MACHINE LEARNING AND COMPUTATION: EXPLORING STRUCTURE-PROPERTY CORRELATIONS IN INORGANIC CRYSTALLINE MATERIALS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

by
Huta Banjade
August 2020

Examiner Committee Members:

Qimin Yan, Advisory Committee Chair, Dept. of Physics

John P Perdew, Examining Committee Chair, Dept. of Physics

Adrienn Ruzsinszky, Committee member, Dept. of Physics

Vincenzo Carnevale, External member of Examining Committee, Institute for Computational Molecular Science
ABSTRACT

by

Huta R. Banjade

Kohn-Sham Density Functional Theory (DFT) has been the most successful tool to probe the electronic structure, mainly the ground-state total energies and densities of many condensed matter systems has led to the development of various databases such as Materials Project (MP), Inorganic Crystal Structure Database (ICSD), and many others. These databases ignited the interest of the material science community towards Machine Learning (ML), leading to the development of a new sub-field in material science called material-informatics, which aims to uncover the interrelation between known features and material properties. ML techniques can handle and identify relationships in complex and arbitrarily high-dimensional spaces data, which are almost impossible for human reasoning. Unlike DFT, the ML approach uses data from past computations or experiments. In many cases, ML models have shown their superiority over DFT in terms of accuracy and efficiency in predicting various physical and chemical properties of materials.

The incorporation of material property data obtained from atomistic simulations is crucial important to make continuous progress in data-driven methods. In this direction, we use DFT with Perdew-Burke-Ernzerhof (PBE), and Heyd–Scuseria–Ernzerhof (HSE) functionals, to introduce a family of mono-layer isostructural semiconducting tellurides
$\text{M}_2\text{N}_2\text{Te}_8$, with $\text{M} = \{\text{Ti}, \text{Zr}, \text{Hf}\}$ and $\text{N} = \{\text{Si}, \text{Ge}\}$. These compounds have been identified to possess direct band gaps that are tunable from 1.0 eV to 1.3 eV, which are well suited for photonics and optoelectronics applications. Additionally, in-plane transport behavior is observed, and small electron and hole (0.11-0.15 m) masses are identified along the dominant transport direction. High carrier mobility is found in these compounds, which shows great promise for applications in high-speed electronic devices. Detailed analysis of electronic structures reveals the presence of metal center bicapped trigonal prism as the structural building blocks in these compounds; a common feature in most of the group V chalcogenides helps to understand the atomic origins of promising properties of this unique class of 2D telluride materials.

Atomistic simulations based on DFT theory played a vital role in the development of data-driven materials discovery process. However, the resource-based constraints have limited the high-throughput discovery process by using DFT. The main motivation of our work towards the application of machine learning in material science is to assist the discovery process using available material property data in various databases. Incorporation of physical principles in a network-based machine learning (ML) architecture is a fundamental step toward the continued development of artificial intelligence for materials science and condensed matter physics. In this work, as inspired by the Pauling’s rule, we propose that structure motifs (polyhedral formed by cations and surrounding anions) in inorganic crystals can serve as a central input to a machine learning framework for crystalline inorganic materials. We demonstrated that an unsupervised learning algorithm Motif2Vec is able to convert the presence of structural motifs and their connections in a
large set of crystalline compounds into unique vectors. The connections among complex materials can be largely determined by the presence of different structural motifs, and their clustering information is identified by our Motif2Vec algorithm. To demonstrate the novel use of structure motif information, we show that a motif-centric learning framework can be effectively created by combining motif information with the recently developed atom-based graph neural networks to form an atom-motif hybrid graph network (AMDNet). Taking advantage of node and edge information on both atomic and motif level, the AMDNet is more accurate than a single graph network in predicting electronic structure related material properties such as band gaps. The work illustrates the route toward the fundamental design of graph neural network learning architecture for complex materials properties by incorporating beyond-atom physical principles.

Due to the limitations in resources, it is not feasible to synthesize hundreds of thousands of materials listed in various databases by experiment or compute their detailed properties by using various electronic structure codes and state-of-the-art computational tools. Hence, the identification of an alternative route to screen such databases is very desirable. If identified, this route would be very helpful in reducing the material search space for any application. Categorizing materials based on their structural building blocks is very important to study the underlying physics and to understand the possible mechanisms for any application. Based on structure motifs, we purpose a novel way to categorize, analyze, and visualize the material space called a material network. The connection between any two nodes in this network is determined by using the calculated similarity value (Tanimoto-coefficient) between each motif and its surrounding information, encoded in terms of a
feature vector of length 64. By mapping a known compound, the network thus constructed can be used to screen compounds for the desired application. All the connections of the mapped compound are identified and extracted as a subgraph for further analysis. In our test screening for the transparent conducting oxides (TCO), the proposed network is successful in identifying compounds that are already listed as TCO in the literature. Thus, this indicates its usefulness in reducing the search space for the new TCO materials and various applications. This motif-based material network can serve as an alternate route for functional material discovery and design.
Dedicated to my late Father and late Brother
ACKNOWLEDGMENTS

I would like to express my sincerest gratitude and warm appreciation to my graduate supervisor Dr. Qimin Yan, for his expertise, guidance, and support during my graduate study. His never-ending suggestions in research methodology helped me a lot to restructure my style of thinking, writing, and speaking. Without his guidance and continuous support, this work would not be a reality. I am grateful for his inspiration, not limited to academic work, but in every aspect of my life. I want to thank my former group members Jinbo Pan and Yanfang Zhang, who have contributed much to helping me shape and reshape my thoughts from the day I joined the group. I would like to thank Jinbo for his valuable critiques, and most especially, inspiring me during my Ph.D. studies.

I am very indebted to my collaborators Dr. Shanshan Zhang, Sandro Hauri, and Prof. Slobodan Vucetic, at the Department of Computer Science at Temple, on the subject of machine learning in material science. The collaboration resulted in the findings of the “Structure motif centric learning framework of the inorganic crystalline materials” section in this work. Their insights and expertise in the subject matter helped me a lot to understand machine learning and its application in physics.

I am very indebted to my parents for bringing me up with love and encouragement to this stage. My father and brother always wanted to see me wearing the graduation cap; unfortunately, both of you are not with me now. I cannot express in words how much you suffered to make me reach this stage. I will always be indebted to this. I want to thank my
Mother Dewaki, Wife Muna, Brothers Surendra and Govinda, and all other family members for all the help and support during my graduate studies. Besides, I want to thank a year and half old Son Aabhas to ignite the feeling of parenthood and made me realize what a parent wants with their children. I hope one day he will read this work and feel proud of his dad. I am obliged to take this opportunity to sincerely thank Dr. Chandra Shahi, who has always been guiding me to work on the right path from my undergraduate studies. I want to express my special thanks to my wonderful friends Dr. Narendra Acharaya, Dr. Ambar Rangu Magar, Mr. Mukesh Pandey, Mr. Lakesh Kansakar, Mr. Santosh Adhikari, Mr. Puskar Bhattrai, Mr. Niraj Nepal, Mr. Kamal Wagle, Mr. Bimal Neupane, Mr. Santosh Gaire, Mr. Yubaraj Gaire, Mrs. Manita Rai, and all other friends at Temple. They have always been with me during my hard times and cherished moments.

As I conclude, though words are not enough to express my gratitude, I am very thankful to all my teachers, family members, group members, and friends who have always been helpful and encouraging throughout my studies.
# TABLE OF CONTENTS

**ABSTRACT** .......................................................................................................................iii

**DEDICATION** ...................................................................................................................vii

**ACKNOWLEDGMENTS** .....................................................................................................viii

**LIST OF FIGURES** ............................................................................................................xiii

**LIST OF TABLES** ..............................................................................................................xix

**CHAPTER 1 DENSITY FUNCTIONAL THEORY AND MACHINE LEARNING: AN OVERVIEW** .........................................................................................................................1

1.1 Material discovery and design: past and present ...............................................................1

1.2 Density functional theory .................................................................................................4

1.2.1 Classification of the exchange-correlation energy ................................................ ......8

1.3 Machine learning .............................................................................................................13

1.3.1 A Brief introduction .................................................................................................. 13

1.3.2 Artificial neural networks (ANNs) .............................................................................16

1.3.3 Machine learning in material science .......................................................................20

1.4 Discovery and design of two-dimensional (2D) materials .............................................24

1.5 Complex oxides ..............................................................................................................26
CHAPTER 2 MONOLAYER 2D SEMICONDUCTING TELLURIDES FOR HIGH-MOBILITY ELECTRONICS

2.1 Introduction ..................................................................................................................28

2.2 Computational details ..................................................................................................33

2.3 Computed electronic properties of semiconducting tellurides ..................................34

2.4 Stability of 2D telluride material ..................................................................................46

2.5 Summary .......................................................................................................................47

CHAPTER 3 STRUCTURE MOTIF CENTRIC LEARNING FRAMEWORK FOR INORGANIC CRYSTALLINE MATERIALS

3.1 Introduction ..................................................................................................................49

3.2 Structure motifs and their identification ......................................................................52

3.3 Motif2Vec: Structure motif vectorization ..................................................................55

3.4 Construction of motif environment matrix ..................................................................56

3.4.1 Analysis of motif environment matrix ...................................................................58

3.5 Incorporation of motif information in graph neural networks ..................................64

3.6 AMDNet: Atom-motif dual graph neural network ......................................................67

3.7 The Training process for atom-motif dual graph neural network .................................69
3.8 Performance comparison between different graph neural network architectures for complex oxides..........................................................71

3.9 Summary and discussion ...........................................................................................................73

CHAPTER 4 MATERIAL NETWORK BASED ON STRUCTURE MOTIFS……75

4.1 Introduction..................................................................................................................................75

4.2 Construction and analysis of motif-based material network .................................79

4.2.1 Finding similarities between compounds.................................................................79

4.2.2 Network density ..................................................................................................................81

4.2.3 Degree ..................................................................................................................................82

4.2.4 Eigenvector centrality ..........................................................83

4.2.5 Betweenness centrality (BC) ............................................................................................84

4.2.6 Matching index ...................................................................................................................85

4.3 Material network for transparent conducting oxide materials ............................86

4.4 Summary ...................................................................................................................................96

CHAPTER 5 SUMMARY AND CONCLUSION.................................................................97

BIBLIOGRAPHY .........................................................................................................................101
LIST OF FIGURES

**Figure 1.1** Architecture of an Artificial Neural Network, showing one input layer, two hidden layers and one output layer. The inputs multiplied by weights ($w_{ij}$) are added to the bias ($b_i$) before passing through the activation function ($f$) in each node in the first hidden layer. The output from the previous layer is again multiplied by the weight and combined with the bias before passing to the next hidden layer and so on……………………………18

**Figure 1.2** Material informatics workflow representing various steps to follow from data to property prediction……………………………………………………….21

**Figure 2.1** (a) Top and (b) side view of the relaxed structure of Ti$_2$Si$_2$Te$_8$. The magenta, cyan and brown balls represent Ti, Si, and Te atoms, respectively…………………………32

**Figure 2.2** Band structures of (a) Ti$_2$Si$_2$Te$_8$, (b) Ti$_2$Ge$_2$Te$_8$, (c) Zr$_2$Si$_2$Te$_8$, (d) Zr$_2$Ge$_2$Te$_8$, (e) Hf$_2$Si$_2$Te$_8$, and (f) Hf$_2$Ge$_2$Te$_8$ calculated using the HSE06 functional. High-symmetry k points in the reciprocal space are presented in the Figure (a)…………………………36

**Figure 2.3** Projected density of states (in arb. Unit) of (a) Ti$_2$Si$_2$Te$_8$, (b) Ti$_2$Ge$_2$Te$_8$, (c) Zr$_2$Si$_2$Te$_8$, (d) Zr$_2$Ge$_2$Te$_8$, (e) Hf$_2$Si$_2$Te$_8$, and (f) Hf$_2$Ge$_2$Te$_8$, showing that the dominant states at VBM are Te $p$-states and CBM are Ti/Zr/Hf $d$-states, respectively…………………37
Figure 2.4 (a) Linear fitting of the energy shift at the band edge as a function of applied strain (specific band edge and direction of strain is shown in the legend) and (b) quadratic fitting of the change in total energy as the function of lattice deformation in Ti2Si2Tes. The linear fitting of energy shift at band edge as a function of strain gives the deformation potential of -1.28 eV at CBM, -9.79 eV at VBM along x and -3.46 eV at CBM, -2.55 eV at VBM along y. With the quadratic fitting and relation mentioned before, 2D elastic modulus of 77.40 Jm\(^{-2}\) and 62.50 Jm\(^{-2}\) along x and y directions are obtained……………………………41

Figure 2.5 Projected crystal orbital Hamiltonian population (-pCOHP) between neighboring atoms of Ti2Si2Tes, with corresponding structure showing the neighboring atoms of which pCOHP is obtained from ………………………………………………………….43

Figure 2.6 Partial charge densities of 2D telluride systems at the VBM at \(\Gamma\) point (left) and partial charge densities at the CBM (top and side view) at \(\Gamma\) point (right) for (a) Ti2Si2Tes, (b) Ti2Ge2Tes, (c) Zr2Si2Tes, (d) Zr2Ge2Tes, (e) Hf2Si2Tes, and (f) Hf2Ge2Tes………………..44

Figure 2.7 Phonon dispersions of the 2D telluride compound Ti2Si2Tes…………………46

Figure 3.1 Statistical distribution of the most frequently occurring motif types in binary and ternary complex oxides in the MP database…………………………………………………………54
**Figure 3.2** Structure motifs in inorganic compounds and the generation of global motif representations by *Motif2Vec* using the motif environment matrix……………………56

**Figure 3.3** An example of the motif-environment matrix (*M*), a sample dataset of 5 compounds is used to generate this matrix. Each entry in *M* represents the frequency of occurrence for the motif-environment pair in the sample dataset……………………57

**Figure 3.4** Distribution of singular values obtained from the motif-environment matrix (with rows having sum >=20) in descending order. The flat tail of non-vanishing values follows a sharp decrease for the first ~50 dimensions. It indicates that the higher dimensions of singular values could also describe the meaningful aspects of motifs. In this work, we choose motif vectors with singular values up to a dimension of 60 for further analysis…………………………………………………………………………………………59

**Figure 3.5** Projection of motif environment matrix spanned by first two principal components (a) and third and fourth principal components (b). The coloring resembles the obtained similarities as encoded in the periodic table (c)……………………………………61

**Figure 3.6** t-SNE projection of motif vectors constructed by using the motif environment matrix. The numbered clusters from 1 to 4 are associated with motif types including (1)
cube, (2) cubooctahedral, (3) octahedral, and (4) a mixture of tetrahedral (in magenta) and square planar (in remanant)…………………………………………………………………………..63

Figure 3.7 Construction of a motif graph based on both atom-level and motif-level information encoded in an inorganic crystal……………………………………………………..66

Figure 3.8 (a) Demonstration of the learning architecture of the proposed atom-motif dual graph network (AMDNet) for the effective learning of electronic structures and other material properties of inorganic crystalline materials. (b) Comparison of predicted and true band gap values (from DFT calculations) and (c) comparison of predicted and true formation energy (from DFT calculations) in the test dataset with 4515 compounds……..68

Figure 4.1 Degree distribution in our network. The red line is the power law fitting. Unlike other networks such as social or real-world networks, it does not follow the power law….82

