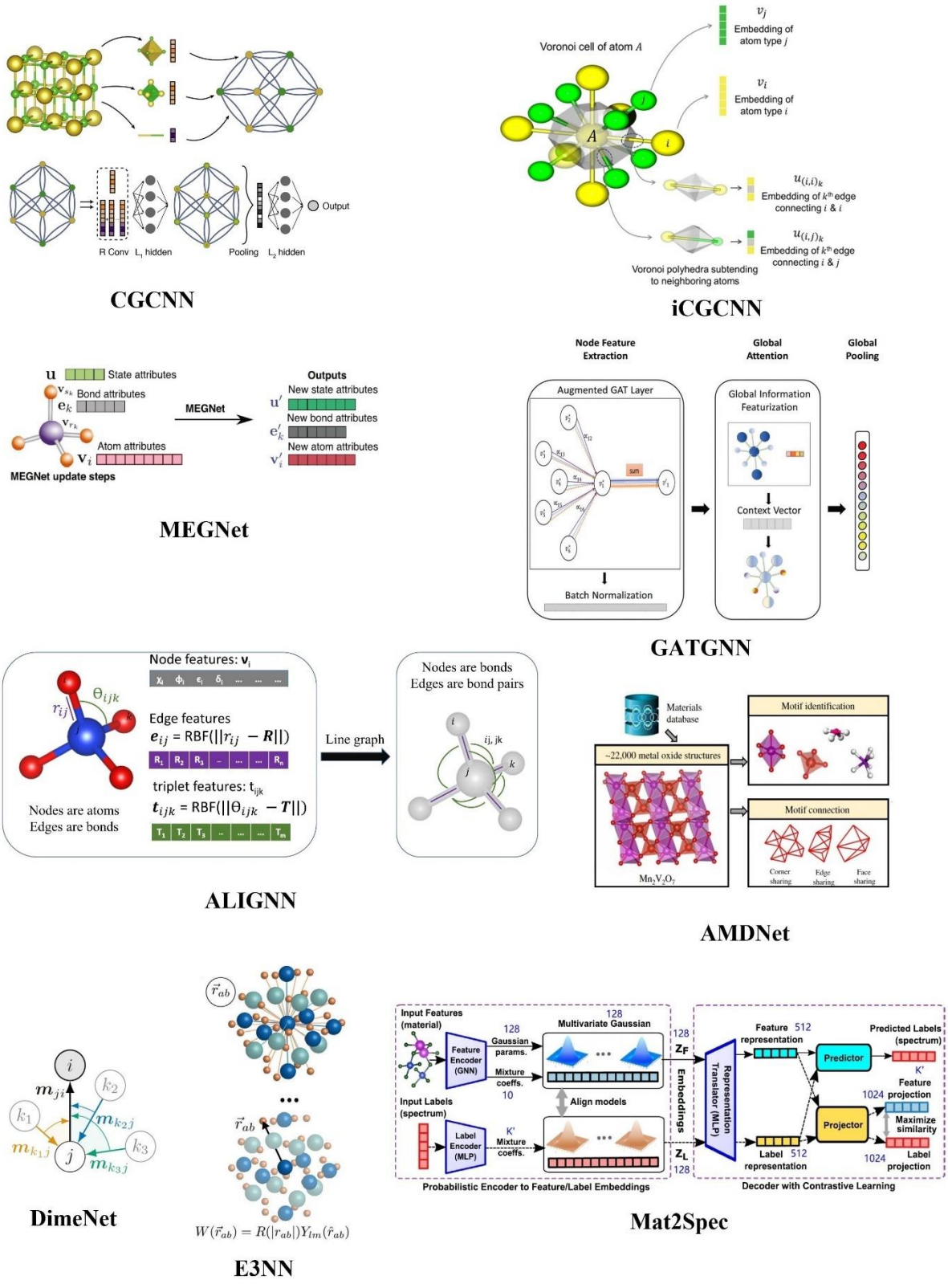

Figure 1-3. Illustration of nine graph neural networks designed for materials with their key innovations. **CGCNN**[71]: crystals converted to graphs with atoms as nodes and bonds as edges. **iCGCNN**[75]: Voronoi neighbors and Voronoi tessellation for edges. **MEGNet**[114]: inclusion of state attributes. **GATGNN**[115]: local and global attention mechanism.

ALIGNN[69]: updating bond angle representations by line graph. **AMDNet**[63]: extraction and use of structure motif information. **DimeNet**[116]: directional message passing. **E3NN**[77]: direct encoding of bond vector by kernels equivariant to 3D translations, rotations and inversion. **Mat2Spec**[76]: exploitation of correlations between spectral properties by probabilistic encoding and contrastive learning.

1.3.2. Datasets of materials

Data is at the heart of data-driven machine learning models. Currently, there are many materials databases containing compositions, structures, and various properties of materials, from both experiments and simulations. Especially, with the development of high-throughput screening, several large computational materials databases have been developed, such as the Materials Project (MP)[6], Open Quantum Materials Database (OQMD)[7], the Automatic Flow of Materials Discovery Library (AFLOW)[8], and the Joint Automated Repository for Various Integrated Simulations (JARVIS)[9]. More materials databases are provided in Ref.[10-14]. Despite the availability of large materials databases, standardization of materials datasets is still under progress for using these databases collectively[122]. Moreover, for machine learning applications in materials science, there are two main limitations on materials datasets: lack of high-quality data and biased dataset.

Lack of high-quality data. Although there are many large computational materials datasets with more than 10^5 data points, these large datasets are mainly based on cheap computation methods, such as DFT with Generalized Gradient Approximation (GGA)[123], and molecular dynamics with classic potentials[124, 125]. Because of the low speed of manual experiments, and the cost-accuracy trade-off of computations, there still lacks high-quality data of materials properties from experiments and expensive computation methods. For example, in the Materials Project database, there are ~10^5 data points of formation enthalpies and band gaps of materials based on GGA, while there are only ~10^3 and ~$3*10^3$ data points of experimentally measured formation enthalpies[97] and band gaps[66] of materials. Another example is, for

lattice thermal conductivity, there are ~$3*10^3$ data points from a semi-empirical formula[126], while there are only ~100 data points from experiments[55]. In addition to experimental data, computational data from expensive methods is also much less than that from cheap methods. For example, there are only ~10^4 data points of formation enthalpies[127] and band gaps[128] from methods more accurate than GGA, and in the database of Li transport behavior in polymer electrolyte[125], the number of properties from long simulations (50 ns) is only ~10% of that from short simulations (5 ns).

Biased datasets. In addition to the size of dataset, whether a dataset covers the materials space widely and evenly is also critical to the usefulness of the dataset to machine learning applications. For molecules, the golden standard QM9 dataset[129] is reported to underrepresent some types of molecules, which contributes to the presence of outliers in predictions[130]. Despite the lack of similar systematic study about distribution of materials in major materials datasets, there are four types of bias in some materials datasets. The first type is bias of presence of elements, such as the bias to oxides in the dataset of solid-state Li-ion conductors[131]. The second type is the bias of number of elements, such as the bias to binary and ternary compounds in the dataset of experimental formation enthalpy[97]. The third type is bias of structure motifs, such as the dataset of Li transport in polymer electrolyte where monomers with aromatic rings are excluded. The fourth type is bias of size of primitive cells, such as the dataset of HSE band gaps[128] where primitive cells with more than 40 atoms are excluded.

1.4. Problem statement and thesis overview

As discussed in Chapter 1.3, currently, machine learning models cannot provide accurate predictions for some problems of materials science. In this thesis, we aim to propose a series

of strategies to improve performance of machine learning models for materials, and apply these strategies to tackle machine learning tasks for realistic materials science problems. Figure 1-4 illustrates the series of strategies in this thesis. With the initial dataset and initial machine learning models, if the prediction is not satisfying, then one should first consider designing more suitable algorithms for the specific learning tasks. On the one hand, one should consider whether the representations of materials capture all the necessary information to determine the output. On the other hand, if there are datasets relevant to the learning tasks, which is very common in materials science, then one should consider transfer information from the relevant datasets to facilitate the learning tasks. More discussions about choosing and tuning general machine learning architectures are provided in Ref.[10, 11, 22]. If algorithm-design alone cannot lead to satisfying prediction, then one might consider expanding the dataset. For better sampling efficiency and mitigation of bias, one might consider sampling the next materials to characterize by active learning or Bayesian Optimization, and for higher speed and lower cost, one might consider accelerating materials characterization by machine learning-accelerated experiments and computations.

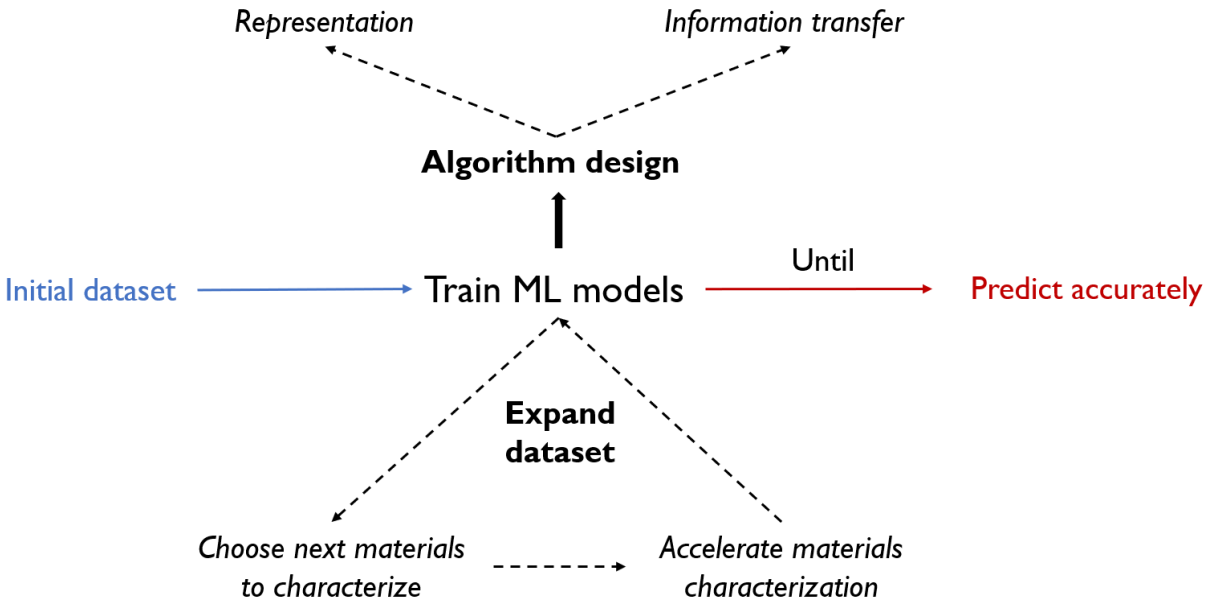

Figure 1-4. Illustration of strategies to improve performance of machine learning models for

materials science.

In Chapter 2, we will study whether the current deep representation learning models can capture knowledge of crystalline materials behind human-designed descriptors, and we will propose a way to improve the prediction performance by hybridizing deep representation learning and human-designed descriptors[132]. In Chapter 3, as a case study of information transfer, we will study how multi-fidelity learning and transfer learning, two information transfer strategies, help to learn the experimentally measured formation enthalpies of materials[97]. In Chapter 4, also as a case study of information transfer, we will use machine learning to study lattice thermal conductivity of materials, and investigate how transfer learning helps to learn the experimentally measured lattice thermal conductivity[55]. In Chapter 5, as an example of choosing the next materials to characterize, we will use Bayesian Optimization to search for the optimal laser-processing parameters for poly(acrylonitrile)[133]. In Chapter 6, as an example of accelerating materials characterization, we will propose a way to predict charge density distributions of materials by graph neural networks[73].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

2. Examining graph neural networks for inorganic crystalline structures

2.1. Introduction

Historically, materials informatics has relied on human-designed descriptors of materials structures. In recent years, graph neural networks (GNNs) have been proposed to learn the representations of crystal structures from the data end-to-end producing vectorial embeddings that are optimized for downstream prediction tasks. However, a systematic scheme is lacking to analyze and understand the limits of GNNs for capturing crystal structures. In this chapter, we propose to use human-designed descriptors as a bank of human knowledge to test whether black-box GNNs can capture knowledge of crystal structures. We find that current state-of-the-art GNNs cannot capture periodicity of crystal structures well, and we analyze the limitations of the GNN models that result in the failure from three aspects: local expressive power, long-range information, and readout function. We propose an initial solution, hybridizing descriptors with GNNs, to improve the prediction of GNNs for materials properties, especially phonon internal energy and heat capacity with 90% lower errors, and we analyze the mechanisms for the improved prediction. All the analysis can be easily extended to other deep representation learning models, human-designed descriptors, and systems such as molecules and amorphous materials.

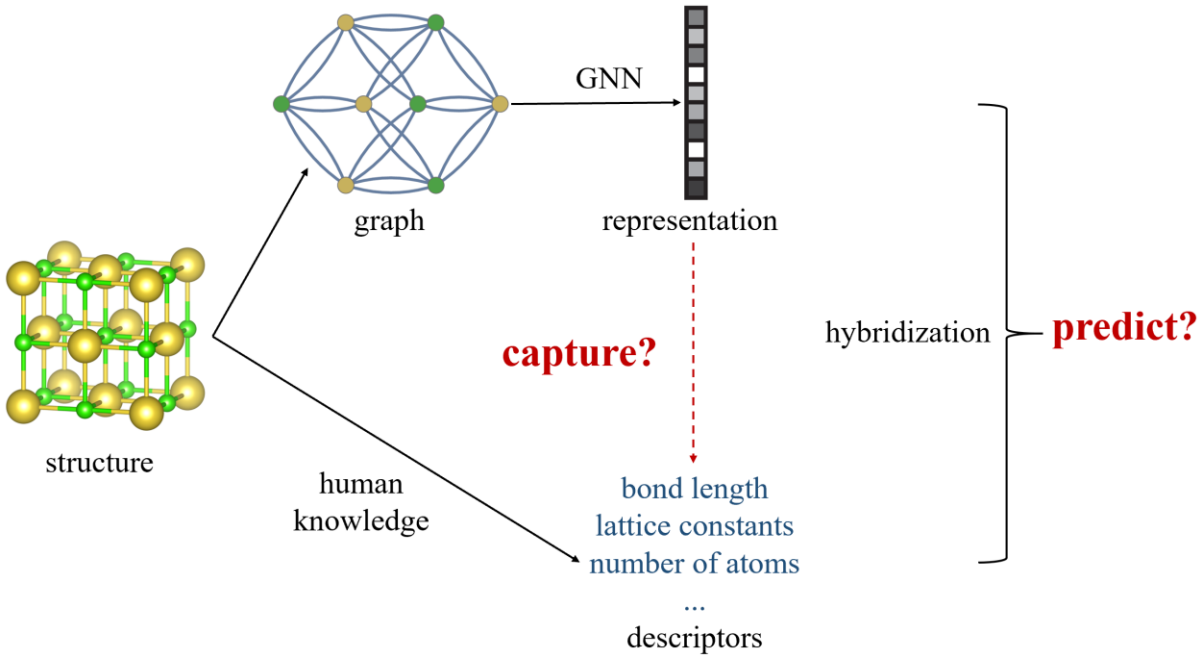

Figure 2-1. Schematic of analyzing whether GNN can capture human knowledge behind human-designed descriptors, and whether hybridization of GNN and human-designed descriptors can improve prediction performance.

Recently, machine learning (ML) has been widely employed to high-throughputly predict properties of materials[25, 46-77]. Conversion of crystalline structures into machine-readable numerical representation is one of the most critical tasks for applications of ML in materials science[11]. Since materials can be intuitively represented as graphs, where atoms forming the nodes and bonds the edges, graph neural networks (GNN) have become the state-of-the-art deep representation learning method for materials science. SchNet[57] and CGCNN[71] are two classic GNN architectures designed for materials. They update the representations of each atom by neighboring atoms and bond length between atoms, and pool all the updated atom representations into an overall representation of each material. In later variants of GNN such as iCGCNN[75], MEGNet[114] and GATGNN[115], bond representations are also updated during the convolution. To explicitly encode many-body interactions, Gasteiger *et al.* proposed DimeNet[116] and GemNet[117] for molecules, and Choudhary *et al.* proposed ALIGNN for periodic materials[69]. Very recently, Batatia *et al.*[121] proposed a general formalism to

encode local atomic environments by GNN with arbitrary body-order. Other efforts have also been made to improve GNN for crystal structures, such as inclusion of state attributes in MEGNet[114], attention mechanism in GATGNN[115], representations equivariant to rotations in E3NN[77, 134], use of structure motifs in AMDNet[63], prediction of tensorial properties in ETGNN[78], and exploitation of correlations in spectral properties in Mat2Spec[76].

Although these variants of GNNs have achieved some success in learning materials' properties, there is still no systematic approach and quantitative metric to analyze and understand the limitations of GNNs for capturing crystal structures, especially for global information of crystal structures. Moreover, the two methods of converting crystal structures into numbers, human-designed descriptors and deep representation learning, are now developed separately, not synergically.

In this chapter, we propose a systematic approach to analyze and quantify the limitations of GNNs for crystal structures, and propose a way to improve the GNNs models for predicting materials properties. As illustrated in Figure 2-1, we use the human-designed descriptors as a bank of human knowledge to test whether the current GNN models can capture certain knowledge about crystal structures. We test the GNNs by employing them to learn and predict the human-designed descriptors, and use the prediction accuracy as a quantitative metric for evaluation. The underlying assumption is that, if the model can accurately predict the descriptor, then the model can capture the knowledge behind the descriptor, otherwise the model may not be able to capture certain pieces of information about crystal structures. We find that the GNNs do not capture the periodicity of crystal structures well, and we analyze the reasons for this failure in some detail. We further hybridize the deep learning models with the human-designed descriptors, and test the descriptors-hybridized models on a range of important materials

properties. We find that hybridization of GNNs and descriptors can result in up to 90% decrease of errors for predictions of phonon-related properties compared with original GNNs.

In this chapter, we choose CGCNN and ALIGNN as two examples of GNNs to investigate their ability to capture human-designed descriptors, and as examples for improving prediction ability by hybridization with descriptors. CGCNN is one of the classic and most frequently used GNNs for materials, while ALIGNN is one of the state-of-the-art models for prediction of materials properties with the best performance on its in-house test set[69] and the open Matbench test set[113]. Both CGCNN and ALIGNN are specifically designed for predicting materials properties and have well-documented open-source codes to use and adapt. CGCNN explicitly encodes two-body interactions, and ALIGNN explicitly encodes three-body interactions. Although there are already GNN models that explicitly encode *n*-body interactions ($n \geq 4$)[118-120], they are not specifically designed for prediction of properties of periodic crystal structures and lack a comprehensive benchmark yet, and are thus not examined in this chapter.

The architecture of CGCNN (https://github.com/txie-93/cgcnn) is summarized in equations (2-1) to (2-3).

$$a_i^{(n+1)} = a_i^{(n)} + \sum_{j,k} \sigma(m_{(i,j)_k}^{(n)} \boldsymbol{W}_{gate}^{(n)}) \odot g(m_{(i,j)_k}^{(n)} \boldsymbol{W}_{message}^{(n)}) \quad \text{..... (2-1)}$$

$$m_{(i,j)_k}^{(n)} = a_i^{(n)} \oplus a_j^{(n)} \oplus b_{(i,j)_k} \quad \text{..... (2-2)}$$

$$\text{Output} = \text{AGG}(a_1^{(n^*)}, a_2^{(n^*)}, \dots, a_N^{(n^*)}) \quad \text{..... (2-3)}$$

Here, $a_i^{(n)}$ denotes the representation of atom *i* at layer *n*, $b_{(i,j)_k}$ representation of the k^{th} bond between atom *i* and *j* at layer *n*, n^* the final convolution layer, $\boldsymbol{W}_{gate}^{(n)}$ the gate matrix at layer *n*, $\boldsymbol{W}_{message}^{(n)}$ the message matrix at layer *n*, $m_{(i,j)_k}^{(n)}$ the message from atom *j* to atom *i* via the

k^{th} bond, $\odot$ element-wise multiplication, $\oplus$ concatenation, σ the sigmoid function, g non-linear activation function, AGG the aggregation function. In CGCNN, the default aggregation function can be written as:

$$\text{Output} = \text{FCN}(\frac{1}{N_a}\sum_{i=1}^{N_a} a_i^{n^*}) \text{ ..... (2-4)},$$

where the output is calculated by first taking the average of all atom representations, then feeding the averaged representations to a fully connected network. Briefly, CGCNN uses neighboring atoms and bond length as messages to each atom, and update each atom representation by feeding the messages into a gate layer and a message processing layer. After convolutions, CGCNN pools all atom representations by taking the average and input the pooled materials representation into a fully connected network to compute the property.

The architecture of ALIGNN (https://github.com/usnistgov/alignn) is summarized in equations (2-5) to (2-10), with equations (2-5) to (2-7) describing the atomistic graph, and equations (2-8) to (2-10) the line graph.

$$a_i^{(n+1)} = a_i^{(n)} + g(a_i^{(n)}\boldsymbol{W}_{self}^{(n)} + \sum_{j,k} g'(b_{(i,j)_k}^{(n)})\boldsymbol{W}_{message}^{(n)} a_j^{(n)}) \text{ ..... (2-5)}$$

$$b_{(i,j)_k}^{(n+1)} = b_{(i,j)_k}^{(n)} + g(m_{(i,j)_k}^{(n)}\boldsymbol{W}_{gate}^{(n)}) \text{ ..... (2-6)}$$

$$m_{(i,j)_k}^{(n)} = a_i^{(n)} \oplus a_j^{(n)} \oplus b_{(i,j)_k}^{(n)} \text{ ..... (2-7)}$$

$$b_i^{(n+1)} = b_i^{(n)} + g(b_i^{(n)}\boldsymbol{W}_{self}^{\prime(n)} + \sum_{j,k} g'(t_{(i,j)_k}^{(n)})\boldsymbol{W}_{message}^{\prime(n)} b_j^{(n)}) \text{ ..... (2-8)}$$

$$t_{(i,j)_k}^{(n+1)} = t_{(i,j)_k}^{(n)} + g(m_{(i,j)_k}^{\prime(n)}\boldsymbol{W}_{gate}^{\prime(n)}) \text{ ..... (2-9)}$$

$$m_{(i,j)_k}^{\prime(n)} = b_i^{(n)} \oplus b_j^{(n)} \oplus t_{(i,j)_k}^{(n)} \text{ ..... (2-10)}$$

Here, t denotes the representation of bond angle, and other symbols share similar meaning to

that of CGCNN. In summary, in each convolution layer, ALIGNN updates atom representations by neighboring atoms and bonds, updates bond representations twice: by connected atoms, and by neighboring bonds and bond angles, and update bond angle representations by connected bonds. After the convolutions, similar to equation (2-4), ALIGNN uses average pooling as the default setting to collect atom representations as the material representation, and calculate property by a feed-forward network.

For building periodic crystal graphs, in their default settings, both CGCNN and ALIGNN use a cut-off radius of 8 Å for 12 nearest neighbors, and both of them use radial basis functions to expand the interatomic distances for initialization of bond representations. ALIGNN also uses radial basis functions to expand cosines of bond angles for initialization of bond angle representations. CGCNN updates atom features by 3 graph convolution layers, and ALIGNN updates atom features by 4 line graph convolution layers (equations (5) to (10)) and 4 normal graph convolution layers (equations (5) to (7)). In Chapter 2.2, we use CGCNN and ALIGNN with the default setting unless otherwise specified.

2.2.Learning and predicting human-designed descriptors

In this chapter, we employ CGCNN and ALIGNN to learn and predict structural descriptors of a subset of crystal structures in the Materials Project database[6] (“MP dataset” as below; details in Chapter 2.5) to examine the ability of the GNNs to capture certain knowledge behind the descriptors. As a baseline, we also use ROOST[23], one of the most powerful composition-only deep learning models, to learn and predict the structural descriptors.

In Figure 2-2a, we show the accuracies of predictions of some of the most basic local structural descriptors calculated by matminer[135] from CGCNN, ALIGNN, and ROOST in

terms of R^2 scores ($R^2 = 1 - \frac{\sum(y_i - y_{i,true})^2}{\sum(y_{i,true} - \bar{y})^2}$, y_i predicted value, $y_{i,true}$ true value, $\bar{y}$ mean of true values). We can see that, for most local structural descriptors, both CGCNN and ALIGNN can properly predict them with R^2 scores close to or higher than 0.8, and both of the two structure-based models outperform the composition-only model (ROOST). Because local descriptors in this chapter are essentially statistics of local environments around each atom, the explicit encoding of bond angles (three-body interaction) in ALIGNN might explain why ALIGNN outperforms CGCNN for learning local structural descriptors as in Figure 2-2a. The cases with lower R^2 scores in Figure 2-2a, such as max_rela_bond_len (maximum relative bond length) and std_avg_bond_ang (standard deviation of average bond angles), can be attributed to the fact that, average pooling (equation (4)) is used by both CGCNN and ALIGNN to obtain the mean statistics of atom representations, while the two descriptors here describe the maximum and standard deviation of a collection of atomic environments.

In addition to basic local descriptors, we also test the ability of CGCNN and ALIGNN to capture knowledge behind more global structural descriptors. In Figure 2-2b, we show the accuracies of predictions of some of the most basic global structural descriptors calculated by matminer[135] and pymatgen[136] from CGCNN, ALIGNN, and ROOST. Both CGCNN and ALIGNN can predict density, vpa (volume per atom), packing fraction, and natoms (number of atoms in the primitive cell; in this chapter, the "primitive cell" is defined as the Niggli reduced cell[137, 138]) with R^2 scores close to or higher than 0.8. However, they cannot predict struct_comp_cell (structural complexity per cell[139]) and lattice constants (a, b, c, α, β, γ; in this chapter, a denotes the length of the longest lattice vector, c the shortest, and α denotes the largest lattice angle, γ the smallest) well. Both structure-based models outperform the composition-only model, and ALIGNN outperforms CGCNN, except for α and γ.

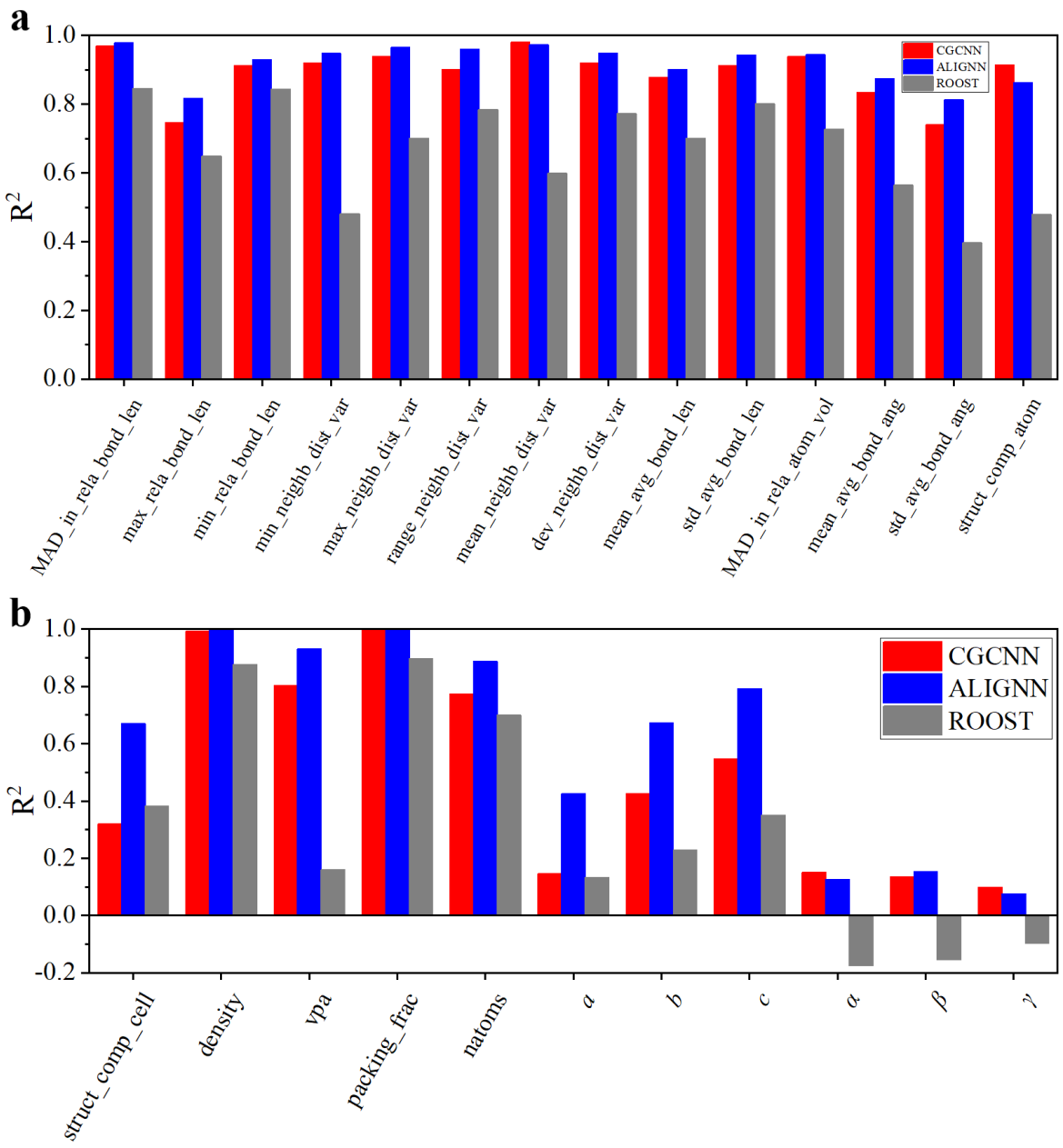

Figure 2-2. Learning and predicting human-designed descriptors to examine whether the GNNs can capture certain human knowledge. **a** and **b** R^2 scores of predictions of human-designed structural descriptors from CGCNN, ALIGNN and ROOST for local and global structural descriptors, respectively. The full names of the descriptors are listed in Table 1.

2.3.Limitations of GNN for capturing periodicity

Although previous works have suggested that lattice constants of crystal structures are learnable based on only compositions[140, 141], the results in this chapter show that even with

structures as input, CGCNN and ALIGNN cannot capture lattice constants well. In this chapter, we analyze the possible reasons for such failure and obtain insights for improving GNNs for crystal structures.

Lattice constants describe the periodicity of atomic structures. If $A(r)$ describes the type of atom at position r ("none" if there is no atom at that position), and if R is a linear combination of lattice vectors, then periodicity requires that:

$$A(r) = A(r + R) \text{.....(2-11).}$$

In 3-dimensional (3D) space, we need 3 linearly independent lattice vectors to describe the periodicity of atomic structures. Lattice constants describe the periodicity by the lengths of lattice vectors (a, b, c) and angles between lattice vectors (α, β, γ). To simplify the analysis, in addition to 3D crystal structures in the MP dataset, we also consider the toy cases of quasi-1D atomic chains as in Figure 2-3a, where periodicity is imposed only along the *x* direction and no constraint is imposed along the other two directions. In this quasi-1D space, we only need the length of the lattice vector (a) to describe the periodicity: $A(r) = A(r + a)$.

For GNNs with average pooling in equation (4), since they use the average local atomic environments to represent the atomic structures, they capture periodicity by learning how equation (2-11) affects the local atomic environments within the receptive fields of atoms in the GNNs. The receptive field of each atom describes the range of the space where information can be propagated to the atom through the GNNs, and it depends on the number of neighbors each atom can connect to and the number of convolution layers in the GNNs:

$$\text{range of receptive field} \propto \text{number of neighbors} * \text{number of convolutions} \text{.....(12).}$$

If the length of the periodicity (length of lattice vector) is smaller than the lengths of the receptive fields of atoms in the GNNs, then the GNNs might be able to capture the short

periodicity; however, if the length of the periodicity is larger than the lengths of the receptive fields of atoms, then in principle the GNNs cannot capture the long periodicity. For example, as in Figure 2-3a, if the periodicity is short, such as the top red arrow which requires that atom 1 and atom 3 (atom *n* and atom *n*+2) have the same type and coordinates in the *y* and *z* directions, then the local atomic environment input to atom i is constrained by such periodicity, and the GNNs might be able to capture the constraint and periodicity. However, if the periodicity is long, such as the bottom red arrow describing that the periodicity is imposed between atom 1 and atom *N*+1 (one atom beyond the receptive field), then there is no constraint inside the receptive field of atom *i*, and the GNNs cannot capture the long constraint and periodicity.

To analyze the behaviors of GNNs on capturing periodicity, in this section, we introduce toy datasets of quasi-1D carbon chains as illustrated in Figure 2-3b ("1D dataset" as below; details in Chapter 2.5), and we create two versions of the 1D datasets: a short dataset where the periodicity of each chain is shorter than the receptive fields of atoms (1D, short), and a long dataset where the periodicity is longer than the receptive fields (1D, long). We use the default CGCNN to learn and predict the length of lattice vector (a) of the two datasets, and in Figure 2-3e and 2-3f, we show the predicted a versus true a of the two datasets. We can see that, for the short chains, CGCNN can predict a well with the R^2 score larger than 0.8, while for the long chains, CGCNN cannot predict a well. The prediction results of a of the quasi-1D carbon chains support our analysis above that GNNs might be able to capture short periodicity while hard to capture long periodicity.

Although the periodicity of most short chains in this chapter can be properly learned as in Figure 2-3e, theoretically, GNNs with limited local expressive power are not able to fully determine the periodicity. Since Chen *et al.*[142] have proved the equivalence between the ability of GNNs to distinguish graphs and approximate graph functions, if a GNN cannot

distinguish two atomic graphs with different periodicity, then the GNN cannot fully determine the graph function describing the periodicity. In Figure 2-3c, we show two cases of 1D chains: a 1D zigzag chain and a 1D armchair chain, which represent structure prototypes of some real crystal structures such as organic crystals[143] and metal chalcogenides[144]. If a GNN uses only diatomic distances to encode local atomic environments (such as CGCNN), and if the GNN only connect to the nearest neighbors (**1** to **2**), then the GNN cannot distinguish different zigzag and armchair 1D chains with the same bond length but different bond angles and cannot capture the angle dependence of a. If the GNN can connect to the second nearest neighbors (**1** to **3**), then the GNN is able to distinguish zigzag and armchair 1D chains with different bond angles; however, it is still not able to distinguish between zigzag and armchair chains with the same bond length and bond angle. The analysis suggests that, to improve the ability of GNNs to capture short periodicity, it might be helpful to increase the local expressive power of GNNs to distinguish structures with different periodicity.

In Figure 2-3g, we show the effects of number of convolution layers and number of neighbors of CGCNN on capturing periodicity of 1D chains and 3D crystal structures. As in equation (2-12), both increasing number of convolution layers and increasing number of neighbors extend the receptive fields of atoms in CGCNN, and as the discussion of zigzag and armchair chains above, increasing number of neighbors can lead to higher local expressive power to distinguish graphs. From Figure 2-3g, we can see that for short chains, increasing the number of neighbors leads to better prediction of a, which supports our suggestion above that improving the local expressive power can help to capture short periodicity, while increasing number of convolution layers results in worse prediction of a, which might be because deeper GNNs are harder to train[145, 146]. For long chains, both increasing number of neighbors and number of convolution layers result in better prediction of a, indicating that extending the receptive fields of atoms in CGCNN can help to capture long periodicity. As for lengths of

lattice vectors of real 3D structures in the MP dataset (mixed with short and long structures), we can see that both increasing number of neighbors and number of convolution layers lead to better prediction of a, b, c. However, we find that increasing number of convolution layers by 133% and number of neighbors by 50% just lead to moderate improvement of prediction of a, b, c. Since the cost of graph convolution operations is proportional to number of convolution layers and neighbors, we suggest that simply increasing number of convolution layers and neighbors might not be an ideal way to improve the ability of GNNs to capture periodicity.

The analysis above is based on average pooling in equation (2-4). If we use sum pooling in equation (2-13) with size extensibility:

$$\text{Output} = \sum_{i=1}^{N_a} \text{FCN}(a_i^{n^*}) \quad \text{..... (2-13)},$$

then the GNNs capture periodicity by summing contributions of each atom to the lattice vectors. In Figure 2-3h, we show the R^2 scores of predictions of a of the 1D chains and natoms, a, b, c of the MP dataset from CGCNN with average pooling and sum pooling, respectively. For 1D short chains, sum pooling can lead to better prediction of a than average pooling, which might be explained by the fact that sum pooling is more expressive than average pooling[147]. For 1D long chains, sum pooling can result in significantly better prediction of a than average pooling, because average pooling requires that each atom encodes information from one end of the long primitive cell to the other end to capture the structural constraint imposed by the periodicity, while sum pooling needs only local contributions of each atom to the lattice vectors. Consistent with the results of a of the 1D chains, for natoms, a, b, c of the MP dataset, sum pooling can also result in better prediction than average pooling. The stronger ability of sum pooling to capture periodicity might lead to better prediction of extensive materials properties, and in Ref.[132], we show that sum pooling can provide better prediction than average pooling for phonon internal energy (U), phonon heat capacity (C_v) and magnetization (M).

Despite the improvement, we suggest that sum pooling is not an ideal solution to the challenge of capturing periodicity. Periodicity and lattice constants of the primitive cells do not scale with supercell size and are intensive characteristics of crystal structures. In principle, sum pooling cannot be employed in machine learning of materials' intensive properties due to the requirement of (supercell) size invariance[71]. The improvement of sum pooling over average pooling in Figure 2-3h is based on the fact that primitive cells of crystals are used as input to the GNNs in this chapter. Even if only primitive cells are input to the GNNs, sum pooling might also fail to capture periodicity in some cases, as periodicity does not always scale with the number of atoms in the primitive cells. For example, in Figure 2-3d we show the case of 1D double chains. Compared with 1D single chains in Figure 2-3b and 2-3c, 1D double chains can have similar periodicity but twice number of atoms. In Ref.[132], we show that, compared with the datasets with only 1D single chains, sum pooling is less powerful to capture the periodicity of the datasets mixed with 1D single and double chains.

From Figure 2-2b, we can see that ALIGNN outperforms CGCNN in the prediction of natoms, a, b, c of the MP dataset. This improved predictive ability could result from two factors: on the one hand, ALIGNN has stronger local expressive power than CGCNN as it explicitly encodes bond angles, and on the other hand, ALIGNN has a larger receptive field than CGCNN, as in each convolution layer in CGCNN, a node receives messages only from the first shell of bonds and neighbors in equation (2-2), while in each convolution layer in ALIGNN, a node also receives messages from the second shell of bonds in equation (2-10). Although with the default settings ALIGNN has 8 convolution layers while CGCNN has only 3 convolution layers, from Figure 2-3g we can see that increasing the number of convolution layers of CGCNN to 8 leads to only moderate improvement and cannot make the predictions of a, b, c from CGCNN as accurate as that of ALIGNN, which shows that different number of convolution layers in CGCNN and ALIGNN with the default settings is not a critical factor on

their relative ability to capture periodicity of the MP dataset.

In Figure 2-2b, we show that both CGCNN and ALIGNN cannot learn the lattice angles of the primitive cell well, and sum pooling, more convolutions, and more neighbors do not improve the prediction. Here we partially attribute these results to the artificial choice of lattice angles. More discussions regarding the determination of the primitive cell are provided in Ref.[132]. In this chapter, we choose the set of six parameters $(a, b, c, \alpha, \beta, \gamma)$ as a widely used rotationally invariant representation of lattice vectors, which might add artificial difficulty to the learning of periodicity. For example, in addition to the problems associated with learning and prediction of a in the 1D cases as above, for learning and prediction of the length of the longest lattice vector of 3D structures the GNNs need to first identify which dimension is associated with the largest length, then determine the largest lattice length. For fairer evaluation, it is necessary to develop representations of periodicity that are equivariant to rotations to avoid this additional difficulty.

According to MLatticeABC[140] and CRYSPNet[141], lattice constants of high-symmetry materials are reported to be learnable based on only compositions of materials, while here we show that lattice constants are not learnable by the GNNs even with structures as input. In the previous works, materials with different symmetry are learned separately, and lattice constants of high-symmetry materials are reported to be more learnable than that of low-symmetry materials, while in this chapter the MP dataset is mixed with different symmetries and is biased to materials with low symmetry. More details about the MP dataset are provided in the Chapter 2.5.

