

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Atomistic Insights Into Material Chemistry: From First Principles to Machine Learning

Permalink

<https://escholarship.org/uc/item/71w2c6fn>

Author

Kwon, Hyuna

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Atomistic Insights Into Material Chemistry: From First Principles to Machine Learning

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical & Environmental Engineering

by

Hyuna Kwon

March 2023

Dissertation Committee:

Dr. De-en Jiang, Co-Chairperson

Dr. Bryan M. Wong, Co-Chairperson

Dr. Juchen Guo

Copyright by
Hyuna Kwon
2023

The Dissertation of Hyuna Kwon is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

COPYRIGHT ACKNOWLEDGEMENT

The text and figures in Chapter 3, in part or full, are reproduced from “Tuning Metal–Dihydrogen Interaction in Metal–Organic Frameworks for Hydrogen Storage,” *J. Phys. Chem. Lett.* 2022, 13 (39), 9129-9133. The coauthor (Dr. De-en Jiang) directed and supervised this research.

The text and figures in Chapter 4, in part or full, are reproduced from “Electron/Hole Mobilities of Periodic DNA and Nucleobases Structures from Large-Scale DFT Calculations,” under review. The coauthor (Dr. Bryan M. Wong) directed and supervised this research.

The text and figures in Chapter 5, in part or full, are reproduced from “Understanding Electrooxidation of Furfural on Cu, Co-Spinel Oxides from Density Functional Theory” (*In preparation*). The coauthor (Dr. De-en Jiang) directed and supervised this research.

The text and figures in Chapter 6, in part or full, are reproduced from “Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per-and Polyfluoroalkyl Substances (PFASs).” *Environ. Sci. Tech. Lett.* 2022, <https://doi.org/10.1021/acs.estlett.2c00530>. The coauthor (Dr. Bryan M. Wong) directed and supervised this research.

The text and figures in Chapter 7, in part or full, are reproduced from “Harnessing Neural Network for Predicting XANES Spectroscopy of Amorphous Carbon Materials” (*In preparation*). The coauthor (Dr. Tuan Anh Pham) directed and supervised this research.

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisors, Dr. De-en Jiang and Dr. Bryan M. Wong, for their unwavering support and guidance throughout my Ph.D. journey. I have learned invaluable lessons about conducting research, from asking scientific questions, providing quantitative evidence for a hypothesis, and effectively presenting the work to the public. My research will be continuously inspired by the knowledge I gained from them.

I would also like to thank Dr. Tuan Anh Pham at Lawrence Livermore National Laboratory (LLNL), who offered me a precious summer internship supported by Computational Chemistry and Materials Science program. With his help and encouragement, I had a life-changing opportunity to participate in exciting projects and collaborate with great researchers at LLNL.

I thank my dissertation committee member, Dr. Juchen Guo, for his guidance. In addition, I want to thank my candidacy exam committee members, Dr. Jianzhong Wu and Dr. Gregory Beran, for their courteous service.

I am also grateful to my colleagues at UC Riverside, Dr. Lu Wang, Dr. Tao Wu, Dr. Kristen Wang Romero, Dr. Tongyu Liu, Ms. Yuqing Fu, Mr. Chuanye Xiong, Mr. Yujing Tong, Mr. Haohong Song, and Ms. Qiuyao Li for their insightful discussion and collaboration.

Lastly, I thank my family and friends for their encouragement throughout my Ph.D. journey. Their love and understanding have been a constant source of strength and motivation.

ABSTRACT OF THE DISSERTATION

Atomistic Insights Into Material Chemistry: From First Principles to Machine Learning

by

Hyuna Kwon

Doctor of Philosophy, Graduate Program in Chemical & Environmental Engineering
University of California, Riverside, March 2023

Dr. De-en Jiang, Co-Chairperson

Dr. Bryan M. Wong, Co-Chairperson

In the last few decades, with the increased water pollution and energy scarcity, there have been attempts to solve these issues by studying chemical reactions and designing new materials. For example, understanding the structural and chemical properties of heterogeneous interfaces is critical in many applications, including water treatment, gas storage, energy storage, or water splitting. In this aspect, ab-initio simulations have been a powerful tool for investigating structural and chemical properties such as charge transfer, chemical stability, hole/electron conduction, and reaction energetics within chemical materials. However, due to the substantial computational cost, large, complex chemical and material systems are challenging to calculate with current density functional theory (DFT) based quantum calculation tools. Therefore, with the help of machine learning (ML) techniques, we can accelerate the prediction of chemical properties and reaction dynamics of intricate systems by providing a large dataset obtained from DFT calculations to train the ML models. My dissertation is composed of two parts. In the first part, we utilize the first principles calculations to explore structure-

property relationships, electronic structure, and reaction energetics of various systems. For instance, we study the hydrogen storage performance of metal-organic framework, the conductivity of DNA strands, and the electrooxidation of biomass on Cu,Co-spinel oxides. In the second part of my dissertation, we address machine-learning-assisted methods to accelerate the investigation of the structure-property relationship of materials for many extended systems. For example, we explore the bioactivities of perfluoroalkyl substances and X-ray absorption spectroscopy of disordered systems such as amorphous carbon systems. We propose the advantages of applying ML in computational chemistry and materials science research from these examples. First, ML accelerates exploring structure-property relationships. Second, ML can also be used to interpret complex systems with a large supercell which would be computationally expensive if we only rely on DFT calculations. In short, ML combined with first-principles calculations enables more efficient and effective investigation of larger systems.

Table of Contents

Chapter 1. Introduction	1
1.1 Density Functional Theory and Machine Learning	1
1.2 Structure-property Relationship.....	2
1.3 Metal-organic Framework	3
1.4 Perfluoroalkyl Substances.....	4
1.5 X-ray Absorption Near Edge Spectroscopy.....	4
1.6 Overview.....	5
Reference	7
Chapter 2. Computational Methods	11
2.1 First Principles Calculations	11
2.1.1 Density functional theory (DFT)	11
2.1.2 Choice of Exchange-Correlation Functional	13
2.1.3 Exchange-correlation Energy.....	13
2.1.4 Dispersion Corrections.....	14
2.2 Molecular Docking	15
2.3 Structural Descriptors	15
2.3.1 Local Many-Body Tensor Representation (LMBTR).....	15
2.3.2 Atom-Centered Symmetry Function (ACSF)	16
2.3.3 Smooth Overlap of Atomic Position (SOAP).....	17
2.4 Machine learning	18
2.4.1 Unsupervised Learning Methods	18
2.4.2 Semi-supervised Metric Learning.....	19
2.4.3 Model Selection	19

Reference	21
Chapter 3. Tuning Metal-Dihydrogen Interaction in Metal-Organic Frameworks.....	27
3.1 Abstract.....	27
3.2 Introduction.....	28
3.3 Results and discussion	29
3.4 Summary and Conclusions	35
3.5 Computational Details	36
Reference	37
Chapter 4. Electron/Hole Mobilities of Periodic DNA and Nucleobases Structures from Large-Scale DFT Calculations.....	41
4.1 Abstract.....	41
4.2 Introduction.....	42
4.3 Computational Method	44
4.4 Results and Discussion	46
4.4.1 Benchmark Calculations on Nucleotide Base Pairs.....	46
4.4.2 Optimized Geometries of Single- and Double-Stranded DNA.....	49
4.4.3 Electron/Hole Mobilities.....	52
4.5 Summary and Conclusions	61
Reference	64
Chapter 5. Understanding Electrooxidation of Furfural on Cu, Co-Spinel Oxides from Density Functional Theory	69
5.1 Abstract.....	69
5.2 Introduction.....	70

5.3 Computational Method	72
5.4 Results and Discussion	75
5.4.1 Bulk CuCo ₂ O ₄	75
5.4.2 Clean Surfaces	76
5.4.3 HMF Adsorption on the CuCo ₂ O ₄ (100) Surface	78
5.4.4 Intermediate adsorption on the CuCo ₂ O ₄ (100) surface	81
5.4.5 HMF to DFF oxidation reaction	83
5.4.6 DFF to FFCA Oxidation Reaction.....	85
5.4.7 HMF to HMFCA.....	87
5.4.8 HMFCA to FFCA	89
5.5 Summary and Conclusions	91
Reference	93
Chapter 6. Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per- and Polyfluoroalkyl Substances (PFASs).....	
6.1 Abstract.....	96
6.2 Introduction.....	97
6.3 Computational Method	99
6.4 Results and Discussion	102
6.4.1 Unsupervised learning	102
6.4.2 Semi-supervised Metric Learning.....	108
6.4.3 Interactions between PFAS and targets	111
6.4.4 Bioactivity Predictions on OECD Dataset.....	112
6.5 Summary and Conclusions	115
Reference	117
Chapter 7. Harnessing Neural Network for Predicting XANES Spectroscopy of Amorphous Carbon Materials.....	
	120

7.1 Abstract	120
7.2 Introduction.....	120
7.3 Computational Method	123
7.3.1 Dataset.....	123
7.3.2 Structure Representation	125
7.3.3 Machine Learning Construction	126
7.4 Results and Discussion	128
7.4.1 Unsupervised Learning	128
7.4.2 Supervised Learning	132
7.4.3 Feature Analysis.....	136
7.4.4 Inverse prediction of local XANES	138
7.4.5 Inverse prediction of global XANES	140
7.5 Summary and Conclusions	140
Reference	143
Chapter 8. Summary and Outlook	148
Appendix A. Publication List	151

List of Tables

Table 3-1. Comparison of the present work with the cluster model and the experiment for V(II)-H ₂ distance (V to the center of mass of H ₂) and the V-Cl distance before and after H ₂ adsorption on V ₂ Cl _{2.8} (btdd).	31
Table 4-1. Comparison of interaction energies predicted by LDA, BLYP, B3LYP, and B3LYP-D against CCSD(T) reference values from the S22 dataset.	48
Table 4-2. Lattice parameters, band gaps, and effective masses of holes/electrons of various single- and double-strand DNA systems computed at the B3LYP-D/6-311g(d,p) level of theory.	53
Table 4-3. Electronic charge per nucleobase and electron/hole mobilities of various single- and double-strand DNA systems computed at the B3LYP-D/6-311g(d,p) level of theory.	54
Table 5-1. Comparison of experimental ²⁸ and calculated lattice parameters of the bulk CuCo ₂ O ₄	76
Table 5-2. HMF adsorption energies (<i>E_{ads}</i>) at the Cu (Co) sites, M-O bond length (<i>r_{M-O}</i>) (M = Cu, Co), and partial atomic charge transfer on Cu (Co) from Bader charge analysis (Δq).	81
Table 5-3. Adsorption energies (<i>E_{ads}</i>) of the intermediates (HMFCA, DFF, FFCA, and FDCA) and M-O bond length (<i>r_{M-O}</i>) (M = Cu, Co).	83
Table 6-1. Cluster number, accuracy, and maximum common structure that are most likely to be found in bioactive molecules toward each target.	105
Table 7-1. MAE of ML model predicting XANES spectra using MBTR, SOAP, and ACSF.	134
Table 7-2. MAE of local structure prediction from XANES	140

List of Figures

Figure 3-1. Optimized structure of H ₂ adsorption in the V ₂ Cl _{2.8} (btdd) MOF: (a) top view; (b) side view; (c) a close-up view of the adsorption site.	30
Figure 3-2. Adsorption energies of H ₂ in M ₂ Cl ₂ (btdd) and M ₂ Cl _{2.8} (btdd) with M being 3d transition metals.	33
Figure 3-3. (a) Schematic of crystal field splitting of V(II) and Cr(II) d orbitals in M ₂ Cl _{2.8} (btdd); (b) projected density of states (PDOS) of the d _{x²-y²} orbital of M(II) cations in M ₂ Cl _{2.8} (btdd) for M=Sc, Ti, V, Cr (top and bottom plots in each panel represent spin-up and spin-down channels, respectively; Fermi level is set as zero).	34
Figure 3-4. Charge density difference plot after H ₂ adsorption: (a) on V(II) in V ₂ Cl _{2.8} (btdd); (b) on V(II) in V ₂ Cl ₂ (btdd). Yello, electron density accumulation; cyan, electron density depletion. Iso-values are +/- 0.140 e Bohr ⁻³ . Net charge changes (Δq measured by Bader charge) after adsorption are also given.	35
Figure 4-1. Molecular structures of DNA nucleobase monomers, stacked pairs, and Watson-Crick base pairs from the S22 dataset used as benchmarks in this work. The carbon, hydrogen, nitrogen, and oxygen atoms are depicted as gray, white, blue, and red spheres, respectively.	47
Figure 4-2. Interaction energies (in kcal/mol) of stacked and Watson-Crick pair configurations of GC and AT calculated at different levels of theory and compared to CCSD(T) benchmark values (denoted as red stars) from the S22 dataset.	48
Figure 4-3. Geometries of periodic poly(G-C) obtained with the LDA, BLYP, B3LYP, and B3LYP-D functionals. Only LDA and B3LYP-D give stable structures, whereas the other functionals give unstable and distorted geometries between adjacent Watson-crick pairs.	51
Figure 4-4. HOCOs and LUCOs of ssDNA obtained with the B3LYP-D functional. The HOCOs and LUCOs were calculated at isovalues of 0.01 and 0.03, respectively.	55
Figure 4-5. HOMOs of A, T, G, and C calculated at the B3LYP-D/6-311g(d,p) level of theory.	56
Figure 4-6. HOCOs and LUCOs of poly(A-T) and poly(G-C) calculated at the B3LYP-D/6-311g(d,p) level of theory.	58
Figure 4-7. Band structures of poly(A), poly(T), poly(A-T), poly(G), poly(C), and poly(G-C) calculated at the B3LYP-D/6-311g(d,p) level of theory. The A- and G-type bands are pushed upwards in the double-stranded poly(A-T) and poly(G-C) and cases, respectively.	60
Figure 5-1. The two possible pathways for the oxidation of HMF to FDCA. (DFF: 2,5-diformylfuran; HMFCA: 5-hydroxymethyl-2-furancarboxylic acid; FFCA: 5-formylfuran-2carboxylic acid)	72
Figure 5-2. Bulk CuCo ₂ O ₄ . Cobalts occupy octahedral sites, and Coppers occupy tetrahedral sites. Cu, green; Co, blue; O, red.	75
Figure 5-3. Top views of clean CuCo ₂ O ₄ surfaces. Co, purple; Cu, green; O, red.	77
Figure 5-4. The surface grand potential of 11 surface models. $\Delta\mu O = -0.27 eV$ at 1 atm, 300K condition is highlighted with yellow.	78

Figure 5-5. HMF adsorption conformation on CuCo ₂ O ₄ surface. (a) O atom of the hydroxyl group adsorbs on Cu and (b) Co sites. (c) O atom of the hydroxyl group and carbonyl group adsorb on the Cu and Co sites, respectively. (d) Top view of (c).	80
Figure 5-6. The most stable adsorption conformations of (a) HMFCa, (b) DFF, (c) FFCA, and (d) FDCA on CuCo ₂ O ₄ (100) surface.	82
Figure 5-7. Two possible reaction pathways for HMF oxidation	84
Figure 5-8. Schematic drawing of the complete catalytic cycle of oxidation of HMF to DFF on the CuCo ₂ O ₄ (100) surface.	85
Figure 5-9. Schematic drawing of the complete catalytic cycle of oxidation of DFF to FFCA on the CuCo ₂ O ₄ (100) surface.....	87
Figure 5-10. Schematic drawing of the complete catalytic cycle of oxidation of HMF to HMFCa on the CuCo ₂ O ₄ (100) surface.	88
Figure 5-11. Schematic drawing of the complete catalytic cycle of oxidation of HMFCa to FFCA on the CuCo ₂ O ₄ (100) surface.	90
Figure 5-12. Pathway-dependent energy profiles for the HMF oxidation into FFCA on CuCo ₂ O ₄ surface	91
Figure 6-1. Machine-learning-based workflow for QSAR construction and application to PFASs.	99
Figure 6-2. Distribution of the molecules in the C3F6 data set. The molecular structures were visualized by PC t-SNE. Each point is a molecule, and the colors of the points are clusters classified using k-means clustering.	104
Figure 6-3. 2-dimensional space distribution of molecules in the CF3 data set. Each point represents a molecule that is either bioactive (red) or inactive (blue) towards (a, b) K18 and (c, d) CYP2C9. The molecules are visualized on a 2-dimensional space using (a, c) PC t-SNE (unsupervised) or (b, d) semi-supervised metric learning.....	108
Figure 6-4. Distribution of molecules in the CF dataset using semi-supervised metric learning. Each point represents a molecule that is either bioactive (red circular edges) or inactive (light blue circular edges) towards (a) CYP2C9, (b) CYP3A4, (c) CYP2D6, and (d) ATXN. The olive green-filled circles represent molecules having the substructure depicted in the plot; i.e., (a, b) ester groups, (c) phenylprimidyl groups, and (d) 4-benzyl-2-(4-fluorophenyl)-1,2-thiazole. The pink-filled circles in (c) represent molecules with phenylethanone. The percentage value represents the ratio of the number of bioactive molecules within the identified substructure. Table S3 lists the predicted substructures for specific targets.	110
Figure 6-5. Clustering/classification of molecules predicted with unsupervised learning (dimension reduction) on CF datasets containing (a) chemical structures and (b) chemical structures and binding affinities with CYP2C9. Each point represents a molecule that is either bioactive (red) or inactive (blue) towards CYP2C9.	112
Figure 6-6. (a) OECD dataset classified by PC t-SNE and clustered based on the k-means clustering method. The orange and yellow dots represent ester-containing molecules. The colors closer to red (yellow) represent a higher (lower) concentration of bioactive molecules. (b) PFAS molecules included in the OECD list are grouped into 40 clusters. Each point represents a molecule, and clusters 13, 25, and 39 denote a high ratio of ester-containing groups.....	114

Figure 7-1. Schematic drawing of procedure.....	124
Figure 7-2. Peaks P1 and P2 present a-C XANES spectra are denoted with red and green arrows, respectively. Grey lines represent each XANES spectrum from individual carbon, and the blue curve represents the average XANES spectrum from carbons with sp, sp ² , and sp ³ hybridization.....	128
Figure 7-3. PCA analysis of XANES spectroscopies. Each point represents XANES spectra and color displays (a) coordination number, (b) P1/P2 intensity ratio, (c) P2 intensity, and (d) minimum bond length.	130
Figure 7-4. PCA analysis of XANES spectra based on bond angles and corresponding local environments of a-C materials. Each point represents XANES spectra, and color represents the minimum bond angle. Grey balls represent surrounding carbon atoms, and yellow balls represent the center carbon atom. The hybridization and the minimum bond angle of the center atom are denoted. The left panels are typical configurations of sp, sp ² , and sp ³ carbons, while the bottom panels demonstrate distorted carbon structures which majorly exist in the a-C systems.	131
Figure 7-5. XANES spectrum prediction using (a) MBTR, (b) SOAP, and (c) ACSF descriptors. Solid lines and dashed lines represent DFT calculated NN predicted XANES spectra, respectively. Lower panels display MSE of predicting P1 and P2 intensities and energies using each of the descriptors.	135
Figure 7-6. Feature analysis using the shuffling method. (a) XANES spectra predicted using limited constrained structural features. Solid lines and dashed lines correspond to DFT calculated, and NN predicted XANES spectra, respectively. (b) MAE of predicting spectroscopic features such as peak intensities and energies.....	137
Figure 7-7. Gradient of (a) P1 and (b) P2 intensity with respect to two body interaction terms (k ₂) in LMBTR.	138
Figure 7-8. Local XANES inverse prediction. (a) Comparison of hybridization prediction accuracies obtained from four different models. (b) Heat map of predicted and actual hybridization obtained from Random Forest classifier.....	139

Chapter 1. Introduction

1.1 Density Functional Theory and Machine Learning

With the increased environmental pollution and energy scarcity, nowadays, many researchers attempt to solve these global issues by studying chemical reactions and designing new materials¹. In this context, understanding the atomistic insights of chemical materials is critical in many applications, including water treatment², gas storage, energy storage³, or water splitting⁴.

First-principles calculations have been a powerful tool for investigating the atomistic level of various properties of chemical materials⁵, such as charge transfer⁶, chemical stability⁷, hole/electron conduction⁸, and reaction energetics⁹. Furthermore, in the last few decades, tremendous progress has been made in developing and applying quantum chemical tools for predicting the properties of chemical, biological, and material systems with the fast growth of computational capabilities¹⁰. These computational tools are helpful not only in computing the properties of known systems but also designing functional materials with desired properties¹¹.

Out of the many quantum chemical techniques currently being utilized, density functional theory (DFT) based methods are now considered one of the most precise and effective methods for predicting various chemical properties in chemistry, physics, and materials science¹². This dissertation also shows that DFT can be successfully applied to diverse systems to compute properties ranging from hydrogen storage performance (Chapter 3) to the hole/electron conductivity of DNA strands (Chapter 4). Moreover, we

also present that DFT can predict chemical reactions by providing thermodynamics of various complex reactions, such as the electrooxidation of biomass on spinel oxides (Chapter 5).

Nevertheless, when applied to large complex systems, DFT-based techniques suffer from high computational costs because they require self-consistent (iterative) computations associated with the Schrödinger-like equation.¹³ Therefore, we anticipate that with the help of machine learning (ML) techniques, the acceleration of the prediction of chemical properties and reaction dynamics of intricate can be achieved. Thus, in the latter part of the dissertation (Chapters 6-7), we propose exploring chemical properties using ML-assisted methodologies.

1.2 Structure-property Relationship

Developing materials with unique and valuable properties is an active and burgeoning field in modern chemistry. Predicting the properties of novel materials and understanding the relationship between properties and structures would be incredibly valuable for materials scientists.¹⁴ Due to the complexity of many new materials, there is a demand for ML techniques to generate reliable and predictive models linking these chemical properties and microscopic structures.¹⁵ This application of ML to model materials properties is known as quantitative structure-property relationship (QSPR) modeling.¹⁶

In Chapters 3 and 6, we explore the structure-property relationship, such as the hydrogen adsorption capability of metal-organic frameworks (MOF) and the bioactivities of perfluoroalkyl substances (PFAS). For MOF, with relatively small periodic cells, we

used first-principles calculations, which resulted in very insightful results and discoveries; however, it required a considerable computational cost. Furthermore, since the btdd MOF has a relatively sizeable periodic cell, the variations on the MOF structure were limited to replacing the metal cation with ten different metals.

In this context, in Chapter 6, we applied machine learning techniques to obtain a more efficient and compelling exploration of the structure-property relationship, which can be utilized for tens of thousands of molecules with diverse structural variations.

1.3 Metal-organic Framework

Fuel cells are gaining much attention as a clean energy source to solve worldwide energy scarcity issues and air pollution.¹⁷ However, one challenge with commercializing fuel cell vehicles is that a very safe and competent hydrogen storage system is required.¹⁸ There are two big categories of hydrogen storage systems; one is physically based, such as compressing and liquifying hydrogen gas.¹⁹ The other is material-based ones such as absorbents like MOF and chemical hydrogens such as ammonia borane.²⁰ Physics-based systems mostly have higher capacity than MOF but require extreme conditions such as high pressure or low temperature.²¹

Therefore, in Chapter 3, we focus on finding the MOF with optimal adsorption energy ranging from 15 to 25 kJ/mol at ambient conditions.²² Also, since hydrogen gas adsorbs on the cation part of MOF, it is critical to investigate metal-dihydrogen interaction to analyze the hydrogen storage performance.²³

1.4 Perfluoroalkyl Substances

PFAS is a well-known water pollutant that causes harmful effects on human health and the environment.²⁴ It is generated from industrial processes such as manufacturing repellents and fire extinguishers²⁵, and they are so-called forever chemicals because breaking down strong C-F bonds is difficult²⁶. PFAS is not a term for one single molecule but refers to a large group of more than 60,000 molecules with diverse molecular structures and chemical properties. Therefore, to classify this wide variety of PFAS molecules based on their chemical properties, in Chapter 6, we present the prediction of the chemical properties of various PFAS molecules, such as C-F bond bioactivities, from a given chemical structure.

1.5 X-ray Absorption Near Edge Spectroscopy

There are many different experimental techniques to probe the chemical structure of materials. For example, X-ray Diffraction (XRD) techniques are widely used to determine crystalline structures.²⁷ However, these conventional measurement techniques are less practical when determining the structure of disordered systems. Therefore, X-ray absorption spectroscopy is extremely valuable for these intricate systems because it is sensitive to local structural environments and can provide information about subtle geometry distortion.²⁸

When an X-ray strikes an atom, one of the core electrons can be excited into an unbound state called the continuum. Then, when electrons are ejected from an atom of solid material, this is essentially the photoelectric effect. Suppose this photoelectron has just enough kinetic energy to escape into the continuum. In that case, multiple scattering

processes occur between surrounding atoms neighboring the absorbing atom. This is why the X-ray near-edge spectrum is sensitive to local structure and has crucial structural information, including formal valence, coordination environment, and subtle geometry distortion, unlike any other spectroscopy.

Having these insights from X-ray spectroscopy, we are specifically interested in XANES of amorphous carbon. Amorphous carbon is recently gaining attention for various applications, such as anode materials for batteries.²⁹ Since there is no long-range order in the amorphous systems, XANES is extremely valuable in identifying chemical structures³⁰; however, interpreting the XANES is not straightforward in most cases. In this context, in Chapter 7, we present the correlation between the structure and the XANES spectrum using ML techniques.

1.6 Overview

This dissertation includes both first-principles simulations and ML-assisted methods. The first part of this dissertation (Chapters 3-5) describes applying the DFT methodology to study chemical properties. In Chapter 3, we investigated the structure-activity relationship of MOF. Specifically, Chapter 3 focuses on hydrogen adsorption performance varying the metal sites of MOF. Chapter 4 utilizes DFT to calculate hole/electron conductivity in biosystems such as DNA. Chapter 5 continues to investigate that DFT can also predict chemical reactions such as the electrooxidation of biomass on the Cu,Co-spinel oxides.

As mentioned earlier, DFT is a handy and powerful tool, yet there is a limitation. It requires tremendous computational cost and cannot be applied to many systems or complex systems with large supercells. Therefore, the next section of the dissertation, Chapters 6-7, explores using ML methods. Chapter 6 discusses employing ML techniques to predict the chemical properties of small molecules such as PFAS. In this way, we present the successful prediction of the bioactivities of more than 60,000 molecules based on their chemical structures. In Chapter 7, we employ ML techniques to predict the X-ray absorption spectroscopy of amorphous carbon systems, which is complicated and has a larger cell size.

Reference

- (1) Cheng, L.; Xiang, Q.; Liao, Y.; Zhang, H. CdS-Based Photocatalysts. *Energy Environ. Sci.* **2018**, *11* (6), 1362–1391. <https://doi.org/10.1039/C7EE03640J>.
- (2) Jenness, G. R.; Koval, A. M.; Etz, B. D.; Shukla, M. K. Atomistic Insights into the Hydrodefluorination of PFAS Using Silylium Catalysts. *Environ. Sci. Process. Impacts* **2022**, *24* (11), 2085–2099. <https://doi.org/10.1039/D2EM00291D>.
- (3) Poizot, P.; Dolhem, F. Clean Energy New Deal for a Sustainable World: From Non-CO₂ Generating Energy Sources to Greener Electrochemical Storage Devices. *Energy Environ. Sci.* **2011**, *4* (6), 2003. <https://doi.org/10.1039/c0ee00731e>.
- (4) Raman, A. S.; Vojvodic, A. Providing Atomistic Insights into the Dissolution of Rutile Oxides in Electrocatalytic Water Splitting. *J. Phys. Chem. C* **2022**, *126* (2), 922–932. <https://doi.org/10.1021/acs.jpcc.1c08737>.
- (5) Wang, L.-F.; Ma, T.-B.; Hu, Y.-Z.; Wang, H. Atomic-Scale Friction in Graphene Oxide: An Interfacial Interaction Perspective from First-Principles Calculations. *Phys. Rev. B* **2012**, *86* (12), 125436. <https://doi.org/10.1103/PhysRevB.86.125436>.
- (6) *First Principles Calculations of Charge Transfer Excitations in Polymer–Fullerene Complexes: Influence of Excess Energy* - Niedzialek - 2015 - *Advanced Functional Materials* - Wiley Online Library. <https://onlinelibrary.wiley.com/doi/full/10.1002/adfm.201402682> (accessed 2023-02-22).
- (7) Wang, L.; Sun, Y. Y.; Lee, K.; West, D.; Chen, Z. F.; Zhao, J. J.; Zhang, S. B. Stability of Graphene Oxide Phases from First-Principles Calculations. *Phys. Rev. B* **2010**, *82* (16), 161406. <https://doi.org/10.1103/PhysRevB.82.161406>.
- (8) Shao, Z.-G.; Ye, X.-S.; Yang, L.; Wang, C.-L. First-Principles Calculation of Intrinsic Carrier Mobility of Silicene. *J. Appl. Phys.* **2013**, *114* (9), 093712. <https://doi.org/10.1063/1.4820526>.
- (9) *First-Principles Calculations for the Energetics of the Hydration Reaction of Acceptor-Doped BaZrO₃* | *Chemistry of Materials*. <https://pubs.acs.org/doi/full/10.1021/acs.chemmater.6b03907> (accessed 2023-02-22).

- (10) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50* (3), 539–543. <https://doi.org/10.1021/acs.accounts.6b00532>.
- (11) Jain, A.; Shin, Y.; Persson, K. A. Computational Predictions of Energy Materials Using Density Functional Theory. *Nat. Rev. Mater.* **2016**, *1* (1), 1–13. <https://doi.org/10.1038/natrevmats.2015.4>.
- (12) Hafner, J. Ab-Initio Simulations of Materials Using VASP: Density-Functional Theory and Beyond. *J. Comput. Chem.* **2008**, *29* (13), 2044–2078. <https://doi.org/10.1002/jcc.21057>.
- (13) Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Putting Density Functional Theory to the Test in Machine-Learning-Accelerated Materials Discovery. *J. Phys. Chem. Lett.* **2021**, *12* (19), 4628–4637. <https://doi.org/10.1021/acs.jpcclett.1c00631>.
- (14) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889–2919. <https://doi.org/10.1021/cr200066h>.
- (15) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10* (3), 2260–2297. <https://doi.org/10.1021/acscatal.9b04186>.
- (16) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks. *J. Phys. Chem. C* **2013**, *117* (15), 7681–7689. <https://doi.org/10.1021/jp4006422>.
- (17) Abdalla, A. M.; Hossain, S.; Petra, P. M.; Ghasemi, M.; Azad, A. K. Achievements and Trends of Solid Oxide Fuel Cells in Clean Energy Field: A Perspective Review. *Front. Energy* **2020**, *14* (2), 359–382. <https://doi.org/10.1007/s11708-018-0546-2>.
- (18) Ren, J.; Langmi, H. W.; North, B. C.; Mathe, M. Review on Processing of Metal–Organic Framework (MOF) Materials towards System Integration for Hydrogen Storage. *Int. J. Energy Res.* **2015**, *39* (5), 607–620. <https://doi.org/10.1002/er.3255>.
- (19) Wang, F.; Swinbourn, R.; Li, C. Shipping Australian Sunshine: Liquid Renewable Green Fuel Export. *Int. J. Hydrog. Energy* **2023**. <https://doi.org/10.1016/j.ijhydene.2022.12.326>.

- (20) Moussa, G.; Moury, R.; Demirci, U. B.; Şener, T.; Miele, P. Boron-Based Hydrides for Chemical Hydrogen Storage. *Int. J. Energy Res.* **2013**, *37* (8), 825–842. <https://doi.org/10.1002/er.3027>.
- (21) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143* (42), 17535–17547. <https://doi.org/10.1021/jacs.1c07217>.
- (22) Jaramillo, D. E.; Jiang, H. Z. H.; Evans, H. A.; Chakraborty, R.; Furukawa, H.; Brown, C. M.; Head-Gordon, M.; Long, J. R. Ambient-Temperature Hydrogen Storage via Vanadium(II)-Dihydrogen Complexation in a Metal–Organic Framework. *J. Am. Chem. Soc.* **2021**, *143* (16), 6248–6256. <https://doi.org/10.1021/jacs.1c01883>.
- (23) Vitillo, J. G.; Regli, L.; Chavan, S.; Ricchiardi, G.; Spoto, G.; Dietzel, P. D. C.; Bordiga, S.; Zecchina, A. Role of Exposed Metal Sites in Hydrogen Storage in MOFs. *J. Am. Chem. Soc.* **2008**, *130* (26), 8386–8396. <https://doi.org/10.1021/ja8007159>.
- (24) Lohmann, R.; Cousins, I. T.; DeWitt, J. C.; Glüge, J.; Goldenman, G.; Herzke, D.; Lindstrom, A. B.; Miller, M. F.; Ng, C. A.; Patton, S.; Scheringer, M.; Trier, X.; Wang, Z. Are Fluoropolymers Really of Low Concern for Human and Environmental Health and Separate from Other PFAS? *Environ. Sci. Technol.* **2020**, *54* (20), 12820–12828. <https://doi.org/10.1021/acs.est.0c03244>.
- (25) Glüge, J.; Scheringer, M.; T. Cousins, I.; C. DeWitt, J.; Goldenman, G.; Herzke, D.; Lohmann, R.; A. Ng, C.; Trier, X.; Wang, Z. An Overview of the Uses of Per- and Polyfluoroalkyl Substances (PFAS). *Environ. Sci. Process. Impacts* **2020**, *22* (12), 2345–2373. <https://doi.org/10.1039/D0EM00291G>.
- (26) *Taking the “F” out of forever chemicals.* <https://doi.org/10.1126/science.add1813>.
- (27) Ali, A.; Chiang, Y. W.; Santos, R. M. X-Ray Diffraction Techniques for Mineral Characterization: A Review for Engineers of the Fundamentals, Applications, and Research Directions. *Minerals* **2022**, *12* (2), 205. <https://doi.org/10.3390/min12020205>.
- (28) Rouff, A. A.; Elzinga, E. J.; Reeder, R. J.; Fisher, N. S. X-Ray Absorption Spectroscopic Evidence for the Formation of Pb(II) Inner-Sphere Adsorption Complexes and Precipitates at the Calcite–Water Interface. *Environ. Sci. Technol.* **2004**, *38* (6), 1700–1707. <https://doi.org/10.1021/es0345625>.

- (29) Xie, L.; Tang, C.; Bi, Z.; Song, M.; Fan, Y.; Yan, C.; Li, X.; Su, F.; Zhang, Q.; Chen, C. Hard Carbon Anodes for Next-Generation Li-Ion Batteries: Review and Perspective. *Adv. Energy Mater.* **2021**, *11* (38), 2101650. <https://doi.org/10.1002/aenm.202101650>.
- (30) Sheng, H. W.; Luo, W. K.; Alamgir, F. M.; Bai, J. M.; Ma, E. Atomic Packing and Short-to-Medium-Range Order in Metallic Glasses. *Nature* **2006**, *439* (7075), 419–425. <https://doi.org/10.1038/nature04421>.

Chapter 2. Computational Methods

This chapter briefly introduces the computational techniques used in this dissertation, along with their primary functions. Subsequently, the succeeding chapters provide more details about the computational methodologies employed in each chapter.

2.1 First Principles Calculations

2.1.1 Density functional theory (DFT)

DFT is a bottom-up approach for simulating atomistic behaviors, including chemical reactions and properties, by calculating the electron-electron interactions using the Hartree-Fock theory¹. A critical assumption in DFT is that the nuclei are seen as fixed (the Born-Oppenheimer approximation), which is highly reasonable considering that the nuclei are several multitudes heavier than electrons.² As a result, a static external potential V is generated in which the electrons move.³ Therefore, with this assumption, the time-independent Schrödinger becomes as follows:

$$\hat{H}\Psi = E\Psi = [K + V + U]\Psi = \left[\sum_{i=1}^N -\frac{\hbar}{2m} \nabla_i^2 + \sum_{i=1}^N V(\mathbf{r}_i) + \sum_{i=1}^N \sum_{j>i}^N U(\mathbf{r}_i, \mathbf{r}_j) \right] \Psi \quad (\text{Eq. 2.1})$$

where N is the number of electrons, \hat{H} is the Hamiltonian, E is the total Energy, K is the kinetic energy, V is the external potential, and U is the electron-electron interaction energy. The simplest way to solve this complex many-electron Schrödinger equation is to include electron density, which is a function of spatial coordinates of the electrons, instead of the wavefunction itself. The formulation for electron density is as follows:

$$n(\mathbf{r}) = N \int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N \int \Psi^*(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \Psi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (\text{Eq. 2.2})$$

Using this electron density, or density functional, calculating the electron-electron interaction becomes remarkably cheaper in terms of computational cost.

Using this density function, modern DFT is based on the Kohn-Sham equations, composed of four terms – kinetic energy (K), Coulombic interaction (V), nuclei-electron interaction (J), and exchange-correlation component (E_{xc}) – to calculate the ground state energy of a system. The formula is as follows:

$$E[n] = K[n] + V[n] + U[n] \quad (\text{Eq. 2.3})$$

Then, Eq. 2.3 can be solved using Kohn-Sham equations, which is a simplified Schrödinger equation:

$$\left(-\frac{1}{2}\nabla^2 + V_s(\mathbf{r})\right)\varphi_i(\mathbf{r}) = \varepsilon_i\varphi_i(\mathbf{r}) \quad (\text{Eq. 2.4})$$

where ε_i is the energy associated with the orbital φ_i and V_s is the effective potential in which electrons are moving, which is the sum of the external potential (V), electron-electron Coulomb repulsion, and exchange-correlation potential (V_{XC}):

$$V_s(\mathbf{r}) = V(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r}' + V_{XC}[n(\mathbf{r})] \quad (\text{Eq. 2.5})$$

As a result of solving the above equation, we can obtain the Kohn-Sham orbital (φ_i) and the sum of the square moduli of the occupied Kohn-Sham orbitals equal to the overall electron density as follows:

$$n(\mathbf{r}) = \sum_{i=1}^n |\varphi_i(\mathbf{r})|^2 \quad (\text{Eq. 2.6})$$

Solving the above equations is self-consistent because the Kohn-Sham orbital (φ_i), density functional ($n(\mathbf{r})$), and the effective potential (V_s) are dependent on each other.