Figure 4.2 N1 is connected with 5 nodes (N3, N4, N6, N7, N8). N2 is connected with 4 nodes (N3, N4, N5, N8). There are three common nodes for N1 and N2 (N4, N8, N3) and in total 6 distinct neighbors (N3, N4, N8, N5, N6, N7). The matching index between nodes 1 and 2 is M12 = 3/6 = 0.5, which indicates that N1 and N2 are functionally similar even though they are not connected. The method of computing matching index is adopted from Ref. 167…86
**Figure 4.3** A material network of known 108 TCOs showing densely connected clusters at the center and various isolated nodes around. The marked clusters and the motifs they share are presented (the coloring in the labels indicates their presence in the marked clusters with respective colors). The size of the nodes indicates their betweenness centrality.............89

**Figure 4.4** Distribution of most occurring motifs in the TCO dataset used in this work.................................................................90

**Figure 4.5** Distribution of motif types in 108 TCOs.........................................................90

**Figure 4.6** Distribution of most occurring motifs in the dataset consisting of 7000 complex oxides. Only those motifs with frequency greater than 1000 are displayed.............91

**Figure 4.7** Distribution of motif types in the dataset with 7000 complex oxides. TP stands for “trigonal prism”, TSF for “three square faces”, and WTF for “with triangular faces”...91

**Figure 4.8** Representation of 7000 complex oxides in a material network. Red nodes are the TCO mapped in the material network.................................................................94

**Figure 4.9** (a) Extracted sub-graph for the compound GeO$_2$ (mp-10913) from the big graph. (b) Band gap distribution in the subgraph (Color: $E_g \leq 0.3$ purple nodes, $0.3 < E_g$
< 1.5 blue nodes, and $E_g \geq 1.5$ orange nodes). This coloring scheme can be more helpful in screening compounds based on band gap from the subgraph. The star-like topology of this subnetwork indicates the subnetwork is centralized and all the compounds have similar connectivity.
LIST OF TABLES

Table 2.1 Relaxed lattice constants ($a$ and $b$) and computed band gaps estimated by using the HSE06 functional………………………………………………………………………………35

Table 2.2 Effective mass $m^* (m_y^*)$, deformation potential $E_{1x} (E_{1y})$, 2D Elastic modulus $c_{2D} (c_{2D})$ and the carrier mobility $\mu_x (\mu_y)$ along the dominant transport direction $\Gamma$-X and $\Gamma$-Y for electron (e) and hole (h) carriers………………………………………………………………………………39

Table 3.1 Performance comparison between various graph architectures for the learning and prediction of electronic band gaps, formation energy, and metal vs. non-metal classification for complex oxides (trained on 18,091 compound and tested on 4,515 compounds)………………………………………………………………………………73

Table 4.1 Compounds (with their mp-id) identified as the neighbors of GeO$_2$ (mp-10913) (in green in Fig. 4.4) in our material network are listed as TCO in the literature. The unique motif types are present in the compounds are presented………………………………………………93
CHAPTER 1
DENSITY FUNCTIONAL THEORY AND
MACHINE LEARNING: AN OVERVIEW

1.1 Material discovery and design: past and present

The behavior of condensed matter phases is controlled by various physical laws, mostly quantum mechanics and statistical mechanics, correlating atomic structures and macroscopic properties. The study of these condensed matters and the correlations between their atomic structures and material properties lead to the significant development in the materials science and technologies, by which the achievement of humankind has been dramatically influenced. The search for better materials in various applications, such as tools and weapons, led to the advancement in material processing techniques. Eras named by the predominant material in use (the stone, iron, and bronze ages) demonstrate the progress in material discovery over time and their importance in human civilization. In addition, the development of various laws in physics eased the understanding of materials and led to new discoveries.

The successful formulation of quantum mechanics in the early 20th century revolutionized the field of physics and chemistry. This made it possible to analyze every system microscopically and changed the thinking of the scientific community on microscopic
particles like atoms, molecules, and subatomic particles such as electrons. This understanding provided a solid theoretical background for the development of physics and chemistry and made it possible to determine most of the properties (such as electromagnetic, mechanical, thermal, and chemical) of matter theoretically. It made it possible to discern the interaction between electrons and atomic nuclei to build up the matter in all its form. However, the major problem lies behind the too much complexity of quantum mechanical equations governing the properties of large molecules and solids. Since that period, significant theoretical and computational progress has been made to solve the complex quantum mechanical equations of molecular and solid-state systems. Much of this progress is based on approximations. The major breakthrough occurred when Pierre Hohenberg and Walter Kohn formulated the density functional theory (DFT)[1], [2]. For this development, Walter Kohn shared the 1998 Nobel Chemistry prize with John Pople, who developed computational methods in quantum chemistry. Based on DFT, the electronic structure can be calculated by using the knowledge of electronic density rather than many-body wavefunction. The success of DFT lies in its most basic assumption; “all the ground state properties of a matter can be determined with the knowledge of charge density distribution.” Atomic nuclei and a dense shower of interacting electrons are considered as the constituent of all material systems. Rather than considering the motion of individual electrons, Walter Kohn's theory proposes the knowledge of an average number of electrons located at any point in space. This assumption leads to a method that is computationally simpler than all other existing models at the time. In addition, DFT became the prominent method to calculate the ground-state properties of large molecules,
because it replaces integral over $3N$ dimensional space by the integral over three-dimensional space that can be solved much faster[3].

The significant progress in the solution of complex quantum mechanical equations started to happen when computers made it possible to handle gigantic numerical works needed to solve such equations. By the end of 1990, the field has observed significant development in the theoretical approach and computational power to solve complex quantum mechanical equations, which revolutionized the field of physics and chemistry. Computer-based calculations mainly served as the companion for the experimental techniques. Hence, rather than solo experimental efforts, the discovery process started to rely on computation as well; since then, it became the well-accepted material discovery process. In the recent past, the advancement in both electronic structure code and computing power has allowed us to use DFT for high-throughput computations of a vast number of materials and their properties, leading to the development of various databases of inorganic materials with their properties and applications listed. Some of those databases include the Materials Project[4], the Automatic-Flow for Material Discovery (AFLOW)[5], the Open Quantum Materials Database (OQMD)[6], and so on. On the other hand, this increase in material datasets ignited the interest of the community towards the acceleration of the material discovery process by using the available data and the current state of the art machine learning (ML) techniques.
ML has been successful in surpassing human performance in various tasks, such as board games (e.g., computer chess), pattern (e.g., facial) recognition, event forecasting, bioinformatics, and many others.[7], [8] Success of the deep learning techniques in identifying hidden patterns from the data made it a promising tool in many research communities that involve data-driven efforts. In addition, these deep learning techniques have the ability to bypass or replace the calculations needed to obtain the data in the first place. There have been various attempts to identify an alternative route to the expensive quantum mechanics-based calculations and optimize computational resources by using ML. For instance, Brockherde et al.[9] used an ML approach to predict the electronic density, which is a form of the Hohenberg-Kohn map from the potential to density. ML has begun to succeed in material science and grasps considerable promise for material discovery[10], it has been successful in shortening the gap between accumulated information (in various databases) and derived knowledge from them.[9], [11], [12] The ML approach of material design and discovery has enormous ability to alleviate the price, risks, and time that is involved in lengthy preparation and testing in experimental approach or time-consuming computing approaches.

1.2 Density functional theory

The development of quantum mechanics made it possible to discern the microscopic properties of the systems. Fascinated by Louis de Broglie's[13] idea about the wave-like
nature of electrons, Erwin Schrödinger devoted himself to finding a wave equation for the electron. His attempt to treat the electron as a non-relativistic particle resulted in the non-relativistic version of the wave equation, which was able to reproduce the correct spectra energies of the Hydrogen atom. This successful attempt gave the famous equation coined after Schrödinger’s name, the time-independent non-relativistic Schrödinger wave equation (SWE) and is given by

\[ \hat{H} \psi = E\psi \quad (1.1) \]

where \( \hat{H} \) is the Hamiltonian is the sum of kinetic energy operator, electron-nuclei interaction, and electron-electron interaction energy operator, and \( E \) is the corresponding energy eigenvalue. In quantum mechanics, \( \psi \) is the system’s wavefunction, which contains all the information we can possibly have about a given system. The solution of Eq. 1.1., provides plenty of information that can be used by physicists or chemists. However, SWE can be solved exactly only for hydrogen-like systems. The equation possesses intrinsic difficulties due to electron-electron interactions, which are present in the system of interest in the chemistry and condensed matter physics. So, solving the many-body Schrödinger equation is the biggest challenge in quantum mechanics. To deal with the many-body effect (exchange and correlation), many wavefunction theories have been developed. Hartree-Fock[14] method is a mean-field theory that takes exchange interaction into account. In this approach, the many-body wavefunction is represented by a single Slater determinant
and so misses the correlation. Post-Hartree-Fock methods (CCSD, CI, etc.) incorporate the correlation effect. The computational cost of these wavefunction methods increases with the number of electrons, and they become impractical for solids and large molecules. DFT is an alternative approach in which the electron density replaces the wavefunction. The Thomas-Fermi\cite{15} model was the first density functional theory proposed in the early 1930s. This model was somehow successful in describing the total energy of atoms qualitatively, however, it is not useful for the systems with valence electrons, which accounts for most of the systems in chemistry and material science\cite{16}. However, it hinted at the possibility of using the electron density as a fundamental variable.

Two simple theorems by Hohenberg-Kohn in 1964 became the foundation of the modern density functional theory, the most widely used atomistic simulation technique to compute the electronic structure of a matter. For a system with $N$ interacting electrons in an external potential $v(\mathbf{r})$, the first theorem states that the external potential $V_{\text{ext}}(\mathbf{r})$ (to within a constant) is a unique functional of the electronic density $n(\mathbf{r})$ and the second theorem is the variational principle for the ground state energy and density. It states that electron density fixes the Hamiltonian and hence all the properties of the system. The total energy is the functional of density $n(\mathbf{r})$, and the second theorem states the ground state energy $E[n]$ is minimal for the ground state density. The energy functional can be expressed as:

$$E[n(\mathbf{r})] = T[n] + V_{ee}[n] + \int d^3r \, v(\mathbf{r})n(\mathbf{r})$$  \hspace{1cm} (1.2)
For the practical application of the theory, we need to resort to approximations. The first approximation used is the Born-Oppenheimer approximation, which decouples the electronic and nuclear wavefunctions, and the second approximation is the Kohn-Sham approximation. The Hohenberg-Kohn theorem tells us that energy is a functional of the density, while Kohn-Sham theory (1964) gives a way to obtain it. It replaces the real system of interacting electrons by an auxiliary system of non-interacting electrons, but both are having the same ground-state density. The energy functional can be expressed in the following form:

\[ E_{tot}[n] = T_s[n] + V_{ext}[n] + U[n] + E_{xc}[n] \]  

(1.3)

where \( T_s[n] \) is the non-interacting kinetic energy, \( V_{ext}[n] \) is the potential energy due to the interaction with the ions, and \( U[n] \) is the classical electrostatic interaction (Hartree) energy, and \( E_{xc}[n] \) is the exchange-correlation energy. The wavefunction for the non-interacting system is the single Slater determinant of the spin-orbitals \( \phi_i \), that satisfies the Schrödinger equation shown below.

\[
\left[ -\frac{1}{2} \nabla^2 + v^{s}(\vec{r}) \right] \phi_i = \epsilon_i \phi_i 
\]  

(1.4)

Here, \( \epsilon_i \) is the energy eigenvalue for the \( i \)th electronic state and the wavefunction \( \phi_i \) is also called Kohn-Sham orbital. The reference potential \( v^{s}(\vec{r}) \) can be shown to be
\[ v^s(\vec{r}) = v_{ext}(\vec{r}) + \int d^3 r' \frac{n(\vec{r})}{|\vec{r} - \vec{r}'|} + v_{xc}(\vec{r}) \]  

(1.5)

In Eq. 1.5., the first term on the right-hand side is the external potential, the second term is the exchange-correlation potential. Clearly, the reference potential depends upon the density \( n(\vec{r}) = \sum_{i=1}^{N} |\phi_i|^2 \). That means Eq. 1.4., must be solved self-consistently. So, one starts with the guessed density \( n(r) \), plugs in a functional form of \( v_{xc} \), and builds the Kohn-Sham potential \( v^s(\vec{r}) \). Next, they obtain the eigenvalues and eigenvector of the Kohn-Sham equations. The electronic density is derived from the set of \( \phi_i \) and the process is repeated until the density stops changing.

1.2.1 Classification of the exchange-correlation energy

As explained in the previous section, the exchange-correlation potential is an exciting and essential piece. Its exact form is generally unknown. Even if we know the correct type, it would be computationally not practical. Although the complexity of the quantum many-body interactions (that are contained in the exchange-correlation energy) is unknown, several suitable approximations can be made. The most successful way to approximate the exchange-correlation energy is the constraint search method that uses the exact mathematical conditions (which are universal) satisfied by the exchange and correlation energies. The approximations to \( E_{xc} \) can be beautifully classified into Jacob's ladder[17] as
defined by J. Perdew. When one climbs up the ladder, the sophistication of the approximation and its accuracy increases as well as the computational cost.

1.2.1.1 The local density approximation (LDA)

The local density approximation LDA is the simplest and yet successful approximation for $E_{xc}[n(r)]$. LDA assumes that the energy depends on the electron density at each point in space. It is considered the mother of all approximations. The key ingredient of this approximation is the uniform electron gas. In this approximation, the exchange-correlation functional can be expressed as

$$E_{xc} = \int \varepsilon_{xc}(n(r))n(r)dr$$  \hspace{1cm} (1.6)

Where, $\varepsilon_{xc}[n(r)]$ is the exchange-correlation energy per particle of a uniform electron gas of density $n(r)$, which can be decomposed into an exchange and a correlation part

$$\varepsilon_{xc}[n(r)] = \varepsilon_x[n(r)] + \varepsilon_c[n(r)]$$  \hspace{1cm} (1.7)

The exchange part ($\varepsilon_x$) is the exchange energy of an electron in a uniform electron gas of particular density and is known precisely. The correlation energy is also known numerically up to two or three significant figures at selected values of the density[18],
which was later fitted for all densities.[19] The success of LDA in slowly varying electronic density, such as in bulk metals, played a crucial role in making it well accepted and popular among physicists[20] during the 1970s. In addition, it is successful in predicting various properties such as structures, vibrational frequencies, elastic moduli, and phase stability of the system. However, it underestimates the binding energy of numerous systems and, to some extent, fails to predict the energy barriers in the chemical reactions, which made it less famous in quantum chemistry. In addition, it can also fail in systems with heavy fermions, so dominated by electron-electron interaction effect. The success of LDA is due to the error cancellation between exchange and correlation.[21] Typically, it underestimates $\epsilon_c$ but overestimates $\epsilon_x$, yielding reasonable $E_{xc}$.