In this chapter, we discuss the limitations of the GNNs on capturing periodicity mainly in three aspects: limited local expressive power, difficulty of capturing long-range information beyond receptive fields of atoms, and average pooling as the readout function. For local

expressive power, advancements of GNNs to capture more structural characteristics, such as ALIGNN-*d* for dihedral angles[148] and equivariant representations for orientation of bond vectors[134, 149], might be helpful to better capture periodicity of structures with lattice vectors shorter than the receptive fields. For long-range information, on the one hand efforts to train very deep GNNs effectively and efficiently, such as DeeperGATGNN[150], are helpful to extend the receptive fields of atoms. On the other hand, the idea of topological message passing[151, 152] might be useful to capture long-range information by connecting nodes in the same cell complex that are far from each other, and the idea of Implicit Graph Neural Networks (IGNN)[153] might also be useful to bypass the problems associated with training very deep graph neural networks by obtaining implicitly defined state vectors from a fixed-point equilibrium equation. It is also necessary to further develop readout functions to collect the long-range information with size invariance, and the whole-graph self-attention based readout function used in GraphTrans[154] might be a good starting point to collect global information of crystals.

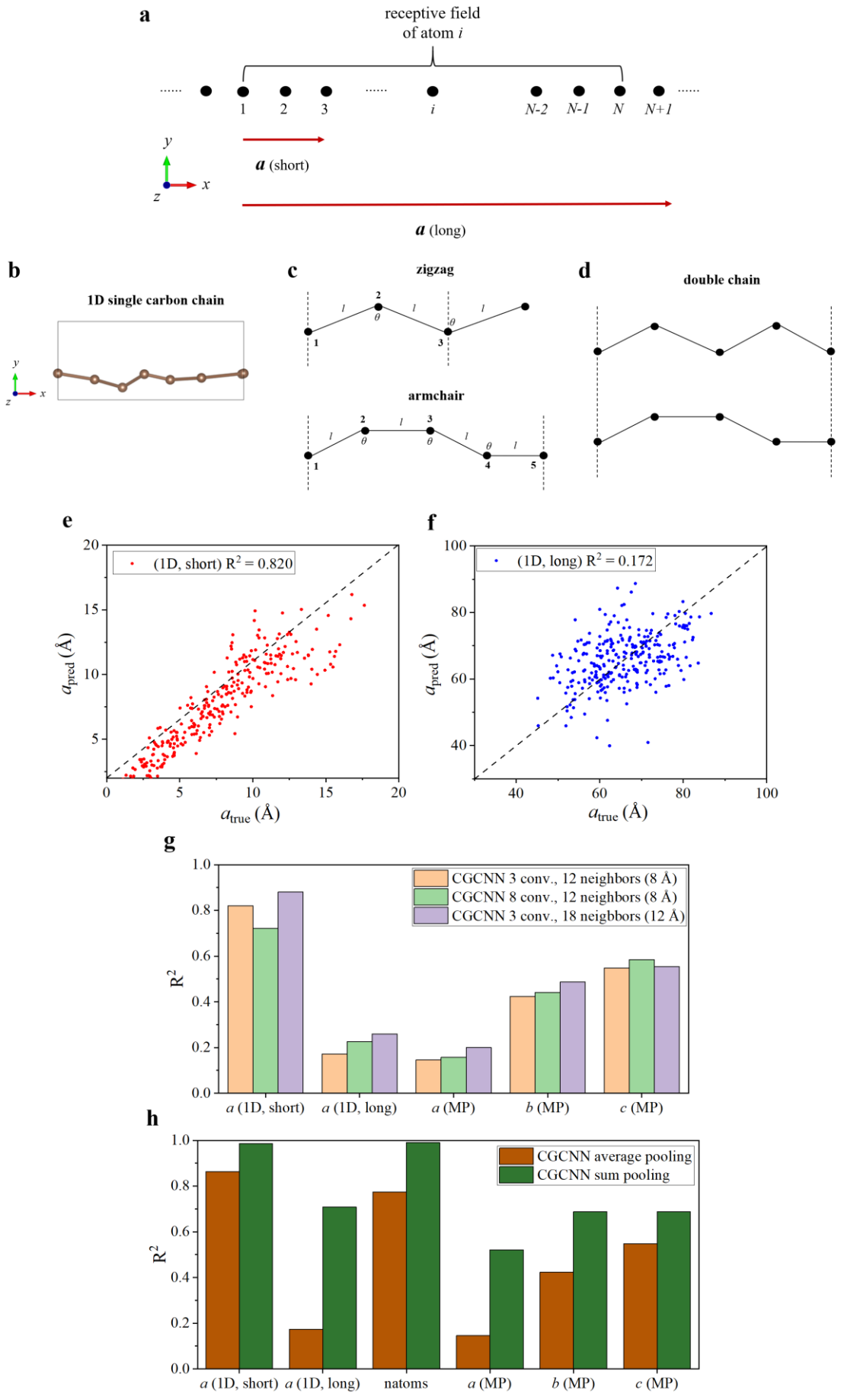
a
receptive field
of atom i
1
2
3
i
N-2
N-1
N
N+1
a (short)
a (long)
b
1D single carbon chain
c
zigzag
armchair
d
double chain
e
(1D, short) R² = 0.820
a_pred (Å)
a_true (Å)
f
(1D, long) R² = 0.172
a_pred (Å)
a_true (Å)
g
CGCNN 3 conv., 12 neighbors (8 Å)
CGCNN 8 conv., 12 neighbors (8 Å)
CGCNN 3 conv., 18 neigbbors (12 Å)
R²
a (1D, short)
a (1D, long)
a (MP)
b (MP)
c (MP)
h
CGCNN average pooling
CGCNN sum pooling
R²
a (1D, short)
a (1D, long)
natoms
a (MP)
b (MP)
c (MP)

Figure 2-3. Limitations of GNNs for capturing periodicity. **a** Illustration of the receptive field of an atom in a GNN and periodicity of a 1-dimensional (1D) structure. Here, atom *i* receives information from atoms 1 to *N*, and two cases of periodicity are plotted: the short periodicity from atom 1 to 3 and the long periodicity from atom 1 to *N*+1. **b** Illustration of 1D single carbon chains as toy structures. The chains are along the *x* direction with periodicity, with random displacement of each atom in the *y* and *z* directions. **c** Illustration of 1D chains with zigzag and armchair configuration, respectively. **d** Illustration of 1D double chain. **e** and **f** a_{true} versus a_{pred} of the datasets of 1D short chains and 1D long chains from default CGCNN, respectively. **g** R^2 scores of predictions of a of 1D short chains and 1D long chains, and a, b, c of the MP dataset, from default CGCNN, CGCNN with 8 convolution layers, and CGCNN connecting 18 nearest neighbors within 12 Å, respectively. **h** R^2 scores of prediction of a of 1D short chains and 1D long chains, and natoms, a, b, c of the MP dataset from CGCNN with average pooling and CGCNN with sum pooling, respectively.

2.4. Descriptors-hybridized deep representation learning

From the results of learning human-designed descriptors, we know that GNNs might not capture all knowledge behind human-designed descriptors. One way to overcome the issue is to design better GNN architectures for specific information, such as long-range information. Another way to overcome the issue is to input the missing knowledge into the deep representation learning models. Although this idea is straightforward and used in previous works[46, 155], such as the incorporation of lattice vectors in GeoCGNN[46], the previous works did not explain the role of the additional information with quantitative evidence. In this chapter we show the mechanisms of how inputting certain knowledge to GNNs improves prediction of materials properties, and we find that the hybridization with descriptors can lead to a large improvement for prediction of some materials properties, especially vibrational properties that largely depend on periodicity.

We construct the descriptors-hybridized graph neural networks as below:

$$\text{Output} = \text{FCN}\left(\frac{1}{N_a}\sum_{i=1}^{N_a} a_i^{n^*} \oplus \text{descriptors}\right) \quad \text{..... (2-14).}$$

In other words, we concatenate the vector of descriptors to the vector of learned representation, and input the hybridized representation vector to the fully-connected network.

In Figure 2-4a, we show the prediction results of descriptors-hybridized CGCNN and ALIGNN (de-CGCNN and de-ALIGNN) on 13 materials properties, with the full names of the abbreviations of properties in Table 2-1, and detailed errors in Ref.[132]. The set of properties includes final energy ($E_{\text{fin.}}$), band gap (E_{g}), bulk and shear modulus (K and G), lattice thermal conductivity (κ), phonon internal energy and heat capacity at 300K (U and C_{v}), Poisson ratio (ν), modulus of the piezoelectric tensor ($||e||_{\infty}$), electronic and total dielectric constant ($\varepsilon_{\text{e.}}$ and $\varepsilon_{\text{t.}}$), refractive index (n) and total magnetization (M). The errors of the machine learning models are presented using the metric $\text{MAE/MAD} = \frac{\sum |y_i - y_{i,true}|}{\sum |y_{i,true} - \bar{y}|}$, which is invariant to scaling and used in the ALIGNN paper[69]. Typically, a model with MAE/MAD smaller than 0.2 is considered a good predictive model[25, 69]. We can see that de-CGCNN has improved prediction performance for most properties compared with the original CGCNN, and de-ALIGNN has close-to or larger than 10% improvement for four properties (κ, U, C_{v}, and M) and similar performance for other properties compared with the original ALIGNN. Both de-CGCNN and de-ALIGNN outperform the descriptors-only model for all properties, regardless of whether CGCNN and ALIGNN outperform the descriptors-only model.

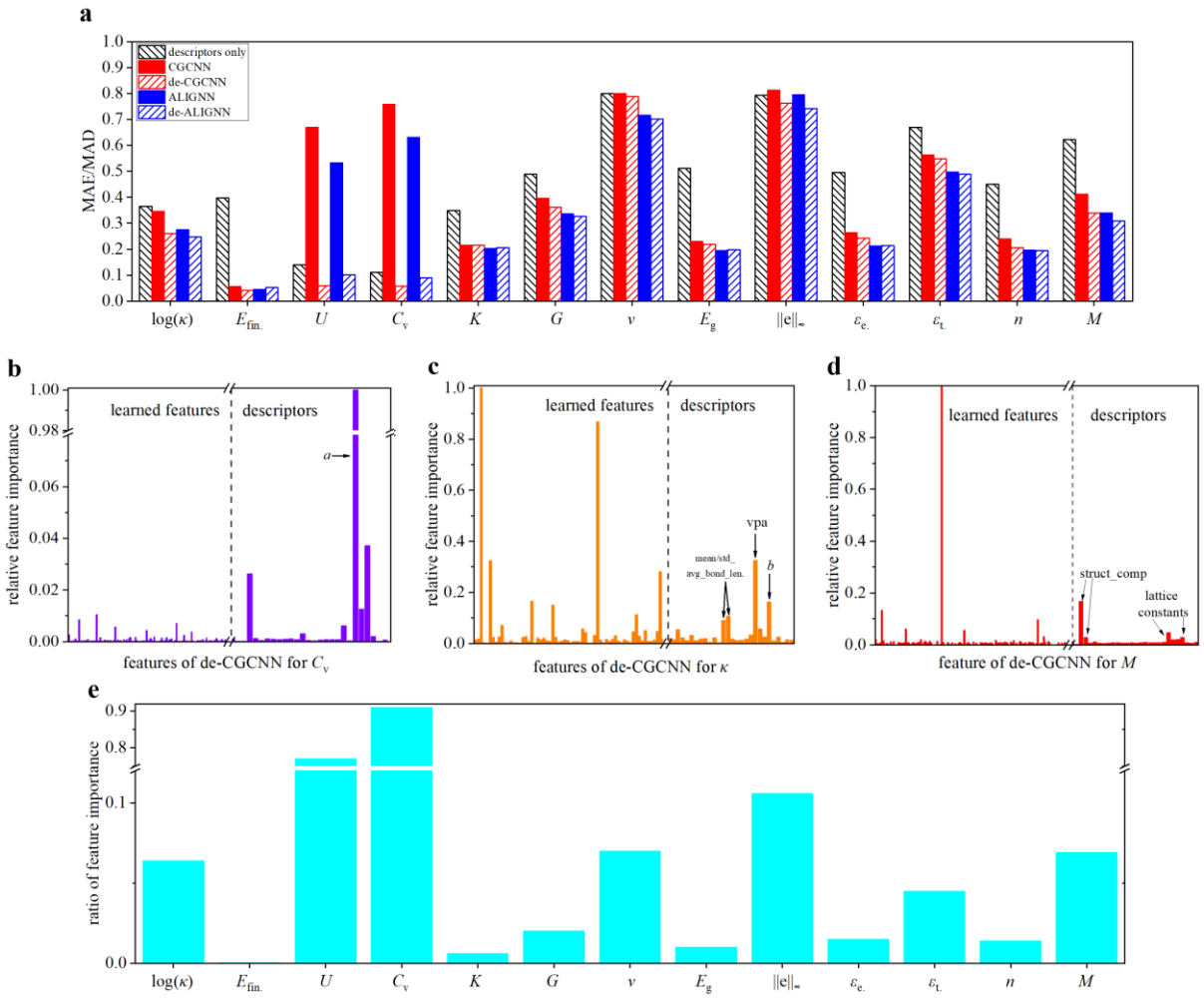

Figure 2-4. Prediction performance of descriptors-hybridized GNNs. **a** MAE/MAD ratio of prediction of 13 materials properties from machine learning models based only on descriptors, CGCNN, ALIGNN and their descriptors-hybridized version (de-CGCNN and de-ALIGNN). **b**, **c** and **d** Relative feature importance of representations from de-CGCNN for C_v, κ, and M, respectively. **e** Ratio of feature importance of input human-designed descriptors to the total feature importance from de-CGCNN for the 13 materials properties.

In Figure 2-4a, we observe that both de-CGCNN and de-ALIGNN have large improvement for the prediction of *U* and C_v, with around 90% lower errors compared with CGCNN and ALIGNN, respectively. To understand the improvement, we show the feature importance spectrum of de-CGCNN for prediction of C_v in Figure 2-4b. We can see that, the human-designed descriptors play important roles in learning C_v, with a being the most important feature, while the learned features are much less important. Therefore, the poor

prediction ability of CGCNN and ALIGNN for U and C_v can be partially explained by the fact that a is important to the two properties but CGCNN and ALIGNN cannot learn a well. The distribution of feature importance agrees well with the phenomenon in Figure 2-4a that, using machine learning model based only on human-designed descriptors can have much lower errors for prediction of U and C_v compared with GNNs. The improved performance for U and C_v also explains why sum pooling outperform average pooling in Figure 3b, as sum pooling can learn a better as in Figure 2-2e.

The importance of primitive cell-level information to U and C_v can be justified physically as below. Approximately, if we only consider the acoustic phonons (collective vibrations for all atoms in the primitive cell), according to the Debye model of density of states, the phonon internal energy (U) and heat capacity (C_v) of a specimen can be written as[156]:

$$U \approx 9Nk_BT\left(\frac{T}{\Theta}\right)^3 \int_0^{x_D} dx \frac{x^3}{e^x-1} \text{.....(2-15)},$$

$$C_v = \left(\frac{\partial U}{\partial T}\right)_v \approx 9Nk_B\left(\frac{T}{\Theta}\right)^3 \int_0^{x_D} dx \frac{x^4e^x}{(e^x-1)^2} \text{.....(2-16)},$$

$$x_D \equiv \frac{\Theta}{T} \text{.....(2-17)},$$

$$\Theta = \frac{\hbar v}{k_B}\left(\frac{6\pi^2N}{V}\right)^{\frac{1}{3}} \text{.....(2-18)},$$

where Θ is the debye temperature, N is the number of primitive cells in the specimen, V is the volume of the specimen, and v is the velocity of sound, which can be approximated by the first-order Hooke's law:

$$v \approx \sqrt{\frac{C}{m}}d \text{.....(2-19)},$$

where C is the effective spring constant, m is the mass of atoms in the primitive cell, and d

is the effective distance between atomic planes along the direction of vibration. Therefore, with the information of $\frac{N}{V}$, C, m, and d, we can estimate acoustic U and C_v per primitive cell at given T within the Debye model. Since the set of descriptors in this chapter includes density and lattice constants, the information of $\frac{N}{V}$, m, and d can be directly obtained by machine learning models from the input descriptors. For C, because it is related to the bonding strength, it can be estimated by the bond length-related descriptors. Consequently, machine learning models based on the set of descriptors in this chapter can approximate U and C_v well within the Debye model, which explains why machine learning model only based on descriptors outperforms CGCNN and ALIGNN in Figure 2-4a, as CGCNN and ALIGNN cannot estimate lattice constants well as in Figure 2-2b.

κ and M are another two properties with around 10% improvement for both de-CGCNN and de-ALIGNN. It is known that κ depends significantly on primitive cell-level information[157], and as shown in Figure 2-4c, some input descriptors, including b, are important to the prediction of κ. As for M, as shown in Figure 2-4d, some descriptors like structural complexity and lattice constants contribute to the prediction of M, which might explain why input of descriptors leads to improved prediction accuracy for M. In Figure 2-4e, we show the ratio of feature importance from the human-designed descriptors to the total feature importance from de-CGCNN. We can see that most properties without significant improvement in Figure 2-4a have low contributions from input human-designed descriptors, with the exception of v and $||e||_\infty$ where all the models perform poorly. The phenomenon that hybridization with descriptors has larger improvement for CGCNN than ALIGNN might be explained by the fact that, CGCNN learns these descriptors worse than ALIGNN as in Figure 2-2a, therefore hybridization with descriptors provides more missing information to CGCNN than ALIGNN.

In addition to providing missing information, hybridization with descriptors might also have

other impacts on the GNNs. In Ref.[132], we show that hybridization of descriptors can bias the learned representations less correlated with the input descriptors, although how such bias affects prediction performance is not clear yet. Other questions worth further investigation include, how the improvement scales with dataset size, and how to choose the set of input descriptors for optimal performance. It will also be important to understand if the two mentioned behaviors (scaling and selection of descriptors) are similar with or different from that of the descriptors-only models and pure deep representation learning models.

2.5. Details of methods

Datasets. In Chapter 2, we choose 25 (in Figure 2-2) human-designed descriptors to test their learnability to CGCNN and ALIGNN, and hybridize 29 descriptors (all descriptors in Table 2-1) with the two models to test the prediction performance. The list of descriptors is provided in Table 2-1, with descriptors after gamma included in the second task but not in the first task. The criterion for choosing the 25 descriptors in the first task is that they are easy to understand and easy to obtain from crystal structures, and the reason for not testing coordination number (CN) in the first task is that we know CN can be learned well given the definition of GNN, and the reason for not testing symmetry in the first task is that we know symmetry cannot be learned as lattice constants cannot be captured. Number of atoms and lattice constants of the primitive cell are determined by the Niggli reduction[137] implemented in the Structure class in pymatgen[136], and other descriptors are calculated by Matminer[135]. For the descriptor “standard deviation average bond length” (and similar descriptors), the calculation procedure is first calculating average bond length for each atom, then calculating the standard deviation for the average bond length of all atoms.

In Chapter 2, most crystal structures and materials properties are downloaded from the

Materials Project database (V2021.03.22)[6], and those for κ are from the TEDesignLab database[126]. U and C_v are calculated by the PhononDos class in pymatgen[136] based on the phonon density of states from the Materials Project database[6]. For machine learning of materials properties in Figure 2-3 and Figure 2-4, we split the datasets into 60%, 20% and 20% as the training, validation and test set.

For the dataset used for testing whether CGCNN and ALIGNN can capture human-designed descriptors of crystal structures, since we know that lattice constants of high-symmetry materials are reported to be more learnable than that of low-symmetry materials[140, 141] based on compositions, we create a subset of the Materials Project database by removing some structures randomly based on their space group number:

$$Probability(\text{removed}) = \frac{\text{Space group number}}{\text{Space group number+15}},$$

where 15 is the space group number of the C2/c group, the last space group in the class of monoclinic Bravais lattice. Consequently, we have a dataset with 47,862 crystal structures biased to materials with low symmetry to test whether CGCNN and ALIGNN can learn human-designed descriptors of crystal structures. To facilitate the analysis about failure of CGCNN to capture lattice constants, we create a dataset of random 1-dimensional carbon chains ("1D dataset"). The random 1D chains are created by the following pseudo-codes in python:

pos = []; for j in range(n): # number of atoms in the chain

if j == 0: pos.append([3*random for 3 dimensions]) # 3 = 2*1.5 Å (approx. C-C bond length).

random: random number between (0, 1)

elif j%2 == 0: pos.append([pos[j-1] + 3*random for 3 dimensions]))

else: pos.append([pos[j-1][0] + 3*random, pos[j-1][1] - 3*random, pos[j-1][2] - 3*random])

a = pos[-1][0]; b = 100; c = 100 # add vacuum for b and c

lattice = Lattice.from_parameters(a, b , c, 90, 90, 90)

structure = pymatgen.core.structure.Structure(lattice, ["C" for _ in range(n)], pos, coords_are_cartesian=True)

For the dataset of (1D, short), the number of atoms is set to be between [2, 9), and for the dataset of (1D, long), the number of atoms is set to be between [37, 51). In total, both datasets have 1,400 data points.For machine learning of human-designed descriptors in Figure 2-2, we split the dataset into 80%, 10% and 10% as the training, validation and test set.

Table 2-1. List of abbreviations of descriptors and properties in Chapter 2.

Abbreviations of descriptors	Full name of descriptors	Abbreviations of properties	Full name of descriptors
MAD_in_rela_bond_len	mean absolute deviation in relative bond length	$\log(\kappa)$	$\log_{10}$ lattice thermal conductivity
max_rela_bond_len	maximum relative bond length	$E_{\text{fin.}}$	final (total) energy per atom
min_rela_bond_len	minimum relative bond length	U	phonon internal energy at 300 K
max_neighb_dist_var	maximum neighbor distance variation	C_{v}	constant volume phonon heat capacity at 300 K
min_neighb_dist_var	minimum neighbor distance variation	K	bulk modulus
range_neighb_dist_var	range neighbor distance variation	G	shear modulus
mean_neighb_dist_var	mean neighbor distance variation	v	poisson ratio
dev_neighb_dist_var	standard deviation neighbor distance variation	E_{g}	band gap
mean_avg_bond_len	mean average bond length	$\|\|e\|\|_{\infty}$	modulus of piezoelectric tensor
std_avg_bond_len	standard deviation average bond length	$\varepsilon_{\text{e.}}$	electronic dielectric constant
MAD_in_rela_atom_vol	mean absolute deviation in atomic volume	$\varepsilon_{\text{t.}}$	total dielectric constant
mean_avg_bond_ang	mean average bond angle	n	refractive index
std_avg_bond_ang	standard deviation average bond angle	M	total magnetization per formula
density	density		
vpa	volume per atom		

packing_frac	packing fraction
struct_comp_atom	structural complexity per atom
struct_comp_cell	structural complexity per primitive cell
natoms	number of atoms per primitive cell
a	the largest lattice length of the primitive cell
b	the second largest lattice length of the primitive cell
c	the smallest lattice length of the primitive cell
alpha	the largest lattice angle of the primitive cell
beta	the second largest lattice angle of the primitive cell
gamma	the smallest lattice angle of the primitive cell
space_group_num	space group number
crys_sys	crystal system
mean_CN	mean coordination number
std_CN	standard deviation coordination number

Models. In Chapter 2, we use the default architecture of CGCNN[115] and ALIGNN[69] for learning human-designed descriptors in Figure 2-2. The reason for using the default architectures is that, as in Figure 3g and 3h, although intentionally revising their architectures can improve learning performance for some descriptors, in this chapter we try to show the representational power and limit of CGCNN and ALIGNN in a setting close to those in real applications. For learning materials properties in Figure 4, hyper-parameter search based on the search spaces in Ref.[132] is conducted. All the neural networks are trained for 300 epochs[69] on a Quadro RTX 6000 GPU. For feature importance in Figure 2-4, since the permutation feature importance of deep neural networks is very expensive to calculate, we estimate feature importance by extracting the representations in equation (2-14), then feed the

representations into a random forest model to calculate the feature importance.

2.6. Chapter summary and outlook

In summary, in Chapter 2, we propose a systematic approach to analyze the representation power of GNNs for crystal structures. We use human-designed descriptors as a bank of knowledge to test whether CGCNN and ALIGNN can capture knowledge of crystal structures behind descriptors. We find that both GNNs can capture basic local structural descriptors well, but cannot capture the periodicity of crystal structures. We analyze the limitations of the GNNs on capturing periodicity from three perspectives: local expressive power, long-range information and pooling function. We also test the idea of hybridization with descriptors to improve the performance of GNN, and show that descriptors-hybridized CGCNN and ALIGNN have better prediction performance for some materials properties than the original CGCNN and ALIGNN, especially phonon internal energy and phone heat capacity with 90% lower errors.

The analysis performed in this chapter can be easily extended to other deep representation learning models, human-designed descriptors, and systems beyond crystals such as molecules and amorphous materials. This chapter shows that the fields of human-designed descriptors and deep representation learning can be developed synergically. For new deep representation learning models, their ability in representation of crystal structures can be tested by learning existing human-designed descriptors, and for new descriptors, they can be used to reveal how well the existing deep representation learning models capture the knowledge behind these descriptors, which can also be hybridized with deep representation learning models for improved prediction performance. We hope this chapter may inspire further development of deep representation learning, human-designed descriptors and hybridized machine learning

models for crystal structures and materials science.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

3. Calibrating DFT formation enthalpy calculations by multi-fidelity learning

3.1.Introduction

In Chapter 3, we present a case study of formation enthalpy of materials to illustrate how to learn small experimental dataset in materials science, with the help of large materials datasets from simulations, and how machine learning can help evaluate the stability of materials.

Machine learning materials properties measured by experiments is valuable yet difficult due to the limited amount of experimental data. In this chapter, we use transfer learning and multi-fidelity machine learning to learn the experimental formation enthalpy of materials. The best machine learning model for this task is a multi-fidelity random forest model with prediction accuracy higher than PBE functional with linear correction and meta-GGA functionals (PBEsol, SCAN and r^2SCAN), and it also outperforms the hotly studied deep neural-network based representation learning and transfer learning. We then use the model to calibrate the DFT formation enthalpy in the Materials Project database, and discover materials with underestimated stability. The multi-fidelity model is also used as a data-mining approach to find how DFT deviates from experiments by explaining the model output.

As discussed in Chapter 1, in order to accelerate the design of new materials, accurate computational methods such as Density Functional Theory (DFT)[158] have been employed to generate large datasets that contain more than 10^5 entries of materials properties. While the availability of such databases has boosted the exploration of novel materials[14, 48, 159-166], it is important to note that most of the data is generated with computationally “cheap” DFT functionals such as PBE[123], that can in turn lead to non-negligible errors when compared

with experimental measurements.

As an example, the formation enthalpy (ΔH_f) is a fundamental property that determines the thermodynamic stability of materials. The mean absolute error (MAE) between the computed ΔH_f in these large DFT databases and experimental measurements are reported to be ~0.1 eV/atom[7, 167]. Due to the sensitivity of phase stability to energy, such a difference (~0.1 eV/atom) might be the difference between a material that is readily synthesizable and one that is almost impossible to realize[67, 168, 169]. In addition, because of the limited amount of available experimental data, currently most machine learning (ML) models applied to materials are trained on DFT datasets[23, 25, 48, 49, 57, 62, 64, 65, 71, 73, 86, 105, 170-173], making any error in the DFT calculations critical to the usefulness of such ML models[14, 24, 25, 174].

To improve the accuracy of formation enthalpy calculations, a number of density functionals have been developed, such as PBEsol[175], SCAN[176], r^{2}SCAN[177] and HSE[178], which have shown significant improvement in accuracy of formation enthalpy calculation[127, 179, 180]. On the other hand, these more accurate functionals are also computationally more expensive, limiting their utility for generation of large databases[180, 181]. Empirical corrections represent another, faster approach to improve the accuracy of prediction of ΔH_f. For example, in the MP dataset, ΔH_f of certain materials (including oxides, phosphates, borates and silicates) is empirically corrected by fitted element corrections[182], and in OQMD ΔH_f is corrected by a chemical-potential fitting[7]. Very recently, Wang *et al.*[183] proposed a linear correction scheme with error of 0.051 eV/atom compared with experimental values on a dataset with 222 materials containing certain anions and transition metals. Yet, despite this recent success in lowering the error for some chemical systems[184], such corrections are based on human understanding of specific chemistries and relatively simple assumptions, and are thus difficult to be transferrable across different chemistries[182, 184]. It would be beneficial to

design prediction schemes that can automatically extract chemistry-property relationship across different chemistries without human intervention, and data-driven ML methods[23, 57, 64, 65, 67, 71, 181] are promising candidates to learn the complex mapping between chemistry and ΔH_f.

One of the biggest challenges in machine learning materials properties is the lack of experimental data[185]. Efforts have been made to improve the performance of learning on small experimental datasets by extracting and transferring information from large DFT datasets. Currently, there are mainly two strategies to achieve the transfer between DFT and experimental datasets, transfer learning[49, 55, 64, 186-188] and multi-fidelity machine learning[52, 53, 66, 181]. The idea of transfer learning (see Figure 3-1a) is first learning large DFT datasets (source) using a large neural network, and then transferring the weights of the network to the machine learning task of small experimental datasets (target). Although transfer learning has achieved success in problems where the source and target datasets are highly correlated[49, 64, 186, 187], the approach is mostly applied to neural network architectures, and if the correlation is not strong enough, transfer learning will not improve and may even deteriorate the learning performance[55]. Different from transfer learning where information is passed by transferring network parameters, in multi-fidelity machine learning (see Figure 3-1b) information of cheap and low-fidelity data is directly passed to the learning task of expensive and high-fidelity data, either in the feature (input) level[66] or in the label (output) level[52, 53, 181, 189]. In other words, the low-fidelity data can be used as feature in the machine learning task of high-fidelity data, or the task of machine learning the high-fidelity data can be converted to the task of machine learning the difference between high-fidelity data and low-fidelity data, which is also known as Δ-Machine Learning[189]. From the handful of previous studies, multi-fidelity machine learning has shown higher predictive power than the single-fidelity ones (directly learning the high-fidelity data) on materials properties like band

gaps and energies from different density functionals[52, 53, 181, 189]. However, there is no previous work that adapt multi-fidelity machine learning in both feature and label level at the same time.

In this chapter, we present a comprehensive machine learning study about ΔH_f^{exp} using transfer learning and multi-fidelity machine learning. For the machine learning architectures, we compare four different models, random forests (RF), multi-layer perceptron (MLP), Representation Learning from Stoichiometry (ROOST)[23] and Crystal Graph Convolutional Neural Network (CGCNN)[71]. We find that multi-fidelity RF in both the feature and label level has the best prediction performance for ΔH_f^{exp} with almost a half reduction in MAE compared with DFT results from MP, and improved performance compared to recent linear correction schemes[183] as well as more sophisticated density functionals like PBEsol[175], SCAN[176] and r^{2}SCAN[177]. We also analyze the effects of machine learning architectures, featurization methods and information transfer strategy on learning ΔH_f^{exp} and ΔH_f^{diff}. Further, the more accurate ΔH_f are applied to re-evaluate the thermodynamic stability of materials, and cases with underestimated stability in the MP database are discovered. We also use the machine learning model to find where current DFT deviates from experiments by explaining the model output.

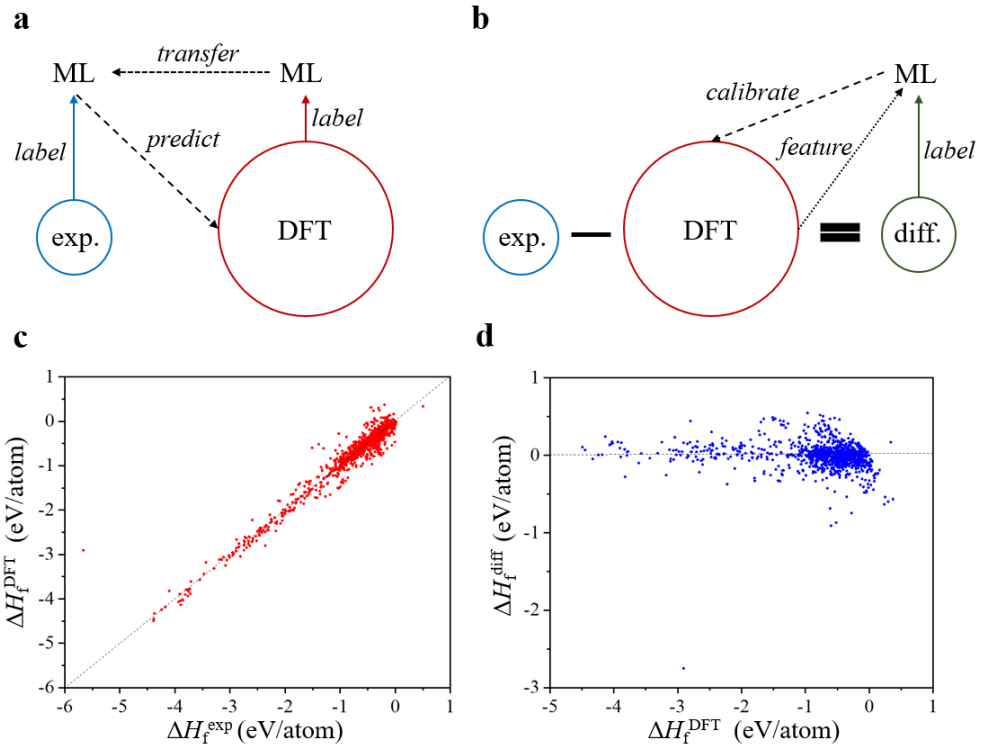

Figure 3-1 Illustrations of the machine learning frameworks and datasets used in Chapter 3. **a** and **b** Schematics of transfer learning and multi-fidelity machine learning in Chapter 3, respectively. In **a**, first the ΔH_f^{DFT} is used as label to train a ML model, then the weights of the first ML model are transferred to initialize a second ML model, and the ΔH_f^{exp} is used as label to train the second model, finally the second model is used to predict ΔH_f^{exp} of all materials in the large DFT dataset. In **b**, first the dataset of the difference between ΔH_f^{exp} and ΔH_f^{DFT} is constructed (ΔH_f^{diff}), then ΔH_f^{diff} is used as label to train a ML model with the ΔH_f^{DFT} as an input feature, and finally the trained model is used to calibrate the different between ΔH_f^{DFT} and ΔH_f^{exp} for all materials in the large DFT dataset. **c** ΔH_f^{DFT} versus ΔH_f^{exp}. **d** ΔH_f^{diff} versus ΔH_f^{DFT}.

3.2. Machine learning frameworks and datasets

In Chapter 3, we use two different strategies to learn ΔH_f^{exp} with the assistance of information from the MP dataset, transfer learning and multi-fidelity machine learning (in the following, "ΔH_f^{DFT}" denotes the empirically-corrected PBE ΔH_f by Jain *et al.*[182] from the MP database, V2021.03.22). As shown in Figure 3-1a, in transfer learning a neural network is first trained on the large MP dataset with more than 10^5 data points of ΔH_f^{DFT}, then weights of the neural network are transferred to initialize a second neural network, and finally part of the weights of the second network are optimized by the small ΔH_f^{exp} dataset. Once trained, the

second neural network can serve to predict ΔH_f^{exp} of materials in the large MP dataset. In multi-fidelity machine learning, as shown in Figure 3-1b, first the dataset of ΔH_f^{diff} (ΔH_f^{exp} - ΔH_f^{DFT}) is built, then machine learning models are trained on ΔH_f^{diff} dataset, and in the training process, ΔH_f^{DFT} can serve as an input feature of each material. Once trained, the machine learning model can serve to calibrate the ΔH_f^{DFT} by adding ΔH_f^{diff} to ΔH_f^{DFT} to get the ΔH_f^{exp}. The key difference between transfer learning and multi-fidelity machine learning is that in the former two networks are trained and information transfer is achieved by transferring network weights, while in the later only one model is trained and information transfer is achieved by learning the difference between two datasets and adding the ΔH_f^{DFT} as one of the input features. In addition to the two basic strategies as shown in Figure 3-1a and b, variants are also tested in this chapter, including combination of transfer learning and multi-fidelity machine learning (initializing a network from one trained on ΔH_f^{DFT} and optimizing the newly initialized network by ΔH_f^{diff}), and multi-fidelity machine learning by only learning ΔH_f^{diff} or only adding ΔH_f^{DFT} as input feature.

As described above, we choose four different machine learning architectures to realize transfer learning and/or multi-fidelity machine learning, which are RF, MLP, ROOST and CGCNN. The choice aims to increase the variety of machine learning architectures to fairly evaluate the effect of transfer learning and multi-fidelity learning, and to enlarge the hypothesis space to search for the best machine learning models for predicting ΔH_f^{exp}. These ML architectures also provide varieties in terms of basic algorithms, input information and featurization: MLP, ROOST and CGCNN are based on neural networks while RF is not; ROOST only needs compositions as input while CGCNN takes both compositions and 3D structures as input, and RF and MLP can be trained either with or without structural information; RF and MLP need human-engineered featurization while ROOST and CGCNN learn fingerprints of materials in the training process.

In Chapter 3, we choose the Materials Project database (MP, V2021.03.22) as the source of ΔH_f^{DFT}, because MP is a widely used large DFT database, and the difference of ΔH_f between MP and other large DFT databases is not large. For example, the difference between ΔH_f of 563 materials from MP and OQMD is reported to be 0.028 eV/atom[7]. As for the experimentally measured ΔH_f, we combine the IIT dataset[167] and SSUB dataset[190] and remove the duplicates, leading to 1143 data points with available ΔH_f^{exp}, ΔH_f^{DFT}, and DFT optimized 3D atomic structures from MP. In addition to the value of ΔH_f^{exp}, there are also uncertainty estimations in the IIT dataset[167], from which one can see that the mean uncertainty of ΔH_f^{exp} based on 499 materials is around 0.023 eV/atom. More details about the data collection procedure are provided in Chapter 3.6. ΔH_f^{DFT} and ΔH_f^{exp} are compared in Figure 3-1c, from which one can see that ΔH_f^{DFT} are already quite close to ΔH_f^{exp} in value, and there is no clear systematic shift between ΔH_f^{DFT} and ΔH_f^{exp}. As shown in Figure 3-1d, the distribution of ΔH_f^{diff} is centered around zero, and there is no obvious correlation between ΔH_f^{diff} and ΔH_f^{DFT}. From Figure 3-1c and 3-1d, one can see that ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp} with the standard deviation of 0.1718 eV/atom and 0.8000 eV/atom for the ΔH_f^{diff} dataset and ΔH_f^{exp} dataset, respectively.