2.1.2 Choice of Exchange-Correlation Functional

The accuracy of DFT calculations largely depends on the choice of exchange-correlation functionals.⁴ The most straightforward functional is Local Density Approximation (LDA)⁵, which assumes homogeneous electron gas and works well only in solid bulk systems⁶. However, since our interest mainly lies in surface chemical reactions, one-step advanced functional, generalized gradient approximation (GGA)⁷ was proposed to incorporate inhomogeneous electron density in the surface and molecules⁸. GGA functionals use the gradient of density ($\nabla\rho$), and BLYP⁹ and PBE¹⁰ are two of the most popular among these types of functionals. GGA functionals are widely used in research due to their affordability and moderate accuracy across diverse systems. Meta-GGA functionals (e.g., SCAN¹¹, ω B97X-d) take one step further, using the second derivative of density ($\nabla^2\rho$) and kinetic energy density¹².

2.1.3 Exchange-correlation Energy

Despite all the efforts to make DFT more accurate, since the standard DFT functionals cannot accurately capture the exchange-correlation energy, there is a well-known defect of the pure DFT methods – band gap problem¹³. In particular, the LDA and GGA functionals tend to produce too much electron delocalization and too little electron localization, resulting in an underestimation of the band gap¹⁴. This issue is particularly

severe for systems with strongly correlated electrons (e.g., transition metal oxides or rare earth compounds), where the exchange-correlation effects are more complex¹⁵. There are two different approaches to solving this problem. The first is the corrective approach, to add Hubbard correction (U)¹⁶, obtained either empirically or from ab initio calculations, which has been proven to be efficient and reasonably reliable.

Another solution is using hybrid functionals (e.g., B3LYP¹⁷), which minimize the error associated with exchange-correlation energy by incorporating a fraction of exact exchange and the exchange-correlation functional used in GGA¹⁸. This exact exchange term, typically obtained from Hartree-Fock theory, helps correct some of the shortcomings of GGA functionals in describing electronic properties, such as the band gap, by better accounting for the exchange energy¹⁹. Furthermore, hybrid functionals often offer a better trade-off between accuracy and computational efficiency than meta-GGA functionals, which can be more computationally expensive²⁰. Lastly, meta-hybrid functionals (e.g., M06-2X²¹) combine meta-functionals and hybrid functionals²².

2.1.4 Dispersion Corrections

To properly characterize systems such as two-dimensional materials held together by van der Waals forces between layers or liquids that form a network via hydrogen bonding, it is essential to describe the van der Waals interaction accurately. Nevertheless, there are challenges in employing DFT to represent van der Waals forces, specifically dispersion²³. The easiest method to incorporate van der Waals correction in DFT is introducing an energy correction to the conventional Kohn-Sham DFT energy²⁴. For this reason, in Chapters 3-5, we utilized the DFT-D3²⁵ approaches proposed by Grimme et al.

2.2 Molecular Docking

AutoDock4²⁶ was used to process the ligand and ligand interaction conformation analysis in the binding affinity analysis. The docking pockets were based on the global search of the protein to determine the optimal binding sites²⁷ AutoDock4 uses a semiempirical free energy forcefield scoring function to perform a quick and accurate evaluation of the binding energies of ligands to proteins using a two-step approach. First, the intramolecular energetics of the transition from the unbound state to the bound form of the protein-ligand complex is estimated. The intermolecular energetics of the bound complex is subsequently evaluated.

2.3 Structural Descriptors

To use molecular structures as an input for machine learning models, we need a quantitative metric to represent molecular structures. In the recent decade, A few molecular representations have been proposed, and here, we introduce three commonly used structural descriptors.

2.3.1 Local Many-Body Tensor Representation (LMBTR)

MBTR²⁸ is a combination of the bag of bonds and coulomb matrix, complementing the limitations such as non-uniqueness, discontinuity, and non-generality.²⁹ More specifically, in Chapter 7, we used Local MBTR (LMBTR), which employs MBTR on a center atom to describe the local environment. LMBTR describes a local atomic structure by 2-body (distances between pairs of atoms) and 3-body (angles in a triplet of atoms) functions³⁰. The mathematical description is as follows:

$$\sum_{i=1}^{N_a} w_k(i) D(x, g_k(x)) \prod_{j=1}^k C_{z_j, z_{ij}} \quad (\text{Eq. 2.7})$$

where index i runs over atoms within the cutoff radius from the center atom, N_a is the number of atoms, D is broadened probabilistic distribution (Gaussian distribution in this work), and g_k is a geometric function that describes k -body terms.

2.3.2 Atom-Centered Symmetry Function (ACSF)

ACSF³¹ is also composed of 2-body and 3-body interaction terms like MBTR, but combined and not separable, unlike MBTR. In ACSF, each atomic environment is encoded into symmetry functions, G^k , where G^1 , G^2 , and G^3 represent 2-body (radial) symmetry functions and G^4 correspond to 3-body (angular) symmetry functions.

G^1 is the sum of the cutoff functions f_c around the center atom i :

$$G_i^1 = \sum_j f_c(R_{ij}) \quad (\text{Eq. 2.8})$$

G^2 calculates atomic density around the center atom by multiplying the exponential damping function with the cutoff function:

$$G_i^2 = \sum_j e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \quad (\text{Eq. 2.9})$$

Parameter η determines the smoothing of the atomic density function. The mathematical description of the rest of the symmetry functions are as follows:

$$G_i^3 = \sum_j \cos(\kappa R_{ij}) f_c(R_{ij}) \quad (\text{Eq. 2.10})$$

$$G_i^4 = 2^{1-\zeta \sum_{j,k,l=i} (1+\lambda \cos \theta_{ijk})^\zeta} e^{-\eta(R_{ij}^2+R_{jk}^2+R_{ki}^2)} f_c(R_{ij}) f_c(R_{jk}) f_c(R_{ki}) \quad (\text{Eq. 2.11})$$

2.3.3 Smooth Overlap of Atomic Position (SOAP)

In SOAP³², the Gaussian density ρ^Z within the sphere of the cutoff radius of the center atom is calculated as follows:

$$\rho^Z(r) = \sum_i^{|Z|} e^{-\frac{|r-R_i|^2}{2\sigma^2}} \quad (\text{Eq. 2.12})$$

where i runs over atoms, Z is the atomic number, and σ is the width of the Gaussian function. Atomic density can also be expressed using spherical harmonics (Y_{lm}) and orthonormal basis function (g_n) as follows:

$$\rho^Z(r) = \sum_{nlm} c_{nlm}^Z g_n(r) Y_{lm}(\theta, \phi) \quad (\text{Eq. 2.13})$$

where c_{nlm} is calculated by taking the volumetric integral of the multiplication of the spherical radial basis function ($g_n(r)$), spherical harmonics, and density function:

$$c_{nlm}^Z = \iiint_{R^3} g_n(r) Y_{lm}(\theta, \phi) \rho^Z(r) dV \quad (\text{Eq. 2.14})$$

The multiplication of atomic densities of each atom is then equivalent to the multiplication of the coefficients c_{nlm}^Z . Therefore, the final form of SOAP of elements Z_1 and Z_2 is as follows:

$$p_{nn'l}^{Z_1 Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c_{nlm}^{Z_1} c_{nlm}^{Z_2} \quad (\text{Eq. 2.15})$$

2.4 Machine learning

Machine learning techniques are generally classified into supervised and unsupervised learning based on data labeling³³. Supervised learning methods use labeled data, while unsupervised learning uses unlabeled data for clustering or dimension reduction purposes³⁴. Supervised learning techniques are powerful tools and provide accurate results for regression or classification purposes; however, they remain a black box, not providing any chemical rationale or insights for their predictions³⁵. On the other hand, unsupervised learning can be better when finding chemical insights from the dataset.

2.4.1 Unsupervised Learning Methods

There are two usual dimension reduction methods, (i) Principal Component Analysis (PCA)³⁶ followed by t-Distributed Stochastic Neighbor Embedding (t-SNE)³⁷, i.e., PC t-SNE, and (ii) UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction)³⁸, were used on our fingerprint data. While t-SNE is a widely-used dimension reduction technique for many types of data analysis, prior studies have shown that UMAP exhibits better clustering performance, especially for large datasets.³⁹

Then, we introduce three representative clustering methods: k-means⁴⁰, Density-Based Spatial Clustering of Applications with Noise (DBSCAN)⁴¹, and Hierarchical DBSCAN (HDBSCAN)⁴². While k-means clustering is relatively more time-efficient, DBSCAN and HDBSCAN can efficiently handle outliers and noisy datasets⁴³.

2.4.2 Semi-supervised Metric Learning

Deep metric learning (DML)⁴⁴ is a machine learning technique to investigate mixed data distributions before building prediction models. DML approaches have been accurately used in state-of-the-art computer vision technologies,⁴⁵⁻⁴⁷ and recently, Na et al. reported their usefulness in cheminformatics⁴⁸. The central concept of DML is to simplify a prediction problem by generating a new vector representation and effectively dividing the data well depending on their target values. The metric-learning algorithms compute/learn Mahalanobis distances, which are given by the expression:

$$D(x, x') = \sqrt{(Lx - Lx')^T(Lx - Lx')} \quad (\text{Eq. 2.16})$$

where x and x' are positions of two data points, and L is the matrix to modify the distance metric. Given L as a unitary matrix, D becomes Euclidean distance. Based on the semi-supervised data, the metric learning problem is generally formulated to optimize Mahalanobis distances.⁴⁹ The metric-learning algorithm seeks to find the parameters of a distance function (L) that optimizes some objective function measuring the agreement with the training data.⁵⁰ Accordingly, in Chapter 6, we used metric learning with a semi-supervised learning algorithm to learn a distance metric that places molecules with similar chemical properties close together and molecules with opposite properties far away.

2.4.3 Model Selection

Once we perform unsupervised learning using multiple models, it is essential to evaluate the performance of the model outputs and determine the best-performing model.

Silhouette score⁵¹ is one criterion to select the best model, which analyzes the distances of each data point to its cluster and neighboring clusters. In short, a higher Silhouette score guarantees better performance in clustering. The Silhouette score $s(i)$ of a data x_i can be calculated by the following expression:

$$s(i) = \frac{b(i) - w(i)}{\max\{b(i), w(i)\}} \text{ with } b(i) = \min_k\{B(i, k)\} \quad (\text{Eq. 2.17})$$

where $w(i)$ is the average distance from the i^{th} point to the other points in its cluster, and $B(i, k)$ is the average distance from the i^{th} point to points in another cluster k . The Silhouette score is also used as an essential statistical method to optimize hyperparameters.⁵²

Reference

- (1) Hafner, J. Ab-Initio Simulations of Materials Using VASP: Density-Functional Theory and Beyond. *J. Comput. Chem.* **2008**, *29* (13), 2044–2078. <https://doi.org/10.1002/jcc.21057>.
- (2) Lonsdale, D. R.; Goerigk, L. The One-Electron Self-Interaction Error in 74 Density Functional Approximations: A Case Study on Hydrogenic Mono- and Dinuclear Systems. *Phys. Chem. Chem. Phys.* **2020**, *22* (28), 15805–15830. <https://doi.org/10.1039/D0CP01275K>.
- (3) Perdew, J. P.; Ruzsinszky, A.; Constantin, L. A.; Sun, J.; Csonka, G. I. Some Fundamental Issues in Ground-State Density Functional Theory: A Guide for the Perplexed. *J. Chem. Theory Comput.* **2009**, *5* (4), 902–908. <https://doi.org/10.1021/ct800531s>.
- (4) Jain, A.; Shin, Y.; Persson, K. A. Computational Predictions of Energy Materials Using Density Functional Theory. *Nat. Rev. Mater.* **2016**, *1* (1), 1–13. <https://doi.org/10.1038/natrevmats.2015.4>.
- (5) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789. <https://doi.org/10.1103/PhysRevB.37.785>.
- (6) Puska, M. J.; Nieminen, R. M. Atoms Embedded in an Electron Gas: Beyond the Local-Density Approximation. *Phys. Rev. B* **1991**, *43* (15), 12221–12233. <https://doi.org/10.1103/PhysRevB.43.12221>.
- (7) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46* (11), 6671–6687. <https://doi.org/10.1103/PhysRevB.46.6671>.
- (8) Moll, N.; Bockstedte, M.; Fuchs, M.; Pehlke, E.; Scheffler, M. Application of Generalized Gradient Approximations: The Diamond– β -Tin Phase Transition in Si and Ge. *Phys. Rev. B* **1995**, *52* (4), 2550–2556. <https://doi.org/10.1103/PhysRevB.52.2550>.
- (9) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. <https://doi.org/10.1103/PhysRevA.38.3098>.

- (10) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868. <https://doi.org/10.1103/PhysRevLett.77.3865>.
- (11) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115* (3), 036402. <https://doi.org/10.1103/PhysRevLett.115.036402>.
- (12) Perdew, J. P.; Constantin, L. A. Laplacian-Level Density Functionals for the Kinetic Energy Density and Exchange-Correlation Energy. *Phys. Rev. B* **2007**, *75* (15), 155109. <https://doi.org/10.1103/PhysRevB.75.155109>.
- (13) Grüning, M.; Marini, A.; Rubio, A. Effect of Spatial Nonlocality on the Density Functional Band Gap. *Phys. Rev. B* **2006**, *74* (16), 161103. <https://doi.org/10.1103/PhysRevB.74.161103>.
- (14) Zheng, X.; Cohen, A. J.; Mori-Sánchez, P.; Hu, X.; Yang, W. Improving Band Gap Prediction in Density Functional Theory from Molecules to Solids. *Phys. Rev. Lett.* **2011**, *107* (2), 026403. <https://doi.org/10.1103/PhysRevLett.107.026403>.
- (15) Cipriano, L. A.; Di Liberto, G.; Tosoni, S.; Pacchioni, G. Band Gap in Magnetic Insulators from a Charge Transition Level Approach. *J. Chem. Theory Comput.* **2020**, *16* (6), 3786–3798. <https://doi.org/10.1021/acs.jctc.0c00134>.
- (16) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. Band Theory and Mott Insulators: Hubbard U Instead of Stoner I. *Phys. Rev. B* **1991**, *44* (3), 943–954. <https://doi.org/10.1103/PhysRevB.44.943>.
- (17) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652. <https://doi.org/10.1063/1.464913>.
- (18) Boese, A. D.; Handy, N. C. New Exchange-Correlation Density Functionals: The Role of the Kinetic-Energy Density. *J. Chem. Phys.* **2002**, *116* (22), 9559–9569. <https://doi.org/10.1063/1.1476309>.
- (19) Xiao, H.; Tahir-Kheli, J.; Goddard, W. A. I. Accurate Band Gaps for Semiconductors from Density Functional Theory. *J. Phys. Chem. Lett.* **2011**, *2* (3), 212–217. <https://doi.org/10.1021/jz101565j>.
- (20) Simón, L.; M. Goodman, J. How Reliable Are DFT Transition Structures? Comparison of GGA, Hybrid-Meta-GGA and Meta-GGA Functionals. *Org. Biomol. Chem.* **2011**, *9* (3), 689–700. <https://doi.org/10.1039/C0OB00477D>.

- (21) Zhao, Y.; Truhlar, D. G. Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2008**, *4* (11), 1849–1868. <https://doi.org/10.1021/ct800246v>.
- (22) Mourik, T. van. Assessment of Density Functionals for Intramolecular Dispersion-Rich Interactions. *J. Chem. Theory Comput.* **2008**, *4* (10), 1610–1619. <https://doi.org/10.1021/ct800231f>.
- (23) Ángyán, J. G.; Gerber, I. C.; Savin, A.; Toulouse, J. Van Der Waals Forces in Density Functional Theory: Perturbational Long-Range Electron-Interaction Corrections. *Phys. Rev. A* **2005**, *72* (1), 012510. <https://doi.org/10.1103/PhysRevA.72.012510>.
- (24) Tsuneda, T.; Hirao, K. Long-Range Correction for Density Functional Theory. *WIREs Comput. Mol. Sci.* **2014**, *4* (4), 375–390. <https://doi.org/10.1002/wcms.1178>.
- (25) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. <https://doi.org/10.1063/1.3382344>.
- (26) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.;Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. <https://doi.org/10.1002/jcc.21256>.
- (27) Feinstein, W. P.; Brylinski, M. Calculating an Optimal Box Size for Ligand Docking and Virtual Screening against Experimental and Predicted Binding Pockets. *J. Cheminformatics* **2015**, *7* (1), 18. <https://doi.org/10.1186/s13321-015-0067-5>.
- (28) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301. <https://doi.org/10.1103/PhysRevLett.108.058301>.
- (29) Iype, E.; Urolagin, S. Machine Learning Model for Non-Equilibrium Structures and Energies of Simple Molecules. *J. Chem. Phys.* **2019**, *150* (2), 024307. <https://doi.org/10.1063/1.5054968>.
- (30) Bürkle, M.; Perera, U.; Gimbert, F.; Nakamura, H.; Kawata, M.; Asai, Y. Deep-Learning Approach to First-Principles Transport Simulations. *Phys. Rev. Lett.* **2021**, *126* (17), 177701. <https://doi.org/10.1103/PhysRevLett.126.177701>.

- (31) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134* (7), 074106. <https://doi.org/10.1063/1.3553717>.
- (32) Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-Density Representations for Machine Learning. *J. Chem. Phys.* **2019**, *150* (15), 154110. <https://doi.org/10.1063/1.5090481>.
- (33) *Supervised and Unsupervised Learning for Data Science*; Berry, M. W., Mohamed, A., Yap, B. W., Eds.; Unsupervised and Semi-Supervised Learning; Springer International Publishing: Cham, 2020. <https://doi.org/10.1007/978-3-030-22475-2>.
- (34) Ando, R. K.; Zhang, T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data.
- (35) Kammeraad, J. A.; Goetz, J.; Walker, E. A.; Tewari, A.; Zimmerman, P. M. What Does the Machine Learn? Knowledge Representations of Chemical Reactivity. *J. Chem. Inf. Model.* **2020**, *60* (3), 1290–1301. <https://doi.org/10.1021/acs.jcim.9b00721>.
- (36) Abdi, H.; Williams, L. J. Principal Component Analysis. *WIREs Comput. Stat.* **2010**, *2* (4), 433–459. <https://doi.org/10.1002/wics.101>.
- (37) Belkina, A. C.; Ciccolella, C. O.; Anno, R.; Halpert, R.; Spidlen, J.; Snyder-Cappione, J. E. Automated Optimized Parameters for T-Distributed Stochastic Neighbor Embedding Improve Visualization and Analysis of Large Datasets. *Nat. Commun.* **2019**, *10* (1), 5415. <https://doi.org/10.1038/s41467-019-13055-y>.
- (38) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv September 17, 2020. <https://doi.org/10.48550/arXiv.1802.03426>.
- (39) Hozumi, Y.; Wang, R.; Yin, C.; Wei, G.-W. UMAP-Assisted K-Means Clustering of Large-Scale SARS-CoV-2 Mutation Datasets. *Comput. Biol. Med.* **2021**, *131*, 104264. <https://doi.org/10.1016/j.combiomed.2021.104264>.
- (40) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28* (1), 100–108. <https://doi.org/10.2307/2346830>.
- (41) Ram, A.; Sharma, A.; Jalal, A. S.; Agrawal, A.; Singh, R. An Enhanced Density Based Spatial Clustering of Applications with Noise. In *2009 IEEE*

- International Advance Computing Conference*; 2009; pp 1475–1478.
<https://doi.org/10.1109/IADCC.2009.4809235>.
- (42) Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2013; pp 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
- (43) Karim, A.; Azam, S.; Shanmugam, B.; Kannoorpatti, K. Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework. *IEEE Access* **2020**, *8*, 154759–154788. <https://doi.org/10.1109/ACCESS.2020.3017082>.
- (44) Kaya, M.; Bilge, H. Ş. Deep Metric Learning: A Survey. *Symmetry* **2019**, *11* (9), 1066. <https://doi.org/10.3390/sym11091066>.
- (45) Farzaneh, A. H.; Qi, X. Facial Expression Recognition in the Wild via Deep Attentive Center Loss; 2021; pp 2402–2411.
- (46) Zhang, Y.; Xiang, T.; Hospedales, T. M.; Lu, H. Deep Mutual Learning; 2018; pp 4320–4328.
- (47) Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; Kompatsiaris, Y. Near-Duplicate Video Retrieval With Deep Metric Learning; 2017; pp 347–356.
- (48) Na, G. S.; Chang, H.; Kim, H. W. Machine-Guided Representation for Accurate Graph-Based Molecular Machine Learning. *Phys. Chem. Chem. Phys.* **2020**, *22* (33), 18526–18535. <https://doi.org/10.1039/D0CP02709J>.
- (49) Hoi, S. C. h.; Liu, W.; Chang, S.-F. Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval and Clustering. *ACM Trans. Multimed. Comput. Commun. Appl.* **2010**, *6* (3), 1–26. <https://doi.org/10.1145/1823746.1823752>.
- (50) Bellet, A.; Habrard, A.; Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. arXiv February 12, 2014. <http://arxiv.org/abs/1306.6709> (accessed 2023-02-27).
- (51) Shahapure, K. R.; Nicholas, C. Cluster Quality Analysis Using Silhouette Score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*; 2020; pp 747–748. <https://doi.org/10.1109/DSAA49011.2020.00096>.

- (52) Panichella, A. A Systematic Comparison of Search-Based Approaches for LDA Hyperparameter Tuning. *Inf. Softw. Technol.* **2021**, *130*, 106411. <https://doi.org/10.1016/j.infsof.2020.106411>.

Chapter 3. Tuning Metal-Dihydrogen Interaction in Metal-Organic Frameworks

3.1 Abstract

Control of metal-dihydrogen interaction is critical to the design and discovery of next-generation materials for hydrogen storage. Inspired by a recently reported vanadium-based metal-organic framework (MOF), $V_2Cl_{2.8}(btdd)$ ($H_2btdd = \text{bis}(1H-1,2,3\text{-triazolo}[4,5-b],[4',5'-i])\text{dibenzo}[1,4]\text{dioxin}$), that shows a significantly improved adsorption enthalpy (-21 kJ/mol), here we employ periodic density functional theory calculations to probe how the number of d electrons and the mixed valences influence the M- H_2 interaction inside the $M_2Cl_x(btdd)$ MOFs ($M = 3d$ transition metals and $x = 2$ and 2.8). We find a cliff in the H_2 adsorption energy: the interaction strength remains strong from Sc to V and then falls sharply at Cr. Our results confirm that $V_2Cl_{2.8}(btdd)$ is one of the best-performing hydrogen adsorbents among first-row 3d transition metals but also predict that $Ti_2Cl_{2.8}(btdd)$ is equally promising and $Sc_2Cl_2(btdd)$ and $Ti_2Cl_2(btdd)$ maybe even better. Our electronic structure analysis reveals that an empty $d_{x^2-y^2}$ orbital is critical to the much stronger binding of H_2 at the open M(II) site ($M=Sc/Ti/V$). In contrast, a partially filled $d_{x^2-y^2}$ orbital in Cr(II) and later M(II) dramatically weakens H_2 binding. Further, the presence of V(III) next to V(II) also facilitates H_2 binding at V(II) by offering more freedom to redistribute electron density. Our findings will be helpful in the design of MOFs to enhance H_2 adsorption.

3.2 Introduction

Hydrogen is an important energy carrier and a vital component of sustainable energy infrastructure.²⁻⁴ For onboard use in fuel cell vehicles, hydrogen storage is a crucial challenge.⁵ Many materials, such as metals/alloys, carbons, inorganic compounds, and organic compounds, have been explored for hydrogen storage in the past fifty years.⁶⁻¹² However, reaching the US Department of Energy's ultimate target of 6.5 wt% H₂ storage at the complete-storage-system level for light-duty fuel cell vehicles remains a challenge.¹³

Due to their excellent chemical tunability, high surface area, and versatile porosity, metal-organic frameworks (MOFs) have been a very active research area in the past twenty years for gas adsorption and separation, including hydrogen storage.¹⁴⁻¹⁸ Several groups reported the synthesis of MOF structure with record-high hydrogen adsorption performance.^{19,20} Especially, the M₂Cl₂(btdd) MOF comprising M(II) cations and bis(1H-1,2,3-triazolo[4,5-b],[4',5'-i])dibenzo[1,4]dioxin (btdd) linkers has gained attention as adsorbents for small gas molecules. For example, Oppenheim et al. found Ni₂X₂(btdd) (X = OH, F, Cl) as CO, H₂, and C₂H₄ adsorbents,²¹ and Rieth et al. reported a record-high ammonia adsorption rate on M₂Cl₂(btdd) MOFs (M = Mn, Co, Ni, Cu).^{22,23}

Recently, Long and coworkers reported that the mixed-valence V₂Cl_{2.8}(btdd) has the ideal hydrogen adsorption enthalpy falling within the optimal range of -15 to -25 kJ/mol at ambient temperature.²⁴ They further used density functional theory, with a cluster model consisting of one V(II) and one V(III) and triazoles as ligands, to understand the V-H₂ interaction.²⁴ Such truncated representation can be a cost-effective

way to approximate the 3D crystalline structure of the real MOF material. Still, it may miss some crucial factors from the actual 3D structure, the pore confinement, and the real linkers. Moreover, it is unclear how varying the transition metal would affect the metal-dihydrogen interaction.

3.3 Results and discussion

To understand how the number of d electrons in the metal M and the presence of mixed valent M ions influence the M-H₂ interaction in the M₂Cl_x(btdd) MOFs, herein, we carry out DFT+*U* calculations using a periodic model representing the real MOF structure to study H₂ adsorption. Both the parent composition of M₂Cl₂(btdd) with M(II) and the V-based composition of M₂Cl_{2.8}(btdd) with both M(II)/M(III) have been considered for M being all 3d transition metals.¹⁵ We also analyze the adsorption trend using projected density of states (PDOS) and Bader charge.

M₂Cl₂(btdd) and M₂Cl_{2.8}(btdd) have the same structure. Their primitive cell contains 18 M ions that are connected by btdd linkers (Figure 3-1a): in M₂Cl₂(btdd), all 18 ions are M(II); in M₂Cl_{2.8}(btdd), there are roughly 11 M(II) and 7 M(III) cations. The M(III) ion is coordinated by an extra Cl⁻ ion, while the M(II) ion has an open site that can adsorb H₂ (Figure 3-1a,b). Using the experimental structure as an initial geometry,²⁴ we have further optimized the structures with and without H₂ adsorption. Figure 1c shows typical H₂ adsorption at the M(II) site, using V₂Cl_{2.8}(btdd) as an example. One can see that H₂ adsorbs side-on at the V(II) side with a distance of 2.08 Å from V to the center of mass of H₂, while the adsorbed H-H distance is 0.77 Å (in comparison with 0.75 Å in the gas phase).

Table 3-1 compares the key distances from our periodic model with the cluster model and the experiment.²⁴ One can see that the values from the present work are in better agreement with the experiment for both the V(II)-H₂ distance and the V-Cl distance before and after H₂ adsorption.

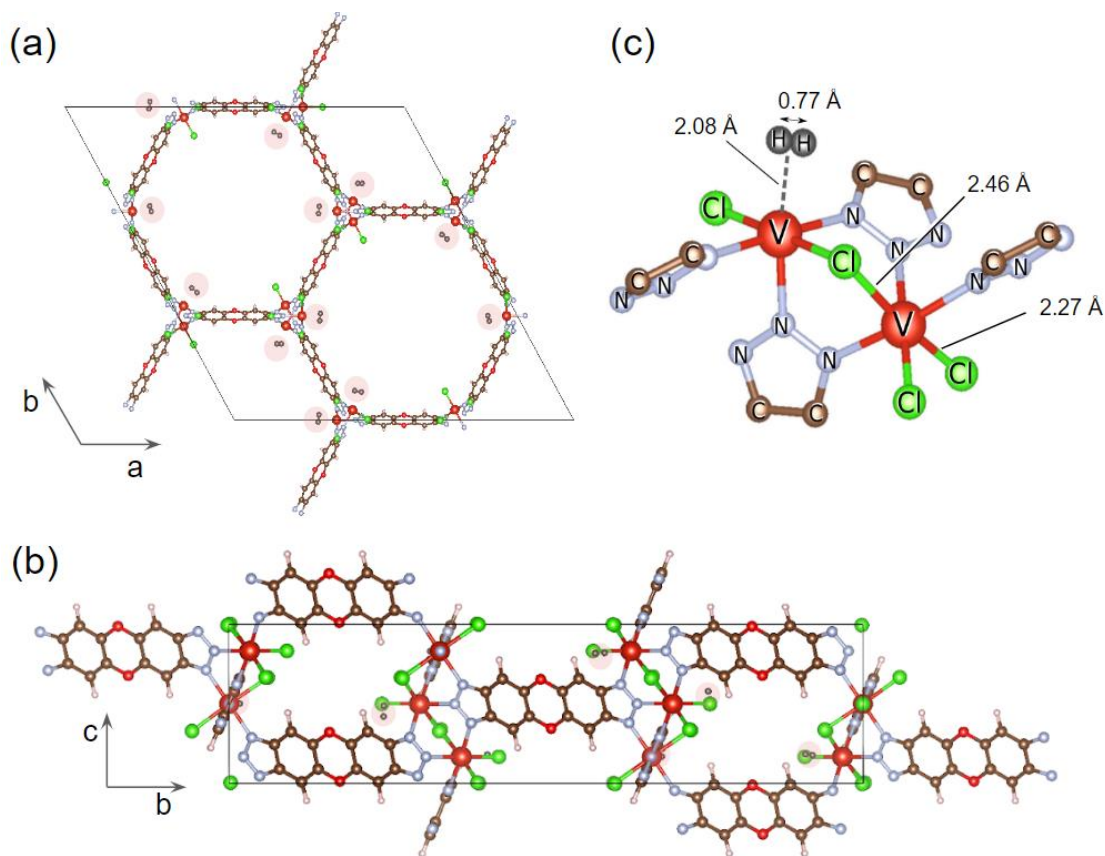


Figure 3-1. Optimized structure of H₂ adsorption in the V₂Cl_{2.8}(btdd) MOF: (a) top view; (b) side view; (c) a close-up view of the adsorption site.

Table 3-1. Comparison of the present work with the cluster model and the experiment for V(II)-H₂ distance (V to the center of mass of H₂) and the V-Cl distance before and after H₂ adsorption on V₂Cl_{2.8}(btdd).

Distances	V-H (after H₂ adsorption)	V-Cl (after H₂ adsorption)	V-Cl (before H₂ adsorption)
Periodic model (present work)	2.08 Å	2.246 Å	2.458 Å
Cluster model ²⁴	2.12 Å	N/A	N/A
Experiment ²⁴	1.97 Å	2.278 Å	2.415 Å

The adsorption energy of H₂ in V₂Cl_{2.8}(btdd) was found to be -39 kJ/mol with PBE+*U* with D3 dispersion correction in the present work. To calculate the adsorption enthalpy at 298 K, we performed frequency calculations on H₂ molecules adsorbed on the V₂Cl_{2.8}(btdd) and obtained thermal energy corrections.²⁵ As a result, the adsorption enthalpy was found to be -28 kJ/mol at 298 K, in reasonable agreement with the experimental enthalpy of adsorption (-20.9 kJ/mol),²⁴ given the usual uncertainties in DFT-computed adsorption energetics.²⁶

The good agreement between our DFT structure/energetics and the experiment for H₂ adsorption in V₂Cl_{2.8}(btdd) justified our use of the crystalline MOF structure to simulate the metal-dihydrogen interaction. Thus, we proceeded further by optimizing adsorption geometry and obtained adsorption energies of H₂ in M₂Cl_{*x*}(btdd) for M being all 3d transition metals and *x* = 2 and 2.8. We assumed a saturated scenario where H₂

molecules adsorb on all open M(II) sites and then obtained the average H₂ adsorption energy. The results are shown in Figure 3-2. First, one can see that the trends and the adsorption energies are similar between x=2 and x=2.8. This can be explained by the fact that in both compositions, H₂ adsorbs on the M(II) site. Second, the adsorption energies of H₂ in Sc₂Cl_{2.8}(btdd) and Ti₂Cl_{2.8}(btdd) are similar to that in V₂Cl_{2.8}(btdd), while those in Sc₂Cl₂(btdd) and Ti₂Cl₂(btdd) are more negative than that in V₂Cl₂(btdd). The strongest adsorption or most favorable adsorption energy is found in Sc₂Cl₂(btdd). Third and more important, there is a drastic drop or cliff between V₂Cl_x(btdd) and Cr₂Cl_x(btdd) in terms of M-H₂ binding: the adsorption energy changes from -0.41 eV in V₂Cl_{2.8}(btdd) to -0.10 eV in Cr₂Cl_{2.8}(btdd), a ~70% decrease in M-H₂ strength. After the drop, the adsorption strength remains weak, despite a slight increase in strength to reach a local maximum at Fe-Co before decreasing again. Figure 3-2 suggests that Sc₂Cl_x(btdd) and Ti₂Cl_x(btdd) are also promising for H₂ storage.

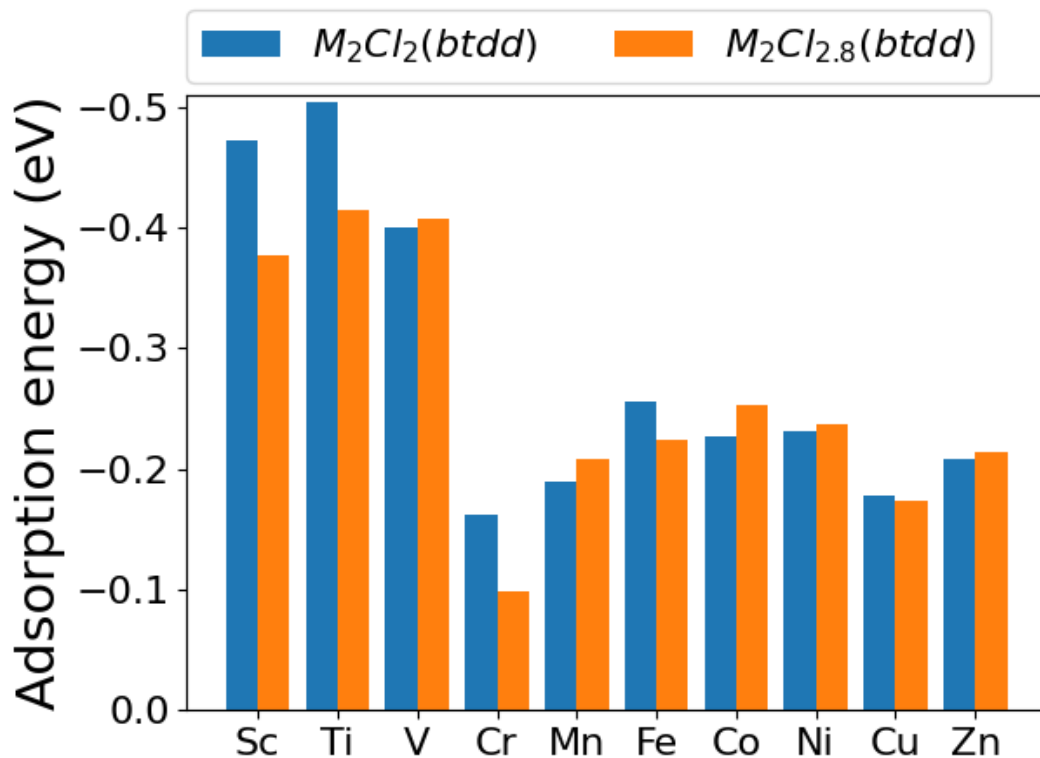


Figure 3-2. Adsorption energies of H_2 in $M_2Cl_2(btdd)$ and $M_2Cl_{2.8}(btdd)$ with M being 3d transition metals.

To elucidate the adsorption energy trend in Figure 3-2 in general and the cliff from V to Cr in particular, we have analyzed the local electronic structure at the M(II) site. Based on the projected density of states and the spin states, we found that the d orbitals of M(II) in $M_2Cl_{2.8}(btdd)$ are split in the crystal field, as shown in Figure 3a. In other words, from $V_2Cl_x(btdd)$ to $Cr_2Cl_x(btdd)$, the main difference is that the $d_{x^2-y^2}$ orbital is empty in V(II) but occupied in Cr(II). Figure 3b depicts the projected density of states (PDOS) for $d_{x^2-y^2}$ orbitals of M(II) cations in $Sc_2Cl_{2.8}(btdd)$, $Ti_2Cl_{2.8}(btdd)$, $V_2Cl_{2.8}(btdd)$, and $Cr_2Cl_{2.8}(btdd)$. Indeed, the M(II) cations of early transition metals (Sc, Ti, V) have empty $d_{x^2-y^2}$ orbitals, while those of Cr and latter transition metals (Mn, Fe,

Co, Ni, Cu, Zn) have filled $d_{x^2-y^2}$ orbitals. This can explain the cliff in the adsorption energy trend in Figure 3-2 because the empty $d_{x^2-y^2}$ orbital can accept electron donation from the σ -bonding orbital of H_2 , while an occupied $d_{x^2-y^2}$ orbital repels the σ -bonding orbital of H_2 .