1.2.1.2 The generalized gradient approximation (GGA)

In LDA, one makes use of the knowledge of density at a point $r$. In practice, any real system is spatially inhomogeneous. Hence, in addition to density, it is very logical to consider the non-homogeneity (i.e., the gradient) of the electron density. With the inclusion of gradient, the exchange-correlation energy functional can be written as

$$E_{xc} = \int \epsilon_{xc}(n(r), \nabla n(r)) n(r) dr$$  \hspace{1cm} (1.8)
GGA significantly improves the binding energy problems, systematic errors such as overestimation of molecular atomization energies, and bond lengths from LDA. Because of this, it was widely accepted by chemists during the early 1990s. Several functionals within GGA has been developed. Some of the most successful GGA functionals are Perdew-Burke-Ernzhof (PBE)[22] and Perdew-Wang (PW91)[19] in physics and Becke-Lee-Yang-Parr (BLYP)[23] in chemistry. Many other types of GGA functionals are available, and new ones continue to be added in the basket of DFT. In general, the currently available GGAs can produce accurate results for main types of chemical bonds such as covalent, ionic, and metallic. However, common GGAs and LDAs fail at capturing van der Walls interactions[24], and more specialized approaches have been developed within DFT to deal with such weak interactions[25]

### 1.2.1.3 Meta-GGA functionals

The Recent development of exchange functionals has led to a more complex exchange-correlation functional. In addition to the density and its derivatives, the new ingredient in these complex functional is the Kohn-Sham kinetic-energy density. Some functionals which depend explicitly on the semi-local information in the Laplacian of the spin density have been developed. These functionals are named as meta-GGA functionals. In this approximation, the exchange-correlation energy functional can be written as
\[ E_{xc} = \int \epsilon_{xc}(n(r), \nabla n(r), \nabla^2 n(r), \tau(r)) n(r) dr \]  

(1.9)

The Kohn-Sham orbital kinetic energy density \( \tau(r) \) is given by

\[ \tau(r) = \frac{1}{2} \sum_i^{ooe} |\nabla \varphi_i(r)|^2 \]  

(1.10)

Meta-GGA functionals show improvements over GGA functionals in many cases, such as atomization energies as well as metal surface energies.[21] The Tao-Perdew-Staroverov-Scuseria functional[26] and the most recently proposed strongly constrained and appropriately normalized (SCAN) functional[27] are the two most popular Meta-GGAs. SCAN is very accurate in predicting the transition pressure[28][29] and is often combined with non-local correction methods like rVV10[30] to supplement the missing van der Walls interactions required to address the adsorption problems.[31], [32]

1.2.1.4 Hybrid exchange functionals

All the approximations discussed so far are based on local (LDA) or semi-local (GGA and meta-GGA) functionals of the density. Among others, the most severe drawback of LDA and GGA is the underestimation of the band gaps of Kohn-Sham band structures. To address the band gap problem, recent developments in exchange-correlation functional
introduce a non-locality in DFT by incorporating a fraction of Fock exchange \(E_x^{Hf}\) with exchange-correlation energy within the GGA.

\[
E_{xc}^{hyb} = (1 - \alpha)E_x^{DFT} + \alpha E_x^{Hf} + E_c^{DFT}
\]  

(1.11)

Here, \(\alpha\) is a mixing parameter and \(E_x^{DFT}\) is an appropriate GGA. The main shortcoming of this high-level functional is the computational requirements. Some of the widely adopted hybrid functionals in physics include PBE0[33], [34], and the Coulomb interaction screened Heyd-Scuseria-Ernzehof (HSE) hybrid functional[35] based on the PBE E\(_{xc}\). In quantum chemistry, the B3LYP functional[36], [37], has been widely used, which introduces mixing as well as other empirical parameters into its precursor BLYP. These functionals are able to make significant improvements in the calculation of many material properties, such as bond lengths, atomization energies, and bang gap in solids.

1.3 Machine learning

1.3.1 A Brief introduction

As a sub-field of artificial intelligence, ML aims to understand the structure of the data and fit that data into models. ML algorithms allow computers to train on input data and use statistical analysis to fetch the output that lies within a specific range. ML task can be
generalized as follows: with the given set of feature space \( X \) and labels \( Y \), we can approximate a function \( y = f(x) \), which maps the features \( x_i \) in \( X \) to their corresponding labels \( y_i \) in \( Y \) by means of supervised learning algorithms. On the other hand, if no labels are present in the data, the task is to find a pattern present in the data, and the learning is named as unsupervised learning. In supervised learning with the known training datasets, the learning algorithm hypothesizes a function to make a prediction about the output values. After enough learning, the model is able to predict the target value for any new input. A well-known example of supervised learning is that of an email spam filter in which with enough data about spam and ham (i.e., not spam) messages, a model is constructed to predict if a new message is a spam or a ham.

A wide range of learning algorithms such as linear regression, logistic regression, support vector machines and kernel methods, neural networks, and many others are considered as supervised learning algorithms[21]. While, in unsupervised learning, the task is to infer a function to describe the hidden structure from unlabeled data. Instead of finding the right output, it explores the data and draws a conclusion that describes the hidden patterns present in the data. Unsupervised learning can be used for feature learning, which allows the machine to automatically discover the representations that are needed to classify the raw data. More often, it is applied in Anomaly detection, Business intelligence problems, Social network analysis, and many others. Other types of ML algorithms include semi-supervised and reinforcement learning. The detailed explanation of such algorithms is beyond the scope of this work.
The ML algorithm consists of three components: representation, optimization, and performance evaluation. Evaluation metrics are used to quantify the model performance. Major ML applications constitute of supervised learning tasks such as classification and regression. Some of the classification metrics include classification accuracy, confusion matrix, logarithmic loss, and area under the curve (AUC). While in regression problems, the most common model evaluating metrics are root mean squared error and mean absolute error, in which the task is to minimize the error which is defined by using the following Loss Functions (L):

\[ L = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \]  \hspace{1cm} (1.12)

or

\[ L = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \]  \hspace{1cm} (1.13)

The summation in Eq. 1.12. and Eq. 1.13. runs from n=1 to the total number of instances N in the dataset, \( y_i \) and \( \hat{y}_i \) are the actual and predicted values respectively. The choice of a learning algorithm depends on the nature of a problem. Details about all the algorithms mentioned are available elsewhere.[21] Due to the interest of this work, a brief overview of the artificial neural networks is presented.
1.3.2 Artificial neural networks (ANNs)

ANNs are a very primitive generalization of biological neurons. They are composed of layers of computational units called neurons, with a connection between different layers through the adjustable weights. The major constituents of ANNs are weights, bias, and the activation function. An excellent choice of the activation function results in the proper accuracy of an ANN model. The most widely used activation functions are Logistic (known as Sigmoid)[38][39] Rectified linear unit,[40] and SoftPlus[41] given by following three equations respectively.

\[ f(x) = \frac{1}{1+e^{-x}} \quad \text{(1.14)} \]

\[ f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases} = \max\{0,x\} \quad \text{(1.15)} \]

\[ f(x) = \ln(1 + e^x) \quad \text{(1.16)} \]

Passage of information along a predetermined path between the neurons is the fundamental idea behind the construction of ANNs. Its architecture is very flexible and various network parameters (such as weights, bias, number of nodes, and number of hidden layers) can be tuned to improve the performance of the network. One can add up the information from
multiple sources to the neurons and apply a non-linear transformation at each node, which helps the network to learn the complexity present in the data. With the application of linear and non-linear transformation in the input data, ANNs transform those initial representations up to a specific outcome. Depending on the learning task, the outcome of the network could be either classification or regression. One crucial aspect of ANNs is that the output is generated instantaneously, and the outputs of ANN neurons are continuously valued. The training of the network in ANNs is usually done by using gradient descent, an optimization algorithm used to determine the gradient of a predefined cost function. During the training, the task is to identify the direction of steepest descent by using the negative of the gradient in various iterations. The goal is to identify a global minimum. Once the minimum if found, the network is considered trained, and no further training is needed. Due to the increasing interest of the scientific community in ANNs and its power in the learning process, various gradient descent algorithms have been invented over time; some of the most prominent algorithms include Stochastic Gradient Descent (SGD)[42] and Adadelta or Adam[43]. A schematic for the neural network is shown in Figure 1.1.
Figure 1.1 Architecture of an Artificial Neural Network, showing one input layer, two hidden layers, and one output layer. The inputs ($x_i$) multiplied by weights ($w_{ij}$) are added to the bias ($b_i$) before passing through the activation function ($f$) in each node in the first hidden layer. The output from the previous layer is again multiplied by the weight and combined with the bias before passing to the next hidden layer and so on.

One of the key features behind the success of ANNs is its iterative learning process. Once the inputs are given to the network, the weights are adjusted during the learning phase. The mathematical foundation of neural networks with weights and differential activation functions is such that all steps from inputs to the outputs are parallelizable, which allows us to run ANN on a large scale at once. Activation functions put together in neural networks are the task to optimize the loss function. The loss function quantifies a gap between prediction and ground truth. For the regression task, the loss function is mean squared error (MSE), and the ground truth is a real number, while in binary classification, it is a cross-
entropy loss, and the ground truth is either 0 or 1. With actual value ‘$y$’ and predicted probability ‘$p$’ the cross-entropy loss ($L$) for the binary classification is given by

$$L = -y(\log(p)) - (1 - y)\log(1 - p)$$

The main advantage of neural networks includes their ability to endure noisy data and classify patterns in the unseen data during the training. The most popular algorithm used by the neural network during the learning is the backpropagation. The differentiable nature of the activation functions made it possible to use backpropagation. Activation functions are used to transform linear input signals into non-linear output signals. The weights are adjusted through backpropagation so that the optimum goal is to calculate the output with minimum error.

Due to the success of ANNs in complex tasks, more and more complex ANNs have been invented over time. Some of them include Convolution Neural Network (CNN)[44] used to generate sequential data such as images and Recurrent Neural Network (RNN) such as texts and videos.[45] In practice, prior knowledge about the problem and the learning goals are the main ingredients behind the success of any ML algorithm. As all ML models rely on data, their integration, selection, cleaning, or pre-processing is another crucial task behind success. Also, the correct interpretation of results and deploying the discovered knowledge in the specific task will measure the success of any ML algorithm.
1.3.3 Machine learning in material science

Since its development, DFT has been the most successful tool in computational physics. In addition to DFT, the development of Monte Carlo Simulations[46] and molecular dynamics code[47] allowed researchers to explore the material space thoroughly. However, due to the increased size of the material search space, DFT calculations remains a computationally demanding task. On the other hand, it’s almost impossible to categorize all the material properties, since there are potentially millions of materials, composed of various elements in the periodic table. The availability and the ongoing development of the material databases leverage the use of a data-driven and ML approach to predict material properties.

The remarkable success of ML in information science, such as bioinformatics, drug design, and many others, is the key motivation behind its application in material science. In contrast to the DFT, the ML approach uses the available data from past computations or experiments and accelerates the learning process. Various ML model has been used and developed for diverse material applications such as predicting band gap[48]–[50], formation energies[51]–[53], thermal conductivity[54], melting temperatures[55], [56], and mechanical properties of materials[57], [58].
The increased use of ML in various tasks in material science leads to the development of a new field called material informatics[10]. The primary mission is to find a mapping between the known standard features of the materials (M1, M2, M3, and so on) and their properties. The material informatics workflow can be realized in Figure 1.2, which shows various steps in mapping crystals to the property of interest.

The success of any ML algorithm mainly lies behind the amount and quality of the data one uses to train the model. Therefore, the representation which serves as the ingredient to the ML models must be well formatted. The learning algorithm will be able to identify the desired relationship only if the necessary variables are sufficiently represented. Extracting suitable features from crystals that well describe the target property is the most challenging task in the application of ML in material science, as the crystal descriptors are not well-

**Figure 1.2** Material informatics workflow representing various steps to follow from data to property prediction.
formatted. This feature extraction or feature engineering process is accountable for most of the time and effort used in any ML project. The goal is to identify the representations that assure prediction accuracy comparable to or better than quantum mechanics-based calculations for a wide range of systems, yet with a lower computational cost. A set of standard features for material includes structure, composition, symmetry, physio-chemical properties like melting point, properties of constituent elements, or any other descriptors like coordination number. The surging interest in the application of ML in material science exacerbates the research in the representation of crystal structure that best correlates the target property. There have been many attempts to represent the crystals; Zhou et al.[59] proposed an idea to represent the crystals based on an atom and its environment in the chemical formula. By applying PCA, such representation is able to classify the elements in a similar way as in the periodic table. Besides, the ML algorithm is able to predict the formation energy with reasonable accuracy by using such representations as input. However, the most important structural information about a compound is missing in such an atom environment vectors.

The advancement in ML algorithms, such as deep learning, made it possible to extract features from structured or unstructured data. By using a fraction of atoms present in a compound as an input to the deep neural network, Jha et al.[60] demonstrated the power of the deep neural network to generalize the problem in material science. Various layers in the network can decode the complexity present in the data. Recent advancements in geometric deep learning made it possible to deal with unstructured data in the non-
Euclidean domain, such as graphs and manifolds[61] in a more elegant way. The growing interest of deep learning on graphs resulted in numerous attempts to apply it in various problems ranging from biochemistry[62] to physics[63], [64]. Xie et al.,[63] and Chen et al.[64] proposed the crystal graph convolution neural network (CGCNN), which realizes the use of graph representations for inorganic crystals. It takes the raw crystal data as an input, transforms it into a graph, and processes it through a series of convolution layers. Cheol et al.[65] proposed the improved CGCNN (iCGCNN) with increased prediction accuracy by incorporating additional descriptors to the CGCNN. All of these models, which are based on ANNs, became the prototype in material science, as they can be used for any material property prediction.

In the search for better representation of the crystal structure that best correlates the target property, we propose the structure motif-based representation for crystals. As motifs (i.e., the coordinated polyhedron) are the building blocks of crystals, they determine various properties of the crystals. Therefore, structure motifs are the robust descriptors of crystalline structures. The stability of a crystal depends on the number of atoms that a motif shared with neighboring motifs in a crystal. When the number of sharing increases, the stability of the material decreases, as a higher number of sharing brings more cations closer together, which results in more electrostatic repulsion that a cation experience. Besides, the number of different kinds of coordination polyhedra in a crystal tends to be small, because the types of polyhedral that can be stacked together with unaltering the crystal symmetry is limited.
1.4 Discovery and design of two-dimensional (2D) materials

The long-debated question of the existence of 2D materials came to an end after the successful exfoliation of graphene from graphite in 2004\[66\]. The importance of this breakthrough in the field was recognized by the Physics Nobel Prize in 2010. Nature offers numerous layered materials that span a wide range of compounds, including oxides, carbides, nitrides, chalcogenides, and many others. The thickness of these stacked layers ranges from one atom to many, which are weakly bound by electrostatic, van der Walls, or hydrogen-bonding interactions. Even though the interactions between these layers are weak, these interactions play a significant role in the electronic structure of the system when one moves from bulk to monolayer limit. For instance, when one moves from bulk to monolayer limit, band gaps may double in many transition metals dichalcogenides\[67\]. Alternatively, the material characteristic of the band gap may completely change from semiconducting to metallic as in graphene\[68\] or palladium sulfide\[69\]. Recent advancement in the various exfoliation techniques has set the ground for the reduction of layered bulk material to the monolayer limit\[70\].

Graphene has been the focus of research for numerous theoretical and experimental studies in the past decade. Graphene possesses an excellent physical property such as high carrier mobility, high electrical and thermal conductivity, and flexibility and strength, it is an auspicious material for nanoelectronics and likewise, a promising candidate to replace
silicon in the future. Even though graphene has great potential, there are still many challenges to overcome; its lack of electronic band gap limits its use in electronic devices acting as a switch. Significant efforts have been devoted to band gap engineering of graphene, and the goal is to make it well accepted in various electronic and optoelectronic applications. However, the success in the band gap engineering in graphene is accompanied by the cost in its mobility and other conducting properties. Along with these efforts to tune the electronic properties of graphene, the material science community has devoted significant time and effort in search of novel functional materials as a graphene alternative. These efforts have resulted in the discovery of many 2D materials such as Transition metal dichalcogenides (TMDCs) with MX₂ stoichiometry[71] (M as Molybdenum (Mo), Tungsten (W), and X such as Selenium (Se), Sulphur (S), and Tellurium (Te)), Phosphorene, and many others. TMDCs family contains compounds that possess a wide variety of electronic properties, including metals, non-metals, insulators, and semiconductors. Band gaps in these compounds span a wide range from direct band gaps to indirect band gaps.[72] Depending on the composition, TMDCs possess a wide range of electron and hole mobilities. Besides, these compounds have a variety of demanding characteristics that affect the charge transport, magnetism, and their catalytic and optical properties. The most prominent and widely studied semiconducting TMDCs for application in electronics include MoS₂, WS₂, MoSe₂, and WSe₂. By stacking confinement and applying the electric field, the band gap can be tuned easily in these materials[73]. High carrier mobility, in-plane stiffness and breaking strength of these materials have made them appealing for applications in high-speed electronics, flexible electronics,[74] and
many others. Recent success in the synthesis[75], [76], and successful fabrication of devices[77] from phosphorene, a semiconducting monolayer of black phosphorous, has provided enormous hope towards the search for novel 2D materials. Because of the anisotropic nature of black phosphorous, it is unique among layered materials, consisting of individual layers buckled, with two in-plane mutually perpendicular direction having armchair and zig-zag edges. The structural anisotropy strongly influences the electronic properties along these directions.