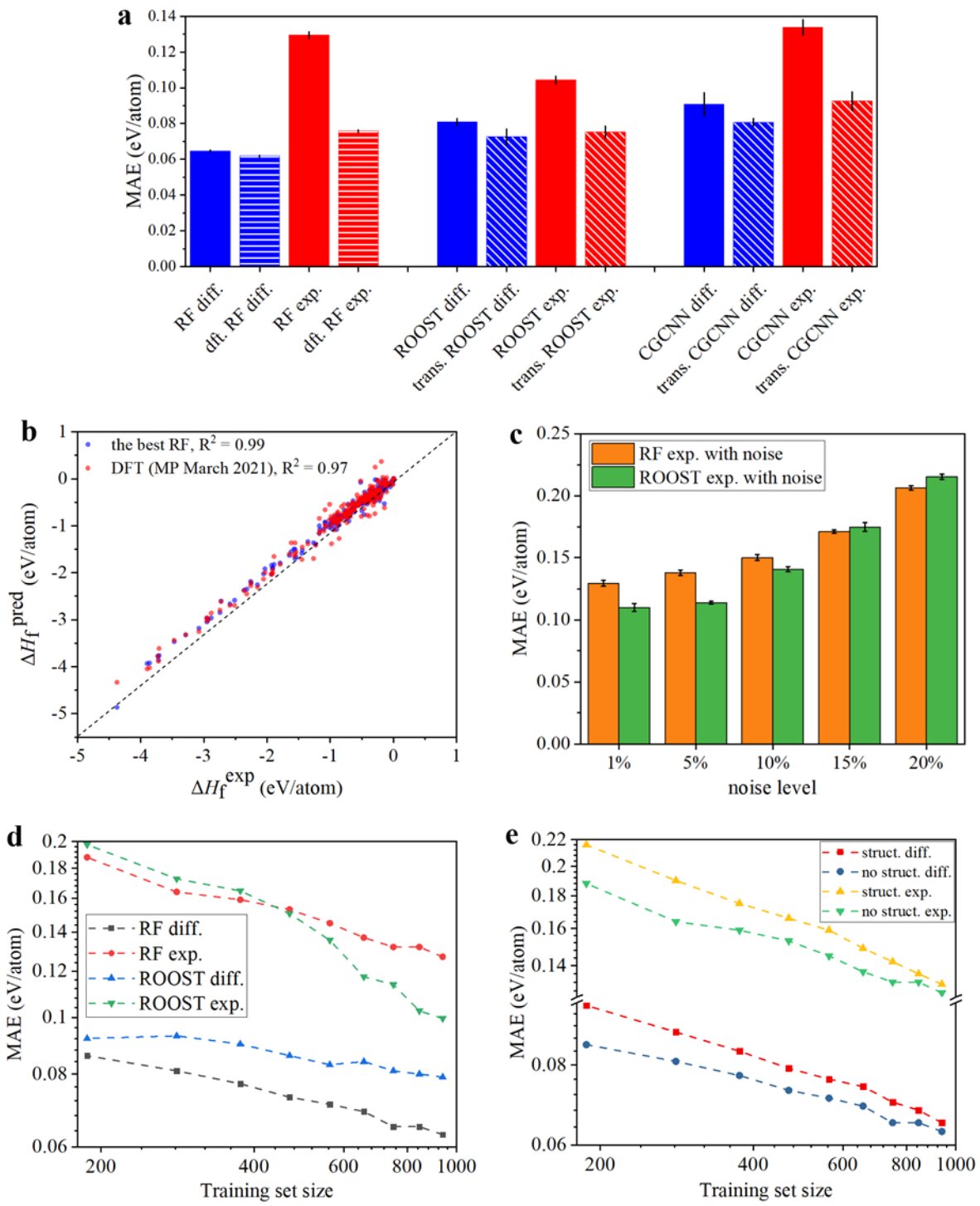

Figure 3-2. Comparison of machine learning models. **a** Mean average errors (MAE) between predictions of ΔH_f from machine learning models and experimental measurements. Each type of machine learning model is trained 10 times to estimate the uncertainty levels. RF denotes random forest, MLP denotes multilayer perceptron, and ROOST[23] and CGCNN[71] are two deep-learning models that automatically extract materials' fingerprints from compositions and structures, respectively. Here, "struct." means the model is trained with structural and compositional features, "no struct." the model is trained with only compositional features, "dft." the model is trained with ΔH_f^{DFT} as an input, "trans." the model is trained in a transfer learning

manner, "diff." the model is trained on ΔH_f^{diff}, "exp." the model is directly trained on ΔH_f^{exp}. The dashed horizontal line corresponds to the MAE of ΔH_f^{DFT}. **b** ΔH_f^{exp} versus ΔH_f^{ML} from the best RF model (the sixth from the left in **a**) and ΔH_f^{DFT}. **c** MAE of predictions of ΔH_f^{exp} with noise from RF and ROOST. Under each noise level, gaussian noises with standard deviation of noise level*0.8 eV/atom (0.8 eV/atom is the standard deviation of the ΔH_f^{exp} dataset) are added to both training set and test set. **d** and **e** Learning curves of different models. The MAE is for the test set. In **e**, all the curves are based on random forest, and "struct." means the model is trained with structural and compositional features, "no struct." the model is trained with only compositional features.

3.3. Predicting ΔH_f^{exp} by machine learning

For the RF and MLP, compositional and structural features are provided from matminer[135] as input features (a list of features is provided in Chapter 3.6), for ROOST only the compositions are provided as input and it automatically learns the fingerprints of materials, and for CGCNN the compositions and 3D atomic structures are provided as input and the fingerprints are learned in the training. In order to test the prediction performance, 20% of the 1143 materials are randomly chosen as the test set. Details about the training procedure are provided in the Chapter 3.6. As a baseline, for the test set, we find that the MAE between ΔH_f^{DFT} and ΔH_f^{exp} is 0.0955 eV/atom. The test results for all machine learning models are shown in Figure 3-2a, and here we analyze the results from the following aspects:

(1). The best performance is achieved with the RF model that is trained on ΔH_f^{diff} and has compositional features and ΔH_f^{DFT} as input features (Figure 3-2a). The error for this best case, 0.0617 eV/atom, is roughly 30% lower than that of ΔH_f^{DFT}. The parity plot of ΔH_f^{DFT} and ΔH_f^{ML} from the best RF model versus ΔH_f^{exp} of the test set is shown in Figure 3-2b, from which one can observe that ΔH_f from the best RF model aligns closer to the ΔH_f^{exp} than ΔH_f^{DFT} within the range from -5 eV/atom to 1 eV/atom. Predictions from the best RF model also have a higher R^2 score (0.99) than that from the DFT calculations in the MP database (0.97).

Recently, Kingsbury *et al.*[127] performed high-throughput calculations for 6,000

materials by PBEsol[175], SCAN[176] and r^{2}SCAN functional[177]. In Table 3-1, MAEs between experimental ΔH_f and ΔH_f from different density functionals with different empirical corrections are listed. Note that, different from Figure 3-2, the reported MAEs in Table 3-1 are based on a dataset with 122 materials that have all the values of ΔH_f from different sources (these materials are in the test set mentioned above). One can observe that, MAE of the best RF model is almost half of that of SCAN[176], PBEsol[175], and also almost half of that of the corrections from Jain *et al.*[182] and Wang *et al.*[183]. The superiority of the best RF model over the meta-GGA functionals (SCAN and r^{2}SCAN) is encouraging, because i) the best RF model provides lower error compared with more sophisticated density functionals, ii) it is much faster than the self-consistent DFT simulations, especially with meta-GGA functionals, enabling one to screen ΔH_f of materials accurately in a high-throughput fashion. For example, for the 10^5 materials in large DFT databases such as MP, more accurate predictions of ΔH_f can be calculated by the RF models within minutes, while that from meta-GGA functionals may take months of calculations. Note that for new materials without low-fidelity ΔH_f predictions yet (such as corrected-PBE), computational cost for the low-fidelity ΔH_f should be added to the total cost of the best RF model.

As for the superiority of the best RF model over the recent linear correction scheme from Wang *et al.*[183] as shown in Table 3-1, there are four possible explanations: i) the RF model takes non-linear effects into account, ii) the compositional descriptors used here capture more information than simple stoichiometry used in Wang *et al.*[183], iii) the learned correction in Wang *et al.*[183] is only from materials with certain anions and transition metals while in the present work there is no such constraint, and iv) the calibration scheme used here is built on empirically corrected PBE results as opposed to uncorrected PBE data in Wang *et al.*[183].

Table 3-1. Comparison of MAEs between ΔH_f^{exp} and ΔH_f from different density functionals with different corrections. Different from Figure 3-2, the reported MAEs here are based on a dataset with 122 materials in the test set that have all the values of ΔH_f from different sources. The two corrections in the cell of "PBE (Jain *et al.*[182], the best RF)" show that the PBE ΔH_f is first corrected by Jain *et al.*[182] then corrected by the best RF model in this chapter. "(no)" in the right three cells at the upper row means that no correction is applied to the ΔH_f from the density functional. "PBE (Jain *et al.*[182])" is the one used in the MP database before May 2021 (V2021.03.22) and is the one used as the low fidelity data in this chapter ("ΔH_f^{DFT}"). "PBE (Wang *et al.*[183])" is the one used in the MP database after May 2021 (V2021.05.13). MAE is in the unit of eV/atom.

Functional (Correction)	PBE (Jain *et al.*[182], the best RF)	PBE (Jain *et al.*[182])	PBE (Wang *et al.*[183])	PBEsol (no)[127]	SCAN (no)[127]	r^{2}SCAN (no)[127]
MAE	0.0542	0.0935	0.0927	0.0973	0.1024	0.0825

(2). Training the machine learning models on ΔH_f^{diff} helps to reduce error compared with training models on ΔH_f^{exp} directly, as under the same condition (architecture and featurization), the models trained on ΔH_f^{diff} always have lower MAE than that trained on ΔH_f^{exp}. Here, we attribute the lower absolute error of learning ΔH_f^{diff} to the fact that ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp} with 5 times smaller standard deviation (0.17 eV/atom versus 0.80 eV/atom). One can imagine that, if ΔH_f^{diff} and ΔH_f^{exp} have the same distribution except a scaling factor of 1/5, then ideally the MAEs of ML models (with proper normalization) trained on ΔH_f^{diff} should also be 1/5 of that trained on ΔH_f^{exp}. However, the MAEs of models trained on ΔH_f^{diff} are all larger than 1/5 of that trained on ΔH_f^{exp}, suggesting that ΔH_f^{diff} is easier to learn absolutely but harder to learn relatively than ΔH_f^{exp}.

In order to further illustrate the above explanation, we use R^2 score, a unitless metric invariant to scaling, to show the relative difficulty of predicting ΔH_f^{diff} and ΔH_f^{exp}. The R^2 of predictions of ΔH_f^{diff} by the best RF model is 0.54 (here R^2 of 0.54 is based on predicted ΔH_f^{diff} versus true ΔH_f^{diff}, while the R^2 of 0.99 in Figure 3-2b is based on predicted ΔH_f^{exp} versus true ΔH_f^{exp}), while the R^2 of predictions of ΔH_f^{exp} by the same RF model is 0.94, suggesting that

ΔH_f^{exp} is easier to learn relatively than ΔH_f^{diff}.

(3). Feeding ΔH_f^{DFT} as one of the input features helps to lower the error. As with the same machine learning architecture (RF or MLP), label, and other features, models with ΔH_f^{DFT} as one of the input features always have lower error than that without ΔH_f^{DFT}. This effect is more significant when the models are trained on ΔH_f^{exp}, because as shown in Figure 3-1c ΔH_f^{DFT} has a strong correlation with ΔH_f^{exp}, while as shown in Figure 3-1d the correlation between ΔH_f^{DFT} and ΔH_f^{diff} is not obvious.

Combining analysis (2) and (3), one can observe that, adapting the strategy of multi-fidelity machine learning might help to significantly lower prediction error, if the difference between the different fidelity datasets has a narrower distribution than the high-fidelity dataset, and/or if there is a strong correlation between the different fidelity datasets. Machine learning models with both the modifications of changing label and adding extra input features might outperform that with either single modification.

(4). Similar to (3), transfer learning helps more when transferring from ΔH_f^{DFT} to ΔH_f^{exp} than from ΔH_f^{DFT} to ΔH_f^{diff} because of the stronger correlation between ΔH_f^{DFT} and ΔH_f^{exp}.

(5) RF with human-engineered features performs better than ROOST and CGCNN, two deep representation learning models, when trained on ΔH_f^{diff}, while RF performs similar to or worse than neural-network based models when trained on ΔH_f^{exp}. Although it is not surprising that neural-network based deep learning algorithms don't show superior performance over RF due to the limited dataset size[55, 191], the effect of learning targets (ΔH_f^{diff} and ΔH_f^{exp}) on prediction performance of different machine learning models is interesting and worth of being discussed.

The different uncertainty level between ΔH_f^{diff} and ΔH_f^{exp} might help to explain why RF performs better than neural network-based models when trained on ΔH_f^{diff} while there is no

such superiority of RF when trained on ΔH_f^{exp}. As discussed above, ΔH_f^{diff} has a narrower distribution than ΔH_f^{exp}. Because $\Delta H_f^{diff} = \Delta H_f^{exp} - \Delta H_f^{DFT}$, if we consider ΔH_f^{exp} and ΔH_f^{DFT} as two independent random variables, then ΔH_f^{diff} should have larger uncertainty than ΔH_f^{exp}. Therefore, the robustness of RF against uncertainty[191, 192] might explain the superiority of RF when trained on ΔH_f^{diff}. The larger uncertainty level of ΔH_f^{diff} might also help to explain why ΔH_f^{diff} is harder to learn relatively than ΔH_f^{exp} as in (2).

In order to further investigate the effect of uncertainty on performance of machine learning models, RF and ROOST are employed to learn ΔH_f^{exp} with random noises, a source of uncertainty. In Figure 3-2a, one can see that RF performs worse than ROOST when trained on ΔH_f^{exp}. In Figure 3-2c, the MAEs of RF and ROOST and the corresponding noise levels are shown. One can see that, under low noise levels the errors of RF are still higher than that of ROOST, while under high noise levels the errors of RF become lower than that of ROOST. The different relative performance of RF and ROOST under different noise levels agrees with the superiority of RF against uncertainty[191, 192], and supports our hypothesis that the different uncertainty levels of the ΔH_f^{diff} dataset and the ΔH_f^{exp} dataset might explain why RF is better on the ΔH_f^{diff} dataset while ROOST is better on the ΔH_f^{exp} dataset.

In Figure 3-2e, we plot the learning curves of RF and ROOST on learning ΔH_f^{diff} and ΔH_f^{exp}, respectively. For learning ΔH_f^{exp}, we observe that with few data points, RF has smaller errors than ROOST, while with more than 400 data points, ROOST outperforms RF, which agrees with previous observations[23, 113] that deep learning is powerful for large datasets while classic machine learning is more suitable for small datasets. However, for learning ΔH_f^{diff}, we observe that RF performs better than ROOST consistently for all dataset sizes. As for the rate of improvement with respect to dataset size (slope of learning curve), we observe that for RF, the slope on learning ΔH_f^{diff} is slightly smaller than that on ΔH_f^{exp}, while for ROOST, the slope

on learning ΔH_f^{diff} is significantly smaller than that on ΔH_f^{exp}, which shows that the slopes of learning curves of machine learning models are affected by the quality of data: higher uncertainty of data, smaller slope of learning curves, and different machine learning models are affected differently: slope of RF is less affected while slope of ROOST is more affected. Further empirical and theoretical studies are necessary to investigate the relation between data quality and slope of learning curve for different machine learning models. From the learning curves, we also expect that, with more ΔH_f^{exp} data points in the future, learning ΔH_f^{exp} directly by ROOST might be more powerful than learning ΔH_f^{diff} by random forest.

Based on the fact that when trained on ΔH_f^{diff}, random forest with human-engineered featurization outperforms neural networks-based models, especially deep representation learning models, we suggest that for machine learning applications in the field of materials science, with limited dataset size and without proof of a low uncertainty level of the dataset, deep neural network-based representation learning algorithms[23, 57, 71, 75] should not be the only type of models employed, and other feature engineering methods and machine learning architectures beyond neural networks should also be tested.

While there are some previous works show that information of local bonding environment can be used to correct formation enthalpies of certain materials like sulfides[193], fluorides[194] and oxides[184, 194], in this chapter, the machine learning models with only compositions as input outperform those with both compositions and structures as input. One of the possible causes of the phenomenon is that there still lacks the data points of polymorphs with the same composition but different ΔH_f^{exp} in the current dataset, which suggests the urgency of building a comprehensive ΔH_f^{exp} dataset with sufficient entries of polymorphs to comprehensively understand the role of structures in determining ΔH_f^{exp}. In Figure 3-2e, we plot the learning curves for random forest (RF) with and without structural features for learning ΔH_f^{diff} and

ΔH_f^{exp}. We can observe that, for learning ΔH_f^{diff} and ΔH_f^{exp}, RF without structural features outperforms RF with structural features, while the slopes of learning curves of the RF models with structural features are larger than that of the RF models without structural features. A possible explanation is that models with structural features have more available information, more degree of freedom and therefore easier to overfit small datasets, while those additional information makes models with structural information more powerful and consequently with steeper learning curves. Based on the learning curves, we expect that with more data in the future, models with structural information might outperform models with only compositional information.

For the models only based on compositional information, such as random forest with only compositional features and ROOST, the corrections are the same to the polymorphs. Since the best model in this chapter is only based on compositional information, in the following sections, analysis is purely based on compositions. However, there are also models based on structural information trained and tested in this chapter, such as random forest with compositional and structural features and CGCNN. Corrections from models with structural information are in principle different for polymorphs. In Ref.[97], we show the corrections from random forest (RF) with both compositional and structural features to three pairs of polymorphs with recorded ΔH_f^{exp} values and not in the training set. We find that, for all materials except $CaSiO_3$ wol., ΔH_f^{RF} is closer to ΔH_f^{exp} than ΔH_f^{DFT}, showing the ability of the RF model to correct DFT prediction of ΔH_f. As for relative phase stability, ΔH_f^{DFT} contradicts with ΔH_f^{exp} for SiO_2 and TiO_2. Unfortunately, for the two systems, corrections from RF cannot reverse the wrong phase stability estimation from DFT. A possible explanation is that, the RF model mainly employs compositional information to learn and predict, as we find that compositional features contribute 80% feature importance, while structural features only contribute 20% feature importance. Therefore, the RF model predicts similar corrections to different structures with

the same composition. More data points of ΔH_f^{exp}, especially that of polymorphs, are necessary to develop machine learning models that rely more on structural information and are capable to reverse the wrong phase stability estimation of polymorphs from DFT.

We summarize the potential drawbacks of using one model versus the others for predicting ΔH_f as below: i) for RF and MLP, they rely on *off-the-shelf* featurization, which means that they cannot capture information unknown to human beings. Therefore, they are typically less powerful than deep representation learning models such as ROOST and CGCNN for large datasets[23, 113]. For predicting ΔH_f, although RF is the best model in this chapter, with more data points in the future, it is likely that RF will be less powerful than ROOST as shown by the learning curves in Figure 3-2d. ii) for ROOST and CGCNN, they are deep representation learning models that learn the features of materials automatically in the training process. Therefore, they are thought to be more powerful than models with *off-the-shelf* featurization, but their prediction performance might be worse with small datasets[23, 113], such as this chapter. iii) for MLP, ROOST and CGCNN, they are neural network-based models. Compared with random forest, which is a decision tree-based ensemble model with hundreds of individual models, ensembles of neural networks are typically only composed of around 10 individual models because of the higher computational cost of neural networks[23]. Therefore, they might be less powerful than RF in cases where number of models in ensembles is important[195], such as the ΔH_f^{diff} with high uncertainty in this chapter[195].

3.4. Discovery of materials with underestimated stability

With the best RF model that can significantly lower the error of ΔH_f from the MP database, we can calibrate ΔH_f of all materials in the MP database. As an application, here we use the calibrated ΔH_f to re-evaluate the thermodynamic stability of all materials in the MP database

by constructing the energy above hull (E_{hull}, the energy difference between the candidate compound and the ground-state phase(s) in a compositional space[196].) However, as Bartel *et al.*[67] pointed out, although sometimes DFT has large errors for prediction of ΔH_f, ΔH_f^{DFT} of similar materials contain similar systematic errors, and when evaluating phase stability, the cancellation of systematic errors makes DFT more useful for evaluating relative stability between compounds than some machine learning models with similar or even better accuracy with respect to ΔH_f^{exp}.

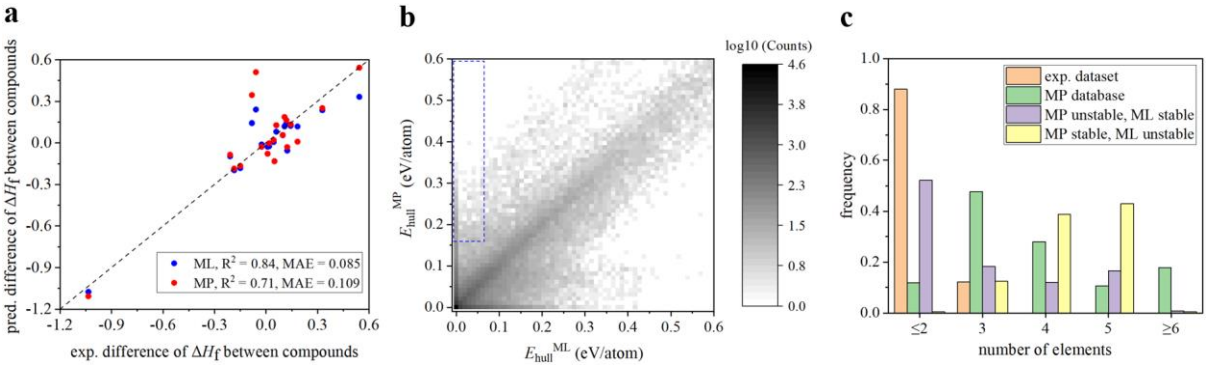

Figure 3-3. Stability evaluation from energy above hull. **a** Difference of ΔH_f between pairs of compounds in the same chemical system from experiments versus that from MP and machine learning. **b** Distribution of energy above hull (E_{hull}, in eV/atom) of all materials in the Materials Project[6] database calculated by the corrected-PBE ΔH_f in MP (E_{hull}^{MP}) versus that calculated by the machine learning ΔH_f in this chapter (E_{hull}^{ML}). Here, E_{hull} is constructed from all materials in the Materials Project database. The color scheme is used to show the (log10 of) number of materials within a range of certain E_{hull}^{ML} and E_{hull}^{MP}, and the red rectangle shows the area with $E_{hull}^{MP} > 0.16$ eV/atom and $E_{hull}^{ML} < 0.06$ eV/atom. **c** Appearance frequencies of number of elements of each material in the datasets. Here, "exp. dataset" is the ΔH_f^{exp} used in this chapter, "MP database" is the set of all materials in the Materials Project database, "MP unstable, ML stable" is the set of materials with $E_{hull}^{MP} > 0.16$ eV/atom and $E_{hull}^{ML} < 0.06$ eV/atom and "MP stable, ML unstable" is the set of materials with $E_{hull}^{MP} < 0.06$ eV/atom and $E_{hull}^{ML} > 0.16$ eV/atom.

Therefore, before screening E_{hull} for the full MP dataset, we first evaluate the performance of ΔH_f^{DFT} and ΔH_f^{ML} for evaluating relative stability between compounds. Since there are only 229 materials in the test set, which are not enough for constructing phase diagrams and E_{hull}, we use the difference between ΔH_f of pairs of compounds in the same chemical system to evaluate relative stability between compounds. We list all 20 pairs of compounds in the same

chemical system in the test set in Table 3-2, and we also plot the difference from experiments versus that from MP and machine learning (ML) in Figure 3-3a. One can see that ML outperforms MP in terms of difference of ΔH_f between compounds in the same chemical system, which shows that the ML model outperforms DFT for relative stability evaluation.

Table 3-2. Difference of ΔH_f between pairs of compounds in the same chemical system from different sources. Difference of ΔH_f is the unit of eV/atom.

Pair of Compounds	Experiment	Materials Project	Machine Learning in This chapter
$TiFe_2$ - TiFe	0.0487	-0.1324	-0.1316
BiI_3 - BiI	0.1075	0.1868	0.1193
$LuIr_2$ - LuIr	-0.1502	-0.1664	-0.1826
LaSi - La_5Si_3	0.143	0.1335	0.1229
BMo_2 - BMo	-0.1858	-0.1856	-0.1972
Na_2O -NaO_2	0.5435	0.5428	0.3328
BW_2 - B_5W_2	-0.0591	0.5108	0.2408
Co_3O_4 -CoO	0.1229	-0.0302	-0.0553
$ZrCo_2$ - Zr_2Co	0.0974	0.0574	0.0553
TmAg - $TmAg_2$	0.1835	0.0088	0.1187
$PrNi_5$ - PrNi	-0.0259	-0.0281	-0.0116
$TiAu_2$ - TiAu	0.0179	-0.0026	-0.0243
NdRh - $NdRh_2$	0.0446	0.0202	0.0064
CaO_2 - CaO	-1.0353	-1.1070	-1.075
Zr_5Si_3 - Zr_5Si_4	-0.2094	-0.0855	-0.0964
Zr_5Si_3 - $ZrSi_2$	0.1181	0.1654	0.1397
Zr_5Si_4 - $ZrSi_2$	0.3275	0.2509	0.2361

Mn_2Sb - MnSb	-0.0824	0.3453	0.1428
CrSi - $CrSi_2$	0.0090	-0.0783	-0.0280
$Mn_{11}Si_{19}$ - Mn_3Si	0.0596	0.1276	0.0809

We next re-evaluate materials stability using ML calibrated ΔH_f to construct E_{hull}^{ML} for all materials in the MP database using all compositions in MP. In chemical intuition, materials with smaller E_{hull} tend to be more thermodynamically synthesizable and stable[168, 169, 197], although $E_{hull} = 0$ is not a hard threshold for successful synthesis and room-temperature and pressure stability of materials because of other factors such as kinetics[198], and in practice empirical heuristics of several room temperature k_BT are used as stability thresholds[168, 169, 197]. In Figure 3-3b the distributions of E_{hull} of all materials in the MP database constructed from ΔH_f^{DFT} and ΔH_f^{ML} of all compositions in the MP database are shown, from which one can see that most materials have similar E_{hull}^{MP} and E_{hull}^{ML}, and majority of materials have close-to-zero E_{hull}^{MP} and E_{hull}^{ML}. More importantly, there are materials with large E_{hull}^{MP} and small E_{hull}^{ML}. These materials might have underestimated stabilities in MP. For example, there are 800 materials in the blue rectangle in the upper-left corner in Figure 3-3b that have $E_{hull}^{MP} > 0.16$ eV/atom and $E_{hull}^{ML} < 0.06$ eV/atom, among which there are around 100 already synthesized materials. (The thresholds are set to be relaxed from 6 times and 2 times of room Temperature k_BT[168, 169]). As examples, we list some interesting materials in Table 3-3 with novel physical properties and/or potential applications with $E_{hull}^{MP} > 0.16$ eV/atom and $E_{hull}^{ML} < 0.06$ eV/atom, where there are both synthesized materials and hypothetical materials. One can see that there are a number of materials with various applications ranging from battery electrodes[199], catalysts[200-202] to optical[203-205], electronic[206, 207], magnetic[208-212] devices and superconductors[213, 214], for which E_{hull}^{ML} succeeds in explaining their synthesizability while E_{hull}^{MP} does not. One extreme example is MnSnIr[215], a stable Half-

Heusler compound synthesized from a peritectic reaction[216], of which E_{hull}^{MP} is considerably high (0.5117 eV/atom) while E_{hull}^{ML} is 0. The large gap between E_{hull}^{MP} and E_{hull}^{ML} is mainly because of the large deviation between ΔH_f^{DFT} (0.2945 eV/atom) and ΔH_f^{ML} (-0.2363 eV/atom) of MnSnIr itself. As a comparison, the ΔH_f^{exp} of MnSnIr is -0.3047 eV/atom[167], which shows that, for this compound, DFT deviates significantly from the experiment, while our machine learning model can calibrate such large difference. A possible reason for the large error of ΔH_f^{DFT} of MnSnIr is that, in Materials Project (V2021.03.22), DFT + *U* correction is only applied to Mn-F, Mn-O and Mn-S systems and not applied to the compound of MnSnIr[182]. Large deviations between ΔH_f^{DFT} and ΔH_f^{exp} are also observed for other compounds containing Mn and Sn, such as MnSnAu (ΔH_f^{DFT}: -0.0488 eV/atom; ΔH_f^{exp}: -0.5016 eV/atom), $MnSn_2$ (ΔH_f^{DFT}: 0.1363 eV/atom; ΔH_f^{exp}: -0.0954 eV/atom), and Mn_2SnRu (ΔH_f^{DFT}: 0.0789 eV/atom; ΔH_f^{exp}: -0.1803 eV/atom), which agrees with the observation in Figure 4b shown later that DFT tends to overestimate ΔH_f (more positive) of compounds with Mn and Sn. As a result, in the phase diagram of Mn-Sn, there is no stable intermetallic compounds according to ΔH_f^{DFT}, which disagrees with the experimental phase diagram where there are several stable intermetallics including Mn_3Sn, Mn_3Sn_2, $MnSn_2$[217].

In addition to the already synthesized materials, those unrealized hypothetical materials provide potential opportunities for energy and environmental materials[99, 218, 219], structural materials[220] and electronic devices[221, 222], and as shown in Table 3-3 and Figure 3-3b, many of these materials that are estimated stable by E_{hull}^{ML} might have underestimated stability in the MP database. Therefore, in the future, if experimentalists intend to realize those materials, large E_{hull}^{MP} alone should not be sufficient for excluding the trial of synthesis if those materials have small E_{hull}^{ML}.

Note that there are also 1,000 materials in the lower-right corner in Figure 3-3b that have $E_{hull}^{MP} < 0.06$ eV/atom and $E_{hull}^{ML} > 0.16$ eV/atom. Details of those materials can be obtained

in the shared online dataset. An extreme example is $LiNbGeO_5$,[223] a synthesized compound with E_{hull}^{MP} of 0 and E_{hull}^{ML} of 0.4334 eV/atom.

In order to further investigate how MP and ML disagree with each other, the appearance frequencies of number of elements in each material in four datasets are plotted in Figure 3-3c. One can see that in the exp. dataset used as the training set in this chapter, around 90% materials are binary compounds and 10% materials are ternary, while in the MP database there are about 40% materials that contain more than 3 elements. Since the training set doesn't cover materials space with more than 3 elements, the ML predictions for materials with more than 3 elements are extrapolations and in general less reliable than that for binary and ternary compounds. For the set of materials unstable by MP and stable by ML, the distribution of number of elements is similar to that of the exp. dataset where the majority of materials are binary or ternary, while in the set of materials stable by MP and unstable by ML, most materials have 4 or 5 elements. Here the lack of materials with more than 3 elements in the current ΔH_f^{exp} dataset suggests that the ML predictions for materials with more than 3 elements should be carefully checked if ML and MP disagree with each other, and it also suggests the urgency of building a comprehensive ΔH_f^{exp} dataset with sufficient entries of materials with more than 3 elements.

Table 3-3. Examples of materials that have novel physical properties and/or potential applications with $E_{hull}^{MP} > 0.16$ eV/atom and $E_{hull}^{ML} < 0.06$ eV/atom. The materials with experiment as one of the characterization methods are synthesized materials, and others are currently only hypothetical. E_{hull} is in the unit of eV/atom.

Materials	MP ID	E_{hull}^{MP}	E_{hull}^{ML}	Characterization method(s)	Comment/ novel physical property/ potential application
MnSnIr	mp-11480	0.5117	0	Experiment	Largest difference between E_{hull}^{MP} and E_{hull}^{ML}.
Ta_3Pb	mp-1187214	0.3386	0	Experiment	Superconductor[214]
AgRh	mp-1183233	0.2633	0.0359	Experiment	Electrocatalyst[200]
FeCoSn	mp-1025124	0.1836	0.0384	Experiment	Tuning phase transitions for isostructural alloying[224]

$SmCo_4Ag$	mp-1219086	0.1797	0.0493	Experiment	Positively correlated magnetization with temperature[208]
$Li_3(FeS_2)_2$	mp-753818	0.1697	0.0180	Experiment	Li-FeS_2 battery electrode[199]
PdRu	mp-1186459	0.2277	0.0032	Experiment	Catalyst[201]
Ni_3Ag	mp-1100764	0.2332	0	Experiment	Dual-frequency absorption[203]
Rb_2NaTaF_6	mp-1114459	0.2038	0	Experiment	Large anisotropic shift from both covalent and polarization spin transfer mechanisms[209]
Nb_3Tl	mp-569366	0.2083	0	Experiment	Superconductor[213]
UPb_3	mp-1184128	0.1621	0	Experiment	Sharp metamagnetic transitions[210]
Cu_3N	mp-1933	0.1865	0.0464	Experiment	Light recording media[204]
$FeNi_2$	mp-1072076	0.1858	0.0292	Experiment	Size-dependent catalytic activity[202]
$HfCo_7$	mp-1105489	0.2098	0.0500	Experiment	Rare-earth-free permanent magnets[211]
MnBi	mp-1185989	0.2078	0	Experiment/DFT	Half-metallic ferromagnetism[207]
Be_2Si	mp-1009829	0.2352	0.0272	Experiment/DFT	Hybrid nodal-line semimetal[206]
Mn_2Hg_5	mp-30720	0.2362	0	Experiment/DFT	π-based covalent magnetism[212]
Ta_3Bi	mp-1187199	0.3442	0	DFT	Topological Dirac semimetal[221]
MnCrSb	mp-1221652	0.2564	0	DFT	Half-metallicity[222]
LiB_{11}	mp-1180507	0.2084	0.0234	DFT	Pseudo-plasticity[220]
$NiAg_3$	mp-976762	0.1850	0	DFT	Acetylene adsorbent[219]
Li_2VN_2	mp-1246112	0.1615	0.0279	DFT	Li-ion battery electrode[218]
$LiGdO_3$	mp-1185401	0.3476	0.0575	Machine learning	Perovskite with high tolerance factor[99]
$LiPmO_3$	mp-1185388	0.2815	0	Machine learning	Perovskite with high tolerance factor[99]

3.5.Data mining of where DFT fails

In addition to predicting more accurate ΔH_f and examining stability of materials, the random forest model trained on ΔH_f^{diff} ($\Delta H_f^{exp} - \Delta H_f^{DFT}$) with human-engineered features can also serve

as a data-mining approach to reveal where and how ΔH_f^{DFT} deviates from ΔH_f^{exp} (as above, "ΔH_f^{DFT}" refers to the empirically corrected PBE ΔH_f by Jain *et al.*[182] in the Materials Project database), which provides clearer trends than machine learning models trained on ΔH_f^{DFT} only. Here, we analyze the relationship between human-understandable features and ΔH_f^{diff} by explaining the model, or for each material, calculating the impact of each feature on the model output (known as the SHAP value[96]). Previously, the error of ΔH_f^{DFT} is mostly discussed in the context of certain anions[7, 178, 182], cations[7] and transition metals[7, 178, 179]. In Figure 3-4a, the impacts of the top 10 compositional features from matminer[135] with the highest sum of absolute SHAP values are shown. One can see that, in addition to anion properties ("max GSbandgap", the detailed explanations of the descriptors are available in the matminer paper[135]) and cation properties ("max GSvolume", "max NdValence", "min CovalentRadius", "min Electronegativity"), mean field of elemental properties ("band center", "mode CovalentRadius") and standard deviation of elemental properties ("std NpUnifilled", "std NdValence") are also among the most impactful properties with respect to ΔH_f^{diff}. For example, with smaller "band center" (geometric mean of electronegativity[135]), ΔH_f^{diff} tends to be larger and ΔH_f^{DFT} tends to be smaller than ΔH_f^{exp}, which means that DFT tends to underestimate ΔH_f of systems with smaller mean electronegativity. A possible explanation for this trend is that, smaller geometric mean of electronegativity, the ability of atoms to bind the electrons near the atomic nuclei is weaker, and electrons tend to be delocalized. Since the GGA approximation tends to overestimate the electron delocalization[225], ΔH_f^{DFT} tends to be more negative for the systems with delocalized bonds (stronger bonding). Another example is, with larger standard deviation of number of *p* valence electrons, ΔH_f^{diff} tends to be smaller and ΔH_f^{DFT} tends to be larger than ΔH_f^{exp}, suggesting that DFT tends to overestimate ΔH_f of systems with more dissimilar *p* valence electron configurations. This trend might be explained by the hypothesis that, with more different *p* electron configuration, in general the compound is more

ionic, and because of the fact that GGA approximation tends to underestimate the electron localization[225], DFT (PBE) ΔH_f tends to be more positive for the systems with localized bonds (weaker bonding).

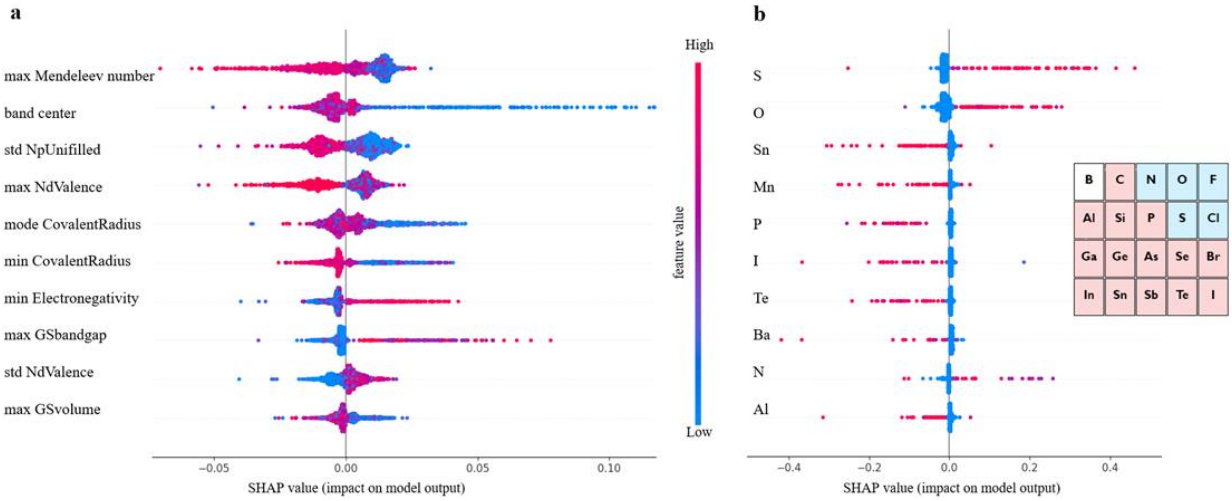

Figure 3-4. Impact of each feature on model output. **a** and **b** Distributions of the impacts (SHAP values[96]) of compositional features and elemental fractions on the model output (ΔH_f^{diff}), respectively. The color represents the feature value (red high, blue low), and here only the top 10 features and elemental fractions with the highest sum of absolute SHAP values are shown. The inserted figure in **b** illustrates the trends of DFT to underestimate or overestimate ΔH_f of materials with certain non-metal elements. Here, the blue shaded elements are those for which DFT tends to underestimate ΔH_f, the red shaded elements are those for which DFT tends to overestimate ΔH_f, and Boron shows a mixed trend.

As for the impacts of certain cations and anions, or impacts of certain elements, we build a decision tree model that takes stoichiometry as input, and the SHAP values of fraction of each element are plotted in Figure 3-4b. One can see that, with higher atomic fraction of S, O and N, DFT tends to underestimate ΔH_f, while for higher atomic fraction of Sn, Mn, P, I, Te, Ba, Al, DFT tends to overestimate ΔH_f. There are more non-metal elements (6) in the top 10 most impactful elements than metals (2) and metalloids (2). Particularly, there is an interesting pattern of how DFT treats different non-metal elements: as shown in Figure 3-4b, for strong oxidizing non-metal elements in the upper-right corner of the periodic table, including F, O, N, S, Cl, DFT tends to underestimate ΔH_f, while for those non-metal elements with weaker

oxidizing ability, DFT tends to overestimate ΔH_f. However, the degree of overestimation or underestimation doesn't simply correlate with the oxidizing ability. For example, F has stronger oxidizing ability than O and S, but the degree of underestimation of ΔH_f^{DFT} for fluorides is less than that of oxides and sulfides. There are two possible sources of errors that would result in the observed trend: on the one hand, the underestimation or overestimation of ΔH_f of materials with certain elements might come from the element type-based empirical corrections[7, 182], and on the other hand, the intrinsic limit of the GGA and GGA + *U* approximation might cause the different deviation patterns. For example, Seo *et al.*[226] proposed that the GGA + *U* method used for transition metal oxides in the MP database[182] overestimates the degree of hybridization between the *d* orbitals of transition metals and *p* orbitals of oxygen, thus makes the calculated ΔH_f more negative.

The trend in Figure 3-4a also agrees with that in Figure 3-4b. For example, for "max GSbandgap" and "max GSvolume", they are calculated in the following procedure: first the ground state band gaps and ground state volumes of all the elements in the compound are listed, then the maximum values of band gaps and volumes are picked up. Therefore, "max GSbandgap" and "max GSvolume" actually relate to the existence of certain elements in the compound. Specifically, "max GSbandgap" describes the presence of specific anion in the compound while "max GSvolume" describes that of cation. Larger "max GSvolume", ΔH_f^{DFT} tends to be larger (more positive) than ΔH_f^{exp}. An explanation for this trend is that with larger "max GSvolume", the cation element tends to have larger ground state volume (closer to the bottom-left of the periodic table with the maximum value at Cs). If the cation is closer to the bottom-left of the table, the compound in general will be more ionic. Therefore, ΔH_f^{DFT} tends to be more positive for the systems with more ionic bonds as mentioned above. On the other hand, larger "max GSbandgap", ΔH_f^{DFT} tends to be smaller (more negative) than ΔH_f^{exp}. This phenomenon might be explained by the fact that, with larger "max GSbandgap", the anionic

element is closer to the upper-right corner of the periodic table the maximum value at N, and according to Figure 3-4b the compound tends to have more negative ΔH_f^{DFT}.

Note that in Wang *et al*.[183] all anionic corrections are negative, which is because their correction is applied to the original PBE results and PBE tends to overestimate the energy of diatomic gas molecules[227], while the trend shown here is based on the empirically corrected PBE energies from MP that already take the effect of overestimated energy of diatomic gas molecules into account.

3.6. Details of methods

Data collection. In Chapter 3, we construct the ΔH_f^{exp} dataset by combining two datasets from IIT[167] and SSUB[190], and we use the Materials Project[6] database (V2021.03.22) to construct the ΔH_f^{DFT} dataset. For the ΔH_f^{diff} dataset, since the ΔH_f^{DFT} values are provided for some materials in the IIT dataset, ΔH_f^{diff} values for those materials are obtained by subtracting the provided ΔH_f^{DFT} from the provided ΔH_f^{exp}, and for materials from the SSUB dataset, since chemical formula is the only identifier, we take the lowest ΔH_f^{exp} for each formula, and for the ΔH_f^{DFT} of these materials, we assign the lowest ΔH_f^{DFT} to each formula. For overlaps between the IIT dataset and SSUB dataset, we take the ΔH_f^{exp} from the IIT database as the IIT database is a more recent one[167]. Note that the mean absolute difference of ΔH_f^{exp} between our dataset and the recent dataset from Wang *et al*.[183] is only 0.007 eV/atom.