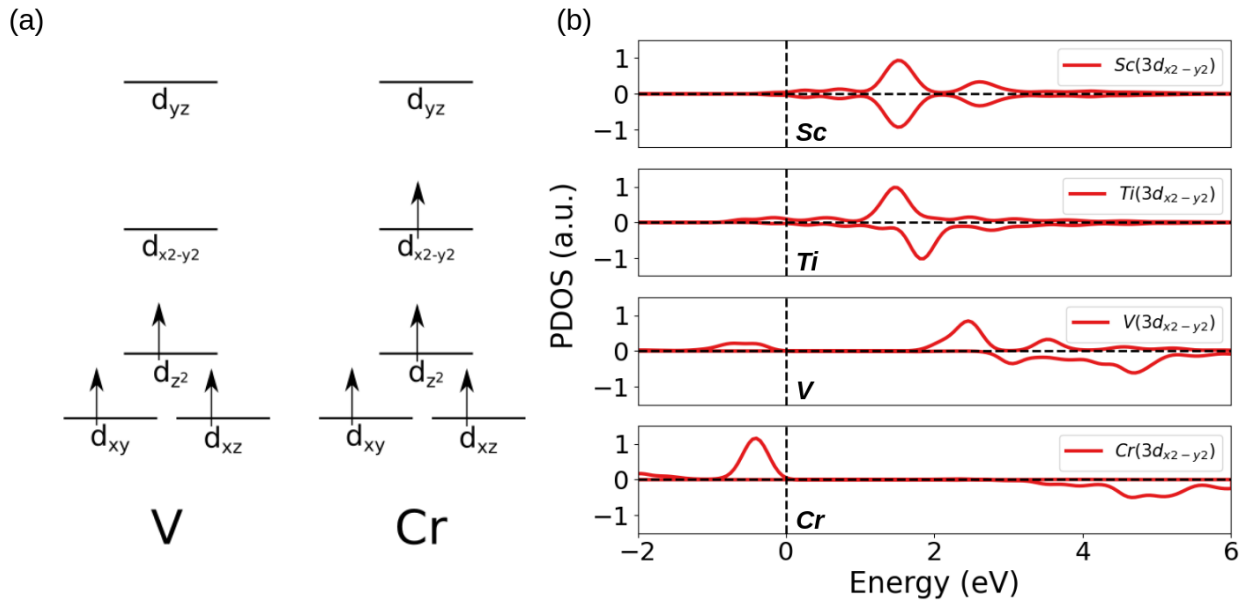


Figure 3-3. (a) Schematic of crystal field splitting of V(II) and Cr(II) d orbitals in $M_2Cl_{2.8}(btdd)$; (b) projected density of states (PDOS) of the $d_{x^2-y^2}$ orbital of M(II) cations in $M_2Cl_{2.8}(btdd)$ for M=Sc, Ti, V, Cr (top and bottom plots in each panel represent spin-up and spin-down channels, respectively; Fermi level is set as zero).

To further understand the nature of the bonding interactions, we have plotted the charge density difference ($\Delta\rho$) induced by the adsorption of H_2 on V(II) in $V_2Cl_{2.8}(btdd)$. As one can see from Figure 4a, electron accumulates between V(II) and H_2 ; more interestingly, there is also significant electron-density re-distribution around V(III) even though it is not directly interacting with H_2 . We have obtained net charge change (Δq) around V(II) and V(III) after H_2 adsorption (Figure 3-4a): V(II) becomes slightly oxidized

($\Delta q=0.11$ e), while V(III) is slightly reduced ($\Delta q=-0.17$ e). In contrast, there is negligible net charge change on V(II) after H₂ in V₂Cl₂(btdd) (Figure 3-4b), where there is no V(III). This difference indicates a more flexible electron response to H₂ adsorption in a mixed-ion MOF such as V₂Cl_{2.8}(btdd).

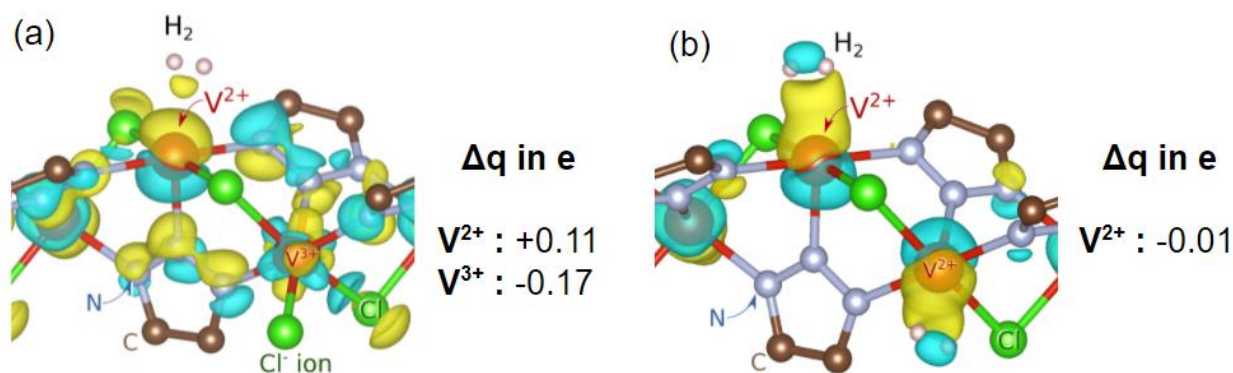


Figure 3-4. Charge density difference plot after H₂ adsorption: (a) on V(II) in V₂Cl_{2.8}(btdd); (b) on V(II) in V₂Cl₂(btdd). Yello, electron density accumulation; cyan, electron density depletion. Iso-values are +/- 0.140 e Bohr⁻³. Net charge changes (Δq measured by Bader charge) after adsorption are also given.

3.4 Summary and Conclusions

In sum, we have investigated H₂ adsorption in M₂Cl_x(btdd) MOFs using periodic density functional theory calculations. We confirmed the structure and energy of H₂ adsorption in V₂Cl_{2.8}(btdd) and further predicted that the Sc and Ti analogs are equally promising for H₂ adsorption (adsorption energy ~ -0.38 to -0.50 eV at the PBE+U+D3 level). Furthermore, we found a drastic drop in H₂ adsorption strength from V to Cr, which is attributed to the occupancy of the d_{x²-y²} orbital. In other words, empty d_{x²-y²} is the key to the much stronger binding of H₂ at the open M(II) site in M₂Cl_x(btdd) MOFs for M being Sc, Ti, and V. Our insights will aim to discover new MOFs with the enhanced H₂ adsorption.

3.5 Computational Details

DFT calculations were performed using the Vienna ab initio simulation package (VASP).²⁷ Electron-ion interactions were described using standard PAW pseudopotentials.^{28,29} The Perdew-Burke-Ernzerhof (PBE) functional³⁰ was combined with the Hubbard U parameter (see Table S2 for the different parameters used for different transition metals and their sources). The empirical correction method (DFT-D3) was chosen to describe the long-range van der Waals (vdW) interactions. Relaxation of the atomic position was carried out using the force criterion of 0.05 eV/Å.

The initial structure of $V_2Cl_{2.8}(\text{btdd})$ was taken from ref²⁴, to which H_2 molecules are added. The adsorption energies (ΔE_{ads}) on different MOFs are calculated using the following equation:

$$\Delta E_{\text{ads}} = (E_{\text{MOF}+nH_2} - E_{\text{MOF}} - n E_{H_2})/n \quad (\text{Eq. 3-1})$$

where $E_{\text{MOF}+H_2}$ is the total energy of the H_2 adsorbed MOF, E_{MOF} , and E_{H_2} represent the total energy for an isolated clean MOF and H_2 molecule, respectively.

The charge density differences were calculated from the individual charge densities for optimized systems of the MOFs, H_2 molecules, and MOF- H_2 complexes using the following equation:

$$\Delta\rho = \rho_{(\text{MOF}+H_2)} - \rho_{\text{MOF}} - \rho_{H_2} \quad (\text{Eq. 3-2})$$

VESTA³¹ software was used for the charge density plot.

Reference

- (1) Poizot, P.; Dolhem, F. Clean Energy New Deal for a Sustainable World: From Non-CO₂ Generating Energy Sources to Greener Electrochemical Storage Devices. <https://doi.org/10.1039/c0ee00731e>.
- (2) Stern, A. G. A New Sustainable Hydrogen Clean Energy Paradigm. *Int J Hydrogen Energy* **2018**, *43* (9), 4244–4255. <https://doi.org/10.1016/J.IJHYDENE.2017.12.180>.
- (3) Abe, J. O.; Popoola, A. P. I.; Ajenifuja, E.; Popoola, O. M. Hydrogen Energy, Economy and Storage: Review and Recommendation. *Int J Hydrogen Energy* **2019**, *44* (29), 15072–15086. <https://doi.org/10.1016/J.IJHYDENE.2019.04.068>.
- (4) Yue, M.; Lambert, H.; Pahon, E.; Roche, R.; Jemei, S.; Hissel, D. Hydrogen Energy Systems: A Critical Review of Technologies, Applications, Trends and Challenges. *Renewable and Sustainable Energy Reviews* **2021**, *146*, 111180. <https://doi.org/10.1016/J.RSER.2021.111180>.
- (5) Staffell, I.; Scamman, D.; Velazquez Abad, A.; Balcombe, P.; Dodds, P. E.; Ekins, P.; Shah, N.; Ward, K. R. The Role of Hydrogen and Fuel Cells in the Global Energy System. *Energy Environ Sci* **2019**, *12* (2), 463–491. <https://doi.org/10.1039/C8EE01157E>.
- (6) Schlapbach, L.; Züttel, A. Hydrogen-Storage Materials for Mobile Applications. *Nature* **2001**, *414* (6861), 353–358. <https://doi.org/10.1038/35104634>.
- (7) Jena, P. Materials for Hydrogen Storage: Past, Present, and Future. *Journal of Physical Chemistry Letters* **2011**, *2* (3), 206–211. https://doi.org/10.1021/JZ1015372/ASSET/IMAGES/MEDIUM/JZ-2010-015372_0006.GIF.
- (8) Yang, J.; Sudik, A.; Wolverton, C.; Siegel, D. J. High Capacity Hydrogen Storage Materials: Attributes for Automotive Applications and Techniques for Materials Discovery. *Chem Soc Rev* **2010**, *39* (2), 656–675. <https://doi.org/10.1039/B802882F>.
- (9) Schneemann, A.; White, J. L.; Kang, S.; Jeong, S.; Wan, L. F.; Cho, E. S.; Heo, T. W.; Prendergast, D.; Urban, J. J.; Wood, B. C.; Allendorf, M. D.; Stavila, V. Nanostructured Metal Hydrides for Hydrogen Storage. *Chem Rev* **2018**, *118* (22), 10775–10839. <https://doi.org/10.1021/ACS.CHEMREV.8B00313>.

- (10) Ahmed, A.; Seth, S.; Purewal, J.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J. Exceptional Hydrogen Storage Achieved by Screening Nearly Half a Million Metal-Organic Frameworks. *Nature Communications* 2019 10:1 **2019**, 10 (1), 1–9. <https://doi.org/10.1038/s41467-019-09365-w>.
- (11) Marques, F.; Balcerzak, M.; Winkelmann, F.; Zepon, G.; Felderhoff, M. Review and Outlook on High-Entropy Alloys for Hydrogen Storage. *Energy Environ Sci* **2021**, 14 (10), 5191–5227. <https://doi.org/10.1039/D1EE01543E>.
- (12) Zheng, J.; Zhou, H.; Wang, C. G.; Ye, E.; Xu, J. W.; Loh, X. J.; Li, Z. Current Research Progress and Perspectives on Liquid Hydrogen Rich Molecules in Sustainable Hydrogen Storage. *Energy Storage Mater* **2021**, 35, 695–722. <https://doi.org/10.1016/J.ENS.M.2020.12.007>.
- (13) Schneemann, A.; White, J. L.; Kang, S.; Jeong, S.; Wan, L. F.; Cho, E. S.; Heo, T. W.; Prendergast, D.; Urban, J. J.; Wood, B. C.; Allendorf, M. D.; Stavila, V. Nanostructured Metal Hydrides for Hydrogen Storage. *Chem Rev* **2018**, 118 (22), 10775–10839. <https://doi.org/10.1021/ACS.CHEMREV.8B00313>.
- (14) Li, J. R.; Kuppler, R. J.; Zhou, H. C. Selective Gas Adsorption and Separation in Metal–Organic Frameworks. *Chem Soc Rev* **2009**, 38 (5), 1477–1504. <https://doi.org/10.1039/B802426J>.
- (15) Furukawa, H.; Cordova, K. E.; O’Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science (1979)* **2013**, 341 (6149). <https://doi.org/10.1126/SCIENCE.1230444>.
- (16) Suh, M. P.; Park, H. J.; Prasad, T. K.; Lim, D.-W. Hydrogen Storage in Metal–Organic Frameworks. *Chem. Rev* **2012**, 112, 782–835. <https://doi.org/10.1021/cr200274s>.
- (17) Dinča, M.; Dailly, A.; Liu, Y.; Brown, C. M.; Neumann, D. A.; Long, J. R. Hydrogen Storage in a Microporous Metal-Organic Framework with Exposed Mn²⁺ Coordination Sites. *J Am Chem Soc* **2006**, 128 (51), 16876–16883. <https://doi.org/10.1021/JA0656853>.
- (18) Lee, K.; Howe, J. D.; Lin, L. C.; Smit, B.; Neaton, J. B. Small-Molecule Adsorption in Open-Site Metal-Organic Frameworks: A Systematic Density Functional Theory Study for Rational Design. *Chemistry of Materials* **2015**, 27 (3), 668–678. <https://doi.org/10.1021/CM502760Q>.
- (19) Kapelewski, M. T.; Geier, S. J.; Hudson, M. R.; Stück, D.; Mason, J. A.; Nelson, J. N.; Xiao, D. J.; Hulvey, Z.; Gilmour, E.; Fitzgerald, S. A.; Head-Gordon, M.; Brown, C. M.; Long, J. R. M₂(m-Dobdc) (M = Mg, Mn, Fe, Co, Ni) Metal-

Organic Frameworks Exhibiting Increased Charge Density and Enhanced H₂ Binding at the Open Metal Sites. *J Am Chem Soc* **2014**, *136* (34), 12119–12129. <https://doi.org/10.1021/JA506230R>.

- (20) Kapelewski, M. T.; Runčevski, T.; Tarver, J. D.; Jiang, H. Z. H.; Hurst, K. E.; Parilla, P. A.; Ayala, A.; Gennett, T.; Fitzgerald, S. A.; Brown, C. M.; Long, J. R. Record High Hydrogen Storage Capacity in the Metal-Organic Framework Ni₂(m-Dobdc) at Near-Ambient Temperatures. *Chemistry of Materials* **2018**, *30* (22), 8179–8189. <https://doi.org/10.1021/ACS.CHEMMATER.8B03276>.
- (21) Oppenheim, J. J.; Mancuso, J. L.; Wright, A. M.; Rieth, A. J.; Hendon, C. H.; Dincă, M. Divergent Adsorption Behavior Controlled by Primary Coordination Sphere Anions in the Metal-Organic Framework Ni₂X₂BTDD. *J Am Chem Soc* **2021**, *143* (40), 16343–16347. <https://doi.org/10.1021/jacs.1c07449>.
- (22) Rieth, A. J.; Dinca, M. Controlled Gas Uptake in Metal-Organic Frameworks with Record Ammonia Sorption. *J Am Chem Soc* **2018**, *140* (9), 3461–3466. <https://doi.org/10.1021/jacs.8b00313>.
- (23) Rieth, A. J.; Tulchinsky, Y.; Dincă, M. High and Reversible Ammonia Uptake in Mesoporous Azolate Metal–Organic Frameworks with Open Mn, Co, and Ni Sites. **2016**. <https://doi.org/10.1021/jacs.6b05723>.
- (24) Jaramillo, D. E.; Jiang, H. Z. H.; Evans, H. A.; Chakraborty, R.; Furukawa, H.; Brown, C. M.; Head-Gordon, M.; Long, J. R. Ambient-Temperature Hydrogen Storage via Vanadium(II)-Dihydrogen Complexation in a Metal-Organic Framework. *J Am Chem Soc* **2021**, *143* (16), 6248–6256. <https://doi.org/10.1021/jacs.1c01883>.
- (25) Wang, V.; Xu, N.; Liu, J. C.; Tang, G.; Geng, W. T. VASPKIT: A User-Friendly Interface Facilitating High-Throughput Computing and Analysis Using VASP Code. *Comput Phys Commun* **2021**, *267*, 108033. <https://doi.org/10.1016/J.CPC.2021.108033>.
- (26) Wellendorff, J.; Silbaugh, T. L.; Garcia-Pintos, D.; Nørskov, J. K.; Bligaard, T.; Studt, F.; Campbell, C. T. A Benchmark Database for Adsorption Bond Energies to Transition Metal Surfaces and Comparison to Selected DFT Functionals. *Surf Sci* **2015**, *640*, 36–44. <https://doi.org/10.1016/J.SUSC.2015.03.023>.
- (27) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput Mater Sci* **1996**, *6* (1), 15–50. [https://doi.org/10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).

- (28) Blöchl, P. E. Projector Augmented-Wave Method. *Phys Rev B* **1994**, *50* (24), 17953. <https://doi.org/10.1103/PhysRevB.50.17953>.
- (29) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys Rev B* **1999**, *59* (3), 1758. <https://doi.org/10.1103/PhysRevB.59.1758>.
- (30) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys Rev Lett* **1996**, *77* (18), 3865. <https://doi.org/10.1103/PhysRevLett.77.3865>.
- (31) Momma, K.; Izumi, F. VESTA 3 for Three-Dimensional Visualization of Crystal, Volumetric and Morphology Data. *urn:issn:0021-8898* **2011**, *44* (6), 1272–1276. <https://doi.org/10.1107/S0021889811038970>.

Chapter 4. Electron/Hole Mobilities of Periodic DNA and Nucleobases Structures from Large-Scale DFT Calculations

4.1 Abstract

Electron/hole transfer mechanisms in DNA and polynucleotide structures continue to garner immense interest as emerging charge-transport systems and molecular electronics. To shed mechanistic insight into these electronic properties, we carried out large-scale DFT calculations (up to 650 atoms and 9,440 basis functions) to systematically analyze the structural and electron/hole transport properties of fully periodic single- and double-stranded DNA. We examined the performance of various exchange-correlation functionals (LDA, BLYP, B3LYP, and B3LYP-D). We found that single-stranded thymine (T) and cytosine (C) are predominantly hole conductors, whereas single-stranded adenine (A) and guanine (G) are better electron conductors. For double-stranded DNA structures, the periodic A-T and G-C electronic band structures undergo a significant renormalization (due to Coulombic repulsion between the nucleobases), which causes hole transport to only occur on the A and G nucleobases. Our calculations (1) constitute the first study of periodic nucleobase structures using dispersion-corrected hybrid functionals with large basis sets (which can be further used as new benchmarks for other coarse-grained methods) and (2) highlight the importance of dispersion effects for obtaining accurate geometries and electron/hole mobilities in these extended systems.

4.2 Introduction

Deoxyribonucleic acid (DNA) and polynucleotide structures continue to garner immense attention in various applications ranging from self-assembled biostructures to building blocks for next-generation electronics.¹⁻⁸ In recent years, DNA has attracted significant attention in nanoelectronics and information storage since it can adopt complex geometries and is inherently stable in many chemical environments.⁹⁻¹⁵ Because of its one-dimensional structure of π -stacked nucleobases, early experimental efforts were intensely focused on the possibility of using DNA as a nanoscale conductor for enhanced electrical conductivity and charge transport.¹⁶⁻²⁰ However, subsequent experiments provided contradictory results, including suggestions that DNA is a conducting wire,²¹ superconductor,²² semiconductor,²³ or a wide-bandgap insulator.²⁴⁻²⁵ The discrepancies in these experimental results were attributed to variations in the DNA structures, such as the specific base sequence and the specific chemical environment used in the experiments.

On the theoretical side, numerous computational studies, including tight-binding models,²⁶⁻²⁷ quantum chemistry calculations,²⁸⁻³⁰ and QM/MM studies³¹⁻³³ have been carried out on DNA and polynucleotide structures to predict their charge transport properties. However, most of these computational studies focused on nucleobase oligomers and did not address band structure properties in a fully periodic geometry. As a result, these oligomer calculations are only appropriate for small molecular-like sections of DNA. They cannot capture the entire electronic transport behavior as a function of electron momentum. There have been a handful of theoretical studies on fully periodic DNA structures; however, these prior studies either employed semilocal exchange-

correlation functionals (known to underestimate bandgaps) with minimal basis sets³⁴ or used Hartree-Fock calculations (which overestimate bandgaps) on idealized geometries extracted from molecular dynamics simulations.³⁵⁻³⁶

To shed additional insight into the electronic properties of DNA and polynucleotide structures, we present large-scale DFT calculations (up to 9,440 basis functions) to systematically analyze their structural and electron/hole transport properties. We also examine the performance of various exchange-correlation functionals, ranging from local (LDA), semilocal (BLYP), hybrid (B3LYP), and dispersion-corrected hybrid (B3LYP-D) methods on the electronic properties of single- and double-stranded DNA structures. It is important to mention that the goal of our current study is not to resolve open issues on charge-transport mechanisms in DNA. Rather, the large-scale calculations presented in this work can serve as new reference benchmarks that are expected to be more accurate than the semilocal or Hartree-Fock calculations discussed previously.³⁴⁻³⁶ Most notably, the B3LYP-D calculations in this work constitute the first study of periodic DNA and nucleobase structures using dispersion-corrected hybrid functionals for both full geometry optimizations and electronic band structures. Using these optimized geometries and band structures, we present electron/hole mobilities for various single- and double-strand DNA structures. Finally, our paper concludes with an analysis of orbitals and charge-transfer mechanisms to rationalize the electronic properties and electron/hole transport mechanisms in these complex nucleobase structures.

4.3 Computational Method

All of the DFT calculations in this study were carried out with a massively parallelized version of the CRYSTAL14 program,³⁷ which can calculate nonlocal Hartree-Fock exchange with all-electron Gaussian basis sets and periodic boundary conditions. While our work focuses on ground-state electronic properties of periodic DNA, previous research by us has shown that the choice of exchange-correlation functional can also strongly affect the accuracy of excitation energies in DNA and RNA nucleobases.³⁸⁻³⁹ As such, we evaluated a wide range of exchange-correlation functionals to understand their effects on DNA electron/hole mobilities, including (1) LDA (local density approximation),⁴⁰ a semilocal functional derived from the exchange-correlation energy of homogeneous electron gas, (2) BLYP (Becke exchange with Lee Yang Parr correlation),⁴¹ a generalized gradient approximation functional without nonlocal exchange, (3) B3LYP,⁴² a popular 3-parameter hybrid functional that contains a 20% fraction of Hartree-Fock exchange, and (4) B3LYP-D,⁴³ a dispersion-corrected version of the B3LYP hybrid functional.

Geometries for all single- and double-stranded DNA structures were optimized using a 6-31G(d,p) all-electron basis set with one-dimensional periodic boundary conditions along the helical axis. Since each of the phosphate groups along the backbone has a -1 charge, a single Na⁺ cation was added near these groups to ensure charge neutrality of the entire periodic system. All the structures examined in this work exhibit a full helical turn (i.e., 360°) with 10 nucleotides in a one-dimensional periodic unit cell. At these optimized geometries, single-point calculations were performed with a larger 6-

311G(d,p) basis set with 100 k-points along the one-dimensional Brillouin zone to obtain the resulting electronic band structures. It is worth noting that the calculations on some of the periodic DNA strands were extremely computationally intensive due to the immense size of these systems. For example, the largest of these structures (poly(A-T)) consists of 650 atoms and 9,440 basis functions and, as such, constitutes one of the most extensive quantum mechanical studies of these periodic biological structures to date.

We briefly outline the deformation potential (DP) formalism⁴⁴⁻⁴⁵ for calculating electron and hole mobilities (μ_e and μ_h , respectively) for each of our DNA structures. Within this formalism, the electron or hole mobilities in a one-dimensional (1-D) periodic system are given by:

$$\mu_{e,h} = \frac{e\hbar^2 C}{(2\pi k_B T)^{1/2} |m_{e,h}^*|^{3/2} E_1^2}, \quad (\text{Eq. 4-1})$$

where e is the charge of an electron, k_B is Boltzmann's constant, \hbar is the reduced Planck constant, T is the temperature (set to 298 K in this study), and E_1 is the DP constant along the 1-dimensional periodic direction. The latter is obtained by calculating the rate of change of the valence/conduction band edge with respect to strain. The elastic modulus of the system is given by $C = 1/a_0 \cdot \partial^2 E / \partial \varepsilon^2$, where E is the total energy of the system and ε is strain. The effective mass of the electrons and holes (m_e and m_h , respectively) was calculated at the conduction band minimum and valence band maximum, respectively, using the expression

$$\frac{1}{m_{e,h}^*} = \pm \frac{1}{\hbar^2} \frac{d^2 \epsilon_{c,v}}{dk^2}. \quad (\text{Eq. 4-2})$$

The positive sign is taken for the (electron) conduction band (ϵ_c), and the negative sign corresponds to the (hole) valence band (ϵ_v). A total of 100 uniformly space points from Γ to the X point were used to calculate $m_{e,h}^*$.

Finally, the elastic constant C was calculated from a contraction-dilation displacement of the entire nucleotide strand using the following expression:

$$C = l_o \left. \frac{\partial^2 E}{\partial l^2} \right|_{l=l_o}, \quad (\text{Eq. 4-3})$$

where l is the length of the DNA strand under tension/compression, E is the total energy per unit cell, and l_o is the equilibrium length. The periodic DNA strand was stretched/compressed at 0.5%, 1.0%, and 1.5% intervals with single-point energies calculated at each step. These seven data points ($\Delta l/l_o = 0, \pm 0.005, \pm 0.01, \pm 0.015$) were then used to generate dilation-energy curves to obtain the elastic constant.

4.4 Results and Discussion

4.4.1 Benchmark Calculations on Nucleotide Base Pairs

Before calculating electron/hole mobilities of the various DNA strands, we first assessed the accuracy of the LDA, BLYP, B3LYP, and B3LYP-D functionals for predicting nucleotide interaction energies when compared to the S22 benchmark dataset.⁴⁶ In particular, the S22 set contains several DNA nucleobase monomers (adenine, cytosine, guanine, and thymine), stacked pair geometries (adenine-thymine and guanine-

cytosine), and a canonical Watson-Crick base pair (adenine-thymine and guanine-cytosine) calculated at a complete-basis-set-extrapolated CCSD(T) level of theory. Figure 4-1 depicts the molecular structures of the various base pair systems considered in this work, and Figure 4-2 compares the interaction energies obtained by the various functionals against the CCSD(T) benchmark values from the S22 dataset (numerical values and root mean square errors (RMSEs) are reported in Table 4-1).

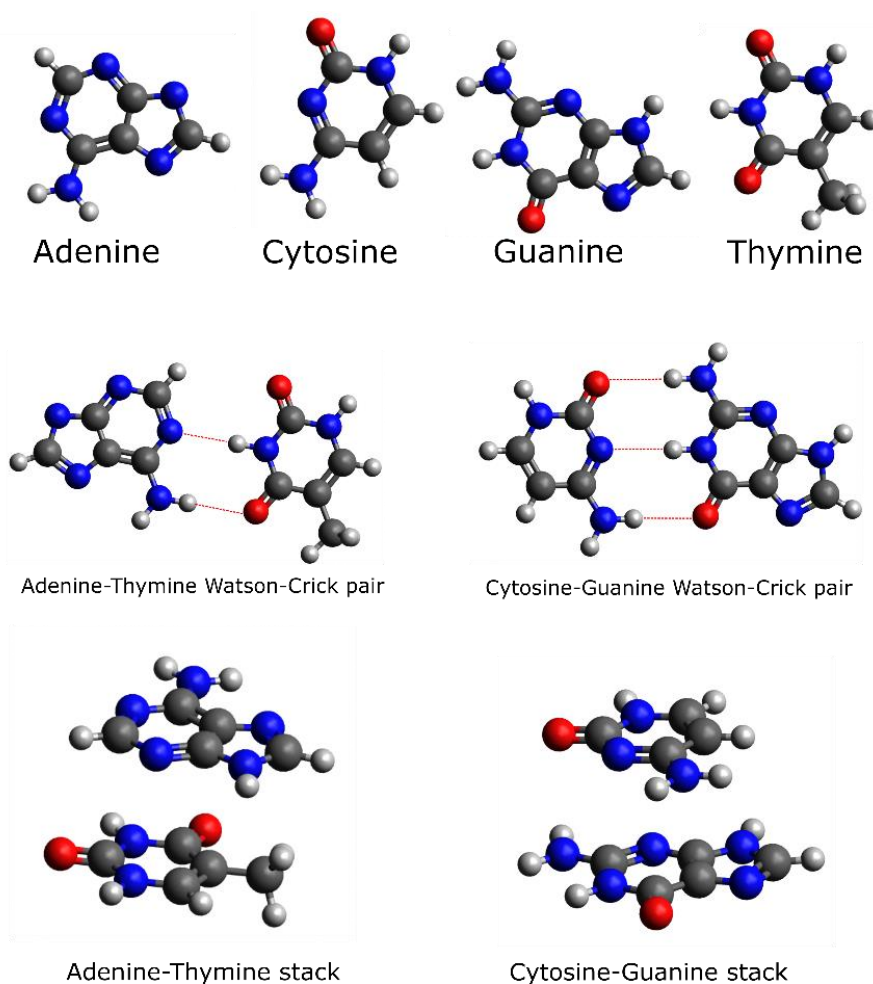


Figure 4-1. Molecular structures of DNA nucleobase monomers, stacked pairs, and Watson-Crick base pairs from the S22 dataset used as benchmarks in this work. The carbon, hydrogen, nitrogen, and oxygen atoms are depicted as gray, white, blue, and red spheres, respectively.

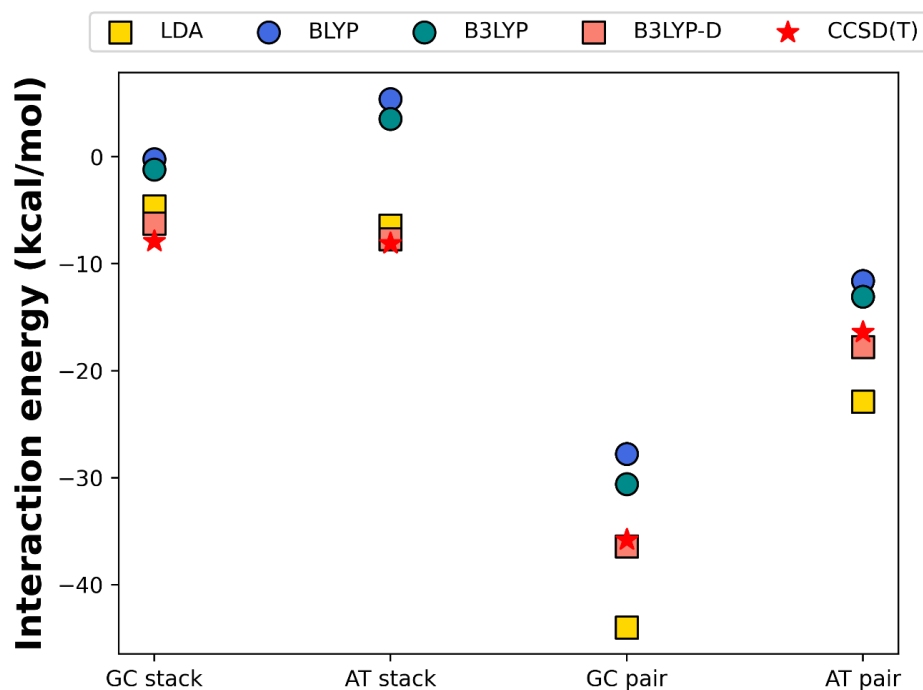


Figure 4-2. Interaction energies (in kcal/mol) of stacked and Watson-Crick pair configurations of GC and AT calculated at different levels of theory and compared to CCSD(T) benchmark values (denoted as red stars) from the S22 dataset.

Table 4-1. Comparison of interaction energies predicted by LDA, BLYP, B3LYP, and B3LYP-D against CCSD(T) reference values from the S22 dataset.

	Interaction Energy (kcal/mol)				
	LDA	BLYP	B3LYP	B3LYP-D	CCSD(T)
AT Stack	-6.42	5.39	3.53	-7.71	-8.10
AT Pair	-22.90	-11.61	-13.09	-17.77	-16.40
GC Stack	-4.65	-0.23	-1.20	-6.24	-7.90
GC Pair	-44.00	-27.77	-30.60	-36.41	-35.80
RMSE	5.54	9.06	7.38	1.14	

As can be seen in Figure 4-2 and Table 4-1, the B3LYP-D and LDA functionals are in excellent agreement with the benchmark values; however, the BLYP and B3LYP methods yield more repulsive energies (i.e., more positive values) with errors larger than 2.0 kcal/mol. BLYP, a GGA functional, significantly reduces the over-binding tendency of LDA, which is exacerbated in monomer pairs (as opposed to stacks) since hydrogen bonding is prevalent in pairs but absent in stacks. The B3LYP and BLYP functionals underestimate binding energies compared to B3LYP-D since they do not include attractive dispersion interactions. While hybrid functionals such as B3LYP have long-range (nonlocal) effects through Hartree-Fock exchange, they remain local in correlation and, therefore, are unable to describe the R^{-6} asymptotic distance-dependence of dispersion forces correctly.⁴⁷

The B3LYP functional gives weaker binding energies than the CCSD(T) benchmarks, and prior work by Zhang et al. suggested that this under-binding becomes more pronounced with increasing molecular size.⁴⁸ As such, B3LYP will incur significant errors for the large periodic strands, as we demonstrate in the next section. Based on these benchmark calculations, the B3LYP-D functional most closely matches the CCSD(T) results, particularly for the van-der-Waals-stacked monomers.

4.4.2 Optimized Geometries of Single- and Double-Stranded DNA

With the individual nucleotide benchmarks calculated, we next optimized the geometries of various ssDNA and dsDNA systems: periodic adenine (poly(A)), thymine (poly(T)), cytosine (poly(C)), guanine (poly(G)), adenine-thymine (poly(A-T)), and guanine-cytosine (poly(G-C)). Figure 4-3 depicts magnified views of the optimized

geometries for the periodic poly(G-C) strands obtained with the LDA, BLYP, B3LYP, and B3LYP-D functionals. As seen in this figure, only LDA and B3LYP-D give structurally stable geometries (Cartesian coordinates for all of the B3LYP-D-optimized ssDNA and dsDNA structures can be found in the Supporting Information). In contrast, both the BLYP and B3LYP functionals produce geometries that are highly distorted in which the individual Watson-Crick base pairs are not even aligned in the same plane.

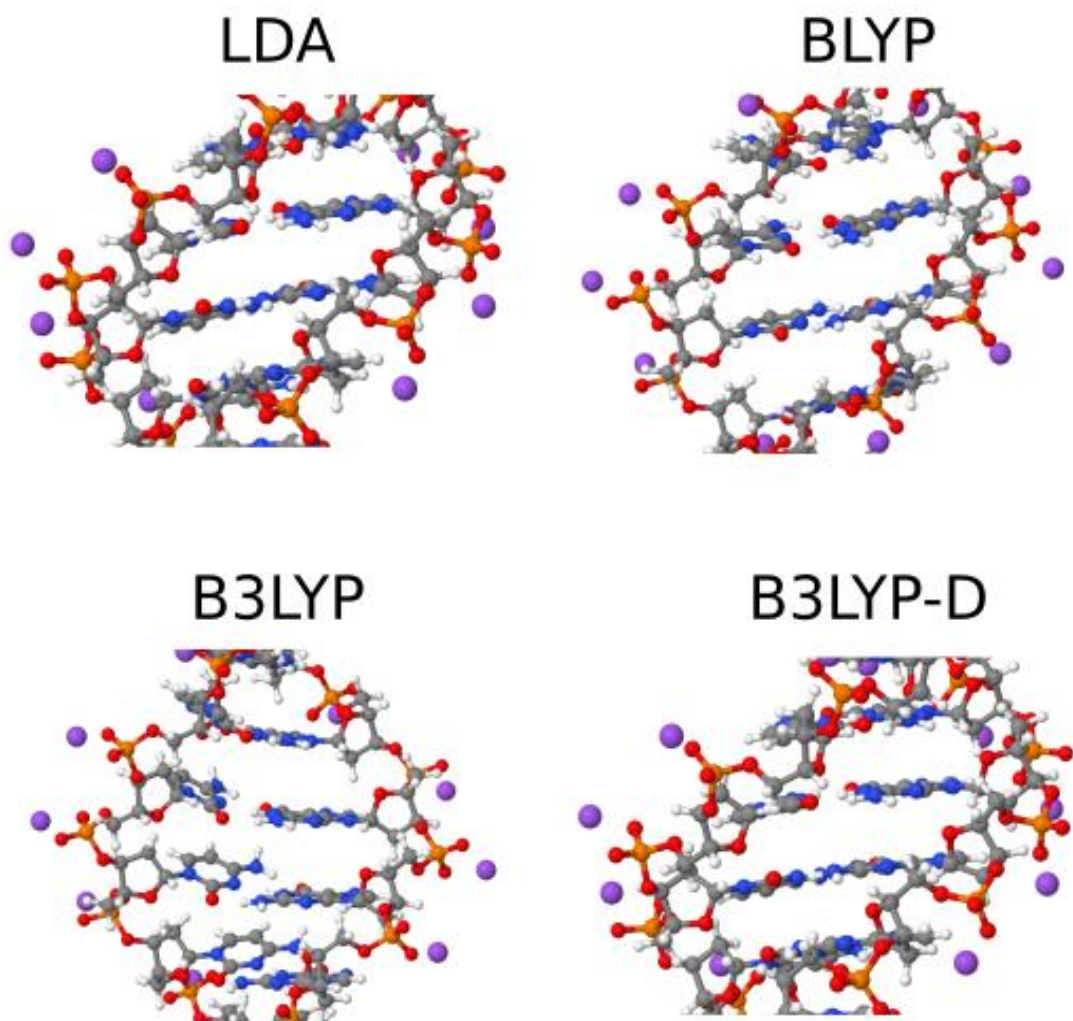


Figure 4-3. Geometries of periodic poly(G-C) obtained with the LDA, BLYP, B3LYP, and B3LYP-D functionals. Only LDA and B3LYP-D give stable structures, whereas the other functionals give unstable and distorted geometries between adjacent Watson-crick pairs.

The structural deformations in these periodic strands are fully consistent with the benchmark calculations described in the previous section. In particular, our benchmark calculations on individual nucleotides showed that only LDA and B3LYP-D predict stable A-T and C-G stacks/Watson-Crick pairs in comparison to the CCSD(T) benchmarks. In contrast, both BLYP and B3LYP considerably underestimate these

interaction energies. As a result, the under-binding tendencies in BLYP and B3LYP become even more pronounced in the periodic systems (since *both* stacking and Watson-Crick pairs are now present in the periodic system), leading to the geometric distortions seen in Figure 4-3. While Figure 4-3 only depicts the poly(G-C) strands for brevity, we observed similar geometric trends in poly(A-T) in which only LDA and B3LYP-D gave stable structures. Our results also corroborate previous molecular dynamics simulations on a DNA dodecamer, suggesting that the double-helical structure is unstable when dispersion interactions are not incorporated.⁴⁹

4.4.3 Electron/Hole Mobilities

With the geometries of all the periodic strands optimized, we now analyze electron/hole mobilities for the various ssDNA and dsDNA systems. For clarity, we only discuss electron/hole mobilities calculated at the B3LYP-D level of theory since this functional simultaneously gives stable geometries and reasonable band gaps⁵⁰⁻⁵² (LDA also gave stable geometries in our study but is well known for underestimating bandgaps). Table 4-2 presents the lattice parameters, bandgaps, and electron/hole masses (required for calculating electron/hole mobilities from (Eq. 4-1)), and

Table 4-3 summarizes the electronic charge per nucleobase and the electron/hole mobilities of all ssDNA and dsDNA structures (electron/hole mobilities for other functionals are given in the Supplementary Information).