Even though these new classes of 2D materials possess many properties such as band gap, which are absent in graphene, their carrier mobilities are thousands of orders less than that of graphene. This provides an extensive playground in the search for novel functional materials as a graphene alternative. With the hope of identifying a better choice, we present a family of semiconducting monolayer 2D tellurides by using DFT in Chapter 2 of this dissertation.

1.5 Complex oxides

In comparison to non-oxides, oxides materials can host a variety of functional properties, mostly owing to the presence of transition metal elements in many compounds of the family.[78] Ranging from simple binary to complex ternary and quaternary oxides, they possess a wide range of electronic band gaps. Besides, most of these oxides exhibit other
exciting phenomena such as ferroelectricity and magnetism. These properties are susceptible and, to some extent, controlled by the type of structural building block and crystal chemistry of these compounds. Also, the detailed analysis of these structural information provides the way of engineering for new functionalities in these complex oxides materials.[78] For instance, in the family of ABO\textsubscript{3} perovskites[79] and related layered compounds, the presence of octahedral motifs plays a significant role in their electronic properties.

Due to the richness in their functional properties, complex metal oxides are a subject of intense interest for physicists, engineers, and material scientists. Besides, these oxide compounds provide an excellent material platform for the creation of various heterostructures, resulting in new heterogeneous oxide systems.[80], [81] Due to the competing phases at the interfaces, novel and unexpected phenomenon may emerge at the interfaces. Owing to these interesting electronic and magnetic properties of these oxides, studying the structural details and their correlations with desired material properties is very demanding. In this work, we present our effort in identifying the structure-property relation in a large set of metal oxides (22606) extracted from the Material Project database. The state-of-the-art machine learning techniques have demonstrated their ability to explore the structure-property relationship in these complicated crystalline materials. More details are presented in Chapter 3.
2.1 Introduction

The synthesis of atomistic simulations based on DFT and supervised/unsupervised machine learning techniques are becoming more and more powerful tools to discover the structure-property correlations in condensed matters and inorganic functional materials, such as two-dimensional (2D) materials and complex oxides. The integration of data-driven approaches in traditional computational physics and materials science involves the extraction of human knowledge based on atomistic simulations in terms of the correlations between local/global symmetry environments and their electronic structures and the incorporation of these physical principles into a machine learning framework by architecture design. From this perspective, atomistic simulations and detailed analysis of structure-property correlations for known material classes are essential.

Since a decade ago, the successful fabrication of graphene,[66], [82] which is the first synthesized 2D material, has paved the way for the discovery and synthesis of many other 2D materials and functional devices. Great efforts have been focused on the discovery and synthesis of novel 2D materials. A number of 2D materials such as transition metal
dichalcogenides (TMDCs),[83], [84] transition-metal oxides,[85], [86] MXene,[87] and black phosphorus (BP),[88], [89] have been synthesized. Successful fabrication of field-effect transistors, photodetectors, and light-emitting diodes using these 2D materials have demonstrated their potential for usage in electronics. These novel 2D materials possess remarkable properties that are absent in their bulk counterparts. For example, the band gap transitions from indirect to direct as one decreases the number of layers in MoS\(_2\).[90] 2D materials have great application potential in optoelectronics, nanoelectronics and ultrathin flexible devices in the future.[91]

The field-effect transistor (FET) is the most successful device concept and a vital component of present semiconductor electronics. Carrier mobility of the channel material and its band gap plays an essential role in the performance of FETs[92] such as on-off ratio and operation speed, and the identification of novel channel materials in the FET technology is a significant task in semiconductor electronics. Graphene exhibits extremely high carrier mobility (10,000 – 15,000 cm\(^2\)V\(^{-1}\)s\(^{-1}\)),[93], which has ignited great excitement in the device community as a possible channel material in FETs. However, the gapless electronic structure and related low on/off ratio of graphene transistors hinder their application for logic operation.[94] Tremendous efforts have been applied to open a sizable band gap in graphene. Unfortunately, the creation of a band gap has always been accompanied by a dramatic decrease in mobility.[93]
With the successful fabrication of high-performance MoS\textsubscript{2} FETs,\cite{95} the search for novel 2D materials for electronics gained more momentum. FETs made from few-layer TMDC’s such as MoS\textsubscript{2} and WSe\textsubscript{2} exhibit high on/off current ratios and excellent current saturation characteristics, but the carrier mobility ($< 200 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$)\cite{96,97,98} is much smaller than that of graphene. Recent studies show that another 2D semiconductor, single-layer black phosphorus (phosphorene) exhibits high carrier mobility ($\sim 1,000 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$),\cite{77,89} but is still orders of magnitude less in comparison to graphene. Hence, the search for stable and semiconducting 2D materials as graphene alternatives that can outperform the silicon-based devices is worthwhile and will be an excellent boon for electronic device community.

At present, the number of experimentally discovered two-dimensional functional materials are minimal compared to their bulk counterparts. Meanwhile, various available material databases such as an Inorganic crystal structure database (ICSD),\cite{99} Automatic flow of material discovery (AFLOWLIB),\cite{100} and the Materials Project (MP)\cite{101} database enabled rapid screening of these materials. Using this data-driven approach, many new functional materials were predicted and later confirmed by experiments.\cite{102} High-throughput computations combined with the experiment have successfully expanded the material space for various applications such as photocatalysis,\cite{103} solar water splitting, and many others.\cite{104}
Our search for 2D functional materials starts with the identification of large interlayer spacings, which are characteristic of weak interlayer bonding that could overcome by mechanical exfoliation.[105] With the goal to enable the accelerated discovery of novel functional 2D materials, we screen the ICSD database and identify ~1500 2D layered compounds by using a layered structure identification algorithm (unpublished work). Among these compounds, ~ 900 of them are classified as monolayer structures and are listed in our 2D material database. By combining this database and a high-throughput screening workflow, we propose a family of M₂N₂X₈ type 2D monolayer semiconductors with a suitable band gap, low carrier effective mass, and potentially high carrier mobility along the dominant propagation direction. We perform the first-principles calculations on a family of 2D ternary group-IV metal chalcogenides consisting of 36 M₂N₂X₈ compounds, with M = (Ti, Zr, Hf) from group IV, N = (Si, Ge) from group XIV, and X = (S, Se or Te) from group XVI of the periodic table. Among those 36 compounds, we found that six telluride compounds with the chemical formula M₂N₂Te₈ are very promising due to their direct band gaps, strong in-plane anisotropy, and low carrier effective masses along the major transport direction.
Compounds with similar structure and composition, including ZrGeTe₄, HfGeTe₄, and TiGeTe₆ have already been synthesized in their bulk form,[106] and the measurement of their structural information and other intrinsic properties such as electrical conductivity and magnetic susceptibility of compounds have been performed.[106] Relaxed structure of Ti₂Si₂Te₈ as a promising compound in our dataset is shown in Figure 2.1. The atomic structure of these semiconducting tellurides (M₂N₂X₈) is characterized by the presence of metal-centered bi-capped trigonal prisms as a basic structural motif, a ubiquitous feature in ternary group-V chalcogenides such as Ta[PS₄|S₂][107]. One of the caps of such bicapped trigonal prism is formed by N atoms, while all other are Te atoms in these 2D tellurides. While viewing the local environment of the N atom, it has four coordination,
one with metal atom M and three with Te atoms. The presence of such structural motifs plays a crucial role in determining the material properties, as the coordination preference does not change when one moves from bulk to monolayer. Details about the coordination environments and their importance in solid-state systems can be found in Chapter 3 of this dissertation. We will show in the following sections that these 2D semiconducting tellurides exhibit many promising electronic properties, which are of vital importance in applications such as optoelectronics,[72] nanoelectronics,[91] and ultra-thin flexible devices.[72], [108]

2.2 Computational details

The high-throughput screening, geometry optimization, and electronic structure calculations are performed by using DFT as implemented in the Vienna ab initio simulation package (VASP).[109] Projected-augmented-wave (PAW) pseudopotentials[110] are used to describe the valence electron and core interactions. Generalized gradient approximation (GGA)[111], [112] and the Perdew-Burke-Ernzerhof (PBE)[22] exchange-correlation functional are used to describe the electron exchange-correlation.

DFT calculations are performed at the PBE level to identify the trends in band structures and extract effective masses, which is considered to be a reasonable approach to finding
such properties. With DFT at the PBE level, we identify six telluride compounds with high mobility utilizing their band structure and effective masses.

It is well known that the local/semi-local functional-based DFT method severely underestimates the band gaps of semiconducting or insulating solids.[113] Alternatively, all the electronic structure calculations on the six telluride compounds are performed by using the Heyd-Scuseria-Ernzerhof hybrid functional (HSE).[35], [114] A $9 \times 3 \times 1 \Gamma$-centered k-point mesh and a plane-wave energy cut-off of 400 eV are used, which provide well-converged results. All atoms are relaxed until the final force exerted on each atom is less than 0.01 eV/Å, and the change in total energy between the two steps is less than 10^{-5} eV.

**2.3 Computed electronic properties of semiconducting tellurides**

The typical electronic band structure and atomic projected density of states (PDOS) of these 2D tellurides are shown in Fig. 2.2, and Fig. 2.3 respectively. Among all these compounds, here we present the detailed analysis of band structure and PDOS of Ti$_2$Si$_2$Te$_8$ as an outstanding example of the ternary 2D telluride family, as shown in Fig. 2.2(a) and 2.3(a). The detailed study of the band structure of this compound shows that the conduction band minimum (CBM) and valence band maximum (VBM) lies at the center of the Brillouin zone, i.e. $\Gamma$ point, indicating a direct transition between the CBM and the VBM.
is plausible. Our hybrid functional calculations show that there are sizable direct band gaps in these 2D compounds, ranging from 1.00 eV in Ti$_2$Ge$_2$Te$_8$ to 1.31 eV in Hf$_2$Si$_2$Te$_8$. Predicted values of band gaps, along with relaxed lattice constants, are presented in Table 2.1. Large dispersion in both conduction and valence bands close to $\Gamma$ point indicates the lower carrier effective mass along the dominant transport direction. Because of the variation in dispersion in the band along the dominant transport direction, the anisotropic effect can be observed in the band structure, and hence the anisotropic transport behavior of charge carriers is expected.

**Table 2.1** Relaxed lattice constants ($a$ and $b$) and computed band gaps estimated by using the HSE06 functional.

<table>
<thead>
<tr>
<th>Compounds</th>
<th>Lattice constant (Å)</th>
<th>Band gap (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>Hf$_2$Ge$_2$Te$_8$</td>
<td>3.99</td>
<td>10.99</td>
</tr>
<tr>
<td>Hf$_2$Si$_2$Te$_8$</td>
<td>3.96</td>
<td>10.85</td>
</tr>
<tr>
<td>Ti$_2$Ge$_2$Te$_8$</td>
<td>3.90</td>
<td>10.76</td>
</tr>
<tr>
<td>Ti$_2$Si$_2$Te$_8$</td>
<td>3.87</td>
<td>10.61</td>
</tr>
<tr>
<td>Zr$_2$Ge$_2$Te$_8$</td>
<td>4.01</td>
<td>11.04</td>
</tr>
<tr>
<td>Zr$_2$Si$_2$Te$_8$</td>
<td>3.97</td>
<td>10.90</td>
</tr>
</tbody>
</table>
Figure 2.2 Band structures of (a) Ti$_2$Si$_2$Te$_8$, (b) Ti$_2$Ge$_2$Te$_8$, (c) Zr$_2$Si$_2$Te$_8$, (d) Zr$_2$Ge$_2$Te$_8$, (e) Hf$_2$Si$_2$Te$_8$, and (f) Hf$_2$Ge$_2$Te$_8$ calculated using the HSE06 functional. High-symmetry k points in the reciprocal space are presented in the Figure (a).
Figure 2.3 Projected density of states (in arb. Unit) of (a) Ti$_2$Si$_2$Te$_8$, (b) Ti$_2$Ge$_2$Te$_8$, (c) Zr$_2$Si$_2$Te$_8$, (d) Zr$_2$Ge$_2$Te$_8$, (e) Hf$_2$Si$_2$Te$_8$, and (f) Hf$_2$Ge$_2$Te$_8$, showing that the dominant states at VBM are Te $p$-states and CBM are Ti/Zr/Hf $d$-states, respectively.
Electron and hole effective masses for these 2D telluride compounds are estimated by using the parabolic fitting of the energy bands around CBM and VBM near \( \Gamma \) point in \( k \) space. Semiconductor effective mass approximation is used to describe the parabolic nature of the bands at VBM and CBM. Electron and hole effective masses are obtained by using the relation, \( m^*_{e(h)} = \frac{\hbar^2}{(\partial^2 E / \partial k^2)} \). The differentiation is done at CBM for an electron and VBM for a hole at \( \Gamma \) point. Effective masses of electrons and holes span a wide range from 0.11 \( m_0 \) to 1.26 \( m_0 \) along the dominant transport direction, where \( m_0 \) is the electron rest mass. Predicted values of effective masses for electrons and holes along the \( \Gamma \)-X and \( \Gamma \)-Y directions are presented in Table 2.2.