Machine learning models training procedure. In Chapter 3, the dataset of the 1143 ΔH_f^{exp} is used for three purposes: 1) hyper-parameters tuning for each machine learning model, 2) model evaluation, and 3) production, or prediction of ΔH_f of all materials in the Materials Project database (MP). For purpose 1) and purpose 2), we first randomly reserve 20% data as the test set for model selection (these 20% data are also excluded in the larger MP dataset for

transfer learning). Then, to determine the best set of hyper-parameters for each model, with the remaining 80% data, we randomly reserve 20% of the remaining data (20%*80% = 16% of total data) as the validation set to evaluate each specific set of hyper-parameters, and use 80% of the remaining data (80%*80% = 64% of total data) to train the machine learning model with the given set of hyper-parameters. We screen hyper-parameters by grid search, and tables of search space of hyper-parameters are provided in Ref.[97]. Finally, with the found best hyper-parameters for each model, we use the 80% of the data (training set + validation set in the hyper-parameter search step) to train machine learning models 10 times with different initialization, and evaluate model performance and uncertainty using the 20% data held out at the very beginning (test set). For purpose 3), production, for best prediction performance, all available 1143 data points are used to train the found best model with the found best hyper-parameter.

We use four different machine learning architectures to realize transfer learning and/or multi-fidelity machine learning, random forest (RF), multi-layer perceptron (MLP), Representation Learning from Stoichiometry (ROOST)[23] and Crystal Graph Convolutional Neural Network (CGCNN)[71]. For ROOST, we feed the compositions of materials as input, and it learns the representations of materials, and for CGCNN, we feed the 3D atomic structures of materials as input, and it also learns the representations. RF and MLP are realized by scikit-learn[19], and we use the descriptors from matminer[135] to feed RF and MLP as features of materials. Modules used to generate compositional features are Element Property, Electron Affinity, Band Center, Cohesive Energy, Miedema, TMetal Fraction, Valence Orbital, Yang Solid Solution, and modules used to generate structural features are Global Symmetry Features, Structural Complexity, Chemical Ordering, Maximum Packing Efficiency, Minimum Relative Distances, Structural Heterogeneity, Average Bond Length, Average Bond Angle, Bond Orientational Parameter, Coordination Number, and Density Features.

Energy above hull. In the Materials Project (MP), the energy above hull (E_{hull}) is defined as the energy of decomposition of a material into the set of most stable materials at this chemical composition[6]. The decomposition is tested against all potential chemical combinations that result in the material's composition. A positive E_{hull} indicates that this material is unstable with respect to decomposition, and a zero E_{hull} indicates that this compound is stable with respect to decomposition. In this chapter, the energy above hull is defined in the same way as MP. Phase Diagram module in Pymatgen[136] is used to calculate the E_{hull}. The inputs required by the Phase Diagram module are the compositions and formation enthalpies, and the corresponding output is the energies vs. compositions diagram, from which the decomposition energies and E_{hull} can be calculated.

3.7. Chapter summary and outlook

In Chapter 3, we conduct a comprehensive machine learning study to learn and predict experimental formation enthalpy of materials. We use two different strategies to transfer information from larger DFT dataset to the smaller experimental dataset, transfer learning and multi-fidelity machine learning, and we use four machine learning architectures to realize the two strategies. We find that the random forest model trained on the difference between experimental and DFT formation enthalpy with DFT formation enthalpy as one of the input features can achieve the lowest error, which is almost half of that of DFT (empirically corrected PBE), and it also outperforms other more accurate but more computationally expansive density functionals, such as meta-GGA functionals. Beyond identifying the best model, we suggest that the deep neural network-based representation learning algorithms and transfer learning should not be the only machine learning architecture and information-transfer strategy considered. Other feature engineering methods such as human-engineered features, machine learning architectures beyond neural networks such as random forest and information-transfer

strategy such as multi-fidelity machine learning should also be tested in machine learning applications for materials science.

As an application, we employ the found best random forest model to calibrate the formation enthalpy of all materials in the Materials Project database, which are then used to construct energy above hull and discover potential important materials that have underestimated stability in the MP database. Further, we use the machine learning model as a data-mining approach to identify patterns in the performance of DFT, for example in its tendency to underestimate the formation enthalpy of materials with elements in the upper-right corner of the periodic table.

Note that Chapter 3 is based on the Materials Project database queried in March, 2021 (V2021.03.22). The methodology of this chapter can also be applied to updated Materials Project database (such as V2021.05.13) and other large DFT databases. It is expected that, with more accurate low fidelity data (DFT formation enthalpy), such as the recent dataset with 6,000 materials calculated by meta-GGA functionals[127], the method in this chapter can be used to provide more accurate calibration (exp. formation enthalpy).

One potential limitation of the multi-fidelity model used in this chapter is that it requires the availability of low-fidelity data for the whole materials space of interest, as in this chapter DFT formation enthalpy is required for learning the difference of formation enthalpy from experiment and DFT. In cases where low-fidelity data is not available to all the materials, transfer learning might be more appropriate to transfer information between different datasets. Another scenario not considered in the current multi-fidelity machine learning scheme is that, for some properties there might be datasets with multiple levels of fidelity available. In such cases, in addition to incorporating different fidelity data into the input, the learning of differences might be conducted multiple times to enlarge the availability of high-fidelity data gradually.

More broadly, for machine learning applications with small datasets, choosing proper models and strategies is critical to the usefulness of the machine learning models. On the one hand, with small datasets, one should carefully compare the performance of deep representation learning and classic machine learning models based on *off-the-shelf* featurization, and make the choice for production. Typically, with more than 10,000 data points, deep representation learning might be more powerful; with less than 500 data points, classic machine learning models might be more suitable; and with more than 500 but less than 10,000 data points, careful comparison is necessary for employing a suitable model for production. On the other hand, if larger low-fidelity datasets are available, then information transfer might be useful to improve the learning and prediction of the high-fidelity data. There are two strategies, transfer learning and multi-fidelity learning, for the information transfer. Although there still lacks theoretical guarantee or quantitative metric to estimate whether information transfer would help or not, empirically, the two strategies worth a try if the high- and low-fidelity datasets are "strongly" correlated.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

4. Charting lattice thermal conductivity for inorganic crystals by machine learning

4.1. Introduction

In Chapter 4, we present a case study of lattice thermal conductivity of materials to show how machine learning can be used to promote the development of functional materials, as well as how transfer learning can help to learn small experimental dataset. Different from Chapter 3, where multi-fidelity machine learning is also employed together with transfer learning to transfer information from larger datasets, here we only employ transfer learning, because as mentioned below, the larger dataset in this case covers only a small portion of the whole materials space. Therefore, multi-fidelity learning is less useful than transfer learning as multi-fidelity learning cannot predict properties of materials without low-fidelity data.

Thermoelectricity produced from usually negative-valued heat is a green and promising candidate on the future energy landscape. The most effective thermoelectric materials exhibit low thermal conductivity κ. However, less than 5% out of about 10^5 synthesized inorganic materials are documented with their κ values, while for the remaining 95% κ values are missing and challenging to predict. In this chapter, by combining graph neural networks, transfer learning, and random forest approaches, we predict the thermal conductivity of all known inorganic materials in the Inorganic Crystal Structure Database, and chart the structural chemistry of κ into extended van-Arkel triangles. Together with the newly developed κ map and our theoretical tool, we identify rare-earth chalcogenides as promising candidates, of which we measured zT exceeding 1.0 as the highest record thus far. Still, we note that the κ chart can be further explored, and our computational and analytical tools are applicable generally for

materials informatics.

The discovery of Seebeck and Peltier effects has enabled driving charge flows by heat and vice versa[228, 229], which powers the Explorer in the deep space and cools medicine/specimen on pharmaceutical sites[149, 230]. Such thermoelectric mechanisms attracted waves of research during the past with the development of condensed matter physics, and the current decade has seen another recurring tide of interest as alternative green energy source for better environment. However, the barrier for the large-scale technical translation of thermoelectrics is its efficiency (e.g., <5% for most thermoelectric materials on the market[228, 231]. On top of the electrical properties, the controlling factor is unstoppable heat flow, which gives rises to irreversibility and is governed by thermal conductivity, higher efficiency entailing lower thermal conductivity.

In fact, solids with both low and high extreme thermal conductivity have been pursued fundamentally and practically for decades. Currently the records are held by diamond (~2000 W/mK)[232] in the upper limit and aerogels (~0.01 W/mK) on the lower end[233], although it remains unclear whether these are hard limits. Regardless, the search for alternative materials that lie at or beyond these extremes is also of practical importance, particularly when multiple constraints are imposed, such as specific mechanical properties for thermal coatings[234] and (opto-) electronic properties for applications in energy conversion[228, 235]. More than thermoelectrics, the diverse applications range from thermal management in electronics and avionics[236], to high-temperature coatings in turbines[234] and human healthcare[237], to name only a few examples.

However, knowledge of the governing physics of (lattice) thermal conductivity (κ) remains incomplete at the atomic scale[238, 239]. Current understanding derives largely from kinetic theory and relates to unit cell properties (e.g., (average) atomic (mass, density),

symmetry)[240]. This understanding has been historically encapsulated into analytical models, such as the Debye-Callaway (D-C) model[241] and its variants[238]. Similarly, analytical models for κ of solid-solution alloys, such as the Klemens model[242], are based on unit cell properties and scattering parameters. These models are explicit, but have parameters either numerically fitted or computed from first principles. For instance, Miller *et al.* developed a modified D-C model with speed of sound and Gruneisen parameter, which are derived from bulk modulus and average coordination number[243].

An emerging approach has been driven by learning from the existing data of κ, benefited from the developments in high-throughput screening and machine learning methods[229, 244-247]. Through high-throughput calculations, databases are growing in size via approaches for computing κ based on Green-Kubo formalism[248, 249] and Boltzmann theory[240, 250]. However, relying on dynamical and/or large-scale first-principles calculations, these methods are often computationally expensive, and most high-throughput studies are limited within certain material families[246]. Alternatively, the above semi-empirical models have also been successfully implemented for high-throughput predictions[126]. Experimental data is even less available. To date, only some hundreds of the total ~10^5 synthesized materials documented in the Inorganic Crystal Structure Database (ICSD) have κ values measured[251]. Thus, while machine learning techniques have shown initial success[54, 239, 252], both more data and novel approaches are needed in order to explore the vast materials space.

Towards this end, general guidelines for navigating and sampling the materials space for κ will be valuable. Existing works for predicting/understanding κ exhibit a catch-22 situation. On the one hand, descriptor-based methods assume *a priori* knowledge of the physics of κ, so that appropriate features could be populated for materials[252]. However, since structural chemistry of κ is largely unknown, the choice of atomic features is currently arbitrary. On the

other hand, techniques based on graph neural networks assume little pre-knowledge of κ, and can predict material properties directly from structure[71]. However, these methods are "black-boxes", and the challenge of interpreting structure- κ relation remains.

In this chapter, we predict κ of all ordered and stoichiometric materials in ICSD (92,919 entries), and then reveal the structural chemistry of κ. Two complementary approaches, neural networks and random forest, are thus combined. While the former predicts κ directly from structures with little need for featurization, the latter extracts the hidden chemistry in the dataset. With resolved important atomic and structural features that govern κ, we are able to chart the structural chemistry of κ using our generalized van-Arkel triangles. Aiming at learning and predicting κ measured by experiments, we build an experimental dataset κ_{exp} collected from the literature, and extend our earlier graph neural networks model[71] with transfer learning (TL-CGCNN). Based on the charts, we identify a set of rare-earth chalcogenides, as a new class of promising thermoelectric materials, of which the figure of merit shows 1.1 at 800 K.

4.2. Machine learning study of lattice thermal conductivity

We start by learning from our recently prepared high-throughput κ_C dataset[126], before moving to the broader ICSD and the underlying structural chemistry. The κ_C dataset contains computed κ_C of 2,668 ordered and stoichiometric inorganic structures from the ICSD. The predicted κ is fairly accurate, with an average factor difference of 1.5 from experimentally measured values, over a range of κ values that span 4 orders of magnitude[243]. In Chapter 4.2, we will show both the transferability and limitation of this dataset, and in Chapter 4.3 we will show its implicit physics. Note that these two purposes suit two separate but complementary machine learning models: crystal graph convolutional neural network (CGCNN)[71], and interpretable random forest. These models are illustrated in Figure 4-1(a), with further details

available in Chapter 4.6. For our high-throughput dataset, we randomly reserve 20% entries as the test set, as plotted in Figure 4-1(b). Both CGCNN and random forest models could predict $\log\kappa_C'$ with MAE<0.15 and R^2>0.85.

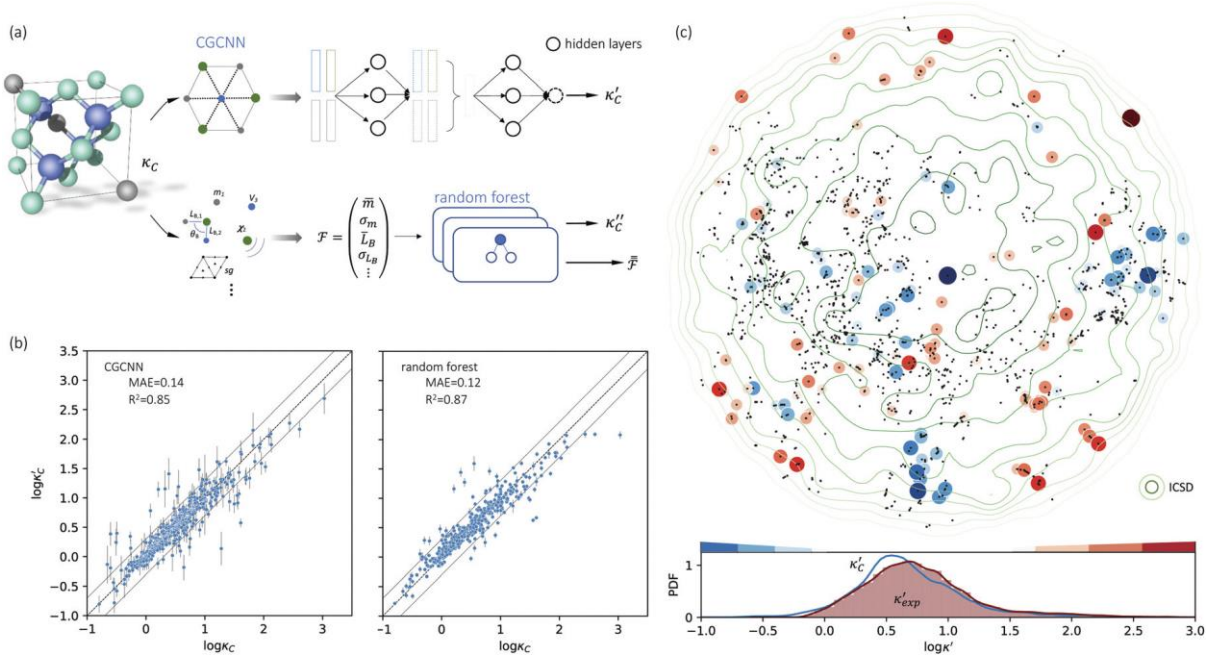

Figure 4-1. a Schematic of two complementary models: CGCNN and random forest. **b** Predicted $\log\kappa_C'$ from these two models. The dashed band denotes a factor of 2. **c** High-throughput $\log\kappa_C'$ for all ordered ICSD structures. The contour denotes the distribution of ICSD materials in the feature space reduced to 2D via PCA/t-SNE, along with the training set denoted by the dots. The histograms are the distribution of predicted $\log\kappa_C'$ and $\log\kappa_{exp}'$. See text for the prediction of $\log\kappa_{exp}'$.

Moreover, different from CGCNN, random forest requires featurization for crystal structures before running through decision trees, which is largely physics-based and in many cases *ad hoc*. Guided by lattice dynamical theory, we choose configurational features from elemental to atomic packing and bonding nature, which are constructed through Matminer[135], Magpie[25], and in-house codes. Since κ is sensitive to both absolute values and variations of atomic properties, our feature engineering leads to a 154-dimensional descriptor, including the statistics (mean $\bar{}$, standard deviation σ, range[245] and mode) of atomic number, covalent radius r_a, atomic mass m, periodic table group and row number, Mendeleev number, volume

per atom from ground state V_{GS}, Pauling electronegativity χ_a, melting point T_m, number of filled N_V and unfilled N_U valence electrons in the *s*, *p*, *d*, and *f* shells of constituting elements, as well as structural features at the cell scale (space group, volume per atom V_a, packing fraction φ, density ρ, bond length L_B, bond angle θ_B, and coordination number *CN*).

To visualize the feature space, we project it onto two dimensions, as shown in Figure 4-1(c). Materials from our high-throughput dataset and the ICSD dataset are considered together, denoted by the scattered points and contour respectively. Note that the *x* and *y* axes are abstract linear combination of all structural features. On this projected materials-feature space, the contour lines show the distribution of all inorganic materials. Deeper color indicates more materials existing in ICSD (we have removed the contour levels though to stress that the magnitude is relative). The contour shows that most materials are populated in the central area, and the distribution varies smoothly, thus friendly for machine learning algorithms. Our high-throughput entries (scattered points) with the highest and lowest κ values highlighted, samples the reduced feature space quite satisfactorily in terms of uniformity. This suggests the potential transferability of our high-throughput dataset to ICSD. We did so using both CGCNN and random forest models. From the histogram in Figure 4-1(c), the distribution of predicted κ_C' follows approximately a normal distribution, with mean $\log\kappa \approx 0.8$ and standard deviation of $\log\kappa \approx 0.5$.

To further validate our machine-learning predictions, we compare them to experimental measurements, and/or to first-principles calculations (see details of experimental and computational methods in Chapter 4.6). More than measurements in the literature, we also chose 12 materials from different structures/compositions/families, and measured their κ. The comparisons are presented in Table 4-1 for several low- and high-κ materials. Overall, our machine learning models can unanimously screen the lowest from the highest, which might be

already sufficient for many materials selection/design scenarios, such as for thermoelectrics and thermal management, where either the lowest or the highest κ values are sought. For instance, in Table 4-1, we have identified rare-earth chalcogenides (REX) as promising thermoelectric materials, which are interesting for future exploration (See below). The other reason that we test our machine-learning models with these extremes is to show their reliability for extrapolation (transferability), which is often more challenging numerically than interpolation.

Table 4-1. The predicted candidates in the lower and upper limits. Note that $\kappa_{exp}{}'$ is from a random forest model for the low regime of κ, and TL-CGCNN for high values. The entries without references are measured/calculated in Chapter 4.

	$\log\kappa_{exp}$	$\log\kappa_{DFT}$	$\log\kappa_{C}{}'$	$\log\kappa_{exp}{}'$
Cu_2HfTe_3	-0.016		0.016	0.022
Cu_3VTe_4	0.19		0.26	0.28
$TaCoTe_2$	-0.21		-0.32	0.052
$AgAlTe_2$	-0.21		-0.36	0.042
$FeIn_2S_4$	0.16		0.58	0.46
$NbTe_4$	0.28		0.30	0.36
TiFeCoGa	0.69		0.86	1.09
Er_2Se_3	0.15		0.21	0.071
Er_2Te_3	0.19		0.18	0.32
Tb_2Te_3	-0.027		0.15	0.21
Dy_2Te_3	0.0056		0.18	0.22
Ho_2Te_3	0.16		0.20	0.32
$Cs_2BiAgCl_6$		-1.2	-0.1	-0.3
$CsTlF_3$		-1.0	0.2	-0.1
$CsTlI_3$		-1.3	-0.1	-0.3
$CsPbI_3$	-0.4[47][253]	-1.0[20][235]	-0.2	-0.2
Tl_3VSe_4	-0.5[9][254]	-0.8[9a][254]	-0.2	-0.3
Be_2C		2.06	2.9	2.6

C_3N_4		2.4	2.5	2.6
BP	2.6[48][255]	2.82[15][240]	2.4	2.6
BAs	3.08[12-14][256]	3.50[15][240]	2.0	2.2
Diamond	3.36[17][232]	3.54[15][240]	3.1	3.4

More quantitatively, the error of our machine learning models is comparable to first-principles calculations based on DFT ($\log\kappa_{DFT}$). For instance, in the case of diamond, the extrapolated values, $\log\kappa$, 3.1 and 3.4, are close to the experimental value 3.36, comparing to 3.54 from DFT calculations. Such level of error applies to all examined entries, except several outlying cases, such as BAs, for which the accuracy is less satisfactory.

Other possible outliers are also observed when experimental values are missing and a substantial difference can be seen between DFT and machine learning, such as $CsTlF_3$ in Table 4-1. However, such possible outliers should be further examined (experimentally preferred) due to the possible underestimation from DFT calculations. In some cases, a difference of 50 - 100% between DFT and experimental values can arise from the relaxation-time approximation up to 3-phonon interactions, which might be resolved by more sophisticated calculations, such as four-phonon and temperature-dependent dispersion[254, 257]. In many other cases, our machine learning prediction can be even more accurate than DFT, such as the iodide perovskite $CsPbI_3$ and the recently studied Tl_3VSe_4 (see Table 4-1). Moreover, note that our above error analyses is based on extrapolation. Even for the highest and lowest values, the machine learning models show satisfactory stability and prediction accuracy.

Nevertheless, our machine learning model is still limited by the quality and finiteness of our dataset. Since the training set used is the largest reliable dataset available, this limitation will be translated to guidelines for future high-throughput calculations. This is discussed further as we extend CGCNN with transfer learning (TL-CGCNN) to predicting experimental values $\log\kappa_{exp}'$ (see Chapter 4.4). The top 50 lowest-κ and highest-κ values are uniformly scattered,

suggesting little knowledge content. However, as we present in Figure 4-2(a), these top 100 points are clustered when we plot without ICSD. This is another indication of the limited transferability to ICSD, but also demonstrates the knowledge content in our known dataset.

4.3. Data mining of lattice thermal conductivity

Such knowledge content can be extracted in the form of ranked features (details in Chapter 4.6). In Fig. Figure 4-2(b), the top 20 features are ranked in decreasing order. These features include the elemental type (V_{GS}, N_V, N_U, m) and structural type. The latter consists of bonding properties (L_B, θ_B, CN), and packing properties (V_a, Dim, φ, SG, ρ). The learning of important features is different from a simple correlation relation. Figure 4-2(c) shows the MAE as a function of increasing number of features, picking from the most important features, from PCA and random forest respectively. As the number of features increases, MAE reduces quickly and reaches our CGCNN accuracy with less than 10 features, and both are lower than PCA. The latter is usually chosen when little pre-knowledge is assumed, and our case shows that such purely data-driven techniques (e.g. PCA for dimensional reduction) could be excelled over by physics-informed approaches. Another interesting application of these important features is to physically categorize/cluster all the training materials. An example is shown in Figure 4-2(d), where high- κ and low- κ values could be separated by the dashed line.

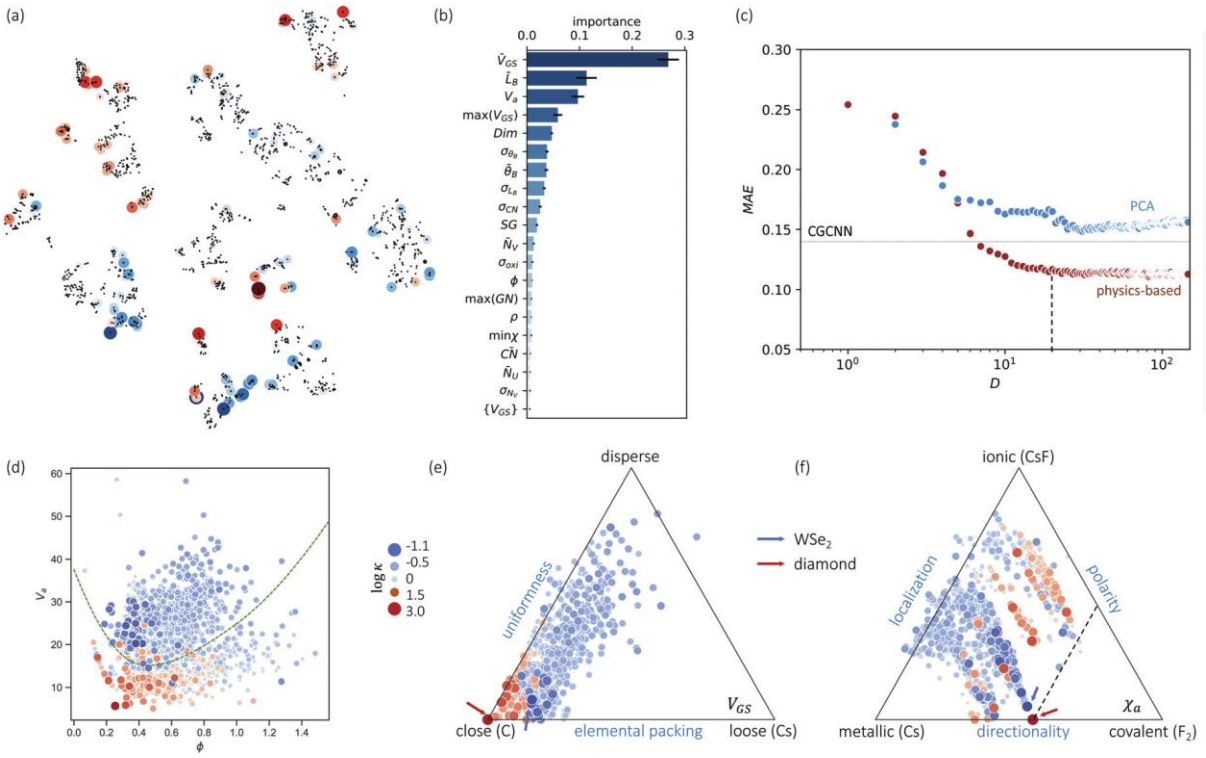

Figure 4-2. **a** Clustering of the high-throughput database using PCA and tSNE, low-κ and high-κ entries are highlighted. **b** Top 20 important features and their F scores. **c** Dimension reduction by random-forest-ranked feature selection lead to even lower than PCA, and MAE approaches to CGCNN around 10 atomic features. Low-κ and high-κ materials can be divided by important features, **d** is an example of using φ-V_a. **e-f** Chemical space illustrated by van-Arkel triangles, examples of elemental (V_{GS}) and bonding (χ_a) information.

Further, phonon transport is sensitive to chemical variations, more than corresponding mean fields. Examples are mass and bond strength: the mean values define mean-field harmonic properties (e.g., group velocity), while the differences determine both harmonic (e.g., phononic bandgap) and anharmonic properties (e.g., higher-order force constants). This is also suggested in Figure 4-2(d), where both mean values and variances are ranked most important, such as L_B, θ_B, *CN*, and N_V. Note that our machine learning models start from different feature list from that of our D-C model. For instance, none of the crucial variances enter the D-C model. This is also true for the past predictions of harmonic properties, such as Debye temperature and vibrational entropy[251, 258]. Despite the partial overlap between our important feature list

and those for harmonic-property predictions, which is expected because κ is determined by both harmonic and more challenging anharmonic properties, the newly revealed variance and how the mean-variance information together impacts κ is unknown. More importantly, other than widely-applied correlograms, an analytical tool to study this is still missing.

Inspired by various forms of van-Arkel-type triangles, we use mean and standard deviation to construct extended triangles and generalize extensively to other atomic features (see Chapter 4.6). Invented originally for binary inorganic compounds, van-Arkel-type triangles were historically constructed to characterize the bonding nature, using the average and difference of the two elements' electronegativity χ_a. In our case, we have multi-component compounds and more dominant quantities than χ_a. Therefore, we extend the original van-Arkel triangle to include more components with mean and standard deviation, and to more physical descriptors important for κ. For instance, the V_{GS} and χ_a triangles shown in Figure 4-2(e-f) characterize packing and bonding information, respectively. Although the extension is straightforward, it helps to chart the structural chemistry of κ. For instance, each of these triangles illustrates a projected materials space, within which all materials should be confined. While the coverage is essential for validating our dataset, it is also interesting to note that many of the chosen features are effective divisors. In other words, given the mean and deviation of any of these features for a unit cell, the relative magnitude of κ can already be estimated.

Note that our work confirms and also enhances our existing understanding of trends in κ. For instance, it is commonly established that low-κ materials often have i) high average atomic mass, and ii) weak interatomic bonding, so that group velocity can be low, and iii) high anharmonicity in order to have short relaxation time (e.g. more scattering channels resulting from complex crystal structures). However, bonding strength and anharmonicity are computationally expensive quantities. Meanwhile, predicting κ directly from atomic structures

was at best qualitative in the literature. With our analysis based on Figure 4-2, we now have proxies for bond strength and even κ. Moreover, our identified structural features have only partial overlap with the past works on learning vibrational properties[251, 258]. In particular, comparing to the learning of harmonic properties, these mean-variance pairs which inspired our extension of van-Arkel triangles also suggest the importance of structural variance and complexity in anharmonicity.

4.4. Transfer learning of experimentally measured thermal conductivity

Because κ_C still deviates from κ_{exp} by an order of 1.5[157], learning κ_C might inherit the error of κ_C. Therefore, directly learning κ_{exp} might avoid learning the error of κ_C. In order to exploit knowledge learned from the larger calculated dataset and promote the learning of the small experimentally measured one, we develop a transfer learning scheme (see Figure 4-3(a)), which is based on the idea that correlated datasets share similar domain knowledge. Here, multi-fidelity learning is less useful than transfer learning, as the larger κ_C dataset only has less 3,000 data points, and multi-fidelity learning cannot predict properties of materials without low-fidelity data.

To take advantage of the knowledge learned from our larger high-throughput dataset, we develop a transfer learning framework demonstrated in Figure 4-3(a). This transfer learning scheme we used to predict experimental conductivity is a two-step modified CGCNN model: i) training a CGCNN model on our high-throughput κ_C dataset to extract knowledge, which has been done in the main text. ii) transferring the parameters of all layers from step i) to initialize a second CGCNN to transfer knowledge, and add one extra layer before the output layer to account for the difference between the two datasets. For the second step, we use the smaller κ_{exp} dataset (132 entries, see Ref.[55]) collected from experimental measurements in the

literature. Since the experimental dataset is very small, in step ii), all the layers other than the last one are frozen to keep the pre-learned knowledge and reduce the degrees of freedom to suppress overfitting.

With this transfer learning scheme, we predict directly experimental values here using CGCNN, but with high MAEs (see Figure 4-3(b)), due to small size of the experimental dataset, $<10^3$ entries. The overall performance is compared with random forest and CGCNN in Figure 4-3(b), using different training datasets, and as can be seen our TL-CGCNN leads to the lowest MAE. Figure 4-3(c) plots the improvement for each data in the test set, defined by the absolute error difference between CGCNN and TL-CGCNN. It can be seen that the accuracy on the high-κ end ($\log\kappa > 1$) is improved, but the accuracy is deteriorated on the low-κ end, even though the overall performance is enhanced. Some example predictions in the high-κ limit from step ii), can be found in Table 4-1. In the $\log\kappa < 1$ region, we recommend $\kappa_{exp}{}'$ from random forest.

To understand the different performance in the high- and low-κ regions of the transfer learning model, we look into the space of crystal features in the neural networks. In Figure 4-3(a), the network before the last hidden layer learns the feature vectors of materials $\boldsymbol{V}_\mathbf{f}$, and the last operation from $\boldsymbol{V}_\mathbf{f}$ to output is simply a regression with softmax activation. Since in TL-CGCNN we freeze $\boldsymbol{V}_\mathbf{f}$ and all layers before the extra layer due to the limited amount of data, we essentially use a one-layer neural network to fine tune $\kappa_{exp}{}'$ learnt from $\kappa_C{}'$. We plot $\boldsymbol{V}_\mathbf{f}$ from the high-throughput and experimental datasets in Figure 4-3(d). Interestingly, we observe a similar distribution between κ_{exp} and κ_C in the $\boldsymbol{V}_\mathbf{f}$ space, showing a strong correlation between the two datasets. However, in the high-κ region, κ_{exp} distributes more smoothly along the V-shape than in the low-κ region, which explains why TL-CGCNN performs better in the high-κ. Such issues in the low-κ region can be tackled from two aspects: i) more experimental data

with low κ should be generated to better understand the κ_{exp} distribution, and ii) future high-throughput calculations should be refined to shrink the difference between κ_C and κ_{exp}, especially the outliers, in order to better sample the experimental $\boldsymbol{V_f}$ space. The observation of data bias indicates the need to expand the current database. Instead of calculating hundreds of candidates in a certain material family each time, feature-space-based sampling techniques may be more computationally efficient to cover the material space.

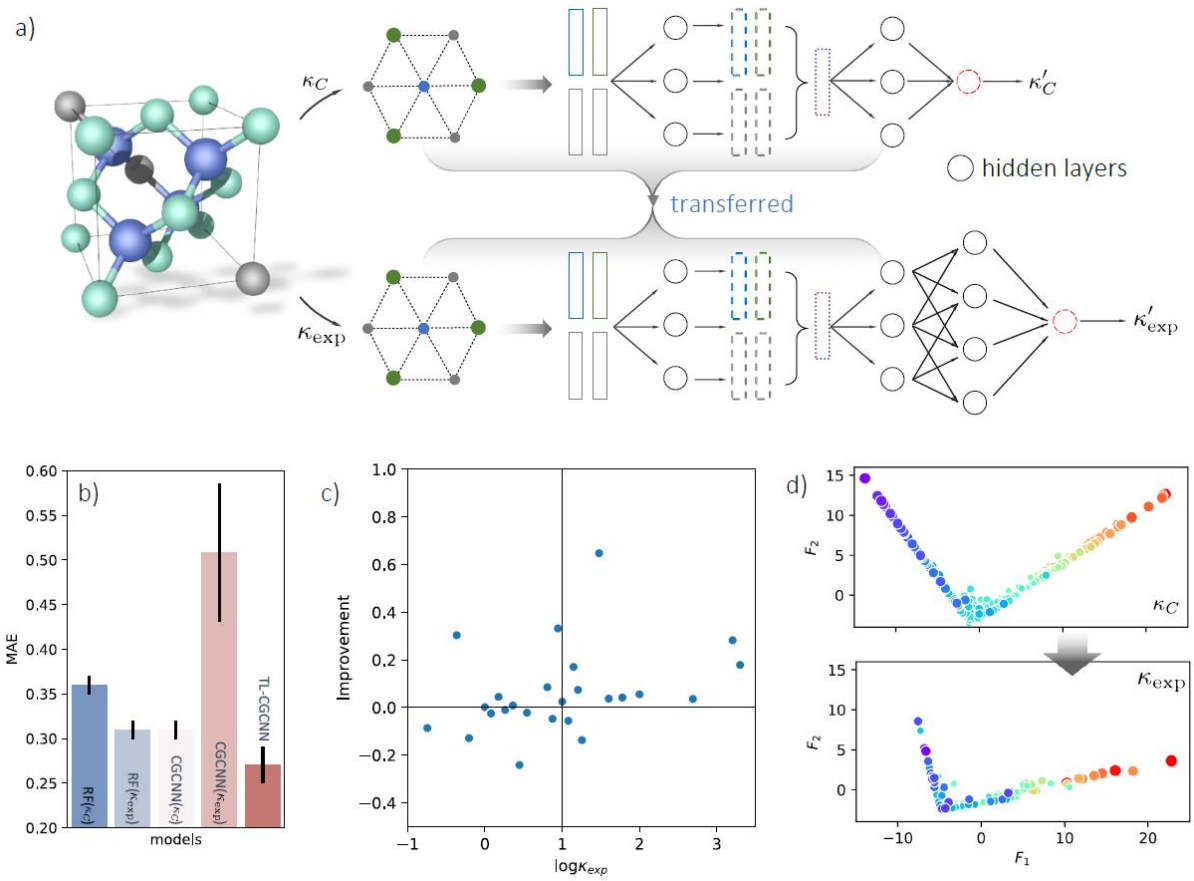

Figure 4-3. **a** This model learns high-throughput dataset κ_C and transfer the knowledge to learning κ_{exp}. **b** Comparison between different machine learning models, including random forest, CGCNN, and TL-CGCNN, trained on κ_C or κ_{exp}. TL-CGCNN exhibits the lowest MAE. **c** A closer look at the improvement of TL-CGCNN compared with CGCNN (κ_C) in prediction on the test set. The region of logκ > 1 is systematically enhanced, while the logκ < 1 region can be better or worse. **d** The distribution of the feature space V_f projected onto two dimensions. The distribution and ranking of κ_C is generally smoother than κ_{exp}, and for κ_{exp} the upper end is smoother than the lower end.

4.5. Discovery of rare earth chalcogenides for thermoelectrics

The structural chemistry of κ can be used to extend the predictions from machine learning. For instance, in the upper limit, machine learning predicts the κ values for BN and diamond to be 764 W/mK and 2225 W/mK, which are close to experimental values. As shown in Figure 4-4(a), from the van-Arkel triangle of χ_a, we notice two candidate materials between BN and diamond: C_3N_4 and B_4C_3. The κ_{exp}' of C_3N_4 ranked in the top 1% in our machine learning predictions over ICSD. In contrast, B_4C_3 is absent from ICSD, and is obtained by reading the van-Arkel triangle. One can also use this approach to search for low-κ materials. Guided by the triangles, we adapt the corner of thalium, and iodine, considering their atomic weight and electronegativity. As shown in Figure 4-4(b), binary and ternary compounds (e.g. TlI, $CsTlF_3$, $CsPbI_3$) are predicted from machine learning. Based on these, we could hypothesize that $CsTlF_3$ would have a low κ, which is also absent from the ICSD and confirmed by our DFT calculations (Table 4-1).

Another group of the least thermally conducting materials are the REX family. As mentioned above, the REX materials rank the lowest 5% in the κ chart. To further confirm their transport properties, we show in Figure 4-4(c) the temperature-dependent thermal conductivity of six compounds (Er_2Se_3, Er_2Te_3, Tb_2Te_3, Dy_2Te_3, Ho_2Te_3, and Y_2Te_3) that belongs to the REX family. Note that the electronic contribution to the thermal conductivity is negligible since these materials are insulators. We obtain fairly low κ of these compounds with minimum values of 0.5 to 0.6 $Wm^{-1}K^{-1}$ at 973 K for several compounds such as Er_2Te_3, Tb_2Te_3, and Dy_2Te_3. The κ values of REX are comparable with Zintl phase $Yb_{14}MnSb_{11}$[259], and lower than SiGe bulk alloy[260] and half-Heusler ZrNiSn[261]. The low κ suggests the potential of these materials for thermoelectric applications. Advanced thermoelectric materials demand decent electronic transport performance, which can be enabled by aliovalent doping to modify the

Fermi level.

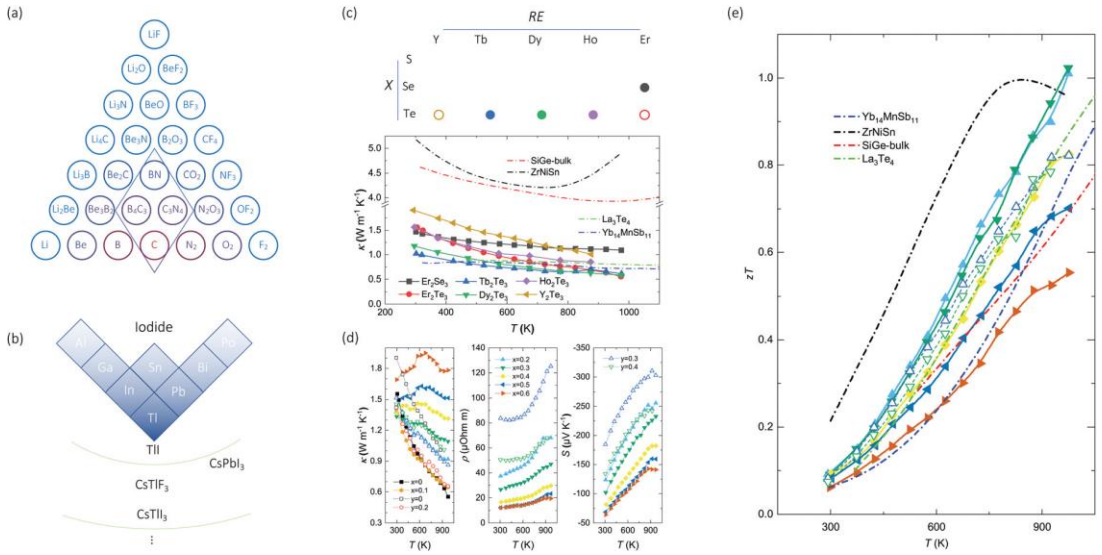

Figure 4-4. Proposed searching directions of **a** high- and **b** low-κ materials. While C_3N_4 exists in ICSD and is recommended by TL-CGCNN, van-Arkel analysis suggests B_4C_3 (absent in ICSD) to have high κ as well. **b** is part of periodic table that m and χ_a are both large, based on which binary/ternary compounds are recommended (TlI, $CsTlF_3$, $CsPbI_3$) and hypothesized ($CsTlF_3$). **c** The proposed REX system, and the temperature-dependent thermal conductivity of 6 chosen REX materials. The materials marked empty are chosen for further thermoelectric measurements. **d** Temperature-dependent thermal conductivity, electrical resistivity, and Seebeck coefficient of compound series $Er_2Te_{3-x}Bi_x$ and $Y_2Te_{3-y}Bi_y$ with x= 0, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6, and y =0, 0.2, 0.3 and 0.4. (e) Temperature-dependent zT of REX, compared to $Yb_{14}MnSb_{11}$ (Zintl phase[259]), ZrNiSn (Half-Heusler[261]), SiGe alloy (bulk alloy[260]), and La_3Te_4 (REX).

