MACHINE-LEARNING MODELS FOR ANALYSIS OF BIOMASS REACTIONS AND PREDICTION OF REACTION ENERGIES

A Dissertation
Presented to
The Academic Faculty

By

Chaoyi Chang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemical & Biomolecular Engineering
Department of Chemical & Biomolecular Engineering

Georgia Institute of Technology

Dec 2021

MACHINE-LEARNING MODELS FOR ANALYSIS OF BIOMASS REACTIONS AND PREDICTION OF REACTION ENERGIES

Thesis committee:

Dr. Andrew J. Medford Chemical & Biomolecular Engineering *Georgia Institute of Technology*	Dr. Fani Boukouvala Chemical & Biomolecular Engineering *Georgia Institute of Technology*
Dr. Matthew Realff Chemical & Biomolecular Engineering *Georgia Institute of Technology*	Dr. Jesse McDaniel Chemistry *Georgia Institute of Technology*
Dr. Carsten Sievers Chemical & Biomolecular Engineering *Georgia Institute of Technology*	

Date approved: October 8, 2021

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Andrew Medford for supporting and advising me on my projects with extraordinary patience. His perseverance, curiosity towards solving the unknown scientific questions, and sharp intuition, new ideas on research has guided me throughout the most challenging problems in my research. He also gave me as much flexibility on studying what I was interested in and always encouraged me to try new techniques. I gained a lot from him, from coding techniques, insight into models to how to approach research questions and his detail guidance has influenced me on my way to become a researcher and a scientist. I appreciate his support for my professional growth, including doing a summer internship and attending workshops and conferences. I am also grateful for my committee members, Dr. Matthew Realff, Dr. Carsten Sievers, Dr. Fani Boukouvala and Dr. Jesse McDaniel for their suggestions about my research. Especially, I would like to thank Dr. Carsten Sievers for sharing experimental data on the molybdenum oxide projects and Dr. Jesse McDaniel for tutorial on OpenMM molecular dynamics. Thanks for GT RBI providing funding for my research and PACE team providing clusters to carry on large scale calculations. In addition, I would like to thank to undergraduate research advisor Dr. Zhongyi Jiang, who had helped me made my way to graduate school and my undergraduate mentor Dr. Guangwei He, who led me into the world of scientific research.

Thanks to Medford group members, Dr. Benjamin Comer, Dr. Xiangyun (Ray) Lei, Dr. Fuzhu Liu, Gabriel Gusmão, Adam Yonge, Yuge (Nicole) Hu, Jagriti Sahoo, Sihoon Choi and Kaylee Tian for listening to my questions and giving me suggestions. Many thanks to Ben for helping me settle down in the group, Ray for helping with AMPtorch coding, Gabriel and Adam for scheduling subgroup meetings and finding interesting paper to read. I would also like to thank all the staff and members of TAPPI student chapter, especially Dr. Christopher Luttegen for arranging of fun activities. Thanks should also go to Dr. Robert

Foley, who answered all my questions about math and stochastic processes in real detail and with extreme patience.

My friends, Jianyuan Zhai, Jianshan Liao, Xiangyun Lei, Yimeng Lyu, Mingxuan Lu, Yutong Wu, Yamin Zhang and Shengtao Jiang, are always a source of support throughout my graduate school time. Especially thanks to Yimeng Song and Yu Li for encouraging me over the difficulties. Finally, I would like to thank my family. My parents have made me through all the most challenging times and they were, are, and will always be a constant and solid support for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Biomass and derived compounds have the potential to form the basis of a sustainable economy by providing a renewable source of many chemicals. The selective synthesis and conversion of biomass compounds are often catalyzed by transition metal catalysts. Computational screening has emerged as a promising tool for discovery and optimization of active and selective catalysts, but most existing examples focus on small molecule reactions. In this study, the density functional theory (DFT) approach is first validated by comparing computational results to experiments for ethanol conversion over molybdenum oxide. Subsequently, DFT is combined with machine-learning approaches to identify and overcome challenges associated with computational screening of biomass catalysts. A recursive algorithm is used to elucidate possible intermediates and chemical bond cleavage reactions are for linear biomass molecules containing up to six carbons. Machine-learning algorithms based on the Mol2Vec embedding are applied to classify reaction types and predict gas-phase reaction energies and adsorption energies on Rh(111) (MAE $\sim$ 0.4 eV). With the workflow, we are able to combine the physics-based density functional tight binding method with the machine learning model to identify the stable binding geometries of biomass intermediates on the Rh (111) surface. Finally, we show preliminary results toward the development of a neural network force field based on the Gaussian multipole feature approach. The results indicate that this strategy is a promising route toward fast and accurate predictions of both energies and forces of hydrocarbons on a range of transition-metal surfaces. The results of this thesis demonstrate the utility of machine-learning techniques for studying biomass reactions, and indicate the potential for further developments in this field.

Contents

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Definition of biomass and the importance of biomass compounds

Biomass compounds are a kind of compounds that can be derived from natural products like lignin, which is the most abundant biopolymer in nature and the side-product of biofuel production. Biomass and their derived compounds are the basis of a sustainable economy by producing renewable sources of many chemicals. In particular, the Department of Energy has reported the top biorefinery chemicals in 2004 [1]. Most of these biorefinery chemicals are able to convert to each other under transition metal catalysts and the ability to selectively convert between biomass compounds are important. For example, sorbitol can be produced directly from glucose, which also has a high volume commercial market and smaller compounds, including xylitol and lactic acid, can be produced from sorbitol [2]. Glycerol, as one of the smallest biorefinery compounds, its catalytic conversion to glyceraldehyde and ethylene glycol is widely studied [3, 4, 5]. It is found that both the structure and reaction pathways of glycerol should be similar to sorbitol [6].

1.2 Prior computational studies of biomass compounds on transition-metal surfaces

1.2.1 Different biomass molecules/intermediates on transition metal surface

Biomass products are the raw material for small molecules like ethanol and are able to convert to other biomass compounds on transition metal surfaces. Selective conversion between those biomass products and other small molecule products are important if biomass compounds are to form the basis of a sustainable economy. Rh is a commonly-used transition metal catalyst for conversion of biomass compounds and has been the subject of numerous experimental studies. For example, Rh catalysts have been used for succinic acid conversion to fumaric acid [7], hydrogenolysis of furfural to 1,2-pentanediol [8] and production of C1 compounds from ethanol [9, 10, 11]. However, there have been relatively few systematic computational studies of biomass intermediates on Rh surfaces [12]. Density functional theory (DFT) is the most common theory used to study the adsorption and reaction of biomass compounds [13, 14, 15]. However, complex molecules have multiple binding sites and various geometries, and DFT calculations of larger molecules are expensive and require significant computational effort to converge. Thus, previous studies often utilize empirical and machine learning (ML) methods to relate reaction properties of biomass molecules or intermediates to their structures, physical properties and even experimental conditions [16, 17, 14, 18]. Besides Rh, there are also study on other transition metals. Vlachos and co-workers have been investigating the small molecule reactions on Pt surfaces by DFT. They studied the elementary bond-cleavage reactions between C-2 species and found that CH_3C is the most stable C-2 species on both surfaces and that ethane dissociation and CH_3CH dehydrogenation are rapid processes on Pt (111) surface [19]. They also developed a group additivity method to estimate the properties on Pt surfaces [20, 21]. Heyden and coworkers have also contributed to the biomass molecule and transition metal catalyst system. They have investigated the cleavage mechanism of the C-O ether bond in the lignin model compound over Pd (111) catalyst surface and proposed the reaction pathway as dehydrogenation first to form ketone and enol, then C-O bond cleavage afterward [22]. They have studied the solvent effects in the hydrodeoxygenation of levulinic acid over Ru (001) surface and their microkinetic models suggest that water facilitates the reaction kinetics significantly than other solvents and the solvent effect is strongest at lower temperature [23].

1.2.2 An introduction to model compound

Most biorefinery molecules contain more than 4 carbon (C) and/or oxygen (O) atoms. The large structure of these molecules creates challenges in studying fundamental properties like adsorption energy and binding configurations due to large system sizes in calculations, similarity of different functional groups in spectroscopic investigations, and the vast number of possible intermediates or spectators that may be present in either experimental or computational studies of reaction kinetics. As an example, Figure 1.1 shows the average number of possible intermediates based on the total number of C and O atoms. A molecule with 12 heavy atoms has more than 10,000 possible intermediates. Clearly, it is not feasible to study all elementary steps since manually setting up experiments or even calculations on a metal surface slab could take years or decades to finish. This fact has driven the use of smaller "model compounds" for studying biomass reactions. A model compound is a compound that has similar structures and properties to the target compound, but is smaller and less complex. A common example would be the use of glycerol to study xylitol/sorbitol. Previous studies have shown the similarity between glyceraldehyde/glycerol and linear glucose/sorbitol [24], and also between C-C scission in glycerol/ethylene glycol and xylitol/ethylene glycol or propylene glycol over transition metals [25, 26]. These results illustrate the effectiveness of glycerol as a model compound, and the similarities are consistent with chemical intuition. Other examples of model compounds are phenolic molecules used to study the hydrodeoxygenation of lignin derivatives on MoO_3, [27] and fufuranic/aromatic compounds used as model feedstocks of biooils on metal catalysts. [28] Despite the widespread use of model compounds to study complex reactions, there is no systematic way of evaluating the efficacy of a model compound, or of identifying model compounds for other complex molecules/reactions.

As mentioned previously, a key challenge with biomass molecules is their complexity, and "model compounds" are often used to simplify systems of interest. A "model compound" refers to a relatively small molecule (typically fewer than 8 heavy atoms) that can be used for studying larger compounds with similar chemical properties and functional groups. Previous experimental studies showed that the reaction pathway of glyceraldehyde <=> glycerol is similar to linear glucose <=> sorbitol [13, 29, 15, 30] and that propionic

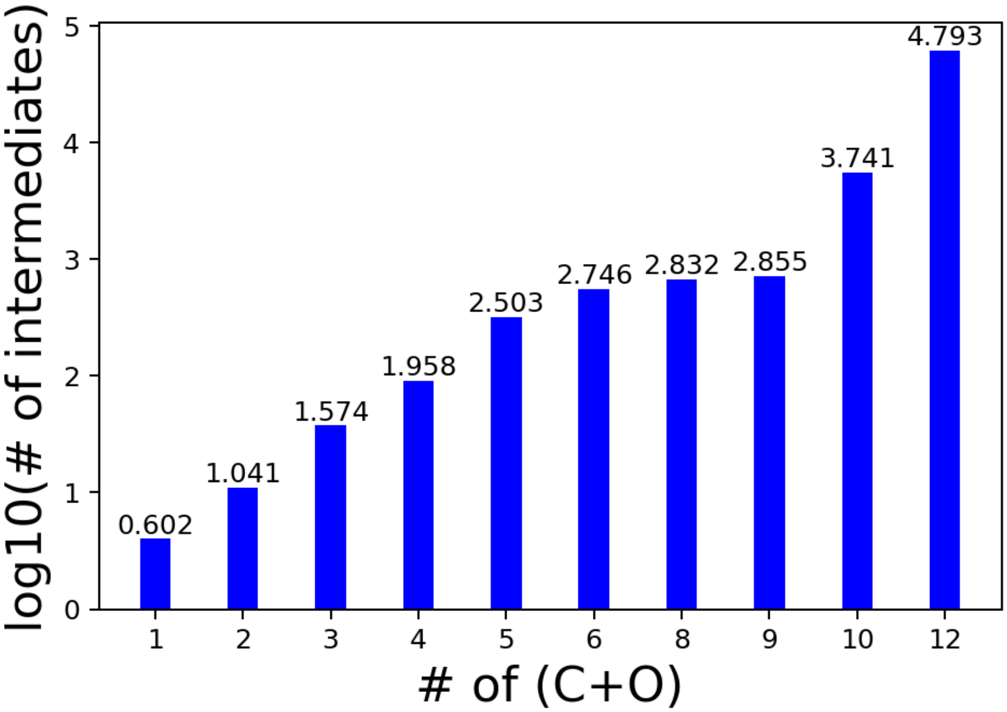


Figure 1.1: Average number of possible intermediates according to the total number of C atoms and O atoms based on the recursive bond-breaking algorithm used in this work.

acid ketonization is similar to the larger carboxylic acid ketonization on Rh surface [14, 31]. While model compounds are commonly used, they are typically identified by heuristics and intuition. However, a data-driven strategy for systematic identification of model compounds is proposed in chapter 4 [32]. It was pointed out that glycerol, glyceraldehyde, propionic acid, and erythrose are good model compounds to study larger molecules like sorbitol and linear-structured glucose. This makes these four molecules a key starting point for computational and experimental studies seeking a more general understanding of the catalytic conversion of biomass derivatives on solid surfaces.

1.3 Review of quantitative and machine learning (ML) methods on predicting properties

1.3.1 Review of softwares and machine learning methods

Various properties can be predicted by combining fingerprints of molecular structures or other encoding methods with regression and classification models. This approach of quantitative structure-property models has been widely used in the pharmaceutical industry. Several software packages are developed to predict the properties based on the potential drug structures. For example, SYNCHEM, an early computer program was created to search synthesis procedures and predict drug-like properties [33]. SYNCHEM is also able to estimate the properties of the compounds not in their database with a substructure matching then estimating based on the most similar compounds in structure. Chematica, which took more than 10 years to be developed, further allows selecting the most promising compounds and the optimal synthesis pathway by optimizing a cost function [34]. Chematica is taking structure information and energies from previous publications for providing a feasible reaction pathway.

Recently, machine learning techniques have also been applied to drug discovery. Deep learning has been implemented for both training and encoding the compounds into vector spaces. ProtVec tool, an open-source python package based on natural language processing (NLP), is developed for embedding proteins based on their biological sequence [35]. Deep neural networks can also be used to predict protein-protein interactions (PPI) [36] and predict the secondary structures of proteins [37]. Random forest (RF) models have been used to predict drug-like interactions with FP2 fingerprints [38] and assess chemical toxicity [39]. Support vector machines (SVM) have been used to predict the effect of ion doping in pharmaceuticals [40] and enzyme selectivity [41]. Artificial neural networks are also widely used in understanding pharmaceutical lipophilicity and other properties [42].

While machine learning techiques have the longest history in pharmaceutical molecules, they have also been applied to numerous other domains including MOFs/zeolites [43, 44], superconductivity [45, 46] and catalysis [16]. RINGS is an open-source software developed for estimating the properties of chemical compounds. RINGS could do the analysis of topological networks based on the part of the structural information which can be represented in the graph theory using nodes for the atoms and links for the bonds. [47]

Specifically for catalysis systems, machine learning is also playing a central role in catalyst screening and property prediction such as atomization [48], formation energies [49], density of states [50], band gaps [51] and melting temperatures [52]. Machine learning is used in finding new materials, scaling relations, and DFT-derived functions. [53] Various machine learning methods have been applied to the crystal structure of AB-type alloys [54] and explored the metal oxide chemical space in a search of previously uncharacterized compounds [55]. Important contributions have been made to establishing scaling relationships between adsorption energies and reactants and intermediates using d-band model [56, 57]. Several reports have also described machine learning predictions of the adsorption energies of reactants and surface intermediates, including DFT-derived adsorption energies on metal and alloy surfaces [58, 59], and to predict d-band centers, [60, 61] which are widely used descriptors for reactions over heterogeneous transition metal surfaces. Heyden and coworkers have been using data-driven model to predict transition energies on Pd surface [12]. CatMAP, [62] an open-source Python package, developed for proposing microkinetic models for catalysts, can also be used to do the screening of different catalysts for specific reactions. Machine learning has aided the design of specific reaction catalysts [63] and has been used for the discovery of inorganic complexes [64]. Gaussian process regression has been used in predicting redox potentials [65] and Bayesian approaches have been used to predict solubility [66]. However, there have been relatively few applications of machine learning techniques to biomass molecules and reaction networks [17, 12]. Most of these ML methods could reach a mean absolute error of adsorption energy/transition state energy within 0.4 eV, [18, 20, 12] with some examples of predictions as accurate as 0.2 eV using a combination of physical and structural features along with feature selection methods [67, 17].

1.3.2 Fingerprinting and encoding methods

Identification and evaluation of model compounds and analysis of biomass molecules, can be facilitated by developing vector representations or "fingerprints" for these molecules [68, 69]. Encoding methods are a strategy for quantitative comparison of different molecular compounds. There are several ways to encode molecular structure including SMILES notation [70], Morgan fingerprint extended connectivity fingerprints (ECFP) [71], FP2 fingerprints[38], group matrices [72], and neural networks [73, 74, 75]. SMILES is a very commonly used linear string-like notation for chemical compounds, and is the basis for many subsequent encoding methods like barcode encoding [76] or artificial neural network encoders [77]. However, SMILES notation can be complicated when the system is large. Morgan fingerprints and ECFP fingerprints generate lists of integers from the molecular structure and properties like atomic number and atomic mass, providing a vector of integers for a given structure, and FP2 fingerprints work similarly but hash the result into a binary representation. Encoding by group is another approach for creating a vector of integers and was originally proposed by Benson [72]. The idea of group encoding is to split the whole molecule into one-heavy-atom-center substructures and encode the number of appearances of each substructure into a matrix. Group additivity has been extended in several ways to increase accuracy and treat more complex systems [78, 79]. For example, Vlachos and co-workers have applied group additivity to predict the adsorption energies of biomass molecules on metal surfaces and in aqueous environments [18, 80, 20]. Group matrices are convenient for regression to molecular properties since matrix algebra is well-established, and the resulting vectors have intuitive meaning (contributions of each group). However, the accuracy is often limited due to the low number of groups, and the method relies on a test set that contains multiple examples of all groups. Encoding molecular structures through neural networks is an alternative approach that leads to a vector of continuous values that corresponds to each molecular structure. These continuous vectors are advantageous because they are natural inputs to machine learning algorithms like regression and classification. Neural network encoders can also be tuned to provide vectors of different sizes, providing a way to arbitrarily adjust the complexity of molecular encoding. Moreover, if an auto-encoding architecture is used, the encoded representation can be transformed back into a molecular representation, although challenges exist with ensuring that the resulting inverse transform is chemically valid [81, 82].

1.4 Summary of the thesis

In this thesis, we first validated DFT approach using a system where experimental data was available. Next, we applied ML methods to the clustering, classification and regression of molecular biomass structures, reactions, and reaction energies. We also proposed new workflows for assessing the model compounds quantitatively and identifying the stable binding geometries on metal surfaces. Finally, we present initial efforts toward the construction of a neural network force field (NNFF) that is generally applicable for biomass and transition metal catalytic systems.

The DFT approach is validated through the energy and vibrational frequency calculation of small biomolecules on MoOx surface. Both energy and frequency are in qualitative agreement with the experimental results then DFT with Quantum ESPRESSO is used for calculation in the following work. In the gas-phase study, we apply the “Mol2Vec” embedding algorithm to analyze the plethora of compounds and elementary steps involved in the conversion of biomass molecules. Unsupervised dimensional reduction and clustering algorithms applied to molecular vectors provide a more detailed approach for classifying elementary steps in biomass reaction networks into a total of 90 possible classes. Linear discriminant analysis (LDA) were applied to the results and found to be able to classify elementary steps into these subclasses with a classification accuracy of 0.99 with a reduced dimensionality of 5 components. The resulting LDA decomposition is also used to assess regression models for gas-phase reaction energy, resulting in a minimum error of 0.59 eV, compared to 2.9 eV for group additivity. Finally, we apply the machine-learned representations to quantitatively assess the ability of different model compounds to represent different reaction types. The results indicate that neural-network based encoding techniques are a promising approach for understanding the complex reaction networks of biomass reactions.