Electronic properties of monolayer a multilayer 2D semiconductors are mainly governed by carrier mobilities. We estimate carrier mobilities theoretically for the family of 2D tellurides along the two in-plane transport directions (\( \Gamma \)-X and \( \Gamma \)-Y) based on the theoretical approach proposed by J. Bardeen et al.[115] with a phonon-limited scattering model in which carrier mobility is primarily limited by the scattering due to phonons.[115] Due to the inverse relationship between mobility and effective mass, a small effective mass is obviously one of the preliminary requirements for high carrier mobility. In addition, to carrier effective mass, other important factors affecting the mobilities include the deformation potentials and the elastic modulus along the propagation direction of longitudinal acoustic waves.[89]
Table 2.2  Effective mass $m_x^*$ ($m_y^*$), deformation potential $E_{1x}$ ($E_{1y}$), 2D Elastic modulus $c_{x2D}$ ($c_{y2D}$), and the carrier mobility $\mu_x$ ($\mu_y$) along the dominant transport direction $\Gamma$-$X$ and $\Gamma$-$Y$ for the electron (e) and hole (h) carriers.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$m_x^*$</th>
<th>$m_y^*$</th>
<th>$E_{1x}$</th>
<th>$E_{1y}$</th>
<th>$c_{x2D}$</th>
<th>$c_{y2D}$</th>
<th>$\mu_x$ (10$^3$) cm$^2$V$^{-1}$s$^{-1}$</th>
<th>$\mu_y$ (10$^3$) cm$^2$V$^{-1}$s$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hf$_2$Ge$_2$Te$_8$</td>
<td>e 0.28</td>
<td>h 0.11</td>
<td>0.35</td>
<td>3.55</td>
<td>3.58</td>
<td>87.25</td>
<td>45.75</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>e 0.25</td>
<td>h 0.12</td>
<td>0.43</td>
<td>2.58</td>
<td>2.29</td>
<td>77.21</td>
<td>53.41</td>
<td>3.10</td>
</tr>
<tr>
<td>Hf$_2$Si$_2$Te$_8$</td>
<td>e 0.30</td>
<td>h 0.12</td>
<td>0.32</td>
<td>3.62</td>
<td>3.03</td>
<td>77.49</td>
<td>39.17</td>
<td>1.36</td>
</tr>
<tr>
<td>Zr$_2$Ge$_2$Te$_8$</td>
<td>e 0.29</td>
<td>h 0.12</td>
<td>0.44</td>
<td>1.68</td>
<td>2.01</td>
<td>77.47</td>
<td>50.46</td>
<td>5.58</td>
</tr>
<tr>
<td>Zr$_2$Si$_2$Te$_8$</td>
<td>e 0.33</td>
<td>h 0.15</td>
<td>0.29</td>
<td>1.28</td>
<td>3.46</td>
<td>77.40</td>
<td>62.50</td>
<td>9.92</td>
</tr>
<tr>
<td>Ti$_2$Si$_2$Te$_8$</td>
<td>e 0.33</td>
<td>h 0.14</td>
<td>1.07</td>
<td>9.79</td>
<td>2.55</td>
<td>77.49</td>
<td>50.46</td>
<td>0.32</td>
</tr>
<tr>
<td>Ti$_2$Ge$_2$Te$_8$</td>
<td>e 0.38</td>
<td>h 0.13</td>
<td>0.22</td>
<td>2.10</td>
<td>4.30</td>
<td>67.83</td>
<td>51.69</td>
<td>2.98</td>
</tr>
</tbody>
</table>

In this work, the carrier mobility of 2D materials is estimated by a simplified relation:[89]

$$\mu_{(y)} = \frac{e \hbar^3 c_{x(y)2D}}{k_B T m_d^* m_d (E_{1y}^*)^2},$$

where, $m_d = \sqrt{(m_x^* m_y^*)}$ is the average effective mass, $i$ represents...
electron for the conduction band or hole for the valence band (subscript $x$ and $y$ indicate the $\Gamma$-X and $\Gamma$-Y transport directions), $m_e^*$ is the carrier effective mass, $T$ is the temperature (room temperature $T = 300K$ is used), $E_{1i}$ stands for the deformation potential along the transport direction, and $C_{x(y)2D}$ is the 2D elastic modulus along the transport direction.

Deformation potential for both electrons and holes along $x$ and $y$ direction is obtained by the linear fitting of band energy at the CBM and VBM with respect to strains along $\Gamma$-X and $\Gamma$-Y. 2D elastic modulus are calculated using the relation, $(E-E_0)/S_0 = C_{x(y)2D} (\Delta l/l_0)^2/2$, where $E_0$ is the total energy and $S_0$ is the lattice area at equilibrium for the 2D system, $l_0$ is the equilibrium lattice constant along the transport direction and $\Delta l$ is its change due to strain. Deformation potential is obtained by using the relation, $E_{1x(y)}^i = \Delta E_i / (\Delta l/l_0)$, where, $\Delta E_i$ is the energy shift at the band edge, VBM for hole, and CBM for electron along the propagation direction $x$ or $y$. $l_0$ and $\Delta l$ are the equilibrium lattice constant and its change under strain. Linear fits for the change in VBM and CBM energies for $\text{Ti}_2\text{Si}_2\text{Te}_8$ is shown in Figure 2.4 (a). As the difference in band edge energy is linear with lattice dilation along $x$ and $y$ directions, deformation potentials are obtained from the slopes of these fitted lines.
Figure 2.4 (a) Linear fitting of the energy shift at the band edge as a function of applied strain (specific band edge and direction of strain is shown in the legend) and (b) quadratic fitting of the change in total energy as the function of lattice deformation in Ti$_2$Si$_2$Te$_8$. The linear fitting of energy shift at band edge as a function of strain gives the deformation potential of $-1.28 \text{ eV}$ at CBM, $-9.79 \text{ eV}$ at VBM along $x$ and $-3.46 \text{ eV}$ at CBM, $-2.55 \text{ eV}$ at VBM along $y$. With the quadratic fitting and relation mentioned before, 2D elastic modulus of 77.40 Jm$^{-2}$ and 62.50 Jm$^{-2}$ along $x$ and $y$ directions are obtained.

Computed values of effective masses, deformation potentials, the 2D elastic modulus, and carrier mobilities for all the 2D telluride compounds are presented in Table 2.2. Carrier mobilities in these compounds range from hundreds to almost tens of thousands of cm$^2$/V-s, which endow this family of 2D compounds a great potential for electronic and
optoelectronic applications. These carrier mobilities exhibit high in-plane directional anisotropy, with electrons being more mobile in general. Even though the electron effective mass (0.33 m\textsubscript{0} in Ti\textsubscript{2}Si\textsubscript{2}Te\textsubscript{8} and 0.19 m\textsubscript{0} in Zr\textsubscript{2}Si\textsubscript{2}Te\textsubscript{8}) along Γ-X direction is slightly larger compared with other known 2D compounds for electronic applications, the computed electron mobility along \textit{x} direction is extremely large in Ti\textsubscript{2}Si\textsubscript{2}Te\textsubscript{8} and Zr\textsubscript{2}Si\textsubscript{2}Te\textsubscript{8} due to rather small absolute deformation potentials for conduction band along \textit{x} direction. The computed results show that holes are more mobile along Γ-Y (except in Hf\textsubscript{2}Ge\textsubscript{2}Te\textsubscript{8}).

As carrier mobilities are highly correlated with the deformation potentials, the understanding of the anisotropy in deformation potentials is essential. Since the computed electron mobility in these compounds is higher along Γ-X direction, here we focused ourselves in understanding the difference in deformation potential related to strain along \textit{x} direction via the VBM and CBM wavefunctions and bonding analysis between neighboring atoms utilizing the projected Crystal orbital Hamiltonian population (pCOHP) as incorporated in the LOBSTER package.\cite{116} pCOHP is a powerful physical quantity to understand the details of bonding, nonbonding, and antibonding interactions between a pair of atoms and their atomic orbitals in a compound.\cite{117} Here we present the -pCOHP in which a positive value corresponds to the bonding state, and a negative value corresponds to the antibonding state.
As shown in Fig. 2.5, taking Ti$_{2}$Si$_{2}$Te$_{8}$ as an example, there exists a strong bonding interaction between Ti/Si atoms and neighboring Te atoms along $x$ direction at the VBM, while a mixture of bonding and antibonding interaction is observed at the CBM. The VBM wavefunction in the left Fig. 2.6(a) shows a substantial overlap along $x$ direction, indicating that a small structural deformation along $x$ direction may have a remarkable effect on this electronic state and hence cause a significant change in its energy, resulting in a large deformation potential.
Figure 2.6 Partial charge densities of 2D telluride systems at the VBM at Γ point (left) and partial charge densities at the CBM (top and side view) at Γ point (right) for (a) Ti$_2$Si$_2$Te$_8$, (b) Ti$_2$Ge$_2$Te$_8$, (c) Zr$_2$Si$_2$Te$_8$, (d) Zr$_2$Ge$_2$Te$_8$, (e) Hf$_2$Si$_2$Te$_8$, and (f) Hf$_2$Ge$_2$Te$_8$.

The situation is rather different for the CBM wavefunction along $x$ direction, which exhibits much weaker orbital overlap in the right in Fig. 2.6(a). As observed in Fig. 2.5, the mixture of bonding and antibonding interaction between neighboring atoms along $x$ direction results in a dramatic decrease in the overall strength of the orbital interaction. Due to this effect, structural deformation has much less effect on the CBM wavefunction along $x$ direction, which results in the smaller values of deformation potential. Along the $y$
direction, the VBM wavefunctions (as shown in the left in Fig. 2.6(a)) are more localized than the CBM wavefunctions (in the right in Fig. 2.6(a)), resulting in smaller values of deformation potential at the VBM than those at the CBM along $y$ direction.

As the nature of anisotropy in the observed deformation potentials is similar in these compounds, we conclude that this anisotropic transport behavior is dominated by the bonding and antibonding interaction between the neighboring atomic orbitals and their different response under strain. Mobility along a transport direction is inversely proportional to the square of the deformation potential along that direction. Therefore, the mobility along a transport direction is strongly controlled (but not completely determined) by the deformation potential in that direction. The observed higher electron mobilities along $x$ direction in these telluride compounds are mostly due to the lower values of deformation potential at the CBM along $x$ direction.

Band decomposed partial charge densities for all compounds at VBM and CBM are shown in Fig. 2.6. These visualizations are done in VESTA and an isosurface level of 0.0064 $e\text{Å}^{-3}$ is used. The magenta, cyan, and brown ball represent atom M (Ti or Zr or Hf), N (Si or Ge), and Te respectively, as shown in Fig. 2.1. The difference in deformation potential along the major transport direction can be manifested by the bonding and antibonding nature of these wavefunctions.
2.4 Stability of 2D telluride material

To study the dynamic stability of these monolayer 2D tellurides, we calculated the phonon dispersion relation by using density functional perturbation theory as implemented in Phonopy code[118] interfaced with VASP. Based on PBE functional, a convergence criterion of $10^{-8}$ eV is used. Computed phonon spectrum for most promising compound $\text{Ti}_2\text{Si}_2\text{Te}_8$ is shown in Figure 2.7; the absence of an imaginary part in the phonon spectra indicates the dynamic stability of this compound. Similar spectra are observed for all other compounds.

![Phonon dispersions](image)

**Figure 2.7** Phonon dispersions of the 2D telluride compound $\text{Ti}_2\text{Si}_2\text{Te}_8$. 
2.5 Summary

In summary, based on PBE and HSE functional calculations, we have identified a family of 2D ternary semiconducting tellurides with the chemical composition $M_2N_2Te_8$ where $M = \{\text{Ti, Zr, Hf}\}$ and $N = \{\text{Si, Ge}\}$. The presence of metal-centered bicapped trigonal prisms, which is a ubiquitous structural feature in many group V chalcogenides, makes these compounds interesting. Considering the similarity in the coordination environments, these two families of compounds might have many similarities in their properties. The detailed analysis of the coordination environments, their types, and their correlation with the material properties such as band gaps and formation energies in the solid-state systems are presented in the upcoming chapters of this dissertation.

Computed band gaps and carrier mobilities of this compound set exhibit great potentials for electronic and optoelectronic applications. All these compounds possess direct band gaps at the center of Brillouin zone (i.e. $\Gamma$ point) with extremely high electron mobilities higher than that of single-layer BP, MoS$_2$, and other 2D MX$_2$ semiconductors. Out of these six compounds, we present two benchmark systems Ti$_2$Si$_2$Te$_8$ with a band gap at 1.03 eV and electron mobility at $9.92 \times 10^3$ cm$^2$V$^{-1}$s$^{-1}$ and Zr$_2$Si$_2$Te$_8$ with a band gap at 1.24 eV and electron mobility at $5.58 \times 10^3$ cm$^2$V$^{-1}$s$^{-1}$, which are promising for logical devices that require high mobility and optimal band gap. Like monolayer BP, the observed anisotropy in mobilities in these compounds is mainly due to the anisotropy in deformation potentials,
which is correlated with the bonding-antibonding interactions between neighboring atomic orbitals and a subsequently different response of the CBM and VBM to strain perturbation.
3.1 Introduction

Being a well-accepted and mostly used tool to compute the material properties, atomistic simulations based on DFT have enabled the so-called data-driven material discoveries while the continued and accelerated development of novel condensed matter systems with more complex properties via high-throughput computations is still facing resource-based constraints.[119] Exploration of the material information and their properties from various databases has been gaining popularity. Machine learning (ML) methods, in combination with massive material data, offers a promising route to accelerate the discovery and rational design of functional solid-state compounds by utilizing a data-driven paradigm.

Supervised learning has been effective in materials property predictions, include phase stability,[120]–[122] crystal structure,[123], [124] effective potential for molecular dynamics simulations,[125] and energy functionals for density functional theory-based simulations.[126] With the recent progress in deep learning, ML has also been applied to inorganic crystal systems to learn from the irregular representation of crystal structures and
identify their complex correlations in those high dimensional data. For instance, band gaps of given classes of inorganic compounds have been predicted using deep learning[127],[128] and ML has been applied on charge densities[129],[130] and Hamiltonian data[131],[132] to predict electronic properties.[133]

The recent development of graph convolutional network (GCN)[132],[134], when combined with domain knowledge, offers a powerful tool to create an innovative representation of crystal structures for inorganic compounds. Within the GCN network, any type of grid and atomic structure can be successfully modeled and analyzed. The flexible graph network structure endows these learning frameworks[132] a large room for improvement by considering more node/edge interactions in the crystal graphs.[135]

Whether ML can efficiently approximate the unknown non-linear map between input and output relies on an adequate representation of solid-state compound systems that captures structure–property relationships, which form the basis of many design rules for functional materials. In inorganic crystalline materials with unit cells that satisfy the periodic boundary condition, bonding environments determined by local and global symmetry are essential components for the understanding of complex material properties.

As stated in the Pauling’s first rule,[136] a coordinated polyhedron of anions is formed about each cation in a compound, effectively creating structure motifs that behave as
fundamental building blocks and are highly correlated with material properties. Structure motifs in crystalline compounds play an essential role in determining the material properties in various scientific domains and technological applications. For instance, the identification of the VO₄ functional motif enabled the discovery of 12 vanadate photoanode materials via high-throughput computations and combinatorial synthesis.[103] In the field of complex oxide devices, MnO₆ octahedral motifs are correlated with small hole polarons that limit electrical conductivity.[137] In battery cathodes for energy storage, high ion mobilities are originated from the local bonding environment of a multivalent ion.[138] For the selective oxidation of hydrocarbons, V⁴⁺ ion related motifs, and the connections of these motifs are found to be important determining factors.[139], [140] The presence of MO₄ tetrahedra (M as Si or Al), can be used to identify the most promising synthetic candidates from the pool of hypothetical zeolites.[141] When designing novel battery materials, it is found that the changing coordination pattern of a migrating ion can be used as a descriptor.[142]

Based on the structure motif information extracted from a large set of inorganic compounds, a recent work comprehensively evaluated the validity and suggested the limited predictive power of the almost one-century-old Pauling rules.[136] The extraction and construction of the dataset of the local environment and their connections is a first step towards building new structure-property correlations that could potentially benefit from the recent growth in the use of ML techniques in materials science and condensed matter physics.
In this work, we propose the incorporation of structure motif information in a machine learning framework by utilizing multiple techniques for artificial intelligence. We show that the presence of structural motifs and their connections extracted from a material structure database can be utilized with unsupervised learning to define unique representations in a high-dimensional space. The dimensional reduction process reveals strong clustering effects, representing the neighborhood properties of elements in the periodic table. By combining the motif information with graph convolutional neural network, we demonstrate the creation of a motif-centric learning framework, atom-motif dual graph neural network (AMDNet), with significant accuracy that surpasses the predictive power of the state-of-the-art atom-based graph network MEGNet for electronic structure predictions.[64]

3.2 Structure motifs and their identification

Governing the structure-property relationship, structure motifs, or coordination environments can be viewed as effective structural descriptors for crystals. The efforts of the identification of local coordination environments were initially focused on structure types[143], [144], or preferential coordination numbers[145] based on simple rules.[146], [147] Very recently, owing to the development of data-driven approaches, systematic and robust way to automatically identify local environments have been developed,[148], [149] which initialized the use of structure motif information for material design in a data-driven
paradigm. For instance, structure motif information has been used to define crystal structure similarity for all the compounds in the Materials Project database (MP).[150]

We identified and analyzed the structural motifs present in the 22606 complex oxides obtained from the MP database. We extract the structure motif information using the local environment identification method developed by Waroquiers et al.[148], as implemented in the pymatgen code.[151] We follow the definition of structure motifs or coordination environments by the International Union of Crystallography[152] and the International Union of Pure and Applied Chemistry[153], as listed in Ref. 143. Note that it is an arduous task to assign exact motif types for all sites in a broad set of compounds using a uniform set of control parameters. One would always expect some deviations from the ground truth by looking at the structures of some compounds through human eyes. We observe that the approach presented in Ref. 143 returns a higher percentage of correct motif types while doing large scale analysis. Statistical analysis of the motif types present in 22606 complex oxides is presented in Figure 3.1. In consistent with the previous studies our statistical analysis shows that octahedral and tetrahedral are the most dominant motif types present in our dataset.
Figure 3.1 Statistical distribution of the most frequently occurring motif types in binary and ternary complex oxides in the MP database.
3.3 Motif2Vec: Sturtecture motif vectorization

In a recent work,[59] it is shown that the unsupervised learning machines (Atom2Vec) can learn the basic properties of atoms in the periodic table from the extensive chemical formula information of known materials, represented in terms of high-dimensional vectors. The clustering of atoms in vector space classifies them into groups that are consistent with human knowledge of the periodic table. Despite the success of unsupervised learning for identifying the correlations among materials, we should note that the atom vectors extracted from chemical formulas do not contain any structural information. In this work, we will demonstrate that structure motifs offer another layer of essential input by carrying both chemical and structural information of crystalline systems. We focus on binary and ternary metal oxides that constitute a vast and diverse material space where local environments well characterize crystal structures through cation-oxygen coordination.