Among the REX compounds with charted thermal conductivity, we select Er_2Te_3, and Y_2Te_3 for case study to investigate their full-thermoelectric properties through partial substitution of Bi at the Te sites. Figure 4-4d shows the temperature-dependent thermal conductivity, which increases with the content of Bi, especially at elevated temperature. Such a thermal-conductivity increase has an electronic origin due to reduced electrical resistivity, which is also shown in Figure 4-4d for compound series $Er_2Te_{3-x}Bi_x$ and $Y_2Te_{3-y}Bi_y$ with x =0.2, 0.3, 0.4, 0.5 and 0.6, and y =0.3 and 0.4, whereas the resistivities of the compounds with x =0, 0.1, and y=0, 0.2 are not shown since they are too high to measure. Generally, the substitution of Bi yields

reduced electrical resistivity for both series, which is accompanied by the reduced Seebeck coefficient (*S*). The combination of $S^2\rho$, termed as the power factor, exhibits a maximum of 1.15 $mWm^{-1}K^{-2}$ for compounds with x =0.3 at 973 K, which is comparable to some advanced TE materials such as Cu_2Se[262] and SnSe[263]. The combination of power factor and thermal conductivity yields the thermoelectric figure-of-merit, *zT*, which shows a peak exceeding 1.0 at 973 K for $Er_2Te_{2.7}Bi_{0.3}$ with an increasing trend, thus suggesting even higher *zT* is possible at higher temperature. The obtained *zT* for $Er_2Te_{2.7}Bi_{0.3}$ is comparable to some advanced TE materials, such as Zintl phase $Yb_{14}MnSb_{11}$[259], Half-Heusler (ZrNiSn[261]), bulk alloy (SiGe[260]), and another REX (La_3Te_4). Our reported *zT* has higher value at either high temperature or the whole temperature range. Er_2Te_3, and Y_2Te_3 are two examples of the REX system, which merits further exploration for high-temperature thermoelectrics.

4.6. Details of methods

Random forest, feature ranking, and dimension reduction. Random forest is an ensemble method that combines multiple decision trees. This model has been used as both classifier and regressor for materials informatics. In contrast to neural networks, random forest models are interpretable by providing an intrinsic metric to evaluate the importance of individual descriptors. We use this advantage of random forest in the main text to extract the important structural features that dominate κ. We use random forest implemented in scikit-learn[19]. The number of trees are set to 50 for all calculations, but the random states are randomly selected when studying the uncertainty in predictions. Dimension reduction has been performed through two approaches: i) principal component analysis (PCA) combined with t-distributed stochastic neighbor embedding (t-SNE). PCA is a linear reduction approach using singular value decomposition, and t-SNE converts similarities between data points to joint probabilities then

minimizes the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. We reduce the 154-dimensional feature space into 20 dimensions using PCA, then visualize the feature space in two dimensions by t-SNE analysis. In essence, this reduces our feature space into 2 dimensions, and enables direct visualization. PCA and t-SNE are both implemented in scikit-learn[19]. ii) Another approach is based on feature selection from random forest. Random forest ranks the importance of features, with which we could reduce the feature space till the performance (e.g., MAE) converges. This process could be more physics-based than the purely data-driven approach in i).

Hyperparameter optimization. In this chapter, we tune the following hyper-paramters by grid-search: number of convolutional layers, length of atom feature vectors, length of hidden layer vectors, learning rate and type of optimizer, and for the last layer of the transfer learning scheme, the length of the layer and regularization term are considered. Descriptions of the hyperparameters for CGCNN are provided in Ref.[71]. For training the larger theoretical dataset, cross-validation is done by randomly selecting 20% of the data as the validation set, and for the small experimental dataset, a 5-fold cross-validation is used. In order to account for the random effect in training neural networks, for each parameter setting the training is repeated 20 times. We used both a Bayesian random search and deterministic grid search to optimize the hyperparameters, and the optimal hyperparameters used in this chapter are listed in Ref.[55].

First-principles validation and other details. The κ_{DFT} values in Table 4-1 are calculated using a supercell perturbation method and the Botlzmann theory implemented in Phono3py[264]. Unit cell sizes are set to be greater than 10 Å, and the magnitude of atomic perturbation to be 0.005 Å. The force constants are extracted from density functional theory with plane-wave basis set through VASP[265]. We employ the generalized gradient

approximation of Perdew, Burke, and Ernzerhof[266], and uniform k-meshes with kpoint density greater than 700 k-points $Å^{-3}$. The plane wave energy cutoff is set to be 1.3 times the maximal ENMAX of elements in the unit cell. The convergence criteria for energy and ionic forces are set to 10^{-6} eV and 0.01 eV/ Å, respectively. Details of experimental measurements are provided in Ref.[55]. Van-Arkel triangles have been used to characterize the bonding nature of binary compounds, in terms of the average and difference of element electronegativity (χ_a).

4.7. Chapter summary and outlook

In summary, we studied the structural chemistry of lattice thermal conductivity κ for inorganic crystals, and predicted κ for a large set of inorganic compounds, directly from their atomic structures. We extended our graph neural network model to include transfer learning, and using as input our recently prepared database of κ. Combining the neural networks model and interpretable random forest, we extract atomic features that dominate the physics of κ, including elemental (V_{GS}, χ_a, r_a) and packing (L_B, V_a). Other features, such as *CN*, are shown to be also important but more complicated than conventionally assumed. With these identified features, we extended van-Arkel triangles as two-dimensional projected materials space. This analytical tool allows the projection and visualization of materials spaces for κ, and could be applied to other materials informatics studies. We also identified rare-earth chalcogenides (REX), which exhibit a zT exceeding 1.0 and are a new promising material system for thermoelectrics. A limitation of the current models is not to fully predict the six tensor components of κ (our current values are polar average of these tensor components), thus the possible anisotropy. However, this will be technically possible with increasing database for anisotropic κ and development of machine learning models to predict tensor properties[78].

As for transfer learning, we show the mechanism of where and how transfer learning

improve the learning performance of the small dataset with help of the large data. Although transfer learning has been increasingly used to predict materials properties with small datasets[49, 61, 64], how to choose the proper larger dataset is still empirical and based on human intuition. It is still not clear fundamentally why and how transfer learning improve learning performance. Practically, it is important to know before the training process whether transfer learning between a pair of datasets would help the learning of the smaller dataset. For this purpose, our analysis in Figure 4-3 might be good start point, although more quantitative metrics are still necessary, such as the *Log Expected Empirical Prediction*[267] and *Log Maximum Evidence*[268].For this purpose, our analysis in Figure 4-3 might be good start point, although more quantitative metrics are still necessary, such as the *Log Expected Empirical Prediction*[267] and *Log Maximum Evidence*[268].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

5. Optimizing laser-processing by Bayesian Optimization

5.1. Introduction

When the dataset is too small or too biased, algorithm design alone might not be sufficient to significantly improve the prediction performance of machine learning models, and expansion of dataset is necessary. For efficient and effective expansion, one should add data points that are not in the same domain as the existing data points, otherwise the added data points cannot provide more information about the whole materials space. For optimizing certain materials properties, one should add data points that are likely to result in the desired properties. Active learning is an approach to add the most uncertain data points to the current machine learning models to the training set, because these uncertain data points are highly unlikely to come from the same domain covered by the current training set. Based on active learning, Bayesian Optimization also add data points that are likely to have the optimal properties to the dataset. In this chapter, we show an example of using Bayesian Optimization to search for the optimal laser-processing parameters, as well as how machine learning enables design and processing of functional materials.

Laser-reduction of polymers has recently been explored to rapidly and inexpensively synthesize high-quality graphitic and carbonaceous materials from commercial polymers. Such easily synthesizable carbonaceous structures hold promise in being utilized for a broad range of electrochemical applications, including in energy storage. However, in previous works, laser-induced graphene has been restricted to semi-aromatic polymers and graphene oxide – in particular, poly(acrylonitrile) (PAN) is claimed to be a polymer that cannot be laser-reduced successfully to form electrochemically-active material. In this chapter, three strategies to

surmount this barrier are employed: (1) thermal stabilization of PAN (resulting in thermally stabilized PAN (TS-PAN)) to increase its sp^2 content for improved laser processability, (2) pre-laser treatment microstructuring to reduce the effects of thermal stresses, and (3) Bayesian Optimization to search the parameter space of laser processing to improve performance and discover new morphologies. Based on these approaches, we demonstrate the lowest reported sheet resistance (6.5 Ω/sq) derived from laser-reduction of any polymer, in addition to demonstrating successful laser reduction of PAN for the first time. The resulting materials are tested electrochemically for activity, and their application as membrane electrodes for vanadium redox-flow batteries is demonstrated. Electrode performances are lower than conventional electrodes, but our approach realizes membrane electrodes that are processed in air, below 300 °C, which are cycled stably over 2 weeks at 40 mA cm^{-2}, motivating further development of laser-reduction of porous polymers for membrane electrode applications such as RFBs.

Graphitic and carbonaceous materials are promising for a variety of energy applications, including electrochemical applications[269, 270], and water treatment[271],. However, these avenues have usually required either extensively processed graphite or mixtures of carbon, or energy-intensive processing such as chemical vapor deposition. Moreover, there is a vast body of literature on the optimization of materials properties for specific applications. However, given the vast number of parameters and approaches involved with optimizing carbonaceous materials, it is evident that such optimization frameworks are challenging to establish and test. In this chapter, we focus on the specific case of porous carbonaceous electrodes, which are typically derived from poly(acrylonitrile) (PAN) that has been manufactured and carbonized / graphitized into micrometric fibers arranged into a free-standing fibrous mat[272]; however, alternative materials sets derived from Rayon and biomass have also been demonstrated[273-275].

To achieve a functional porous carbon electrode, careful design of its microstructure, in addition to carbonization methods involving thermal processing, are required. Recent work demonstrated that the bottom-up synthesis of PAN-derived, non-fibrous electrode microstructures using non-solvent induced phase separation (NIPS) provide multimodal porous structure consisting of interconnected large pores with interspersed small pores, which can outperform conventional fibrous electrodes[276]. In the case of PAN, the polymer scaffold undergoes a multi-step heat-treatment, whereby the material is thermally treated in air at lower temperatures (typically a maximum of 300 °C) and then pyrolyzed at higher temperatures to increase graphitization content. The thermal stabilization is essential for crosslinking the PAN and improving the ensuing mechanical properties of the electrode after carbonization[272, 277]. The overall process is also crucial to preserve the structure and hence electrochemical performance of the carbon electrode[278]. However, conventional means to pyrolyze PAN-based materials through thermal processing is cumbersome, costly, and time-consuming – especially in the case of thermal processing which becomes increasingly challenging above 1000°C, which is the temperature range needed for effective carbonization, but also a temperature which approaches the melting points of many common metals, which then requires specialized, custom-built graphite ovens. These hardware limitations motivate versatile, low temperature, and high-throughput manufacturing routes.

Over the past several years, laser-annealing has been shown to be a promising means to rapidly generate high-quality carbonaceous and graphitic material from polymer precursors with lower energy requirements and higher throughput than standard thermal processing[279, 280]. Laser-annealing leverages the strong optical absorption of polymers in specific wavelength ranges (for example, at 10.6 μm), which causes them to experience temperatures of over 2500°C under very short timescales[280]. With conventional means, accessing temperatures above 2000°C poses both significant challenges and opportunities in materials

processing. Previous works have aimed to identify criteria for “laseability” of materials[281]; unfortunately, PAN has repeatedly been identified as a polymer that cannot be laser-annealed effectively[279, 280]. Recent work has shown that even reportedly laseable polymers may need pre-treatments to surmount phase transitions that might cause melting or ablation[282]. In this chapter, we demonstrate that for PAN, the aforementioned thermal stabilization step is essential to prevent melting or ablation from the rapid material changes induced by laser processing.

However, with a completely new materials system which was previously thought to be unlaseable, the process of optimizing numerous variables that govern the laser-annealing process is daunting. It is time-consuming to explore the full search space to discover new morphologies, or even better optima, which are not biased by the initial search criteria. Thus, we leverage Bayesian Optimization (BO): a promising tool to optimize the laser-annealing conditions for PAN. Optimization of chemical reactions is usually challenging because there are often too many possible conditions for an experiment, which makes it impossible to exhaustively measure outcomes of all possible conditions.

BO is an uncertainty-guided optimization method for complex black-box objective function[283]. It consists of two basic steps: exploitation and exploration. In exploitation, BO suggests that the candidate point that has the optimal value predicted by BO should be priorly tested, while in exploration it suggests that the candidate point that are uncertain to the BO should be priorly tested. The balance between exploitation and exploration is controlled by the choice of acquisition function. Turner *et al.*[283] have shown that BO is superior to random search for black-box optimization problems such as machine learning hyperparameter tuning, and Shields *et al.* has even shown that BO outperforms human decision-making for optimization of some chemical reactions[18]. Consequently, BO has been increasingly used to tune the experimental conditions in the field of chemistry and materials science[18, 41, 284-

287]. Given the massive amount of possible laser conditions, BO seems to be useful to find appropriate laser conditions with reasonable times of trials, and Wahab *et al.*[40] have employed BO to guide the lasering of graphene.

However, lasing PAN might be a different or even a more challenging problem for BO than most previous studies. In previous studies, BO seldom yields a counter-intuitive result (with the exception Dave *et al.*[41]). Here, the task of lasing PAN to make it conductive is itself counter-intuitive, as previous works note the unlaseability of PAN[279, 280]. In this sense, BO is employed to discover new paradigms than to accelerate a well-known physical and chemical process. More specifically, to tune laser parameters for PAN by BO, human intervention might be necessary in the operation of BO. Recently, BO has been increasingly combined in the autonomous platform of experiments where BO is conducted in a human-defined search space without human-intervention during the optimization[18, 41, 285]. The search of lasering conditions for graphene is also conducted in this autonomous way[40]. Pre-defining the search space is important for such autonomous optimization, because there is a trade-off between efficiency and effectiveness: if the search space is too large, then the steps to reach the optimum might be larger; otherwise, the optimum might not be included. Unlike polymers such as polyimide and poly(ether sulfone) which are known to be laseable[40, 279, 280], there is limited information about lasing PAN, so it is possible that the pre-defined search space might be too large or too small and adjustment of search space might be necessary during the operation of BO – where the initial search parameters in this study are chosen on the basis of an optimum derived from numerous experiments on a completely different material system (graphene oxide)[288]. Second, different PAN lasing conditions are shown to result in different carbon properties (as a spectrum between graphitic and carbonized), and consequently the impact of different lasing conditions on the morphology and property of PAN might be highly nonlinear and non-smooth, which is challenging for the standard BO based on Gaussian

Process (GP). Third, in most cases, BO focuses on a single aspect of the materials system, while in our case, during the optimization, other aspects of materials might also change and impact the merits of the materials. Fourth, unlike lasing conventional monolithic polymers, where resulting structures are physically more predictable and can thus be investigated easily through Raman spectroscopy – we use the linear resistance (R) to rapidly probe the overall progression of the lasing process to approach a conductivity regime where electrochemical properties can be further investigated.

Since the initial materials system and resulting morphology/properties are not as well known, we are not able to use automation approaches as described in previous works. Our approach thus employs BO driven to optimize R – an easily measured parameter – to arrive at suitable candidates for electrochemical performance. We arrive at two parameter configurations with low R which yield either highly graphitic membranes, or highly disordered/carbonized membranes. We show that an intermediate between these two properties yields the best results in terms of electrochemical performance and stability. Thus, the BO discovers two vastly different parameter sets to yield different chemical and physical properties, of which one was found to represent a good set of properties for electrochemical applications.

In summary, this chapter employs an array of strategies, namely: (a) microstructuring of PAN, (b) thermal stabilization of PAN, and (c) a BO-driven optimization of laser annealing parameters, to demonstrate successful laser-annealing of PAN for the first time. Our study paves the way for future work to use BO as a means of accessing new material morphologies, rather than simply accelerating, or optimizing known processes, and demonstrates the ability for laser-annealing to create PAN-derived membrane electrodes without an energy-intensive high temperature carbonization step, which yields the lowest sheet resistance demonstration (6.5 Ω/sq) for a laser-reduced polymer.

5.2.*A priori* knowledge and insights needed to start Bayesian Optimization

To start approaching the problem of laser-processing PAN, we begin by verifying the current knowledge. After testing lasing parameters on as-prepared PAN membranes (as described in Chapter 5.6), we verify that the polymer is either unaffected when below a certain power threshold, or completely melts or burns above the threshold, which does not result in a conductive structure as previously described. Therefore, we identify a need to further process PAN from its native state.

PAN is often thermally stabilized to transition it into a morphology that can be effectively carbonized and graphitized, and this property has been well-explored especially through carbon-fiber synthesis (Figure 5-1a). However, a purely sp^2-character precursor is not ideal for the process to occur – in fact, a balance of sp^2 and sp^3 nature in a material has been shown to promote graphitization. Therefore, the stabilization process in air is shown to promote this morphology[281].

This process causes rapid thermal expansion and contraction associated with the fast timescales of the lasing process, and the high temperatures experienced due to the absorption process and subsequent chemical changes (such as graphitization) of the polymer, which can lead to ablation and mechanical stresses that prevent bulk structures from being realized through the lasing process. We thus identify micro-structuring to mitigate complete ablation of the polymer, and this is shown to yield high-quality graphitized material with minimal ablation.

As an initial input into the system, we explore a sample space of previously probed parameters (such as laser scan speed, laser power, focal point height (Z) and image density (ID)), which are each described in the Methods section. Based on intuition from graphene-oxide, we choose constraints on each of the variables to limit the initial exploration. The direct

physical effects of these parameters can be condensed into a physically meaningful expression known as dynamic fluence (in units of J/mm^2), which is expressed as follows[289, 290], assuming $Z/Z_0 >> 1$ (where $Z_0 = 0.02$ inches) (equation 5-1):

$$Dynamic\ fluence = \frac{Laser\ Power}{Scan\ speed \times Beam\ diameter} \propto \frac{Laser\ power}{Scan\ speed \times Z} \text{.....(5-1)}$$

However, as evidenced by previous work[289, 290], such parameters, with the rapidly changing properties of the host polymer itself as it chemically reduces, ablates, or changes in volume, quickly increases the complexity of the problem. This lends itself to modelling the system as a simple set of input machine parameters and a rapidly testable output.

With this information in mind, we designed our procedure. We start with making small, square test-areas using a set of 20 input parameters, and we probe the linear resistance of the samples across a small square to roughly obtain the bulk resistance of the lased sample (where unlased, TS-PAN is non-conductive). The resulting R values are input into the BO algorithm (Fig 5-1b) to suggest parameters to further explore the search space, or to exploit a specific set of successful parameters. The process is repeated for several cycles until optimized parameters are found (Fig 5-1c). Then, the optimized parameters are used on both sides of the PAN membrane to allow for a fully-conductive slab (Fig 5-1d), which is then used for electrochemical testing (Fig 5-1e).

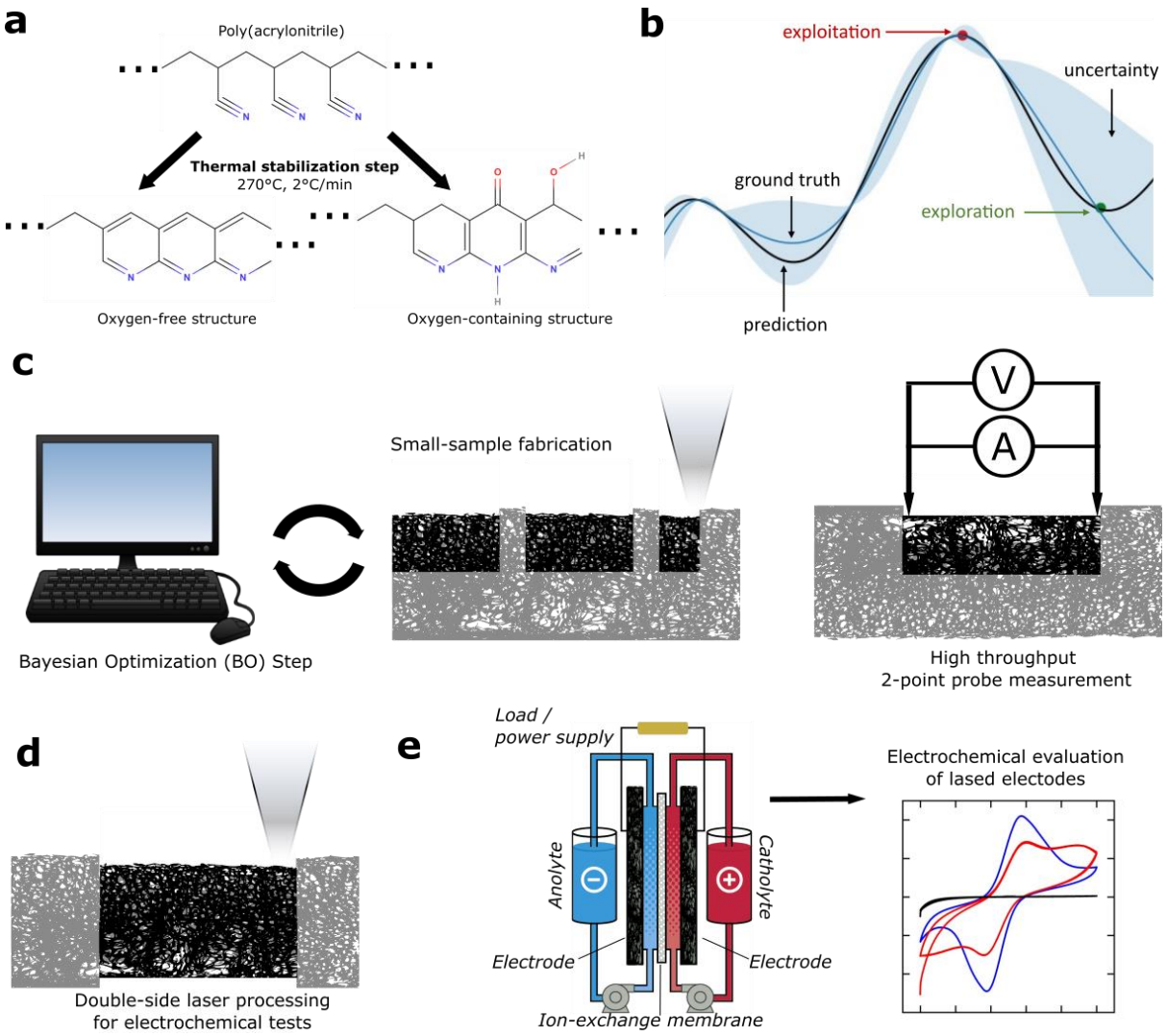

Figure 5-1. a Chemical structure of PAN and possible structures in TS-PAN, showing increased conjugation in thermally-stabilized polymer. **b** Illustration of the Bayesian Optimization process, showing the exploration and exploitation process in the Bayesian optimization algorithm, around the ground truth of ideal parameters. **c** Experimental cycle for initial optimization. Samples are fabricated with input parameters, then tested for linear resistance across a 1cm square, followed by feeding results into the Bayesian optimization algorithm, and this is repeated for 8 cycles. **d** For electrochemical tests, the optimized parameters are used to lase both sides of the TS-PAN. **e** General schematic of the redox-flow battery test (left) and cyclic voltammetry test (right).

5.3. Searching for sheet resistance optimum

In this chapter, we use BO to optimize the lasering conditions for PAN. Figure 5-2a shows the evolution of the lowest R during the BO, and we can see that the lowest R is achieved at iteration #5 with R = 10 Ω, which is 40% of that with the initial human-designed parameters at

iteration #-1. Note that such a significant improvement is achieved with 204 tests out of 1 million possible combinations of parameters, showing the efficiency of the optimization. Figure 5-2b shows the evolution of the parameters that result in the lowest R in each iteration. The evolutions of the best parameters show the combination of exploration (large jumps) and exploitation (small changes) in the optimization process. The full set of 204 points explored in this study is displayed in terms of power and speed in Figure 5-2c, from which one can see the exploration direction, from low power, medium z region to medium power, highest and lowest Z region. Moreover, Figure 5-2c shows that, after iteration #4, BO extensively exploits the region with highest Z (the region with lowest R).

A few phenomena and morphologies occurred during the optimization. The first phenomenon that emerged during the experiment was called the "burnout" regime. This was observed at high powers and low speeds, which were in the direction that the BO algorithm initially optimized for, where a high degree of ablation was observed, which removed a large fraction of the material, but in some cases, left a conductive sheet of material which was brittle. With even higher powers, this material was discontinuous, and therefore R could not be measured. A "burnout" sample was therefore designated as an undesirable result, since it would not be useful for most electrochemical applications where both high conductivity and electrochemical activity are necessary, rather than just conductivity at the expense of electrode active surface area. Starting from iteration #3, we begin to observe burnout. Since we aim to avoid burnout, we set the R of the burnout cases to be 10^6 Ω to force the BO optimize away from the burnout region. However, R of the burnout cases are quite low strictly from a resistance measurement perspective. For example, for the parameter set of (Z = 0.14, ID = 5, Power = 35, Speed = 10), although it results in burnout, the measured R is 9.3 Ω, which is even lower than the lowest R reported above.

Because of the assignment of large R to the conductive burnout cases as mentioned above, and because of the non-linear relation between each individual parameter and R as shown below, the underlying function between the four parameters and R is highly non-linear. Therefore, the GP used in standard BO might not capture the non-smooth function very well. Designing GP for non-smooth functions is still in development[291-293]. Here, after iteration #3, we use neural networks (NN) with high expressive power for non-linear functions[21] to help the exploitations in the high Z regions. After each iteration, we use the updated dataset to train a GP by the standard BO package, and also train a NN, and we replace some of suggested parameters from GP by that from the NN to avoid points which are intuitively likely to result in burnout (i.e., replacing high power, low speed points suggested by GP with high Z points suggested by NN). As a result, the parameter that results in the lowest R is discovered by the NN. More discussions about the role of NN in this chapter are provided in Chaper 5.4.

At iteration #3, we expand the search space of Z from [0.03", 0.10"] to [0.02", 0.14"]. The motivation for the expansion is that, at iterations from #0 to #2, the lowest Rs are all observed at the boundary of Z = 0.100 inches. As a result, we observe lower Rs after the expansion, and the lowest R is observed at Z = 0.138 inches. As previously noted, the dynamic fluence of a particular set of parameters is proportional to the Power:Z ratio, which can remain fixed even as Z is expanded. Therefore, to maintain energy efficiency of the process, we limit further expansion of the Z limit. Despite the current trend of autonomous optimization of experiments, here we suggest that human monitoring and modification of the optimization of experiments might still be necessary. This is evident due to the following interventions: (1) assignment of large R to burnout cases, (2) the use of NN for exploitation, and (3) the expansion of search space during the optimization. However, the benefit of the mixed BO and NN process is the discovery of two parameter regimes which unlock distinct morphologies as observable by Raman spectroscopy. The evolution of the parameter set resulting in the lowest R in each batch

of 20 samples is shown in Figure 5-2d.

To further understand the role of each parameter in determining R, we plot the impact of each parameter on 1/ R (SHAP value[96]) for all the 204 data points in Figure 5-2e. Based on empirical observation informed by the physical insight of the dynamic fluence, we see that higher speed leads to higher R, and higher values of Power and ID have either a strongly positive or strongly negative effect on the final R. From a physical perspective, high power and ID samples are more likely to approach a regime of full graphitization / carbonization which is desirable from the conductivity perspective. However, they are also more likely to result in ablation with the incorrect corresponding parameters. The speed result corresponds specifically to the fact that slower speeds ensure a more even morphology across the sample, which improves the electrical connectivity between regions.

Since the 2-point linear resistance of samples is correlated, but not necessarily equivalent, to the sheet resistance metric, select samples were tested with a van der Pauw (vdP) method to obtain the electrical properties of the network without series and contact resistance corresponding to the poor interface between the probes and the porous network (procedure described in the Methods section). The lowest demonstrated sheet resistance obtained through this method was 6.5 Ω/sq, which is the lowest reported sheet resistance to date for laser-reduced polymers.

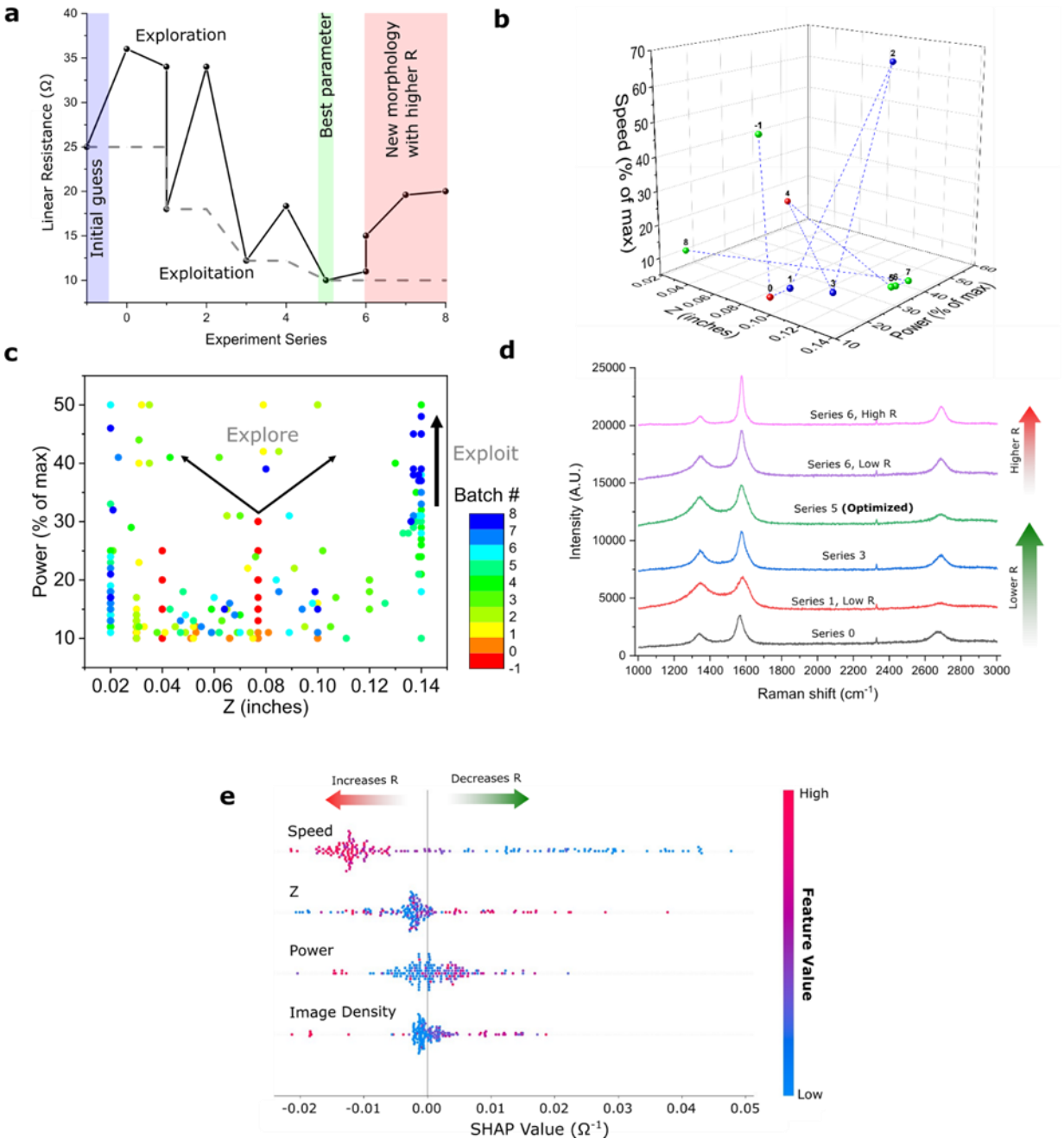

Figure 5-2. a Illustration of the full Bayesian Optimization process to find the lowest resistance – where illustrated points show the best resistance measured over a set of 20 samples. **b** Illustration of the full 3D space of exploration, where all image densities are explored in each point. **c** Full illustration of explored power vs. Z, showing the exploration of a wide space of parameters to find the overall minima across the imposed boundary conditions. **d** Representative Raman spectra of select points in the optimization process, showing a progression towards an intermediate between highly-graphitic / carbonized electrode. **e** Effect of parameter values on the overall R outcome, where positive SHAP value represents a parameter expecting to reduce R, while a negative SHAP value represents an expected increase in R.

5.4. Neural networks for exploitation

Because of the assignment of large R to the burnout cases, and because of the non-linear relation between each individual parameter and R, the underlying function between the four parameters and R is highly non-linear. Therefore, the Gaussian Process (GP) used in standard BO might not capture the non-smooth function very well. We show the fitting results of the GP implemented in EDBO at iteration #7 in Figure 5-3, from which one can see that the GP cannot fit the dataset very well. Especially, the GP shows a very high false positive rate, or in other words, there are many points predicted to have low R which in turn do not have low R measured in experiments. Such high false positive rate might lower the efficiency of the exploitation step, as many of the predicted low R parameters would result in high R in measurements. In Figure 5-3c, we compare other surrogate models available in EDBO for fitting the dataset at iteration #7, including GP with different length-scale priors and nu parameters for the Matern kernel, and Bayesian Linear Model and Random Forest Model. We can see that switching from the automatic setting by EDBO to other models does not improve the fitting performance significantly.

In this chapter, we use neural networks (NN) to help the exploitation step. We use 3-layer networks with 16 neurons in each layer, and use the "relu" function as the non-linear activation. 10 networks with different random initialization are used as an ensemble, and the predicted values of Rs are the mean of predictions from the ensemble. We choose the "relu" function as the activation function, because the dataset has step-function behavior, which partly results from the fact that we assign $R = 10^6\ \Omega$ to all non-conductive samples and burnout cases. As shown in Figure 5-3b, the NNs have R^2 scores close to 1.0 at all iterations. As a comparison, with the same dataset and other settings of NN, if we switch from "relu" activation to linear activation, the R^2 scores would drop to around 0.4, and if we switch to logistic activation, the R^2 scores would drop to less than 0.9.

In Figure 5-3b, we show the fitting performance of GP and NN during the optimization. For GP, with the addition of new data, the fitting performance first drops quickly and then slowly improves, while for NN, the fitting performance is excellent over the whole process. We argue that, the combination of GP and NN in this chapter is critical to the optimization, because of the following reasons:

i) Both GP and NN contribute to the discovery of the parameters that result in the lowest R at each iteration. Specifically, NN suggests such parameters with lowest R at iterations #3, #5, #6, and #7, and GP suggests such parameters at iterations #0, #1, #2, #4, and #8. NN discovers the parameters with the lowest R during the whole optimization at iteration #5.

ii) Because of its strong fitting power, NN is very helpful for exploiting the high Z region, as shown in the argument i). However, we cannot rely solely on NN, because even if we use the standard deviation of NN ensembles as an estimation of uncertainty, all the parameters NN suggests to test are concentrated in the high Z region. Therefore, in order to continue the exploration for whole parameter space, GP is still necessary even after the introduction of NN.

iii) Although GP cannot fit the dataset well after several iterations, at the initial iterations, GP alone successfully lowers R from 25 to 18 Ω, and it suggests the exploration directions from medium Z to highest and lowest Z, which pushes us to expand the search space and identify the regions with low R.

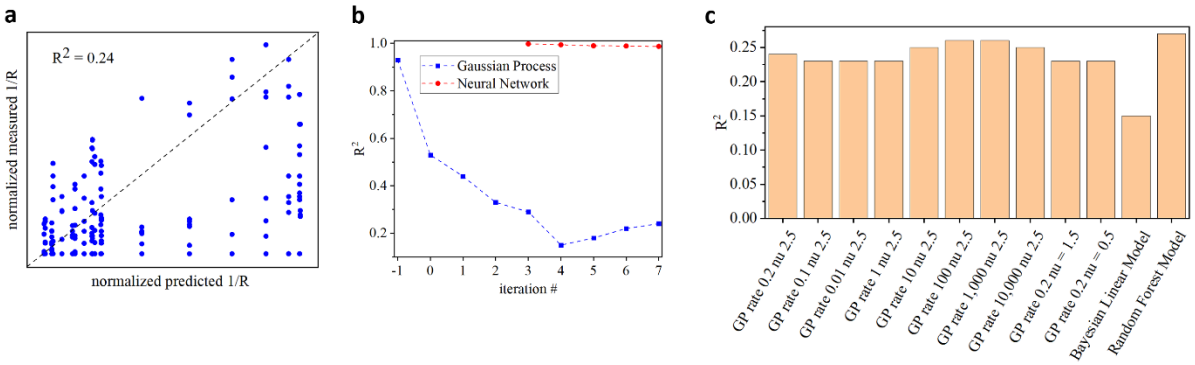

Figure 5-3. a. Normalized measured 1/R versus predicted 1/R from GP in EDBO at #7. **b.** Evolution of R^2 scores of predictions of 1/R from GP in EDBO and NN in this chapter. **c.** R^2 scores of different surrogate models in EDBO for fitting 1/R at #7.

5.5. Properties of laser-processed poly(acrylonitrile)

While the material sheet resistance and electronic conductivity are indicators of electron transfer capabilities[294], we sought to confirm the electrochemical performance of lased electrodes optimized for sheet resistance. To this end, cyclic voltammetry (CV) was performed in aqueous iron chloride solution, as it has moderately fast kinetics and is a redox couple that holds promise as a low-cost, abundant electroactive material. In these experiments, the working electrode was one of the lased electrodes optimized for lowest sheet resistance (Low ID or High ID). The electrolyte composition was 50 mM $Fe^{2+/3+}$ in 1 M KCl at 50%. Figure 5-4a shows representative voltammograms for two optimized parameters, Param A and Param B, at a 5 mV s^{-1} scan rate. Encouragingly, both samples show electrochemical activity as evinced by well-defined peak currents in the CVs. While the location of the prominent Fe^{2+} oxidation and Fe^{3+} reduction peaks (ca. 0.59 V and 0.37 V vs Ag/AgCl in 3 M NaCl, respectively) are similar for both samples, Param A exhibits sharper and more distinct currents at redox peaks than Param B, indicating higher electrochemical activity. We note that the results of the CVs are to be taken semi-quantitatively due to convoluting factors that complicate interpretation of definite electron transfer rates for non-planar and porous substrates during potentiodynamic measurements[295-297].

To rationalize the differences in electrochemical activity, we performed *ex situ* Raman spectroscopy on Low ID and High ID samples. Comparison of the Raman signatures of the top and bottom of the lased electrodes in Figure 5-4b reveals that the electrodes exhibited different degrees of graphitized and carbonized content; in particular, Low ID had an intermediate of graphitic and amorphous content relative to High ID, which was highly carbonized at both edges. The combination of graphitic and amorphous physicochemical property in carbon-based materials has been shown in previous works to improve electrochemical activity[298, 299].[42,43]

Figure 5-4c shows high-resolution XPS scans of the laser-reduced electrodes, showing a marked difference in carbon signatures. In order of increasing binding energy, the peaks are attributable to carbides, C=C, C-C, C-O, and C=O bonds. The main peaks of interest are the C=C and C-C compared with other binding states, which clearly show a high degree of reduction and carbonization for high ID electrodes, while the XPS denotes a lower degree of overall reduction for the low ID electrodes. This observation is further justified by observed dendritic structures at the surface of lased portions which give rise to the low D and pronounced 2D peaks which are observable in Raman, and also evident from the higher current density from the CV plot. Higher ID values tend to cause more ablation, and therefore are less likely to preserve highly graphitic features.