Table 4-2. Lattice parameters, band gaps, and effective masses of holes/electrons of various single- and double-strand DNA systems computed at the B3LYP-D/6-311g(d,p) level of theory.

System	Lattice Parameter (Å)	Band Gap (eV)	m_e^* (m_0)	m_h^* (m_0)
poly(A)	32.22	3.92	7.39	10.13
poly(T)	29.70	3.12	32.02	2.00
poly(G)	31.39	3.66	10.12	50.36
poly(C)	30.74	3.22	17.86	3.25
poly(A-T)	31.29	3.23	8.86	17.74
poly(G-C)	32.45	1.39	22.74	2.69

Table 4-3. Electronic charge per nucleobase and electron/hole mobilities of various single- and double-strand DNA systems computed at the B3LYP-D/6-311g(d,p) level of theory.

System	Electronic Charge per Nucleobase (e)	Electron mobility (cm ² /V·s)	Hole mobility (cm ² /V·s)
poly(A)	-0.26 (A)	18.22	9.72
poly(T)	-0.20 (T)	4.08	22.33
poly(G)	-0.24 (G)	15.77	0.54
poly(C)	-0.18 (C)	4.09	39.81
poly(A-T)	-0.23 (A), -0.24 (T)	40.23	5.14
poly(G-C)	-0.21 (G), -0.17 (C)	9.59	19.38

Figure 4-4 shows that the highest occupied crystal orbitals (HOCOs) are localized on the nucleobase (regardless of nucleobase species) in all of the B3LYP-D-optimized ssDNA structures. In contrast, the lowest unoccupied crystal orbitals (LUCOs) are primarily found on the Na⁺ cations and the phosphate backbone. Both of these HOCO and LUCO localization patterns are consistent with previous work³⁵⁻³⁶ which used Hartree-Fock calculations and small basis sets. Moreover, since the spatial distribution of the HOCO influences the hole mobility, our calculations predict that hole transport in ssDNA occurs intramolecularly across the nucleobase stacks, whereas electron transport (which is determined by the LUCO) occurs across the Na⁺ cations and phosphate backbone.

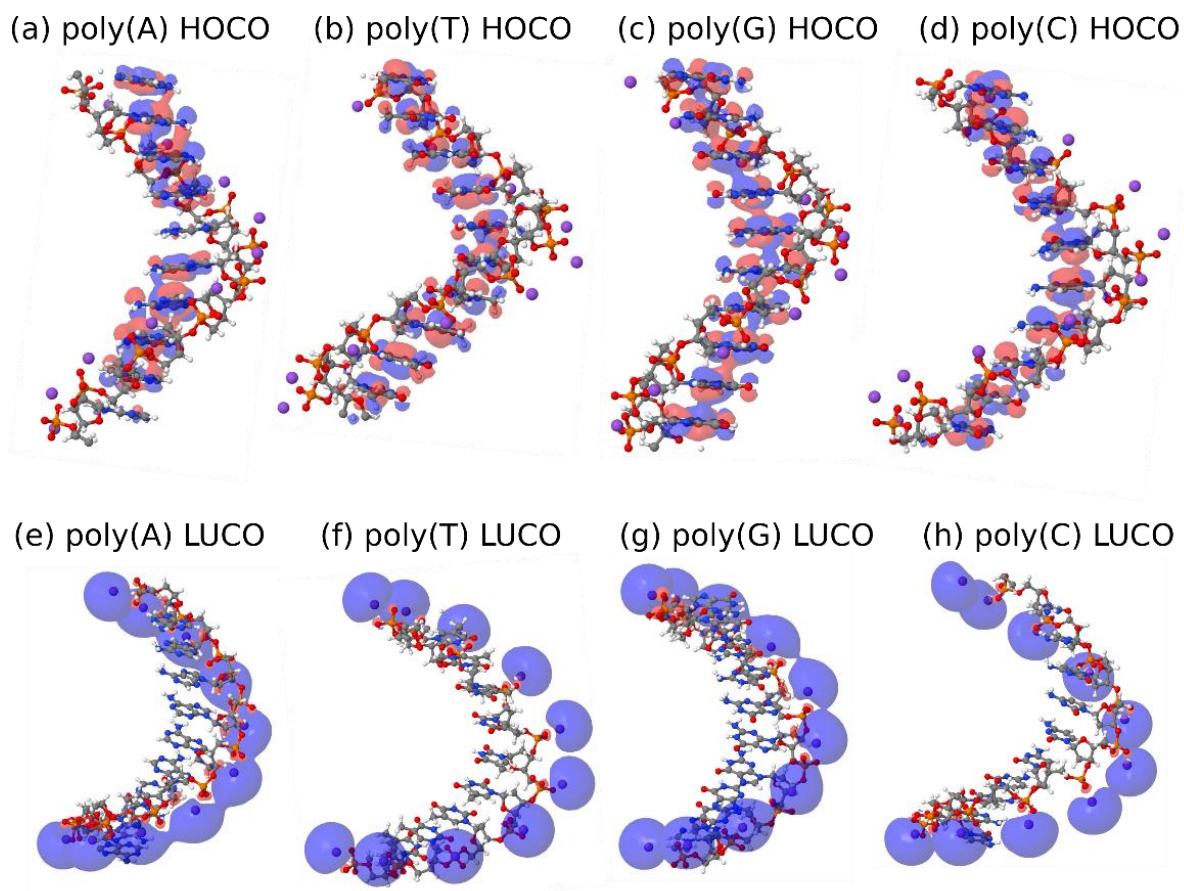


Figure 4-4. HOCOs and LUCOs of ssDNA obtained with the B3LYP-D functional. The HOCOs and LUCOs were calculated at isovalues of 0.01 and 0.03, respectively.

It is worth noting that the ssDNA structures with purine nucleobases (i.e., A and G) have a significantly lower hole mobility than the corresponding structures with pyrimidine nucleobases (i.e., T and C). This is due to the HOCOs in poly(A) and poly(G) being formed from the highest occupied molecular orbitals (HOMOs) of A and G (cf. Figure 4-5), which have an *antibonding* interaction in the helical ssDNA stacked geometry.³⁴ Conversely, the ssDNA structures with purine nucleobases have a significantly higher electron mobility than the corresponding structures with pyrimidine nucleobases. This trend arises from the LUCOs in poly(A) and poly(G) having a larger overlap than the corresponding LUCOs in poly(T) and poly(C), as can be seen in Figure

4-4. It is also interesting to note that the electronic charge per nucleobase (cf. Figure 4-3) is also correlated with the electron/hole mobility in the ssDNA structures. While our DFT calculations indicate that the nucleobases in all four of the ssDNA structures are negatively charged, nucleobases with the most negative charge (A and G) exhibit the highest electron mobilities, whereas nucleobases with the least negative charge (T and C) have the highest hole mobilities. Taken together, the HOCO/LUCO interactions and electronic charges in these ssDNA structures result in poly(T) and poly(C) being hole conductors, whereas poly(A) and poly(G) structures are better electron conductors.

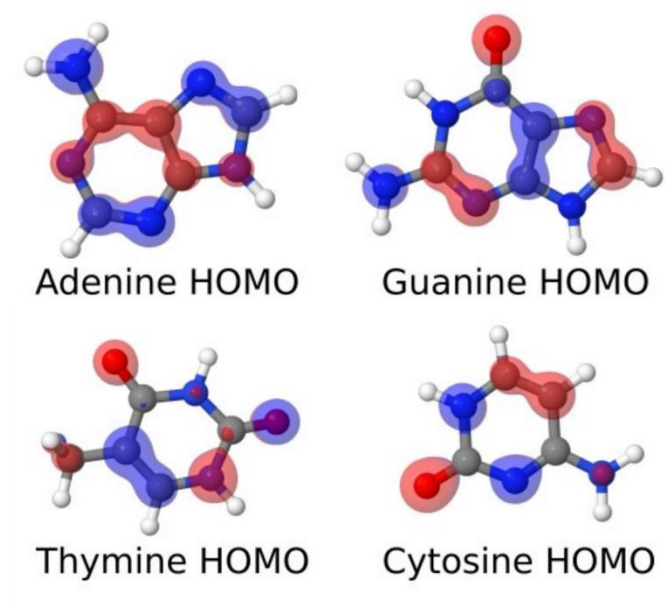
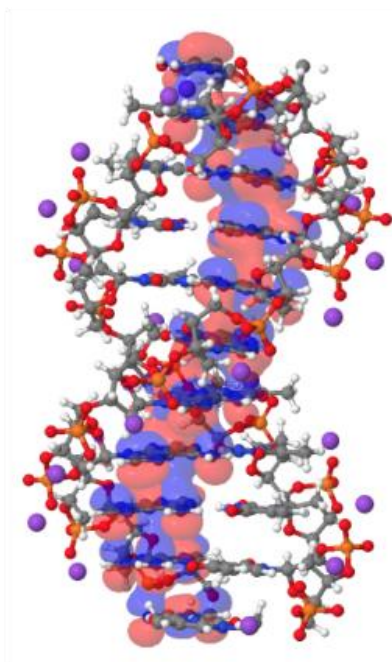


Figure 4-5. HOMOs of A, T, G, and C calculated at the B3LYP-D/6-311g(d,p) level of theory.

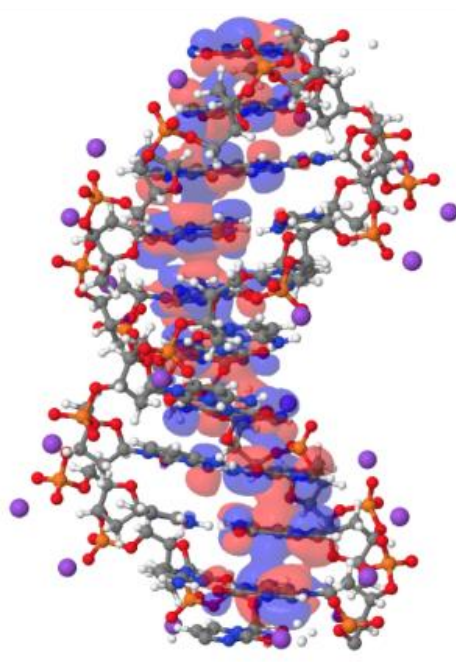
Turning our attention to the dsDNA structures, Figure 4-6 shows that the HOCOs on these systems are localized on the A and G nucleobases in the poly(A-T) and poly(G-C) structures, respectively. As such, our calculations predict that hole transport in dsDNA occurs only across the purine (and not the pyrimidine) nucleobases in both these

structures. In contrast, the LUCOs are localized on the Na^+ cations and the phosphate groups (proximal to the pyrimidine bases, C and T), indicating that electron transport occurs along the backbone in both structures. To visualize the HOCOs and LUCOs more easily, the Supporting Information provides 3D animations of these orbitals for the poly(A-T) structure.

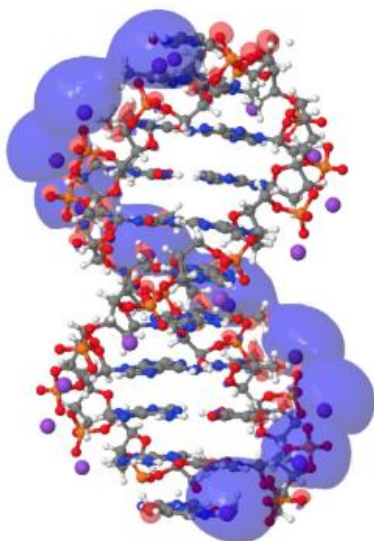
(a) poly(A-T) HOCO



(b) poly(G-C) HOCO



(c) poly(A-T) LUCO



(d) poly(G-C) LUCO

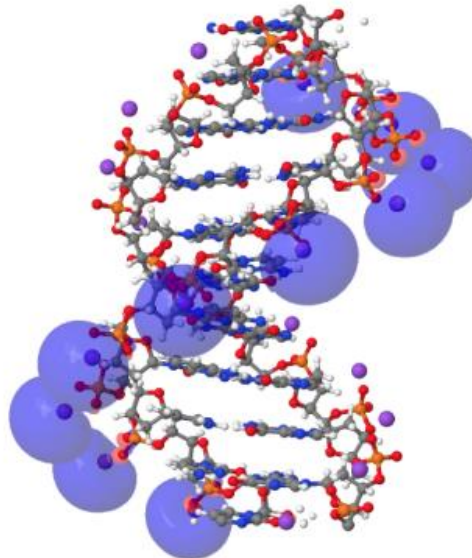


Figure 4-6. HOCOs and LUCOs of poly(A-T) and poly(G-C) calculated at the B3LYP-D/6-311g(d,p) level of theory.

To understand the electronic interactions in these dsDNA structures more closely, Figure 4-7 plots the electronic band structures of poly(A), poly(T), poly(A-T), poly(G), poly(C), and poly(G-C) calculated at the B3LYP-D/6-311g(d,p) level of theory. It is worth noting that the electronic properties of poly(A-T) and poly(G-C) are more complex than their constituents and *are not* merely superpositions of the individual ssDNA poly(A)+poly(T) or poly(G)+poly(C) band structures. In particular, the right-most column of Figure 4-7 shows that the A and G purine-type bands are pushed significantly upwards in energy within the double-stranded poly(A-T) and poly(G-C) structures. As mentioned previously, the nucleobases in all four of the ssDNA structures are negatively charged, and the Coulombic repulsion between these nucleobases within the compact dsDNA structure results in an upward shift (i.e., a destabilization) of the A and G bands. As such, the renormalization of the dsDNA band structures causes the highest-filled orbitals to be only localized on the A and G nucleobases, which corroborates the HOCO-localization trends depicted in Figure 4-6.

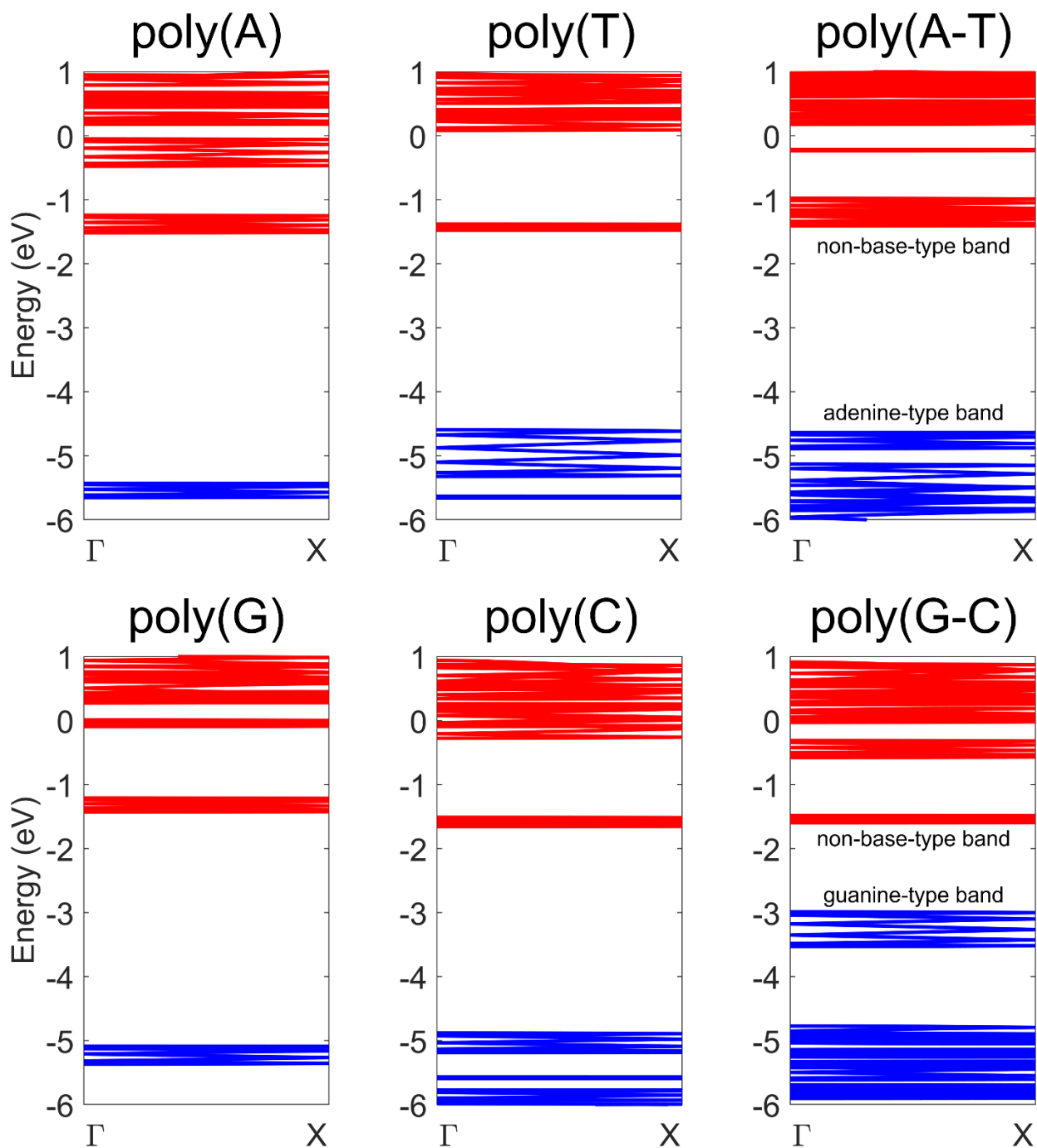


Figure 4-7. Band structures of poly(A), poly(T), poly(A-T), poly(G), poly(C), and poly(G-C) calculated at the B3LYP-D/6-311g(d,p) level of theory. The A- and G-type bands are pushed upwards in the double-stranded poly(A-T) and poly(G-C) and cases, respectively.

Finally, it is worth mentioning that our calculations predict poly(A-T) to be an electron conductor, whereas poly(G-C) is a better hole conductor. This trend can be seen in Figure 4-6(c) and (d), which show the LUCOs in poly(A-T) having a larger overlap than the corresponding LUCOs in poly(G-C). The electronic charge per nucleobase in the dsDNA structures (cf. Table 3) also corroborates these trends in the electron/hole mobilities. As mentioned in our previous analysis of ssDNA structures, nucleobases with the most negative charge exhibit the highest electron mobilities, whereas nucleobases with the least negative charge have higher hole mobilities. The total charge per A-T and G-C Watson-Crick pair in the poly(A-T) and poly(G-C) structures is $-0.47e$ and $-0.38e$, respectively, which reflects the trends in electron/hole mobilities discussed previously. It is interesting to point out that we also observed similar electron/hole mobility trends with LDA since this functional also predicts accurate dsDNA geometries (even though LDA predicts severely underestimated bandgaps compared to B3LYP-D). In contrast, the BLYP and B3LYP results give spurious results for electron/hole mobilities since these functionals produced deformed dsDNA geometries (cf. Figure 4-3). As such, these results emphasize the importance of including dispersion effects when calculating electronic properties on self-consistent optimized geometries (using the same functional) for these systems.

4.5 Summary and Conclusions

In conclusion, we have carried out large-scale DFT calculations to systematically analyze the structural and electron/hole transport properties of fully periodic single- and double-stranded DNA. To understand how these periodic nucleobase structures are

affected by their optimized geometry and underlying electronic structure, we examined the performance of various exchange-correlation functionals, ranging from local (LDA), semilocal (BLYP), hybrid (B3LYP), and dispersion-corrected hybrid (B3LYP-D) methods. Most notably, the latter calculations constitute the first study of periodic DNA and nucleobase structures using dispersion-corrected hybrid functionals for both full geometry optimizations and electronic band structures.

With these optimized geometries and band structures, we used the deformation potential formalism to calculate electron/hole mobilities for all of the various ssDNA and dsDNA structures. Our analysis showed that poly(T) and poly(C) are hole conductors, whereas poly(A) and poly(G) structures are better electron conductors. For the dsDNA structures, we found that the poly(A-T) and poly(G-C) band structures are more than just the “sum of their parts.” Specifically, Coulombic repulsion between the nucleobases results in a significant renormalization of the band structure, which causes the highest-filled orbitals (and, hence, hole transport) to be only localized on the A and G nucleobases. Further analyses of the B3LYP-D orbitals and electronic charges in the dsDNA structures show that poly(A-T) is an electron conductor, whereas poly(G-C) is a better hole conductor. Our calculations also highlight the importance of including dispersion effects when calculating electronic properties for these systems since functionals without dispersion will produce deformed dsDNA geometries with spurious electron/hole mobilities.

Finally, while our work focused on optimized structures and electron/hole mobilities of single- and double-stranded DNA, we anticipate that our calculations could

also be applied to other DNA-based materials and applications. In particular, the self-consistent geometries, band structures, and electron/hole mobilities from our B3LYP-D calculations could serve as new reference benchmarks to parameterize other coarse-grained DNA models or QM/MM studies,⁵³ which require accurate electronic properties as input parameters to enable larger-scale simulations.

Reference

1. Green, L. N.; Subramanian, H. K. K.; Mardanlou, V.; Kim, J.; Hariadi, R. F.; Franco, E., Autonomous dynamic control of DNA nanostructure self-assembly. *Nature Chemistry* **2019**, *11* (6), 510-520.
2. Li, Z.; Wang, J.; Li, Y.; Liu, X.; Yuan, Q., Self-assembled DNA nanomaterials with highly programmed structures and functions. *Materials Chemistry Frontiers* **2018**, *2* (3), 423-436.
3. Ke, Y.; Ong, L. L.; Shih, W. M.; Yin, P., Three-Dimensional Structures Self-Assembled from DNA Bricks. *Science* **2012**, *338* (6111), 1177-1183.
4. Keren, K.; Berman, R. S.; Buchstab, E.; Sivan, U.; Braun, E., DNA-Templated Carbon Nanotube Field-Effect Transistor. *Science* **2003**, *302* (5649), 1380-1382.
5. Marini, M.; Piantanida, L.; Musetti, R.; Bek, A.; Dong, M.; Besenbacher, F.; Lazzarino, M.; Firrao, G., A Reversible, Autonomous, Self-Assembled DNA-Origami Nanoactuator. *Nano Letters* **2011**, *11* (12), 5449-5454.
6. O'Brien, E.; Holt, M. E.; Thompson, M. K.; Salay, L. E.; Ehlinger, A. C.; Chazin, W. J.; Barton, J. K., The [4Fe4S] cluster of human DNA primase functions as a redox switch using DNA charge transport. *Science* **2017**, *355* (6327), eaag1789.
7. Genereux, J. C.; Barton, J. K., Mechanisms for DNA Charge Transport. *Chemical Reviews* **2010**, *110* (3), 1642-1662.
8. Genereux, J. C.; Boal, A. K.; Barton, J. K., DNA-Mediated Charge Transport in Redox Sensing and Signaling. *Journal of the American Chemical Society* **2010**, *132* (3), 891-905.
9. Dai, X.; Li, Q.; Aldalbahi, A.; Wang, L.; Fan, C.; Liu, X., DNA-Based Fabrication for Nanoelectronics. *Nano Letters* **2020**, *20* (8), 5604-5615.
10. Taniguchi, M.; Kawai, T., DNA electronics. *Physica E: Low-dimensional Systems and Nanostructures* **2006**, *33* (1), 1-12.
11. Tapio, K.; Leppiniemi, J.; Shen, B.; Hytönen, V. P.; Fritzsche, W.; Toppari, J. J., Toward Single Electron Nanoelectronics Using Self-Assembled DNA Structure. *Nano Letters* **2016**, *16* (11), 6780-6786.
12. Vittala, S. K.; Han, D., DNA-Guided Assemblies toward Nanoelectronic Applications. *ACS Applied Bio Materials* **2020**, *3* (5), 2702-2722.

13. Church, G. M.; Gao, Y.; Kosuri, S., Next-Generation Digital Information Storage in DNA. *Science* **2012**, *337* (6102), 1628-1628.
14. Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B.; Birney, E., Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **2013**, *494* (7435), 77-80.
15. Lin, K. N.; Volkel, K.; Tuck, J. M.; Keung, A. J., Dynamic and scalable DNA-based information storage. *Nature Communications* **2020**, *11* (1), 2981.
16. Murphy, C. J.; Arkin, M. R.; Jenkins, Y.; Ghatlia, N. D.; Bossmann, S. H.; Turro, N. J.; Barton, J. K., Long-Range Photoinduced Electron Transfer Through a DNA Helix. *Science* **1993**, *262* (5136), 1025-1029.
17. Hall, D. B.; Holmlin, R. E.; Barton, J. K., Oxidative DNA damage through long-range electron transfer. *Nature* **1996**, *382* (6593), 731-735.
18. Dandliker, P. J.; Holmlin, R. E.; Barton, J. K., Oxidative Thymine Dimer Repair in the DNA Helix. *Science* **1997**, *275* (5305), 1465-1468.
19. Núñez, M. E.; Hall, D. B.; Barton, J. K., Long-range oxidative damage to DNA: Effects of distance and sequence. *Chemistry & Biology* **1999**, *6* (2), 85-97.
20. Kelley, S. O.; Barton, J. K., Electron Transfer Between Bases in Double Helical DNA. *Science* **1999**, *283* (5400), 375-381.
21. Holmlin, R. E.; Dandliker, P. J.; Barton, J. K., Charge Transfer through the DNA Base Stack. *Angewandte Chemie International Edition in English* **1997**, *36* (24), 2714-2730.
22. Kasumov, A. Y.; Kociak, M.; Guéron, S.; Reulet, B.; Volkov, V. T.; Klinov, D. V.; Bouchiat, H., Proximity-Induced Superconductivity in DNA. *Science* **2001**, *291* (5502), 280-282.
23. Porath, D.; Bezryadin, A.; de Vries, S.; Dekker, C., Direct measurement of electrical transport through DNA molecules. *Nature* **2000**, *403* (6770), 635-638.
24. de Pablo, P. J.; Moreno-Herrero, F.; Colchero, J.; Gómez Herrero, J.; Herrero, P.; Baró, A. M.; Ordejón, P.; Soler, J. M.; Artacho, E., Absence of dc-Conductivity in λ -DNA. *Physical Review Letters* **2000**, *85* (23), 4992-4995.
25. Gómez-Navarro, C.; Moreno-Herrero, F.; de Pablo, P. J.; Colchero, J.; Gómez-Herrero, J.; Baró, A. M., Contactless experiments on individual DNA molecules show no

evidence for molecular wire behavior. *Proceedings of the National Academy of Sciences* **2002**, *99* (13), 8484-8487.

26. Hawke, L. G. D.; Kalosakas, G.; Simserides, C., Electronic parameters for charge transfer along DNA. *The European Physical Journal E* **2010**, *32* (3), 291-305.

27. Klotsa, D.; Römer, R. A.; Turner, M. S., Electronic Transport in DNA. *Biophysical Journal* **2005**, *89* (4), 2187-2198.

28. Tassi, M.; Morphis, A.; Lambropoulos, K.; Simserides, C., RT-TDDFT study of hole oscillations in B-DNA monomers and dimers. *Cogent Physics* **2017**, *4* (1), 1361077.

29. Deng, A.; Li, H.; Bo, M.; Huang, Z.; Li, L.; Yao, C.; Li, F., Understanding atomic bonding and electronic distributions of a DNA molecule using DFT calculation and BOLS-BC model. *Biochemistry and Biophysics Reports* **2020**, *24*, 100804.

30. Olofsson, J.; Larsson, S., Electron Hole Transport in DNA. *The Journal of Physical Chemistry B* **2001**, *105* (42), 10398-10406.

31. Biswas, P. K.; Chakraborty, S., Targeted DNA oxidation and trajectory of radical DNA using DFT based QM/MM dynamics. *Nucleic Acids Research* **2019**, *47* (6), 2757-2765.

32. Woźniak, A. P.; Leś, A.; Adamowicz, L., Theoretical modeling of DNA electron hole transport through polypyrimidine sequences: a QM/MM study. *Journal of Molecular Modeling* **2019**, *25* (4), 97.

33. Landi, A.; Capobianco, A.; Peluso, A., The Time Scale of Electronic Resonance in Oxidized DNA as Modulated by Solvent Response: An MD/QM-MM Study. *Molecules* **2021**, *26* (18), 5497.

34. Taniguchi, M.; Kawai, T., Electronic structures of A^{S} - and B^{S} -type DNA crystals. *Physical Review E* **2004**, *70* (1), 011913.

35. Bende, A.; Bogár, F.; Ladik, J., Model calculations of the energy band structures of double stranded DNA in the presence of water and Na^+ ions. *Solid State Communications* **2011**, *151* (4), 301-305.

36. Bende, A.; Bogár, F.; Ladik, J., Hole mobilities of periodic models of DNA double helices in the nucleosomes at different temperatures. *Chemical Physics Letters* **2013**, *565*, 128-131.

37. Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; De La Pierre, M.; D'Arco, P.; Noël, Y.; Causà, M.; Rérat,

- M.; Kirtman, B., CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry* **2014**, *114* (19), 1287-1317.
38. Foster, M. E.; Wong, B. M., Nonempirically Tuned Range-Separated DFT Accurately Predicts Both Fundamental and Excitation Gaps in DNA and RNA Nucleobases. *Journal of Chemical Theory and Computation* **2012**, *8* (8), 2682-2687.
39. Raeber, A. E.; Wong, B. M., The Importance of Short- and Long-Range Exchange on Various Excited State Properties of DNA Monomers, Stacked Complexes, and Watson–Crick Pairs. *Journal of Chemical Theory and Computation* **2015**, *11* (5), 2199-2209.
40. Hohenberg, P.; Kohn, W., Inhomogeneous Electron Gas. *Physical Review* **1964**, *136* (3B), B864-B871.
41. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38* (6), 3098-3100.
42. Becke, A. D., Density-Functional Thermochemistry .3. The Role of Exact Exchange. *Journal of Chemical Physics* **1993**, *98* (7), 5648-5652.
43. Grimme, S., Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* **2004**, *25* (12), 1463-1473.
44. Bardeen, J.; Shockley, W., Deformation Potentials and Mobilities in Non-Polar Crystals. *Physical Review* **1950**, *80* (1), 72-80.
45. Shuai, Z.; Wang, L.; Song, C., Deformation Potential Theory. In *Theory of Charge Transport in Carbon Electronic Materials*, Shuai, Z.; Wang, L.; Song, C., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 67-88.
46. Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P., Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **2006**, *8* (17), 1985-1993.
47. Burns, L. A.; Mayagoitia, Á. V.-.; Sumpter, B. G.; Sherrill, C. D., Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *The Journal of Chemical Physics* **2011**, *134* (8), 084107.
48. Zhang, I. Y.; Wu, J.; Xu, X., Extending the reliability and applicability of B3LYP. *Chemical Communications* **2010**, *46* (18), 3057-3070.

49. Černý, J.; Kabeláč, M.; Hobza, P., Double-Helical → Ladder Structural Transition in the B-DNA is Induced by a Loss of Dispersion Energy. *Journal of the American Chemical Society* **2008**, *130* (47), 16055-16059.
50. Wong, B. M.; Ye, S. H., Self-assembled cyclic oligothiophene nanotubes: Electronic properties from a dispersion-corrected hybrid functional. *Physical Review B* **2011**, *84* (7), 075115.
51. Wong, B. M.; Cordaro, J. G., Electronic Properties of Vinylene-Linked Heterocyclic Conducting Polymers: Predictive Design and Rational Guidance from DFT Calculations. *The Journal of Physical Chemistry C* **2011**, *115* (37), 18333-18341.
52. Allec, S. I.; Wong, B. M., Inconsistencies in the Electronic Properties of Phosphorene Nanotubes: New Insights from Large-Scale DFT Calculations. *The Journal of Physical Chemistry Letters* **2016**, *7* (21), 4340-4345.
53. Chen, W.; Sun, L.; Tang, Z.; Ali, Z. A.; Wong, B. M.; Chang, C.-e. A., An MM and QM Study of Biomimetic Catalysis of Diels-Alder Reactions Using Cyclodextrins. *Catalysts* **2018**, *8* (2), 51.

Chapter 5. Understanding Electrooxidation of Furfural on Cu, Co-Spinel Oxides from Density Functional Theory

5.1 Abstract

Co-based spinel oxides are promising catalysts for the electrooxidation of biomass-derived 5-hydroxymethylfurfural (HMF) to attain high-value chemicals. However, the atomistic-level mechanism of the reaction on Co-based spinel oxides is not yet well understood due to the complex nature of the HMF electrooxidation reaction (HMFOR) involving the combined adsorption of organic molecules and hydroxides on the electrode surface. Inspired by a recently reported Cu, Co-spinel oxide (CuCo_2O_4), which shows greatly improved HMFOR activity, we employ periodic density functional theory (DFT) calculations to study the atomistic level mechanism of HMFOR and the site-specific roles of CuCo_2O_4 . From ab initio atomistic thermodynamics, we find that the stabilities of the surfaces at 1 atm H_2 and 300K follow the trend of $(100) > (110) > (111)$. More importantly, tetrahedral Cu sites are found to adsorb HMF with higher adsorption energy than Co sites. Furthermore, we present and compare the energy profile of two different pathways of HMFOR at the atomistic level and explain the rate differences in the intermediate production shown in the experiment. Our work provides important insights into the excellent HMFOR activity of CuCo_2O_4 and a basis for a further mechanistic understanding of HMFOR on CuCo_2O_4 and other spinel oxides.

5.2 Introduction

Oxidation of abundantly available biomass-derived 5-hydroxymethylfurfural (HMF) into value-added chemicals, such as 2,5-furan dicarboxylic acid (FDCA), enables the manufacture of a wide variety of valuable chemical products (e.g., medicines, polymers^{1,2}, or fine chemicals) without petroleum-based ingredients.³⁻⁷ Furthermore, HMF electrooxidation reaction (HMFOR) can be coupled with hydrogen evolution reaction (HER), replacing oxygen evolution reaction (OER).⁸ The coupling of HMFOR and HER enables achieving hydrogen production and HMF oxidation simultaneously, increasing the economic value of the overall electrochemical process.⁹ For these reasons, HMF is considered one of the crucial bio-renewable platform chemicals listed by the U.S. Department of Energy.¹⁰ There has been an immense interest in developing efficient electrocatalysts for HMFOR over the past decade.¹¹

In this context, Co-based spinel oxides have emerged as HMFOR electrocatalysts due to their abundant active sites and tunable defect structures.^{12,13} Co_3O_4 is well known and widely used electrocatalyst for various oxidation reactions¹⁴⁻¹⁶, and many researchers attempted to further enhance the electrocatalytic activity of Co-based spinel oxides by tuning their structures. For example, replacing tetrahedral Co^{2+} of Co_3O_4 with highly electronegative metal cations such as Ni^{2+} can induce octahedral Co^{3+} to lose its coordination with oxygen, thus enhancing the catalytic activity for HMFOR.^{11,17}

Recently, Lu et al. systematically investigated the catalytic activity of various Co-based spinel oxides, including Co_3O_4 , CuCo_2O_4 , CoAl_2O_4 , and ZnCo_2O_4 , and concluded that CuCo_2O_4 has record-high catalytic activity with four folds enhanced catalytic activity

compared to Co_3O_4 .¹⁸ However, despite the defined reaction pathway of HMF oxidation, in-depth mechanistic insight is still not yet clear without an understanding of the atomistic-level reaction mechanism.

As a step toward elucidating the atomistic-level mechanism and site-specific roles of CuCo_2O_4 for HMFOR, herein, we employ density functional theory (DFT), starting by investigating the most stable CuCo_2O_4 surface structures. Then we explore the adsorption conformation of HMF on the surface, which is a critical step in the catalytic process. After that, we further employ ab initio atomistic thermodynamics to calculate the energy profile and the atomistic level mechanism of the two possible reaction pathways for HMFOR. (Figure 5-1) We elucidate the nature of O-H and C-H activations to produce intermediates, water formation, and desorption throughout the catalytic cycle. These insights will provide a better understanding of the geometric site-specific roles of surface atoms in determining the selectivity and yield of the HMFOR pathway that may allow one to design an optimal electrocatalyst.

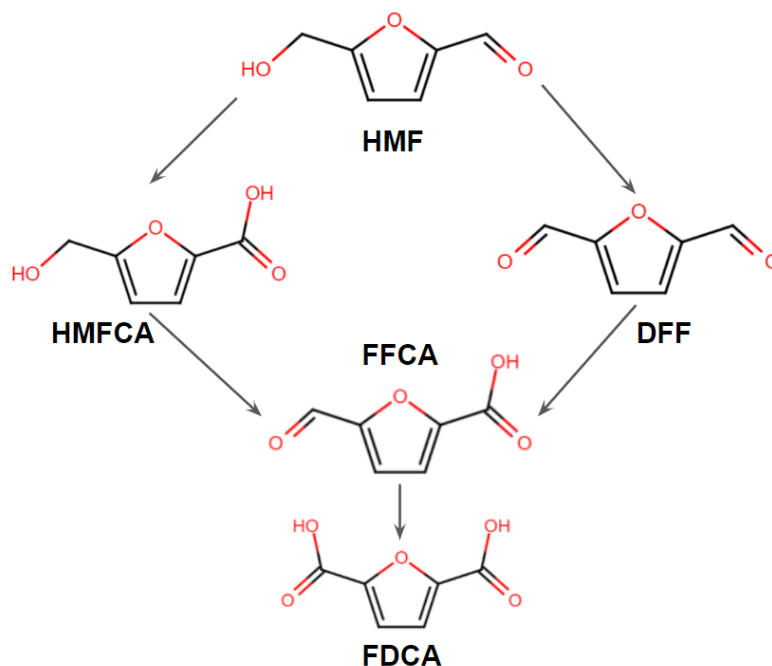


Figure 5-1. The two possible pathways for the oxidation of HMF to FDCA. (DFF: 2,5-diformylfuran; HMFCFA: 5-hydroxymethyl-2-furancarboxylic acid; FFCA: 5-formylfuran-2-carboxylic acid)

5.3 Computational Method

We perform spin-polarized DFT calculations using the Vienna *ab initio* simulation package (VASP)^{19,20} and the ion-electron interaction is described with the projector augmented wave (PAW) method²¹. The PBE functional^{22,23} with dispersion correction (PBE-D3)²⁴ is employed, and a cutoff energy of 520 eV is used. We employ Hubbard U (DFT+U) corrections for Co, and U = 3.32 eV is chosen from the Materials Project.²⁵

Three surface cleavages, (111), (110), and (100), with 11 different terminations, are examined. Each model consists of 7-9 layers of slabs with 20 Å of vacuum gap along the z-direction, and the Brillouin zone was sampled by (2×2×1) Monkhorst-Pack k-point

mesh. When calculating surface grand potentials (SGP), the middle three layers of slabs are fixed, while the top and bottom layers are allowed to relax.