Next, by using robocrystallography[154] module combined with pymatgen code, we identify the connections between a motif and its neighboring motifs based on the number of oxygen atoms shared by two motifs. Three different types of connectivity may exist, from which we identify the connections as corner-sharing (if only one atom is shared), edge-sharing (if two atoms are shared), and face sharing (if three or more than three atoms are shared). Material properties, such as orbital interactions within a crystal, are known to be related to bond lengths as well as bonding angles. We extract the distance between the
cation center of a motif $M_1$ and its neighboring motif center ($M_2$). The extracted motif connection information will be an essential input for the learning process utilizing a graph convolutional network, as described below. A high-level realization of the Motif2Vec machine is shown in Figure 3.2.

**Figure 3.2** Structure motifs in inorganic compounds and the generation of global motif representations by Motif2Vec using the motif environment matrix.

### 3.4 Construction of motif environment matrix

By using the *pymatgen code* and the method described in Ref. 143, all the motif types associated with cations in a complex oxide are determined; in addition, the connectivity type of each motif site is assigned. The motif environment is defined by the neighboring motifs and the kind of motif connection it has. By iterating over all the compounds in our dataset (22606 complex oxides), all the motif-environment pairs are identified, and a motif
environment matrix \((M)\) is constructed. Each entry \(M_{ij}\) in \(M\) represents the frequency of connection of the \(i\)th motif-type with its \(j\)th motif environment. A schematic of \(M\) for a small set of compounds is shown in Figure 3.3. Each column in \(M\) gives the count of different environments with a single motif type, and each row provides the count for different motif types with a unique environment. Two motif types will behave similarly if their corresponding row vectors are close to each other in the high-dimensional vector space. As each motif type is related to only a small portion of all environments, \(M\) is hugely sparse and contains a very high dimension (our motif matrix contains 4373 rows and 10264 columns).

\begin{verbatim}
<table>
<thead>
<tr>
<th>Motif &amp; its type</th>
<th>(\text{VO}_2)(tbp-corner)</th>
<th>(\text{VO}_2)(tbp-edge)</th>
<th>(\text{VO}_2)(tet-corner)</th>
<th>(\text{MnO}_6)(oct-edge)</th>
<th>(\text{MnO}_6)(oct-corner)</th>
<th>(\text{TeO}_6)(oct-edge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{MnO}_6)(oct)</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(\text{VO}_2)(tbp)</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>(\text{VO}_2)(tet)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(\text{SiO}_4)(tet)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(\text{TeO}_6)(oct)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(\text{FeO}_6)(oct)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(\text{CoO}_6)(oct)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
\end{verbatim}

\textbf{Figure 3.3} An example of the motif-environment matrix \((M)\), a sample dataset of 5 compounds are used to generate this matrix. Each entry in \(M\) represents the frequency of occurrence for the motif-environment pair in the sample dataset.
The motif environment matrix, thus constructed, is analyzed and preprocessed before performing any further analysis. First, the entry $M_{ij}$ which contains the motif type and the environment with the same element as the center of the motif (For example MnO$_6$(oct) as the motif type and MnO$_6$(oct-corner) as an environment) is assigned a zero value as it introduces the unevenness in the dataset. In the next step, to choose the dominant motif types in the material set, we consider the sum of entries of any row in the matrix to be greater than or equal to 20. With this criterion, we obtain 610 most dominant motif types, which are used for the unsupervised learning task. The sum of counts over any row gives the total frequency of occurrence for a motif type with all other environments available, which differs significantly among all motif types. Such unevenness in a frequency distribution is treated by a normalization process, $M_{i,j} = M_{i,j}/(\sum_j M_{i,j}^p)^{\frac{1}{p}}$ on row vectors.

The relative importance between dense and sparse environments can be tuned by using the hyperparameter $p$. Among different choices for $p$, we choose $p = 2$ as it gives the natural distance measure between different vectors.

### 3.4.1 Analysis of motif environment matrix

We aim to identify patterns and clustering information for these high-dimensional motif vectors that, in turn, influence the complex material properties of oxide compounds. By using various linear and non-linear transformations, dimensionality reduction algorithms serve this purpose by creating a low-dimensional embedding that best preserves the overall
variance of the original dataset. To demonstrate the clustering of the motif vectors from Motif2Vec, we visualize the high dimensional data by using the t-SNE (stochastic neighbor embeddings student t-distribution),[155] a recently developed non-linear dimensionality reduction technique. t-SNE models’ similarities in the input space and embedding space as probability densities. In the original, i.e., high-dimensional space, Gaussian distribution is used to measure the similarities. In contrast, in the low dimensional (embedding) space, student’s t-distribution is used to model the similarities.

Figure 3.4 Distribution of singular values obtained from the motif-environment matrix (with rows having sum >=20) in descending order. The flat tail of non-vanishing values follows a sharp decrease for the first ~50 dimensions. It indicates that the higher
dimensions of singular values could also describe the meaningful aspects of motifs. In this work, we choose motif vectors with singular values up to a dimension of 60 for further analysis.

As suggested by L. van der Maaten et al. [155] in the original work, we perform the singular value decomposition (SVD) [156] to project the original high-dimensional representation of materials to low dimensions before performing t-SNE. By using SVD, we reduce the high dimension motif matrix to 60-dimensions, corresponding to the most significant 60 singular values obtained by using the scree plot in Figure 3.4. We also test the performance of principal component analysis (PCA) [157], which is a widely used dimensionality reduction technique through linear projections of high dimensional data to fewer dimensions. Even though the PCA plot in Figure 3.5 shows a certain degree of clustering of some motifs based on their types and the group of cation elements in the periodic table, the clustering trend is less clear. We expect the use of a non-linear projection technique, such as t-SNE, is more useful to identify a piece of more substantial cluster information.
Figure 3.5 Projection of motif environment matrix spanned by the first two principal components (a) and third and fourth principal components (b). The coloring resembles the obtained similarities as encoded in the periodic table (c).

Figure 3.6 shows the projected motif vector data in two dimensions (2D) obtained through the t-SNE process. Different motif types are represented by various colors, reflecting the
preserved similarity information among motifs in 2D space. First, it is clear when projected to 2D that the unsupervised learning by Motif2Vec creates a high dimensional representation of motif types in a similar way the elements are classified in the periodic table. Elements having similar chemical properties with different motif types are clustered together. For instance, all the lanthanide-based motifs are grouped based on metal species and the variety of motifs they are associated with. It is interesting to see Yttrium-based motifs always stay close to the Lanthanide-based ones, as the chemical properties of Yttrium are known to be similar to Lanthanides. Besides, the same types of motifs associated with Zn and Mg always cluster together, which is consistent with the fact that Zn is chemically similar to Mg in that both of them exhibits only one normal oxidation state (+2) and their ions (Zn$^{2+}$, Mg$^{2+}$) are identical in size.

As shown in Figure 3.6, cluster 1 contains cubic motifs associated with the Lanthadnides, while the cuboctahedral motifs associated primarily with main group elements appear in cluster 2. The clustering of motifs determined by elements, as discussed above, is per the grouping pattern in the periodic table, although no information about the periodic table was used in this process. It is interesting to see that octahedral motifs associated mostly with the transition metal elements occur together in cluster 3. In contrast, the tetrahedral and square planer motifs associated with transition metal elements are grouped but well separated in cluster 4. This separation of various clusters of motifs demonstrates the power of Motif2Vec to capture local structural information as well as elemental information. These findings achieved by unsupervised learning strongly support our intuition that
structure motifs can serve as essential fingerprints for crystalline compounds that carry both elemental and structural information.

**Figure 3.6** t-SNE projection of motif vectors constructed by using the motif environment matrix. The numbered clusters from 1 to 4 are associated with motif types, including (1) cube, (2) cubooctahedral, (3) octahedral, and (4) a mixture of tetrahedral (in magenta) and square planar (in remnant).
3.5 Incorporation of motif information in graph neural networks

Next, we will demonstrate the incorporation of the structure motif information as an essential layer of input in a graph neural network architecture. Following the notations presented in the graph neural network (GNN) framework,[158] we represent an attributed graph as $G = (V, E)$, where $V = \{v_i\}_{i=1,2,...,N_v}$ is a set of nodes of cardinality $N_v$, $v_i$ is the node attribute vector of the $i$th node. $E = \{(e_k, r_k, s_k)\}_{k=1,2,...,N_e}$ is a set of edges of cardinality $N_e$, $e_k$ is the attribute vector for edge $k$ between nodes $s_k$ and $r_k$. Several graph neural networks[63]–[65] have been proposed that formulate the task of predicting chemical properties of materials as learning a mapping $f(G; W) \rightarrow y$, where $W$ is a set of learnable parameters and $y$ is a target property.

Most of the graph networks applied to crystalline materials[64], [134], [135] are based on graphs on the atomic level $G_0^{atom}$ as input for the network. Such atomic graphs contain information about atoms (such as atomic number, electronegativity, and many others) and their bonds. For instance, in the $G_0^{atom}$ of atomic graph network MEGNet,[159] $v_i$ is a vector representing the $i$th atom in a unit cell and is represented by the atomic number of the element, $e_{ij}$ is a vector representing a bond between atom $i$ and atom $j$.

As illustrated above, structure motifs are the above-atomic-level building blocks of crystals, and motif-wise interactions within a crystal strongly influence the material property.
properties as well. In this work, we propose that atom-motif dual graph networks (AMDNet) can be constructed to accelerate the learning process and improve the prediction accuracy for the electronic structure properties of metal oxides.

We follow the procedure introduced in existing atomic graph networks[63], [64], by representing the edges using the spatial distance, expanded with Gaussian base $\exp[-4(r-r_0)^2]$ centered at 100 points between 0 and the maximum edge distance. Two atoms are connected by an edge when they are no more than 5 Å apart, and we consider all possible bonds in a periodic crystal structure when constructing the graph. We propose to represent the complex oxides as motif graphs, where a node represents each motif in a crystal, and each connection between two motifs is represented by an edge, as shown in Figure 3.7. Motif graphs represent the same materials as in atom graphs, with higher granularity than atom graphs, but more comprehensive information is encoded in each node. In a motif graph, we use the same edge representation as an atom graph, with the exception that the maximum distance is set as 11 Å (because the motif-motif connections are measured from the center atom of one motif to the next). In the motif graph, a combination of atom-level and motif-level information is encoded in each node. We adopt the atom-level node representations by combining two existing approaches to form a 103-dimensional vector that uses the information of atoms within the motif. The first 86 dimensions represent the fractional encoding of the atoms proposed by Meredig et al.,[119] and the next 17 dimension is for physical properties proposed by Ward et al.[49]. On the other hand, we define the motif fingerprints by order parameters (of dimension 61), which
describes the numerical measure of the local environment around an atom relative to a standard target motif.\cite{149,160}

Figure 3.7 Construction of a motif graph based on both atom-level and motif-level information encoded in an inorganic crystal

Detailed descriptions about various types of order parameters and methods to compute such parameters are presented in work by Zimmermann et al.\cite{161} All the structural information used to construct the motif graph, including extended connectivity, angle, distance, and order parameters for each motif, are computed by using the python package robocrystallography combined with the pymatgen code. By combining atomic-level and
motif-level information, we utilize a 164-dimensional vector to represent each motif in the graph.

3.6 AMDNet: Atom-motif dual graph neural network

A high-level illustration of our proposed atom-motif dual graph neural network (AMDNet) architecture is shown in Figure 3.8. To incorporate the motif information acquired above into the graph network learning framework, the central concept in the proposed architecture is to generate both motif graphs and atom graphs representing the same compounds, with different cardinality of edges and nodes, and combine the representation information before making predictions.

For each material, we generate an atom graph and a motif graph (Figure 3.8). We adopt the convolution structure of the MEGNet proposed by Chen et al.[64] when constructing the atom-level graph network. The choice of graph network structure is only for a benchmark purpose, and many other types of crystal graph convolution networks could be used to take advantage of the motif-level graph information.[134], [135] As a preliminary test, we use the same architecture as that for the atom graphs in MEGNet to generate $\mathcal{G}_0^{\text{motif}}$ by utilizing the 164-dimensional atom-motif-mixed vector input for the nodes in the network.
Figure 3.8 (a) Demonstration of the learning architecture of the proposed atom-motif dual graph network (AMDNet) for the effective learning of electronic structures and other material properties of inorganic crystalline materials. (b) Comparison of predicted and actual band gap values (from DFT calculations) and (c) comparison of predicted and actual formation energy (from DFT calculations) in the test dataset with 4515 compounds.

Note that MEGNet can be interpreted as a neural network that encodes the whole crystal graph input to a low dimensional vector of dimension 16, upon which a final single-value prediction is made. Taking advantage of this fixed-dimension representation of any
MEGNet graph convolution network, we can effectively combine the information from motif and atom dual graphs by concatenating the two low-dimensional representations generated from the motif graph and atom graph, respectively. This concatenated vector is then fed to a small feed-forward neural network for single-value predictions.