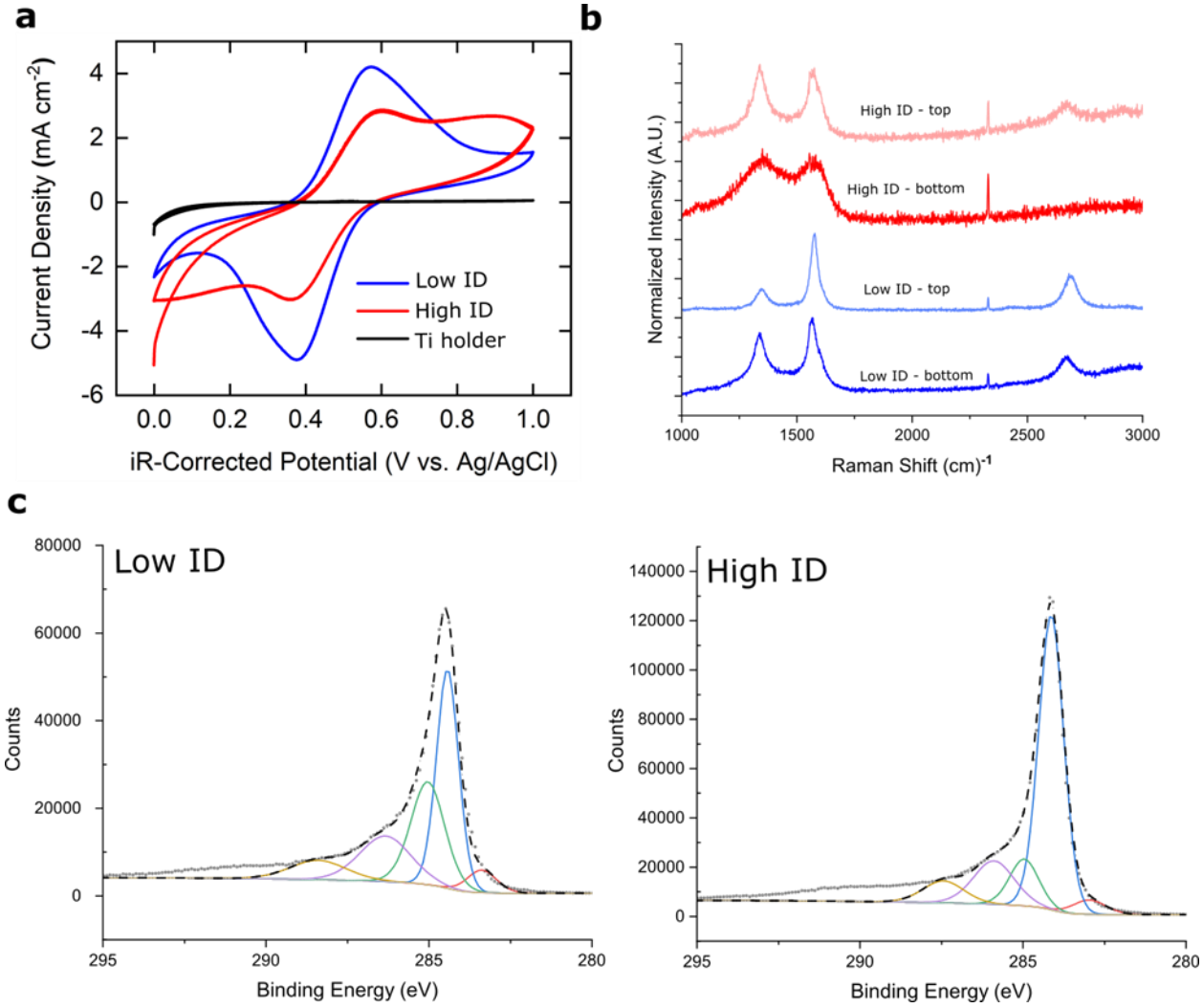

Figure 5-4. Exploration of the physiochemical morphology of optimized electrodes. **a** Current-voltage plot of lased electrodes in 50 mM $Fe^{2+/3+}$ in 1M KCl, showing the higher electrochemical activity corresponding to the lower-image-density electrodes. **b** Raman spectra of both electrodes, showing that lower image densities preserve graphitic features which improve electrochemical activity, but reduce sheet resistance. **c** C1s XPS scans showing the changes in degree of reduction in the electrodes after lasing.

As shown above, the best parameters tend to have an intermediate between low ID and high ID, where a continuum between the optimized parameters explored in Figure 5-4 were explored for Parameters 1, 2, and 3, respectively (exact Parameter values defined in Chapter 5.6). The cross-sectional images in Figure 5-5a highlight the stark morphological differences that result from modulating the different laser parameters. In each parameter, graphitic dendritic structures are visible at each electrode edge, indicating a possible increase in edge sites, which

can increase the electrochemically active surface area, as well as lower the redox overpotential of a certain electrochemical reactions[300]. However, the final performance in redox flow batteries depends on a variety of factors, which will be discussed in the following section. The morphologies of the electrodes as a result of using different Z, Power, and Speed parameters shows that there are differences in morphology and final electrode thickness which result from lasing the porous electrode, further complicating the final insights that could be drawn from the single variable used in BO. Figure 5-5b shows the similarity in overall oxidation states in each of the lased electrodes, and Figure 5-5c shows the progression of the 3 Parameters, from more carbonized to more graphitic. Due to the high degree of rapid chemical reduction experienced by PAN through this process, there is a marked degree of nitrogen content (as shown in the SI), but this effect has been investigated in previous work with other N-containing polymers and can be minimized with further parameter optimization[280].

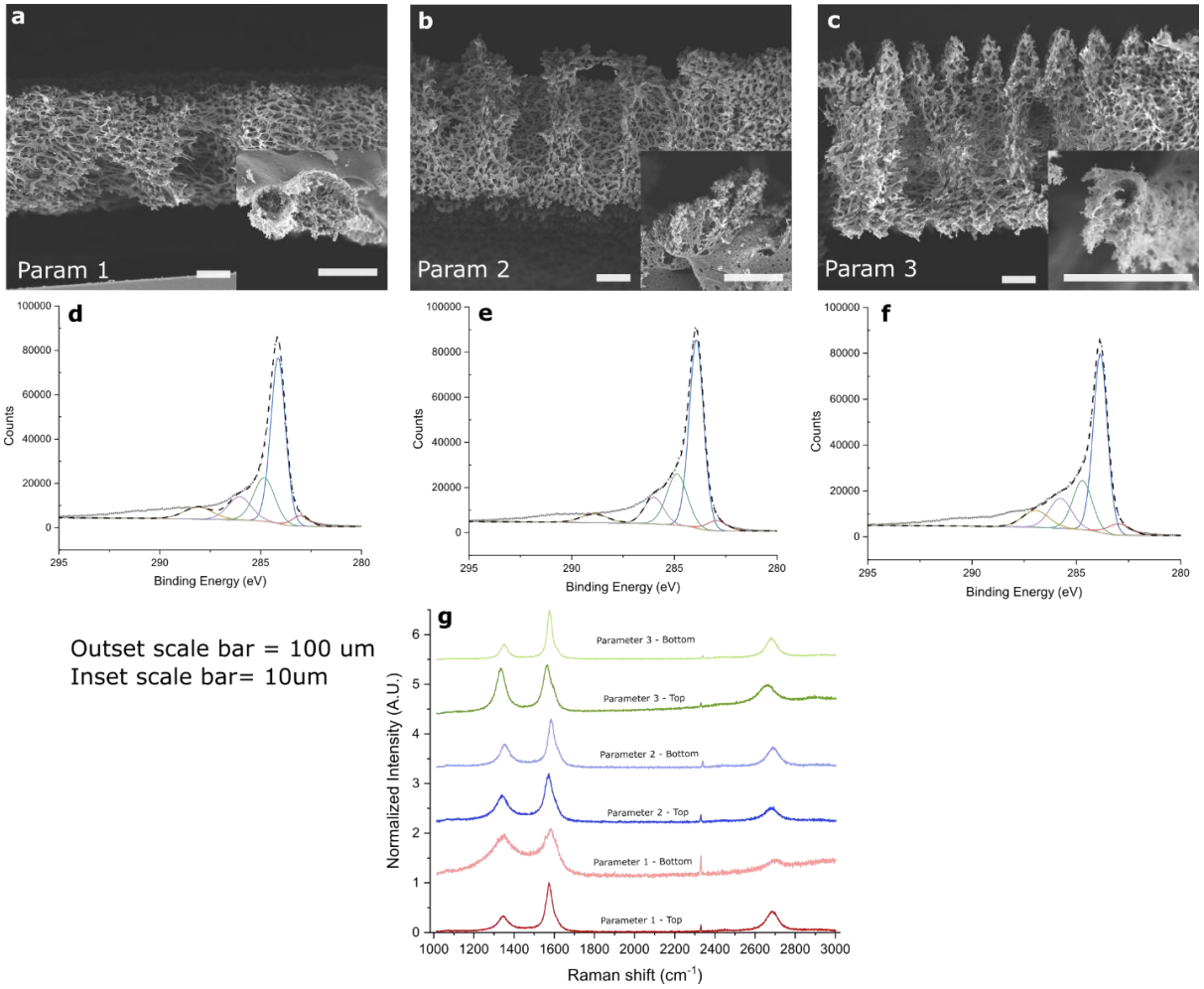

Figure 5-5. Scanning electron images of cross sections of **a** Parameter 1, **b** Parameter 2, and **c** Parameter 3, with high-magnification insets. Scale bars are 100 μm, and 10 μm for insets. Parameter values used for each are listed in the SI. X-ray photoelectron spectra for the Carbon binding energy range, with deconvoluted peaks are shown for **d** Parameter 1, **e** Parameter 2, and **f** Parameter 3. Each spectrum is deconvoluted to resolve contributions from specific carbon chemistries. **g** Raman spectra from the top and bottom of each electrode, showing the range of electrode surface morphologies achievable despite similar R values in the BO optimization step.

5.6. Details of methods

Synthesis of phase separated membranes and electrodes The synthesis of the phase separated electrodes follows the methods described in previous reports[276]. Briefly, polyacrylonitrile (PAN, MW ~ 150,000 g mol^{-1}, Sigma Aldrich), polyvinypyrrolidone (PVP, MW ~ 1,300,000 mol^{-1}, Alfa Aesar), and N,N-dimethylformamide (DMF, for HPLC, ≥99.9%) were mixed together in a glass reservoir. A typical composition consisted of 6.4 g PAN, 9.6 g

PVP, and 80 mL of DMF, leading to 17.5 polymer weight percent, or 0.20 g polymer per mL of solvent[276]. To ensure uniform mixing of the reagents, the mixture was heated and stirred at 70 °C until a homogeneous, viscous, and clear polymer solution was obtained. Three aluminum molds, each machined to contain notches 10 × 5 cm wide and 0.1 cm deep, were arranged onto a glass plate. Polymer mixture was poured into each aluminum mold and dispersed evenly across the notches using a glass slide. The casted polymers were then rested in ambient conditions for ca. 15 min; during this step, humidity from the ambient environment leads to vapor induced phase separation at the non-solvent/solvent interface, preventing an impenetrable non-porous dense layer from forming. After resting for the prescribed time, the glass plate with the aluminum molds is submerged into a coagulation bath consisting of 3 L of deionized water to initiate the phase separation process. After phase separating overnight, the membranes are removed from the aluminum molds, and repeatedly soaked in fresh boiling water until the added water becomes completely clear; this process maximizes the likelihood that PVP and DMF are eliminated from the pores of the PAN membrane.

Following phase separation, the membranes were dried under vacuum overnight at ca. 80 °C to remove residual non-solvent. Then, the electrodes were thermally stabilized in a Barnstead Thermolyne muffle furnace. The temperature was ramped at 2 °C min^{-1} from room temperature (ca. 23 °C) to 270 °C, where it was held for 1 h, and allowed to cool back to room temperature without further intervention. For the furnace carbonized samples, the thermally stabilized materials were inserted into a Carbolite Gero GHA 12/300 and carbonized under flowing nitrogen with the following programming sequence: ramp 5 °C min^{-1} from room temperature to 850 °C, hold for 40 min at 850 °C, ramp from 850 °C to 1050 °C, hold for 40 min at 1050 °C, cool down to room temperature without further intervention.

Laser processing of membranes and physiochemical characterization Electrodes are

laser processed with a VersaLaser VLS2.30 (Universal Laser Systems) with a 10.6 um CO_2 laser. Parameters which are optimized include (1) Laser Power, which is modulated as a percentage of 25W; (2) Laser speed, which is a percentage of 1270 mm/s; (3) Image Density, which indicates the vertical line density in a given scan – where an Image Density of 6 corresponds to 1000 DPI (the horizontal pixel density is fixed at 1000 DPI); (4) Z-height, which indicates the degree of defocus relative to the sample. The samples are approximately 0.02 inches thick, which means this is where the surface of the sample is perfectly in focus, and any higher settings indicate a defocusing of the laser spot. The initial search space for Z was restricted to 0.10 inches, which was expanded to 0.14 inches in the final procedure. A constraint was set, since defocusing necessitates an increase in power at a specific laser point, which reduces the energy efficiency of the process. Once the membrane areas were patterned, the resistance of the samples was measured from the edges of the patterned area with a multimeter. The measurements are taken across opposite edges and the final R value is listed as the average of the two readings. This is also illustrated in the lower power requirements for a Z-height of 0.02 inches, compared to 0.139 inches. All samples were patterned on the "Top" surface, which was designated as the dense layer of the PAN membrane. For double-side patterned membranes, both the "Top" and "Bottom" surfaces were patterned. The list of best explored parameters from each Experiment Series are listed below in Table 5-1:

Table 5-1. List of all parameters explored using BO (also depicted partially in Figure 5-2).

Experiment Series	Z-height (inches)	Image Density	Power *(% of 25W)*	Speed *(% of 1270 mm/s)*	R (Ohms)
-1	0.077	5	20	25	25
0	0.051	6	10	10	36
1	0.100	6	50	65	34
1	0.100	6	16	10	18
2	0.100	6	50	65	34
3	0.120	6	20	10	12

4	0.043	4	41	20	18
5	0.138	5	30	10	10
6	0.139	5	31	10	11
6	0.140	7	18	10	15
7	0.137	7	31	30	20
8	0.020	5	17	10	20

A set of 20 measurements is performed, and the sample with the lowest resistance was further characterized with Raman spectroscopy (Renishaw Invia Reflex Raman Confocal Microscope, 50 mW, 532nm laser, 10X objective lens). At later series, multiple samples were analyzed due to similar R measurements resulting from very different parameter sets. The final membranes were additionally characterized using Scanning Electron Microscopy (Zeiss Gemini 450) and high-resolution X-ray Photoelectron Spectroscopy (Thermo Fischer Nexsa) with a flood gun for charge correction, and Shirley background correction on obtained spectra. For electrochemical evaluation, the following parameters were used to prepare membranes for testing (Table 5-2):

Table 5-2. List of Parameters tested electrochemically (as represented in Figures 5-3 and 5-4).

Label	Lased Side	Z-height *(inches)*	Image Density	Power *(% of 30W)*	Speed *(% of 1270 mm/s)*
Low ID	Top	0.139	5	31	10
	Bottom	0.02	4	15	10
High ID	Top	0.14	7	10	15
	Bottom	0.14	7	10	15
Parameter 1	Top	0.139	5	29	10
	Bottom	0.139	5	29	10
Parameter 2	Top	0.02	5	17	10
	Bottom	0.02	5	17	10

Parameter 3	Top	0.077	5	20	25
	Bottom	0.077	5	20	25

To compare the linear resistance measurement to previously reported values, we perform a sheet resistance measurement through the vdP method[301]. 1 cm^2 laser reduced samples were prepared with the following parameters ([Z, ID, Power, Speed]) –Parameter A: [0.077, 5, 20, 25]); Parameter B: [0.138, 5, 27, 10]. Then, the sample was cut to shape and secured onto a 1" x 3" glass slide with double-sided tape. Then, 4 copper tape strips were pasted to the 4 corners (1-2mm from the edge of the lased area) and the contact was reinforced with silver paste (DuPont 4922N-100). This resulted in the following [R_{lin}, R_{sh}] combinations: Parameter A: [35 Ω, 16.3 Ω/sq]; Parameter B: [15 Ω, 6.5 Ω/sq]. Thus, the linear resistance values are generally overestimates of the sheet resistance values in this study, but still serve as a suitable measurement to optimize sheet resistance.

Bayesian Optimization In this chapter, because of the large range of R (10 Ω to 10^6 Ω), we maximize 1/R as the optimization target. The package EDBO[18] is used to perform the standard Bayesian Optimization. Gaussian Process is used as the surrogate model in EDBO with the Matern Kernel as the covariance function. Expected Improvement is used as the acquisition function. The BO_express module of EDBO is used to conduct the Bayesian Optimization, as this module automatically featurizes the reaction space, preprocesses the data and selects the priors for Gaussian Process. For the neural networks used to help exploitation, we use Scikit-Learn[19] to construct 3-layer networks with 16 neurons in each layer, and use the "relu" function as the non-linear activation. 10 networks with different random initialization are used as an ensemble, and the predicted values of R are the mean of predictions from the ensemble. For the SHAP values, we first build a decision tree model to fit the dataset, then use the Package SHAP[96] to derive the impact of parameters to model output.

The search space for BO is defined as follows: Z ∈ [0.03, 0.10] inches, with the spacing of 0.01 inch; ID ∈ (4,5,6,7); power ∈ [10, 50] % of 30 W, with the spacing of 1 % of 30 W; speed ∈ [10, 60] % of 1270 mm/s, with spacing of 5% of 1270 mm/s. Therefore, the number of combinations of parameters is 593,844; After expansion, the Z space is expanded to [0.02, 0.14], and the number of parameters increases to 1,012,044.

5.7. Chapter summary and outlook

This chapter demonstrates several avenues of fundamental progress in processing PAN for electrochemical applications. We demonstrate that using the strategies of microstructuring, thermal stabilization, and BO, are crucial to arriving at a suitable set of parameters to realize porous carbon electrodes processed at <300°C. Our approach starts with a careful investigation of the initial conditions and previous knowledge of laser-reduction of polymers and graphene oxide, followed by a BO which explores the parameter space in ways that yield unexpected morphologies at parameter combinations that would have otherwise been thought to lead to undesirable results. However, we also show that in our specific study which only uses linear resistance as a rapidly-testable measurement to inform BO, we need additional intervention through expanding testing boundaries and introducing a NN to exploit successful parameters. The resulting parameters yield the lowest sheet resistance value reported to date (6.5 Ω/sq) for any laser-reduced polymer, using PAN: a polymer that was previously reported to be unlaseable in its native form. Overall, this chapter motivates future studies on BO used for exploration of parameter spaces to discover new morphologies, as well as continued optimization of porous laser-reducible polymer scaffolds for electrochemical applications.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

6. Predicting charge density distributions by graph convolutional network

6.1. Introduction

In Chapter 6, we present a method to predict charge density distributions of materials by graph neural network, which can be further employed to accelerate density functional theory calculations (DFT). As stated in the Hohenberg-Kohn theorem[302], ground state energy is determined by the electron charge density distributions of materials. In modern DFT calculations, charge density distributions are first obtained by solving the Kohn-Sham equation[158] self-consistently, then other properties are calculated based on the charge density. However, the relatively high computational cost and high memory demands of DFT[86] limits its use for large systems with more than several hundred atoms. Rapid and direct prediction of charge density is critical to the acceleration of DFT calculations. Meanwhile, charge densities are increasingly used as input features of machine learning (ML) models to predict other materials properties[303-305]. Therefore, it is important to develop methods capable of accurately predicting charge density with less computational demand, to "by-pass the Kohn-Sham equations"[84], and machine learning (ML) is a promising tool for this goal.

In principle, an ideal ML algorithm should meet three requirements: high accuracy, high transferability and low computational cost. Very recently, there have been attempts [84, 87] to employ ML to predict the charge density of molecules by expanding the density as a sum of atom-basis functions. For the case of periodic systems, Schmidt *et al.*[306] employed basis functions, summing over the contributions from only neighboring atoms to achieve transferability between different cell sizes and lower memory demands, while Chandrasekaran *et al.*[86] encoded the position of each grid-point to neighboring atoms by well-designed

invariants to predict charge density. However, compositional and structural transferability remains a challenge, as these methods account for variations in one structure at a time (i.e., strained lattices or different molecular dynamics snapshots).

In this chapter, we develop a ML-based approach that can predict charge density for different structures with varying compositions, structural features and defects for a given class of materials in a single training, which is necessary for application on systems such as amorphous hydrocarbons or glasses where local structures are highly complex. In previous works, a three-step process was followed: 1) record the distance between each grid point and all neighboring atoms, 2) add all distances together to form a feature vector, and 3) compute charge density by regression on the final feature vector. For multi-elemental systems, the first two steps are repeated for each element type and the feature vectors are concatenated together. In order to build upon this approach with increasing transferability between different structures, in addition to recording the distance between grid-points and atoms, we propose to both explicitly encode the geometry of the cluster formed by neighboring atoms, and account for all elements simultaneously as opposed separately. Encoding the geometry, on the one hand, avoids the problem of different local environments leading to a similar sum of atom contributions, on the other hand, enables the model to learn from the geometry of existing structural features and speculate new ones. Greater structural transferability should also lead to improved accuracy in the prediction of charge density for defect structures, as new structural features can form during the formation of defects. To accommodate different elements, the dimension of the final feature vector should be independent of composition, otherwise the regression process (matrix-vector multiplication) cannot be done.

A graph representation, encoding both nodes and bonds, has a number of advantages that meet the requirements listed above. Graph representations have been used recently to encode information on both the level of atom and geometry with high accuracy and transferability

across composition, structure and property space[57, 69, 71, 76, 77, 105, 114], and the feature vectors can be of the same dimension for different compositions if properly designed. In this chapter, we encode environments of grid-points as graphs and employ the crystal graph convolution neural network (CGCNN)[71] to find a relationship between local environment and charge density. We train and test our scheme on two classes of crystalline materials, polymers and zeolites. For each case training data is from some structures and the model is applied to others to test transferability, and the accuracy of the predicted charge density is evaluated through statistics and visualization.

6.2. Model architecture

As shown in Figure 6-1, we encode three-dimensional space in the unit cell using CGCNN by placing an imaginary atom at each grid-point in the unit cell. The local environment is computed for a given grid-point by identifying atoms within a cut-off radius (R_{cut}) from the imaginary atom, as shown in Figure 6-1b. Next as shown in Figure 6-1c, atoms outside R_{cut} are removed, and the remaining structure is placed in a larger cell to avoid interactions between periodic images. Here R_{cut} is 4 Å, larger than typical bond lengths for the materials considered in this chapter, and the lattice parameters of the larger cell are set to be no less than $3\times R_{cut}$. Finally, the remaining structure together with the imaginary atom are converted into a graph representation as shown in Figure 6-1d by connecting neighbors. The CGCNN is then trained on the local-environment-based graphs with the charge density on the grid-points from DFT calculations as the target property (with units of $e/Å^3$). The neural network structure is summarized in Figure 6-1d. Details of the DFT calculations and representation of the imaginary atom are given in Chapter 6.5. This process meets both of the requirements as mentioned above, since after convolution the atom feature vector for the imaginary atom encodes the distances between one grid-point and neighboring lattice atoms, while that for atoms of materials encodes

their position with respect to not only other atoms of materials but also the imaginary atom. The pooling process incorporates all the information together and make the final feature vector of the same dimension for materials with different compositions.

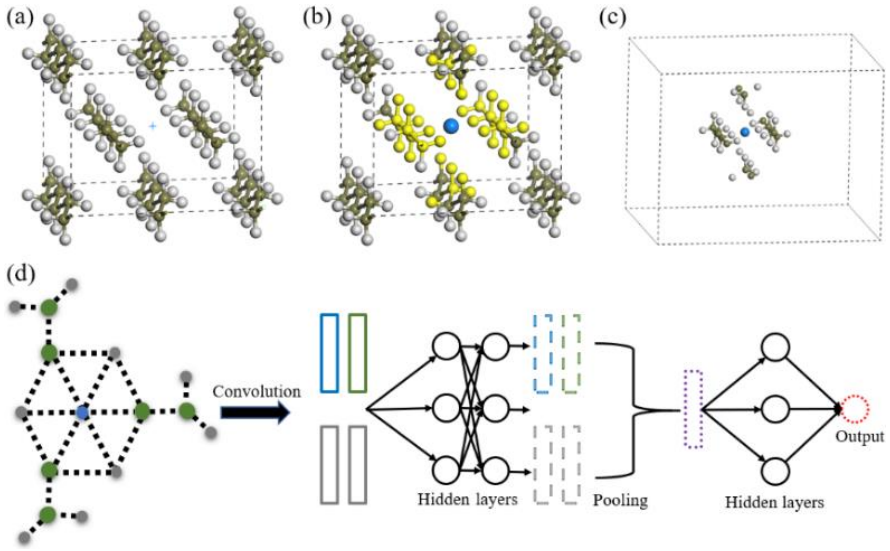

Figure 6-1 a Crystal structure of crystalline ethylene. The blue plus symbol in the center denotes a grid point we are interested in. **b** Crystalline ethylene with the imaginary atom. Highlighted atoms are those within the cut-off radius. **c** Local environment around the imaginary atom. **d** Sketch of local-environment-based graph and CGCNN architecture. Color coding: green: carbon; grey: hydrogen; blue: imaginary atom; yellow: highlighted atoms within the cut-off radius.

6.3. Prediction of charge density distribution

In the case of crystalline polymers, we extract 30,000 graphs (grid-points) from 37 different structures as training data, while in the case of zeolites, 8,000 graphs are generated from 5 different structures for training. In order to test the degree of transferability towards different structures, we apply our model to predict the charge density of 17 crystalline polymers and 9 zeolites not included in the training sets, as shown in Table 6-1. In both cases, the nomex polymer and NPO zeolite, also have versions with explicitly created defect structures (denoted as nomex_defect and NPO_defect) in order to represent additional chemical complexity. These

materials are not subsets of the training sets in terms of structure or size. Structural features are represented by coordinations of skeleton atoms (C/O in the case of polymer/zeolite). For example, C_2H_2 means there are 2 C atoms and 2 H atoms coordinated with the central atom. For polymers, in Figure 6-2a the frequency of different coordinations for carbon atoms is shown for both the training and test sets, from which one can see that nearly 20 different coordinations appear, showing considerable bonding complexity. More importantly, there are three coordinations in the test set that are not included in the training set (H_4, C_1H_1 and C_4). For zeolites, the training set is simpler than the polymer set in terms of structure, as only two coordinations exist, and in the test set only the structure with a defect, NPO_defect, has the coordination of Si_1, while all other structures have coordination Si_2. From the perspective of size, for polymers, structures in the training set span a range from 8 to 288 atoms in the unit cell, while the structures in the test set span a range from 24 to 504 atoms, and for zeolites the size ranges are 120 to 366 atoms and 18 to 576 atoms for the training and test set, respectively.

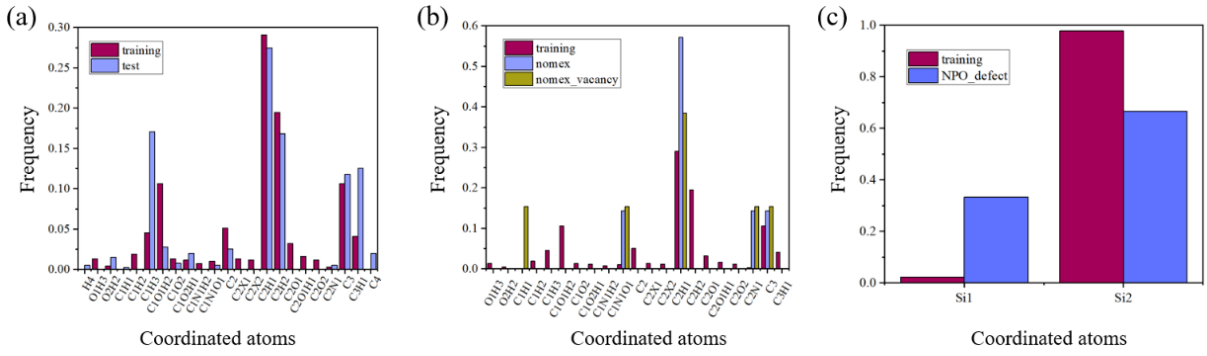

Figure 6-2 a and **b** Appearance frequency of coordinated atoms of carbon atoms in the training set for the case of crystalline polymers versus the test set as a whole and nomex and nomex_defect, respectively. Here 'X' denotes rare elements in our case (Cl, F, S, Si, Hg). **c** Appearance frequency of oxygen coordinated atoms in the training set for the case of zeolites versus the structure of NPO_defect.

Table 6-1. Root mean square errors (RMSE) and coefficients of determination (R^2) of the ML predicted charge density (ρ, in e/Å^3). For each structure, the error metrics are computed over all grid-points in the unit cell. The last nine structures with 3-letter abbreviations are zeolites, and others are crystalline polymers.

name	formula (inside the cell)	RMSE (ρ)	R^2 (ρ)
1,3-dioxolane-II	$C_{24}H_{48}O_{16}$	0.0628	0.9933
acetaldehyde	$C_{32}H_{64}O_{16}$	0.0818	0.9848
cis-1,4-butadiene	$C_{16}H_8$	0.0902	0.9805
glycolide	$C_8H_8O_8$	0.0681	0.9943
gutta-percha-alpha	$C_{20}H_{32}$	0.0369	0.9953
i-4m1p	$C_{168}H_{336}$	0.0666	0.9729
i-alpha-vnaph	$C_{192}H_{160}$	0.0661	0.9816
i-ortho-mths	$C_{144}H_{160}$	0.0593	0.9831
i-propylene-alpha	$C_{36}H_{72}$	0.0491	0.9881
isobutylene	$C_{64}H_{128}$	0.0910	0.9569
nomex	$C_{14}H_{10}O_2N_2$	0.0626	0.9926
nomex_defect	$C_{13}H_9O_2N_2$	0.0665	0.9913
oxymethylene	$C_4H_8O_4$	0.0786	0.9926
p-xylylene	$C_{16}H_8$	0.0580	0.9890
s-propylene-1	$C_{24}H_{12}$	0.0523	0.9835
tetramtht	$C_{12}H_{12}O_4$	0.0502	0.9960
trans-decenamer	$C_{10}H_{18}$	0.0309	0.9970
NPO	Si_6O_{12}	0.0977	0.9893
NPO_defect	Si_5O_{12}	0.1798	0.9745
JBW	Si_6O_{12}	0.0847	0.9914
CAN	$Si_{12}O_{24}$	0.0831	0.9906
AFY	$Si_{16}O_{32}$	0.0778	0.9894
JSN	$Si_{16}O_{32}$	0.0785	0.9911
MTN	$Si_{136}O_{272}$	0.0821	0.9903
TUN	$Si_{192}O_{384}$	0.0754	0.9920
UOV	$Si_{176}O_{352}$	0.0912	0.9881

Here, we choose two metrics, root mean square errors (RMSE) and coefficients of determination (R^2), to quantify errors in the ML predicted charge density. These metrics, also used in Schmidt *et al.*[306], provide insights on both the magnitude of absolute errors (by RMSE) and relative performance of the predictions (by R^2). As shown in Table 6-1, the RMSE of the predicted charge densities are all less than 0.2 e/Å^3, which are comparable to the errors in Schmidt *et al*. [306], and the level of accuracy was demonstrated to be sufficient for most applications relying on the accuracy of the density representation[307]. The RMSEs of test structures are also close to that of the training sets (0.067 e/Å^3 and 0.064 e/Å^3 for crystalline polymers and zeolites, respectively), indicating little overfitting. More importantly, the R^2 are larger than 0.95 for all test structures, suggesting a high prediction performance. The results for the case of zeolites show that for such a simple materials class, accurate prediction of the charge density can be achieved with a relatively small training set (less than 10,000 training data in this case). In addition to these general trends, we highlight the cases with different coordination environments (i-4m1p, isobutylene, and the nomex_defect). Although larger errors are observed in these cases, they are not far from other structures, suggesting good transferability to unseen structural features.

In order to visualize the performance and transferability of our model, we compare the ML computed charge densities and difference between charge densities from ML and DFT of pristine nomex, nomex with a C-H vacancy, pristine NPO and NPO with a Si vacancy in Figure 6-3. In all the cases, the building blocks of structures (e.g., the C six-ring and Si-O six-ring) are well presented. For defect structures, although there are more significant differences between ML and DFT, the magnitude of the difference is still low compared with the charge density itself, suggesting high transferability towards defect structures.

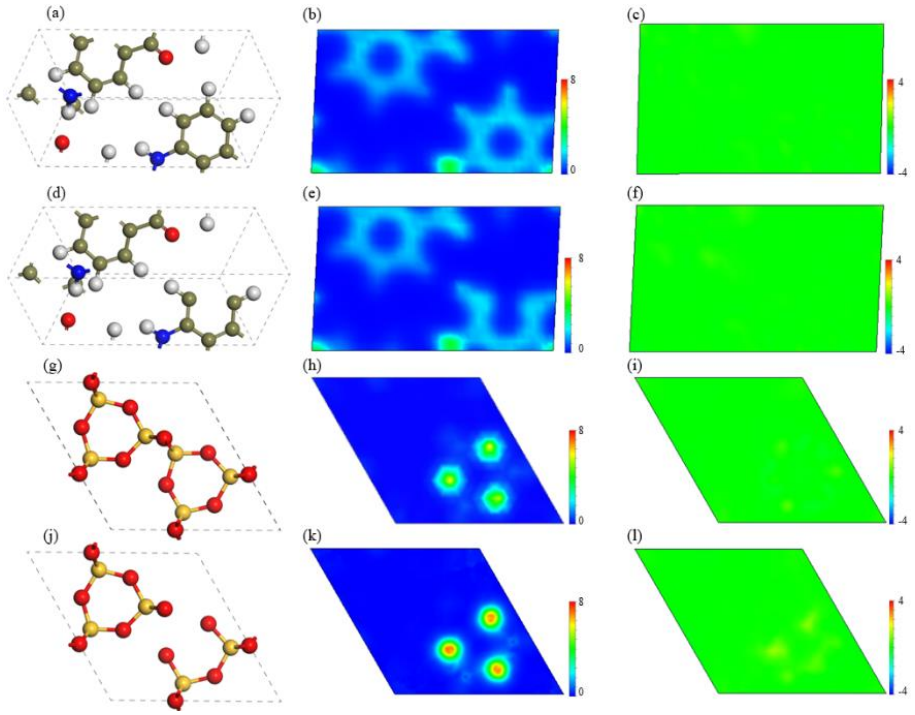

Figure 6-3. Visualization of electron charge density (ρ, in e/Å^3). **a, b, c** and **d, e, f** crystal structure, ML predicted ρ, and difference between ML predicted ρ and DFT calculated ρ on the C six-ring plane of pristine nomex and nomex with a carbon and a hydrogen vacancy, respectively. **g, h, i** and **j, k, l** crystal structure, ML predicted ρ, and difference between ML predicted ρ and DFT calculated ρ on the Si-O six-ring plane of pristine NPO and NPO with a Si vacancy, respectively. Atom color coding: green: carbon; grey: hydrogen; red: oxygen; blue: nitrogen; yellow: silicon.

We further compare the value of ML predicted ρ versus DFT calculated ρ as shown in Figure 6-4. The ML model successfully captures the charge densities in most regions for the four structures with well alignment. As shown in Figure 6-4b and 6-4d, our ML model is able to accurately capture the charge density of a vacancy even though no defect structures were present in the training sets. Meanwhile, we can see that most of the deviation in the ML approach compared with DFT is from regions with ultrahigh charge density (near atom cores as shown in Figure 6-3).

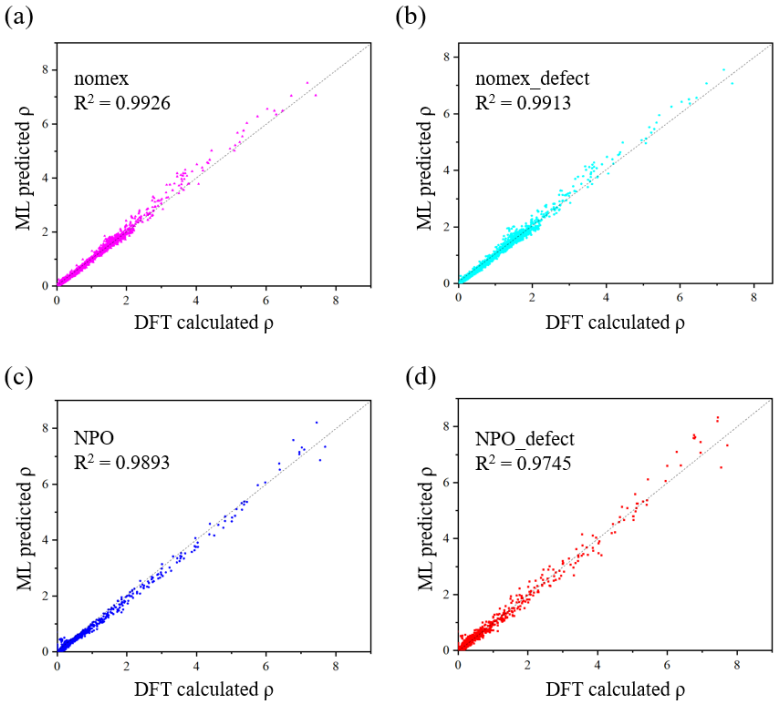

Figure 6-4. a, b, c and **d** ML predicted charge density (ρ, in e/Å^3) versus DFT calculated ρ for pristine nomex, nomex_defect, pristine NPO and NPO_defect, respectively.

6.4. Discussion about transferability

In order to probe the origin of the transferability of our model, we propose that the difficulty of transferability between different structures arises from both training and prediction: in training, the model has to distinguish between environments that seems to be 'similar' but have very different values of charge, and in prediction, the model has to find similarities between new and existing features. Here, the geometry of neighboring atoms contained in our graph representation simultaneously provides the information for the two tasks, leading to the improved transferability of our model. On the one hand, encoding the geometry makes the local environments more distinguishable; on the other hand, learning the geometry enables the model to speculate new structural features from existing ones, which also helps to predict the shape of charge density around the defects from the shape of structural features.

In order to illustrate the impact of encoding the geometry of neighboring atoms for

distinguishing local environments, we sketch two local environments in Figure 6-5a. If the environments of grid-points are simply described by considering distances to each atom separately and then summing atom contributions as in the previous models, the two environments would appear to be very similar. However, they are actually quite different, and the difference can be explicitly encoded by the distance between the two atoms. For speculating new structural features from existing ones, we plot the geometries of central carbon atoms with coordinated C_1H_3, C_2H_2, C_3H_1 and C_4 atoms in Figure 6-5b. When predicting charge density around C_4, our model can learn from the geometries of C_1H_3, C_2H_2, C_3H_1 in the training set that the tetrahedral shape of C_4 corresponds to a sp^3-hybridized central carbon atom, which gives key information for charge distribution around the central carbon atom. As for transferability to defect-induced structural features, although in the nomex_defect case there is a structural feature (C_1H_1) that doesn't exist in the training set with all pristine structures, as shown in Figure 6-5c, the shape of C_1H_1 (C-C-H, an obtuse angle) is very similar to that of C-O-H in the training set. Therefore, the charge distributions around the two structural features should be both in a shape of obtuse angle. With the information of geometries, our model can capture such similarity and predict the obtuse-angle-like charge density around C_1H_1.

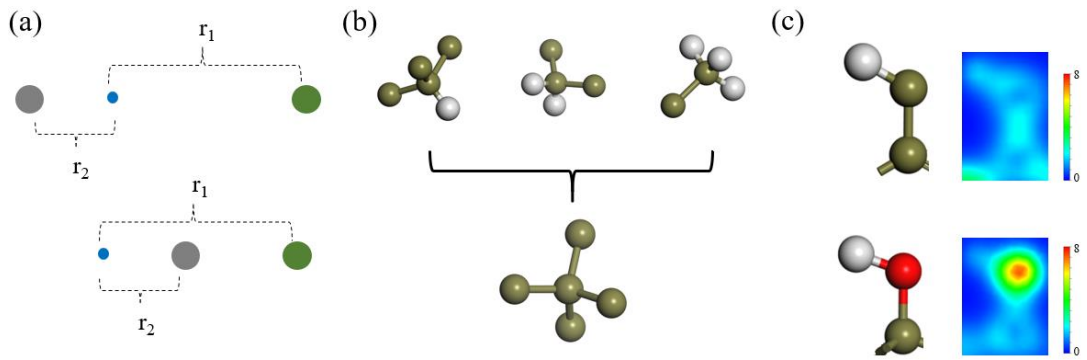

Figure 6-5. a Sketch of two different local environments with similar sum of atom contributions. **b** Geometries of central carbon atoms with coordinated C1H3, C2H2, C3H1 and C4 atoms. **c** Shape of C-C-H and C-O-H and their charge density distributions (ρ, in e/Å3).