Then, we extend the previously-developed embedding models of gas-phase formation energies to adsorbed surface intermediates on the Rh(111) surface. A total of 171 intermediates from the first 2 bond-breaking recursions of erythrose, glyceraldehyde, glycerol and propionic acid are studied. Mol2Vec is used for generating vector descriptors and 83 clusters based on 6 reaction types (C-C, C-O, O-H, C-H, C-M, O-M) are obtained from single-group "radius zero" (R0) Mol2Vec descriptors. Linear discriminant analysis (LDA) and partial least squares (PLS) are used for dimensional reduction of two- and three-group "radius one" (R1) Mol2Vec descriptors, providing low-dimensional vector descriptors for each adsorbate. These vectors are combined with a linear least-squares regression model, yielding mean absolute errors (MAE) as low as 0.39 eV. Finally, pre-optimization via density-functional tight binding (DFTB) is combined with our embedding models to establish a workflow for rapidly identifying stable adsorption geometries. We show that this workflow identifies 20 new lowest-energy geometries for 171 adsorbates studied, indicating that systematic approaches for identifying the lowest-energy structures of large adsorbed molecules are a necessary addition to the tool set of computational catalysis.

Further more, a NNFF based on AMPtorch is in development and the test errors (MAE) of the present NNFF are 0.11 eV for for energy and 0.06 for force on atoms. The dataset contains DFT calculations of smaller molecules consisting of no more than 4 C and 4 O on transiton metal (211) surfaces. Adsorbates on Rh are first tested with AMPtorch and then transferred to multiple metals. The NNFF based on AMPtorch could be a potential substitution of Hotbit as a pre-optimization tool.

1.5 Summary of the introduction

Biomass molecules pose a significant challenge due to their size and complexity. As mentioned above, the number of possible intermediates is increasing exponentially with the size of molecule and there exists more than 10,000 possible intermediates for a molecule containing more than 10 heavy atoms. Also, the binding geometry become more complex as the number of atoms increases. Machine learning approaches have been useful for making rapid property predictions in numerous other fields of chemistry and biochemistry (drug discovery for example) and machine learning could be a potential method for studying the biomass systems. However, there have been relatively few efforts to apply machine learning to biomass molecules, making this a promising research direction.

CHAPTER 2
ALGORITHMS AND METHODS

2.1 DFT and DFTB

2.1.1 Schrödinger's equation

The famous Schrödinger's equation [83] is the equation that all quantum chemistry calculations start with to solve:

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi \tag{2.1}$$

and its original form is:

$$\hat{H}\Psi = E\Psi \tag{2.2}$$

Both the $\hat{H}$ (Hamiltonian) and the Ψ (wavefunction) are complex and depend on each other. The Hamiltonian is an operator that is related to the described physical system [84]. It is commonly expressed as the sum of operators corresponding to the kinetic ($\hat{T}$) and potential energies ($\hat{V}$) of a system. The potential energies can be expressed by the summation of electron-nucleus ($\hat{V_{en}}$) and electron-electron energy ($\hat{V_{ee}}$). Catalysis systems typically contain adsorbates and metal atoms with multiple nuclei and free electrons. Thus, we need to solve for the equation with Hamiltonian that can represent many-body system. Initially,

$$\hat{H} = \hat{T} + \hat{V_{en}} + \hat{V_{ee}} \tag{2.3}$$

and $\hat{T}$, $\hat{V_{en}}$, $\hat{V_{ee}}$ can be calculated by :

$$\hat{T} = -\frac{1}{2}\sum_{i}^{Ne}\nabla_i^2 \tag{2.4}$$

$$\hat{V_{en}} = \sum_{i}^{Ne}\sum_{I}^{Nn}\frac{-Z_I}{|\overrightarrow{r_i} - \overrightarrow{R_I}|} \tag{2.5}$$

$$\hat{V_{ee}} = \frac{1}{2}\sum_{i}^{Ne}\sum_{j\neq i}^{Ne}\frac{1}{|\overrightarrow{r_i} - \overrightarrow{r_j}|} \tag{2.6}$$

where Ne or Nn is the number of electrons or nuclei in the system, ∇^2 is the Laplacian operator, $\overrightarrow{r_i}$ is the spatial coordinate of electron i, $\overrightarrow{R_I}$ is the spatial coordinate of nucleus I [85] and Z_I is the charge of nucleus I.

The wavefunction (Ψ) is depending on the coordinates of all the electrons in the system at a specific time ($\Psi = \Psi(\overrightarrow{r_1}, ..., \overrightarrow{r_{Ne}}) = \Psi(\overrightarrow{r_i})$) with $3Ne$ dimensions as each electron should have 3 spatial coordinates. The wavefunction is complex-valued and in this case it could describe all the electrons.

However, solving for the position of electrons and the energies simultaneously are not possible for a many-body system. Thus, analytical solutions to the Schrödinger's equation is not possible and different numerical methods with different accuracy are used. For example, Hartree-Fock and perturbation theory are both the approximations to the wavefunction and Hamiltonian. While the accuracy of the approximation methods increase, the complexity and cost also increases. For most of these methods, for example, Hartree-Fock, requires solving N-coupled equations for N spin orbitals and a recursive solution for those equations known as “self-consistent field method” (SCF) until convergence although it approximates the N-body wave function by a single Slater determinant. So approximation methods to wavefunction are still very computationally costly and require a constraint on the size of studied system.

2.1.2 Kohn-Sham equations

In a many-body system, only the electron density but not the wavefunction could be observed. The electron density means the expectation value of the number of electrons at a specific point in the system space and the expectation value could be calculated by the probability of an electron being found at that point. Indeed, the wavefunction could be used to fully describe a many-body system for interacting electrons, but the wavefunction is not observable as the electron probability or electron density. But the electron density can be obtained from the wavefunction:

$$n(\overrightarrow{r}) = N \int \cdots \int_{R_3} \Psi^*(\overrightarrow{r}, \overrightarrow{r_2} \ldots, \overrightarrow{r_{Ne}}) \Psi(\overrightarrow{r}, \overrightarrow{r_2} \ldots, \overrightarrow{r_{Ne}}) d\overrightarrow{r_2} \ldots d\overrightarrow{r_{Ne}} \tag{2.7}$$

$$= N \int \cdots \int_{R_3} |\Psi(\overrightarrow{r}, \overrightarrow{r_2} \ldots, \overrightarrow{r_{Ne}})|^2 d\overrightarrow{r_2} \ldots d\overrightarrow{r_{Ne}} \tag{2.8}$$

where $|\Psi|^2$ is the probability of the wavefunction density, which is normalized to 1 by definition and Ψ^* is the complex conjugate of Ψ. The electron density only contains 3 spatial dimensions and is real-valued, also observable. Thus it is much easier to deal with than the wavefunction with complex-value and $3Ne$ dimensions.

The electron density is the foundation of DFT as well as the two theorems proved by Hohenberg and Kohn [86]. The conclusion of the two Hohenberg-Kohn theorems is that the exact electron density could be obtained by optimizing (usually minimizing) the energy with respect to the electron density ($\frac{\partial E[n_{exact}(\overrightarrow{r})]}{\partial n(\overrightarrow{r})} = 0$) given the exact energy functional ($E_{exact}[n(\overrightarrow{r})]$). This leads to a 3-dimensional optimization (minimization) problem and could be solved by numerical methods. But solving for the exact energy and defining the energy functional still remains a challenge. Equation 2.3 can be expanded in the form of electron density as energy is the eigenvalue (expectation) of Hamiltonian:

$$E_{ne} = -\int_{R_3} n(\overrightarrow{r}) \sum_{i}^{Nn} \frac{Z_I}{|\overrightarrow{r_i} - \overrightarrow{R_I}|} d\overrightarrow{r} \tag{2.9}$$

$$= -\int_{R_3} n(\overrightarrow{r}) V_{ne}(\overrightarrow{r}) d\overrightarrow{r} \tag{2.10}$$

$$E_{ee} = \frac{1}{2} \int_{R_3} \int_{R_3} \frac{n_2(\overrightarrow{q}, \overrightarrow{r})}{|\overrightarrow{q} - \overrightarrow{r}|} d\overrightarrow{q} d\overrightarrow{r} \tag{2.11}$$

where n_2 is the density correlation function and provides the probability given that the other electron exists at $\overrightarrow{r}$ with an electron exists at $\overrightarrow{q}$. Some approximations are necessary since the density of the two-electron system cannot be obtained by the single-electron density. The two-electron density could be obtained by just taking the product of two single-electron densities if the two particles are independent. And this independent correlation gives us (some correction is implemented):

$$n_2(\overrightarrow{q}, \overrightarrow{r}) = n(\overrightarrow{q}) n(\overrightarrow{r}) + \delta n_2(\overrightarrow{q}, \overrightarrow{r}) \tag{2.12}$$

which suggests:

$$E_{ee} = \frac{1}{2} \int_{R_3} \int_{R_3} \frac{n(\overrightarrow{q}) n(\overrightarrow{r})}{|\overrightarrow{q} - \overrightarrow{r}|} d\overrightarrow{q} d\overrightarrow{r} + \triangle E_{ee} \tag{2.13}$$

So the conclusion here is that both Coulombic potential energy and electron potential energy are functionals of electron density:

$$E_{en}(n(\overrightarrow{r})) = -\int_{R_3} n(\overrightarrow{r}) V_{ne}(\overrightarrow{r}) d\overrightarrow{r} \tag{2.14}$$

$$E_{ee}(n(\overrightarrow{r})) = \frac{1}{2} \int_{R_3} \int_{R_3} \frac{n(\overrightarrow{q}) n(\overrightarrow{r})}{|\overrightarrow{q} - \overrightarrow{r}|} d\overrightarrow{q} d\overrightarrow{r} \tag{2.15}$$

Above calculations are the estimations of potential energies. We still have not talked about the kinetic energy. To estimate the kinetic energy $\hat{T}$, one key assumption of DFT is used: there exists some collections of Ne non-interacting electrons which have the same ground-state density as the true interacting system.

$$n(\overrightarrow{r}) \overset{\text{assume}}{=} \sum_{i}^{Ne} |\phi_i(\overrightarrow{r})|^2 \tag{2.16}$$

where the $\phi_i(\overrightarrow{r})$ are the Kohn-Sham orbitals [87] and these orbitals are unspecified initially. The assumption here is that the true density could be expressed by these non-interacting orbitals while actually this might not be the case. The kinetic energy density can be expressed as (if the assumption is true):

$$E_T = -\frac{1}{2}\sum_{i}^{Ne}\int_{R_3}\phi_i^*(\overrightarrow{r})\nabla^2\phi_i(\overrightarrow{r})d\overrightarrow{r} + \triangle E_T \tag{2.17}$$

$$E_T(\phi_i(\overrightarrow{r})) = -\frac{1}{2}\sum_{i}^{Ne}\int_{R_3}\phi_i^*(\overrightarrow{r})\nabla^2\phi_i(\overrightarrow{r})d\overrightarrow{r} \tag{2.18}$$

where $\triangle E_T$ is the term representing the difference between the true kinetic energy and the kinetic energy derived from non-interacting Kohn-Sham orbitals. After this substitution, the energy can be expressed as:

$$E = E_T[\phi_i(\overrightarrow{r})] + E_{en}[n(\overrightarrow{r})] + E_{ee}[n(\overrightarrow{r})] + \triangle E_{ee}[n_2(\overrightarrow{q}, \overrightarrow{r})] + \triangle E_T[\Psi(\overrightarrow{r_1}, \ldots, \overrightarrow{r_{Ne}})] \tag{2.19}$$

where the first 3 terms (E_T, E_{en}, E_{ee}) are known while the last 2 terms ($\triangle E_{ee}$, $\triangle E_T$) are unknown. $\triangle E_{ee}$ and $\triangle E_T$ are representing the compensate for the approximations in E_{ee} and E_T. It is proven by the first Hohenberg-Kohn theorem that the energy is a functional of the density. This suggests that these energies (E_{ee}, E_T and E_{en}) are also functionals of the density. Therefore we can combine the two energy correction terms together and the combined term is known as "exchange correlation" energy:

$$E_{xc}[n(\overrightarrow{r})] = \triangle E_{ee}[n_2(\overrightarrow{q}, \overrightarrow{r})] + \triangle E_T[\Psi(\overrightarrow{r_1}, \ldots, \overrightarrow{r_{Ne}})] \tag{2.20}$$

Various different approximations have been developed for the exchange-correlation ($E_{xc}[n(\overrightarrow{r})]$) since it is impossible to calculate the true exchange-correlation energy analytically. With the exchange-correlation part, the energy functional can be written as:

$$E[\phi_i(\overrightarrow{r})] = E_T[\phi_i(\overrightarrow{r})] + E_{en}[n(\overrightarrow{r})] + E_{ee}[n(\overrightarrow{r})] + E_{xc}[n(\overrightarrow{r})] \tag{2.21}$$

The above equation is a functional of Kohn-Sham orbitals and the electron density and the electron density can be estimated by Kohn-Sham orbitals given our previous assumption when deriving the exchange-correlation. So minimizing the function with respect to Kohn-Sham or the electron density should be equivalent. Taking the derivative, the functional can be minimized and gives:

$$[-\frac{1}{2}\nabla^2(\phi_i(\overrightarrow{r})) + V_{eff}(n(\overrightarrow{r}))]\phi_i(\overrightarrow{r}) = \epsilon_i\phi_i(\overrightarrow{r}) \tag{2.22}$$

$$V_{eff}(n(\overrightarrow{r})) = V_{ne}(n\overrightarrow{r}) + \int_{R_3} \frac{n(\overrightarrow{q})}{|\overrightarrow{q} - \overrightarrow{r}|} d\overrightarrow{q} + \frac{\delta E_{xc}}{\delta n(\overrightarrow{r})} \tag{2.23}$$

The above equation is known as the famous “Kohn-Sham” equation and is also the foundation of DFT. The original $3Ne$-dimensional wavefunction transforms to Ne 3-dimensional non-interacting Kohn-Sham orbitals.

It is still a challenge to solve the Kohn-Sham equation although the equation is already much simpler than the well-known Schrödinger’s equation. There are many different packages for solving it, for example, VASP, Quantum ESPRESSO, and GPAW [88, 89, 90, 91]. In this thesis, the calculations are carried out with Quantum ESPRESSO and all details are provided in the corresponding chapters. This section is a high-level and very general introduction of the equations, theorem, and approximations to solve the equations. Usually, the numerical method goes through an “SCF” cycle to be converged:

- Choose an initial guess for trial density ($n(\overrightarrow{r})$)
- Solve Equation 2.23 for $\phi_i(\overrightarrow{r})$ using the trial density in V_{eff}

- Calculate the electron density of these solutions as $\tilde{n}(\overrightarrow{r}) = \sum_i^{Ne} |\phi_i(\overrightarrow{r})|^2$

- Compare the trial density $n(\overrightarrow{r})$ with the solution density $\tilde{n}(\overrightarrow{r})$. If the difference is smaller than the cutoff, this density is considered the ground-state electron density. Otherwise, method like gradient-descent is applied to obtain a new trail density and Step 2 - Step 4 are repeated.

The electronic ground state is determined by the "SCF cycle" but one assumption that the exact coordinates of the nuclei are known is applied here. However, this is not the case in reality but this can be overcome. The Hellman-Feymann theorem says $\overrightarrow{F_I} = \overrightarrow{F_I}[n(\overrightarrow{r})]$, where F_I is the force on nucleus I [92]. So we know that the forces on nuclei can be obtained from the electron density. The local minima can be determined given the initial guess:

- Choose an initial guess for the atomic coordinates $\overrightarrow{R_I}$

- SCF cycle to calculate the ground truth energy for the given initial guess coordinates/optimized coordinates

- Force calculated by taking the derivative of energy

- Check force error. If the maximum force error is within the tolerance then a local minima is found. If not, the atom are moved to new coordinates and Step 2- Step 4 are repeated until convergence

This local minimum obtained is highly depending on the initial guess and it is not necessarily guaranteed as a global minimum. Several initial guesses should be attempted here to get a "global minimum".

2.1.3 Selection of DFT parameters

DFT provides a systematic way to solve the Schrödinger's equation, it still has the limitation that the Hohenberg-Kohn theorems only prove that the ground-state energy is a functional of the electron density. However, most chemistry can be captured by the ground-state calculations. Also, Kohn-Sham orbitals are not equivalent to the true wavefunctions, thus the eigenvalues are not the true eigenvalues of the system, although the practical results show they are reasonable. In addition, the practical limitation of the true exchange correlation ($E_{xc}[n(\overrightarrow{r})]$) is not known. The PBE functional is one of the earliest and most successful of the surface-tailored functionals, while the BEEF-vdW functional allows uncertainty estimation and performs slightly better for adsorption systems. Another strategy is to abandon the constraint that the exchange correlation is a functional of density and instead compute the approximate form of exchange correlation based on wavefunction method. For example, Hartree-Fock can be included in the exchange-correlation energy.

When solving the Kohn-Sham equations numerically, there are many numerical parameters that must be selected through convergence testing. The common parameters of DFT calculations include the plane-wave cutoff, k-point sampling, and the exchange correlation functionals. There is a trade-off between cost and the accuracy of the estimated energy. The cost is determined by numerical accuracy that controlled by plane-wave cutoff and k-points. The higher the value of planewave cutoff and the k-points, the higher the calculation cost and the higher the accuracy of the estimated energy. Here we are using planewave cutoff of 400 eV - 450 eV, k-point sampling[93] of $4\times4\times1/5\times5\times1$ and PBE/BEEF-vdW as exchange-correlation functionals. All density functional calculations are performed with Quantum ESPRESSO [88]. A BFGS algorithm provided by Atomic Simulation Environment (ASE)[90] was applied to the geometry optimization until the maximum force was no more than 0.05 eV/Å.