The surface termination stability is quantified using SGP energies calculated from the following previous work by Wang et al.²⁶ First, SGP is calculated using the following equation:

$$\omega_i = \frac{1}{2S} [E_{slab}^i + PV - TS - N_{Co}\mu_{Co} - N_{Cu}\mu_{Cu}N_O\mu_O] \quad (\text{Eq. 5-1})$$

where ω_i is the SGP of the termination, and S represents the surface area of the slab model. N_i and μ_i is the number of elements present in the model and the chemical potential of the element, respectively.

Assuming that the PV-TS term is negligible in ambient conditions, (Eq. 5-1 is rewritten as:

$$\omega_i = \frac{1}{2S} [E_{slab}^i + N_{Co}\mu_{Co} - N_{Cu}\mu_{Cu}N_O\mu_O] \quad (\text{Eq. 5-2})$$

The chemical potential of each element is calculated from the following equations:

$$\mu_{CuCo_2O_4} = E_{CuCo_2O_4}^{bulk} = \mu_{Cu} + 2\mu_{Co} + 4\mu_O \quad (\text{Eq. 5-3})$$

$$\Delta\mu_{Cu} = \mu_{Cu} - E_{Cu}^{bulk} \quad (\text{Eq. 5-4})$$

$$\Delta\mu_{Co} = \mu_{Co} - E_{Co}^{bulk} \quad (\text{Eq. 5-5})$$

$$\Delta\mu_O = \mu_O - E_{O_2}^{gas} \quad (\text{Eq. 5-6})$$

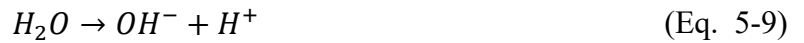
Substituting Eq. 5-3 – 5-6 into (Eq. 5-2 yields the following equation:

$$\omega_i = \varphi_i - \frac{1}{2S} [(N_{Co} - 2N_{Cu})\Delta\mu_{Co} + (N_O - 4N_{Cu})\Delta\mu_O] \quad (\text{Eq. 5-7})$$

$$\varphi_i = \frac{1}{2S} [(E_{stab}^i - N_{Cu}E_{CuCo_2O_4}^{bulk}) - (N_{Co} - 2N_{Cu})E_{Co}^{bulk} - (N_O - 4N_{Cu})\frac{E_{O_2}^{gas}}{2}] \quad (\text{Eq. 5-8})$$

On the other hand, when calculating the adsorption energy and thermodynamics of the reaction, the bottom four layers are fixed. In comparison, the top three layers are allowed to relax together with the adsorbed HMF molecule. The adsorption energies are defined by $E_{ads} = E_{surface+adsorbate} - (E_{surface} + E_{adsorbate})$. The energy of the adsorbate, $E_{adsorbate}$ is calculated by placing an adsorbate molecule in a cubic cell with a side length of 20 Å. Partial atomic charges are obtained using Bader charge analysis as implemented by Henkelman and coworkers.²⁷

All the electrochemical reactions are assumed to occur in an alkaline condition, according to the experiment.¹⁸ Under assumptions of equilibrium of the following reaction,



The chemical potential of the elements is as follows:

$$\mu_{H_2O} = \mu_{OH^-} + \mu_{H^+} \quad (\text{Eq. 5-10})$$

With the Computational Hydrogen Electrode (CHE) approach,

$$\mu_{H^+} + \mu_{e^-} = \frac{1}{2}\mu_{H_2} - eU_{RHE} \quad (\text{Eq. 5-11})$$

Combining equations, we obtain the following equation:

$$\mu_{OH^-} - \mu_{e^-} = \mu_{H_2O} - \left(\frac{1}{2}\mu_{H_2} - eU_{RHE}\right) \quad (\text{Eq. 5-12})$$

5.4 Results and Discussion

We start with optimizing the bulk structure of CuCo_2O_4 , followed by the clean surfaces, including (111), (110), and (100). We then examine and compare the stabilities of HMF adsorption conformations on the most stable surface.

5.4.1 Bulk CuCo_2O_4

Bulk CuCo_2O_4 has a space group of $Fd\bar{3}m$, with Cu and Co occupying tetrahedral and octahedral sites, respectively (Figure 5-2). The experimental²⁸ and calculated lattice parameters of bulk CuCo_2O_4 are listed in Table 1. One can see that the calculated values agree well with the experimental values.

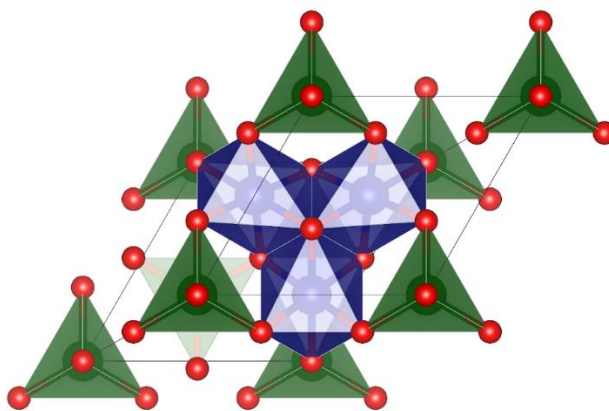


Figure 5-2. Bulk CuCo_2O_4 . Cobalts occupy octahedral sites, and Coppers occupy tetrahedral sites. Cu, green; Co, blue; O, red.

Table 5-1. Comparison of experimental²⁸ and calculated lattice parameters of the bulk CuCo₂O₄.

Lattice Parameter	a (Å)	b (Å)	c (Å)
Experimental²⁸	8.14	8.07	8.14
Calculated	8.12	8.12	8.12

5.4.2 Clean Surfaces

Low Miller-index surfaces are usually considered first in surface science studies due to their higher stabilities than higher index surfaces. Here, we consider three surface cleavages: (111), (110), and (100), and Figure 5-3 shows the structures of these clean CuCo₂O₄ surfaces.

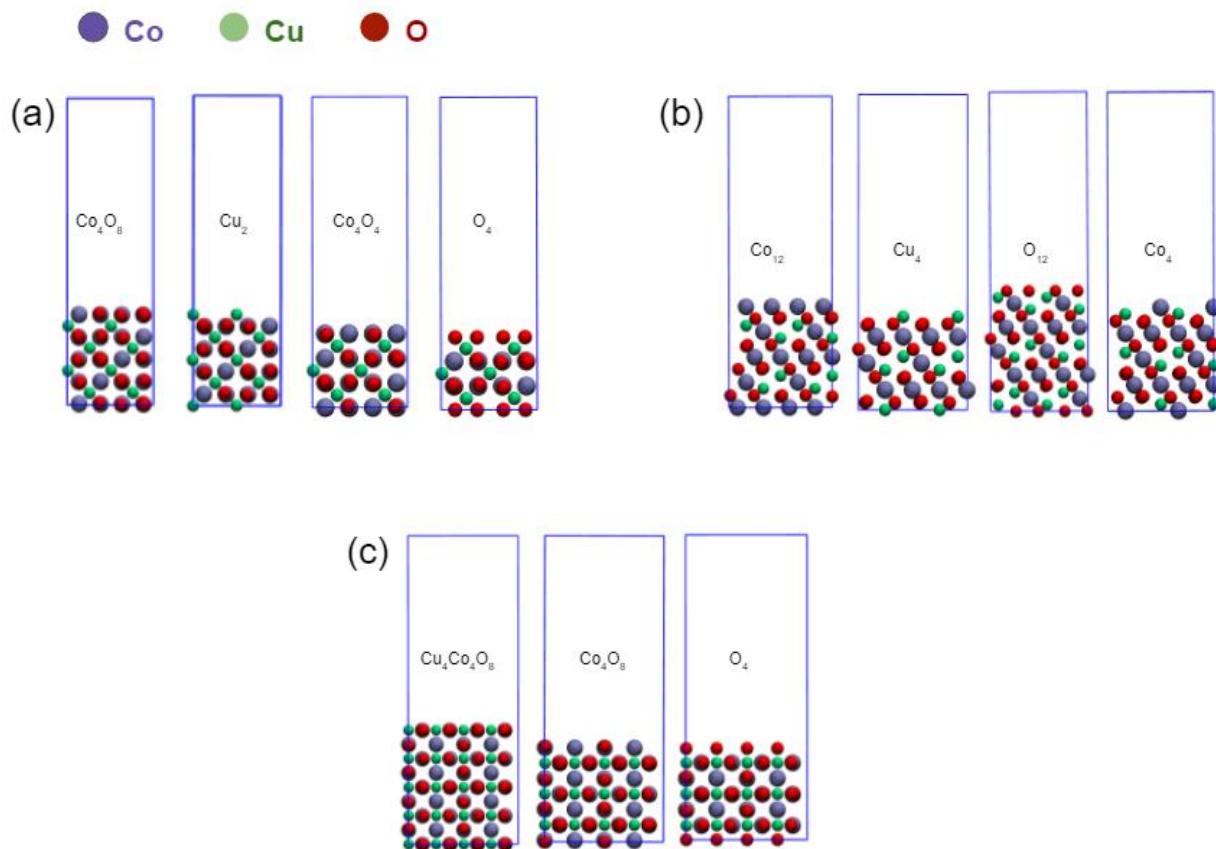


Figure 5-3. Top views of clean CuCo_2O_4 surfaces. Co, purple; Cu, green; O, red.

To evaluate the most stable surface cleavage and termination of CuCo_2O_4 at the experimental HMFOR conditions (1 atm, 300 K), we used ab initio atomistic thermodynamics to determine Surface Grand Potential, $SGP(T, p)$ as a function of cobalt chemical potential ($\Delta\mu_{\text{Co}}$) and hydrogen chemical potential ($\Delta\mu_{\text{H}}$) (or pressure at $T=300\text{K}$). Figure 5-4 shows the calculated $SPG(T, p)$ vs. $\Delta\mu_{\text{Co}}$ relationship: each line represents a given surface termination and $\Delta\mu_{\text{O}} = -0.27 \text{ eV}$ at 1 atm, 300K condition is highlighted yellow. One can see that (100) surfaces are the most stable, and (111)

surfaces are the least stable (Figure 5-4). More specifically, (100) cleavage with O_4 termination is the most stable among all 11 slab models. However, since HMF cannot adsorb on Oxygen sites, we focused on the second most stable slab model, Co_4O_4 termination, for the HMFOR mechanism.

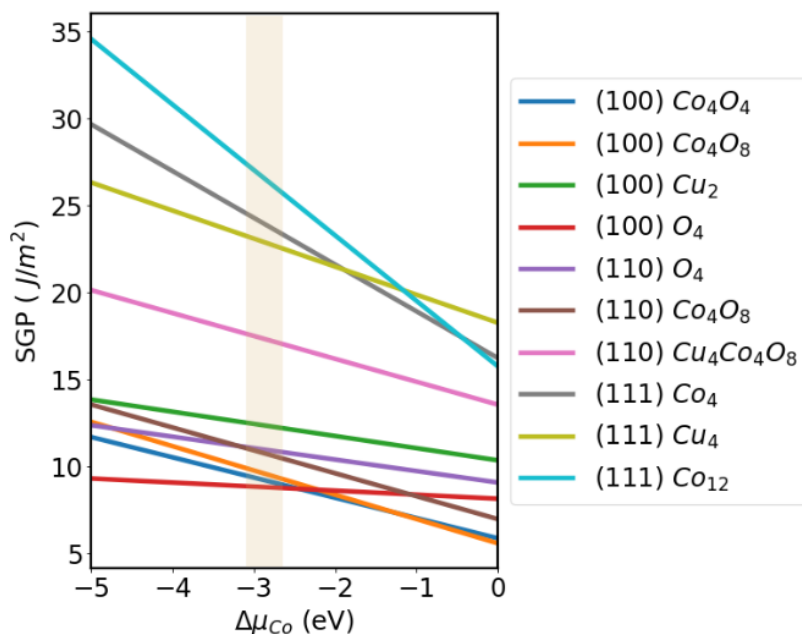


Figure 5-4. The surface grand potential of 11 surface models. $\Delta\mu_O = -0.27 eV$ at 1 atm, 300K condition is highlighted with yellow.

5.4.3 HMF Adsorption on the $CuCo_2O_4(100)$ Surface

Since the HMFOR process includes the oxidation of hydroxyl and aldehyde groups, different adsorption energies and hydroxide sites on catalysts significantly influence the HMFOR activity. Thus, the adsorption of HMF on Cu, Co-spinel oxide is a crucial starting point of the HMFOR mechanism. Furthermore, due to the presence of both hydroxyl and carbonyl groups, there are a few different conformations that HMF can bind to the $CuCo_2O_4$ surface.

Figure 5-5 shows the optimized structures of HMF adsorbed on CuCo_2O_4 (100) surface. After investigating eight different adsorption conformations, we find the three most stable HMF adsorption conformations on $\text{CuCo}_2\text{O}_4(100)$. The adsorption is mainly due to the coordination effect of O atoms of hydroxyl and carbonyl groups with the Co and Cu atoms of the surface. The first two conformations exhibit the hydroxyl group adsorbs on the Co and Cu sites (Figure 5-5a and b). In contrast, the final one shows dual adsorption, in which the O atom of the hydroxyl group adsorbs on the Cu site and the O atom of the carbonyl group adsorbs on the Co site simultaneously (Figure 5-5c). It is important to note that in the case of dual adsorption conformation, the hydroxyl group is readily dehydrogenated without the assistance of any active oxygen species, demonstrating that the activation of the O-H bond is highly favorable in the dual-adsorbed conformation. (Figure 5-5d)

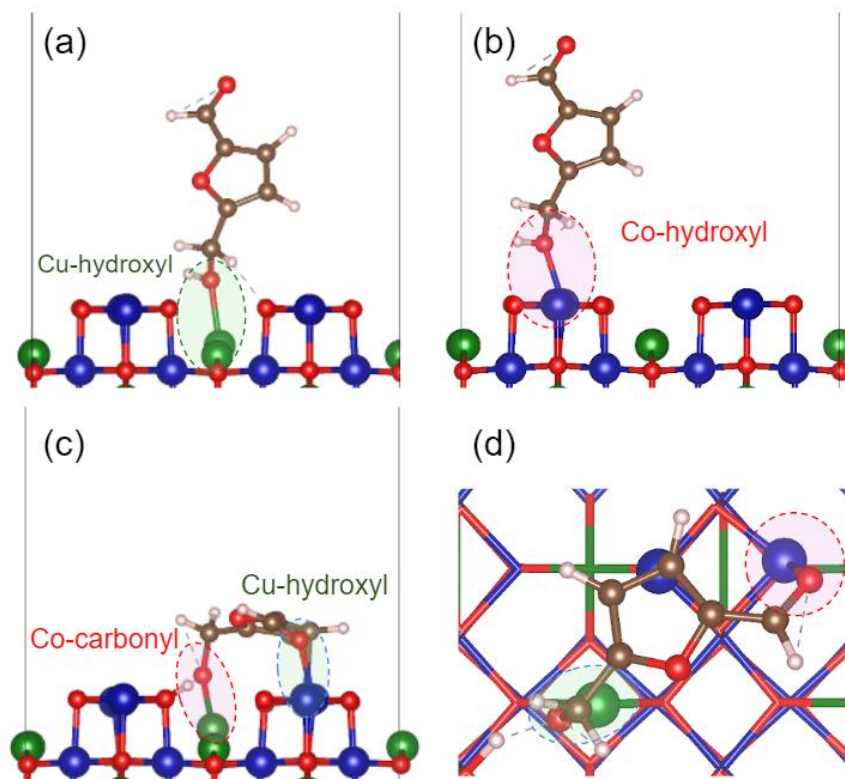


Figure 5-5. HMF adsorption conformation on CuCo₂O₄ surface. (a) O atom of the hydroxyl group adsorbs on Cu and (b) Co sites. (c) O atom of the hydroxyl group and carbonyl group adsorb on the Cu and Co sites, respectively. (d) Top view of (c).

Table 5-2 summarizes adsorption energies, the distances between the Cu (Co) site and the oxygen atom of HMF, and the number of electron transfers (Δq) from Cu (Co) from the Bader charge analysis. Adsorption energy is more significant when the hydroxyl group of HMF adsorbs on the subsurface Cu site (-1.31 eV) than on the surface Co site (-0.71 eV), which suggests that Cu sites are more favorable for HMF adsorption. This trend is consistent with the experimental observation from ref 18 that tetrahedral sites play an essential role in adsorption. Three possible explanations exist for the difference in adsorption energy on different metal sites. First, Cu sites have a lower coordination number (2) than Co (3). Second, Cu is more electronegative than Co, so Cu forms a stronger bond with the electron-rich O atom of the hydroxyl group. Third, Cu-adsorbed conformation allows hydrogen atoms on the $-\text{CH}_2$ group to form a hydrogen bond with adjacent surface O.

Table 5-2. HMF adsorption energies (E_{ads}) at the Cu (Co) sites, M-O bond length (r_{M-O}) (M = Cu, Co), and partial atomic charge transfer on Cu (Co) from Bader charge analysis (Δq).

Adsorption conformation	Cu-adsorbed	Co-adsorbed	Co, Cu-adsorbed
E_{ads} (eV)	-1.31	-0.71	-0.52
r_{M-O} (Å)	r_{Cu-O} : 2.02	r_{Co-O} : 1.97	r_{Co-O} : 1.90 r_{Cu-O} : 1.99
Δq (e)	Cu: +0.28	Co: +0.10	Co: +0.087 Cu: +0.095

5.4.4 Intermediate adsorption on the $\text{CuCo}_2\text{O}_4(100)$ surface

Using the same method, we calculate the most stable adsorption conformation of intermediates such as HMFCA, DFF, FFCA, and FDCA on the $\text{CuCo}_2\text{O}_4(100)$ surface. We systematically investigate all the possible adsorption conformations and conclude that

the conformations displayed in Figure 5-6 are the most stable conformation with the highest adsorption energies. Most intermediates are most stable when adsorbed on subsurface Cu and surface Co sites, except for DFF, which is the most stable when adsorbed on two surface Co sites. Unlike other intermediates, DFF has a planar structure due to the conjugated π bonds causing higher steric hindrance. Therefore, it is more stable when adsorbed on surface Co sites—one carbonyl group is adsorbed on the bridge site of Co, and another carbonyl group is adsorbed on the top of Co (Figure 5-6b). Also, it is worth noting that DFF has higher adsorption energy than HMFCFA, which affects the preference between two different HMFOR reaction pathways, which will be discussed in more detail in the next section.

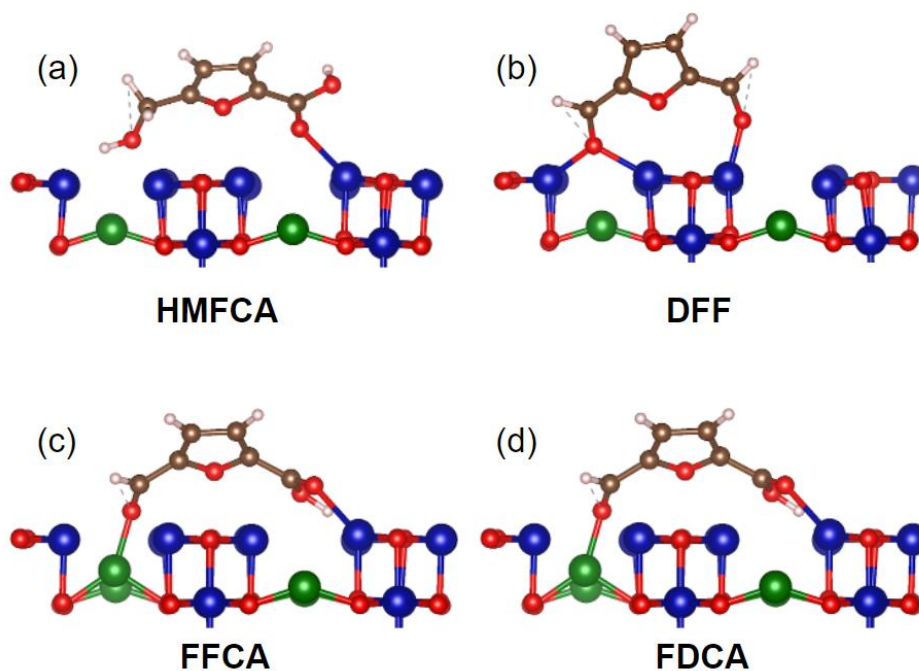


Figure 5-6. The most stable adsorption conformations of (a) HMFCFA, (b) DFF, (c) FFCA, and (d) FDCA on CuCo_2O_4 (100) surface.

Table 5-3. Adsorption energies (E_{ads}) of the intermediates (HMFCa, DFF, FFCA, and FDCA) and M-O bond length (r_{M-O}) (M = Cu, Co).

	HMFCa	DFF	FFCA	FDCA
E_{ads} (eV)	-0.86	-1.20	-1.89	-1.83
r_{M-O} (Å)	r_{Co-O} : 2.30	r_{Co-O} : 1.91	r_{Co-O} : 2.28	r_{Co-O} : 2.11
	r_{Cu-O} : 2.97		r_{Cu-O} : 1.92	r_{Cu-O} : 2.15

5.4.5 HMF to DFF oxidation reaction

Because HMF possesses both hydroxyl and carbonyl groups, HMFOR comprises two steps of oxidation: oxidation of the hydroxyl group into the carbonyl group and oxidation of the carbonyl group into the carboxylic acid group. Accordingly, there are two different pathways of HMF oxidation into FDCA, depending on which reaction step occurs first. For example, hydroxyl group oxidation can occur first, forming DFF, followed by carbonyl group oxidation, oxidizing into FFCA, which we refer to as pathway I. On the other hand, pathway II refers to the reaction that carbonyl group oxidation occurs first, forming HMFCa, and after that, the hydroxyl group is subsequently oxidized to form FFCA (Figure 5-7). To reflect the experimental conditions (1.0 M KOH) from ref 18, we assume reactions occur in alkaline conditions, allowing hydroxides to be the oxygen source.

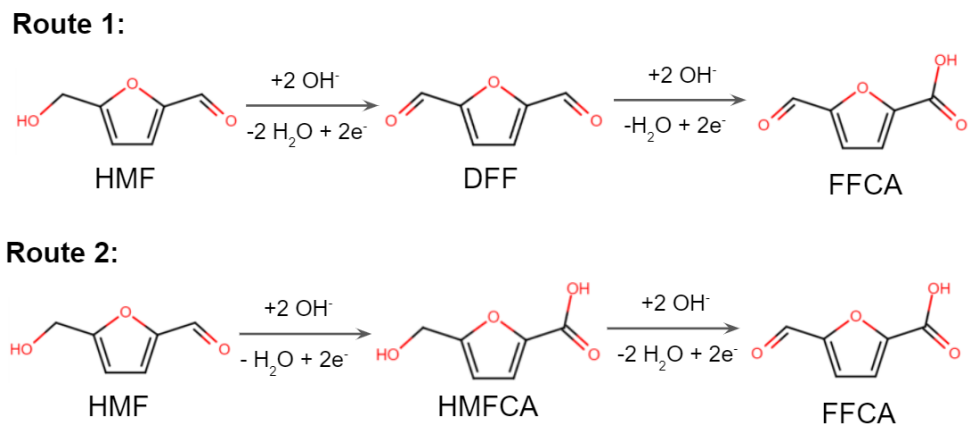


Figure 5-7. Two possible reaction pathways for HMF oxidation

We start investigating the first part of pathway I, HMF oxidizes into DFF. As mentioned in the previous section, the most stable HMF adsorption conformation is the hydroxyl group adsorbed on the Cu site.

In the first dehydrogenation step, as shown in Figure 5-8, adsorption of HMF is formed by adsorption of the O atom of the $-\text{CH}_2\text{OH}$ group on the top side of the Cu site. Then the OH^- adsorbs on an adjacent Co site and captures the dissociated hydrogen from O-H in the $-\text{CH}_2\text{OH}$ group and H_2O is generated. The first step of O-H bond scission ($\text{HMF}^* + \text{OH}^* \rightarrow \text{DFF-H}^* + \text{H}_2\text{O}^*$) on the Cu site, with the assistance of OH^* , is endothermic by 1.33 eV.

In the second dehydrogenation step, the $-\text{CH}_2\text{O}$ group of generated DFF-H^* is attacked by OH^- , which could capture the dissociated hydrogen from the $-\text{CH}_2\text{O}$ group and, similarly to the first step, generate an H_2O molecule. The C-H bond scission ($\text{DFF-H}^* + \text{OH}^* \rightarrow \text{DFF} + \text{H}_2\text{O}$) is exothermic by 1.67 eV. It is found that the oxidation of HMF to DFF on the CuCo_2O_4 surface is an exothermic process by 0.34 eV.

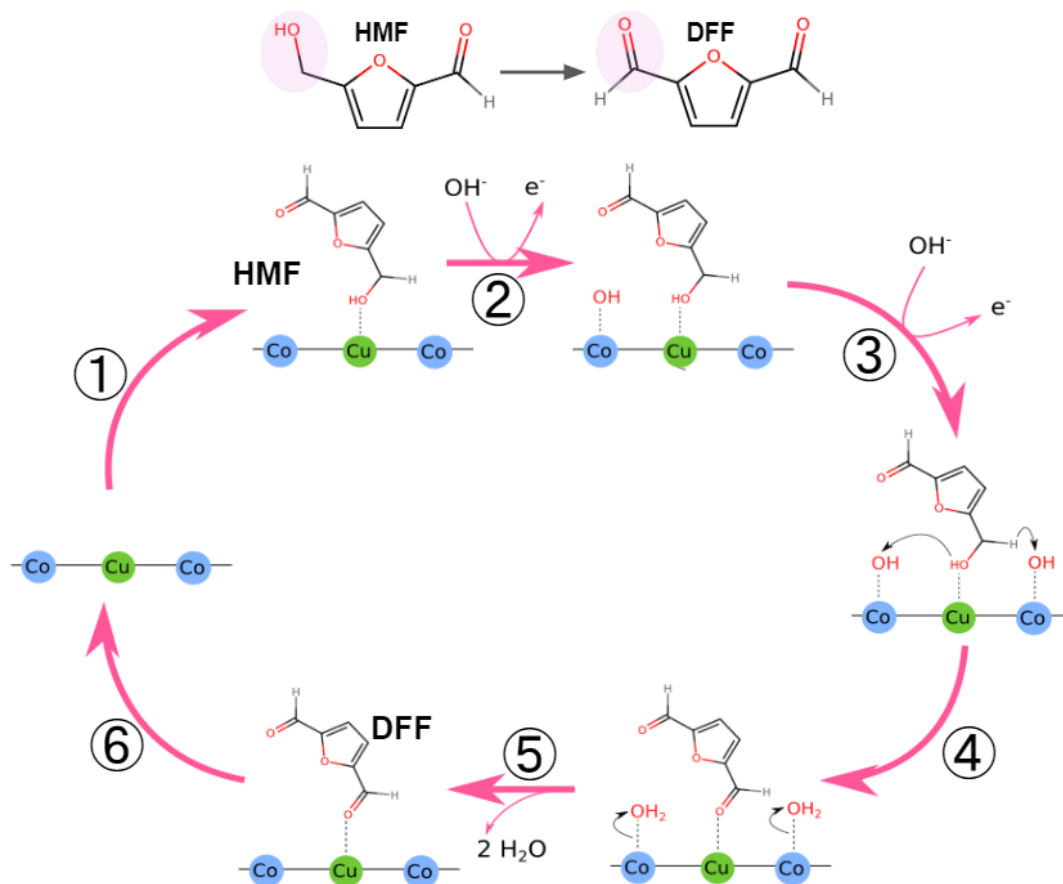


Figure 5-8. Schematic drawing of the complete catalytic cycle of oxidation of HMF to DFF on the CuCo₂O₄ (100) surface.

5.4.6 DFF to FFCA Oxidation Reaction

Then we move on to the second part of pathway I, DFF oxidizes into FFCA. As mentioned in the previous section, the most stable adsorption conformation of DFF is of which both carbonyl groups adsorb on the Co sites.

As shown in Figure 5-9, the OH⁻ adsorbs on an adjacent Co site and attacks the carbonyl group in the first oxidation step. The first step of OH addition (DFF* + OH* → FFCA-H*) on the Co site, with the assistance of OH*, is exothermic by 0.05 eV.

In the second dehydrogenation step, the -CHOOH group of generated FFCA-H* is attacked by OH⁻, which could capture the dissociated hydrogen from the -CHOOH group and, similarly to the first step, generate an H₂O molecule. The C-H bond scission (FFCA-H* + OH* → FFCA + H₂O) is endothermic by 0.14 eV. It can be found that the oxidation of DFF to FFCA on the CuCO₂O₄ surface is an exothermic process by -1.94 eV.

It is important to note that the DFF formation reaction pathway is mainly endothermic. In contrast, the DFF oxidation into the FFCA reaction pathway is mostly exothermic except for the C-H dehydrogenation step, which indicates that DFF is oxidized into FFCA as soon as formed. Therefore, our results agree well with experimental results showing that a negligible amount of DFF is detected throughout the HMF oxidation reaction.¹⁸

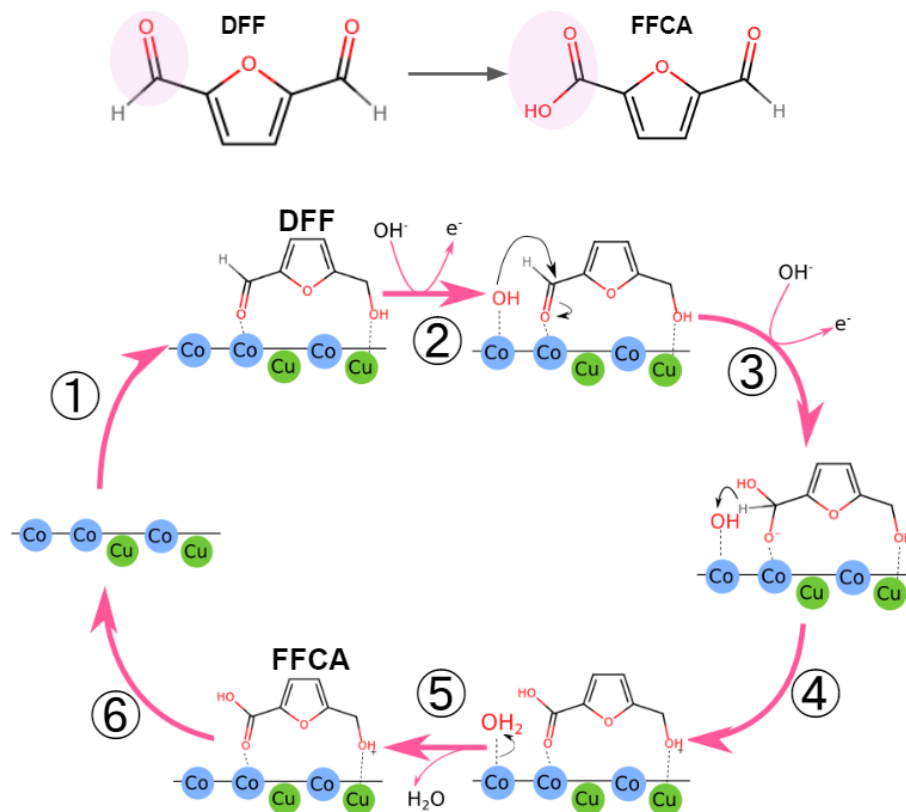


Figure 5-9. Schematic drawing of the complete catalytic cycle of oxidation of DFF to FFCA on the CuCo_2O_4 (100) surface.

5.4.7 HMF to HMFCFA

Next, we move to the first part of pathway II, where HMF oxidizes to HMFCFA. Since HMFCFA formation involves oxidation of the carbonyl group, the HMFCFA pathway starts from dual adsorption conformation—the hydroxyl group adsorbs on the Cu site, and the carbonyl group adsorbs on the Co site simultaneously.

The overall reaction pathway is similar to the second part of pathway I. As shown in Figure 5-10, the OH^- adsorbs on an adjacent Co site and attacks the carbonyl group in the first oxidation step. The first step of OH addition ($\text{HMF}^* + \text{OH}^* \rightarrow \text{HMFCFA-H}^*$) on the Co site, with the assistance of OH^* , is exothermic by 0.07 eV.

In the second dehydrogenation step, the -CHOOH group of generated HMFCa-H* is attacked by OH⁻, which could capture the dissociated hydrogen from the -CHOOH group and, similarly to the first step, generate an H₂O molecule. The C-H bond scission (HMFCa-H* + OH* → FFCA + H₂O) is exothermic by 1.19 eV.

The reaction energies of the first hydroxide addition step are both marginal (-0.05 eV and +0.07 eV) in both pathways I and II. However, the second dehydrogenation step reaction energy is -1.19 eV for pathway II, which is highly exothermic compared to pathway I (+0.14 eV). This energy profile indicates that pathway II is more thermodynamically favorable in this reaction.

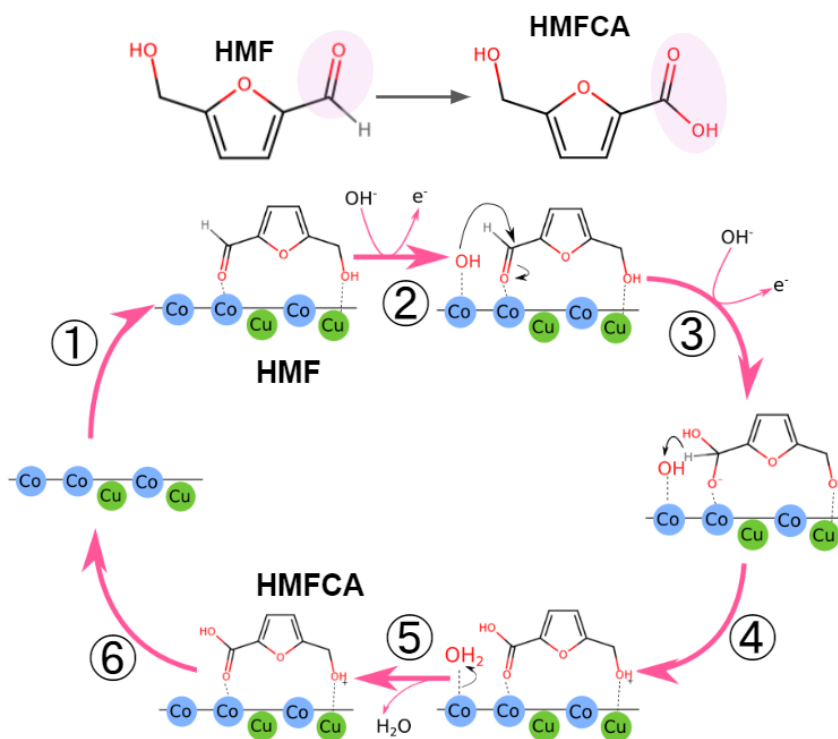


Figure 5-10. Schematic drawing of the complete catalytic cycle of oxidation of HMF to HMFCa on the CuCo₂O₄ (100) surface.

5.4.8 HMFCA to FFCA

Lastly, we move on to the second part of pathway II, HMFCA oxidizes to FFCA (Figure 5-11). The overall reaction pathway is similar to the first step of pathway I. However, the hydroxyl group is readily dehydrogenated when HMFCA is adsorbed on the surface. Therefore, in the first dehydrogenation step, the OH^- available in the solvent captures the dissociated hydrogen from O-H in the $-\text{CH}_2\text{OH}$ group, and H_2O is generated. The first step of O-H bond scission ($\text{HMFCa}^* + \text{OH}^* \rightarrow \text{FFCA-H}^* + \text{H}_2\text{O}^*$) on the Co site, with the assistance of OH^* , is exothermic by 0.39 eV.

In the second dehydrogenation step, the $-\text{CH}_2\text{O}$ group of generated DFF-H^* is attacked by OH^- , which could capture the dissociated hydrogen from the $-\text{CH}_2\text{O}$ group and generate an H_2O molecule. The C-H bond scission ($\text{HMFCa-H}^* + \text{OH}^* \rightarrow \text{FFCA} + \text{H}_2\text{O}$) is endothermic by 0.46 eV. It can be found that the oxidation of HMF to DFF on the CuCo_2O_4 surface is an exothermic process by 0.64 eV.

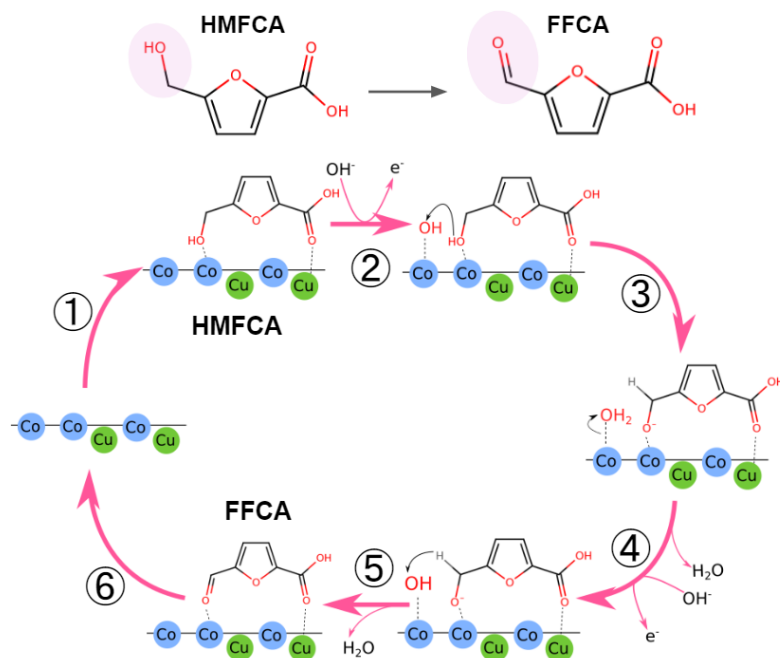


Figure 5-11. Schematic drawing of the complete catalytic cycle of oxidation of HMFCFA to FFCA on the CuCo₂O₄ (100) surface.