3.7 The Training process for atom-motif dual graph neural network

In the AMDNet with $L$ layers, the module generates a sequence of atomic graph representation \{ $G_1^{atom}$, $G_2^{atom}$, ..., $G_L^{atom}$ \} and motif graph representations \{ $G_1^{motif}$, $G_2^{motif}$, ..., $G_L^{motif}$ \}, where each graph has the same number of nodes and edges as in the input graphs, $G_0^{atom}$ and $G_0^{motif}$, respectively. Through a graph convolutional process called AtomNet Block for atom graphs and MotifNet Block for motif graphs, information of each edge and its respective connecting nodes are passed through a dense neural network with a non-linear activation function (we use the shifted softplus function), which creates a new edge representation. To generate the new node representation, the node information together with the knowledge of the latest incident edges is passed through a separate dense neural network with the same non-linear activation function. Each graph convolutional block has a hidden dimension of 64 for both node and edge convolution. In our work, we use three graph convolutional blocks to apply the graph convolution, which creates an output graph representation. The graph representation is transformed into vector form by averaging over all nodes and edges, respectively, which is denoted as set2set(E).
and set2set(V) in Figure 3.8(a). These set2set vectors are concatenated before going through two densely connected layers, as shown in Figure 3.8(a). This results in a low-dimensional vector representation of the original atom and motif graph representation of the crystal. These representations are concatenated again and passed through two densely connected layers to make a single real-valued prediction. For the training and test process, we choose a 60-20-20 train-validation-test split. Figure 3.8(c) shows the mean square error (MSE) on the training and validation set during the training process. The same hyperparameters that resulted in the best prediction performance in MEGNet are used to train our neural network. All deep models are trained with ADAM optimizer with an initial learning rate $\alpha = 0.001$. Training formation energy prediction was slower to converge, with the same parameters for the band gap prediction task. To make the training process faster, we adjusted some parameters. We stop training when the validation error doesn’t improve for 20 and 100 epochs to train band gap prediction and formation energy prediction, respectively. We save the model with the lowest observed validation error and use it to evaluate the models on the test data. We use 64 compounds per minibatch for band gap prediction and 32 compounds per minibatch for formation energy prediction.
3.8 Performance comparison between different graph neural network architectures for complex oxides

We use 22606 complex oxides from the MP database to evaluate the efficiency of our proposed model for the prediction of band gaps, which is one of the complex electronic structure problems. Complex oxides are a class of solid-state compounds that are challenging for both \textit{ab initio} quantum simulations and machine learning in general. Chen \textit{et al.}'s work focused on the compounds with the non-zero band gap for the band gap prediction task, while our focus is complex oxides, including both metals and non-metals. This involves utilizing significantly different material datasets, and a direct comparison between our proposed model with the published results from Chen \textit{et al.}'s work for the band gap prediction task is not possible. To ease the comparison, we split the material dataset that contains the band gap information used in Chen \textit{et al.}'s work into three parts: non-metallic non-oxides (26455), non-metallic oxides (15434) and metallic oxides (7172). The non-metal oxides are the overlap data between ours and Chen \textit{et al.}'s work. We generate the train-validation-test splits into each of those groups, which allows us to compare performance on a different subset of the data. For the comparison purpose, we create a motif graph network model, MNet, which uses motif graphs ($G^\text{motif}_0$) as the only input to the network.
Table 3.1 shows a comparison between MEGNet, MNet, and our proposed AMDNet on the prediction accuracy of band gaps, metal vs. non-metal classification, and the formation energy for all complex oxides used in this work. Given the same test and training data, AMDNet shows its superiority over the state-of-the-art baseline model, in the band gap prediction task. MNet performs worse than MEGNet, which is expected as it uses much smaller graph representations. AMDNet outperforms MEGNet on the band gap prediction task, which illustrates that motif representation enhances the effective learning of material properties. We investigated metal (with band gap less than and equal to 0.2) vs. non-metal (with a band gap greater than 0.2) classification and found that our model achieves an accuracy of 82.1%. In comparison, MEGNet only achieves 75.3% accuracy.

On the other hand, MEGNet and AMDNet show an identical performance in the formation energy prediction task. This indicates that atom level representation is sufficient for the formation energy prediction task, a more straightforward material problem than the band gap prediction task. Figure 3.8(b) shows the comparison of the predicted band gaps on the test datasets with the actual band gaps. A large portion of the compounds’ band gap is clustered close to the diagonal, indicating an excellent performance of our model on the band gap prediction task. The fact that the combination of atom and motif graphs outperforms MEGNet on the band gap prediction task in the oxides illustrates the additional material information encoded by motif representations, which in turn enhances the effective learning of material properties.
Table 3.1 Performance comparison between various graph architectures for the learning and prediction of electronic band gaps, formation energy, and metal vs. non-metal classification for complex oxides (trained on 18,091 compounds and tested on 4,515 compounds).

<table>
<thead>
<tr>
<th>Model</th>
<th>Band gap MAE (eV)</th>
<th>Formation energy MAE (eV/atom)</th>
<th>Metals vs. non-metal classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEGNet (atom graph)</td>
<td>0.542</td>
<td>0.0469</td>
<td>75.3%</td>
</tr>
<tr>
<td>MNet (motif graph)</td>
<td>0.639</td>
<td>0.1214</td>
<td>74.7%</td>
</tr>
<tr>
<td>AMDNet (dual graph)</td>
<td>0.443</td>
<td>0.0470</td>
<td>82.1%</td>
</tr>
</tbody>
</table>

3.9 Summary and discussion

We demonstrate in this work how structure motifs in crystal structures can be combined with both unsupervised and supervised machine learning techniques to enhance the effective representation of solid-state material systems. As a step forward from Atom2Vec to Motif2Vec, motif vectors learned from motif environments in 22606 complex metal oxides using unsupervised learning effectively capture the motif similarities and their
clustering properties. To enhance the learning of solid-state crystalline systems for complex electronic structures, structure motif and connection information are incorporated as an essential input in a dual graph neural network model (AMDNet), which outperforms the state-of-the-art atom graph neural network model for the prediction of electronic band gaps and metal vs. non-metal classification task. Besides, the AMDNet model is able to predict the formation energy in close agreement with the existing state-of-the-art atom graph-based models. AMDNet is the general learning framework for solid-state atomistic systems that can be used to predict many other materials properties. Several directions related to the motif-centric learning methods are worthy of exploring in the future. Although we perform the test on perfect crystalline systems, through the addition of special types of local motif information, the motif-enhanced graph network framework can be expanded for the learning and prediction of surface and defected material systems. Besides the use of a dual graph network architecture, motif information and the physical principles behind it can be incorporated into a learning framework in other manners, such as through a motif-enhanced convolutional process in an atom-based graph convolutional network or different novel algorithms that are actively developing in the graph theory including graph attention. Moving one step forward to demonstrate the useful applications of motif-based representation in a data-driven material design framework, we will discuss the construction of motif-based material networks in the next chapter.
4.1 Introduction

First-principles computations based on DFT has significantly accelerated the discovery and design of novel materials. Hundreds of thousands of structures with their fundamental material properties have been reported and listed in various databases, as described in the introduction and Chapter 3 of this dissertation. However, only a tiny fraction of these materials has been studied in detail. Due to the limitation in resources, it is not feasible to synthesize and characterize each of them experimentally or compute their detailed properties accurately by using state-of-the-art computational tools. Identifying materials with desired physical and chemical properties from such a large dataset is still an outstanding challenge.

To find a way to address such a challenge, many knowledge-based structure-property relationships have been proposed to correlate materials properties with straightforward structural information and to build connections among a broad set of materials. In Chapter 3, we presented our efforts in using machine learning techniques to help alleviate this challenge by introducing structure motifs as an essential tier of material information
encoded in crystal structures. The ability to visualize the materials space via connections to discover the underlying pattern in the materials data and reduce the search space of materials for any application is expected to close the gap between computational and experimental efforts in the material discovery process.

There have been several pioneering works to define the similarities among a large number of inorganic crystals. For instance, Isayev et al.[162] measured the similarity between inorganic compounds by converting their electronic structure diagrams into numerical descriptors and constructed a material cartogram. Aykol et al.[163] used the dynamics of the material stability network to discover new materials that are closely connected with those already known. The above work paved the way for the evaluation of the similarity or dissimilarity of crystals and represented them in terms of graph-based material networks.

Graphs have been a potent tool in analyzing complex relational data.[164] Graphs can be used to represent diverse relations present in the physical, biological, and information systems. As described in Chapter 3, mathematically stating a graph G(V, E) can be defined as a pair (V, E), where V is the set of vertices representing the node, and E is set of edges representing the connection between nodes. The existence of an edge E = {(u, v) | u, v ∈ V}) between two nodes u and v determine either they are neighbors or not. In a directed graph, the relation between two nodes (i.e., E) is non-symmetric, while in the undirected graph, the relation E is symmetric. In many physical and real-world applications, it is
desired that the edges of the graph support some weights, i.e., real numbers signifying a particular property of the edge. In such a case, the graph is called a weighted graph. The effectiveness of graphs lies in its capability to model many diverse circumstances.

In a real-world application, the term “network” is used to define a graph in which various attributes can be incorporated at the nodes and edges. In this sense, a network is a graph with details node and edge information encoded by a given application. Networks are used in a variety of disciplines such as computer science, physics, social science, and many others because they not only capture the relationships but also support the inferences. Studying the properties of networks help to solve problems related to the various complex phenomenon that lies around us. Some networks such as social networks,[165] flight networks,[166] and roadmap networks[167] have been a part of human life. A simple example of a network can be considered as the connection between two cities, such as New York and Washington, D.C.. These two cities act as nodes, while all the possible connections between these cities are considered as links. Understanding networks - especially the bigger ones and obtaining meaningful information from them is itself a challenging task. To understand complex networks, recent trends have been focused on studying mediators, hubs, and interconnected structures.

By combining graph theory and the motif-based similarity searching algorithms, we present a unique way of categorization, evaluation, and visualization of the complex oxides
present in the Material Project database (MP). Compounds in our explored material space act as a node of the network, while the edges between two compounds carry information in the form of their similarity counted as the weight of the link. As demonstrated in chapter 3, properties of a compound depend on the nature of the atomic arrangement, types of the coordination environment (motifs), and the number of atoms shared by two motifs in a compound. We use all the structural details and the connection types with the neighboring motifs as a feature vector to represent a motif site in a compound. Apart from global information, these descriptors carry easily interpretable local details of the crystals. In this work, these feature vectors are used to compute the similarity between different motifs in two compounds, which are then combined to obtain a similarity value between two compounds. Here we used the Tanimoto coefficient (TC)[168] as a similarity measure to quantify the similarity between different compounds. The detailed procedure of similarity measure between two compounds is presented in the upcoming section.

In our purposed network, the edges do not have any directions, i.e., there is a symmetric relationship between two compounds. By using the symmetric and undirected relationship between two compounds acting as nodes, we are able to find their sub-communities and the compounds which are essential to those communities. In general, the network has a topology that could be centralized or decentralized, dense or sparse, cyclic or linear. In many cases, network shape and its basic properties mainly determine the reasonable way of analyzing a network. The prior information about the number of nodes, edges, and the structure of the network is critical in the network analyzing process. Also, other details on
the networks, such as, the arrangement of the clusters and their complexity, are crucial to understanding the networks. Various quantitative metrics such as network density, degree, eigenvector centrality, betweenness centrality are useful in differentiating the networks, in learning their topologies, and are very useful to unfold the mixture of nodes and edges into something we can learn.

4.2 Construction and analysis of motif-based material network

Here we present our preliminary analysis of a network obtained for 7000 complex oxides (shown in Fig. 4.3) chosen randomly from the Materials Project database. This dataset contains ~10% compounds from the MP databases and ~30% of the complex oxides in the MP database. Our preliminary result indicates that the proposed network is very promising to identify the clusters of materials based on their similar properties,

4.2.1 Finding similarities between compounds

We begin our screening by identifying a shared motif between any two compounds in our dataset. If they share at least one motif, then those compounds are considered for further consideration. As each compound is composed of motifs, the site fingerprints[150] vector for each motif carries information about the coordination environment and perfection of the motif, i.e., the local structure order parameters. As stated in Pauling’s third law, the
surrounding information and the connectivity of a motif with the next nearest motif is also a critical factor determining the properties of a compound. A large number of atoms shared by two different cations are usually associated with the low stability of a compound. To construct the surrounding feature vector, we identified the connection between a motif and its neighboring motifs. As described in Chapter 3, the connection type between a motif and its surrounding is identified as corner, edge, and face. The surrounding feature vector is of length three contains the frequency of connection type a motif has with its neighboring motif. To make this uniform with the site feature vector, it is normalized. In the next step, the site and the environment feature vectors are concatenated to form a single feature vector of length 64 for each site. This is a well-accepted way to obtain a single feature vector by using many feature vectors in data-driven guided studies. Identifying the better measure to quantify the similarity between two compounds is itself a challenging task. Bajusz et al.[168] and the references therein made the detailed comparison of various similarity metrics in cheminformatics and related fields. These studies conclude the Tanimoto coefficient as the best measure to quantify the molecular similarity in cheminformatics. Isayev et al.[162] used the Tanimoto coefficient to quantify the similarity measure between two crystals by using their band structure and density of states fingerprints. By borrowing the idea from these works, the Tanimoto coefficient is used to quantify the similarity measure between two compounds. Also, we compare the performance of various similarity measures like Euclidean distance (similarity = \(\exp(-\text{distance})\) ) and the square root of the dot product of two vectors, as described by N. Zimmermann et al.[150] After all, we decided to use the Tanimoto coefficient as it quantifies the similarity reasonably compared
to others in our dataset. By iterating over all sites, the Tanimoto coefficient is calculated, and their average is taken to obtain a similarity value between different sites in two compounds. The average of such similarity values is retaken to get the final similarity value. This measured similarity value acts as a weight between two nodes. In this way, we represent all compounds in our dataset in a graph $G (V, E)$, where $V$ represents the vertex of the network (a compound), and an edge $E$ exists between two compounds with a weight equal to the value of similarity between two nodes. Edges are weighted based on the similarity value; if the similarity is higher than 0.75, then the weight of 1 is assigned, if it is between 0.3 and 0.7 weight of 0.5 is assigned while all other values are not considered for the connection purpose. These weights determine the density of the clusters observed in the network. All the network properties are calculated by using a networkx package[169], and the visualization is done by using Gephi.[170] A force-directed layout (force atlas 2 algorithm)[171] is used for network stabilization. In force-directed layout, nodes repel each other like magnets, while edge attracts their nodes like springs, with a spring constant equal to the assigned similarity value for the edges. These forces create a movement that converges to a balanced state. This final configuration is presented for visualization purposes (in Figure 4.8) in this work.

4.2.2 Network density
To analyze our network, we begin with the network density, a good metric to start with network analysis. A network density is simply the ratio of actual edges in the network to all possible edges in the network. Network density is measured in the range from 0 to 1, with a density close to 0 means that there are no connections at all, and a 1 would indicate that all possible edges are present, i.e., a network is perfectly connected. Our network measures the density of 0.0336 (on 7000 compounds), which is on the lower end of the scale 0 to 1, but it is still far from 0.

4.2.3 Degree

![Graph showing degree distribution with number of nodes on the y-axis and degree on the x-axis.](image-url)
**Figure 4.1** Degree distribution in our network. The red line represents the power-law fitting. Unlike other networks such as social or real-world networks, it does not follow the power law.

The degree distribution is a measure of the topology of the network. In our network, a degree of a node corresponds to the number of nearest neighbors a node has. Figure 4.1 shows the degree distribution in our network. The degree distribution indicates the biggest hubs (node with the highest degree) present in the network. The biggest hub has a degree of 913 for compound Li$_4$SiO$_4$ (mp-1223129) in our network. The connectivity of this compound with many others is due to the different motifs (LiO$_4$, LiO$_6$, LiO$_5$, and SiO$_4$) present in this compound, which is presented as the first criterion in our similarity finding algorithm.

### 4.2.4 Eigenvector centrality

This measure can be viewed as an extension to the degree, which is not able to discern additional information about the rest of the nodes. Centrality measure can give us more information than just hubs, by combining the information about a node’s edge and the edges of that node’s neighbors. Higher the eigenvector centrality ranks the nodes that are connected to the important neighbors. Like network density, eigenvector centrality is measured in range 0 to 1. The value close to 1 indicates the higher eigenvector centrality.
This measure is very useful to identify a node (as a compound) that can pass information to many other nodes quickly. In our material network, the compound Mn₄P₂O₉ (mp-770499) and ~10 other compounds composed of Mn, P, and O have eigenvector centrality of 0.0481.