2.1.4 Density functional based tight binding theory

Density functional tight-binding (DFTB) provides a rapid physics-based route that provides relatively accurate energies and geometry of reactive surfaces [94, 95]. DFTB theory is first proposed by G. Seifert. The formalism of optimized linear combination of atomic orbitals as introduced by Eschng and Bergert [96] for band-structure calculations are introduced. The Kohn-Sham orbitals ψ_i of the system are expanded in terms of atom-centered localized basis functions ϕ_ν:

$$\psi_i(r) = \sum_k^K \sum_\nu C_{\nu i} \phi_\nu(\boldsymbol{r} - \boldsymbol{R_k}) \tag{2.24}$$

where $C_{\nu i}$ is the linear coefficient and is a constant [96], K is the number of unit cells, R_k is position. For orbital ψ_i, solve the following Schrödinger-like equation in a linearized way:

$$\hat{H}\psi_i(\boldsymbol{r}) = \epsilon_i \psi_i(\boldsymbol{r}) \tag{2.25}$$

To find ϵ_i and the linear coefficient $C_{\nu i}$, multiply both sides of the Schrödinger's equation, integrate over all space, and we obtain equations that could be solved in a matrix form.

$$\alpha_1 C_{1i} + \beta_{1\mu} \sum_{nn,\mu \neq 1} C_{\mu i} = \epsilon_i C_{1i} \tag{2.26}$$

$$\alpha_2 C_{2i} + \beta_{2\mu} \sum_{nn,\mu \neq 2} C_{\mu i} = \epsilon_i C_{2i} \tag{2.27}$$

$$\vdots \tag{2.28}$$

$$\alpha_i C_{\nu i} + \beta_{\nu\mu} \sum_{nn,\mu \neq \nu} C_{\mu i} = \epsilon_i C_{\nu i} \tag{2.29}$$

where $\alpha_i = \int \phi_i^* \hat{H} \phi_i$ and $\beta_{ij} = \int \phi_i^* \hat{H} \phi_j = \int \phi_j^* \hat{H} \phi_i$ and nn representing nearest neighbors. By representing the linear equations in a matrix form, it is easy to find that the energy ϵ is the eigenvalue and the $C_{\mu i}$s are the eigenvectors. The total energy of the system can be approximated as the summation of the band-structure energy and a short-range repulsive two-body potential:

$$E_{tot}(\boldsymbol{R_k}) = E_{BS}(\boldsymbol{R_k}) + E_{rep}(|\boldsymbol{R_k} - \boldsymbol{R_l}|) \tag{2.30}$$

$$= \sum_i n_i \epsilon_i(\boldsymbol{R_k}) + \sum_k \sum_{<l} V_{rep}(|\boldsymbol{R_l} - \boldsymbol{R_k}|) \tag{2.31}$$

where n_i is the occupation number of orbital i, ϵ_i is the eigenvalue of the non-self-consistent Schrödinger-like equation (Equation 2.25) and V_{rep} is a short-range pairwise repulsion between the atoms at R_l and R_k. In DFTB, the electronic-electronic, nuclear-nuclear repulsions are assumed to be pariwised. In this case, the ϵ_i is now the solutions of non-self-consistent Schrödinger's equation but not the self-consistent one. The original DFTB ignores the self-consistency and and assumes that all the important nonpairwise behavior in the interatomic forces comes from the sum of the one-electron eigenvalues.

G. Seifert and coworkers proposed a polynomial approach to estimate the repulsive potential (V_{rep}) [94]. The "actual" repulsive potentials were obtained by taking the difference of the total energy (calculated by self-consistent approach) and the band-structure energy ϵ with different values of interatomic distances R:

$$V_{rep}(R) = E^{sc}(R) - \epsilon \tag{2.32}$$

Finally, the $V_{rep}(R)$ could be written as a sum of polynomials:

$$V_{rep}(R) = \begin{cases} \sum_{n=2}^{NP} d_n (R_c - R)^n & (R < R_c) \\ 0 & \text{otherwise} \end{cases} \tag{2.33}$$

In the same paper, the authors determined the coefficients of the polynomial expansion for short-range repulsive potential of C and H. After the very first polynomial expansion, there also exists other methods to estimate the repulsive and non-repulsive potential more accurately, for example, G. Seifert later published a paper taking electron-nuclear long-distance energy into consideration [97] and method based on the two-center parametrization of Slater and Koster also Harrison's tight binding theory [98].

We use the open-source Hotbit [99] Python package with a previously-developed parameterization for Rh/C/H/O [100] for performing DFTB to get pre-optimized local minimas. The Hotbit calculator is used with constrained minima hopping [101] where metal slab atoms have fixed positions and adsorbate bonds have fixed lengths [102] to generate local minima geometries for each adsorbate.

2.2 Molecular structure generation

Gas-phase species data generation is based on a chemical-bond-breaking recursion algorithm. Figure 2.1 shows an example of CH_3OH group generation from the original CH_3OH molecule: each chemical bond is broken for the input substructures in each recursion, and only the unique groups are kept; the process is recursed until no new groups are generated. The recursive bond breaking of CH_3OH leads to the generation of 5 unique substructures of the 1st recursion, 4 unique substructures of the 2nd recursion, 3 unique substructures of the 3rd recursion and 2 unique substructures of the 4th recursion. The recursive bond breaking stops at the 4th recursion since no more unique substructures will be generated (C and O) further or only single atoms remain, i.e. C. The unique reactions are generated following the same algorithm as the substructures. The 1st recursion of CH_3OH is taken as an example here. The algorithm generates 3 unique reactions in the 1st recursion, which are: 1) C-H bond breaking: $CH_3OH \rightarrow CH_2OH + H$; 2) O-H bond breaking: $CH_3OH \rightarrow CH_3O + H$; 3) C-O bond breaking: $CH_3OH \rightarrow CH_3 + OH$. Then each of the generated substructures goes through the same process and finally 14 substructures are generated from CH_3OH. Double bonds are broken simultaneously, for example, $C{=}C \rightarrow C + C$. 14 small biomass molecules are taken as its original input, finally obtaining more than 90k possible intermediates and chemical bond breaking reactions. More details please see chapter 4.

The advantages of this recursive bond breaking algorithms include: 1) obtaining all possible intermediates of a given molecule; 2) having a clear stop criterion; and 3) intermediates from any level (recursive number) could be obtained and this algorithms could be useful where intermediates from the 1st and 2nd recursions are extracted. However, the algorithm also has some limitations, which are 1) it is a brute force method so larger molecules would take an unexpected longer time to finish all recursions and 2) it cannot

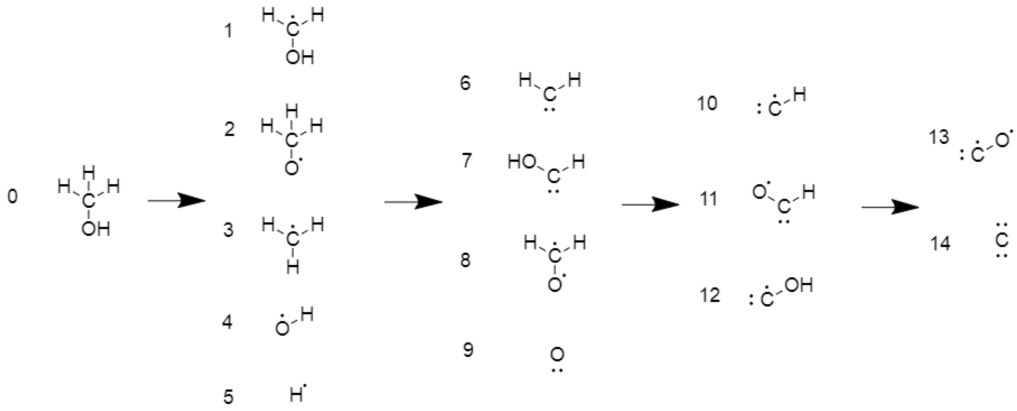


Figure 2.1: Recursion bond breaking of CH_3OH with the generation of 5 unique groups of the 1st recursion, 4 unique groups of the 2nd recursion, 3 unique groups of the 3rd recursion and 2 unique groups of the 4th recursion. The recursive bond breaking stops at the 4th recursion due to no more unique substructures will be generated further or only single atoms left.

deal with the mechanisms of chemical bond formation.

The above algorithm is for the data generation of the molecules themselves. Surface species are also needed to study. However, surface species cannot be obtained by an approach similar to the gas-phase species since we cannot obtain the system of adsorbates and surfaces directly from any existing database. A surface species generation algorithm is needed here. Surface species data generation is based on a metal-adding algorithm (see Figure 2.2) and yield elementary surface reactions similarly to gas-phase reactions. The process is first finding the unsaturated atoms and saturate these atoms one by one. See chapter 5 for detailed examples. Also, there does not exist an algorithm to convert the atom coordination files (e.g., .xyz file, .traj file) to SMILES notations. We also write an algorithm to convert the coordination files to SMILES notations (which is the input of Mol2Vec). First, metal atoms within the covalent bond distance of the adsorbate atoms are kept and their coordinations are saved. Next, whether a chemical bond is formed is based on the distance between the atoms and we assign single chemical bonds here since there are intermediates with free electrons thus conjugate occurs and chemical bond order cannot represent the actual situation. (Please see https://github.com/cchang373/molecule_rxn)

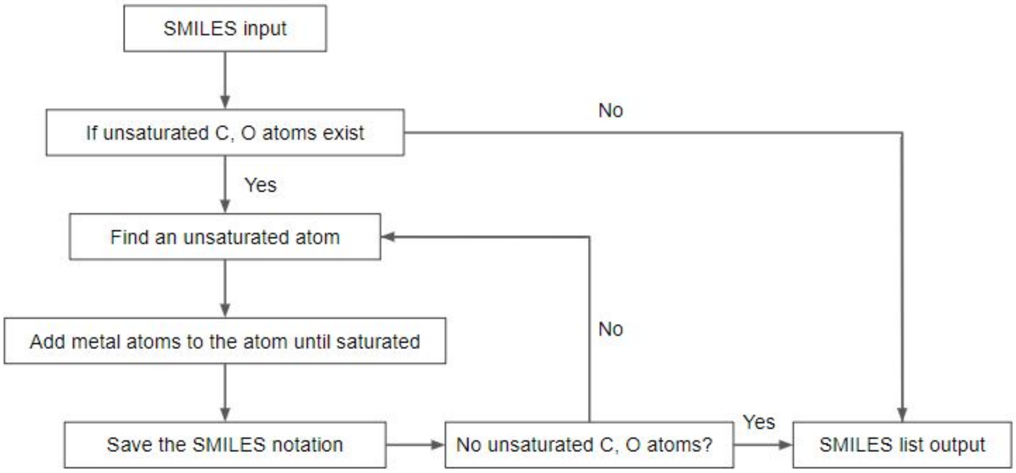

SMILES input
If unsaturated C, O atoms exist
No
Yes
Find an unsaturated atom
Add metal atoms to the atom until saturated
No
Save the SMILES notation
No unsaturated C, O atoms?
Yes
SMILES list output

Figure 2.2: Generating algorithms of SMILES notation of adsorbate with metal atoms

2.3 Molecule encoding and Mol2Vec

The Mol2Vec[73] algorithm is used to generate the unsupervised features of substructures obtained from the recursive bond-breaking algorithm and metal-adding algorithm described above. The two hyper-parameters of the algorithm are: dimension and radius; dimension is the size of vector generated by Mol2Vec for each substructure and radius refers to how large each group is. Mol2Vec is an embedding method for chemical structures based on Word2Vec [103], an NLP (natural language process) tool that developed by Google to extract word meaning. The skip-gram is a shallow two-layer neural network. It takes a large amount of context corpus and outputs a vector space (weight) based on the nearby words and Figure 2.3 shows skip-gram structure.

Different modified extended connectivity fingerprint (ECFP) [71] are implemented in Mol2Vec when generating vector descriptors for gas-phase and surface species (see chapter 4 and chapter 5) respectively. The steps of ECFP algorithm includes: 1) an initial stage assigned to each atom with an integer identifier; 2) a second iterating stage where the identifier of each atom is updated from the information of the atom's neighbors, including

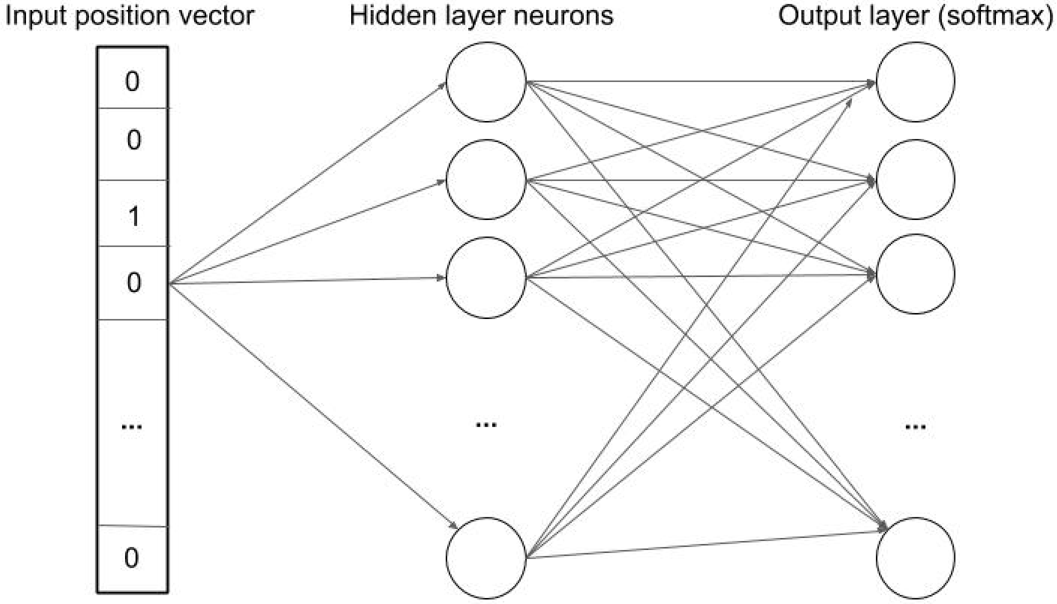


Figure 2.3: Word2Vec skip-gram neural network structure

identification of whether it is a structural duplicate of other features; and 3) a final stage where the duplicate identifiers are removed but the occurrence count is retained in the final feature list. The original ECFP contains 6 invariants: atomic number, number of heavy atoms around, number of hydrogen atoms around, charge, mass, and whether the atom is in rings. Valence and electronegativity are added to the invariants for a more clear classification of C and O atoms, and the invariant of the atom is in a ring is removed since this analysis is restricted to linear-structures.

Methanol is still taking as an example to illustrate how ECFP works. In radius 0, CH_3OH can be divided to 2 substructures: CH_3 and OH and the atom dictionary containing the neighbors and properties are hashed to integers shown in the green OH and the red CH_3. Next step, in radius 1, the whole molecule is included and only counted once since we do not want duplicate structures, the information of hashed integers of radius 0 are as the values or inputs of the radius 1 dictionary and the dictionary is hashed into another integers. For methanol, the first 2 iterations (radius 0 and radius 1) cover every piece of the molecule. The iteration would stop when either all pieces (group size assigned by the radius input) are covered or it reaches the assigned radius.

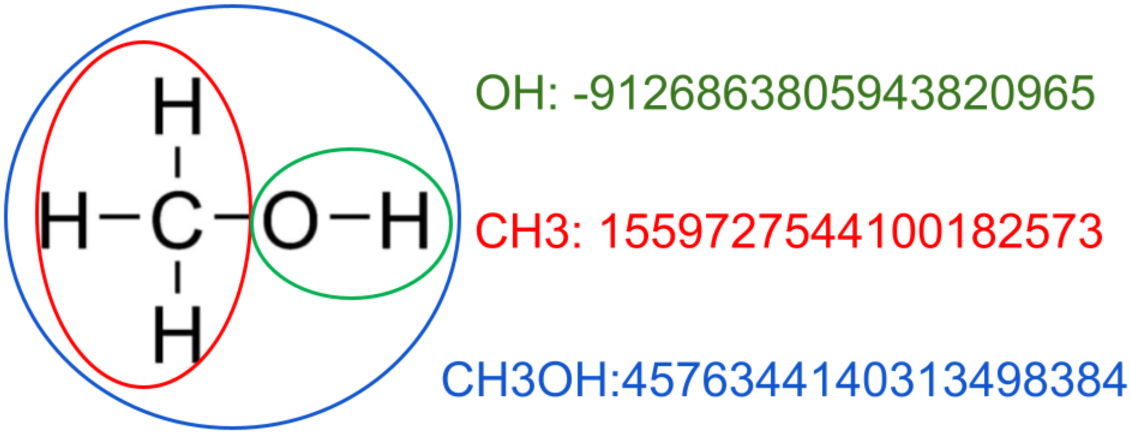


Figure 2.4: Illustration of how ECFP works with the example of methanol, radius 0 contains 2 substructures: CH_3 and its hashed integer in red, OH and its hashed integer in green; radius 1 contains 1 more (sub)structure: CH_3OH and its hashed integer in blue

As Figure 2.4 shows, methanol with radius 1 (or more than radius 1) should result in three hashed integers: 1559727544100182573, -9126863805943820965 and $457634414-0313498384$. Each of these integer descriptors is equivalent to one "word" in Word2Vec and all three words form a "sentence". The projection from Mol2Vec to Word2Vec is that each group of the structure is similar to a word and the structure containing multiple groups is similar to a sentence. "paragraph" in Mol2Vec is the multiple structures contained in the input file. Finally, several "sentences" form a "paragraph" and the structure "paragraph" is the input of Mol2Vec.

2.4 Dimensional reduction, regression, and clustering

Mol2Vec gives us vector descriptors of 200 dimensions and descriptors with high dimension are not easy to conceptualize or visualize. ML methods are used here for reducing the dimension and analyzing the results. Principal component analysis (PCA) [104] is used for reducing dimensions and generating unsupervised features. PCA is a linear and unsupervised dimension reduction method. It is based on singular value decomposition (SVD) and eigenvalue/eigenvector calculation to project the original high-dimension vectors to lower-dimension vectors with orthogonal columns. Linear discriminant analysis (LDA) [105] is a linear supervised classification method and performs dimension reduction accordingly. The algorithm of LDA is also based on SVD and the difference with PCA is that LDA is supervised while PCA in unsupervised. It is used for the supervised classification of bond breaking reactions and dimension reduction for all radii. LDA projections from high-dimension to low-dimension are implemented for the vector descriptor of reaction and adsorption and then followed by regression method to predict the energy. Ordinary least squares (OLS) and partial least squares (PLS) are used for reaction energy and adsorption energy regression. OLS and PLS are both regression methods. OLS takes all the input dimensions while PLS does a feature extraction first and then regression based on the feature extraction. PLS is also a dimension reduction method but totally supervised and limited to the vector and the target to be regressed.

A mean shift [106] algorithm with bandwidths from 0.1 to 15 is applied to identify clusters, with the optimal bandwidth determined by the maximum silhouette score. Meanshift is an unsupervised cluster method and the algorithm is to converge $m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$ where $N(x)$ is the neighborhood of s and $K(x_i - x)$ is the kernel. Silhouette score is an assessment of the validation of consistency within clusters of data. It is based on Euclidean distance and is measuring the intra-cluster distance difference with inter-cluster difference. All machine-learning packages are based on the `scikit-learn` implementations [107].

CHAPTER 3
DFT APPROACH VALIDATION: SMALL BIO-MOLECULES ON MOOX CALCULATION

3.1 Introduction

Catalytic process plays an important role in lignocellulosic biomass conversion. Density functional theory(DFT) can help understand the mechanism, transitional state and further design of catalysts. Molybdenum oxides have been the subject of a limited number of computational investigations [108, 109, 110, 111, 112] with varying levels of theory including GGA[111, 112] and hybrids. Previous work have shown that MoO_3 is the most commonly used catalyst due to the moderate temperatures[110]. Vlachos et al have studied the mechanism of the retro-aldol fragmentation of fructose on MoO_3 (010) using DFT [109]. They have found that fructose could undergo epimerization to hamamelose followed by retro-aldol fragmentation and further converted to alkyl. It was through the C-C bond activation by proton abstraction β-OH group with a terminal oxygen and followed by the C-C bond scission. Besides fructose there are still other biomass DFT studies reported for MoO_3. The Formox process is reported by Rellan-Pineiro and Lopez [108]. Methanol is oxidised to formaldehyde, in which molybdenum oxides doped with iron are the catalysts. They have found that Mo(VI) cycles to Mo(IV) and serves as the active center if forming an O-vacancy and Fe doping could increase the activity by means of DFT. The O-vacancy reconstruction is a classical reaction pathway that most MoO_3 oxidation reaction would follow. DFT calculations performed on the multistep hydrodeoxygenation of acetone on MoO_3(010) followed the O-vacancy pathway [111]. Green and Shetty also illustrated the creation of the oxygen vacancy would make the α-MoO_3 surface approach metallic be-

[0] All experimental results were contributed by S. Najmi and C. Sievers.

havior. And another DFT calculation on acetaldehyde hydrodeoxygenation supports that [112].