Atom color coding: green: carbon; grey: hydrogen; red: oxygen.

We further conduct a toy-model experiment to verify the above statement regarding geometry-induced transferability, shown in Figure 6-6a. First, a CGCNN model is trained on 3000 grid-points within 4Å of a linear C-C-C molecule, and then used to predict the charge density of an orthogonal C-C-C molecule. To examine the effect of geometry towards predicting new structural features, we sample a new set of 3000 grid-points equally from both the linear C-C-C molecule and an orthogonal C-O-C molecule and train another CGCNN model, and we find that after incorporating the orthogonal geometry into the training set, the prediction error to the orthogonal C-C-C molecule decreases and is lower than that of the two single-training molecule cases, which shows that encoding geometry helps to predict new structural features. Another insight from this experiment is that currently transferability between elements is still limited in the sense that it is difficult to predict the ratio of charge density between C-C only from that of C-O, which can be attributed to the poor design of the element feature vector, a subject of further investigation in future work.

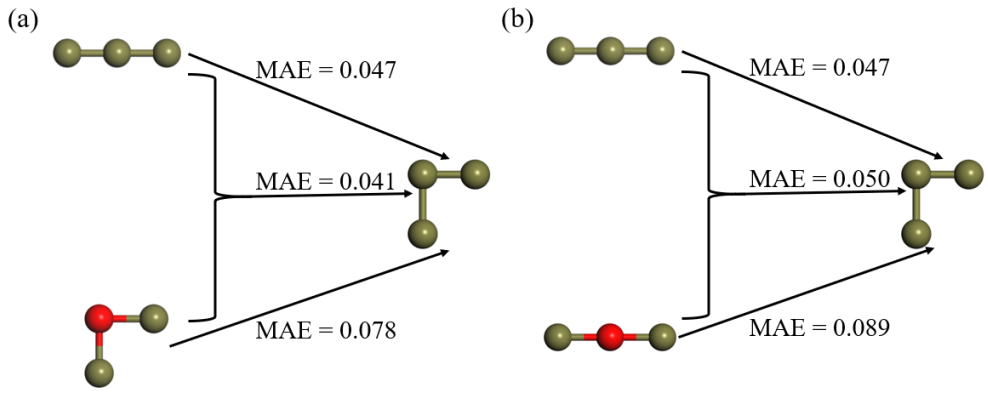

Figure 6-6. a Illustration of the first toy-model experiment. The top and bottom MAEs (in e/Å^3) are from the predictions to the orthogonal C-C-C molecule by one of the two training molecules (linear C-C-C and orthogonal C-O-C), while the middle one is from the prediction trained on both of the training molecules. **b** Illustration of the second toy-model experiment. The top and bottom MAEs (in e/Å^3) are from the predictions to the orthogonal C-C-C molecule by one of the two training molecules (linear C-C-C and linear C-O-C), while the middle one is from the prediction trained on both of the training molecules (linear C-C-C and linear C-O-C). Atom

color coding: green: carbon; red: oxygen.

In Chapter 6.1, we also mention the effect of the same dimension of the features. We do believe that the same-dimensional features facilitates the training process, since if the dimension of features scales linearly with the number of elements, then the time of training will also largely depend on it, which is undesirable in multi-elemental systems. However, the same dimension of the features is not the most fundamental origin of transferability, and it is less important than encoding geometry, the main origin of transferability as proposed above. In order to further verify the dominant role of geometry in transferability, we perform another toy-model experiment and illustrate it in Figure 6-6b. In this experiment, we sample a set of 3000 grid-points from both the linear C-C-C molecule and a new linear C-O-C molecule and train a CGCNN model with the same settings as the first experiment. Therefore, the dramatic increase of the prediction error to the orthogonal C-O-C molecule and the fact that it is higher than that of the case with the linear C-C-C molecule as the single training molecule can only be attributed to the geometry of training molecules, which shows that the transferability achieved in the first experiment is the result of only geometry, not other conditions including the dimension of features.

6.5. Details of methods

Details of DFT calculations. DFT calculations to obtain charge density distributions are implemented in the Vienna *Ab initio* Simulation Package (VASP)[265]. The exchange-correlation is approximated by Perdew-Burke-Ernzerh functional (PBE)[123]. For the calculation of time scaling, the first Brillouin zone is sampled by a $2 \times 2 \times 2$ k-point grid, while that for other calculations is of ~0.5 $Å^{-1}$. In order to account for van der Waals forces, the DFT-D2[308] dispersion-correlation is used.

Discussion about how to represent the imaginary atom. In principle, any representation of the imaginary atom that is different from those for elements in our system is acceptable. Since CGCNN constructs a representation for atoms based on elemental properties, here for simplicity we use the representation of the He atom in CGCNN to represent the imaginary atom, as He doesn't exist in our cases nor most periodic systems, and Table 6-2 shows that different representations of the imaginary atom would lead to similar performance. Nevertheless, when necessary one can always construct other representations different from all existing elements such as adding additional dimensions to tag the imaginary atom.

Table 6-2. Mean average errors (MAEs, in e/Å^3) of the training set in the zeolite case versus the choice of representation of imaginary atom.

Choice of imaginary atom	He	Li	Ne	Cs	Xe
MAE	0.030	0.041	0.037	0.035	0.036

Dataset construction and grid spacing. For the case of crystalline polymers, initially 52 structures were downloaded from the database in *Materials Studio*, and then randomly split into training set and test set with the ratio of 70% and 30% (36 and 16), respectively. A defect structure was generated to test the transferability from pristine structures. One elemental crystal (graphite) was added to the training set to increase its complexity, giving a training set with 37 structures and test set with 17 structures.

For the case of zeolites, 5 structures with intermediate size are randomly selected from the database of *Structure Commission of the International Zeolite Association* as the training set. As for the design of test sets, 5 small zeolite structures are manually included to test the

transferability from large structures to small while 3 structures larger than that in the training set are also included with similar intention. One defect structure is also manually created to test the transferability from pristine structures.

After collecting structures, for each structure in the training sets, all the symmetrically inequivalent grid-points inside the unit cell with a given spacing (~0.5 Å for polymers and ~0.75 Å for zeolites) are converted into graphs as discussed in the main text. In order to avoid bias towards certain structures, in the pool of graphs from all the structures, the maximum number of graphs from one structure is set be 2,000. Then, some graphs are randomly picked from the pool as the training data, on top of which CGCNN is trained. When calculating the error statistics of each test structure, all the grid-points in that structure are considered.

For training sets, the grid spacing for polymers is set to ~0.5 Å and for zeolites ~0.75 Å. For test sets, for crystalline polymers and the six zeolites with small unit cells, the charge density is predicted on a grid of ~0.5 Å while for the three large zeolites it is set to ~0.75 Å. For visualization, a refined grid of ~0.25 Å was used.

6.6. Chapter summary and outlook

In summary, we have developed a machine learning model to predict electron charge density distribution of materials based on graph convolutional neural networks with. In the case studies of crystalline polymers and zeolites, local-environment-based graphs are extracted from some structures and features learned, and the learned models are applied to structures different from the training sets. The accuracy and usability of our model has been evaluated by statistical errors and visualization. The most important benefit of our model is high transferability between different structures, which can be attributed to the ability of the graph representation to explicitly encode the geometry of local environment.

Future efforts will be applied to further improve the scheme presented in four aspects.

First, we will optimize the algorithm to achieve lower computational cost. One of the possible directions is switching from the sequential prediction of each grid points to parallel predictions of many grid points simultaneously. Second, as mentioned we will design architectures to efficiently generate more materials properties based on charge density, especially the total energy of the unit cell, for which both traditional methods (e.g. Kohn-Sham equations[158] or embedded-atom method[309]) and machine learning approaches[303-305] are options under consideration. Third, as discussed above regions near nuclei possess the highest deviations, and to improve the sensitivity of our model for small distances between imaginary and real atoms, transformations to weight small distances during the learning can be designed. Last, as mentioned we aim to develop new atom feature vectors that can achieve better transferability between different elements, with one possible approach to learn atomic features back from charge density distributions around each type of atom.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

7. Conclusion and outlook

7.1. Summary of the thesis

In summary, methodologically, this thesis is centered around the question of how to improve the prediction performance of machine learning models for materials science. In this thesis, we propose and apply a series of strategies to improve performance of machine learning models from different perspectives: representation of materials, information transfer, and expansion of datasets. For expanding datasets, two strategies are proposed, machine learning-guided sampling and machine learning-accelerated simulations.

First of all, we exam whether the current representation of materials can fully represent the materials, or in other words, capture all knowledge of materials. In Chapter 2, we exam whether two graph neural networks, CGCNN and ALIGNN, can capture knowledge behind human-designed descriptors. We find that both of them can capture local atomic environments well, but both cannot capture periodicity of crystal structures well. As an initial solution, we hybridize the descriptors with the GNNs, which leads to large improvement of prediction accuracy for properties where uncaptured long-range information is critical.

Then, for situations where the dataset associated with the learning task is small while there exist large relevant datasets, which is very common in materials, we propose the idea of information transfer between the large and small datasets to improve the learning performance of the small dataset. In Chapter 3, we study how multi-fidelity learning and transfer learning, two information transfer strategies, help to learn the experimentally measured formation enthalpies of materials, from which we obtain qualitative insights about where and why multi-fidelity learning and transfer learning improves. In Chapter 4, we investigate how transfer

learning helps to learn the experimentally measured lattice thermal conductivity, where we visualize the origin of improvement of transfer learning.

Finally, if algorithm-design cannot lead to satisfying prediction performance, we suggest researchers to expand the dataset. During the expansion, for sampling efficiency and mitigation of bias, we propose to use active learning or Bayesian Optimization. In Chapter 5, we use Bayesian Optimization to search for the optimal laser-processing parameters for poly(acrylonitrile). If the data collection procedure (experiment or computation) is too expansive, then we suggest researchers to use machine learning to accelerate the collection. In Chapter 6, we propose a way to predict charge density distributions of materials by machine learning, which can potentially accelerate DFT calculations.

From the perspective of materials discovery, machine learning models developed in this thesis help to propose new materials systems, processes, and insights. In Chapter 3, we find hundreds of materials that might have underestimated stability from the cheap DFT functionals (PBE). In Chapter 4, we propose the system of rare-earth-chalcogenides (REXs) as promising thermoelectric materials, which is verified through experiments. In Chapter 5, we find laser-processing parameters that can transform insulative poly(acrylonitrile) sheet into conductive porous carbon electrodes with desirable electrochemical properties.

7.2. Future directions

Despite the progresses made in this thesis, there are still many challenges for machine learning applications in materials science that require further development and understanding of machine learning models. Here we summarize two most emergent and critical challenges about representation of materials and information transfer, followed by some challenges for active learning/Bayesian Optimization and machine learning accelerated DFT, as well as a

challenge for machine learning applications in materials science for more distant future.

More powerful and efficient representations of materials: as suggested in Chapter 3, current representation of materials still cannot capture all human knowledge. Although Batatia *et al.*[121] have proposed a general formalism to encode the local atomic environments equivariantly to E(3) symmetry group by GNN or atomistic cluster expansion with arbitrary body-order, how to capture long-range information is still challenging. Although deeper GNN and larger receptive field are natural solutions to encode long-range information, there are still practical challenges for training deeper GNNs such as bottleneck[145] and over-smoothing[146]. Even if deep GNNs can be effectively trained, there is a fundamental trade-off between number of convolutions, number of neighbors and computational cost. For applications sensitive to cost, such as molecular dynamics, more efficient representations of materials are necessary to achieve low cost as well as to capture long-range information.

Quantitative metric to estimate whether information transfer will help: despite the recent success of transfer learning and multi-fidelity learning for improving learning performance of small materials datasets[49, 55, 61, 64, 66, 97] such as in Chapter 3 and 4, information transfer is still empirical in materials science. In other words, for materials science, there is no quantitative metric about whether information transfer will help to improve prediction of target compared with training models from scratch. Currently, for materials datasets, the criterion for choosing the source and target dataset for information transfer is whether the two datasets are "strongly" correlated, and choosing the correct source datasets for specific target sets is still based on trial-and-error. Since training machine learning models based on gradient is still expensive, the lack of quantitative metric to estimate the usefulness of information transfer before training limits the application of information transfer. For the goal of quantitative estimation, further studies are necessary to develop quantitative metrics

that depend only on the distribution of the two datasets and do not require the pre-trained models to obtain.

For active learning/Bayesian optimization, as suggested in Chapter 5, the incapability of gaussian process for highly nonlinear functions is still a big challenge. Although neural networks have stronger power to fit nonlinear functions, how to estimate uncertainty of neural networks is still an open question. On the other hand, the question of whether lowering uncertainty guarantees lowering prediction error is still not fully answered. For machine learning accelerated DFT, although there are already methods such as that in Chapter 6 to generate charge density distributions of materials, the current methods that generate and predict charge density are still not as transferable as physics-based simulation methods such as DFT. It is necessary to develop machine learning assisted-DFT methods that can be applied to very different systems once trained.

Since there will be inevitably more materials data in the future, data-driven machine learning models will be more powerful and more widely used in materials science. With more data, the prediction performance of supervised learning will be inevitably improved. In the future, in addition to prediction of well-defined materials properties, we hope that machine learning can be used to learn the pattern of "ambiguously" defined materials properties, such as synthesizability, processability and toxicity of materials, which are critical yet lack quantitative metrics to evaluate.

Bibliography

1. Https://Www.Cpp.Edu/~Jbputhoff/History.Html.
2. Hu, A.; Levis, S.; Meehl, Gerald A.; Han, W.; Washington, Warren M.; Oleson, Keith W.; van Ruijven, Bas J.; He, M.; Strand, Warren G., Impact of Solar Panels on Global Climate. *Nat. Clim. Change* 2015, *6* (3), 290.
3. Zeng, X.; Li, M.; Abd El-Hady, D.; Alshitari, W.; Al-Bogami, A. S.; Lu, J.; Amine, K., Commercialization of Lithium Battery Technologies for Electric Vehicles. *Adv. Energy Mater.* 2019, *9* (27).
4. Richard, J., World's Slowest-Moving Drop Caught on Camera at Last. *Nature* 2013, *https://doi.org/10.1038/nature.2013.13418*, 1783.
5. Agrawal, A.; Choudhary, A., Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* 2016, *4* (5).
6. A. Jain*, S. P. O., G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* 2013, *1* (1), 011002.
7. Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C., The Open Quantum Materials Database (Oqmd): Assessing the Accuracy of Dft Formation Energies. *npj Comput. Mater.* 2015, *1* (1).
8. Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O., Aflowlib.Org: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* 2012, *58*, 227.
9. Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Walker, A. R. H.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A. J. a. p. a., The Joint Automated Repository for Various Integrated Simulations (Jarvis) for Data-Driven Materials Design. *npj Comput Mater* 2020, *6*, 173.
10. Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C., Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Comput. Mater.* 2022, *8* (1).
11. Batra, R.; Song, L.; Ramprasad, R., Emerging Materials Intelligence Ecosystems Propelled by Machine Learning. *Nat. Rev. Mater.* 2020, *6* (8), 655.
12. Moosavi, S. M.; Jablonka, K. M.; Smit, B., The Role of Machine Learning in the Understanding and Design of Materials. *J Am Chem Soc* 2020.
13. Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P., Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv Sci (Weinh)* 2019, *6* (21), 1900808.
14. Horton, M. K.; Dwaraknath, S.; Persson, K. A., Promises and Perils of Computational Materials Databases. *Nat. Comput. Sci.* 2021, *1* (1), 3.
15. Attia, P. M.; Grover, A.; Jin, N.; Severson, K. A.; Markov, T. M.; Liao, Y. H.; Chen, M. H.; Cheong, B.; Perkins, N.; Yang, Z.; Herring, P. K.; Aykol, M.; Harris, S. J.; Braatz, R. D.; Ermon, S.; Chueh, W. C., Closed-Loop Optimization of Fast-Charging Protocols for Batteries with Machine Learning. *Nature* 2020, *578* (7795), 397.
16. Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* 2016, *533* (7601), 73.

17. Segler, M. H. S.; Preuss, M.; Waller, M. P., Planning Chemical Syntheses with Deep Neural Networks and Symbolic Ai. *Nature* 2018, *555* (7698), 604.
18. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G., Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* 2021, *590* (7844), 89.
19. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, *12*, 2825.
20. Lakshminarayanan, B.; Pritzel, A.; Blundell, C., Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *31st Conference on Neural Information Processing Systems*, 2017; pp 6402.
21. Chen, T.; Chen, H., Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and Its Application to Dynamical Systems. *IEEE trans. neural netw.* 1995, *6* (4), 911.
22. Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L., Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* 2019, *5* (1), 1.
23. Goodall, R. E. A.; Lee, A. A., Predicting Materials Properties without Crystal Structure: Deep Representation Learning from Stoichiometry. *Nat Commun* 2020, *11* (1), 6280.
24. Jha, D.; Ward, L.; Paul, A.; Liao, W. K.; Choudhary, A.; Wolverton, C.; Agrawal, A., Elemnet: Deep Learning the Chemistry of Materials from Only Elemental Composition. *Sci Rep* 2018, *8* (1), 17593.
25. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C., A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* 2016, *2* (1).
26. Xie;, T.; Fu;, X.; Ganea;, O.-E.; Barzilay;, R.; Jaakkola, T., Crystal Diffusion Variational Autoencoder for Periodic Material Generation. *International Conference on Learning Representations* 2022.
27. Cheng, G.; Gong, X. G.; Yin, W. J., Crystal Structure Prediction by Combining Graph Network and Optimization Algorithm. *Nat Commun* 2022, *13* (1), 1492.
28. Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y., Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent Sci* 2020, *6* (8), 1412.
29. Jonathan, S.; Love, P.; Verdozzi;, C.; Botti, S.; Marques, M. A. L., Crystal Graph Attention Networks for the Prediction of Stable Materials. *Sci. Adv.* 2021, *7* (49), eabi7948.
30. Fung, V.; Hu, G.; Ganesh, P.; Sumpter, B. G., Machine Learned Features from Density of States for Accurate Adsorption Energy Prediction. *Nat Commun* 2021, *12* (1), 88.
31. Modarres, M. H.; Aversa, R.; Cozzini, S.; Ciancio, R.; Leto, A.; Brandino, G. P., Neural Network for Nanoscience Scanning Electron Microscope Image Recognition. *Sci Rep* 2017, *7* (1), 13282.
32. Kusche, C.; Reclik, T.; Freund, M.; Al-Samman, T.; Kerzel, U.; Korte-Kerzel, S., Large-Area, High-Resolution Characterisation and Classification of Damage Mechanisms in Dual-Phase Steel Using Deep Learning. *PLoS ONE* 2019, *14* (5), e0216493.
33. Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M., Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction. *Appl. Phys. Rev.* 2020, *7* (4).
34. Hiszpanski, A. M.; Gallagher, B.; Chellappan, K.; Li, P.; Liu, S.; Kim, H.; Han, J.; Kailkhura, B.; Buttler, D. J.; Han, T. Y., Nanomaterial Synthesis Insights from Machine Learning of Scientific

Articles by Extracting, Structuring, and Visualizing Knowledge. *J. Chem. Inf. Model.* 2020, *60* (6), 2876.
35. Kim, E.; Jensen, Z.; van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H. S.; Strubell, E.; McCallum, A.; Jegelka, S.; Olivetti, E., Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* 2020, *60* (3), 1194.
36. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A., Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* 2019, *571* (7763), 95.
37. Iacomi, P.; Llewellyn, P. L., Data Mining for Binary Separation Materials in Published Adsorption Isotherms. *Chem. Mater.* 2020, *32* (3), 982.
38. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G., Opportunities and Challenges of Text Mining in Aterials Research. *iScience* 2021, *24* (3), 102155.
39. Bessa, M. A.; Glowacki, P.; Houlder, M., Bayesian Machine Learning in Metamaterial Design: Fragile Becomes Supercompressible. *Adv Mater* 2019, *31* (48), e1904845.
40. Wahab, H.; Jain, V.; Tyrrell, A. S.; Seas, M. A.; Kotthoff, L.; Johnson, P. A., Machine-Learning-Assisted Fabrication: Bayesian Optimization of Laser-Induced Graphene Patterning Using in-Situ Raman Analysis. *Carbon* 2020, *167*, 609.
41. Dave, A.; Mitchell, J.; Kandasamy, K.; Wang, H.; Burke, S.; Paria, B.; Póczos, B.; Whitacre, J.; Viswanathan, V., Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning. *Cell Rep. Phys. Sci.* 2020, *1* (12).
42. Li, C.; Rubin de Celis Leal, D.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.; Height, M.; Venkatesh, S., Rapid Bayesian Optimisation for Synthesis of Short Polymer Fiber Materials. *Sci Rep* 2017, *7* (1), 5683.
43. Yu, M.; Yang, S.; Wu, C.; Marom, N., Machine Learning the Hubbard U Parameter in Dft+U Using Bayesian Optimization. *npj Comput. Mater.* 2020, *6* (1).
44. Tavadze, P.; Boucher, R.; Avendaño-Franco, G.; Kocan, K. X.; Singh, S.; Dovale-Farelo, V.; Ibarra-Hernández, W.; Johnson, M. B.; Mebane, D. S.; Romero, A. H., Exploring Dft+U Parameter Space with a Bayesian Calibration Assisted by Markov Chain Monte Carlo Sampling. *npj Comput. Mater.* 2021, *7* (1).
45. Wang, W.; Gómez-Bombarelli, R., Coarse-Graining Auto-Encoders for Molecular Dynamics. *npj Comput. Mater.* 2019, *5* (1).
46. Cheng, J.; Zhang, C.; Dong, L., A Geometric-Information-Enhanced Crystal Graph Network for Predicting Properties of Materials. *Commun. Mater.* 2021, *2* (1).
47. Karamad, M.; Magar, R.; Shi, Y.; Siahrostami, S.; Gates, I. D.; Barati Farimani, A., Orbital Graph Convolutional Neural Network for Material Property Prediction. *Phys. Rev. Mater.* 2020, *4* (9).
48. 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 Cent Sci* 2018, *4* (8), 996.
49. Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R., Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent Sci* 2019, *5* (10), 1717.
50. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem Sci* 2017, *8* (7), 5137.
51. Zheng, X.; Zheng, P.; Zhang, R. Z., Machine Learning Material Properties from the Periodic Table Using Convolutional Neural Networks. *Chem Sci* 2018, *9* (44), 8426.
52. Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R., A Multi-

Fidelity Information-Fusion Approach to Machine Learn and Predict Polymer Bandgap. *Comput. Mater. Sci.* 2020, *172*, 109286.
53. Pilania, G.; Gubernatis, J. E.; Lookman, T., Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* 2017, *129*, 156.
54. Chen, L.; Tran, H.; Batra, R.; Kim, C.; Ramprasad, R., Machine Learning Models for the Lattice Thermal Conductivity Prediction of Inorganic Materials. *Comput. Mater. Sci.* 2019, *170*.
55. Zhu, T.; He, R.; Gong, S.; Xie, T.; Gorai, P.; Nielsch, K.; Grossman, J. C., Charting Lattice Thermal Conductivity for Inorganic Crystals and Discovering Rare Earth Chalcogenides for Thermoelectrics. *Energy Environ. Sci.* 2021, *14* (6), 3559.
56. Jang, J.; Gu, G. H.; Noh, J.; Kim, J.; Jung, Y., Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. *J Am Chem Soc* 2020, *142* (44), 18836.
57. Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R., Schnet - a Deep Learning Architecture for Molecules and Materials. *J Chem Phys* 2018, *148* (24), 241722.
58. Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W., Convolutional Neural Network of Atomic Surface Structures to Predict Binding Energies for High-Throughput Screening of Catalysts. *J Phys Chem Lett* 2019, *10* (15), 4401.
59. Liang, J.; Zhu, X., Phillips-Inspired Machine Learning for Band Gap and Exciton Binding Energy Prediction. *J Phys Chem Lett* 2019, *10* (18), 5640.
60. Gong, S.; Wang, S.; Zhu, T.; Chen, X.; Yang, Z.; Buehler, M. J.; Shao-Horn, Y.; Grossman, J. C., Screening and Understanding Li Adsorption on Two-Dimensional Metallic Materials by Learning Physics and Physics-Simplified Learning. *JACS Au* 2021.
61. Gupta, V.; Choudhary, K.; Tavazza, F.; Campbell, C.; Liao, W. K.; Choudhary, A.; Agrawal, A., Cross-Property Deep Transfer Learning Framework for Enhanced Predictive Analytics on Small Materials Data. *Nat Commun* 2021, *12* (1), 6595.
62. Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A., Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat Commun* 2017, *8*, 15679.
63. Banjade, H. R.; Hauri, S.; Zhang, S.; Ricci, F.; Gong, W.; Hautier, G.; Vucetic, S.; Yan, Q., Structure Motif Centric Learning Framework for Inorganic Crystalline Systems. *Sci. Adv.* 2021, *7* (17), eabf1754.
64. Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W. K.; Choudhary, A.; Campbell, C.; Agrawal, A., Enhancing Materials Property Prediction by Leveraging Computational and Experimental Data Using Deep Transfer Learning. *Nat Commun* 2019, *10* (1), 5316.
65. Ye, W.; Chen, C.; Wang, Z.; Chu, I. H.; Ong, S. P., Deep Neural Networks for Accurate Predictions of Crystal Stability. *Nat Commun* 2018, *9* (1), 3800.
66. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P., Learning Properties of Ordered and Disordered Materials from Multi-Fidelity Data. *Nat. Comput. Sci.* 2021, *1* (1), 46.
67. Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G., A Critical Examination of Compound Stability Predictions from Machine-Learned Formation Energies. *npj Comput. Mater.* 2020, *6* (1).
68. Kailkhura, B.; Gallagher, B.; Kim, S.; Hiszpanski, A.; Han, T. Y.-J., Reliable and Explainable Machine-Learning Methods for Accelerated Material Discovery. *npj Comput. Mater.* 2019, *5* (1).
69. Choudhary, K.; DeCost, B., Atomistic Line Graph Neural Network for Improved Materials Property Predictions. *npj Comput. Mater.* 2021, *7* (1).
70. Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M., Big Data of Materials Science: Critical Role of the Descriptor. *Phys Rev Lett* 2015, *114* (10), 105503.
71. Xie, T.; Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and

Interpretable Prediction of Material Properties. *Phys Rev Lett* 2018, *120* (14), 145301.
72. Choudhary, K.; DeCost, B.; Tavazza, F., Machine Learning with Force-Field Inspired Descriptors for Materials: Fast Screening and Mapping Energy Landscape. *Phys Rev Mater* 2018, *2* (8).
73. Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C., Predicting Charge Density Distribution of Materials Using a Local-Environment-Based Graph Convolutional Network. *Phys. Rev. B* 2019, *100* (18), 184103.
74. Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U., How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B* 2014, *89* (20), 205118.
75. Park, C. W.; Wolverton, C., Developing an Improved Crystal Graph Convolutional Neural Network Framework for Accelerated Materials Discovery. *Phys. Rev. Mater.* 2020, *4* (6), 063801.
76. Kong, S.; Ricci, F.; Guevarra, D.; Neaton, J. B.; Gomes, C. P.; Gregoire, J. M., Density of States Prediction for Materials Discovery Via Contrastive Learning from Probabilistic Embeddings. *Nat Commun* 2022, *13* (1), 949.
77. Chen, Z.; Andrejevic, N.; Smidt, T.; Ding, Z.; Xu, Q.; Chi, Y. T.; Nguyen, Q. T.; Alatas, A.; Kong, J.; Li, M., Direct Prediction of Phonon Density of States with Euclidean Neural Networks. *Adv. Sci.* 2021, *8* (12), e2004214.
78. Zhong;, Y.; Yu;, H.; Gong;, X.; Xiang, H., Edge-Based Tensor Prediction Via Graph Neural Networks. *arXiv:2201.05770* 2022.
79. Behler, J., Perspective: Machine Learning Potentials for Atomistic Simulations. *J Chem Phys* 2016, *145* (17), 170901.
80. Mueller, T.; Hernandez, A.; Wang, C., Machine Learning for Interatomic Potential Models. *J Chem Phys* 2020, *152* (5), 050902.
81. Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W., Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys Rev Lett* 2018, *120* (14), 143001.
82. Batzner;, S.; Musaelian;, A.; Sun;, L.; Geiger;, M.; Mailoa;, J. P.; Kornbluth;, M.; Molinari;, N.; Smidt;, T. E.; Kozinsky, B., E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic. *Nat Commun* 2022, *13*, 2453.
83. Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noe, F.; Clementi, C., Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent Sci* 2019, *5* (5), 755.
84. Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Muller, K. R., Bypassing the Kohn-Sham Equations with Machine Learning. *Nat Commun* 2017, *8* (1), 872.
85. Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O., Artificial Intelligence-Enhanced Quantum Chemical Method with Broad Applicability. *Nat Commun* 2021, *12* (1), 7022.
86. Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R., Solving the Electronic Structure Problem with Machine Learning. *npj Comput. Mater.* 2019, *5* (1), 1.
87. Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M., Transferable Machine-Learning Model of the Electron Density. *ACS Cent Sci* 2019, *5* (1), 57.
88. Kasim, M. F.; Vinko, S. M., Learning the Exchange-Correlation Functional from Nature with Fully Differentiable Density Functional Theory. *Phys Rev Lett* 2021, *127* (12), 126403.
89. Tsubaki, M.; Mizoguchi, T., Quantum Deep Field: Data-Driven Wave Function, Electron Density Generation, and Atomization Energy Prediction and Extrapolation with Machine Learning. *Phys Rev Lett* 2020, *125* (20), 206401.
90. Kirkpatrick;, J.; McMorrow;, B.; Turban;, D. H. P.; Gaunt, A. L.; Spencer, J. S.; Matthews, A.

G. D. G.; Obika, A.; Thiry, L.; Fortunato, M.; Pfau, D.; Castellanos, L. R.; Petersen, S.; Nelson, A. W. R.; Kohli, P.; Mori-Sánchez, P.; Hassabis, D.; Cohen, A. J., Pushing the Frontiers of Density Functionalsby Solving the Fractional Electron Problem. *Science* 2021, *374*, 1385.
91. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* 2018, *4* (2), 268.
92. Long, T.; Fortunato, N. M.; Opahle, I.; Zhang, Y.; Samathrakis, I.; Shen, C.; Gutfleisch, O.; Zhang, H., Constrained Crystals Deep Convolutional Generative Adversarial Network for the Inverse Design of Crystal Structures. *npj Comput. Mater.* 2021, *7* (1).
93. Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A., Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Machine Intell.* 2021, *3* (1), 76.
94. Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y., Inverse Design of Solid-State Materials Via a Continuous Representation. *Matter* 2019, *1* (5), 1370.
95. Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M., Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction. *Applied Physics Reviews* 2020, *7* (4).
96. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. I., From Local Explanations to Global Understanding with Explainable Ai for Trees. *Nat Mach Intell* 2020, *2* (1), 56.
97. Sheng Gong, S. W., Tian Xie, Woo Hyun Chae, Runze Liu, Yang Shao-Horn, Jeffrey C. Grossman, Calibrating Dft Formation Enthalpy Calculations by Multi-Fidelity Machine Learning. *JACS Au* 2022, *10.1021/jacsau.2c00235*.
98. Gallegos, L. C.; Luchini, G.; St John, P. C.; Kim, S.; Paton, R. S., Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc Chem Res* 2021, *54* (4), 827.
99. Li, X.; Dan, Y.; Dong, R.; Cao, Z.; Niu, C.; Song, Y.; Li, S.; Hu, J., Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning. *Appl. Sci.* 2019, *9* (24).
100. Loftis, C.; Yuan, K.; Zhao, Y.; Hu, M.; Hu, J., Lattice Thermal Conductivity Prediction Using Symbolic Regression and Machine Learning. *J Phys Chem A* 2021, *125* (1), 435.
101. Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W. J., Simple Descriptor Derived from Symbolic Regression Accelerating the Discovery of New Perovskite Catalysts. *Nat Commun* 2020, *11* (1), 3513.
102. Udrescu, S.-M.; Tegmark, M., Ai Feynman: A Physics-Inspired Method for Symbolic Regression. *Sci. Adv.* 2020, *6*, eaay2631.
103. Iten, R.; Metger, T.; Wilming, H.; Del Rio, L.; Renner, R., Discovering Physical Concepts with Neural Networks. *Phys Rev Lett* 2020, *124* (1), 010508.
104. Mrdjenovich, D.; Horton, M. K.; Montoya, J. H.; Legaspi, C. M.; Dwaraknath, S.; Tshitoyan, V.; Jain, A.; Persson, K. A., Propnet: A Knowledge Graph for Materials Science. *Matter* 2020, *2* (2), 464.
105. Xie, T.; Grossman, J. C., Hierarchical Visualization of Materials Space with Graph Convolutional Neural Networks. *J Chem Phys* 2018, *149* (17), 174111.
106. 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. *Nat Commun* 2019, *10*

(1), 2667.
107. Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R., Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* 2015, *115* (16), 1094.
108. Kajita, S.; Ohba, N.; Jinnouchi, R.; Asahi, R., A Universal 3d Voxel Descriptor for Solid-State Material Informatics with Deep Convolutional Neural Networks. *Sci Rep* 2017, *7* (1), 16991.
109. Mukherjee, K.; Colón, Y. J., Machine Learning and Descriptor Selection for the Computational Discovery of Metal-Organic Frameworks. *Mol. Simul.* 2021, *47* (10-11), 857.
110. Muraoka, K.; Sada, Y.; Miyazaki, D.; Chaikittisilp, W.; Okubo, T., Linking Synthesis and Structure Descriptors from a Large Collection of Synthetic Records of Zeolite Materials. *Nat Commun* 2019, *10* (1), 4459.
111. Schindler, P.; Antoniuk, E. R.; Cheon, G.; Zhu, Y.; Reed, E. J., Discovery of Materials with Extreme Work Functions by High-Throughput Density Functional Theory and Machine Learning. *arXiv:2011.10905v1* 2020.
112. Janet, J. P.; Kulik, H. J., Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J Phys Chem A* 2017, *121* (46), 8939.
113. Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A., Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput. Mater.* 2020, *6* (1), 138.
114. Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* 2019, *31* (9), 3564.
115. Louis, S. Y.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; Hu, J., Graph Convolutional Neural Networks with Global Attention for Improved Materials Property Prediction. *Phys Chem Chem Phys* 2020, *22* (32), 18141.
116. Gasteiger;, J.; Groß;, J.; Günnemann, S., Directional Message Passing for Molecular Graphs. *International Conference on Learning Representations* 2020.
117. Gasteiger, J.; Becker, F.; Günnemann, S., Gemnet: Universal Directional Graph Neural Networks for Molecules. In *35th Conference on Neural Information Processing Systems* 2021.
118. Maron, H.; Litany, O.; Chechik, G.; Fetaya, E., On Learning Sets of Symmetric Elements. In *37th International Conference on Machine Learning*, Hal, D., III; Aarti, S., Eds. 2020; Vol. 119, pp 6734.
119. Morris, C.; Rattan, G.; Mutze, P., Weisfeiler and Leman Go Sparse: Towards Scalable Higher-Order Graph Embeddings. *34th Conference on Neural Information Processing Systems* 2020.
120. Omee, S. S.; Louis, S.-Y.; Fu, N.; Wei, L.; Dey, S.; Dong, R.; Li, Q.; Hu, J., Scalable Deeper Graph Neural Networks for High-Performance Materials Property Prediction. *arXiv:2109.12283v1* 2021.
121. Batatia, I.; Batzner, S.; Kovacs, D. P.; Musaelian, A.; Simm, G. N. C.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csanyi, G., The Design Space of E(3)-Equivariant Atom-Centered Interatomic Potentials. *arXiv:2205.06643v1* 2022.
122. Hammer, A. J. S.; Leonov, A. I.; Bell, N. L.; Cronin, L., Chemputation and the Standardization of Chemical Informatics. *JACS Au* 2021, *1* (10), 1572.
123. Perdew, J. P., Burke, K., & Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* 1996, *77* (18), 3865.
124. Lee, S.; Kim, B.; Kim, J., Predicting Performance Limits of Methane Gas Storage in Zeolites with an Artificial Neural Network. *J. Mater. Chem. A* 2019, *7* (6), 2709.
125. Xie, T.; France-Lanord, A.; Wang, Y.; Lopez, J.; Stolberg, M. A.; Hill, M.; Leverick, G. M.; Gomez-Bombarelli, R.; Johnson, J. A.; Shao-Horn, Y.; Grossman, J. C., Accelerating Amorphous

Polymer Electrolyte Screening by Learning to Reduce Errors in Molecular Dynamics Simulated Properties. *arXiv:2101.05339v2* 2022.
126. Gorai, P.; Gao, D.; Ortiz, B.; Miller, S.; Barnett, S. A.; Mason, T.; Lv, Q.; Stevanović, V.; Toberer, E. S., Te Design Lab: A Virtual Laboratory for Thermoelectric Material Design. *Comput. Mater. Sci.* 2016, *112*, 368.
127. Kingsbury, R.; Gupta, A. S.; Bartel, C. J.; Munro, J. M.; Dwaraknath, S.; Horton, M.; Persson, K. A., Performance Comparison of R2scan and Scan Metagga Density Functionals for Solid Materials Via an Automated, High-Throughput Computational Workflow. *Phys. Rev. Mater.* 2022, *6* (1).
128. Kim, S.; Lee, M.; Hong, C.; Yoon, Y.; An, H.; Lee, D.; Jeong, W.; Yoo, D.; Kang, Y.; Youn, Y.; Han, S., A Band-Gap Database for Semiconducting Inorganic Materials Calculated with Hybrid Functional. *Sci Data* 2020, *7* (1), 387.
129. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci Data* 2014, *1*, 140022.
130. Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B., Dataset's Chemical Diversity Limits the Generalizability of Machine Learning Predictions. *J Cheminform* 2019, *11* (1), 69.
131. Muy, S.; Voss, J.; Schlem, R.; Koerver, R.; Sedlmaier, S. J.; Maglia, F.; Lamp, P.; Zeier, W. G.; Shao-Horn, Y., High-Throughput Screening of Solid-State Li-Ion Conductors Using Lattice-Dynamics Descriptors. *iScience* 2019, *16*, 270.
132. Gong, S.; Xie, T.; Shao-Horn, Y.; Gomez-Bombarelli, R.; Grossman, J. C., Examining Graph Neural Networks for Crystal Structures: Limitation on Capturing Periodicity. *arXiv:2208.05039* 2022.
133. Patil, J.; Wan, C. T.-C.; Gong, S.; Chiang, Y.-M.; Brushett, F. R.; Grossman, J. C., Bayesian-Optimization-Assisted Laser-Reduction of Poly(Acrylonitrile) for Electrochemical Applications *in preparation* 2022.
134. Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B., E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat Commun* 2022, *13* (1), 2453.
135. Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A., Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* 2018, *152*, 60.
136. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* 2013, *68*, 314.
137. Grosse-Kunstleve, R. W.; Sauter, N. K.; Adams, P. D., Numerically Stable Algorithms for the Computation of Reduced Unit Cells. *Acta Crystallogr. A* 2004, *60* (Pt 1), 1.
138. WOLFF;, P. M. D.; GRUBER, B., Niggli Lattice Characters. *Acta Cryst.* 1991, *A47*, 29.
139. Sabirov, D. S.; Shepelevich, I. S., Information Entropy in Chemistry: An Overview. *Entropy (Basel)* 2021, *23* (10).
140. Li, Y.; Yang, W.; Dong, R.; Hu, J., Mlatticeabc: Generic Lattice Constant Prediction of Crystal Materials Using Machine Learning. *ACS Omega* 2021, *6* (17), 11585.
141. Liang, H.; Stanev, V.; Kusne, A. G.; Takeuchi, I., Cryspnet: Crystal Structure Predictions Via Neural Networks. *Phys. Rev. Mater.* 2020, *4* (12).
142. Chen, Z.; Villar, S.; Chen, L.; Bruna, J., On the Equivalence between Graph Isomorphism Testing and Function Approximation with Gnns. *33rd Conference on Neural Information Processing Systems* 2019, 15868.
143. Saelim, T.; Chainok, K.; Kielar, F.; Wannarit, N., Crystal Structure of a Novel One-