4.2.5 Betweenness centrality (BC)

Betweenness centrality is used to represent the number of edges, any one node or set of nodes has. BC can be defined as the percentage of the shortest path that passes through a given node. BC is measured on a scale of 0 to 1, it is good at finding nodes that connect two otherwise disparate parts of a network and monitoring the flow of information or other things through networks. If there is a single node connecting two clusters, then all the possible communication in those separated clusters pass through that node. In contrast to a hub, this sort of node is often referred to as a broker as it connects two disparate, unknown clusters of the networks. Betweenness centrality is the only measure that is able to identify such nodes, and it gives a sense of which nodes are important not because they have lots of connections themselves but because they stand between groups, providing the network connectivity and cohesion. In short, betweenness centrality measure the degree to which a node is essential in the network. It is likely to have nodes that have a high degree and high betweenness centrality. However, nodes that have a low degree can have a high betweenness centrality measure as well.
In our material network, this means that compounds having higher betweenness centrality shares common properties or motifs with two disparate parts of the graph. We identified that the compound Si₅P₆O₂₅ (mp-3273) (composed of unique motifs SiO₆, SiO₄, and PO₄) with a degree of 833 has the highest betweenness centrality value of 0.003. The compound Zr₄N₂O₅ (mp-758272) has the second-largest betweenness centrality of 0.0028. The compound with the highest degree, Li₄SiO₄ (mp-1223129), has a betweenness centrality of 0.0026, ranked fourth in terms of betweenness centrality. This set of network measure analysis indicates that, in our material network, a compound with a higher degree does not mean that it has a higher importance in the network.

4.2.6 Matching index

While obtaining a node similarity in a graph, a matching index can be a handy measure. This measure shows how similar two nodes are within the network. Two functionally similar nodes do not always have to be connected. The matching index $M_{uv}$ measures the “similarity” of two nodes based on the number of common neighbors shared by nodes $u$ and $v$. [172] The two nodes are considered to be more similar if the matching index is higher. Figure 4.2 shows how the matching index between two nodes is computed. In our material network, this matching index brings an effective way to identify indirect connections between two compounds. For example, two compounds KPO₃ (mp-1196581) and U(PO₃)₄ (mp-1202567) are not connected directly in our material network graph in
Figure 4.8, but the matching index between them is 0.74, indicating a higher similarity of the neighbors of these two compounds. While compounds Li_{4}SiO_{4} (mp-1223129) and Si_{5}P_{6}O_{25} (mp-3273) has a matching index of 0.24, indicating less similarity in their environments and the compounds themselves.

Figure 4.2 \( N_{1} \) is connected with 5 nodes (\( N_{3}, N_{4}, N_{6}, N_{7}, N_{8} \)). \( N_{2} \) is connected with 4 nodes (\( N_{3}, N_{4}, N_{5}, N_{8} \)). There are three common nodes for \( N_{1} \) and \( N_{2} \) (\( N_{4}, N_{8}, N_{3} \)) and in total 6 distinct neighbors (\( N_{3}, N_{4}, N_{8}, N_{5}, N_{6}, N_{7} \)). The matching index between nodes 1 and 2 is \( M_{12} = 3/6 = 0.5 \), which indicates that \( N_{1} \) and \( N_{2} \) are functionally similar even though they are not connected. The method of computing a matching index is adopted from Ref. 167.

4.3 Material network for transparent conducting oxide materials

Because of their high electrical conductivity and optical transparency, transparent conducting materials (TCMs)[147], [173]–[177] are an essential class of materials from
both a fundamental point of view and for diverse applications such as solar cells, flat panel displays, touchscreen, and many others. Among various types of TCMs, here we devoted ourselves to the identification of transparent conducting oxide (TCO) materials, which are the most studied and widely used TCMs. The identification and development of an efficient TCO material is a very challenging task. Mostly the discoveries are made either by chemical intuition or by laborious experimental efforts.[175] In comparison to \( n \)-type TCO, identification of high mobility \( p \)-type TCO is more challenging because of the localized oxygen \( p \)-states at the VBM. This localization leads to flat valence bands and hence large hole effective mass, which results in low hole mobility. Significant efforts have been focused on the development of such \( p \)-type TCO materials.

With the available knowledge of TCO, we screened complex oxides from the MP database by using our purposed material network. As an initial proof of concept, we complied 108 TCO materials from previous studies.[173]–[176], [178] As data processing is very crucial in a quantitative materials structure-property relationship study, in the beginning, we obtain a small network of these TCO, as shown in Figure 4.3 (all the connected and isolated nodes are presented).

The size of the nodes in the network (in Figure 4.3) represents its importance in the network defined by its betweenness centrality. We observe the clustering of the compounds with higher similarity in their motifs. For example, the compounds in the cluster marked with a
black circle (top right) share TiO₅ as the common structural motif, and the similarity coefficient between those compounds is high (TC = 0.98). At the same time, this small cluster is connected to the rest of the network through the compound MgIn₂O₄ (mp-7831), which shares the MgO₄ motif with the compound MgTiO₂ (mvc-458). The computed value of the similarity coefficient between these two compounds is low (TC = 0.3), and hence they form a well-separated cluster. The SnO₆ octahedral motif well connectes the compounds marked with a green circle. Within each group, the measured similarity between compounds is very high (TC > 0.9), indicating the similarity in their motifs or vice versa.

The compound Cd₃TeO₆ (mp-14243) has the highest betweenness centrality (shown by the biggest node in the network in Fig. 4.3), as it connects various clusters in the network. It has three different types of motifs: CdO₇ square face capped trigonal prism, CdO₆ octahedral, and TeO₆ octahedral. Multiple groups of compounds sharing similar motifs are shown in Figure 4.3. The frequency of occurrence of the motifs and their types in these TCO is shown in Figure 4.4 and Figure 4.5, respectively. Most of these transparent conducting oxides are composed of octahedral and tetrahedral motifs. This is consistent with the statistical analysis of motif types present in 7000 complex oxides shown in Figure 4.7. Detailed descriptions about the identification of motifs and their types are presented in Chapter 3 of this dissertation.
Figure 4.3 A material network of known 108 TCOs showing densely connected clusters at the center and various isolated nodes around. The marked groups and the motifs they share are presented (the coloring in the labels indicates their presence in the marked clusters with respective colors). The size of the nodes indicates their betweenness centrality.
Figure 4.4 Distribution of most occurring motifs in the TCO dataset used in this work.

Figure 4.5 Distribution of motif types in 108 TCOs.
Figure 4.6 Distribution of most occurring motifs in the dataset consisting of 7000 complex oxides. Only those motifs with a frequency higher than 1000 are displayed.

Figure 4.7 Distribution of motif types in the dataset with 7000 complex oxides. TP stands for “trigonal prism,” TSF for “three square faces,” and WTF for “with triangular faces.”
With this prior knowledge, in the next step, we mapped the known TCOs in the big graph of 7000 complex oxides, which are represented by the red colored nodes in Figure 4.8. The key idea is to identify nearest and next-nearest neighbors of these TCOs in the graph network, from which a set of compounds are obtained sharing the similar structure motifs with the known TCOs. These neighbors can be analyzed in further detail by comparing their electronic and optical properties with known TCOs. Thus, we can reduce the compound search space from hundreds of thousands to few, which is much more accessible for further screening with higher-level computational methods. As a test case for this mapping, we choose a compound GeO$_2$ (mp-19012), a known transparent conducting oxide in Ref. 161-171. The extracted subgraph with the near neighbor connection of this compound (in green) is shown in Figure 4.9. This compound is composed of GeO$_6$ motifs and has a total of 55 compounds as its neighbors sharing GeO$_6$ as a common motif.

These 55 compounds we identified as the neighbors of GeO$_2$ (mp-10913) (the green node in Fig. 4.9) are composed of a mixture of Ge with another element such as Cd, Zn, In or Ga, whose binary oxide is a good TCO as noted in the high-throughput computational analysis by Hautier et al.[177] This new list of compounds provides a starting ground for the search of novel TCO that share similar properties with GeO$_2$ (mp-10913). We compared these compounds with the ones listed in Ref. 161-171., and “rediscovered” the compounds listed in Table 4.1, which are already known as TCOs. Identification of these previously known TCO materials boosts our confidence in the data-driven materials discovery approaches by using the motif-based material network we proposed. One can
map other known TCOs and identify a different set of compounds for the further screening process. As a next step, the rest of the compounds from the list can be studied by computing their relevant properties to identify promising transparent conducting oxides. The screening for TCO can be made simpler by identifying the nodes with specific band gap values, as shown in Figure 4.9 (b).

**Table 4.1** Compounds (with their mp-id) identified as the neighbors of GeO$_2$ (mp-10913) (in green in Fig. 4.4) in our material network are listed as TCO in the literature. The unique motif types present in the compounds are presented.

<table>
<thead>
<tr>
<th>Compounds</th>
<th>Unique motif presented in the compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeO$_2$ (mp-10913)</td>
<td>GeO$_6$</td>
</tr>
<tr>
<td>CdGeO$_3$ (mp-13003)</td>
<td>CdO$_8$, GeO$_4$, GeO$_6$</td>
</tr>
<tr>
<td>Cd$_2$Ge$<em>7$O$</em>{16}$ (mp-29213)</td>
<td>CdO$_8$, GeO$_4$, GeO$_6$</td>
</tr>
<tr>
<td>CdGeO$_3$ (mp-8275)</td>
<td>CdO$_6$, GeO$_6$</td>
</tr>
<tr>
<td>ZnGeO$_3$ (mp-8285)</td>
<td>ZnO$_6$, GeO$_6$</td>
</tr>
</tbody>
</table>
Figure 4.8 Representation of 7000 complex oxides in a material network. Red nodes are the TCO mapped in the material network.
Figure 4.9 (a) Extracted sub-graph for the compound GeO$_2$ (mp-10913) from the big graph. (b) Band gap distribution in the subgraph (Color: $E_g$ $\leq$ 0.3 purple nodes, 0.3 $< E_g$ $< 1.5$ blue nodes, and $E_g$ $\geq$ 1.5 orange nodes). This coloring scheme can be more helpful in screening compounds based on the band gap from the subgraph. The star-like topology of this subnetwork indicates the subnetwork is centralized, and all the compounds have similar connectivity.

We would like to emphasize that this material network is very flexible, as we can use material properties such as formation energy, energy above-hull, and many others as the node attributes. For example, if one wants to identify the stable compounds by using a known stable one, one can use the energy above-hull as the node attribute. This flexibility of the network will help in screening compounds with a desirable property and thus ease the material screening process. As the identification of $p$-type TCO is more challenging,
our proposed material network can be used for the initial screening purpose by using a known *p*-type TCO material as the starting node.

### 4.4 Summary

Driven with the hypothesis that the properties of materials are mostly (but not entirely) determined by their structural details and the materials with similar structures are in some way identical in their physical and chemical properties, we propose a novel approach to categorize and visualize the material space by using the local structural information (structure motifs) of the crystals. This method provides a good starting ground to explore different material systems and measure material similarities. Moreover, this study complements previous efforts in the high-throughput material discovery process by reducing the material search space dramatically. The intuitions acquired from the network visualization can assist in discovering patterns from a large pool of material space, which is almost impossible to examine separately and essential for the discovery and design of novel materials. We showed the case of TCOs as a benchmark of the role of our material network in a material design framework. Beyond our test for the TCOs, this method can possibly be useful to study additional material properties for the examination of novel functional materials for various applications.
CHAPTER 5
SUMMARY AND CONCLUSION

Structure-property correlation is the central research topic for both condensed matter physics and materials science at different scales. Increasing accessibility of computational resources made chemists, physicists, and material scientists to focus their efforts more and more on automatized tools for material analysis and machine learning to help accelerate the material discovery process. These techniques are helpful for the study of a massive amount of material data efficiently and added a footstep towards the accelerated progress in material science. Simple statistical models such as regression and classification, have shown their success in the prediction of fundamental properties, such as electronic band gaps, formation energies, and crystalline structures. In this thesis, DFT-based first-principles computations are combined with machine learning to enable a paradigm shift towards the data-driven discovery and design of inorganic materials, including 2D functional materials and complex oxides.

Employing the high-throughput computations based on DFT, we identify a set of 2D semiconducting tellurides possessing high carrier mobilities and tunable band gap from 1.0 eV to 1.3 eV. Detailed structural analysis shows the presence of a unique structure motif metal--centered bicapped trigonal prism - in these compounds. As a ubiquitous feature in many group V chalcogenides, this common structure motif is correlated with many
similarities in terms of electronic properties. These compounds are promising for high-speed electronics applications. Among these 2D tellurides, two compounds Ti$_2$Si$_2$Te$_8$ and Zr$_2$Si$_2$Te$_8$ are most promising for their band gaps and carrier mobilities. These compounds can be a good starting point in search of novel 2D materials for optoelectronic applications. Besides, their mechanical properties can be studied further to determine their breaking strength, which might open a gateway for flexible electronics. The computational identification of these 2D telluride compounds provides a novel material platform for experiments in the search for functional materials that enable future 2D electronic devices.

With the motivation to assist the shortcomings of DFT and explain the success of ML models in the prediction of material property, we dedicated ourselves to identifying the new representation for the crystals. By moving one step forward from the existing Atom2Vec learning framework, we proposed a novel Motif2Vec learning algorithm, which is able to learn the motif and its environments present in 22606 complex oxides present in the Material Project database. Using a clustering algorithm, Motif2Vec demonstrates its ability to capture the similarity in motif types and the atoms they are associated with, in the form of clustering information. In the next step, by incorporating detailed information about the structural motifs and their environments, we proposed the motif graph representation for complex oxides. By combining motif graphs with existing atom graphs, we introduced the atom-motif dual graph network (AMDNet). Using information from both atomic and motif levels, the AMDNet is able to predict electronic structure related properties such as band gaps with a higher level of accuracy than the existing atom-level
graph representation. This work demonstrates the success of graph neural networks learning architecture and its usefulness in determining the complex electronic structure problem in condensed matter physics and material science.

The increasing gap between the demand for materials and the experimental discovery is due to the cost and time-consuming procedure of experimental synthesis. High-throughput (HT) computational material design and available databases have been very successful in alleviating this gap. However, the HT approach needs lots of computational efforts, and one needs to interrogate the database intellectually in search of materials with desired properties. With the motivation to close this gap, we proposed the motif-based material network, which can be used to build connections among materials and accelerate the material screening process for any application by reducing the material search space significantly. The proposed material network is very flexible in terms of the encoded material properties at the nodes, which makes it applicable for other domain problems in condensed matter physics and material science. Our preliminary success in the screening of transparent conducting oxides boosts our confidence in the use of the proposed material network in other applications.

To conclude, recent trends in various research field shows a paradigm shift towards data-driven science. The development of a well-accepted route in this field is still an ongoing work. Multiple possibilities and limitations of data-driven approaches are being analyzed.
The continued growth of data-driven discovery and design of novel materials calls for the combinatorial efforts from theoretical, computational, and experimental scientists to achieve a deeper understanding of structure-property correlations, with the help of both humans and machines.
BIBLIOGRAPHY


[41] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” in


[58] M. de Jong *et al.*, “A Statistical Learning Framework for Materials Science:


[87] Z. Ling *et al.*, “Flexible and conductive MXene films and nanocomposites with high


[99] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme, “Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and


[152] J. Lima-de-Faria, E. Hellner, F. Liebau, E. Makovicky, and E. Parthé,