Here, we report DFT calculations on some small biomass molecules, ethanol, methanol, acetaldehyde, formaldehyde, glycolaldehyde and crotonaldehyde for example using MoO_x as the slab. We report adsorption energies and the possible reconstruction involved in the adsorption process. In addition, the impact of O-vacancy formation is also studied. However, the O-vacancy defect concentration in the theoretical calculated MoOx surface is higher and the vacancies are more organized than would be expected for the experimental surfaces (R-MoO_3 and O-MoO_3) due to the periodic boundary conditions. Nonetheless, the local atomic structure of the defect should remain similar, making it possible to calculate the adsorption energies and vibrational frequencies at these sites. Therefore, the theoretical calculated energies and vibrational frequencies are expected to show semi-quantitative agreement with experimental results for surfaces under various states of reduction, and the comparison with oxygen vacancy results should qualitatively correlate with the extent of surface reduction, since the oxygen vacancy concentration will increase as the surface is more reduced.

3.2 DFT approach

Exchange-correlation interactions are treated with the BEEF-vdW functional [113] and a Monkhorst-Pack k-point sampling [93] of $5 \times 5 \times 1$ and a planewave cutoff of 450 eV are used [108]. A Fermi-Dirac electron smearing width of 0.1 eV is used for convergence and spin polarization and dipole correction [114] are included to achieve more accurate results. The slab was modeled using 10Å of vacuum unit cell with periodic boundary condition in x and y direction. The lower two layers were fixed to facilitate adsorption energy calculation but upper layers were relaxed to observe the surface change during the adsorption process. A quasi-Newton algorithm provided by Atomic Simulation Environment (ASE) [90] is applied to the geometry optimization until the maximum force is no more than 0.05 eV/Å.

Initial structures are obtained from previous work [108] to achieve energy-minimized structures. The adsorption energies are calculated as follow:

$$E = E_{slab+ads} - E_{slab} - E_{ads} \tag{3.1}$$

where E is the adsorption energy, $E_{slab+ads}$, E_{slab} and E_{ads} are the energy of the system(slab and adsorbate), slab and adsorbate respectively.

Vibrational modes were estimated with the normal mode analysis by using a finite difference approximation with difference delta of 0.01Å of the Hessian matrix provided by ASE. Adsorbate gas molecules were calculated based on the optimized structure previously obtained from DFT calculation. The imaginary and low frequency modes are not compared with experimental results because they arise due to transitional/rotational modes that are neglected in the harmonic approximation, or cannot be clearly classified.

3.3 Adsorption energy on MoOx surface calculation

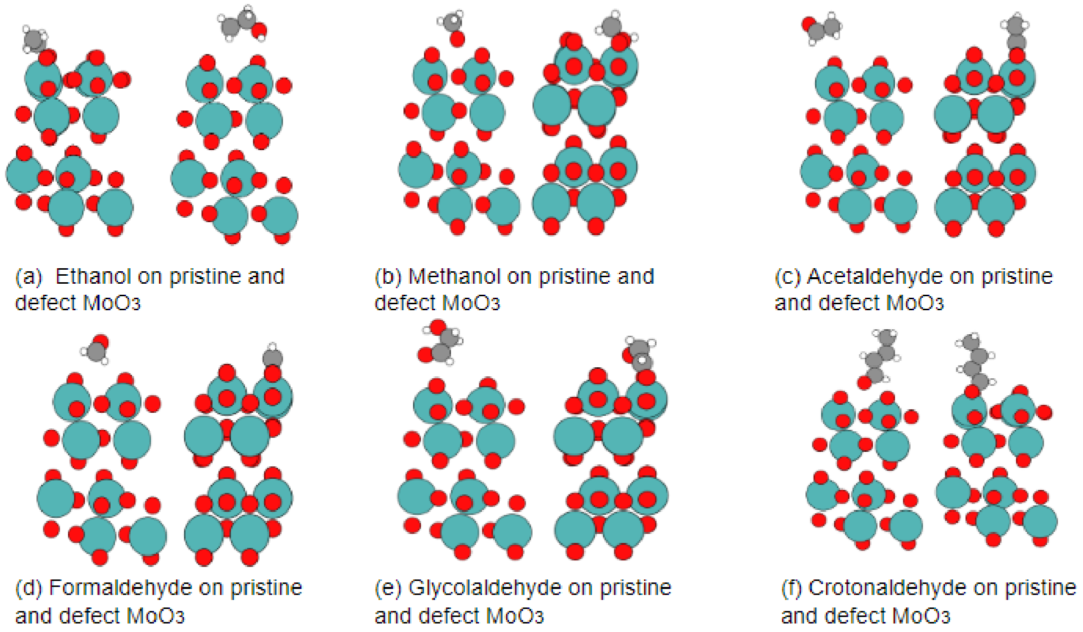
DFT calculations were used to investigate how the presence of oxygen vacancies on MoO_3 influences the binding of small biomass compounds. The optimized adsorption geometries for ethanol, methanol, acetaldehyde, formaldehyde, glycolaldehyde and crotonaldehyde on O —MoO_3 indicate only nondirectional bonds and relatively low adsorption energies of roughly —0.20 eV (see Figure 3.1 and Table 3.1). These results provide evidence that all molecules are physisorbed on O —MoO_3. In contrast, there was a clear directed interaction between the alcohol/aldehyde group and the oxygen vacancy site, and binding energies were substantially more exergonic (< —0.75 eV), indicating strong chemisorption at the vacancy sites. This suggests that defect sites are critical to activating alcohol and aldehyde functionalities. The results that ethanol and acetaldehyde physi-sorbed on the pristine MoO_3 surface and chemi-sorbed on the defect MoO_3 surface via an O through the O vacancy on the surface are in qualitative agreement with the experiment results. Our experimental collaborators provided the DRIFTS of acetaldehyde adsorbed on to different MoOx surfaces Figure 3.2 [115]. The overtone peaks in DRIFTS of acetaldehyde adsorbed on R-MoOx suggest that the acetaldehyde are coordinated onto the surface via oxygen atom and through the vacancy site. Whereas, these overtone peaks do not appear on the O —MoO_3, supporting the physi-adsorbion. The reason that DRIFTS results of O-MoO_3 are in much difference with R-MoOx and I-MoOx is that O-MoO_3 is the oxidized surface. Our experimental collaborator also suggests that the oxygen vacancy sites could also contribute to the activity and selectivity of the catalyst. Dr. Sievers' group (our collaborator) also proposed potential reaction mechanisms for ethanol on the MoO_3 surface (see Figure 3.3), suggesting that the oxygen vacancy through the loss of oxygen in M=O provides possiblity for the formation of anion-cation sites. The sites are necessary for further dehydrogenation of the ethanol since the adsorption of ethanol is dissociative into a proton and an ethoxide ion. The proposed pathways by the experimental results also implies that the oxygen vacancy is crucial for adsorbates binding. This qualitative agreement between theory and experiment validate the DFT approach as the method to calculate adsorption energies.

Table 3.1: Adsorption energy of different molecules on pristine MoO_3 and defect MoO_3

Molecules	Pristine MoO_3 (eV)	Defect MoO_3 (eV)
Ethanol	-0.26	-0.75
Methanol	-0.26	-0.80
Acetaldehyde	-0.19	-0.92
Formaldehyde	-0.26	-0.90
Glycoaldehyde	-0.21	-0.71
Crotonaldehyde	-0.25	-1.14

3.4 Vibrational frequency calculation

Vibrational frequencies are shown in Table 3.4 and Table 3.2 for both types of surfaces and the adsorbate molecules. For the convenience of visualization, we only keep the vibrational frequencies that are larger than 1000 cm^{-1} in the computational study. Intensity values require density functional perturbation theory, which is unavailable in the code used for this



(a) Ethanol on pristine and defect MoO3

(b) Methanol on pristine and defect MoO3

(c) Acetaldehyde on pristine and defect MoO3

(d) Formaldehyde on pristine and defect MoO3

(e) Glycolaldehyde on pristine and defect MoO3

(f) Crotonaldehyde on pristine and defect MoO3

Figure 3.1: Adsorption geometries on pristine MoO_3 (100) (L) and defect MoO_3 (R) of (a) ethanol, (b) methanol, (c) acetaldehyde, (d) formaldehyde, (e) glycolaldehyde and (f) crotonaldehyde with Mo atoms (teal) and bulk O atoms (red) are represented by their ionic radii, while H (white), C (gray)

work, therefore we only report the frequencies. Ethanol and acetaldehyde vibrational frequency calculated by DFT are compared to experimental results. The experimental results show that for most of the modes, the defect site always caused a slight blueshift of 0 —50 cm^{1}, while the O —H stretch is significantly more red-shifted by around 150 cm^{-1}. However, the effects of the defect site for acetaldehyde are more ambiguous in that most modes are slightly blue-shifted (0 —50 cm^{1}), while the C —C and C —H stretch of the CH group was red-shifted by 45 and 116 cm^{1} respectively. Although the blueshift and redshift are a little bit different in number, the qualitative agreement is still achieved between the theoretical calculation and the experiment.

The experimental vibrational frequency results of ethanol, acetaldehyde and crotonaldehyde are shown in Figure 3.4 and Table 3.5 (also from our experimental collaborator). The region above 3200 cm^{-1} is classified as the hydroxyl stretching region. In Figure 3.4A, bands at 1479 and 1447 cm^{-1} suggested the presence of acetate species on the surface. We

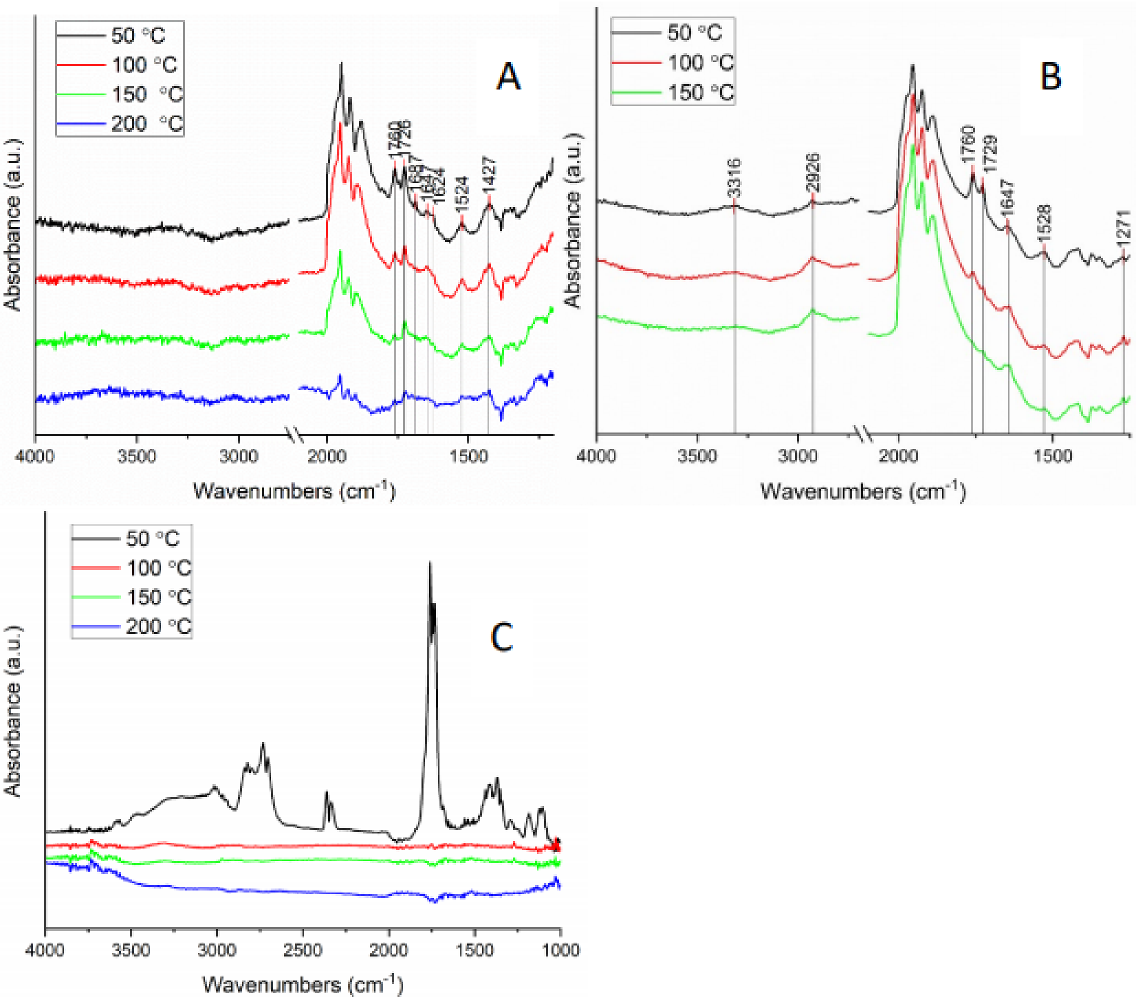


Figure 3.2: DRIFTS spectra of acetaldehyde adsorbed onto (A) R-MoOx (B) I-MoOx (C) O-MoO_3

could also observe 1489.2 cm^{-1}, 1457.2 cm^{-1} in computational results for ethanol. 1625 cm$^{(} - 1)$ on the R–MoOx that appeared around 1645 cm$^{(} - 1)$ is most likely the $\nu(CC)$ mode of crotonaldehyde corresponding to 1634.5 cm$^{(} - 1)$ in computational result. In Figure 3.4B, the are peaks similar to Figure 3.4A. And in Figure 3.4C, the broad peak at 1427 cm^{-1} is assigned to $\delta(CHx)$ modes of various species (acetate species) of acetaldehyde on reduced surface can be seen and 1634.5 cm^{-1} of $\nu(C = C)$ for crotonaldehyde is also in agreement with experimental. Though the computational vibrational frequencies are not perfectly coincide with the experiment, the qualitative agreement suggests that the calculation could still have some insights into the vibrational mode.

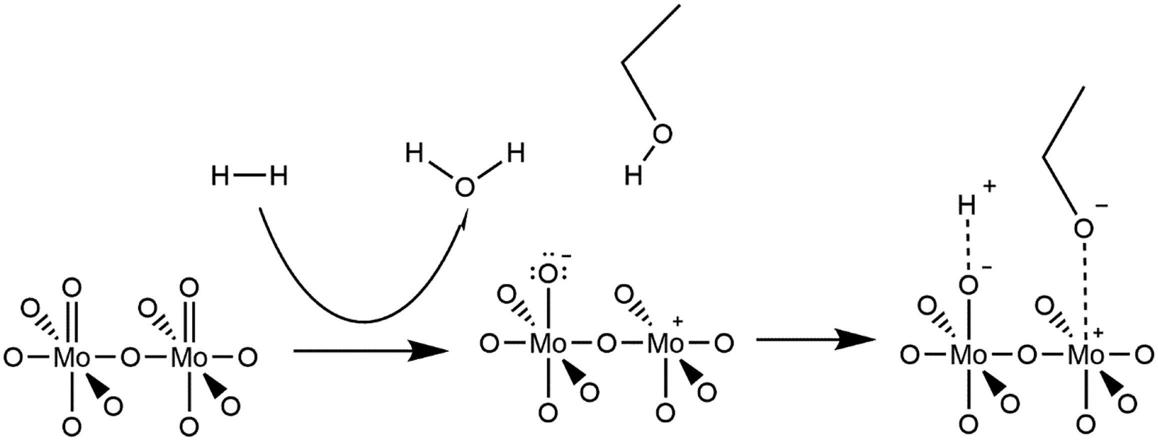


Figure 3.3: Potential mechanism for formation of oxygen vacancies on MoO_3 via hydrogen treatment and the stabilization of the ethoxide ion on the surface

3.5 Conclusion

The computational results are in qualitative agreement with experimental results in both adsorption energy, geometry and vibrational frequency calculations. Thus, DFT calculation with Quantum ESPRESSO is a validated approach of the following energy calculations for gas-phase reaction energy and surface adsorption energy systems.

Table 3.2: Vibrational frequencies of ethanol on pristine MoO_3 and defect MoO_3 (intensity is not available)

Pristine MoO_3 (cm^{-1})	Defect MoO_3 (cm^{-1})
1013.9	
1024.7	1012.9
1093.1	1120.0
1256.7	1174.6
1352.1	1294.2
1380.2	1387.2
1407.6	1409.8
1457.2	1468.6
1489.7	1479.7
1511.0	1496.6
2988.4	3013.7
3010.3	3015.9
3029.4	3064.0
3075.8	3085.1
3135.9	3103.6
3691.9	3545.7

Table 3.3: Vibrational frequencies of crotonaldehyde on defect MoO_3 (intensity is not available from simulations).