The first part (oxidation of HMF* forming DFF*) in the DFF pathway and the second step (oxidation of HMFCFA* forming *FFCA) in the HMFA pathway are highly possible to be the rate-determining step for each reaction route in the viewpoint of thermodynamics. (Figure 5-12) One can note that a slightly lower energy difference for the rate-determining step in the reaction of the HMFCFA pathway than that in the DFF pathway may indicate that the oxidation of HMF is more inclined to the HMFCFA pathway (HMF-HMFCFA-FFCA-FDCA). These results agree with experimental observations showing that only a negligible amount of DFF but a more considerable amount of HMFCFA is detected throughout the HMFOR.¹⁸

Furthermore, one of the most significant endothermic steps of both pathways is shown to be OH⁻ adsorption, which presents 1.15 eV (HMFCFA pathway) and 1.61 eV

(DFF pathway). According to experimental results⁴⁹, the overpotential needed for HMF oxidation on CuCo_2O_4 is 1.23-1.35 eV. Therefore, we can anticipate that the overpotential is not enough to overcome the barrier of the DFF pathway, which again agrees well with the experimental result showing limited DFF production.

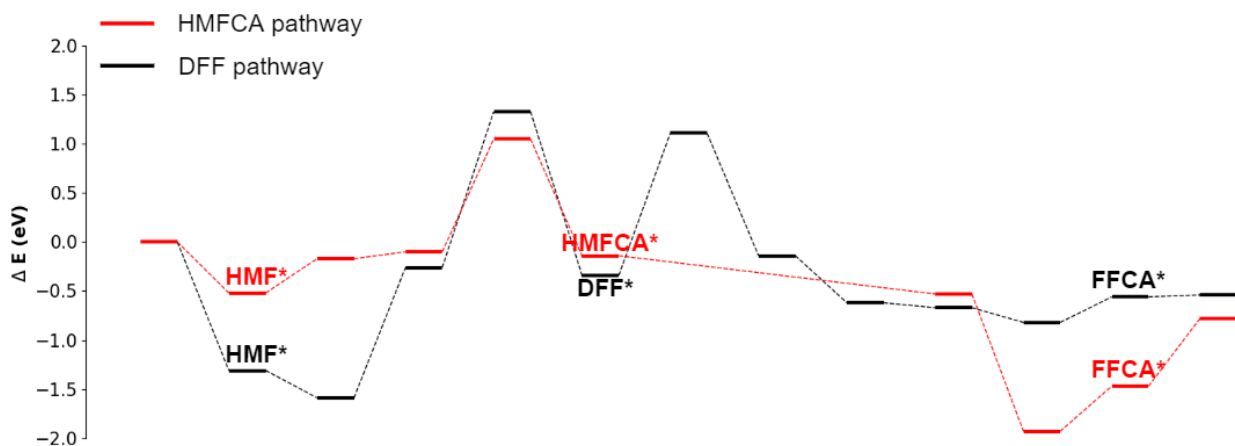


Figure 5-12. Pathway-dependent energy profiles for the HMF oxidation into FFCA on CuCo_2O_4 surface

5.5 Summary and Conclusions

To shed light on the activity of HMF oxidation reaction (HMFOR) on CuCo_2O_4 , we have studied the adsorption of atomic hydrogen on the low-Miller-index surfaces of CuCo_2O_4 , including (111), (110), and (100), by using periodic DFT. From the calculated surface grand potential, we predict that (111), (110), and (100) surfaces are the most stable. Furthermore, from ab initio atomistic thermodynamics, we find that the stabilities of the surfaces at 1 atm H_2 and 300K follow the trend of (100) > (110) > (111). The critical feature of HMF adsorption on the (100) is that Cu sites can adsorb the O atom of the hydroxyl group of HMF.

First, we revealed that Cu plays a vital role in adsorption by calculating the adsorption energies of HMF on various conformations. Furthermore, by calculating the energy profile of HMFOR on CuCo_2O_4 following both HMFCA and DFF pathways, we reveal why only a negligible amount of DFF is detected in the experiment. First, DFF has high adsorption energy on the surface. Second, the HMFCA pathway is thermodynamically more favorable. Lastly, DFF is likely to be oxidized into FFCA as soon as DFF is formed because HMFCA \rightarrow DFF pathway is exothermic, while the DFF \rightarrow FFCA pathway is highly endothermic. These insights open a door for a further mechanistic understanding of HMFOR on CuCo_2O_4 and other spinels.

Reference

- [1] A.F. Sousa, C. Vilela, A.C. Fonseca, M. Matos, C.S.R. Freire, G.J.M. Gruter, J.F.J. Coelho, A.J.D. Silvestre, Biobased polyesters and other polymers from 2,5-furandicarboxylic acid: a tribute to furan excellency, *Polym Chem.* 6 (2015) 5961–5983. <https://doi.org/10.1039/C5PY00686D>.
- [2] G.Z. Papageorgiou, D.G. Papageorgiou, Z. Terzopoulou, D.N. Bikiaris, Production of bio-based 2,5-furan dicarboxylate polyesters: Recent progress and critical aspects in their synthesis and thermal properties, *Eur Polym J.* 83 (2016) 202–229. <https://doi.org/10.1016/J.EURPOLYMJ.2016.08.004>.
- [3] Y.-C. Lin, G.W. Huber, The critical role of heterogeneous catalysis in lignocellulosic biomass conversion, (2008). <https://doi.org/10.1039/b814955k>.
- [4] L.D. Schmidt, P.J. Dauenhauer, Chemical engineering: Hybrid routes to biofuels, *Nature.* 447 (2007) 914–915. <https://doi.org/10.1038/447914a>.
- [5] P. Gallezot, Conversion of biomass to selected chemical products, *Chem Soc Rev.* 41 (2012) 1538–1558. <https://doi.org/10.1039/C1CS15147A>.
- [6] S.A. Akhade, N. Singh, O.Y. Gutiérrez, J. Lopez-Ruiz, H. Wang, J.D. Holladay, Y. Liu, A. Karkamkar, R.S. Weber, A.B. Padmaperuma, M.-S. Lee, G.A. Whyatt, M. Elliott, J.E. Holladay, J.L. Male, J.A. Lercher, R. Rousseau, V.-A. Glezakou, Electrocatalytic Hydrogenation of Biomass-Derived Organics: A Review, *Chem Rev.* 120 (2020) 11370–11419. <https://doi.org/10.1021/ACS.CHEMREV.0C00158>.
- [7] M. Garedew, F. Lin, B. Song, T.M. DeWinter, J.E. Jackson, C.M. Saffron, C.H. Lam, P.T. Anastas, Greener Routes to Biomass Waste Valorization: Lignin Transformation Through Electrocatalysis for Renewable Chemicals and Fuels Production, *ChemSusChem.* 13 (2020) 4214–4237. <https://doi.org/10.1002/CSSC.202000987>.
- [8] H.G. Cha, K.-S. Choi, Combined biomass valorization and hydrogen production in a photoelectrochemical cell, (2015). <https://doi.org/10.1038/NCHEM.2194>.
- [9] X. Han, H. Sheng, C. Yu, T. W. Walker, G. W. Huber, J. Qiu, S. Jin, Electrocatalytic Oxidation of Glycerol to Formic Acid by CuCo₂O₄ Spinel Oxide Nanostructure Catalysts, *ACS Catal.* 10 (2020) 6741–6752. <https://doi.org/10.1021/acscatal.0c01498>.
- [10] B. Program, T. Werpy, G. Petersen, Top Value Added Chemicals from Biomass Volume I-Results of Screening for Potential Candidates from Sugars and

Synthesis Gas Produced by the Staff at Pacific Northwest National Laboratory (PNNL) National Renewable Energy Laboratory (NREL) Office of Biomass Program (EERE) For the Office of the Energy Efficiency and Renewable Energy, (n.d.). <http://www.osti.gov/bridge> (accessed April 12, 2022).

- [11] M.J. Kang, H. Park, J. Jegal, S.Y. Hwang, Y.S. Kang, H.G. Cha, Electrocatalysis of 5-hydroxymethylfurfural at cobalt based spinel catalysts with filamentous nanoarchitecture in alkaline media, *Appl Catal B*. 242 (2019) 85–91. <https://doi.org/10.1016/J.APCATB.2018.09.087>.
- [12] Y. Lu, T. Liu, C.-L. Dong, Y.-C. Huang, Y. Li, J. Chen, Y. Zou, S. Wang, Tuning the Selective Adsorption Site of Biomass on Co₃O₄ by Ir Single Atoms for Electrosynthesis, *Advanced Materials*. 33 (2021) 2007056. <https://doi.org/10.1002/ADMA.202007056>.
- [13] J. Ren, K. he Song, Z. Li, Q. Wang, J. Li, Y. Wang, D. Li, C.K. Kim, Activation of formyl C–H and hydroxyl O–H bonds in HMF by the CuO(1 1 1) and Co₃O₄(1 1 0) surfaces: A DFT study, *Appl Surf Sci*. 456 (2018) 174–183. <https://doi.org/10.1016/j.apsusc.2018.06.120>.
- [14] X. Xie, Y. Li, Z.Q. Liu, M. Haruta, W. Shen, Low-temperature oxidation of CO catalysed by Co₃O₄ nanorods, *Nature* 2009 458:7239. 458 (2009) 746–749. <https://doi.org/10.1038/nature07877>.
- [15] M. Chun Yan, Z. Mu, J.J. Li, Y.G. Jin, J. Cheng, G.Q. Lu, Z.P. Hao, S.Z. Qiao, Mesoporous co₃o₄ and AU/CO₃o₄ catalysts for low-temperature oxidation of trace ethylene, *J Am Chem Soc*. 132 (2010) 2608–2613. <https://doi.org/10.1021/JA906274T>.
- [16] L. Hu, Q. Peng, Y. Li, Selective synthesis of Co₃O₄ nanocrystal with different shape and crystal plane effect on catalytic property for methane combustion, *J Am Chem Soc*. 130 (2008) 16136–16137. <https://doi.org/10.1021/JA806400E>.
- [17] L. Gao, Y. Bao, S. Gan, Z. Sun, Z. Song, D. Han, F. Li, L. Niu, Hierarchical Nickel–Cobalt–Based Transition Metal Oxide Catalysts for the Electrochemical Conversion of Biomass into Valuable Chemicals, *ChemSusChem*. 11 (2018) 2547–2553. <https://doi.org/10.1002/CSSC.201800695>.
- [18] Y. Lu, C.L. Dong, Y.C. Huang, Y. Zou, Z. Liu, Y. Liu, Y. Li, N. He, J. Shi, S. Wang, Identifying the Geometric Site Dependence of Spinel Oxides for the Electrooxidation of 5-Hydroxymethylfurfural, *Angewandte Chemie - International Edition*. 59 (2020) 19215–19221. <https://doi.org/10.1002/anie.202007767>.

- [19] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys Rev B.* 59 (1999) 1758. <https://doi.org/10.1103/PhysRevB.59.1758>.
- [20] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput Mater Sci.* 6 (1996) 15–50. [https://doi.org/10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).
- [21] P.E. Blöchl, Projector augmented-wave method, *Phys Rev B.* 50 (1994) 17953–17979. <https://doi.org/10.1103/PhysRevB.50.17953>.
- [22] J.P. Perdew, M. Ernzerhof, K. Burke, Rationale for mixing exact exchange with density functional approximations, *J Chem Phys.* 105 (1998) 9982. <https://doi.org/10.1063/1.472933>.
- [23] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys Rev Lett.* 77 (1996) 3865. <https://doi.org/10.1103/PhysRevLett.77.3865>.
- [24] S. Grimme, J. Antony, S. Ehrlich, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, *J. Chem. Phys.* 132 (2010) 154104. <https://doi.org/10.1063/1.3382344>.
- [25] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* 1 (2013) 011002. <https://doi.org/10.1063/1.4812323>.
- [26] Y. Wang, J. Cheng, M. Behtash, W. Tang, J. Luo, K. Yang, First-principles studies of polar perovskite KTaO_3 surfaces: structural reconstruction, charge compensation, and stability diagram, *Physical Chemistry Chemical Physics.* 20 (2018) 18515–18527. <https://doi.org/10.1039/c8cp02540a>.
- [27] W. Tang, E. Sanville, G. Henkelman, A grid-based Bader analysis algorithm without lattice bias, *J. Phys.: Condens. Matter.* 21 (2009) 84204–84211. <https://doi.org/10.1088/0953-8984/21/8/084204>.
- [28] S. Liu, D. Ni, H.F. Li, K.N. Hui, C.Y. Ouyang, S.C. Jun, Effect of cation substitution on the pseudocapacitive performance of spinel cobaltite MCo_2O_4 (M = Mn, Ni, Cu, and Co), *J Mater Chem A Mater.* 6 (2018) 10674–10685. <https://doi.org/10.1039/c8ta00540k>.

Chapter 6. Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per- and Polyfluoroalkyl Substances (PFASs)

6.1 Abstract

Due to their bioactive and persistent bioaccumulative properties, many per- and polyfluoroalkyl substances (PFASs) pose significant health hazards. However, assessing the bioactivities of PFASs is both time-consuming and costly due to the sheer number and expense of in vivo and in vitro biological experiments. To this end, we harnessed new unsupervised/semi-supervised machine learning models to automatically predict the bioactivities of PFASs in various human biological targets, including enzymes, genes, proteins, and cell lines. Our semi-supervised metric learning models were used to predict the bioactivity of PFASs found in the recent Organisation of Economic Co-operation and Development (OECD) report list, which contains 4730 PFASs used in a broad range of industries and consumers. Our work provides the first semi-supervised machine learning study of structure–activity relationships for predicting possible bioactivities in various PFAS species.

6.2 Introduction

Since the 1930s,¹ per- and polyfluoroalkyl substances (PFASs) have been used in several consumer products (including fire-fighting foams) due to their outstanding stability and water/oil-repellant properties.² However, these compounds pose significant risks to the environment and biosystems. The presence of PFASs in surface water and groundwater can result in exposure to organisms, subsequently leading to accumulation in the body, with adverse effects on the liver, kidneys, blood, and immune system.^{2,3} Because of these harmful effects, there is a pressing need to identify and understand the bioactivity of PFAS-based compounds that can adversely affect human health.

For these reasons, several international groups, including the Organization for Economic Cooperation and Development (OECD), United States Environmental Protection Agency, Food and Drug Administration, European Chemicals Agency, European Food Safety Authority, and Ministry of Ecology and Environment (China) continue to monitor PFASs that are produced in the global market.^{4,5} According to a 2018 OECD report, more than 4,700 PFASs currently exist as manufacturers bring new forms of PFASs into industrial and consumer products (it is worth pointing out, however, that not all 4,700 structures exist in commerce). Nevertheless, among the vast varieties of PFAS molecules, the potential hazards of these new forms remain largely unknown.

Due to the sheer number of PFAS species, *in vivo* and *in vitro* biological experiments are both time-consuming and costly. As such, the construction of predictive and reliable quantitative-structure activity relationship (QSAR) models⁶⁻⁸ is essential for assessing the bioactivities of these contaminants (even for PFAS species that are yet to be

made). Specifically, a QSAR model that can accurately predict the bioactivities of PFASs can be harnessed to screen several of these contaminants, saving immense time and experimental resources. While there have been prior machine learning studies on PFAS molecules,^{65,66} most of these approaches used supervised learning techniques to suggest *general* structure-bioactivity correlation trends based on their target-specific predictions on bioactivity. (i.e., the focus was on aggregate data for all targets as opposed to analyzing chemical trends specific to each target).

In this work, we present a new QSAR model using semi-supervised metric learning techniques to assess which functional groups affect bioactivities toward specific biological targets. Semi-supervised learning is a different machine learning approach that has the advantages of both supervised and unsupervised learning. It can be used on a dataset with primarily unlabeled data and only a few labeled data. Like unsupervised learning, it can also automatically cluster unlabeled data. Our approach is integrated with molecular docking calculations to predict possible bioactivities of PFAS molecules based on their chemical functional groups and specific biological targets (e.g., genes, proteins, or cell lines). Our approach first combines dimension reduction methods with clustering methods to classify PFASs based on their molecular structures. We then apply a semi-supervised metric learning method to improve clustering performance. Finally, we use a molecular docking approach to shed light on the physicochemical reasons for their bioactivity. Our study provides the first unsupervised/semi-supervised learning approach for screening potentially bioactive PFAS molecules beyond conventional supervised learning or QSAR approaches.

6.3 Computational Method

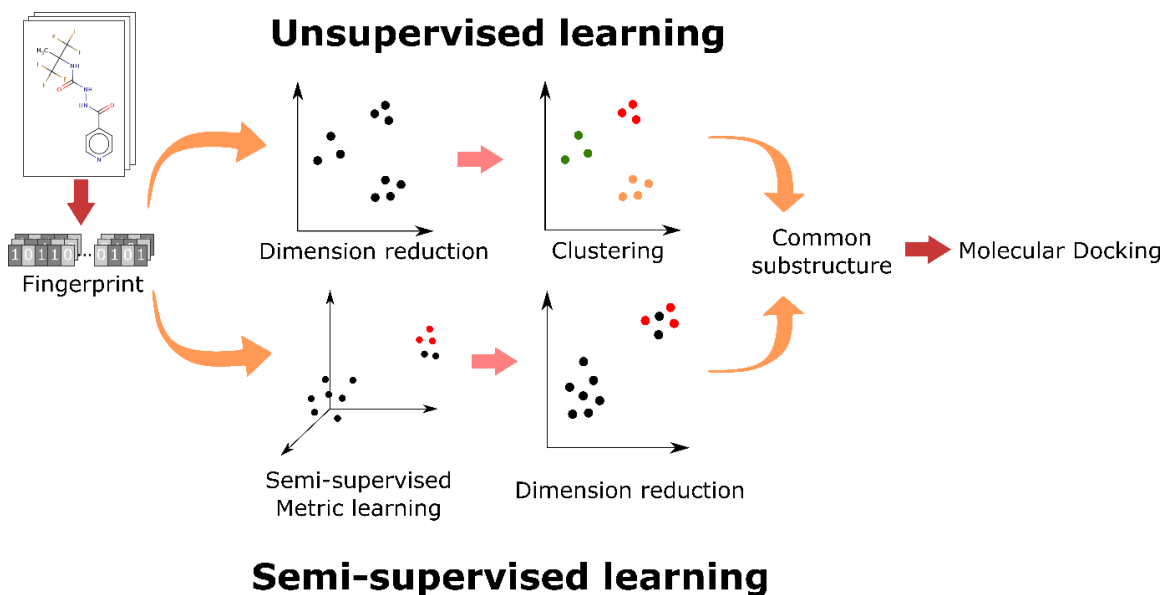


Figure 6-1. Machine-learning-based workflow for QSAR construction and application to PFASs.

Our QSAR machine-learning framework utilizes four sequential steps followed by a reasoning step: (1) collecting a training data set from verified open-source databases, (2) encoding those compounds into molecular fingerprints, (3) clustering the model to predict chemical properties based on the molecular fingerprints and assessing the performance of the models, (4) evaluating the clustering by choosing the optimal model and finding the molecular groups that play essential roles in developing bioactivity based on the clustering, and (5) molecule docking to rationalize the role of the molecular groups.

Starting with our first step, we obtained our data sets from comprehensive open-source databases, including PubChem's BioAssay,¹¹ Maximum Unbiased Validation,¹² Toxicology in the 21st Century,¹³ beta-secretase 1,¹⁴ and Blood-brain barrier penetration

data sets,¹⁵ which are available from the Supporting Information of ref 10. We used two different data sets with varying screening criteria as provided in ref 10: (1) CF data set includes substances containing at least one –CF– moiety (62043 molecules), and (2) C3F6 data set includes substances containing perfluoroalkyl moiety with three or more carbons (1012 molecules). For both data sets, we used bioactivity data against 26 biological targets.

Encoding the compounds to molecular fingerprints followed next in our framework. We used the extended connectivity fingerprint (ECFP) featurization¹⁶ with a default diameter of 4 (i.e., ECFP4), thus considering a maximum of four neighbors. ECFPs are topological molecular characterization that was developed for substructure and similarity searching. By encoding molecular structures into fingerprints, we obtained a binary array with a constant length of 2048, making it a convenient input for the unsupervised/semi-supervised learning models. Furthermore, since the simplified molecular-input line-entry system (SMILES) sequences for all PFAS molecules are available in the form of a line notation for describing the structure of chemical species, they can be readily converted into fingerprint-based representations using RDKit.¹⁷

Then, we applied (1) **unsupervised learning** and (2) **semi-supervised learning** methods to the generated fingerprints. Our QSAR method trained machine learning models to predict the bioactivities of PFAS molecules by first (a) *reducing the fingerprint data sets dimension* and then (b) *clustering them*.

For the first half of this step, (1) **unsupervised learning model**, we used (a) *dimension reduction* methods to lay the high-dimensional fingerprint input data into a 2-dimensional space. Two different dimension reduction methods, (i) Principal Component Analysis (PCA) followed by t-Distributed Stochastic Neighbor Embedding (t-SNE) and (ii) UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction),¹⁸ were used on our fingerprint data. While t-SNE is a widely-used dimension reduction technique for many types of data analysis, research shows UMAP exhibits better clustering performance, especially for large data sets.¹⁸ Thus, we decided to use both and compared the clustering performance of the techniques.

After the dimension reduction procedure, we used the scikit-learn¹⁹ library to execute three different (b) *clustering methods*: k-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Hierarchical DBSCAN (HDBSCAN). Computational models use all three clustering methods to group data points into clusters based on similarity. While k-means clustering is relatively more time-efficient, DBSCAN and HDBSCAN can efficiently handle outliers and noisy data sets. Since each method has clear advantages, we used all three techniques and evaluated their performances. By combining dimension reduction and clustering methods, we grouped various PFAS structures based on their similarities and visualized them in a 2-dimensional space.

The second half of our QSAR model used (2) a **semi-supervised metric learning** algorithm to group/classify molecules with similar bioactivities automatically. Metric learning has two main advantages: (i) its predictions are more efficient/accurate since the model distinctly separates new molecular representations according to their bioactivities;

(ii) it automatically generates a vector-shaped representation from the molecular fingerprint and can be directly integrated with conventional dimension reduction methods.

The final clusters were selected based on the best Silhouette score, which analyzes the distances of each data point to its cluster and neighboring clusters.²⁰ In short, a higher Silhouette score guarantees better performance in clustering. Then we identified which substructures or molecular functional groups played essential roles in determining the bioactivity of the molecules.

Lastly, we conducted several molecular docking calculations using Autodock²¹ to elucidate the physicochemical reasons for the bioactivity trends obtained from our QSAR model (i.e., using ligand-protein binding conformations to rationalize the role of chemical substructures in inducing bioactivity on biological targets.)

6.4 Results and Discussion

6.4.1 Unsupervised learning

In this section, we compare the performance of each machine-learning algorithm in classifying/clustering the PFAS molecules based on their bioactivity. The parameters for each model are optimized for the best performance (i.e., highest Silhouette score). As mentioned in the methods section, we tested six unsupervised learning models by combining two different dimension reduction methods (PC t-SNE and UMAP) with three different clustering methods (k-means, DBSCAN, and HDBSCAN). The combination of

k-means clustering and the UMAP model exhibited better performance (with a Silhouette score of 0.577) than other machine-learning methods.

Figure 6-2 displays the distribution of bioactivities based on the molecular structures. First, the molecular structures were converted into constant-length arrays by ECFP. Next, the vector representations were projected onto a 2-dimensional space using the PC t-SNE dimension reduction methods (molecules grouped closer together indicated structural similarity). These molecules were subsequently clustered into ten groups using k-means clustering, an algorithm that classifies data into a given number of distinct clusters.

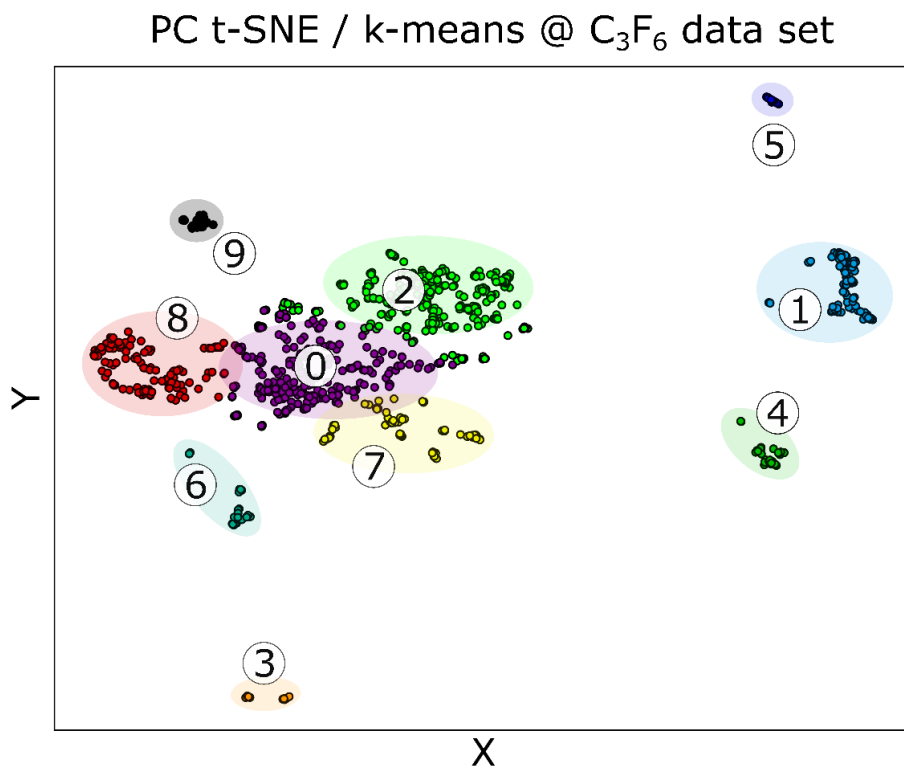
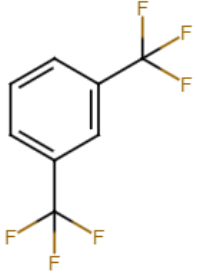
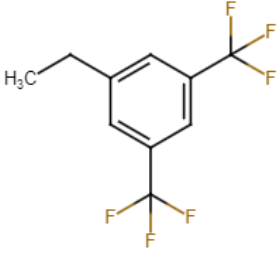


Figure 6-2. Distribution of the molecules in the C3F6 data set. The molecular structures were visualized by PC t-SNE. Each point is a molecule, and the colors of the points are clusters classified using k-means clustering.

We found that 90.6%, 76.9%, and 70.0% of molecules in Cluster 1 are bioactive on CYP2C9, CYP2D6, and CYP3A4, respectively. CYP2C9, CYP2D6, and CYP3A4 are members of Cytochrome p450 enzymes (Cyps) involved in metabolism by oxidation of xenobiotics in the human body. Cyps are major phase-I xenobiotic-metabolizing enzymes induced by many environmental xenobiotics and drugs.²² Also, 42.9% of molecules in Cluster 4 are bioactive on ATXN, a DNA-binding protein.^{23,24} We only demonstrate targets with the top 4 true-positive rates; 90.6%, 76.9%, 70.0%, and 42.9%. True positive rate is the probability that an actual positive (bioactive molecule) will be predicted

positive (bioactive). Table 6-1 shows the maximum common substructures of the clusters. Both clusters, 1 and 4, have the common functional group 1,3-Bis(trifluoromethyl) benzene.

Table 6-1. Cluster number, accuracy, and maximum common structure that are most likely to be found in bioactive molecules toward each target.

Target	Cluster	True-Positive Rate	Common Substructure
CYP2C9	1	90.6%	
CYP2D6	1	76.9%	
CYP3A4	1	70.0%	
ATXN	4	42.9%	

Even though our unsupervised learning approaches could automatically classify PFASs into a reasonable number of clusters, it is essential to note that not all 26 targets gave a high performance in clustering (i.e., only ATXN, CYP2C9, CYP3A4, and CYP2D6 showed reasonable accuracy). From this result, we were able to infer that the bioactivity of PFAS against ATXN and Cyps have the strongest correlation with the molecular structures.

Furthermore, a larger data set (CF3 data set: 62,043 molecules) did not have a higher clustering performance than a smaller data set (C3F6 data set: 1,012 molecules)

since traditional unsupervised learning clustering methods, including k-means clustering, work better when applied to smaller data sets.²⁵ Because the CF3 data set is 50 times larger than the C3F6 data set and screened based on less rigorous criteria, we expect it to be more challenging to handle with unsupervised learning techniques.

For example, Figure 3a shows the bioactivities of molecules toward keratin18 (K18), an intermediate filament protein.²⁶ Our unsupervised machine learning failed to successfully cluster the PFAS molecules based on the bioactivity toward K18, demonstrating a true-positive rate of only 55.9%. Similarly, Figure 3c shows the bioactivity toward CYP2C9 using unsupervised learning. Only 4.0% of the cluster showed bioactivity. These results highlight the inherent limitation of purely unsupervised learning for molecular structures. Since bioactivities are sensitive to minimal changes in molecular structures, they exhibit a mixed distribution in the molecular structure space. In other words, a pair of similar-structured molecules can have entirely different bioactivities, which are commonly referred to as activity cliffs in cheminformatics.²⁷⁻²⁹

To overcome our PC t-SNE and UMAP results limitations, we utilized a combination of metric and semi-supervised learnings to produce a more distinct separation between bioactive and inactive molecules toward K18, which can be seen in Figure 3b, demonstrating a true-positive rate of 79.2%. Similarly, Figure 3c shows the bioactivity toward CYP2C9 using unsupervised learning. Again, only 4.0% of the cluster showed bioactivity. Finally, Figure 3d shows the semi-supervised clustering, where 99.7% of a cluster's molecules are bioactive (i.e., a true positive rate of 99.7%).

It is not surprising that semi-supervised learning demonstrated significantly higher performance for PFAS QSAR tasks since it uses partially labeled data. In contrast, an unsupervised learning model takes unlabeled data as input. Furthermore, the performance difference is more drastic with larger data, considering that traditional unsupervised learning clustering methods, including k-means clustering, work better when applied to a smaller data set.²⁵ Metric learning combined with semi-supervised learning uses partially labeled bioactivity data as input data, placing molecules with similar bioactivities closer and opposite bioactivities further away. On the other hand, unsupervised learning tries to predict bioactivities solely based on the chemical structure, making it extremely difficult to complete an appropriate QSAR task.

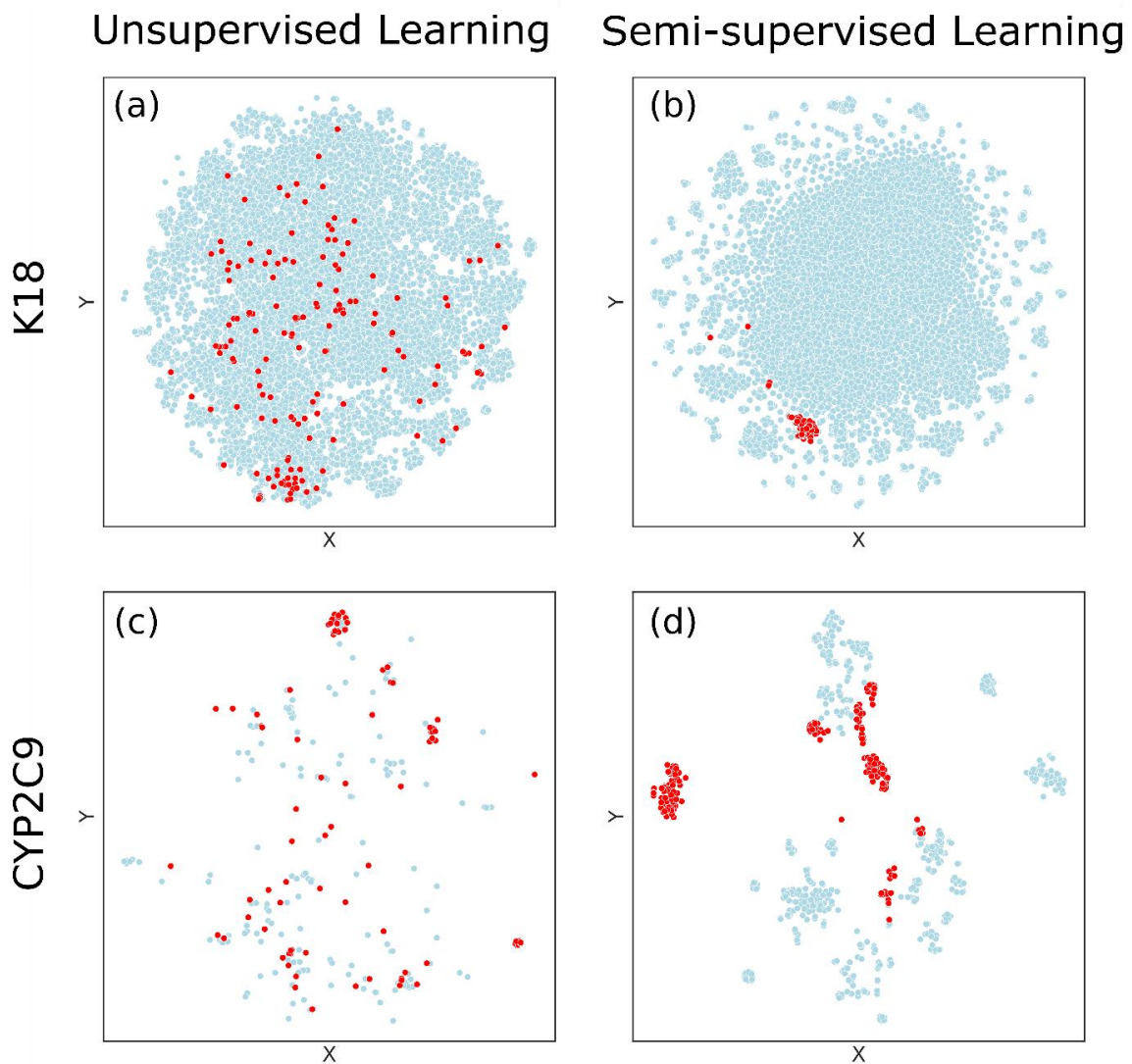


Figure 6-3. 2-dimensional space distribution of molecules in the CF3 data set. Each point represents a molecule that is either bioactive (red) or inactive (blue) towards (a, b) K18 and (c, d) CYP2C9. The molecules are visualized on a 2-dimensional space using (a, c) PC t-SNE (unsupervised) or (b, d) semi-supervised metric learning.

6.4.2 Semi-supervised Metric Learning

Figure 6-4 displays true-positive ratios and classifications between bioactive/inactive molecules on four representative targets that show the best performance in the CF dataset using semi-supervised metric learning (for example, in

Figure 6-4**Error! Reference source not found.**a, we obtain a true-positive ratio of 97.3% by computing $\frac{\text{number of molecules containing esters and are also bioactive}}{\text{number of ester-containing molecules in the cluster}}$. Using the Maximum Common Structure (MCS) module in the RDKit software package on bioactive molecules, we found that the ester functional group is the critical substructure that causes bioactivity on Cyps (Figure 6-4a, b, and c) and ATXN (Figure 6-4d).

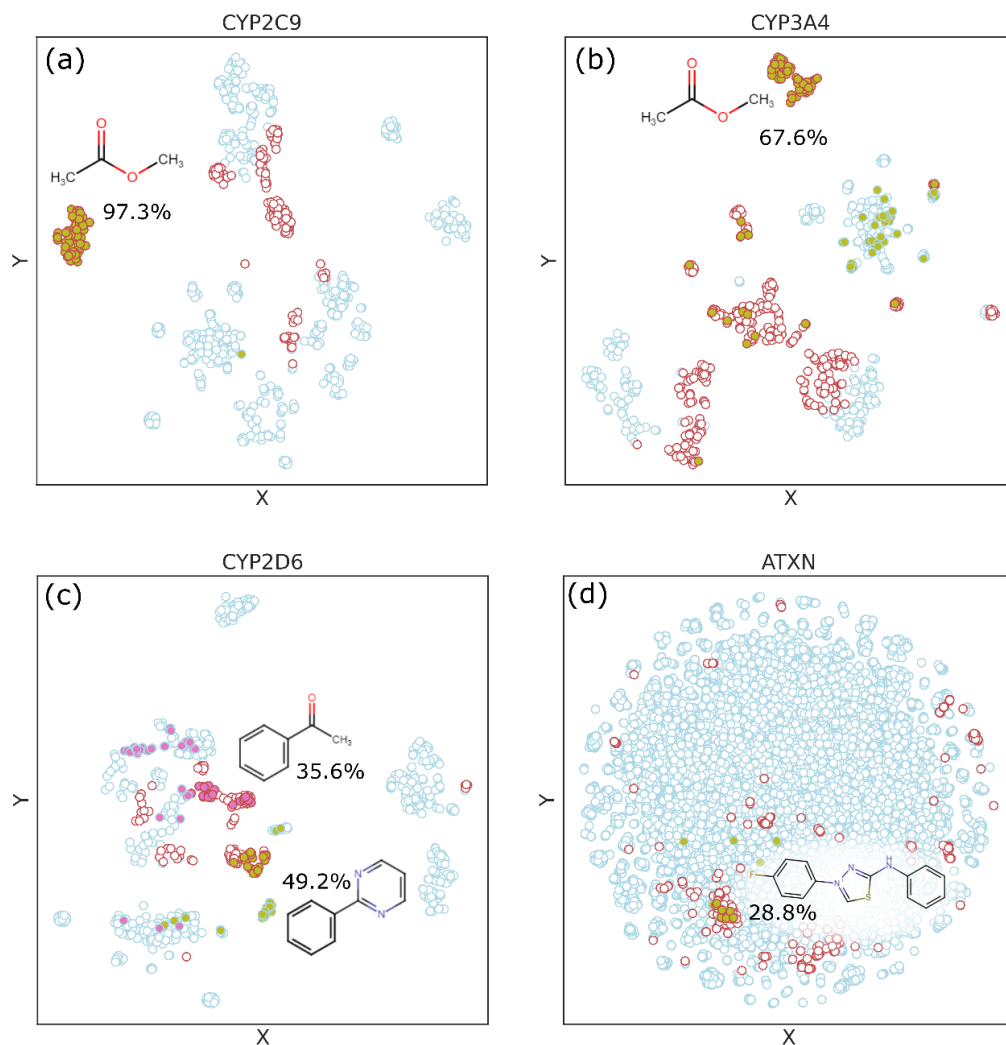


Figure 6-4. Distribution of molecules in the CF dataset using semi-supervised metric learning. Each point represents a molecule that is either bioactive (red circular edges) or inactive (light blue circular edges) towards (a) CYP2C9, (b) CYP3A4, (c) CYP2D6, and (d) ATXN. The olive green-filled circles represent molecules having the substructure depicted in the plot; i.e., (a, b) ester groups, (c) phenylpyridyl groups, and (d) 4-benzyl-2-(4-fluorophenyl)-1,2-thiazole. The pink-filled circles in (c) represent molecules with phenylethanone. The percentage value represents the ratio of the number of bioactive molecules within the identified substructure. Table S3 lists the predicted substructures for specific targets.