Dimensional Zigzag Chain-Like Cobalt(Ii) Coordination Polymer Constructed from 4,4'-Bi-Pyridine and 2-Hy-Droxy-Benzoate Ligands. *Acta Crystallogr E Crystallogr Commun* 2020, *76* (Pt 8), 1302.
144. Sarkar, A. S.; Stratakis, E., Recent Advances in 2d Metal Monochalcogenides. *Adv. Sci.* 2020, *7* (21), 2001655.
145. Alon, U.; Yahav, E., On the Bottleneck of Graph Neural Networks and Its Practical Implications. *International Conference on Learning Representations* 2021.
146. Li; , Q.; Han; , Z.; Xiao-MingWu, Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. *AAAI Conference on Artificial Intelligence* 2018, 3538.
147. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S., How Powerful Are Graph Neural Networks? *International Conference on Learning Representations* 2019.
148. Hsu, T.; Pham, T. A.; Keilbart, N.; Weitzner, S.; Chapman, J.; Xiao, P.; Qiu, S. R.; Chen, X.; Wood, B. C., Efficient and Interpretable Graph Neural Network Representation for Angle-Dependent Properties and Its Application to Optical Spectroscopy. *npj Comput Mater* 2022, *8*, 151.
149. Mao, J.; Chen, G.; Ren, Z. J. N. M., Thermoelectric Cooling Materials. *Nat. Mater.* 2021, *20* (4), 454.
150. Omee, S. S.; Louis, S.-Y.; Fu, N.; Lai Wei; Dey, S.; Dong, R.; Li, Q.; Hu, J., Scalable Deeper Graph Neural Networks for High-Performance Materials Property Prediction. *Patterns* 2022, *3* (5), 100491.
151. Bodnar, C.; Frasca, F.; Wang, Y. G.; Otter, N.; Mont´ufar, G.; Li`, P.; Bronstein, M. M., Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks. *38 th International Conference on Machine Learning* 2021, 139.
152. Bodnar, C.; Frasca, F.; Otter, N.; Wang, Y. G.; Liò, P.; Montúfar, G.; Bronstein, M., Weisfeiler and Lehman Go Cellular: Cw Networks. *35th Conference on Neural Information Processing Systems* 2021.
153. Gu, F.; Chang, H.; Zhu, W.; Sojoudi, S.; Ghaoui, L. E., Implicit Graph Neural Networks. *34th Conference on Neural Information Processing Systems* 2020.
154. ZhanghaoWu; Jain, P.; Wright, M. A.; Mirhoseini, A.; Gonzalez, J. E.; Stoica, I., Representing Long-Range Context for Graph Neural Networks with Global Attention. *35th Conference on Neural Information Processing Systems* 2021.
155. <B-Doped-Graphite-Li-Anode.Pdf>.
156. Kittel, C., *Introduction to Solid State Physics Eighth Edition*. John Wiley & Sons.: 2004.
157. Yan, J.; Gorai, P.; Ortiz, B.; Miller, S.; Barnett, S. A.; Mason, T.; Stevanović, V.; Toberer, E. S., Material Descriptors for Predicting Thermoelectric Performance. *Energy Environ. Sci.* 2015, *8* (3), 983.
158. Kohn, W.; Sham, L. J., Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* 1965, *140* (4A), A1133.
159. Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K.-A. N.; Cui, Y.; Reed, E. J., Holistic Computational Structure Screening of More Than 12 000 Candidates for Solid Lithium-Ion Conductor Materials. *Energy Environ. Sci.* 2017, *10* (1), 306.
160. Yang, T.; Zhou, J.; Song, T. T.; Shen, L.; Feng, Y. P.; Yang, M., High-Throughput Identification of Exfoliable Two-Dimensional Materials with Active Basal Planes for Hydrogen Evolution. *ACS Energy Lett.* 2020, *5* (7), 2313.
161. Zhou, J.; Shen, L.; Costa, M. D.; Persson, K. A.; Ong, S. P.; Huck, P.; Lu, Y.; Ma, X.; Chen, Y.; Tang, H.; Feng, Y. P., 2dmatpedia, an Open Computational Database of Two-Dimensional Materials from Top-Down and Bottom-up Approaches. *Sci Data* 2019, *6* (1), 86.
162. Yan, Q.; Yu, J.; Suram, S. K.; Zhou, L.; Shinde, A.; Newhouse, P. F.; Chen, W.; Li, G.;

Persson, K. A.; Gregoire, J. M.; Neaton, J. B., Solar Fuels Photoanode Materials Discovery by Integrating High-Throughput Theory and Experiment. *Proc Natl Acad Sci U S A* 2017, *114* (12), 3040.
163. Zhu, H.; Hautier, G.; Aydemir, U.; Gibbs, Z. M.; Li, G.; Bajaj, S.; Pöhls, J.-H.; Broberg, D.; Chen, W.; Jain, A.; White, M. A.; Asta, M.; Snyder, G. J.; Persson, K.; Ceder, G., Computational and Experimental Investigation of Tmagte2and Xyz2compounds, a New Group of Thermoelectric Materials Identified by First-Principles High-Throughput Screening. *J. Mater. Chem. C* 2015, *3* (40), 10554.
164. Dunstan, M. T.; Jain, A.; Liu, W.; Ong, S. P.; Liu, T.; Lee, J.; Persson, K. A.; Scott, S. A.; Dennis, J. S.; Grey, C. P., Large Scale Computational Screening and Experimental Discovery of Novel Materials for High Temperature Co2 Capture. *Energy Environ. Sci.* 2016, *9* (4), 1346.
165. Li, S.; Xia, Y.; Amachraa, M.; Hung, N. T.; Wang, Z.; Ong, S. P.; Xie, R.-J., Data-Driven Discovery of Full-Visible-Spectrum Phosphor. *Chem. Mater.* 2019, *31* (16), 6286.
166. Cooley, J. A.; Horton, M. K.; Levin, E. E.; Lapidus, S. H.; Persson, K. A.; Seshadri, R., From Waste-Heat Recovery to Refrigeration: Compositional Tuning of Magnetocaloric Mn1+Xsb. *Chem. Mater.* 2020, *32* (3), 1243.
167. Kim, G.; Meschel, S. V.; Nash, P.; Chen, W., Experimental Formation Enthalpies for Intermetallic Phases and Other Inorganic Compounds. *Sci Data* 2017, *4*, 170162.
168. Aykol, M., Dwaraknath, S. S., Sun, W., & Persson, K. A., Thermodynamic Limit for Synthesis of Metastable Inorganic Materials. *Sci. Adv.* 2018, *4* (4), eaaq0148.
169. Sun, W., Dacek, S.T., Ong, S.P., Hautier, G., Jain, A., Richards, W.D., Gamst, A.C., Persson, K.A. and Ceder, G., The Thermodynamic Scale of Inorganic Crystalline Metastability. *Sci. Adv.* 2016, *2* (11), e1600225.
170. Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J., Machine Learning-Assisted Discovery of Solid Li-Ion Conducting Materials. *Chem. Mater.* 2018, *31* (2), 342.
171. Cubuk, E. D.; Sendek, A. D.; Reed, E. J., Screening Billions of Candidates for Solid Lithium-Ion Conductors: A Transfer Learning Approach for Small Data. *J Chem Phys* 2019, *150* (21), 214701.
172. Gong, S.; Wu, W.; Wang, F. Q.; Liu, J.; Zhao, Y.; Shen, Y.; Wang, S.; Sun, Q.; Wang, Q., Classifying Superheavy Elements by Machine Learning. *Phys. Rev. A* 2019, *99* (2).
173. Shi, Z.; Tsymbalov, E.; Dao, M.; Suresh, S.; Shapeev, A.; Li, J., Deep Elastic Strain Engineering of Bandgap through Machine Learning. *Proc Natl Acad Sci U S A* 2019, *116* (10), 4117.
174. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C., Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* 2014, *89* (9).
175. Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K., Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys Rev Lett* 2008, *100* (13), 136406.
176. Sun, J.; Ruzsinszky, A.; Perdew, J. P., Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys Rev Lett* 2015, *115* (3), 036402.
177. Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J., Accurate and Numerically Efficient R(2)Scan Meta-Generalized Gradient Approximation. *J Phys Chem Lett* 2020, *11* (19), 8208.
178. Chevrier, V. L.; Ong, S. P.; Armiento, R.; Chan, M. K. Y.; Ceder, G., Hybrid Density Functional Calculations of Redox Potentials and Formation Energies of Transition Metal Compounds. *Phys. Rev. B* 2010, *82* (7).
179. Sarmiento-Perez, R.; Botti, S.; Marques, M. A., Optimized Exchange and Correlation Semilocal Functional for the Calculation of Energies of Formation. *J Chem Theory Comput* 2015, *11*

(8), 3844.
180. Zhang, Y.; Kitchaev, D. A.; Yang, J.; Chen, T.; Dacek, S. T.; Sarmiento-Pérez, R. A.; Marques, M. A. L.; Peng, H.; Ceder, G.; Perdew, J. P.; Sun, J., Efficient First-Principles Prediction of Solid Stability: Towards Chemical Accuracy. *npj Comput. Mater.* 2018, *4* (1).
181. Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M., Multifidelity Statistical Machine Learning for Molecular Crystal Structure Prediction. *J Phys Chem A* 2020, *124* (39), 8065.
182. Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G., A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci.* 2011, *50* (8), 2295.
183. Persson, K.; Dwaraknath, S.; Ong, S. P.; Jain, A.; Horton, M.; McDermott, M.; Kingsbury, R.; Wang, A., A Framework for Quantifying Uncertainty in Dft Energy Corrections. *Sci. Rep.* 2021, *11*, 15496.
184. Friedrich, R.; Usanmaz, D.; Oses, C.; Supka, A.; Fornari, M.; Buongiorno Nardelli, M.; Toher, C.; Curtarolo, S., Coordination Corrected Ab Initio Formation Enthalpies. *npj Comput. Mater.* 2019, *5* (1).
185. Zhang, Y.; Ling, C., A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* 2018, *4* (1), 1.
186. Xie, T., Bapst, V., Gaunt, A.L., Obika, A., Back, T., Hassabis, D., Kohli, P. and Kirkpatrick, J., Atomistic Graph Networks for Experimental Materials Property Prediction. *arXiv:2103.13795* 2021.
187. Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E., Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat Commun* 2019, *10* (1), 2903.
188. Kong, S.; Guevarra, D.; Gomes, C. P.; Gregoire, J. M., Materials Representation and Transfer Learning for Multi-Property Prediction. *Appl. Phys. Rev.* 2021, *8* (2).
189. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *J Chem Theory Comput* 2015, *11* (5), 2087.
190. Hurtado, I.; Neuschutz, D., Thermodynamic Properties of Inorganic Materials, Compiled by Sgte, Vol. 19. Springer, Berlin (1999–2005): 1999.
191. Liu, T.; Abd-Elrahman, A.; Morton, J.; Wilhelm, V. L., Comparing Fully Convolutional Networks, Random Forest, Support Vector Machine, and Patch-Based Deep Convolutional Neural Networks for Object-Based Wetland Mapping Using Images from Small Unmanned Aircraft System. *Glsci Remote Sens* 2018, *55* (2), 243.
192. Folleco, A., Khoshgoftaar, T.M., Van Hulse, J. and Bullard, L., Identifying Learners Robust to Low Quality Data. *2008 IEEE International Conference on Information Reuse and Integration* 2008, 190.
193. Yu, Y.; Aykol, M.; Wolverton, C., Reaction Thermochemistry of Metal Sulfides with Gga Andgga+Ucalculations. *Phys. Rev. B* 2015, *92* (19).
194. Aykol, M.; Wolverton, C., Local Environment Dependentgga+Umethod for Accurate Thermochemistry of Transition Metal Compounds. *Phys. Rev. B* 2014, *90* (11).
195. Lakshminarayanan, B.; Pritzel, A.; Blundell, C., Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems* 2017, 6402.
196. Nolan, A. M.; Zhu, Y.; He, X.; Bai, Q.; Mo, Y., Computation-Accelerated Design of Materials and Interfaces for All-Solid-State Lithium-Ion Batteries. *Joule* 2018, *2* (10), 2016.
197. Sun, W.; Holder, A.; Orvañanos, B.; Arca, E.; Zakutayev, A.; Lany, S.; Ceder, G., Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides. *Chem. Mater.* 2017, *29* (16),

6936.
198. Jones, H. J. R. o. P. i. P., Splat Cooling and Metastable Phases. *Rep. Prog. Phys.* 1973, *36* (11), 1425.
199. Takeuchi, T.; Kageyama, H.; Nakanishi, K.; Inada, Y.; Katayama, M.; Ohta, T.; Senoh, H.; Sakaebe, H.; Sakai, T.; Tatsumi, K. J. J. o. t. E. S., Improvement of Cycle Capability of Fes2 Positive Electrode by Forming Composites with Li2s for Ambient Temperature Lithium Batteries. *J. Electrochem. Soc.* 2011, *159* (2), A75.
200. Darling, A. J.; Stewart, S.; Holder, C. F.; Schaak, R. E. J. C., Bulk-Immiscible Agrh Alloy Nanoparticles as a Highly Active Electrocatalyst for the Hydrogen Evolution Reaction. *ChemNanoMat* 2020, *6* (9), 1320.
201. Wang, H.; Li, Y.; Li, C.; Deng, K.; Wang, Z.; Xu, Y.; Li, X.; Xue, H.; Wang, L., One-Pot Synthesis of Bi-Metallic Pdru Tripods as an Efficient Catalyst for Electrocatalytic Nitrogen Reduction to Ammonia. *J. Mater. Chem. A* 2019, *7* (2), 801.
202. Wu, K.-L.; Yu, R.; Wei, X.-W. J. C., Monodispersed Feni 2 Alloy Nanostructures: Solvothermal Synthesis, Magnetic Properties and Size-Dependent Catalytic Activity. *CrystEngComm* 2012, *14* (22), 7626.
203. Lee, C.-C.; Cheng, Y.-Y.; Chang, H. Y.; Chen, D.-H. J. J. o. A., Synthesis and Electromagnetic Wave Absorption Property of Ni–Ag Alloy Nanoparticles. *J. Alloys Compd.* 2009, *480* (2), 674.
204. Asano, M.; Umeda, K.; Tasaki, A. J. J. J. o. A. P., Cu3n Thin Film for a New Light Recording Media. *Jpn. J. Appl. Phys.* 1990, *29* (10R), 1985.
205. Ono, S.; El Ouenzerfi, R.; Quema, A.; Murakami, H.; Sarukura, N.; Nishimatsu, T.; Terakubo, N.; Mizuseki, H.; Kawazoe, Y.; Yoshikawa, A. J. J. j. o. a. p., Band-Structure Design of Fluoride Complex Materials for Deep-Ultraviolet Light-Emitting Diodes. *Jpn. J. Appl. Phys.* 2005, *44* (10R), 7285.
206. Li, Z.; Wang, W.; Zhou, P.; Ma, Z.; Sun, L. J. N. J. o. P., New Type of Hybrid Nodal Line Semimetal in Be2si. *New J. Phys.* 2019, *21* (3), 033018.
207. Xu, Y.-Q.; Liu, B.-G.; Pettifor, D. J. P. R. B., Half-Metallic Ferromagnetism of Mnbi in the Zinc-Blende Structure. *Phys. Rev. B* 2002, *66* (18), 184435.
208. Gjoka, M.; Panagiotopoulos, I.; Niarchos, D. J. J. o. m. p. t., Structure and Magnetic Properties of Sm (Co1− Xmx) 5 (M= Cu, Ag) Alloys. *J. Mater. Process. Technol.* 2005, *161* (1-2), 173.
209. McGarvey, B. R.; Reuveni, A., 19 F Nmr Studies of Rare Earth Elpasolites. In *Magnetic Resonance and Related Phenomena*, Springer: 1979; pp 121.
210. Sugiyama, K.; Iizuka, T.; Aoki, D.; Tokiwa, Y.; Miyake, K.; Watanabe, N.; Kindo, K.; Inoue, T.; Yamamoto, E.; Haga, Y. J. J. o. t. P. S. o. J., High-Field Magnetization of Usn 3 and Upb 3. *J. Phys. Soc. Japan* 2002, *71* (1), 326.
211. Balamurugan, B.; Das, B.; Zhang, W.; Skomski, R.; Sellmyer, D. J. J. J. o. P. C. M., Hf–Co and Zr–Co Alloys for Rare-Earth-Free Permanent Magnets. *J. Phys. Condens. Matter* 2014, *26* (6), 064204.
212. Yannello, V. J.; Lu, E.; Fredrickson, D. C. J. I. C., At the Limits of Isolobal Bonding: Π-Based Covalent Magnetism in Mn2hg5. *Inorg. Chem.* 2020, *59* (17), 12304.
213. Kammerdiner, L. W. Film Deposition of Nb-Based A15 Superconductors. California Univ. San Diego, 1975.
214. Volodin, V.; Zhakanbaev, E.; Tuleushev, A. Z.; Tuleushev, Y. J. V. N. n. Y. T. R. K., Synthesis and Structure of New Intermetallic Compound Ta 3 Pb. *Vestnik Natsional'nogo Yadernogo Tsentra Respubliki Kazakhstan* 2005, *4* (24), 49.
215. Masumoto, H.; Watanabe, K. J. J. o. t. P. S. o. J., New Compounds of the Clb, Cl Types of

Rhmnsb, Irmnsn and Irmnal, New L21 (Heusler) Type of Ir2mnal and Rh2mnal Alloys, and Magnetic Properties. *J. Phys. Soc. Jpn.* **1972,** *32* (1), 281.
216. Yin, M.; Nash, P. J. T. J. o. C. T., Standard Enthalpies of Formation of Selected Xyz Half-Heusler Compounds. *J. Chem. Thermodyn.* **2015,** *91*, 1.
217. Aljarrah, M.; Obeidat, S.; Fouad, R. H.; Rababah, M.; Almagableh, A.; Itradat, A., Thermodynamic Calculations of the Mn–Sn, Mn–Sr and Mg–Mn–{Sn, Sr} Systems. *IET Sci. Meas.* **2015,** *9* (6), 681.
218. Xu, J.; Wang, D.; Liu, Y.; Lian, R.; Gao, X.; Chen, G.; Wei, Y. J. J. o. M. C. A., Theoretical Prediction and Atomic-Scale Investigation of a Tetra-Vn 2 Monolayer as a High Energy Alkali Ion Storage Material for Rechargeable Batteries. *J. Mater. Chem. A* **2019,** *7* (47), 26858.
219. Zhou, Y.; Sun, W.; Chu, W.; Zheng, J.; Gao, X.; Zhou, X.; Xue, Y. J. A. S. S., Adsorption of Acetylene on Ordered Nixag1-X/Ni (111) and Effect of Ag-Dopant: A Dft Study. *Appl. Surf. Sci.* **2018,** *435*, 521.
220. Dudenkov, I.; Solntsev, K. J. R. j. o. i. c., Theoretical Prediction of the New High-Density Lithium Boride Lib 11 with Polymorphism and Pseudoplasticity. *Russ. J. Inorg. Chem.* **2009,** *54* (8), 1261.
221. Hou, W.; Liu, J.; Zuo, X.; Xu, J.; Zhang, X.; Liu, D.; Zhao, M.; Zhu, Z.-G.; Luo, H.-G.; Zhao, W., Prediction of Crossing Nodal-Lines and Large Intrinsic Spin Hall Conductivity in Topological Dirac Semimetal Ta3as Family. *npj Comput. Mater.* **2021,** *7* (1).
222. Guan-Nan, L.; Ying-Jiu, J. J. C. P. L., First-Principles Study on the Half-Metallicity of Half-Heusler Alloys: Xyz (X= Mn, Ni; Y= Cr, Mn; Z= as, Sb). *Chin. Phys. Lett.* **2009,** *26* (10), 107101.
223. Manaa, H.; Moncorgé, R.; Butashin, A. V.; Mill, B.; Kaminskii, A. A. In *Luminescence Properties of Cr-Doped Linbgeo5 Laser Crystal*, ASSL, New Orleans, Louisiana, 1993/02/01; Pinto, A.; Fan, T., Eds. Optical Society of America: New Orleans, Louisiana, 1993; p TL13.
224. Li, Y.; Dai, X.-F.; Liu, G.-D.; Wei, Z.-Y.; Liu, E.-K.; Han, X.-L.; Du, Z.-W.; Xi, X.-K.; Wang, W.-H.; Wu, G.-H. J. C. P. B., Structural, Magnetic Properties, and Electronic Structure of Hexagonal Fecosn Compound. *Chinese Phys. B* **2018,** *27* (2), 026101.
225. Gerrits, N.; Smeets, E. W. F.; Vuckovic, S.; Powell, A. D.; Doblhoff-Dier, K.; Kroes, G. J., Density Functional Theory for Molecule-Metal Surface Reactions: When Does the Generalized Gradient Approximation Get It Right, and What to Do If It Does Not. *J Phys Chem Lett* **2020,** *11* (24), 10552.
226. Seo, D.-H.; Urban, A.; Ceder, G., Calibrating Transition-Metal Energy Levels and Oxygen Bands in First-Principles Calculations: Accurate Prediction of Redox Potentials and Charge Transfer in Lithium Transition-Metal Oxides. *Phys. Rev. B* **2015,** *92* (11).
227. Grindy, S.; Meredig, B.; Kirklin, S.; Saal, J. E.; Wolverton, C., Approaching Chemical Accuracy with Density Functional Calculations: Diatomic Energy Corrections. *Phys. Rev. B* **2013,** *87* (7).
228. He, J.; Tritt, T. M., Advances in Thermoelectric Materials Research: Looking Back and Moving Forward. *Science* **2017,** *357* (6358), eaak9997.
229. Gorai, P.; Stevanović, V.; Toberer, E. S. J. N. R. M., Computationally Guided Discovery of Thermoelectric Materials. *Nat. Rev. Mater.* **2017,** *2* (9), 1.
230. DiSalvo, F. J. J. S., Thermoelectric Cooling and Power Generation. *Science* **1999,** *285* (5428), 703.
231. Snyder, G. J.; Ursell, T. S. J. P. r. l., Thermoelectric Efficiency and Compatibility. *Phys. Rev. Lett.* **2003,** *91* (14), 148301.
232. Wei, L.; Kuo, P.; Thomas, R.; Anthony, T.; Banholzer, W. J. P. r. l., Thermal Conductivity of Isotopically Modified Single Crystal Diamond. *Phys. Rev. Lett.* **1993,** *70* (24), 3764.

233. Lu, X.; Arduini-Schuster, M.; Kuhn, J.; Nilsson, O.; Fricke, J.; Pekala, R. J. S., Thermal Conductivity of Monolithic Organic Aerogels. *Science* 1992, *255* (5047), 971.
234. Clarke, D. R.; Phillpot, S. R. J. M. t., Thermal Barrier Coating Materials. *Mater. Today* 2005, *8* (6), 22.
235. Zhu, T.; Ertekin, E. J. E.; Science, E., Mixed Phononic and Non-Phononic Transport in Hybrid Lead Halide Perovskites: Glass-Crystal Duality, Dynamical Disorder, and Anharmonicity. *Energy Environ. Sci.* 2019, *12* (1), 216.
236. Schelling, P. K.; Shi, L.; Goodson, K. E. J. M. T., Managing Heat for Electronics. *Mater. Today* 2005, *8* (6), 30.
237. Peng, Y.; Cui, Y. J. J., Advanced Textiles for Personal Thermal Management and Energy. *Joule* 2020, *4* (4), 724.
238. Toberer, E. S.; Zevalkink, A.; Snyder, G. J. J. J. o. M. C., Phonon Engineering through Crystal Chemistry. *J. Mater. Chem.* 2011, *21* (40), 15843.
239. Seko, A.; Togo, A.; Hayashi, H.; Tsuda, K.; Chaput, L.; Tanaka, I. J. P. r. l., Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization. *Phys. Rev. Lett.* 2015, *115* (20), 205901.
240. Lindsay, L.; Broido, D.; Reinecke, T. J. P. r. l., First-Principles Determination of Ultrahigh Thermal Conductivity of Boron Arsenide: A Competitor for Diamond? *Phys. Rev. Lett.* 2013, *111* (2), 025901.
241. Morelli, D. T.; Slack, G. A., High Lattice Thermal Conductivity Solids. In *High Thermal Conductivity Materials*, Springer: 2006; pp 37.
242. Gurunathan, R.; Hanus, R.; Snyder, G. J. J. M. H., Alloy Scattering of Phonons. *Mater. Horiz.* 2020, *7* (6), 1452.
243. Miller, S. A.; Gorai, P.; Ortiz, B. R.; Goyal, A.; Gao, D.; Barnett, S. A.; Mason, T. O.; Snyder, G. J.; Lv, Q.; Stevanovic´, V. J. C. o. M., Capturing Anharmonicity in a Lattice Thermal Conductivity Model for High-Throughput Predictions. *Chem. Mater.* 2017, *29* (6), 2494.
244. Toher, C.; Plata, J. J.; Levy, O.; De Jong, M.; Asta, M.; Nardelli, M. B.; Curtarolo, S. J. P. R. B., High-Throughput Computational Screening of Thermal Conductivity, Debye Temperature, and Grüneisen Parameter Using a Quasiharmonic Debye Model. *Phys. Rev. B* 2014, *90* (17), 174107.
245. Carrete, J.; Li, W.; Mingo, N.; Wang, S.; Curtarolo, S. J. P. R. X., Finding Unprecedentedly Low-Thermal-Conductivity Half-Heusler Semiconductors Via High-Throughput Materials Modeling. *Phys. Rev. X* 2014, *4* (1), 011019.
246. van Roekeghem, A.; Carrete, J.; Oses, C.; Curtarolo, S.; Mingo, N. J. P. R. X., High-Throughput Computation of Thermal Conductivity of High-Temperature Solid Phases: The Case of Oxide and Fluoride Perovskites. *Phys. Rev. X* 2016, *6* (4), 041061.
247. Wang, S.; Wang, Z.; Setyawan, W.; Mingo, N.; Curtarolo, S. J. P. R. X., Assessing the Thermoelectric Properties of Sintered Compounds Via High-Throughput Ab-Initio Calculations. *Phys. Rev. X* 2011, *1* (2), 021012.
248. McGaughey, A. J.; Kaviany, M. J. A. i. h. t., Phonon Transport in Molecular Dynamics Simulations: Formulation and Thermal Conductivity Prediction. *Adv. Heat Transfer* 2006, *39*, 169.
249. Schelling, P. K.; Phillpot, S. R.; Keblinski, P. J. P. R. B., Comparison of Atomic-Level Simulation Methods for Computing Thermal Conductivity. *Phys. Rev. B* 2002, *65* (14), 144306.
250. McGaughey, A. J.; Jain, A.; Kim, H.-Y.; Fu, B. J. J. o. A. P., Phonon Properties and Thermal Conductivity from First Principles, Lattice Dynamics, and the Boltzmann Transport Equation. *J. Appl. Phys.* 2019, *125* (1), 011101.
251. Gaultois, M. W.; Sparks, T. D.; Borg, C. K.; Seshadri, R.; Bonificio, W. D.; Clarke, D. R. J. C.

o. M., Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations. *Chem. Mater.* 2013, *25* (15), 2911.
252. Wei, H.; Zhao, S.; Rong, Q.; Bao, H. J. I. J. o. H.; Transfer, M., Predicting the Effective Thermal Conductivities of Composite Materials and Porous Media by Machine Learning Methods. *Int. J. Heat Mass Transfer* 2018, *127*, 908.
253. Lee, W.; Li, H.; Wong, A. B.; Zhang, D.; Lai, M.; Yu, Y.; Kong, Q.; Lin, E.; Urban, J. J.; Grossman, J. C. J. P. o. t. N. A. o. S., Ultralow Thermal Conductivity in All-Inorganic Halide Perovskites. *Proc Natl Acad Sci U S A* 2017, *114* (33), 8693.
254. Mukhopadhyay, S.; Parker, D. S.; Sales, B. C.; Puretzky, A. A.; McGuire, M. A.; Lindsay, L. J. S., Two-Channel Model for Ultralow Thermal Conductivity of Crystalline Tl3vse4. *Science* 2018, *360* (6396), 1455.
255. Kumashiro, Y.; Mitsuhashi, T.; Okaya, S.; Muta, F.; Koshiro, T.; Takahashi, Y.; Mirabayashi, M. J. J. o. a. p., Thermal Conductivity of a Boron Phosphide Single-Crystal Wafer up to High Temperature. *J. Appl. Phys.* 1989, *65* (5), 2147.
256. Li, S.; Zheng, Q.; Lv, Y.; Liu, X.; Wang, X.; Huang, P. Y.; Cahill, D. G.; Lv, B. J. S., High Thermal Conductivity in Cubic Boron Arsenide Crystals. *Science* 2018, *361* (6402), 579.
257. Xia, Y.; Pal, K.; He, J.; Ozoliņš, V.; Wolverton, C. J. P. r. l., Particlelike Phonon Propagation Dominates Ultralow Lattice Thermal Conductivity in Crystalline Tl 3 Vse 4. *Phys. Rev. X* 2020, *124* (6), 065901.
258. Legrain, F.; Carrete, J.; van Roekeghem, A.; Curtarolo, S.; Mingo, N., How Chemical Composition Alone Can Predict Vibrational Free Energies and Entropies of Solids. *Chem. Mater.* 2017, *29* (15), 6220.
259. Brown, S. R.; Kauzlarich, S. M.; Gascoin, F.; Snyder, G. J. J. C. o. m., Yb14mnsb11: New High Efficiency Thermoelectric Material for Power Generation. *Chem. Mater.* 2006, *18* (7), 1873.
260. Rowe, D. M., *Crc Handbook of Thermoelectrics*. CRC press: 2018.
261. Chen, S.; Lukas, K. C.; Liu, W.; Opeil, C. P.; Chen, G.; Ren, Z. J. A. E. M., Effect of Hf Concentration on Thermoelectric Properties of Nanostructured N-Type Half-Heusler Materials Hfxzr1–Xnisn0. 99sb0. 01. *Adv. Energy Mater.* 2013, *3* (9), 1210.
262. Yang, D.; Su, X.; Li, J.; Bai, H.; Wang, S.; Li, Z.; Tang, H.; Tang, K.; Luo, T.; Yan, Y. J. A. M., Blocking Ion Migration Stabilizes the High Thermoelectric Performance in Cu2se Composites. *Adv Mater* 2020, *32* (40), 2003730.
263. Chang, C.; Wu, M.; He, D.; Pei, Y.; Wu, C.-F.; Wu, X.; Yu, H.; Zhu, F.; Wang, K.; Chen, Y. J. S., 3d Charge and 2d Phonon Transports Leading to High out-of-Plane Zt in N-Type Snse Crystals. *Science* 2018, *360* (6390), 778.
264. Togo, A.; Chaput, L.; Tanaka, I. J. P. R. B., Distributions of Phonon Lifetimes in Brillouin Zones. *Phys. Rev. B* 2015, *91* (9), 094306.
265. Kresse, G.; Furthmu¨ller, J., Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* 1996, *54* (16), 11169.
266. Perdew, J. P.; Burke, K.; Ernzerhof, M. J. P. r. l., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* 1996, *77* (18), 3865.
267. Nguyen, C. V.; Hassner, T.; Seeger, M.; Archambeau, C., Leep: A New Measure to Evaluate Transferability of Learned Representations. *arXiv:2002.12462v2*.
268. You, K.; Liu, Y.; Wang, J.; Long, M., Practical Assessment of Pre-Trained Models for Transfer Learning. *38 th International Conference on Machine Learning* 2021, *139*.
269. Besenhard, J. O.; Schöllhorn, R., Chromium Oxides as Cathodes for Secondary High Energy Density Lithium Batteries. *J. Electrochem. Soc.* 1977, *124*, 968.

270. Melot, B. C.; Tarascon, J.-M., Design and Preparation of Materials for Advanced Electrochemical Storage. *Acc. Chem. Res.* 2013, *46* (5), 1226.
271. Bhardwaj, T.; Antic, A.; Pavan, B.; Barone, V.; Fahlman, B. D., Enhanced Electrochemical Lithium Storage by Graphene Nanoribbons. *J. Am. Chem. Soc.* 2010, *132* (36), 12556.
272. Fitzer, E.; Frohs, W.; Heine, M., Optimization of Stabilization and Carbonization Treatment of Pan Fibres and Structural Characterization of the Resulting Carbon Fibres. *Carbon* 1986, *24* (4), 387.
273. Jensen, W. B., Electronegativity from Avogadro to Pauling_ Ii. Late Nineteenth- and Early Twentieth-Century Developments. *J. Chem. Educ.* 2003, *80* (3), 279.
274. Wang, R.; Li, Y., Twin-Cocoon-Derived Self-Standing Nitrogen-Oxygen-Rich Monolithic Carbon Material as the Cost-Effective Electrode for Redox Flow Batteries. *J. Power Source* 2019, *421*, 139.
275. Wan, C. T. C.; López Barreiro, D.; Forner-Cuenca, A.; Barotta, J. W.; Hawker, M. J.; Han, G.; Loh, H. C.; Masic, A.; Kaplan, D. L.; Chiang, Y. M.; Brushett, F. R.; Martin-Martinez, F. J.; Buehler, M. J., Exploration of Biomass-Derived Activated Carbons for Use in Vanadium Redox Flow Batteries. *ACS Sustain. Chem. Eng.* 2020, *8* (25), 9472.
276. Wan, C. T. C.; Jacquemond, R. R.; Chiang, Y. M.; Nijmeijer, K.; Brushett, F. R.; Forner-Cuenca, A., Non-Solvent Induced Phase Separation Enables Designer Redox Flow Battery Electrodes. *Adv Mater* 2021, *33* (16).
277. Rahaman, M. S. A.; Ismail, A. F.; Mustafa, A., A Review of Heat Treatment on Polyacrylonitrile Fiber. In *Polym. Degrad. Stab.*, 2007; Vol. 92, pp 1421.
278. Xu, C.; Yang, X.; Li, X.; Liu, T.; Zhang, H., Ultrathin Free-Standing Electrospun Carbon Nanofibers Web as the Electrode of the Vanadium Flow Batteries. *J. Energy Chem.* 2017, *26* (4), 730.
279. Chyan, Y.; Ye, R.; Li, Y.; Singh, S. P.; Arnusch, C. J.; Tour, J. M., Laser-Induced Graphene by Multiple Lasing: Toward Electronics on Cloth, Paper, and Food. *ACS Nano* 2018, *12* (3), 2176.
280. Lin, J.; Peng, Z.; Liu, Y.; Ruiz-Zepeda, F.; Ye, R.; Samuel, E. L. G.; Yacaman, M. J.; Yakobson, B. I.; Tour, J. M., Laser-Induced Porous Graphene Films from Commercial Polymers. *Nat. Commun.* 2014, *5* (1), 1.
281. Zang, X.; Jian, C.; Ingersoll, S.; Li, H.; Adams, J. J.; Lu, Z.; Ferralis, N.; Grossman, J. C., Laser-Engineered Heavy Hydrocarbons: Old Materials with New Opportunities. *Sci. Adv.* 2020, *6* (17).
282. Bergsman, D. S.; Getachew, B. A.; Cooper, C. B.; Grossman, J. C., Preserving Nanoscale Features in Polymers During Laser Induced Graphene Formation Using Sequential Infiltration Synthesis. *Nat. Commun.* 2020, *11* (1), 1.
283. Turner, R.; Eriksson, D.; McCourt, M.; Kiili, J.; Xu, V. Z.; Escalante, H. J.; Hofmann, K. In *Bayesian Optimization Is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020*, 34th Conference on Neural Information Processing Systems, 2021; pp 3.
284. Ueno, T.; Rhone, T. D.; Hou, Z.; Mizoguchi, T.; Tsuda, K., Combo: An Efficient Bayesian Optimization Library for Materials Science. *Mater. Discov.* 2016, *4*, 18.
285. Rohr, B.; Stein, H. S.; Guevarra, D.; Wang, Y.; Haber, J. A.; Aykol, M.; Suram, S. K.; Gregoire, J. M., Benchmarking the Acceleration of Materials Discovery by Sequential Learning. *Chem. Sci.* 2020, *11* (10), 2696.
286. Bessa, M. A.; Glowacki, P.; Houlder, M., Bayesian Machine Learning in Metamaterial Design: Fragile Becomes Supercompressible. *Adv. Mater.* 2019, *31* (48).
287. Li, C.; Rubín De Celis Leal, D.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.; Height,

M.; Venkatesh, S., Rapid Bayesian Optimisation for Synthesis of Short Polymer Fiber Materials. *Sci. Rep.* 2017, *7* (1).
288. Straub, A. P.; Bergsman, D. S.; Getachew, B. A.; Leahy, L. M.; Patil, J. J.; Ferralis, N.; Grossman, J. C., Highly Conductive and Permeable Nanocomposite Ultrafiltration Membranes Using Laser-Reduced Graphene Oxide. *Nano Lett.* 2021, *21* (6), 2429.
289. Tiliakos, A.; Ceaus, C.; Iordache, S. M.; Vasile, E.; Stamatin, I., Morphic Transitions of Nanocarbons Via Laser Pyrolysis of Polyimide Films. *J. Anal. Appl. Pyrolysis* 2016, *121*, 275.
290. Murray, R.; Burke, M.; Iacopino, D.; Quinn, A. J., Design of Experiments and Optimization of Laser-Induced Graphene. *ACS Omega* 2021, *6* (26), 16736.
291. Calandra, R.; Peters, J.; Rasmussen, C. E.; Deisenroth, M. P., Manifold Gaussian Processes for Regression. *2016 International Joint Conference on Neural Networks* 2016.
292. Pang, G.; Yang, L.; Karniadakis, G. E., Neural-Net-Induced Gaussian Process Regression for Function Approximation and Pde Solution. *J. Comput. Phys.* 2019, *384*, 270.
293. Bruinsma, W. P.; Tegnér, M.; Turner, R. E., Modelling Non-Smooth Signals with Complex Spectral Structure. *25th International Conference on Artificial Intelligence and Statistics* 2022.
294. McCreery, R. L., Advanced Carbon Electrode Materials for Molecular Electrochemistry. *Chem. Rev.* 2008, *108* (7), 2646.
295. Streeter, I.; Wildgoose, G. G.; Shao, L.; Compton, R. G., Cyclic Voltammetry on Electrode Surfaces Covered with Porous Layers: An Analysis of Electron Transfer Kinetics at Single-Walled Carbon Nanotube Modified Electrodes. *Sens. Actuators B Chem.* 2008, *133* (2), 462.
296. Punckt, C.; Pope, M. A.; Aksay, I. A., On the Electrochemical Response of Porous Functionalized Graphene Electrodes. *J. Phys. Chem. C.* 2013, *117* (31), 16076.
297. Friedl, J.; Stimming, U., Determining Electron Transfer Kinetics at Porous Electrodes. *Electrochim. Acta* 2017, *227*, 235.
298. Alkire, R. C.; Bartlett, P. N.; Lipkowski, J., *Advances in Electrochemical Science and Engineering*. wiley: 2016; Vol. 16, p 1.
299. Ambrosi, A.; Chua, C. K.; Bonanni, A.; Pumera, M., Electrochemistry of Graphene and Related Materials. *Chem. Rev.* 2014, *114* (14), 7150.
300. Velický, M.; Toth, P. S.; Woods, C. R.; Novoselov, K. S.; Dryfe, R. A. W., Electrochemistry of the Basal Plane Versus Edge Plane of Graphite Revisited. *J. Phys. Chem. C.* 2019, *123* (18), 11677.
301. van der Pauw, L. J., A Method of Measuring the Resistivity and Hall Coefficient on Lamellae of Arbitrary Shape. *Philips Techn. Rev.* 1958, *20*, 220.
302. Hohenberg, P.; Kohn, W., Inhomogeneous Electron Gas. *Phys. Rev.* 1964, *136* (3B), B864.
303. Choi, H.; Sohn, K.-S.; Pyo, M.; Chung, K.-C.; Park, H., Predicting the Electrochemical Properties of Lithium-Ion Battery Electrode Materials with the Quantum Neural Network Algorithm. *J. Phys. Chem. C.* 2019, *123* (8), 4682.
304. Kolb, B.; Lentz, L. C.; Kolpak, A. M., Discovering Charge Density Functionals and Structure-Property Relationships with Prophet: A General Framework for Coupling Machine Learning and First-Principles Methods. *Sci Rep* 2017, *7* (1), 1192.
305. Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R., Accelerating Materials Property Predictions Using Machine Learning. *Sci Rep* 2013, *3*, 2810.
306. Schmidt, E.; Fowler, A. T.; Elliott, J. A.; Bristowe, P. D., Learning Models for Electron Densities with Bayesian Regression. *Comput. Mater. Sci.* 2018, *149*, 250.
307. Pichon-Pesme, V.; Lecomte, C.; Lachekar, H., On Building a Data Bank of Transferable Experimental Electron Density Parameters Applicable to Polypeptides. *J. Phys. Chem.* 1995, *99* (16), 6242.

308. Grimme, S., Semiempirical Gga-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006,** *27* (15), 1787.
309. Daw, M. S.; Baskes, M. I., Embedded-Atom Method: Derivation and Application to Impurities, Surfaces, and Other Defects in Metals. *Phys. Rev. B* **1984,** *29* (12), 6443.