Defect MoO_3 (cm^{-1})
1108.3
1197.7
1274.7
1293.3
1311.0
1393.2
1446.7
1484.5
1513.7
1634.5
2829.7
2992.4
3062.8
3072.1
3102.2
3118.6

Table 3.4: Vibrational frequencies of acetaldehyde on pristine MoO_3 and defect MoO_3 (intensity is not available)

Pristine MoO_3 (cm^{-1})	Defect MoO_3 (cm^{-1})
1069	
1106.8	1105.0
1353.7	1356.1
1395.9	1396.3
1447.9	1445.4
1678.5	1564.4
2858.7	2979.3
3022.0	3022.9
3093.1	3040.9
3119.8	3114.0

Table 3.5: Vibrational mode assignments from labeled peaks in Figure 3.4 for species other than ethanol

Mode (cm^{-1})	Vibration (eV)	Species
1268	$\nu(C-O)$	Enol
1393	$\delta_{as}(CH_3)$	Acetaldehyde
1427	$\delta(CH_x)$	Various
1447	$\delta_{as}(CH_3), \nu_s(COO)$	Various
1479	$\delta_{as}(CH_2)$	Various
1524	$\nu_{as}(COO)$	Acetate
1624	$\nu(C=C)$	Enol
1645.7	$\nu(C=C)$	Crotonaldehyde
1726	$\nu(C=O)$	Acetaldehyde (Chemisorbed)
1735	$\nu(C=O)$	Acetaldehyde
1760	$\nu(C=O)$	Acetaldehyde (Physisorbed)

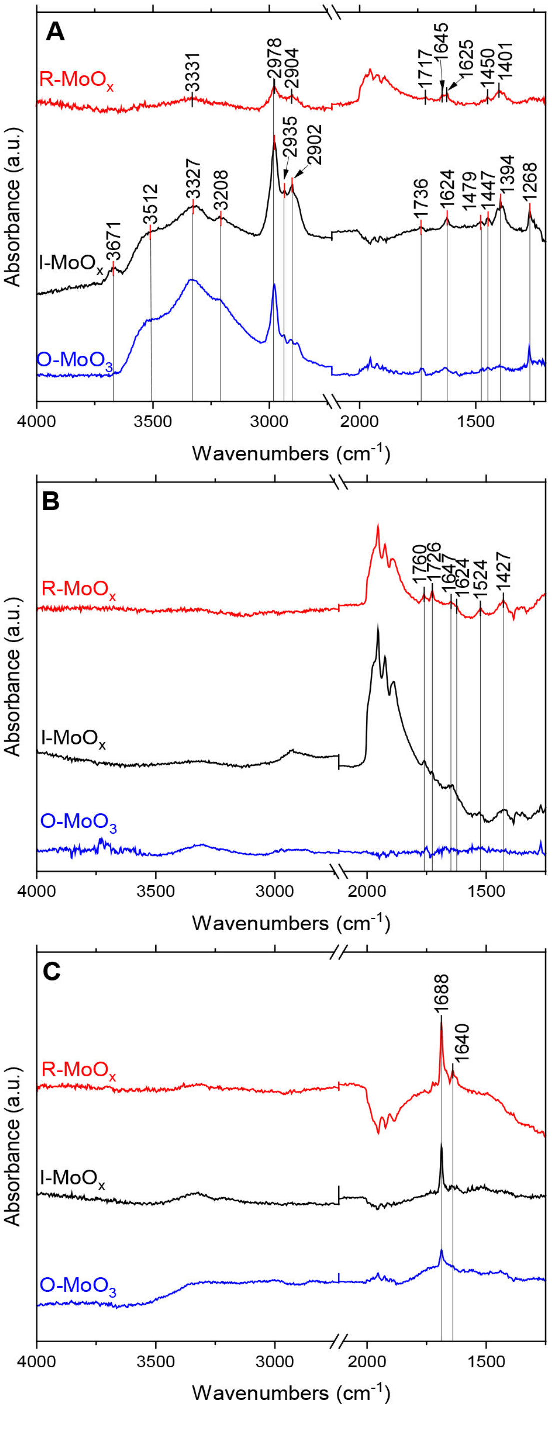


Figure 3.4: Spectral subtraction DRIFTS of (A) ethanol adsorbed onto R–MoOx, I–MoOx, and O–MoO_3 sample at 50 °C. (B) acetaldehyde adsorbed onto R–MoOx, I–MoOx, and O–MoO_3 sample at 100 °C. (C) crotonaldehyde adsorbed onto R–MoOx, I–MoOx, and O–MoO_3 sample at 100 °C

CHAPTER 4

CLASSIFICATION AND PREDICTION OF GAS-PHASE BOND ENERGIES AND MODEL COMPOUND ASSESSMENT

4.1 DFT calculations, gas-phase data generation and Mol2Vec

The gas-phase energies of 189 substructures are calculated with density functional theory (DFT) via the Quantum ESPRESSO[88] package, and the exchange-correlation interactions are treated with the BEEF-vdW functional[113]. Calculations are performed with a planewave cutoff of 400 eV in a Γ-point calculation. A Fermi-Dirac electron smearing width of 0.1 eV is used for convergence and spin polarization was included to account for unpaired electrons. All gas phase groups were modeled using 5Å of vacuum unit cell with periodic boundary condition. A quasi-Newton algorithm provided by the Atomic Simulation Environment (ASE)[90] is applied with a maximum force of 0.05 eV/Å. Reaction energies are calculated as:

$$E_{reaction} = \sum E_{product} - \sum E_{reactant} \tag{4.1}$$

where $E_{reaction}$ is the reaction energy, $E_{product}$ are the energies of products and $E_{reactant}$ represents the energy of reactants, and all energies are computed directly from DFT.

The data contain 10 linear structured biomass molecules from top biorefinery chemicals listed by DOE[1], which are: succinic acid, fumaric acid, malic acid, propionic acid, itaconic acid, levulinic acid, sorbitol, xylitol, glycerol and glucose and 5 biomass chemicals of sorbic acid, muconic acid, 2-hexenedioic acid, 3-hexenedioic acid and erythrose and 33 small molecules that contains no more than 5 carbon atoms. The full list is in Table 4.1 and Table 4.2. Figure 4.1 shows an example of HC(=O)OH group generation from the original HC(=O)OH molecule: each chemical bond is broken for the input substructures in each recursion, and only the unique groups are kept; the process is recursed until no new groups are generated. The recursive bond breaking of HC(=O)OH leads to the generation of 7 unique substructures of the 1st recursion, 4 unique substructures of the 2nd recursion. The recursive bond breaking stops at the 2nd recursion since no more unique substructures will be generated (C, O and H) further or only single atoms remain, i.e. C. The unique reactions are generated following the same algorithm as the substructures. The 1st recursion of HC(=O)OH is taken as an example here. The algorithm generates 4 unique reactions in the 1st recursion, which are: 1) C-H bond breaking: HC(=O)OH → [C](=O)OH + [H]; 2) O-H bond breaking: HC(=O)OH → [CH](O)=O+[H]; 3) C-O bond breaking: HC(=O)OH → [O] + [CH]O and HC(=O)OH → [OH] + [CH][O]. Then each of the generated substructures goes through the same process and finally 11 substructures are generated from HC(=O)OH. Double bonds are broken simultaneously, for example, C=C → C + C. 14 small biomass molecules are taken as its original input, finally obtaining more than 90k possible intermediates and chemical bond breaking reactions. The training set contains 91098 unique substructures with full recursions and 13422 reactions from the first 5 recursions.

Table 4.1: Compounds from DOE report

	Smile Notaion	Rxn Recursion
1	C(CC(=O)O)C(=O)O	5
2	C(=CC(=O)O)C(=O)O	5
3	C(C(C(=O)O)O)C(=O)O	5
4	CCC(=O)O	Full
5	C=C(CC(=O)O)C(=O)O	5
6	CC(=O)CCC(=O)O	5
7	C(C(C(C(C(CO)O)O)O)O)O	5
8	C(C(C(C(CO)O)O)O)O	5
9	C(C(CO)O)O	Full
10	C(C(C(C(C(C=O)O)O)O)O)O	5
11	CC=CC=CC(=O)O	5
12	C(=CC(=O)O)C=CC(=O)O	5
13	C(CC(=O)O)C=CC(=O)O	5
14	C(C=CCC(=O)O)C(=O)O	5
15	C(C(C(C=O)O)O)O	5

As mentioned previously in chapter 2, Mol2Vec with modified ECFP is applied on the

Table 4.2: Small compounds

	Smiles Notation	Rxn Recursion
1	C	Full
2	CC	Full
3	C=C	Full
4	C(C(C=O)O)O	Full
5	CCC	Full
6	CC=C	Full
7	CCCC	Full
8	CCCCC	Full
9	CCC=CC	Full
10	CCCCCO	Full
11	CCC(CC)O	Full
12	CCCC(C)O	Full
13	CCCCO	Full
14	CCC(C)O	Full
15	CC(C)C(=O)O	Full
16	CCCO	Full
17	C=CCO	Full
18	C=CC=O	Full
19	CC=CO	Full
20	CC(C)O	Full
21	CCC=O	Full
22	CC(C)CC	Full
23	CCO	Full
24	C=CO	Full
25	CC=O	Full
26	CCOC	Full
27	CC(=O)C	Full
28	C(CO)O	Full
29	CC(=O)O	Full
30	C(C=O)O	Full
31	CO	Full
32	C=O	Full
33	C(=O)O	Full

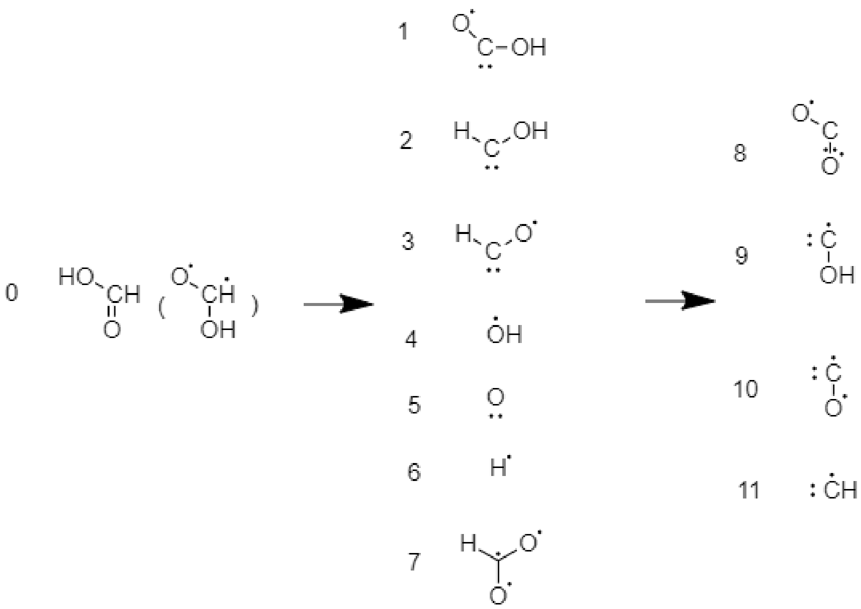


Figure 4.1: Recursion bond breaking of HC(=O)OH with the generation of 7 unique groups of the 1st recursion, 4 unique groups of the 2nd recursion. The recursive bond breaking stops at the 2nd recursion due to no more unique substructures will be generated further or only single atoms left.

91098 structures to generate vector unsupervised features. A modified version of ECFP is implemented into Mol2Vec. The original ECFP contains 6 invariants: atomic number, number of heavy atoms around, number of hydrogen atoms around, charge, mass and whether the atom is in rings. Valence and electronegativity are added to the invariants for more clear classification of C and O atoms, and the invariants of the atom is in a ring is removed since this analysis is restricted to linear-structures. Unless otherwise stated, 200 dimensions are used, and both radius 0 and radius 1 are used as described in section 4.3.

4.2 Vector descriptor generation by Mol2Vec

4.2.1 Substructure vector descriptor calculation

Vector representations of molecules provide a straightforward route to training machine learning models. The Mol2Vec algorithm is an unsupervised approach for converting molecular substructures to vectors, but it requires a large “corpus” of molecular structures to establish context for molecular groups [73]. For this study, the corpus of training structures is created through a recursive bond breaking algorithm, which starts from 15 common biomass molecules and generates 91135 unique molecular substructures and 13422 corresponding elementary steps (details in chapter 2). These substructures provide a corpus of molecular structures that are commonly observed in biomass chemistry, and are used to train the unsupervised Mol2Vec algorithm.

The Mol2Vec algorithm [73] uses a shallow neural network to learn the joint probability distribution of the occurrence of different types of groups within a substructure, similar to the widely used Word2Vec algorithm from natural language processing [103]. One hyper-parameter of the Mol2Vec algorithm is the radius that defines the group size, where every heavy atom in a substructure is considered as a ”center”, and the radius determines the number of centers considered in a group. Vectors for the substructure are obtained by summing up vectors representing the groups that are generated by Mol2Vec (see Figure 4.2b). The dimension of these vectors is determined by a second hyper-parameter of the Mol2Vec algorithm which determines the number of nodes used in the shallow neural network, and equivalently the dimension of the resulting vector. The vectors generally have a large number of dimensions (10-1000), and are hence difficult to visualize. Principal component analysis (PCA) [104] is used to reduce the vectors to 2 dimensions for the convenience of visualization and analysis. For example, propionic acid (CH_3CH_2COOH) has 5 heavy atoms, corresponding to 5 unique radius 0 groups (CH_3, CH_2, C , O and OH), and 5 unique radius 1 groups (CH_3CH_2, CH_3CH_2C, CH_2COOH, C=O and COH), as illustrated in Figure 4.2. Figure 4.2a shows the first 2 principal components of one-heavy-atom R0 groups and Figure 4.2b shows multi-heavy-atom R1 groups with the center atom circled with the same color as the group. Each circled group is analogous to a “word” in Word2Vec, the propionic acid substructure is analogous to a “sentence” in Word2Vec and the whole training set of substructures is analogous to a “paragraph” or “corpus” in Word2Vec.

Ultimately, the vectors for groups and substructures must be combined to establish a vector representation for particular molecules of interest. The overall radius 0 vector representing the CH_3CH_2COOH molecule is calculated as:

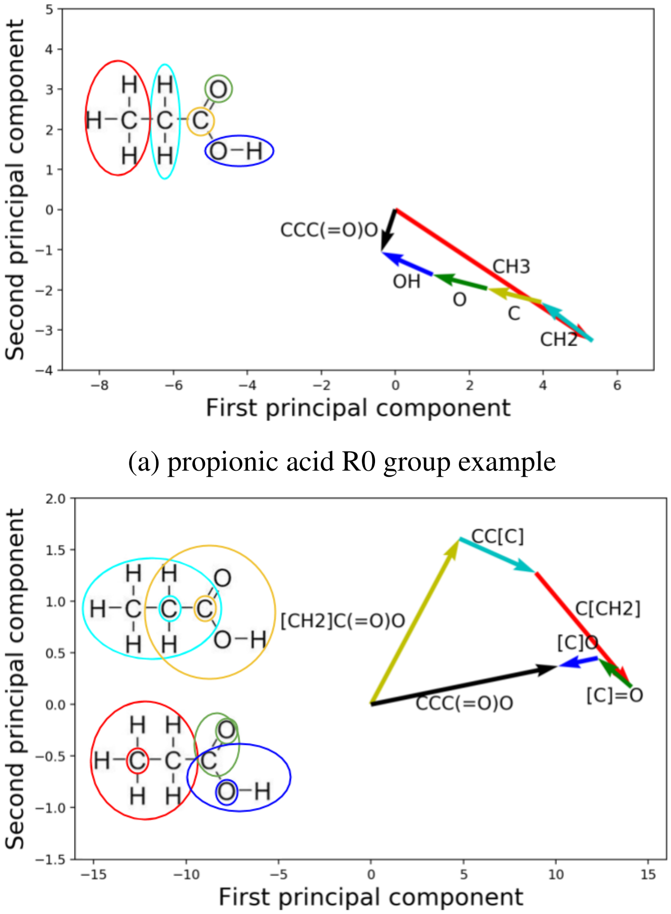
$$V_{CH_3CH_2COOH} = V_{CH_3} + V_{CH_2} + V_C + V_O + V_{OH} \tag{4.2}$$

$$\begin{aligned} V_{CH_3CH_2COOH} = V_{CH_3CH_2} + V_{CH_3CH_2C} \\ + V_{CH_2COOH} + V_{C=O} + V_{COH} \end{aligned} \tag{4.3}$$

where V_i is the vector for group i. Similarly, glycerol contains 2 CH_2, 3 OH, 1 CH at radius 0 and 2 CH_2OHCH, 2 CH_2OH, 1 CHOH and 1 $CH_2CHOHCH_2$ at radius 1. By considering the different vectors of each group contained in a substructure, Mol2Vec vectors provide a route to quantify the difference between different types of molecules. For example, the distance between vectors should be able to measure the similarity of molecules. Figure 4.3a shows the relative distance between vectors representing glycerol and propionic acid with different hyper-parameters. The vectors representing glycerol are aligned to the unit vector along the x-axis and the vectors representing propionic acid are projected using the projection matrix obtained from the same hyperparameter combination. A visual analysis reveals that the vectors change between different molecules indicating that the algorithm is able to distinguish between these molecules. However, the relative orientation of the vectors changes as the radius and dimension hyper-parameters are varied. The results suggest that a convergence criterion is needed to identify the optimum choice of hyper-parameters (discussed in section 4.3).

4.2.2 Reaction vector descriptor calculation

Biomass reaction networks are composed of elementary steps that connect various intermediates species. Vector representations of these reactions can enable data-driven classification of reaction types, or predictions of reaction energies. Reaction vectors are computed



(a) propionic acid R0 group example

(b) propionic acid R1 group example

Figure 4.2: Illustration of propionic acid groups in R0 (a) and R1 (b) with 50-dimension descriptors. The substructure vector is black, and group vectors are color-coded according to the heavy atom center (small circles) and group (large circles).

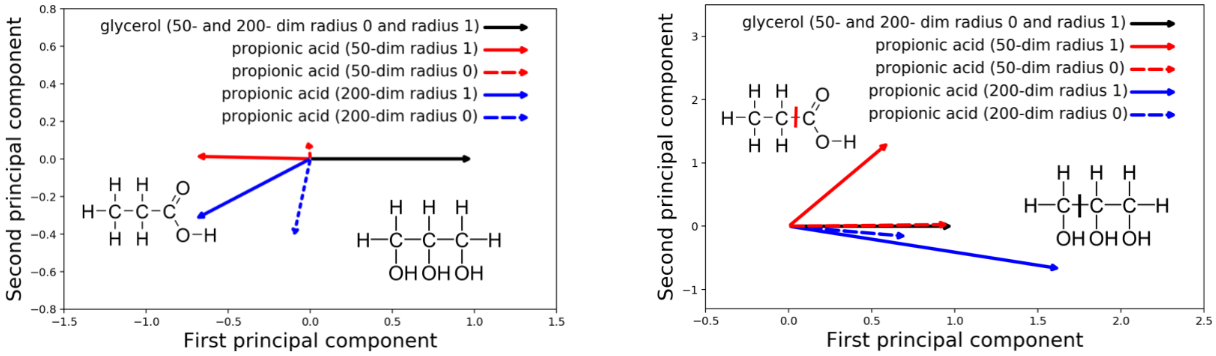
from substructure vectors as:

$$V_{reaction} = \sum V_{product} - \sum V_{reactant} \tag{4.4}$$

where $V_{reaction}$ is the reaction vector descriptors and $V_{product}$ and $V_{reactant}$ are the vector descriptors for products and reactants.In this work, we define reaction vectors as the sum of the vectors representing the product substructures minus the sum of the vectors representing reactant substructures (equation Equation 4.4). For example, one of the C-C bond breaking reactions for propionic acid can be calculated as:

$$V_{reaction} = V_{CH_3CH_2} + V_{COOH} - V_{CH_3CH_2COOH} \tag{4.5}$$

where V_i is the vector for substructure i (computed following equation Equation 4.2 or equation Equation 4.3 for radius 0 or radius 1 respectively) and $V_{reaction}$ is the vector of reaction.An alternative interpretation is that this vector represents the chemical C-C bond, and the direction of the vector will determine if the bond is being broken or formed. Similar to substructure vectors, it is expected that the reaction vector should have the ability of measuring similarity in bond breaking reactions. Figure 4.3b shows PCA projections of examples of C-C bond breaking reaction vectors for glycerol and propionic acid with varied hyper-parameters (vector dimension and radius). Similar to the case of substructures, the length and direction of vectors are not robust to the change of radius and dimension, which means not only the vector representation itself changes but the relative position between two vectors also changes. Moreover, in this case the vectors are not able to distinguish between the reactions at 50-dimensions with radius 0. This indicates that cancellation of error does not occur, and emphasizes the need for converging the dimension and radius (discussed in section 4.3).