We used structural alerts to cross-check the validity of the predicted substructures that play a crucial role in bioactivity. Within the bioinformatics community, structural alerts are molecular functional groups associated with a particularly adverse outcome, in our case, bioactivity.^{67,68} We cross-referenced the ChEMBL dataset to our machine learning results since it contains structural alert information for some PFAS molecules.⁶⁹ As mentioned previously, the ester group was found to be the critical structure that induces interaction with Cyps.^{70,71}

6.4.3 Interactions between PFAS and targets

We carried out molecular docking calculations with Autodock²¹ to rationalize the underlying molecular causes of bioactivities in PFAS and predict their interaction with target enzymes. The Supporting Information gives additional details of our molecular docking calculations. We successfully docked all PFASs into the active sites of the targets and binned the binding affinity results based on their bioactivity with the target. Figure S5 displays one of the bioactive structures with the ester group of the CYP2C9-PFAS complex, methyl 4-[2-propyl-1-({[4-trifluoromethyl]phenyl}sulfonyl}amino)-2-hexen-1-yl]benzoate.

To verify the correlation between the Autodock binding affinities and their bioactivity, we performed a dimension reduction procedure using unsupervised learning on the CF dataset, which consists of molecular structures with binding affinity data (see Figure 6-5). We used unsupervised learning here to make the point that unsupervised learning underperforms when only structural data is provided. Specifically, if the

classification accuracy is improved with additional feature inputs, those features must contain some information to discriminate among the population.^{72,73} In other words, if the inclusion of binding affinity data enhances the clustering accuracy, it provides another co-descriptor for bioactivity. Indeed, Figure 6-5b and a show that descriptors consisting of chemical structures *and* binding affinity data give a better separation/distinction between active and inactive molecules compared to the unsupervised learning results based only on chemical structures.

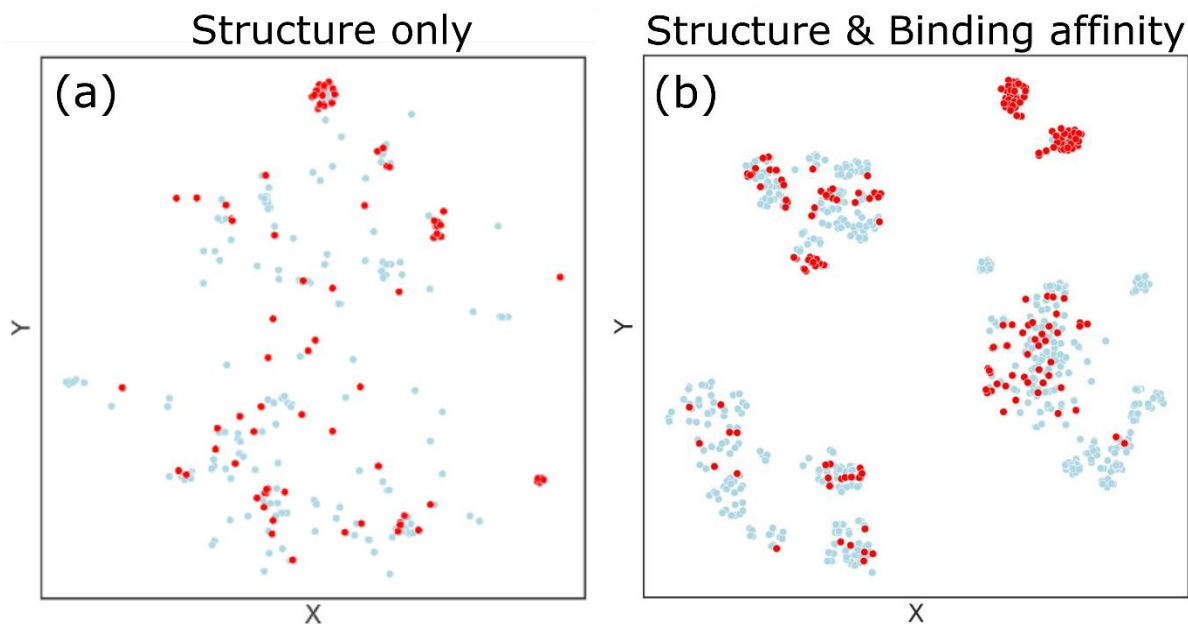


Figure 6-5. Clustering/classification of molecules predicted with unsupervised learning (dimension reduction) on CF datasets containing (a) chemical structures and (b) chemical structures and binding affinities with CYP2C9. Each point represents a molecule that is either bioactive (red) or inactive (blue) towards CYP2C9.

6.4.4 Bioactivity Predictions on OECD Dataset

In 2018, the Global Perfluorinated Chemicals Group⁷⁴ within the OECD published a list of 4,730 PFASs to develop regulatory approaches for reducing the use of

perfluorinated substances in products. However, researchers have yet to discover the bioactivities of the molecules in the list. Using the QSAR model developed in this work, we give predictions and a rationale for the bioactivities of molecules in the OECD list.

We performed molecular docking calculations on molecules containing the ester group among the OECD list to verify similar binding conformations. Of the 4,730 PFASs in the OECD list, 414 have an ester functional group. In particular, the ester-containing molecules in the OECD list bind strongly with Fe^{2+} of the HEME group (an active site of Cyp enzyme), which is similar to the binding interactions that we observed in the CF dataset. Therefore, we expect a large portion of the 414 ester-containing molecules among the OECD list to form strong bonds with Fe^{2+} of the HEME group with a similar conformation, leading to bioactivity toward Cyp enzymes. Furthermore, based on our docking calculations, 87.7% of these 414 molecules have a stronger binding affinity than -5 kcal/mol (the average binding affinity is -5.77 kcal/mol), which falls in the range of the mean binding affinity of the bioactive molecules from the CF dataset.

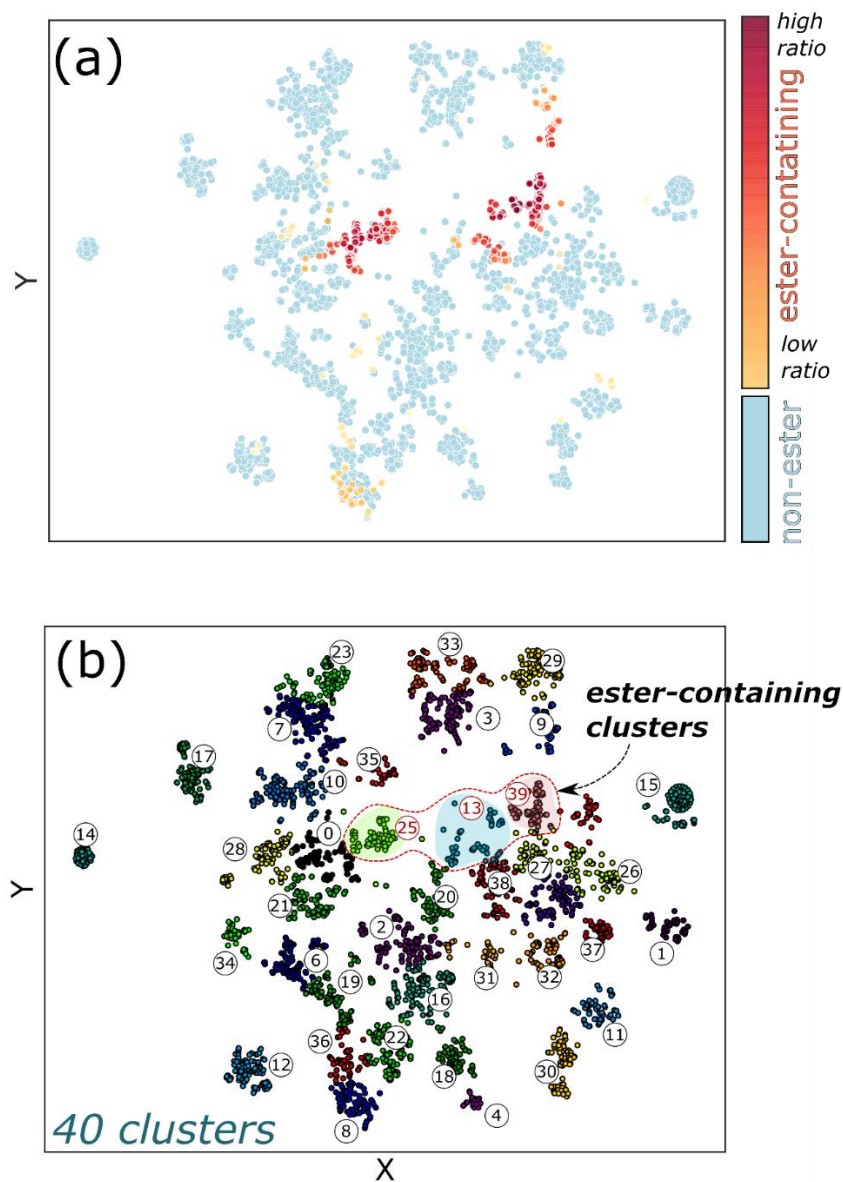


Figure 6-6. (a) OECD dataset classified by PC t-SNE and clustered based on the k-means clustering method. The orange and yellow dots represent ester-containing molecules. The colors closer to red (yellow) represent a higher (lower) concentration of bioactive molecules. (b) PFAS molecules included in the OECD list are grouped into 40 clusters. Each point represents a molecule, and clusters 13, 25, and 39 denote a high ratio of ester-containing groups.

We then clustered the OECD dataset into 40 clusters using the k-means clustering method. Using both the clustered results (Figure 6-6b) and the distribution of ester-group-containing molecules (Figure 6-6a), we found that clusters 13, 25, and 39 contain ester functional groups. Analyzing the CF dataset, we found that the ester group plays a possible role in bioactivity toward Cyp enzymes; that is, molecules in these clusters have a high probability of being bioactive against CYP2C9 and CYP3A4.

6.5 Summary and Conclusions

In summary, we have developed a new QSAR model validated with ChemMLB structural alerts and molecular docking calculations, which constitutes the first application of semi-supervised metric learning for predicting/rationalizing bioactivities in PFASs. Using a semi-supervised metric learning algorithm, our machine-learning-based QSAR model accurately identified specific substructures, such as ester-containing groups, that play a possible role in determining bioactivities. With our semi-supervised learning approach, we obtained a distinct classification between bioactive and inactive molecules, resulting in an accuracy of up to 97.3% in the CF dataset. We also used semi-supervised metric learning to automatically classify/cluster and predict functional groups that could possibly play a role in bioactivity.

In addition, our machine learning model proposed a few significant substructures that could induce bioactivity, which were subsequently examined with molecular docking calculations. Most importantly, our machine learning predictions on bioactivities can provide a more efficient screening of potentially bioactive PFASs that can be used to complement *in vitro* assessments. All of our machine learning algorithms are publicly

available, and we anticipate that researchers can further extend our methodology to screen other contaminants or analyze the potential bioactivity of PFAS molecules.

Reference

- (1) Hepburn, E.; Madden, C.; Szabo, D.; Coggan, T. L.; Clarke, B.; Currell, M. Contamination of Groundwater with Per- and Polyfluoroalkyl Substances (PFAS) from Legacy Landfills in an Urban Re-Development Precinct. *Environ. Pollut.* **2019**, *248*, 101–113.
- (2) Blake, B. E.; Pinney, S. M.; Hines, E. P.; Fenton, S. E.; Ferguson, K. K. Associations between Longitudinal Serum Perfluoroalkyl Substance (PFAS) Levels and Measures of Thyroid Hormone, Kidney Function, and Body Mass Index in the Fernald Community Cohort. *Environ. Pollut.* **2018**, *242*, 894–904.
- (3) Guillette, T. C.; McCord, J.; Guillette, M.; Polera, M. E.; Rachels, K. T.; Morgeson, C.; Kotlarz, N.; Knappe, D. R. U.; Reading, B. J.; Strynar, M.; Belcher, S. M. Elevated Levels of Per- and Polyfluoroalkyl Substances in Cape Fear River Striped Bass (*Morone saxatilis*) Are Associated with Biomarkers of Altered Immune and Liver Function. *Environ. Int.* **2020**, *136*, 105358.
- (4) OECD. 033-066-C609-51.Pdf. *Series on Risk Management* **2018**, No. 39. (39), 1–24.
- (5) Cousins, I. T.; Dewitt, J. C.; Glüge, J.; Goldenman, G.; Herzke, D.; Lohmann, R.; Miller, M.; Ng, C. A.; Scheringer, M.; Vierke, L.; Wang, Z. Strategies for Grouping Per-and Polyfluoroalkyl Substances (PFAS) to Protect Human and Environmental Health. *Environ. Sci.: Process. Impacts.* **2020**, *22*, 1444–1460.
- (6) Hansch, Corwin.; Fujita, Toshio. P- σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **2002**, *86*, 1616–1626.
- (7) Cherkasov, A.; N. Muratov, E.; Fourches, D.; Varnek, A.; I. Baskin, I.; Cronin, M.; Dearden, J.; Gramatica, P.; C. Martin, Y.; Todeschini, R.; Consonni, V.; E. Kuz'min, V.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.
- (8) Neves, B. J.; Braga, R. C.; Melo-Filho, C. C.; Moreira-Filho, J. T.; Muratov, E. N.; Andrade, C. H. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **2018**, *9*, 1275.
- (8) Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S. S. R. K. C.; Lian, C.; Kwon, H.; Wong, B. M. A Machine Learning Approach for Predicting Defluorination of Per- And Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environ. Sci. Technol. Lett.* **2019**, *6*, 624-629.

- (10) Cheng, W.; Ng, C. A. Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List. *Environ. Sci. Technol.* **2019**, *53*, 13970–13980.
- (11) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 Update. *Nucleic Acids Res.* **2014**, *42*, 1075–1082.
- (12) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.
- (13) Krewski, D.; Acosta, D.; Andersen, M.; Anderson, H.; Bailar, J. C.; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V. G.; Green, S.; Kelsey, K. T.; Kerkvliet, N. I.; Li, A. A.; McCray, L.; Meyer, O.; Patterson, R. D.; Pennie, W.; Scala, R. A.; Solomon, G. M.; Stephens, M.; Yager, J.; Zeise, L. Toxicity Testing in the 21st Century: A Vision and a Strategy. *J. Toxicol. Environ. Health. B. Crit. Rev.* **2010**, *13*, 51–138.
- (14) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β -Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936–1949.
- (15) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686–1697.
- (16) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (17) *RDKit*. <http://www.rdkit.org/> (accessed 2021-06-29).
- (18) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
- (19) Morris, G. M.; Ruth, H.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Software News and Updates AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- (20) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 147.

- (21) Yang, H.; Lou, C.; Li, W.; Liu, G.; Tang, Y. Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. *Chem. Res. Toxicol.* **2020**, *33*, 1312–1322.
- (22) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620.
- (23) Cheng, X.; Klaassen, C. D. Perfluorocarboxylic Acids Induce Cytochrome P450 Enzymes in Mouse Liver through Activation of PPAR- α and CAR Transcription Factors. *Toxicol. Sci.* **2008**, *106*, 29–36.
- (24) Miners, J. O.; Birkett, D. J. Cytochrome P4502C9: An Enzyme of Major Importance in Human Drug Metabolism. *Br. J. Clin. Pharmacol.* **1998**, *45*, 525–538.
- (25) Ashburner, J.; Klöppel, S. Multivariate Models of Inter-Subject Anatomical Variability. *Neuroimage* **2011**, *56*, 422–439.
- (26) Chu, C.; Hsu, A. L.; Chou, K. H.; Bandettini, P.; Lin, C. P. Does Feature Selection Improve Classification Accuracy? Impact of Sample Size and Feature Selection on Classification Using Anatomical Magnetic Resonance Images. *Neuroimage* **2012**, *60*, 59–70.
- (27) *OECD Portal on Per and Poly Fluorinated Chemicals - OECD Portal on Per and Poly Fluorinated Chemicals.* <https://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/> (accessed 2021-07-01).

Chapter 7. Harnessing Neural Network for Predicting XANES Spectroscopy of Amorphous Carbon Materials

7.1 Abstract

X-ray absorption spectroscopy (XAS) provides a wealth of information about the local structure of materials. Recently, significant advancement has been made in the development of machine learning models for predicting local environments of absorbing atoms from XAS. However, existing studies have primarily focused on crystalline systems, while less has been paid to more complex and disordered systems. In this work, we develop a neural network model for predicting XAS spectra of amorphous carbon (a-C) using the local structural descriptor as a sole input. In addition, we compared the performance of different structural descriptors, including the Local Many-Body Tensor Representation (LMBTR), Atom-Center Symmetry Function (ACSF), and Smooth Overlap of Atomic Positions (SOAP). We find that the use of LMBTR yields the highest accuracy, and the inclusion of both bond length and bond angle information is necessary for accurate XAS prediction. Furthermore, among the three representations, LMBTR offers a unique advantage in interpretability as the input components can be easily associated with specific structural features. We also extend our model to predict not only local structure features of a-C, such as bond lengths and angles but also its global chemical composition from XAS spectra.

7.2 Introduction

Historically, the interpretation of XANES spectra is primarily qualitative and

relies on semi-empirical rules.^{1,2} Due to the desire to rapidly characterize spectra for arbitrary local environments, data-driven methods for XAS are now enjoying great interest across various communities; for a general overview, we recommend a review by Timoshenko and Frenkel. These methods attempt to exploit all of the information contained within a spectrum, as opposed to the subset that a heuristic describes, and are enabled by the high availability of theoretical data and the promise of high-throughput experimental XAS data. ML models have been used to automate the analysis of experimental XANES and EXAFS data to gain insights into system properties and behavior. Previous work has demonstrated the feasibility of classifying specific structural properties, such as oxidation number and coordination, from said spectra via ensemble learning. Recent work has also used artificial neural networks and random forests to focus on coordination alone. For example, Timoshenko et al. developed a NN model to predict the coordination number of Pt nanoparticles³. They used the same approach to predict the radial distribution function of metals from XANES spectra.⁴ There were also several attempts to extract the coordination environment of the absorbing site from XANES spectra. For instance, Zheng et al. trained an ensemble of weak learners to predict the coordination environment of metals from XANES spectra.⁵ Carbone et al. applied a Convolutional NN classifier to predict the coordination number of transition metal oxides from XANES spectra,⁶ and Liu et al. used a similar model for copper oxide clusters.⁷ Torrisi et al. used a random forest model to extract coordination number, nearest neighbor distance, and Bader charges from XANES spectra of transition metal oxides.⁸

To interpret XANES spectra, one needs to address two classes of problems. First, the forward problem simulates XANES spectra from given the atomic arrangement. Using electronic structure theory, the forward problem can be solved by simulating XANES spectra which is a powerful tool yet requires immense computational cost. Therefore, with the recent increased popularity of Machine Learning (ML) in materials science, ML techniques opened a new way to perform inexpensive simulations of XANES spectroscopy. For example, Rankine et al. used a Deep NN to predict XANES of arbitrary Fe systems based on local structural features.⁵ Carbone et al. used a Message-Passing NN to simulate XANES of Oxygen or Nitrogen in small molecules from the QM9 database.⁶ These efforts present ML models which accurately simulate XANES spectra of a given chemical structure with minimum computational cost.

Most of the inverse problem ML approaches aim to predict coordination number because they are mainly focused on metals,^{8,9} metal oxides,¹⁰ nanoparticles,¹³ or small molecules⁶ of which the coordination number identifies local atomic arrangements. In contrast, only a handful of studies focused on amorphous systems,¹⁴ which require detailed information (i.e., bond length, angles, and coordination numbers) to describe the local environment fully. Furthermore, it is worth emphasizing that the ML approach is particularly beneficial for amorphous materials. The immense computational cost is required to calculate XANES spectra from first principles due to the large lattice box.

Among many amorphous materials, we focus on amorphous carbon (a-C) materials in this work. For the last few decades, a-C materials have been extensively

used in various applications such as surface protective film or coatings¹⁵⁻¹⁷ due to their low cost and high mechanical and chemical stability.¹⁸ More recently, in addition to the properties mentioned earlier, there has been emerging attention to its high electronic conductivity,¹⁹ which makes it an attractive candidate for anode material for batteries.²⁰⁻²² However, the biggest hurdle of utilizing a-C materials is that it is difficult to identify the structure due to a lack of long-range order, having XANES as an ideal tool for determining the detailed structure.^{23,24} In this work, we first apply ML techniques to simulate XANES spectra, given structures of a-C materials. First, we constructed and trained an NN model using a database with 12,528 a-Cs and various densities to predict the XANES spectra of given structural features. Second, we introduce the inverse ML algorithm to predict global and local structural components, such as bond length, bond angles, coordination numbers, and chemical composition, given XANES spectra. Lastly, we discuss the performance of three different structural descriptors (i.e., Local Many-Body Tensor Representation (LMBTR), Smooth Overlap of Atomic Position (SOAP), and Atom-Centered Symmetry Function (ACSF)) by evaluating the performance of ML models with a detailed analysis of feature importance.

7.3 Computational Method

7.3.1 Dataset

The overall procedure we followed is as described in Figure 1. We first generated a dataset of a-C materials from molecular dynamics (MD) simulations to train the NN model. Then, melt-quench MD simulations with Gaussian approximation

potential (GAP)²⁵ were performed to obtain a-C structures at different densities. The simulation protocol from Ref 26 was employed. First, a simple-cubic lattice of 216 carbon atoms was generated at the densities of 1.5, 2.0, 2.5, 3.0, and 3.5 g/cm³. Next, the structures were heated to 9000 K and held at a constant temperature for 3 ps. Then the liquid carbon was cooled and held at 5000 K for 3 ps. Finally, the structures were quenched and annealed at 300 K for 3 ps. The simulations were carried out using LAMMPS²⁷ linked with the QUIP package.²⁸ A timestep of 1 fs and a Nose-Hoover thermostat was used in all the simulations.

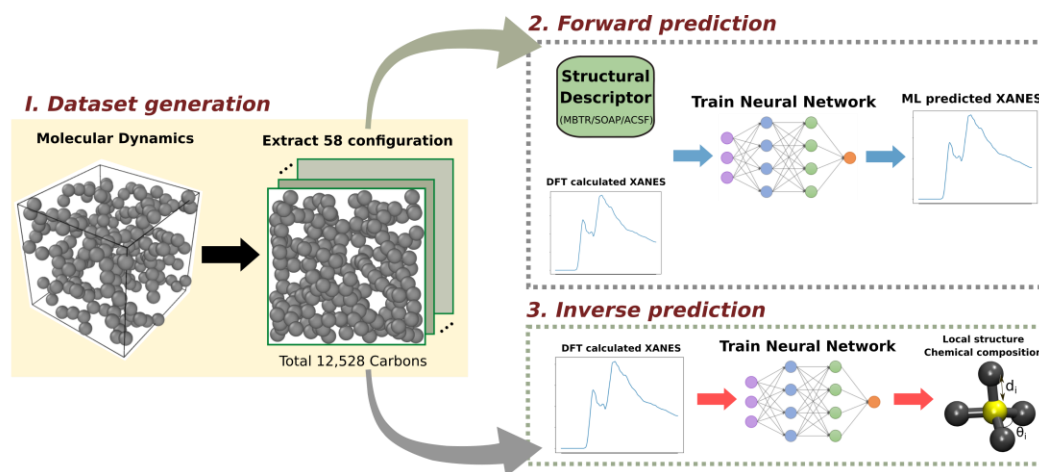


Figure 7-1. Schematic drawing of procedure

A dataset of 58 configurations of a-C structures with various densities was sampled from the previously mentioned molecular dynamics simulations. Each structure contains 216 C atoms, achieving 12,528 local Carbon structures. The corresponding XANES spectrum was generated from the first principles for each system, applying Fermi's golden rule approximations.²⁹ We employed constrained-occupancy DFT calculations within the excited electron and core-hole (XCH) approach.^{29,30} We

simulated X-ray photon absorption and production of a core hole by placing the associated excited electron in the lowest available empty state of the system. We approximated the higher-energy excited states utilizing the unoccupied part of the Kohn-Sham DFT eigenspectrum within the consequent XCH self-consistent field. DFT-calculated-XANES spectra were preprocessed as we shifted the energy levels and smoothed the spectra to represent spectra from experiments using well-known carbon structures (i.e., diamond and highly oriented pyrolytic graphite) as a benchmark. We used a gaussian filter for smoothing, and the parameters are optimized to achieve the best fitting for the benchmark.

7.3.2 Structure Representation

When choosing a structural descriptor, the most important criterion is whether the ML model can accurately predict the desired property using the descriptor as an input. The optimal descriptor vastly differs depending on many factors, such as types of systems, models, and complexity of the problem. To rigorously determine the most suitable structural representation for a-C, we constructed ML models using each descriptor to compare the mean absolute error (MAE). Among many proposed ways of representing periodic structures such as Coulomb Matrix (CM),³¹ Bag of Bonds (BoB),³² Smooth Overlap of Atomic Positions (SOAP),³³ Many-Body Tensor Representation (MBTR),³⁴ and Atom-Center Symmetry Function (ACSF),³⁵ we selected SOAP, Local MBTR (LMBTR), and ACSF because we are only interested in local properties of absorbing sites. In other words, for the prediction of local XANES spectra, local structure representations (i.e., LMBTR, SOAP, and ACSF) are the most

suitable. In brief, ACSF represents distance and angular symmetry functions for each atom in a system. SOAP describes the local environment of a center atom by placing Gaussian-smearred atomic densities on atoms. Lastly, LMBTR is a descriptor that groups interactions by interatomic distances and angles.

7.3.3 Machine Learning Construction

First, we employed the Artificial NN model for the forward prediction (i.e., simulating the XANES spectrum given the chemical structure). The NN model is constructed using the Tensorflow³⁶ library with the Keras³⁷ framework. NN is a well-known composite function that can be trained to represent a relationship between inputs and outputs by tuning the weights of the nodes.³⁸ Due to its feasibility, NN has been used in many applications to predict various chemical properties.^{11,39}

The forward NN model we employed comprises an input layer of 1400 neurons (size of LMBTR array) and an output layer of 100 neurons (size of XANES array). In addition, two dense hidden layers exist between the input and output layers, with 700 and 400 neurons. Each hidden layer uses the soft plus activation function, and batch normalization and dropout are applied in each hidden layer.

Second, for the inverse prediction (i.e., predicting the local structure of a given XANES spectrum), we extracted three critical structural features of an a-C: hybridization, bond lengths, and bond angles. Therefore, we first utilized a classifier model to predict the coordination number (hybridization) of each carbon using XANES spectra. We employed four different models (i.e., Random Forest Classifier, Linear Support Vector Classifier (SVC), Multinomial Naive Bayes (NB), and Logistic

Regression) from sklearn⁴⁰ to find the optimal classifier model and compared the accuracy. Then, based on the predicted hybridization, we developed a NN model to predict all bond lengths and angles of absorbing sites. The inverse NN model comprises an input layer of 100 neurons (size of LMBTR array) and an output layer of 1-6 neurons (depending on the coordination number of the center atom). In addition, there is one dense hidden layer between the input and output layers, with 50 neurons using the Scaled Exponential Linear Unit (SELU) activation function. Our forward and inverse NN uses gradients of MAE as loss functions according to the adaptive moment estimation (ADAM) algorithm calculated within the minibatch size of 32 samples. Finally, we evaluated the model performance using K-fold cross-validation with an 80:20 split.

Lastly, we made a separate model which predicts the chemical composition of the system given global XANES spectra (i.e., the sum of all the local XANES spectra in the system). Since various critical physical properties such as hardness or Young's modulus of a material depend on the sp^2/sp^3 ratio, there have been a few attempts to predict the hybridization composition in carbon materials based on XANES spectra.^{24,41} However, the traditional σ^*/π^* ratio method overestimates sp^3 concentrations because it does not consider the existence of sp carbon.⁴² Therefore, in this paper, we developed a NN model that predicts the whole composition of sp , sp^2 , and sp^3 carbons in an a-C material. Due to a lack of global

XANES data (58 global XANES), we constructed 22,000 synthetic global XANES by randomly combining 216 local XANES, and our NN model was trained based on the

synthesized XANES spectra.

7.4 Results and Discussion

7.4.1 Unsupervised Learning

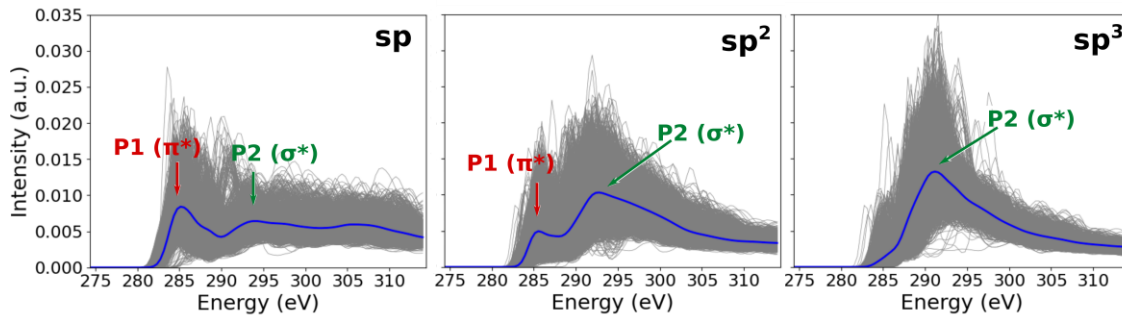


Figure 7-2. Peaks P1 and P2 present a-C XANES spectra are denoted with red and green arrows, respectively. Grey lines represent each XANES spectrum from individual carbon, and the blue curve represents the average XANES spectrum from carbons with sp , sp^2 , and sp^3 hybridization.

To understand the structure-XANES feature relationship, we first analyzed the general characteristics of XANES spectra of a-C materials. For example, Figure 7-2 indicates that all the XANES spectra we obtained from DFT calculation mainly exhibit two peaks: P1 (~ 286 eV) and P2 (~ 292 eV). In this context, we focused on the quantitative features of P1 and P2, such as peak energies, intensities, and intensity ratios.

Then we performed Principal Component Analysis (PCA) for dimension reduction on the XANES spectra. After the PCA dimensionality reduction, each XANES spectrum (array size 100) is reduced to a point on a 2-dimensional latent space and aggregated into three major groups according to their coordination number (Figure 7-3a). Then we further analyzed the PCA results using spectroscopic features (Figure 7-3b and c) and

structural features (Figure 7-3c and Figure 7-4).

One can see from Figure 7-3a and b that P1/P2 intensity ratio is a crucial factor that differentiates sp and sp^2 carbons. Also, the bond length distribution on the PCA plot (Figure 7-3d) is inversely proportional to the P1/P2 peak intensity ratio. As such, sp carbons have a high P1/P2 intensity ratio (~ 3.0) and short bond length (~ 1.2 Å; $C\equiv C$), while sp^2 carbons present low P1/P2 intensity (<1) and long bond length (~ 1.4 Å; $C=C$).

The close relationship between hybridization and P1/P2 intensity ratio is not surprising, considering that P1 and P2 are associated with σ^* and π^* states, respectively.^{41,43} As such, sp carbons have a stronger P1 intensity than P2, while sp^3 carbons have a pronounced P2 peak. In addition, only P2 exists in the XANES spectra of sp^3 carbons.

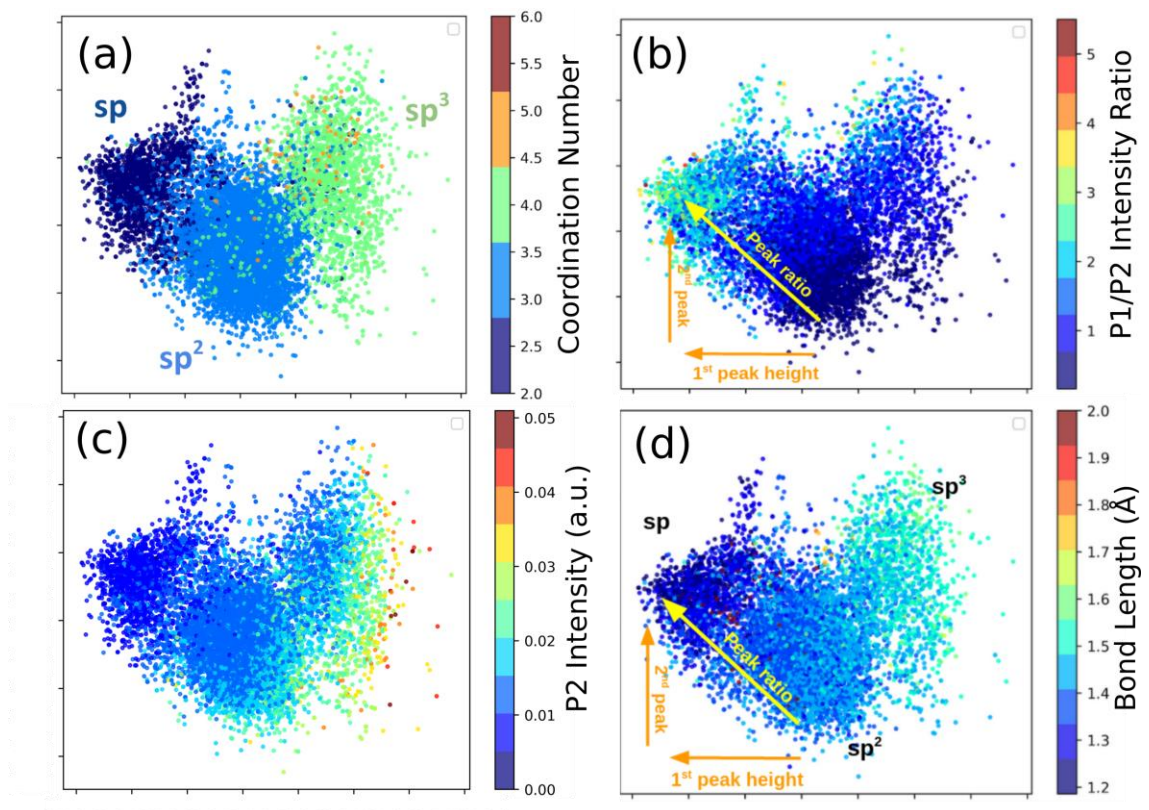


Figure 7-3. PCA analysis of XANES spectroscopies. Each point represents XANES spectra and color displays (a) coordination number, (b) P1/P2 intensity ratio, (c) P2 intensity, and (d) minimum bond length.

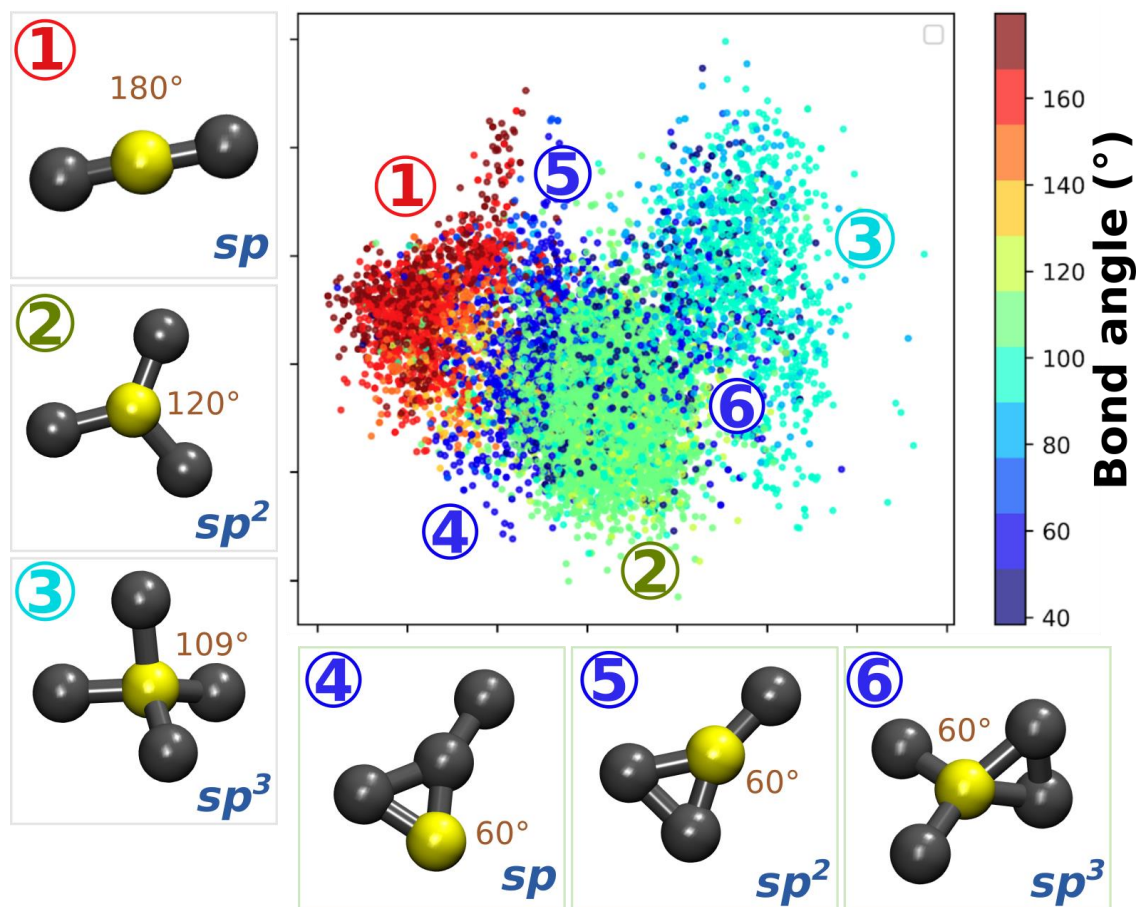


Figure 7-4. PCA analysis of XANES spectra based on bond angles and corresponding local environments of a-C materials. Each point represents XANES spectra, and color represents the minimum bond angle. Grey balls represent surrounding carbon atoms, and yellow balls represent the center carbon atom. The hybridization and the minimum bond angle of the center atom are denoted. The left panels are typical configurations of sp , sp^2 , and sp^3 carbons, while the bottom panels demonstrate distorted carbon structures which majorly exist in the a-C systems.