(a) PCA decomposed 2-dim vector of glycerol (normalized on x-axis) with and propionic acid with 50-, 200- dim and radius 0, 1

(b) PCA decomposed 2-dim vector of CC bond breaking in glycerol and propionic acid with 50-, 200- dim and radius 0, 1

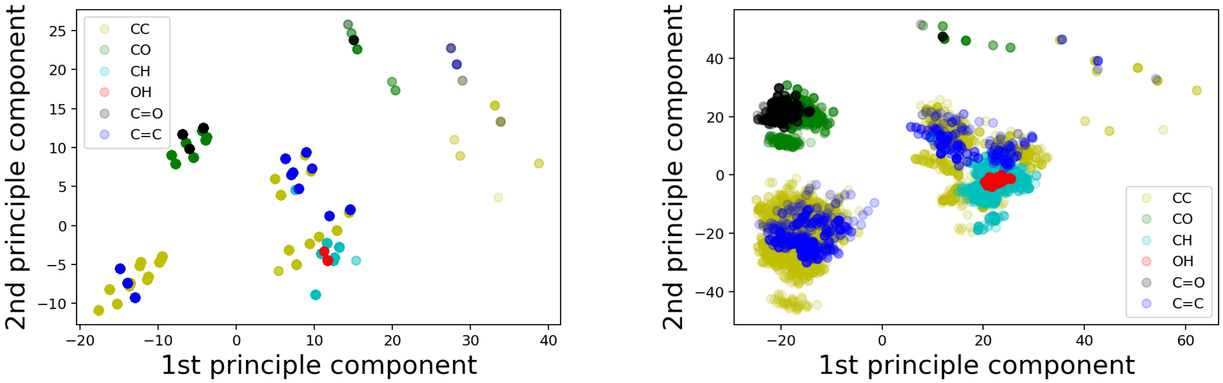
Figure 4.3: PCA decomposed 2-dim vector of (a) glycerol and propionic acid, (b) C-C bond breaking reactions in glycerol and propionic acid with different size and radii

4.3 Classification of gas-phase bond-breaking reactions

4.3.1 unsupervised feature generation

In addition to comparing different reaction/substructure vectors individually, it is useful to compare all reactions/substructures to each other simultaneously to identify different classes of reactions or molecular substructures. This is facilitated by visualizing reaction vectors as points defined by their corresponding vector. Figure 4.4 shows the vector point data of the bond breaking reaction sample set (13422 reactions in total) in radius 0 and radius 1 with 200 dimensions projected onto 2 dimensions using PCA. There are clear discrete clusters in radius 0, while radius 1 consists of a more continuous distribution with significant scatter. The clusters in radius 0 are separated based on the number of hydrogen and heavy atoms around a given atomic center. The color indicates that these clusters generally follow the classification expected based on the nature of each atom center in the bond (e.g. C-C, C-O, C=C, C=O, C-H, O-H). For example, glycerol C-C bond breaking reactions (C(C(CO)O)O $\rightarrow$ O[CH]CO+[CH2]O in SMILES notation) and xylitol edge C-C bond breaking reactions (C(C(C(C(CO)O)O)O)O $\rightarrow$ C(O)C(O)C(O)[CH]O+[CH2]O in SMILES notation) appear in the same cluster located at [-10.53, 0.92]. In contrast, the results for radius 1 do not exhibit clear separation between clusters or bond type in the PCA projection.

The clusters are quantitatively identified using a mean-shift clustering algorithm in the original 200-dimensional space. This results in 70 clusters for radius 0, and more than 4500 clusters for radius 1. We assess the quality of these clusters by computing the average silhouette score, which is $>$0.99 for radius 0 and 0.82 for radius 1. The poor clustering in radius 1 is consistent with the low-dimensional visualization shown in Figure 4.4. Moreover, the quality of the clusters is assessed by assigning a bond type to each cluster based on the mode of bond types in that cluster. This model is then assessed by computing the accuracy toward assigning/predicting bond type, as shown in Table. Table 5.1. The results confirm that the R0 clusters are better able to predict the bond type, although there are some mis-classifications for C-C/C=C, C-O/C=O. The single/double bond mis-classifications may be due to conjugated bonds, which are often assigned as single/double in SMILES notation by the distance between the two atoms. The bond order is determined by the distance between two atoms within the original molecule structure obtained from the PubChem database. We use the combination of the two classifications to create a new set of 90 reaction classes, where 20 of the 70 clusters are split based on single/double bonds. These 90 bond types represent a richer classification than the simple A-B type bond classification, and present a data-driven route to automatically identify different types of elementary steps in biomass reaction networks. We note that this scheme is restricted to reactions that only involve a single bond, and concerted reactions such as retro-aldol condensation and Grob fragmentation are not included. However, the framework could be extended to include concerted reactions by summing the reaction vectors for the bonds involved.



(a) 1st vs. 2nd principal components in radius 0

(b) 1st vs. 2nd principal components in radius 1

Figure 4.4: PCA decomposition to 2 dimensions of the original 200-dimension reaction vectors in radius 0 and radius 1

Table 4.3: 70 cluster with atomic environment (number of hydrogen atom and heavy atom surrounding) Note: reaction type with numbers contains both type of reactions in the same cluster

n_cluster	n_hydrogen_0	n_heavy_atom_0	n_hydrogen_1	n_heavy_atom_1	label
1	2	3	1	4	C-C
2	2	3	0	4	C-C
3	0	4	0	3	C-C
4	1	4	1	3	C-C
5	1	3	0	3	C-C 433 (C=C) 247
6	0	4	1	3	C-C
7	2	3	1	3	C-C
8	1	4	0	3	C-C
9	0	3	2	3	C-C
10	1	3	1	3	C-C 176 (C=C) 243
11	0	3	0	3	C-C 204 (C=C) 47
12	0	4	0	4	C-C
13	2	3	2	3	C-C
14	1	4	0	4	C-C
15	1	4	1	4	C-C
16	0	2	1	4	C-C
17	0	2	1	3	C-C 80 (C=C) 36
18	0	2	2	3	C-C
19	0	2	0	4	C-C 26 (C=C) 9
20	0	2	0	3	C-C 51 (C=C) 14
21	0	2	0	2	C-C 3 (C=C) 1
22	1	2	2	2	C-C

Continued on next page

Table 4.3 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
23	2	2	1	4	C-C
24	2	2	0	3	C-C 56 (C=C) 18
25	0	4	2	2	C-C 55 (C=C) 67
26	2	2	1	3	C-C
27	2	2	2	3	C-C
28	0	2	2	2	C-C 7 (C=C) 1
29	2	2	2	2	C-C
30	0	2	1	2	C-C 5 (C=C) 4
31	1	2	1	4	C-C
32	1	2	0	4	C-C 34 (C=C) 33
33	1	2	1	3	C-C 57 (C=C) 66
34	1	2	0	3	C-C 52 (C=C) 44
35	1	2	2	3	C-C
36	1	2	1	2	C-C 5 (C=C) 3
37	3	2	0	4	C-C
38	3	2	0	3	C-C
39	3	2	2	3	C-C
40	1	3	3	2	C-C
41	3	2	0	2	C-C
42	2	2	3	2	C-C
43	1	2	3	2	C-C
44	1	4	1	2	C-O
45	1	3	1	2	C-O
46	0	4	1	2	C-O

Continued on next page

Table 4.3 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
47	0	3	1	2	C-O
48	2	3	1	2	C-O
49	2	2	1	2	C-O
50	0	2	1	2	C-O
51	2	2	0	2	C-O
52	1	3	0	2	C-O 41 (C=O) 20
53	1	4	0	2	C-O
54	0	3	0	2	C-O 162 (C=O) 320
55	0	4	0	2	C-O 313 (C=O) 823
56	2	3	0	2	C-O
57	1	2	1	2	C-O
58	1	2	0	2	C-O 2 (C=O) 1
59	0	2	0	2	C-O 9 (C=O) 8
60	3	1	1	1	C-H
61	3	2	1	1	C-H
62	1	1	1	1	C-H
63	1	4	1	1	C-H
64	1	3	1	1	C-H
65	1	2	1	1	C-H
66	2	2	1	1	C-H
67	2	3	1	1	C-H
68	2	1	1	1	C-H
69	1	2	1	1	O-H
70	1	1	1	1	O-H

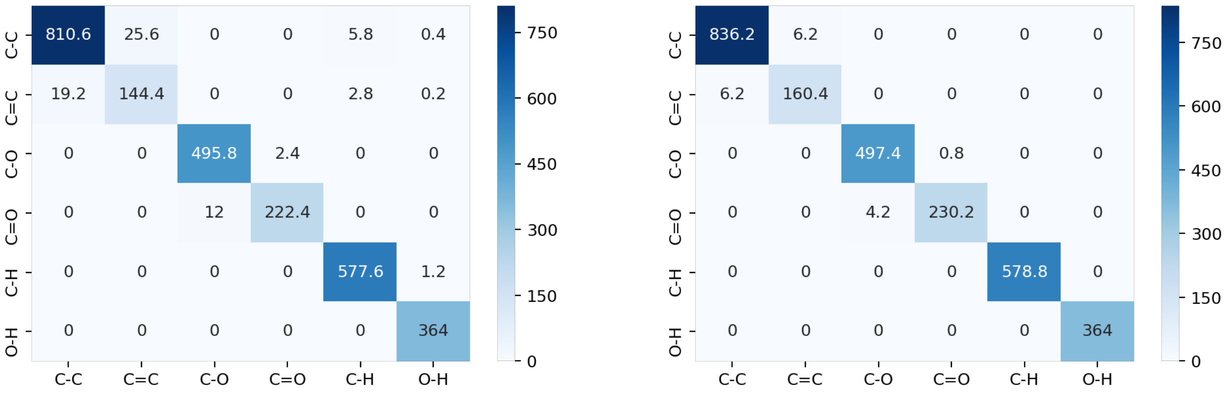
In summary, the results of unsupervised data analysis algorithms have led to the following conclusions: (1) the relationship between different substructures or reaction vectors (e.g. propionic acid and glycerol, see Figure 4.3) is sensitive to the hyper-parameters used by Mol2Vec, and (2) the radius 0 vectors can be used to robustly identify 70 sub-classes of elementary steps, which can be combined with standard bond types to generate 90 total classes of elementary biomass reactions. The first finding leads to the question of how to optimize hyper-parameters, which is addressed by using the results of the second finding along with supervised classification in the subsequent section to arrive at a compact and robust vector representation of biomass molecules and reactions.

4.3.2 Supervised feature generation

The results of unsupervised dimension reduction through PCA indicate that reaction classes are not well-separated in radius 1, and that convergence of Mol2Vec hyperparameters is difficult without an accuracy metric. These shortcomings can be overcome through supervised classification. Specifically, linear discriminant analysis (LDA) is used to converge and estimate the optimal dimension and radius based on classification accuracy of 6 A-B (C-C, C-H, C-O, O-H, C=C, C=O) type of reactions. Figure 4.7a shows the 5-fold prediction accuracy (with error bars from 5-fold cross validation) from 10- to 400-dimension Mol2Vec vectors with radius 0, 1 and 2. Based on the 5-fold cross validation, R0 accuracy is highly sensitive to the data set split for training/testing as the error bars are much larger than R1 and R2. R1 accuracy reaches the highest average of 0.97 at 200-dimension and decreases slightly at 400-dimension. R2 follows the same trend with R1, but the average accuracy is lower than R1 before 200-dimension and only slightly higher than R1 at 400-dimension. The 200-dimension vectors at R1 are selected as giving the best tradeoff between accuracy and complexity, and are used for all subsequent analysis.

Distinguishing different reactions is a key test of the vector representation. From the above results, 200-dimension is used for the analysis and identification of A-B type reactions, resulting in an accuracy of 0.97. However, 200-dimension vectors are difficult to conceptualize and computationally inefficient with respect to both the classification itself and subsequent energy regression. Thus, supervised dimension reduction is used to find a reduced dimensional representation that is capable of capturing basic reaction types. Specifically, LDA provides a convenient route to dimensional reduction, since the classification naturally reduces the dimension from 200 to 5 since the total number of reaction types is 6. Figure 4.5a shows the 5-fold average confusion matrix of ground-truth (row)/prediction (column) at radius 1. The train/test ratios are 80%/20% in both Fig. 6a and 6b and test sets are randomly selected. The process is repeated 5 times. The major mis-classification in radius 1 is in C-C/C=C, while there are also some mis-classifications between C-O/C=O, C-C/C-H, C=C/C-H, and C-H/O-H. The C-C/C=C and C-O/C=O mis-classifications are chemically intuitive because the conjugation of electrons can make the boundary between double and single bonds unclear. For example, reaction [O]C[C][CH]O → [C]C[O] + [CH]O / [C][C]=CC=CC → [CH]C=CC + [C][C] and reaction [CH][O] → [CH] + [O] / [CH]=O → [CH] + [O] are in the same classes according to LDA results since the two C atoms in the C-C (C=C) bond breaking reactions have the same atomic environment (2 heavy atoms, 0 hydrogen and 2 heavy atoms, 1 hydrogen). The bond breaking type is not clearly separated here since the free electrons and bonds conjugate in these radicals and similarly with C, O bond breaking reactions. However, the mis-classifications between other bond types is more problematic, and reflects an inability of the model to linearly separate these bond classes in a 5-dimensional space.

As discussed in subsection 4.3.1, the reactions separate into 70 distinct clusters in R0 space (Fig. Figure 4.4a), and can be further classified into 90 distinct reaction classes by labeling single/double bonds within a given cluster. However, the R1 vectors do not exhibit this clustering in the reduced PCA space, likely due to loss of variance related to the information contained in the original high dimension on reduction to 2 dimensions. We address this by using the 90 reaction classes identified in R0 as inputs to LDA, enabling

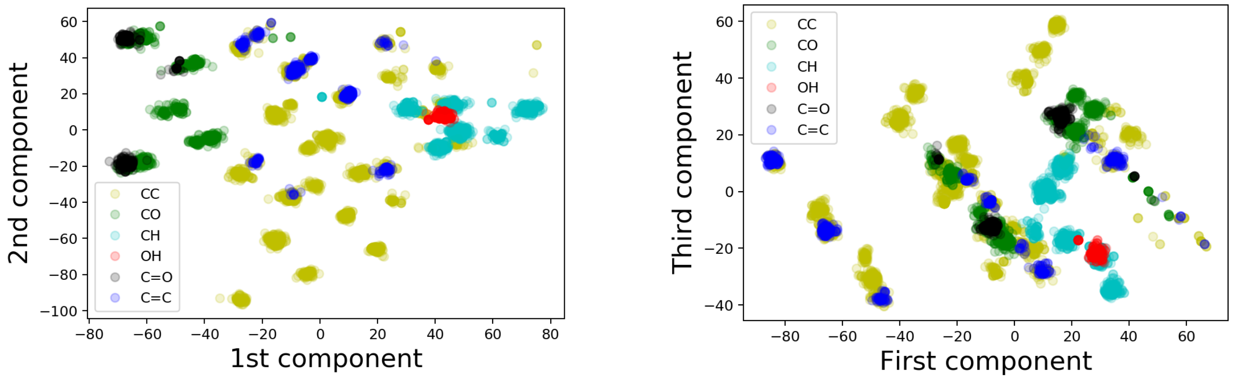


(a) A-B type radius 1 confusion matrix

(b) 90 classes radius 1 confusion matrix

Figure 4.5: Averaged confusion matrix for 100 times random train-test split of 6 A-B type reactions (C-C, C=C, C-O, C=O, C-H, O-H) in radius 1 and of 5 classes and 90 classes based on radius 0 clusters in radius 1

classification in up to 90 dimensions. Figure 4.6a shows a visualization of the first 2-dimensions of the LDA vectors labelled by the 6 A-B type reaction classes. Separation among different classes is much clearer than in the case of PCA (Figure 4.4b), indicating a more chemically-meaningful dimensional reduction. The accuracy of the classification is also promising, with an overall accuracy of 0.99 when classifying reactions between the 90 R0 classes. Figure 4.7b shows the classification accuracy vs. number of LDA dimensions from 5 to 89, showing that a 15-dimensional space achieves an accuracy of 0.99. Moreover, this higher-dimensional space can be used to separate reactions based on the simpler A-B type bonds. Figure 4.5b shows the confusion matrix of 6 A-B type reactions based on the classification into the 90 more specific classes in the 15-dimensional space. The accuracy is increased to 0.99 as compared to an accuracy of 0.97 from direct classification on the 6 A-B classes (see Figure 4.5a). This improvement in accuracy is attributed to the higher dimension of the sub-space. The mis-classifications are also improved, with the only confusion occurring between C-C/C=C and C-O/C=O bonds, which is unavoidable due to the single/double bonding in conjugated systems, as previously discussed.



(a) LDA classification 2D plot of 1st vs. 2nd in radius 1 with 90 classes

(b) LDA classification 2D plot of 1st vs. 3rd components in radius 1 with 90 classes

Figure 4.6: LDA classification results in radius 1 with 90 classes of (a) 1st and 2nd LDA components with and (b) 1st and 3rd LDA components

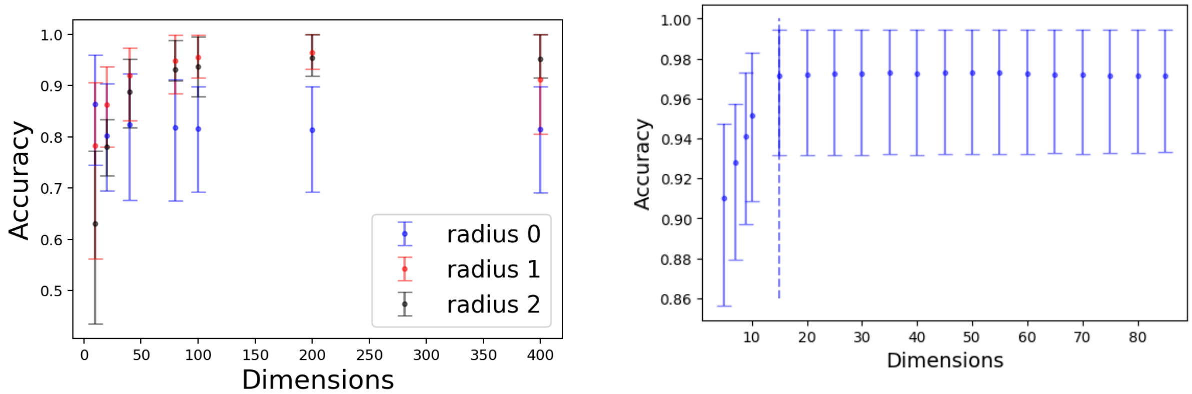
4.4 Feature validation via linear regression

The prior sections describe generation of molecular feature vectors obtained by applying the unsupervised Mol2Vec algorithm followed by supervised dimensional reduction based on classification of different bonding types. This approach is powerful because it is based solely on molecular structure, and does not require any first-principles calculations. Therefore, the ability of these vectors to predict the results of first-principles calculations is an independent validation of their description of key molecular properties. In this section, the energies of 189 substructures are computed with DFT and used to determine reaction energies for 1117 bond breaking reactions. The features computed from LDA are used as inputs to predict the reaction energies, and the results are compared with the widely-used group additivity approach. The 15 single-heavy-atom groups are used for group additivity, and more information is available in Table 4.4.

Reaction energies are predicted using simple linear regression with the LDA feature vectors for each reaction as input:

$$D\vec{x} = \vec{p} \tag{4.6}$$

where D is a $m \times n$ matrix for m reactions with $n =$ 5- to 89-dimensional features ex-



(a) 10 dimensions to 400 dimensions were implemented with R0, R1, R2 and 200 dimensions is the lowest dimension that reaches a high accuracy of 0.96 (prediction accuracy) in R1

(b) prediction accuracy of 5 dimensions to 85 dimensions and 15 dimension as the lowest dimension that reaches a high accuracy of 0.97

Figure 4.7: Accuracy of (a) 6-type classification of reactions with vector descriptors of different radius and dimensions and (b) 90-class classification of reactions with radius 1 and different dimensions of LDA projected original 200-dim vectors

tracted from LDA, $\vec{x}$ are the weights determined by linear regression, and $\vec{p}$ is the property vector.

Regression analysis for gas-phase energies is also performed using both Benson's original group additivity approach [72] and the regression to the vector-based descriptors introduced in this work (details in section 4.4). Comparisons are made using identical training and testing sets, with training sets chosen to ensure that examples of all groups defined by group additivity are present. In the case of group additivity, substructure energies are predicted and reaction energies are computed using Equation 5.1, while for the vector-based regression introduced here the reaction energies are directly predicted. By comparison, the group additivity method is defined as:

$$G\vec{c} = \vec{p} \tag{4.7}$$

where G is a $j \times k$ matrix for j molecules with k group types, $\vec{c}$ are the contribution of each group, and $\vec{p}$ are substructure energies. Reaction energies are computed by subtracting the predicted energies of each substructure.In this case, 15 group types are selected based on all single-heavy-atom groups present in the molecular structures [72]. Group additivity can also be extended by selecting more groups, but moving beyond single-heavy-atom groups requires selecting appropriate groups and is beyond the scope of this work. Here, we select the simplest version of group additivity, and control for complexity by comparing it to a machine-learned model with the same number of linearly-fitted parameters.

Table 4.4: Group Additivity Groups

Hydrogen count	Heavy count	SMILES (X for heavy atom)
0	2	[O]X
0	3	[C](X)X
1	2	OX
1	2	[CH]X
1	3	[CH](X)X
0	4	[C](X)(X)X
0	2	[C]X
0	1	[C]
1	1	[CH]
1	1	[OH]
0	1	[O]
2	3	[CH2](X)X
2	2	[CH2]X
1	4	[CH](X)(X)X
2	1	[CH2]

The results are shown in Figure 4.8. As the number of dimensions increases from 5 to 89, the mean/max absolute error goes from 1.59 eV/5.1 eV down to 0.59 eV/2.9 eV. By comparison, the group additivity model (shown in blue/red cross sign for mean/max absolute error) has a larger mean/max error (3.1 eV/5.9 eV) than the machine-learned model at 15-dimensions. Moreover, the machine-learning model can be systematically extended by increasing the number of dimensions. In contrast to extending the group additivity models, this does not require any arbitrary selection of which groups to include or not include, providing a systematic way to increase model complexity. Furthermore, the results of the cross-validation suggest that the model does not suffer from significant over-fitting, even at large numbers of dimensions. However, the variance of the max error does increase beyond 80 dimensions suggesting that the results become sensitive to the selection of training/testing data beyond 80 dimensions and over-fiting may begin to become an issue.

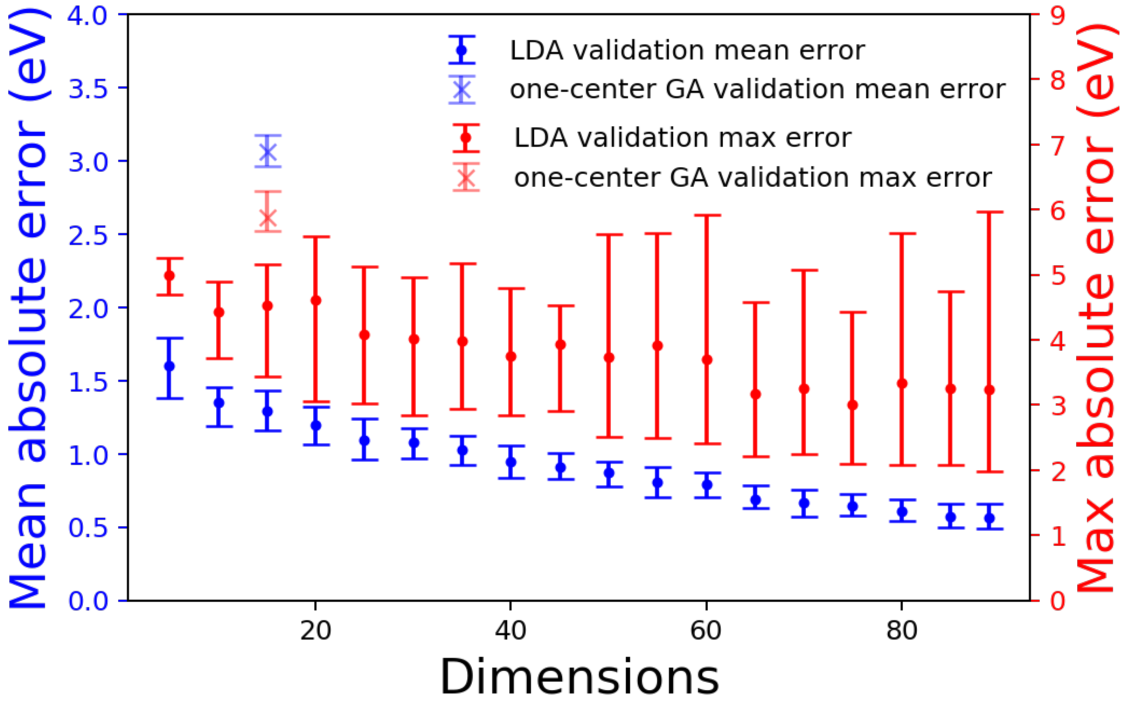


Figure 4.8: Linear regression maximum and mean validation error on LDA components vs. number of dimensions compared to one-center group additivity results

4.5 Model compound assessment

4.5.1 Quantitative assessment of model compound illustration

As mentioned in chapter 1, glycerol is often used as a model compound to study the reaction of more complex and larger chemicals (e.g. glucose, xylitol, sorbitol), which contain more than 10 carbon or oxygen atoms. Although the use of glycerol as a model compound may seem obvious based on chemical intuition, there is no quantitative way of assessing other model compounds or evaluating the efficacy of glycerol as a model compound for larger biomass molecules. In this section we assess model compounds with up to 4 carbons and 4 oxygen atoms (8 total heavy atoms) using comparisons elementary steps and the descriptors described in chapter 2.

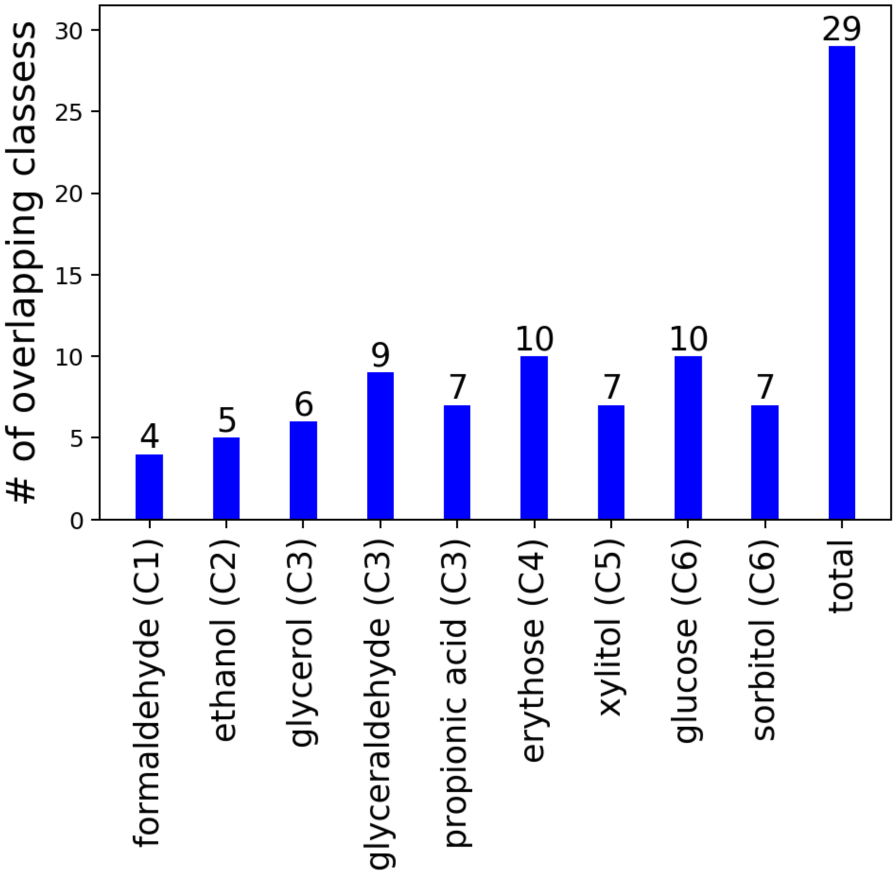


Figure 4.9: number of classes model compounds represent in 1st recursion bond breaking with 29 classes in total

One way of evaluating the efficacy of a model compound is by comparing the types of elementary steps (or chemical bonds) between the model compound and the compound of interest. The chemical bonds in a given compound are described by the first recursion of the bond breaking algorithm (see Figure 2.1), and the types of these bonds can be classified using the 90 classes determined from analysis of the R0 reaction vectors (see Fig. Figure 4.4a). Fig. Figure 4.9 shows the number of reaction types (bond types) present for various biomass molecules. This contains only the bonds present in the compound (i.e. bonds present in substructures are not included), and the results indicate that only 29 of the 90 total classes are present in the compounds studied (see chapter 2). As molecules become larger, the number of reaction types generally increases. However, there is some variation within molecules with the same size, and the number of reaction types does not increase beyond C5. This suggests that the approach of using model compounds is generally an efficient route to reduce complexity without reducing the types of chemistry present. For example, the number of reaction types in erythrose is equal to the number of reaction types in glucose, and in fact the reaction types are the same, suggesting that ethyrose is an excellent model compound for glucose. Moreover, glyceraldehyde contains 9 different reaction types, indicating that it is also an excellent model compound for studying general biomass chemistry. While this analysis only considers single-bond reactions, it could be extended to identify compounds that exhibit concerted reactions by examining the first two recursions and comparing the sum of these vectors to important concerted reactions like retro-aldol condensation or Grob fragmentation. This comparison of reaction types provides insights into which compounds contain the most diverse types of bonds, but comparison between specific model compounds and larger target compounds is less intuitive (i.e. just because glyceraldehyde contains the most bonds doesn't mean it is the most appropriate model compound for all larger molecules).

4.5.2 Example of model compound

An alternative approach is to directly compare the similarity of different molecules. This can be achieved using the vectors obtained from Mol2Vec, and computing the Euclidean distance between different compounds in the 200-dimensional space. Figure 4.10 shows the distance between some common compounds in the 200-dimensional vector space. The results provide a route to identify a small model compound that is most chemically-similar to a larger compound of interest, and also enable validation of the chemical vectors against chemical intuition. Glyceraldehyde/glycerol and glucose/sorbitol reactions are two of the most widely studied compounds in previous research, which has shown similar mechanisms between the two reactions: both are absorbed onto the transition metal surface via =O, and the rate between the adsorbed species and the dissociated hydrogen is rate determining. [24, 116] So as an example, we consider the reaction for conversion of glucose to sorbitol. Figure 4.10 shows the Euclidean distance between possible model compounds and sorbitol/glucose. The C1 and C2 compounds are the farthest away, which is consistent with the fact that many of the bond types on sorbitol/glucose are not present in these small molecules. Of the C3 compounds considered, it is clear that glycerol is closest to sorbitol while glyceraldehyde is closest to glucose, and propionic acid is relatively far away in comparison. This is consistent with chemical intuition, since glyceraldehyde contains an aldehyde group, similar to glucose, and glycerol contains alcohol groups similar to sorbitol. In contrast, propionic acid contains an acid group that is not present in either sorbitol or glucose, and is hence farther away. If C4 and C5 compounds are considered, erythrose is found to be very close to glucose, while xylitol is very close to sorbitol, also consistent with chemical intuition. There is also evidence that the bond cleavage of erythrose to glyceraldehyde and of glucsoe to arabinose follows the same pathway.[117] A full distance matrix between all compounds studied in this work is available in the Figure 4.10. This comparison may assist with the identification of model compounds for studying arbitrary biomass conversion reactions.

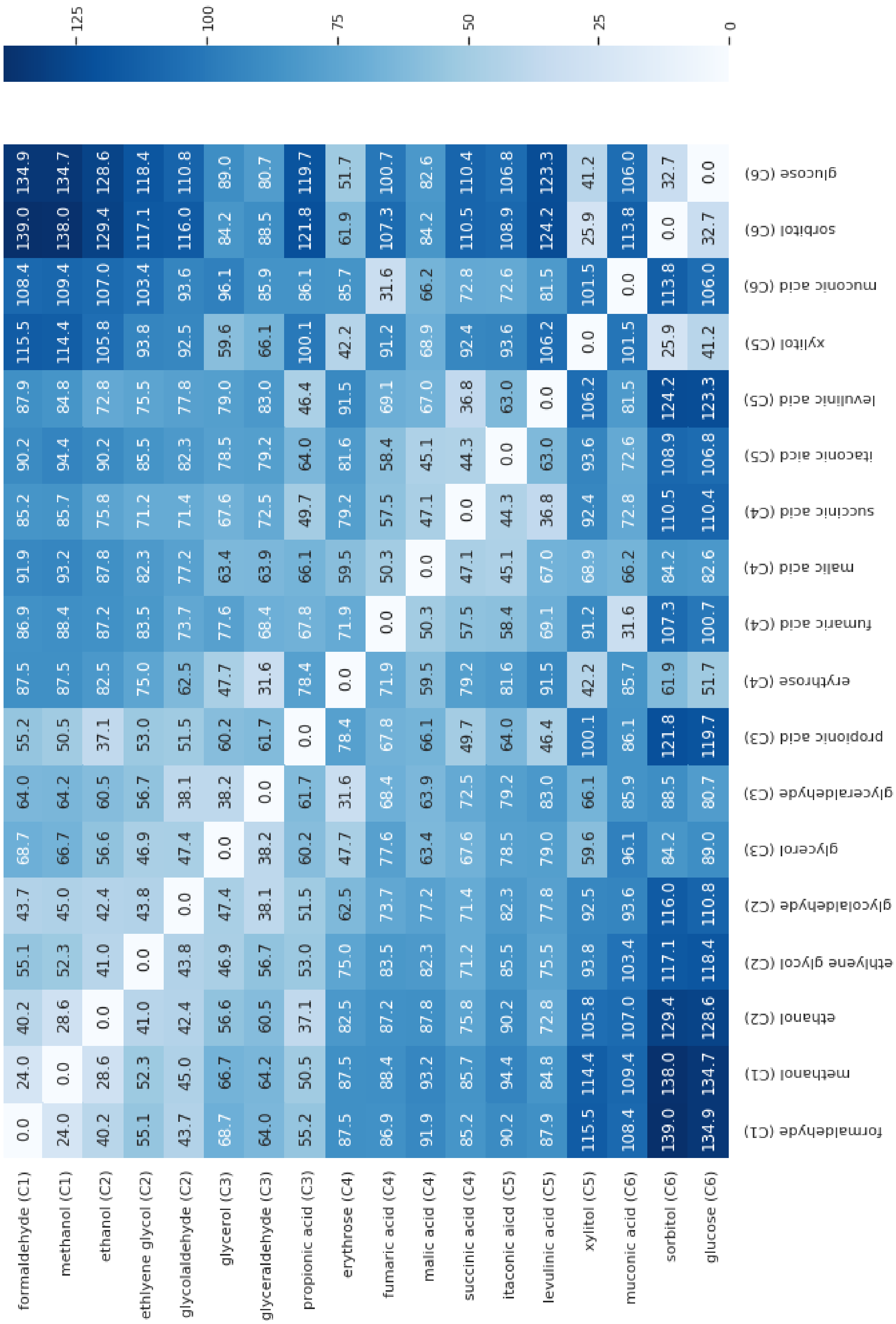
4.6 Conclusion and future work

In this section, we investigated a new vector-based descriptor for complex biomass molecules. The descriptor is able to accurately classify the 6 general types (C-C, C-O, C-H, O-H, C=C, C=O) of biomass reactions, provide a new classification scheme encompassing 90 types of bonds in biomass compounds, quantitatively assess the similarity of molecules commonly used as model compounds, and predict the gas phase reaction energy. The validation accuracy of classification using the vector descriptor reached 0.97, and the mean absolute error for linear regressed projected vector could reach a minimum of 0.59 eV with 85 dimensions. Moreover, we provide a quantitative method to analyse simple model compounds for complex compounds/reactions that is consistent with chemical intuition.

The vector descriptors are generated by first using unsupervised training of biomass substructures/molecules in Mol2Vec with 200-dimensions with both radius 0 and radius 1 used to control the locality of the resulting descriptors. The radius 0 vectors lead to 70 discrete clusters, which are used along with simple bond types to arrive at 89 sub-classes of bond breaking reactions. Linear discriminant analysis (LDA) is used along with these 90 classes to create a 15-dimensional representation of each compound capable of classifying the 6 basic bond types with and accuracy of 0.97 and the 90 more complex bond types with an accuracy of 0.99. The errors in classification were found to primarily arise due to single/double bonds in conjugated systems, suggesting that the classification is consistent with chemical intuition.

The vector descriptors from this combination of Mol2Vec and LDA are further used as inputs to regression models capable of predicting reaction energies for gas-phase reactions with mean absolute errors as low as 0.59 eV, compared to errors of 3.06 eV for the commonly-used group additivity approach. The vector representations are also used to assess similarities between different commonly-studied biomass compounds, and it was found that the Euclidean distance between these vectors provides a good metric of compound similarity that is consistent with chemical intuition.