On the other hand, bond angle distribution is relatively complicated due to distorted structures in a-C materials (Figure 7-4). The majority of the distribution is easily interpretable for regular coordination, such as sp carbons with 180° , sp^2 carbons with 120° , or sp^3 carbons with 109° , as shown in the left panels of Figure 7-4. However, as previously reported both theoretically^{26,44-49} and experimentally,⁵⁰ our a-C structures

present various small carbon rings. For example, carbons associated with three-membered-ring, shown in the bottom panels of Figure 7-4, demonstrate a minimum bond angle of 60° (blue points in Figure 7-4). These three-membered carbon rings have a significant strain, but the presence of larger carbon rings (i.e., five- to eight-membered rings) fused with three-membered rings provides stability. It is worth noting that a-Cs in three-membered ring conformations display distinctive features in XANES spectra. For example, sp carbons in three-membered rings (panel 4 in Figure 7-4) exhibit exceptionally strong P2 intensity (Figure 7-3) compared to ordinary sp carbons (panel 1 in Figure 7-4).

On the other hand, sp³ carbons associated with three-membered rings (panel 6 in Figure 7-4) have exceptionally weak P2 intensity (Figure 7-3) compared to ordinary sp³ carbons (panel 3 in Figure 7-4). In short, unlike crystalline systems, the coordination number of the center atom does not decisively determine the local environment or XANES spectra in the case of amorphous systems. Therefore, bond angle plays a vital role in identifying the XANES spectra of a-C materials.

7.4.2 Supervised Learning

Using the critical structural and spectroscopic features determined from unsupervised learning, we constructed a supervised learning model that can accurately simulate the XANES spectrum of a given chemical structure. We employed the NN algorithm for our forward prediction based on advantages such as implicitly detecting complex nonlinear relationships between dependent and independent variables.

A descriptor that can qualitatively describe the structural feature is needed to use atomic structural information as an input for the NN model. To find the optimal descriptor that can effectively describe a-C materials, we compared the performance of ML models using three different structural representations—LMBTR, SOAP, and ACSF. As shown in Figure 7-5 and Table 7-1, the MAE of prediction for the overall spectra and the spectroscopic feature (i.e., peak energies/intensities) both indicate that LMBTR outperforms the other two.

While ACSF demonstrates lower but comparable performance to LMBTR, SOAP significantly underperforms due to the absence of the angle feature. While MBTR and ACSF include explicit angle features (angular function), SOAP employed from the DDescribe package only consists of the interatomic distance feature of the atoms within the cutoff radius from the center atom. Furthermore, since there are a variety of distorted local environments in a-C systems, various bond angles are possible within the same hybridization (Figure 7-4), unlike crystal structures. Therefore, even though SOAP has demonstrated outstanding performance in predicting the chemical properties of crystalline systems,⁵¹ we concluded that it is not the best descriptor for amorphous systems.

In addition to the high performance, LMBTR has the advantage of high interpretability because the features can be easily visualized and correspond to specific structural properties of the system. More specifically, MBTR is composed of three different k-body terms, and each of the terms corresponds to atom species (k=1), bond lengths (k=2), and bond angles (k=3). Therefore, each geometric feature can be

separately analyzed and interpreted. On the other hand, since ACSF is composed of multiple symmetry functions, its interpretation based on geometric features is not as simple as LMBTR. Thus, we chose LMBTR as the optimal representation for a-C systems.

As a result, the performance of the NN model on the LMBTR-converted dataset is promising. For example, the MAE of the test set maintains 5.74×10^{-4} a.u. while requiring only 2 hours on a single-node GPU. Furthermore, the LMBTR-based-NN model accurately predicts spectrum features such as peak intensities and energies. For instance, P1 intensity prediction achieves MAE smaller than 1.5×10^{-4} a.u. Finally, it is worth emphasizing that our NN model can accurately predict significant peaks of XANES spectra of random systems such as diamond and graphene, even though they are not included in the dataset.

Table 7-1. MAE of ML model predicting XANES spectra using MBTR, SOAP, and ACSF

	LMBTR	ACSF	SOAP
MAE (a.u.)	5.74×10^{-4}	7.79×10^{-4}	1.184×10^{-3}

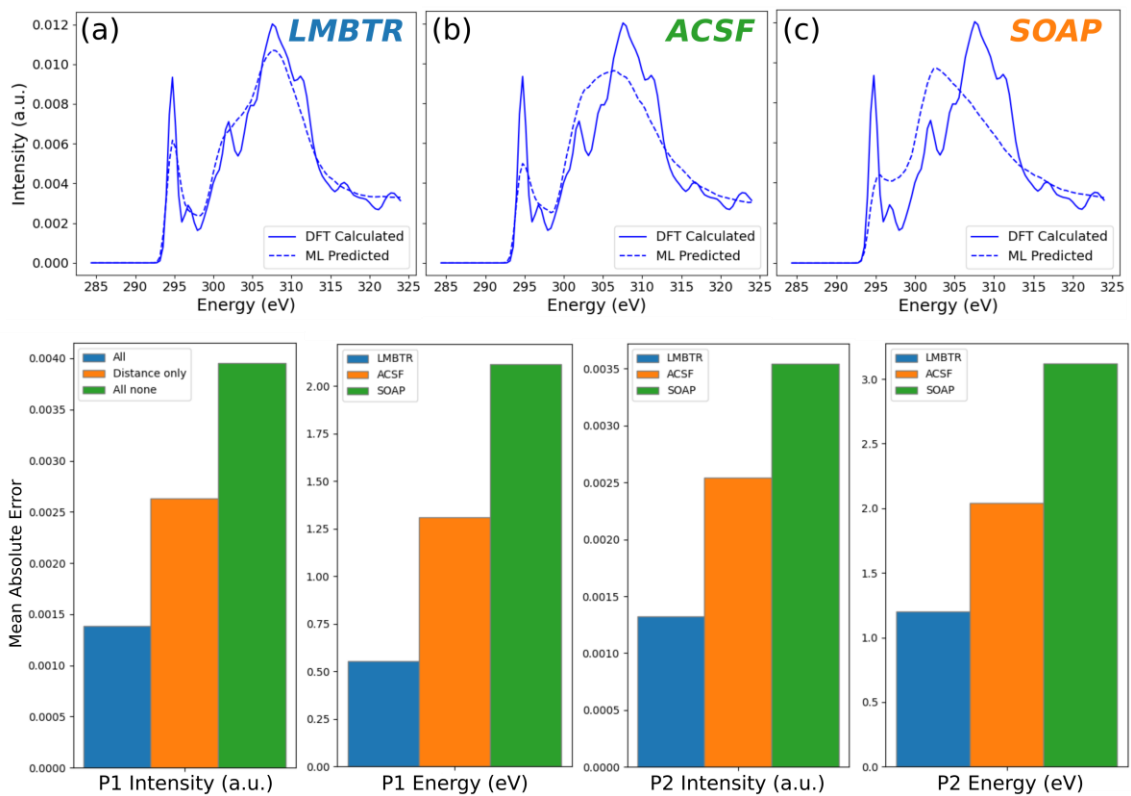


Figure 7-5. XANES spectrum prediction using (a) MBTR, (b) SOAP, and (c) ACSF descriptors. Solid lines and dashed lines represent DFT calculated NN predicted XANES spectra, respectively. Lower panels display MSE of predicting P1 and P2 intensities and energies using each of the descriptors.

7.4.3 Feature Analysis

We proceeded with feature importance analysis to further strengthen our hypothesis that the explicit angle feature of MBTR plays a vital role in XANES prediction. Feature importance analysis was performed using the optimized ML models to understand the relative importance of the structural features predicting XANES spectra. The importance of each feature was defined as the change in the MAE of predicted XANES spectra after randomly shuffling the values of this feature and keeping other descriptors unchanged. In this process, we took benefit of the interpretability of MBTR. Since MBTR can be separated into two parts, the bond length feature and bond angle feature, we could modify features and evaluate the accuracy of each modification.

Figure 7-6 shows the feature importance analysis results. Even though bond length information solely offers reasonable accuracy in simulating general peak appearance and peak intensities, the bond angle feature is necessary to achieve a highly accurate prediction. Examples of predicted XANES spectra are shown in Figure 7-6a. MAE values shown in the figure are the average values obtained from the test set. We further calculated MAE values to predict spectroscopic features such as peak intensities and positions (Figure 7-6b). Again, the general trend is the same: most accurate when all the features are provided and less accurate when only the distance feature is maintained. From this result, again, we strengthen our hypothesis that the angle feature plays a vital role in predicting XANES spectra; thus, MBTR overperforms SOAP because MBTR has an explicit angle feature, while SOAP does not.

The shuffling method provides the relative importance of geometric features but does not imply which specific values are significantly correlated with spectra features. Therefore, thanks to the interpretability of LMBTR, we further analyzed feature importance using the automatic gradient method to look closely into the specific range of geometric feature values that plays a vital role in determining XANES spectra. Thanks to the high interpretability of MBTR, we employed Gradient Tape available in the Tensorflow package³⁶ for automatic differentiation. We took XANES features such as peak intensity and position (L) and the gradient of L with respect to MBTR (dL/dx). The magnitude of gradient values corresponds to the importance of the specific geometric value (i.e., bond lengths) toward the XANES features.

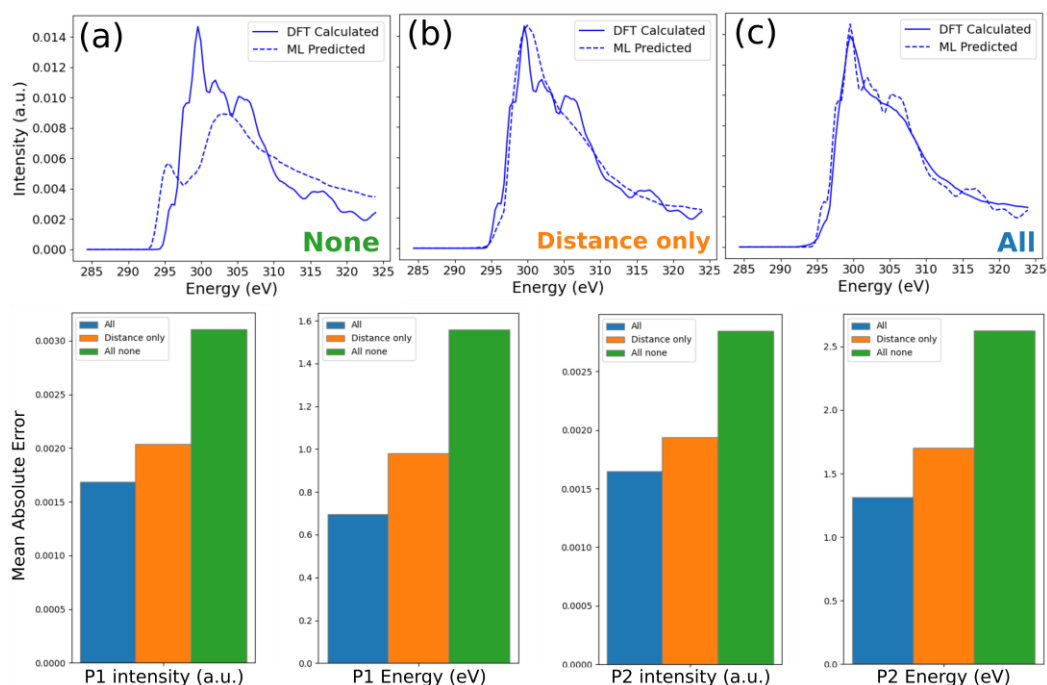


Figure 7-6. Feature analysis using the shuffling method. (a) XANES spectra predicted using limited constrained structural features. Solid lines and dashed lines correspond to DFT calculated, and NN predicted XANES spectra, respectively. (b) MAE of predicting spectroscopic features such as peak intensities and energies.

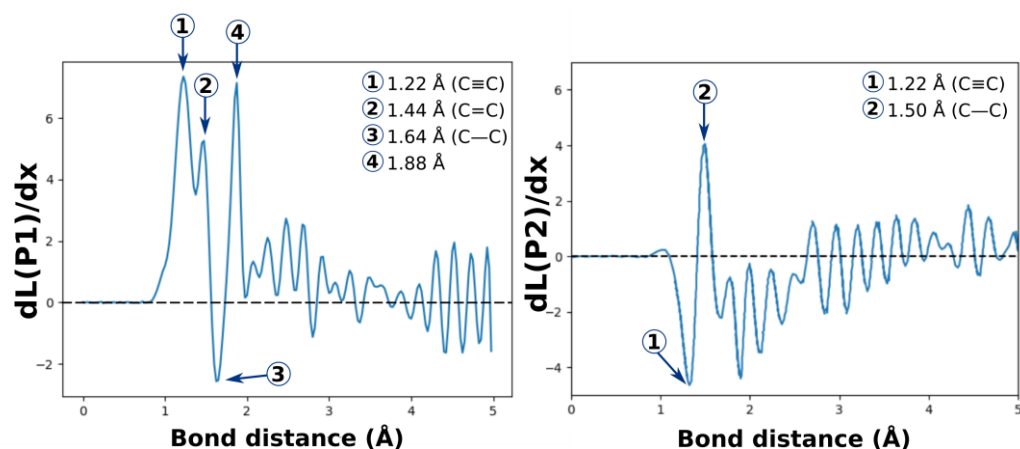


Figure 7-7. Gradient of (a) P1 and (b) P2 intensity with respect to two body interaction terms (k_2) in LMBTR.

Figure 7-7 shows that P1 intensity is positively proportional to the probability function of an atom existing at a distance of 1.22 Å (C≡C) and 1.44 Å (C=C) from the center atom. On the contrary, P1 intensity is negatively proportional to the probability of an atom existing at 1.64 Å (C-C) from the center atom. On the other hand, P2 intensity is positively proportional to the probability of an atom existing at 1.50 Å (C-C) and negatively proportional to 1.22 Å (C≡C). For both cases, the probability of an atom located farther than 2.0 Å from the center atom does not play an essential role in determining XANES peak intensity. This result again emphasizes that the bond order of an absorbing site is a critical factor in determining XANES peak intensities.

7.4.4 Inverse prediction of local XANES

For the local structure prediction from XANES spectra (i.e., inverse prediction), we first predicted the coordination number (hybridization) of each carbon using XANES spectra. It is already shown from PCA analysis that the coordination number

is directly correlated with the P1/P2 intensity ratio. Then, we performed four different classification methods (i.e., Random Forest Classifier, Linear SVC, Multinomial NB, and Logistic Regression) and concluded that Random Forest Classifier offers the most accurate prediction (Figure 7-8a).

The classification results for the test set are shown in Figure 7-8b, indicating that most a-C coordination numbers are accurately predicted. However, many sp^3 carbons are wrongly predicted to be sp^2 . This may be because of the outlier type (VI) shown in Figure 7-4, which is sp^3 carbon, but still indicates XANES spectra comparable to sp^2 carbons due to their distorted structures.

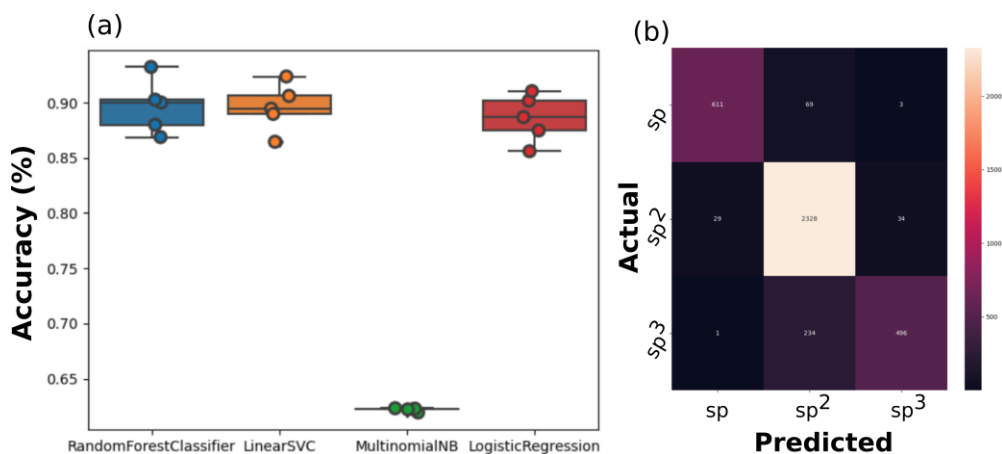


Figure 7-8. Local XANES inverse prediction. (a) Comparison of hybridization prediction accuracies obtained from four different models. (b) Heat map of predicted and actual hybridization obtained from Random Forest classifier.

Next, based on the coordination number predicted by the Random Forest classifier, we generated a NN model to predict bond lengths and angles. We first need to separate the dataset based on their hybridization. They must be trained separately because each has different bond lengths and angles. For example, sp carbons need 2

bond lengths and 1 bond angle to be predicted, while sp^3 carbons have 4 bond lengths and 6 bond angles. The results are shown in Table 7-2.

Table 7-2 shows that local bond lengths can be predicted with high accuracy with less than 2% error, which is high enough to indicate the bond order. In contrast, bond angle prediction is reasonable but less accurate (< 4%) than distance prediction. It is unsurprising, considering that many distorted a-Cs exist in the system, and many outliers present complicated angle distributions (Figure 7-4).

Table 7-2. MAE of local structure prediction from XANES

	sp	sp^2	sp^3
Bond Angle MAE (°)	4.43	4.60	4.07
Bond length MAE (Å)	0.0136	0.0266	0.0249

7.4.5 Inverse prediction of global XANES

Since we predicted local carbon structures given local XANES with high accuracy, we made another attempt to predict global structural features given global XANES spectra. Therefore, we constructed another NN model trained using 22,000 synthetic global XANES. The $sp:sp^2:sp^3$ ratios are predicted within 0.017% of error. In contrast, when using an analytical method to find the optimal ratio of the linear combination of average sp, sp^2 , and sp^3 XANES spectra, the composition ratios were predicted with 21.8% of MAE, which is significantly inaccurate compared to the NN predicted ratio.

7.5 Summary and Conclusions

Simulating XANES spectroscopy of each absorbing site of amorphous materials

is laborious and time-consuming, with the absence of symmetry and a large lattice box. This chapter presents a simple yet powerful NN model that can accurately simulate local XANES of a-C given only chemical structure information that requires marginal computational cost compared to ab initio methods. Furthermore, the predicted XANES spectra agree with DFT calculated XANES with an MAE of only 5.74×10^{-4} a.u. Also, spectra features such as peak positions and intensities are accurately predicted.

We chose LMBTR as a structural descriptor to be used as an input for the NN model because (1) MBTR-based NN presents the highest accuracy, and (2) MBTR is easily interpretable. In addition, one attractive characteristic of amorphous systems is that SOAP significantly underperforms in describing a-C systems due to the absence of angular features. From PCA analysis and feature importance analysis, we show that the bond angle feature is critical in determining XANES spectra of a-C due to the distorted local environment of which structures cannot be solely determined by hybridization or interatomic distance features.

In addition, The most significant advantage of using LMBTR is that interatomic distance and angle features can be analyzed separately. Using this advantage, we performed feature importance analysis using random shuffling and automatic gradient methods, which can only be applied to LMBTR among all structural descriptors.

We also present a NN model that can predict local structure features such as coordination number, bond lengths, angles, and chemical composition given XANES (inverse prediction). Unlike crystalline systems or small molecules, all coordination numbers, bond lengths, and bond angles are required to describe the local structure of

a-C deliberately. In addition, we present the $sp:sp^2:sp^3$ ratio extracted from XANES spectra using the NN model, which overcomes the limitation of the traditional σ^*/π^* ratio method. Through this process, we successfully tackled interpreting the complex XANES-structure relationship of amorphous systems and extended the application of ML-XANES predictions to amorphous systems.

Future work could improve accuracy by generating a larger and more balanced dataset (balanced number of sp , sp^2 , and sp^3 carbons). In addition, we anticipate that our findings can be used to explore the XANES spectra of amorphous materials. Finally, we also expect our results to ultimately offer the possibility of building an ML model for the fully automated conversion of XANES to atomic positions at a minimum computational cost.

Reference

- (1) Bianconi, A.; Dell'Ariceia, M.; Gargano, A.; Natoli, C. EXAFS and near edge structure; Springer, 1983; pp 57–61
- (2) Westre, T. E.; Kennepohl, P.; DeWitt, J. G.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. A multiplet analysis of Fe K-edge $1s \rightarrow 3d$ pre-edge features of iron complexes. *Journal of the American Chemical Society* 1997, 119, 6297–6314
- (3) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *The journal of physical chemistry letters* 2017, 8, 5091–5098
- (4) Timoshenko, J.; Anspoks, A.; Cintins, A.; Kuzmin, A.; Purans, J.; Frenkel, A. I. Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy. *Physical review letters* 2018, 120, 225502
- (5) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y.; Tang, H.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F., et al. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Computational Materials* 2018, 4, 1–9.
- (6) Carbone, M. R.; Yoo, S.; Topsakal, M.; Lu, D. Classification of local chemical environments from x-ray absorption spectra using supervised machine learning. *Physical Review Materials* 2019, 3, 033604.
- (7) Liu, Y.; Marcella, N.; Timoshenko, J.; Halder, A.; Yang, B.; Kolipaka, L.; Pellin, M. J.; Seifert, S.; Vajda, S.; Liu, P., et al. Mapping XANES spectra on structural descriptors of copper oxide clusters using supervised machine learning. *The Journal of Chemical Physics* 2019, 151, 164201.
- (8) Torrisi, S. B.; Carbone, M. R.; Rohr, B. A.; Montoya, J. H.; Ha, Y.; Yano, J.; Suram, S. K.; Hung, L. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Computational Materials* 2020, 6, 1–11.
- (9) Timoshenko, J.; Frenkel, A. I. “Inverting” X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *Acs Catalysis* 2019, 9, 10192–10211.
- (10) Carbone, M. R.; Topsakal, M.; Lu, D.; Yoo, S. Machine-learning X-ray absorption spectra to quantitative accuracy. *Physical Review Letters* 2020, 124, 156401.
- (11) Guda, A. A.; Guda, S. A.; Martini, A.; Kravtsova, A.; Algasov, A.; Bugaev, A.; Kubrin, S.; Guda, L.; Sot, P.; van Bokhoven, J., et al. Understanding X-ray

- absorption spectra by means of descriptors and machine learning algorithms. *npj Computational Materials* 2021, 7, 1–13
- (12) Rankine, C. D.; Madkhali, M. M.; Penfold, T. J. A deep neural network for the rapid prediction of X-ray absorption spectra. *The Journal of Physical Chemistry A* 2020, 124, 4263–4270.
- (13) Erdemir, A.; Donnet, C. Tribology of diamond-like carbon films: Recent progress and future prospects. *Journal of Physics D: Applied Physics* 2006, 39
- (14) Wei, J.; Guo, P.; Liu, L.; Li, H.; Li, H.; Wang, S.; Ke, P.; Saito, H.; Wang, A. Corrosion resistance of amorphous carbon film in 3.5 wt% NaCl solution for marine application. *Electrochimica Acta* 2020, 346, 136282
- (15) Zhao, Q.; Mou, Z.; Zhang, B.; Zhang, X.; Wang, Z.; Wang, K.; Gao, K.; Jia, Q. Revealing the corrosion resistance of amorphous carbon films under heat shock via annealing. *Diamond and Related Materials* 2020, 102, 107692.
- (16) Mangolini, F.; Krick, B. A.; Jacobs, T. D.; Khanal, S. R.; Steller, F.; McClimon, J. B.; Hilbert, J.; Prasad, S. V.; Scharf, T. W.; Ohlhausen, J. A.; Lukes, J. R.; Sawyer, W. G.; Carpick, R. W. Effect of silicon and oxygen dopants on the stability of hydrogenated amorphous carbon under harsh environmental conditions. *Carbon* 2018, 130, 127–136
- (17) Yi, P.; Zhang, D.; Peng, L.; Lai, X. Impact of Film Thickness on Defects and the Graphitization of Nanothin Carbon Coatings Used for Metallic Bipolar Plates in Proton Exchange Membrane Fuel Cells. *ACS Applied Materials and Interfaces* 2018, 10, 34561–34572
- (18) Wang, Y.; Tian, W.; Wang, L.; Zhang, H.; Liu, J.; Peng, T.; Pan, L.; Wang, X.; Wu, M. A Tunable Molten-Salt Route for Scalable Synthesis of Ultrathin Amorphous Carbon Nanosheets as High-Performance Anode Materials for Lithium-Ion Batteries. *ACS Applied Materials and Interfaces* 2018, 10, 5577–5585
- (19) Li, Y.; Hu, Y.-S.; Li, H.; Chen, L.; Huang, X. A superior low-cost amorphous carbon anode made from pitch and lignin for sodium-ion batteries. *Journal of Materials Chemistry A* 2016, 4, 96–104
- (20) Lu, P.; Sun, Y.; Xiang, H.; Liang, X.; Yu, Y. 3D Amorphous Carbon with Controlled Porous and Disordered Structures as a High-Rate Anode Material for Sodium-Ion Batteries. *Advanced Energy Materials* 2018, 8

- (21) Brandes, J. A.; Cody, G. D.; Rumble, D.; Haberstroh, P.; Wirick, S.; Gelinas, Y. Carbon K-edge XANES spectromicroscopy of natural graphite. *Carbon* 2008, 46, 1424–1434.
- (22) Watanabe, H.; Okino, R.; Hanamura, K. Structural evolution of carbon deposition on a Ni/YSZ cermet of a SOFC analyzed by soft x-ray XANES spectroscopy. *International Journal of Hydrogen Energy* 2019, 44, 24028–24035
- (23) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters* 2010, 104, 136403
- (24) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Physical Review B* 2017, 95, 094203
- (25) Plimpton, S.; Kohlmeyer, A.; Thompson, A.; Moore, S.; Berger, R. LAMMPS Stable release 29 October 2020. 2020; <https://doi.org/10.5281/zenodo.4157471>.
- (26) Csányi, G.; Winfield, S.; Kermode, J. R.; De Vita, A.; Comisso, A.; Bernstein, N.; Payne, M. C. Expressive Programming for Computational Physics in Fortran 95+. *IoP Comput. Phys. Newsletter* 2007, Spring 2007
- (27) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics* 1984, 81, 511–519.
- (28) Prendergast, D.; Galli, G. X-ray absorption spectra of water from first principles calculations. *Physical review letters* 2006, 96, 215502.
- (29) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I., et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* 2009, 21, 395502
- (30) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* 1996, 77, 3865
- (31) Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Physical review B* 1990, 41, 7892.
- (32) Prendergast, D.; Louie, S. G. Bloch-state-based interpolation: An efficient generalization of the Shirley approach to interpolating electronic structure. *Physical Review B* 2009, 80, 235126

- (33) Shirley, E. L. Ab initio inclusion of electron-hole attraction: Application to x-ray absorption and resonant inelastic x-ray scattering. *Physical review letters* 1998, 80, 794
- (34) Vinson, J.; Kas, J. J.; Vila, F.; Rehr, J.; Shirley, E. Theoretical optical and x-ray spectra of liquid and solid H₂O. *Physical Review B* 2012, 85, 045101
- (35) Linear-response and real-time time-dependent density functional theory studies of core-level near-edge x-ray absorption. *Journal of chemical theory and computation* 2012, 8, 3284–3292
- (36) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics* 2011, 134, 074106
- (37) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* 2016, 18, 13754–13769.
- (38) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. 2017; <https://arxiv.org/abs/1704.06439>.
- (39) Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; <https://www.tensorflow.org/>, Software available from tensorflow.org
- (40) Chollet, F., et al. Keras. 2015; <https://github.com/fchollet/keras>
- (41) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830
- (42) Aarva, A.; Deringer, V. L.; Sainio, S.; Laurila, T.; Caro, M. A. Understanding X-ray spectroscopy of carbonaceous materials by combining experiments, density functional theory, and machine learning. Part I: Fingerprint spectra. *Chemistry of Materials* 2019, 31, 9243–9255
- (43) Sainio, S.; Wester, N.; Aarva, A.; Titus, C. J.; Nordlund, D.; Kauppinen, E. I.; Leppanen, E.; Palomaki, T.; Koehne, J. E.; Pitkanen, O., et al. Trends in carbon, oxygen, and nitrogen core in the X-ray absorption spectroscopy of carbon nanomaterials: a guide for the perplexed. *The Journal of Physical Chemistry C* 2020, 125, 973–988
- (44) Chen, W.-T.; Hsu, C.-W.; Lee, J.-F.; Pao, C.-W.; Hsu, I.-J. Theoretical analysis of Fe K-edge XANES on iron pentacarbonyl. *ACS omega* 2020, 5, 4991–5000

- (45) Osswald, S.; Yushin, G.; Mochalin, V.; Kucheyev, S. O.; Gogotsi, Y. Control of sp^2/sp^3 carbon ratio and surface chemistry of nanodiamond powders by selective oxidation in air. *Journal of the American Chemical Society* 2006, 128, 11635–11642.
- (46) Marks, N.; McKenzie, D.; Pailthorpe, B.; Bernasconi, M.; Parrinello, M. Microscopic structure of tetrahedral amorphous carbon. *Physical review letters* 1996, 76, 768.
- (47) Ma'zdziarz, M.; Mrozek, A.; Ku's, W.; Burczy'nski, T. Anisotropic-Cyclicgraphene: A New Two-Dimensional Semiconducting Carbon Allotrope. *Materials* 2018, 11.
- (48) Ranganathan, R.; Rokkam, S.; Desai, T.; Keblinski, P. Generation of amorphous carbon models using liquid quench method: A reactive molecular dynamics study. *Carbon* 2017, 113, 87–99.
- (49) Jiang, X.; Arhammar, C.; Liu, P.; Zhao, J.; Ahuja, R. The R3-carbon allotrope: a pathway towards glassy carbon under high pressure. *Scientific Reports* 2013, 3, 1–9.
- (50) Clark, S.; Crain, J.; Ackland, G. Comparison of bonding in amorphous silicon and carbon. *Physical Review B* 1997, 55, 14059.
- (51) Drabold, D.; Fedders, P.; Stumm, P. Theory of diamondlike amorphous carbon. *Physical Review B* 1994, 49, 16415.
- (52) Dutta, S.; Wakabayashi, K. Magnetization due to localized states on graphene grain boundary. *Scientific reports* 2015, 5, 1–9.
- (53) Karamad, M.; Magar, R.; Shi, Y.; Siahrostami, S.; Gates, I. D.; Farimani, A. B. Orbital graph convolutional neural network for material property prediction. *Physical Review Materials* 2020, 4, 093801.
- (54) Diaz, J.; Monteiro, O.; Hussain, Z. Structure of amorphous carbon from near-edge and extended x-ray absorption spectroscopy. *Physical Review B* 2007, 76, 094201.

Chapter 8. Summary and Outlook

In this dissertation, we utilized a combination of density functional theory (DFT) and machine learning (ML) techniques to investigate the chemical properties of various materials. Initially, we focused on studying periodic systems, specifically the metal-organic framework (MOF), and determined the factors contributing to its hydrogen storage capacity. Our DFT calculations revealed the correlation between charge delocalization, d orbital occupancy, and hydrogen adsorption energy. We anticipate that our insights aim to discover new MOFs with enhanced hydrogen adsorption.

Additionally, we analyzed the structural and electron/hole transport properties of single- and double-stranded DNA. We utilized the deformation potential formalism to compute the electron/hole mobilities of various single-stranded and double-stranded DNA structures. Our findings indicated that poly(T) and poly(C) are effective hole conductors, while poly(A) and poly(G) structures are better at conducting electrons. Moreover, by analyzing the B3LYP-D orbitals and electronic charges in the double-stranded DNA structures, we discovered that poly(A-T) functions as an electron conductor, while poly(G-C) is a more effective hole conductor. We anticipate that our findings can help develop DNA-based microelectronics.

We then investigated the HMF electrooxidation reaction on Cu,Co-spinel oxides and discovered the role of Cu in the adsorption process. Also, by calculating the energy profile of HMFOR on CuCo_2O_4 following both HMFCA and DFF pathways, we explained why only a tiny amount of DFF is detected in experiments. We anticipate that our findings can help develop an electrocatalyst for HMF oxidation.

In the latter part of the dissertation, we employed ML techniques to extend our study to a more significantly large number of molecules with diverse variations. Firstly, we trained deep neural networks using the Density Functional Theory (DFT) data of PFAS, allowing for the accurate and efficient prediction of their bioactivities based on their chemical structure, thus advancing our understanding of the structure-property relationship. Then, using a semi-supervised metric learning algorithm, we automatically classify and cluster functional groups that could play a role in bioactivity prediction. As a result, we anticipate that our model can provide a more efficient screening for PFASs that can be used to complement experimental assessments.

We also developed a simple yet powerful NN model that can accurately simulate local XANES of amorphous carbons, given only chemical structure information. The model only requires information about the chemical structure and is computationally inexpensive compared to ab initio methods. Furthermore, the predicted XANES spectra are consistent with those obtained from DFT calculations, and it can accurately predict spectral characteristics such as peak positions and intensities. We anticipate that our finding provides a more efficient analysis of spectroscopy that can be used to experimental measurements.

In future works, we anticipate solving inverse prediction using machine learning to predict materials' atomic structures from desired chemical properties. For example, we expect that directly predicting possible amorphous carbon structures from XANES spectroscopy can be achieved using graph neural network techniques.

Also, the ML approach can be beneficial for accelerating molecular dynamics calculations by training the deep neural network model using the ab initio molecular dynamics simulations. Using this deep neural network-based molecular dynamics, we can explore chemical reactions in large systems such as batteries with more extended time scales.

Overall, this dissertation comprehensively investigates various materials using DFT and ML techniques, providing insights into their chemical properties and structure-property relationships. We anticipate that the multiple roles of ML applied in computational chemistry and materials science research shown in this dissertation will be far more extended in future works. In this dissertation, we present that ML can be applied to accelerate exploring structure-property relationships. It can also be used to interpret complex systems with a large supercell which would be computationally expensive if we only rely on DFT calculations. In short, machine learning combined with first-principles calculations enables investigation of larger systems with cheaper computational costs.

Appendix A. Publication List

1. **Kwon, H.**, Jiang, D. E. Tuning Metal–Dihydrogen Interaction in Metal–Organic Frameworks for Hydrogen Storage. *The Journal of Physical Chemistry Letters*, **13**, 9129-9133, (2022)
2. **Kwon, H.**, Ali, Z.A., Wong, B.M. Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per-and Polyfluoroalkyl Substances (PFASs). *Environmental Science & Technology Letters*, <https://doi.org/10.1021/acs.estlett.2c00530>
3. Yang, S.D., Ali, Z.A., **Kwon, H.**, Wong, B.M. Predicting Complex Erosion Profiles in Steam Distribution Headers with Convolutional and Recurrent Neural Networks. *Industrial & Engineering Chemistry Research*, **61**, 8520-8529, (2022)
4. Biswas, P., Ghildiyal, P., **Kwon, H.**, Wang, H., Alibay, Z., Xu, F., Wang, Y., Wong, B.M., Zachariah, M.R. Rerouting Pathways of Solid-State Ammonia Borane Energy Release. *The Journal of Physical Chemistry C*, **126**, 48-57, (2021)
5. Mondal, S., Powar, N.S., Paul, R., **Kwon, H.**, Das, N., Wong, B.M., In, S.-I., Mondal, J. Nanoarchitectonics of Metal-Free Porous Polyketone as Photocatalytic Assemblies for Artificial Photosynthesis. *ACS Applied Materials & Interfaces*, **14**, 771-783, (2021)
6. S.S.R.K.C., Yamijala, **Kwon, H.**, Guo, J., Wong, B.M., Stability of Calcium Ion Battery Electrolytes: Predictions from Ab Initio Molecular Dynamics Simulations. *ACS Applied Materials & Interfaces*, **13**, 13114-13122, (2021)
7. Bentel, M.J., Yu, Y., Xu, L., **Kwon, H.**, Li, Z., Wong, B.M., Men, Y., Liu, J. Degradation of perfluoroalkyl ether carboxylic acids with hydrated electrons: Structure–reactivity relationships and environmental implications. *Environmental Science & Technology*, **54**, 2489-2499, (2020)
8. Biswas, S., **Kwon, H.**, Barsanti, K.C., Myllys, N., Smith, J.N., Wong, B.M., Ab initio metadynamics calculations of dimethylamine for probing pK_b variations in bulk vs. surface environments. *Physical Chemistry Chemical Physics*, **22**, 26265-26277, (2020)
9. Raza, A., Bardhan, S., Xu, L., Yamijala, S.S.R.K.C., Lian, C., **Kwon, H.**, Wong, B.M., A Machine Learning Approach for Predicting Defluorination of Per-and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environmental Science & Technology Letters*, **6**, 624-629, (2019)

10. Lian, C., Ali, Z.A., **Kwon, H.**, Wong, B.M. Indirect but Efficient: Laser-Excited Electrons Can Drive Ultrafast Polarization Switching in Ferroelectric Materials. *The Journal of Physical Chemistry Letters*, **10**, 3402-3407, (2019)