The ability to construct vector descriptors for complex chemical compounds such as biomass molecules and intermediates provides a new route to constructing machine learning models for complex chemical systems. The approach developed here can be directly applied to predict properties of biomass compounds, or can be applied to new types of chemical compounds like hydrocarbons or biomolecules. The approach utilizes unsupervised learning, making it especially advantageous for situations where limited data is available. This may improve the ability to develop data-driven models of chemical properties in the future. The future energies for reactions or adsorptions on transition metal surface species can also help with prediction of the energies and further obtaining the probability of intermediates in recursions.



	formaldehyde (C1)	methanol (C1)	ethanol (C2)	ethylene glycol (C2)	glycolaldehyde (C2)	glycerol (C3)	glyceraldehyde (C3)	propionic acid (C3)	erythrose (C4)	fumaric acid (C4)	malic acid (C4)	succinic acid (C4)	itaconic aicd (C5)	levulinic acid (C5)	xylitol (C5)	muconic acid (C6)	sorbitol (C6)	glucose (C6)
formaldehyde (C1)	0.0	24.0	40.2	55.1	43.7	68.7	64.0	55.2	87.5	86.9	91.9	85.2	90.2	87.9	115.5	108.4	139.0	134.9
methanol (C1)	24.0	0.0	28.6	52.3	45.0	66.7	64.2	50.5	87.5	88.4	93.2	85.7	94.4	84.8	114.4	109.4	138.0	134.7
ethanol (C2)	40.2	28.6	0.0	41.0	42.4	56.6	60.5	37.1	82.5	87.2	87.8	75.8	90.2	72.8	105.8	107.0	129.4	128.6
ethlyene glycol (C2)	55.1	52.3	41.0	0.0	43.8	46.9	56.7	53.0	75.0	83.5	82.3	71.2	85.5	75.5	93.8	103.4	117.1	118.4
glycolaldehyde (C2)	43.7	45.0	42.4	43.8	0.0	47.4	38.1	51.5	62.5	73.7	77.2	71.4	82.3	77.8	92.5	93.6	116.0	110.8
glycerol (C3)	68.7	66.7	56.6	46.9	47.4	0.0	38.2	60.2	47.7	77.6	63.4	67.6	78.5	79.0	59.6	96.1	84.2	89.0
glyceraldehyde (C3)	64.0	64.2	60.5	56.7	38.1	38.2	0.0	61.7	31.6	68.4	63.9	72.5	79.2	83.0	66.1	85.9	88.5	80.7
propionic acid (C3)	55.2	50.5	37.1	53.0	51.5	60.2	61.7	0.0	78.4	67.8	66.1	49.7	64.0	46.4	100.1	86.1	121.8	119.7
erythrose (C4)	87.5	87.5	82.5	75.0	62.5	47.7	31.6	78.4	0.0	71.9	59.5	79.2	81.6	91.5	42.2	85.7	61.9	51.7
fumaric acid (C4)	86.9	88.4	87.2	83.5	73.7	77.6	68.4	67.8	71.9	0.0	50.3	57.5	58.4	69.1	91.2	31.6	107.3	100.7
malic acid (C4)	91.9	93.2	87.8	82.3	77.2	63.4	63.9	66.1	59.5	50.3	0.0	47.1	45.1	67.0	68.9	66.2	84.2	82.6
succinic acid (C4)	85.2	85.7	75.8	71.2	71.4	67.6	72.5	49.7	79.2	57.5	47.1	0.0	44.3	36.8	92.4	72.8	110.5	110.4
itaconic aicd (C5)	90.2	94.4	90.2	85.5	82.3	78.5	79.2	64.0	81.6	58.4	45.1	44.3	0.0	63.0	93.6	72.6	108.9	106.8
levulinic acid (C5)	87.9	84.8	72.8	75.5	77.8	79.0	83.0	46.4	91.5	69.1	67.0	36.8	63.0	0.0	106.2	81.5	124.2	123.3
xylitol (C5)	115.5	114.4	105.8	93.8	92.5	59.6	66.1	100.1	42.2	91.2	68.9	92.4	93.6	106.2	0.0	101.5	25.9	41.2
muconic acid (C6)	108.4	109.4	107.0	103.4	93.6	96.1	85.9	86.1	85.7	31.6	66.2	72.8	72.6	81.5	101.5	0.0	113.8	106.0
sorbitol (C6)	139.0	138.0	129.4	117.1	116.0	84.2	88.5	121.8	61.9	107.3	84.2	110.5	108.9	124.2	25.9	113.8	0.0	32.7
glucose (C6)	134.9	134.7	128.6	118.4	110.8	89.0	80.7	119.7	51.7	100.7	82.6	110.4	106.8	123.3	41.2	106.0	32.7	0.0

Figure 4.10: Distance of vectors representing C1-C6 compounds

CHAPTER 5
COMBINING PHYSICAL AND MACHINE LEARNING MODELS TO IDENTIFY STABLE GEOMETRIES ON RH (111) SURFACE

5.1 Data generation and algorithms

5.1.1 Data generation for surface species

We establish an approach to assign detailed classes for each bond type of adsorbed biomass intermediates, similar to the previously-developed approach for gas-phase molecules chapter 4. Mol2Vec [73] is used for generating vector descriptors for intermediates and reactions with 200 dimensions in R0 and R1 chapter 4. The traditional extended-connectivity fingerprint (ECFP) is modified to include surrounding heavy atom number, surrounding hydrogen atom number, valence, electric negativity and mass as invariants for atoms contained in a structure. Metal atoms are still considered to be heavy atoms but all the other properties are considered as NaN so that metal atoms are a general rather than specific type (e.g. Rh is indistinguishable from any other metal). An algorithm (see Figure 2.2) to add metal atoms to unsaturated C, O atoms one-at-a-time is applied to the 171 intermediates generated from the first 2 bond-breaking recursions of erythrose, glyceraldehyde, glycerol and propionic acid, and additional structures are generated by DFTB minima hopping calculations, yielding a total of 2,498 adsorbed structures. An example of [CH2][O] (Figure 5.1) is taken here to illustrate how the adding metal algorithms actually work. [CH2][O] is the starting point and at the first iteration, C atom from [CH2] is found as the unsaturated atom and with 1 chemical bond unsaturated, and then 1 [Rh] is added to the C atom and the SMILES notation after adding [Rh] would be C([Rh])[O], which is the starting point of the second iteration. Similar to the first iteration, C([Rh])O[Rh] is obtained after the second iteration and the algorithm stops here since all atoms are saturated. The adsorbed structures are combined with 91,098 gas-phase species, resulting in a total of 93,569 structures that are used as the corpus for training the Mol2Vec model. We also utilize 6 basic types of reactions for visualization and checking intuition. These 6 types include 4 types of intra-adsorbate reactions, C-C (C=C and C-C), C-H, C-O (C=O and C-O), OH, and 2 types of elementary reactions with metal atoms, C-M (C-Metal), O-M (O-Metal). The single and double bond breaking reactions are considered as a single class since the distinction between them becomes ambiguous for adsorbed species due to partial bond orders and conjugation. A total of 13,422 gas-phase reactions from our previous work chapter 4 and 1,666 additional elementary surface reaction steps are included for the analysis of reactions.

[CH2][O] → C([Rh])[O] → C([Rh])O[Rh]

Figure 5.1: Generating algorithms of SMILES notation of adsorbate with metal atoms: an example with [CH2][O],[CH2][O] is the starting point and at the first iteration, C atom from [CH2] is found as the unsaturated atom and with 1 chemical bond unsaturated, and then 1 [Rh] is added to the C atom and the SMILES notation after adding [Rh] would be C([Rh])[O], which is the starting point of the second iteration. Similar to the first iteration, C([Rh])O[Rh] is obtained after the second iteration and the algorithm stops here since all atoms are saturated.

5.1.2 Physics-based methods

As mentioned in chapter 2, Hotbit package based on DFTB theory with minima-hopping is used for obtaining various pre-optimized local-minimas of binding geometries. Rh slab atoms have fixed positions and adsorbate bonds have fixed lengths [102] to generate local minima geometries for each adsorbate. For more accurate energy estimations, DFT calculations are performed using Quantum ESPRESSO [88].A Monkhorst-Pack k-point sampling[93] of $4\times4\times1$ and a planewave cutoff of 450 eV were used. All surface species were modeled using 3.8034 as lattice constant and vacuum of 10.0 Åwith periodic condition. A BFGS algorithm provided by Atomic Simulation Environment (ASE)[90] was applied to the geometry optimization until the maximum force was no more than 0.05 eV/Å. The adsorption energy is calculated as follow:

$$E_{adsorption} = E_{system} - E_{surface} - E_{adsorbate} \tag{5.1}$$

where $E_{adsorption}$ is the adsorption energy, E_{system} is the total energy of the adsorbate and the Rh slab, $E_{surface}$ is the energy of Rh slab and $E_{adsorbate}$ is the reference energy of adsorbate relative to CH_4, H_2O and H_2. Since DFT and DFTB do not use a common reference, it is necessary to align the energies based on the stochiometry of each adsorbate:

$$E_{DFT} = E_{DFTB} + \left(\sum_{i \in [C,H,O]} c_i * n_i \right) + \epsilon \tag{5.2}$$

where E_{DFT} is the DFT adsorption energy, E_{DFTB} is the DFTB adsorption energy, n_i is the number of C, H, O atoms in the adsorbate, c_i are fitted coefficients and ϵ is the residual error.

5.1.3 ML algorithms

Principal component analysis is used to reduce the dimension of the original vector descriptors for the convenience of visualization. We also apply the semi-supervised LDA on the classification of 83-type of reactions and the LDA projections are saved for further adsorption energy regression. We apply ordinary least squares (OLS) and PLS on the original Mol2Vec vector descriptors, and LDA projection on original vector descriptors followed by OLS with adsorption energies calculated from DFT to do the regression. In each case we utilize a 75/25 train/test split with 4 random repeats with the results from the 328 DFT calculations as the target. The `scikit-learn` package [107] is used for each algorithm. We use the Spearman's correlation coefficient from scipy package to evaluate the ability of DFTB and ML methods to correctly order the energies of different geometries for a given adsorbate.

5.2 Cluster and classification of elementary step reactions

For the convenience of visualization, PCA is used for reducing the 200-dimensional reaction vectors to 2 dimensions, and each type of reaction is represented by a different color (yellow for C-C, green for C-H, cyan for C-O, red for O-H, black for C-M, blue for O-M). Figure 5.2 shows an example of how the 6 different types of chemical bonds are defined. Figure 5.3a and Figure 5.3b show the visualization of R0 and R1 reaction vectors for the full corpus of intermediates and reactions.

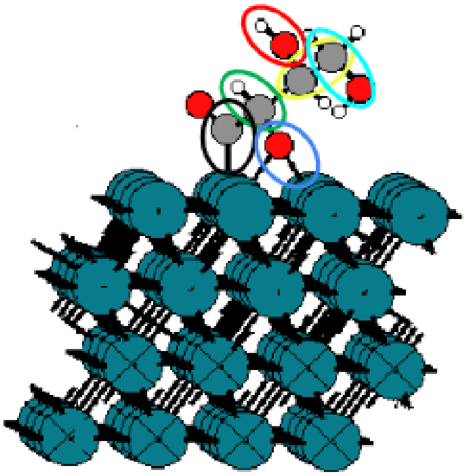


Figure 5.2: Example of different types of chemical bonds (yellow for C-C, green for C-H, cyan for C-O, red for O-H, black for C-M, blue for O-M)

The PCA result in R0 indicates that there are discrete well-defined clusters within the 15,088 reactions. The Euclidean distance between reactions of the original 200-dim R0 vectors are calculated to identify distinct clusters, and the cutoff to separate clusters is set to be 0.05. A total of 83 clusters in R0 are obtained. Each of the 83 clusters contain reactions with different bond-breaking types and different atomic environments of the the atoms within the elementary reaction. The atomic environment refers to the surrounding heavy atoms and hydrogen atoms of the 2 reacting atoms. Table Table 5.1 shows all details of the atomic environments for each of these reaction types.

5.2.1 Reaction types identified from vector clustering

Table 5.1: 83 clusters from R0 vectors with atomic environment (# of hydrogen atom and heavy atom surrounding)

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
1	2	3	1	4	C-C
2	2	3	0	4	C-C
3	0	4	0	3	C-C
4	1	4	1	3	C-C
5	1	3	0	3	C-C
6	0	4	1	3	C-C
7	2	3	1	3	C-C
8	1	4	0	3	C-C
9	0	3	2	3	C-C
10	1	3	1	3	C-C
11	0	3	0	3	C-C
12	0	4	0	4	C-C
13	2	3	2	3	C-C
14	1	4	0	4	C-C
15	1	4	1	4	C-C
16	0	2	1	4	C-C
17	0	2	1	3	C-C
18	0	2	2	3	C-C
19	0	2	0	4	C-C
20	0	2	0	3	C-C
21	0	2	0	2	C-C

Continued on next page

Table 5.1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
22	1	2	2	2	C-C
23	2	2	1	4	C-C
24	2	2	0	3	C-C
25	0	4	2	2	C-C
26	2	2	1	3	C-C
27	2	2	2	3	C-C
28	0	2	2	2	C-C
29	2	2	2	2	C-C
30	0	2	1	2	C-C
31	1	2	1	4	C-C
32	1	2	0	4	C-C
33	1	2	1	3	C-C
34	1	2	0	3	C-C
35	1	2	2	3	C-C
36	1	2	1	2	C-C
37	3	2	0	4	C-C
38	3	2	0	3	C-C
39	3	2	2	3	C-C
40	1	3	3	2	C-C
41	3	2	0	2	C-C
42	2	2	3	2	C-C
43	1	2	3	2	C-C
44	1	4	1	2	C-O
45	1	3	1	2	C-O

Continued on next page

Table 5.1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
46	0	4	1	2	C-O
47	0	3	1	2	C-O
48	2	3	1	2	C-O
49	2	2	1	2	C-O
50	0	2	1	2	C-O
51	2	2	0	2	C-O
52	1	3	0	2	C-O
53	1	4	0	2	C-O
54	0	3	0	2	C-O
55	0	4	0	2	C-O
56	2	3	0	2	C-O
57	1	2	1	2	C-O
58	1	2	0	2	C-O
59	0	2	0	2	C-O
60	3	1	1	1	C-H
61	3	2	1	1	C-H
62	1	1	1	1	C-H
63	1	4	1	1	C-H
64	1	3	1	1	C-H
65	1	2	1	1	C-H
66	2	2	1	1	C-H
67	2	3	1	1	C-H
68	2	1	1	1	C-H
69	1	2	1	1	O-H

Continued on next page

Table 5.1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
70	1	1	1	1	O-H
71	2	3			C-M
72	1	4			C-M
73	0	4			C-M
74	0	5			C-M
75	1	3			C-M
76	0	3			C-M
77	2	2			C-M
78	3	2			C-M
79	0	2			C-M
80	1	2			C-M
81	0	3			O-M
82	0	2			O-M
83	1	2			O-M

The obtained 83 R0 clusters are then used as labels for a supervised classification on R1. Linear discriminant analysis (LDA), is applied on the 200-dimension R1 vector descriptors together with the 83 classes as class labels. Figure 5.3c shows the first and the second LDA components of LDA-projected R1 reaction vectors, and the 6 reaction types are represented by 6 different colors for the convenience of visualization. The R1 vectors are reduced to 1-82 dimensions via the LDA projection. The LDA projections represent an unsupervised vector space for analyzing reactions and intermediates, and are used to predict adsorption energies on Rh (111) surface.

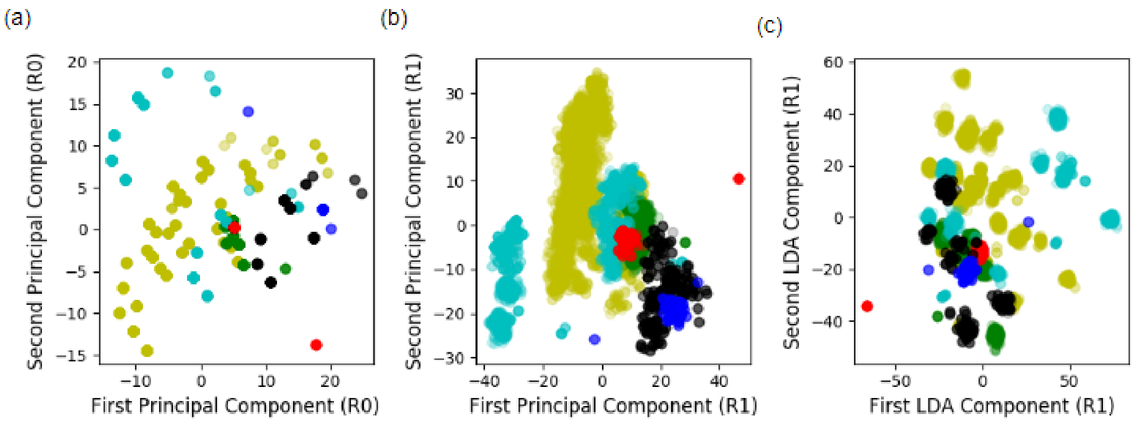


Figure 5.3: Visualization of reaction vector clusters and 6 reaction types (yellow for C-C, green for C-H, cyan for C-O, red for O-H, black for C-M, blue for O-M): (a) 1st vs. 2nd component of PCA (R0), (b) 1st vs. 2nd component of PCA (R1) and (c) 1st vs. 2nd component of LDA (R1 with class labels from R0 clusters).

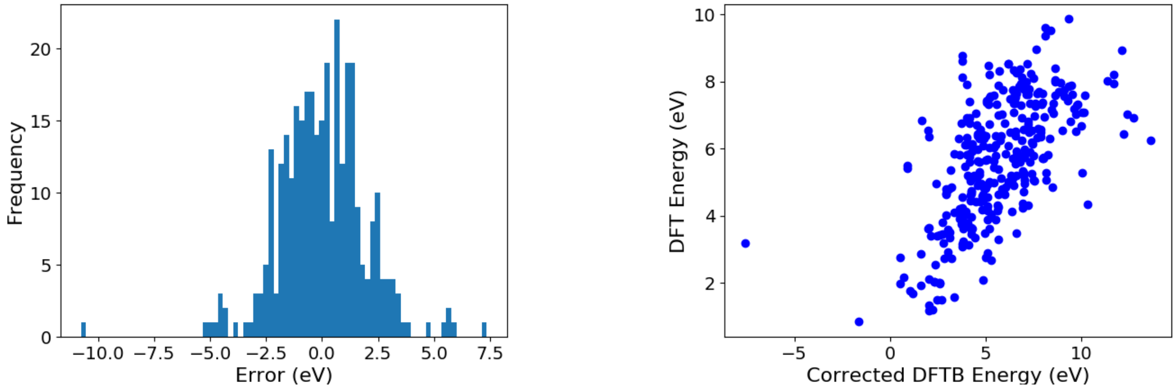
5.3 Adsorption energy calculation and prediction

5.3.1 DFTB and DFT comparison

Ultimately, the goal is to predict adsorption energies of biomass species. The 171 adsorbates, ranging from C1 - C4 species included in this study are generated from the first two bond-breaking recursions chapter 4 of erythrose, glyceraldehyde, glycerol and propionic acid. In general, each species can have multiple adsorption energies due to differences in molecular configuration and binding sites. This makes it challenging to directly predict the most stable binding site even with DFT. For this reason, Hotbit Python package based on DFTB theory is used to generate pre-optimized binding geometries for each adsorbate.

The DFTB-based minima hopping process is used to generate up to 50 adsorption geometries for each of the 171 adsorbates, yielding a total of 857 different adsorbate structures. Geometries that are within 1 eV of the lowest Hotbit energies are then calculated by DFT. This process yields 328 unique geometries and associated adsorption energies, which are used as inputs for supervised training of the regression models.

Adsorption energies with both DFT and DFTB are computed for all 328 unique geometries. This data is used to assess the accuracy of the DFTB energies. The corrected error is calculated by Equation 5.2. Figure 5.4a shows the distribution of the residual DFTB error in energy calculation and Figure 5.4b shows the parity plot of DFT and corrected DFTB energies. The mean absolute error (MAE) of DFTB is 1.46 eV, with notable outliers that can have errors of >5 eV. The results, shown in Figure 5.5b, reveal that DFTB yields incorrect ordering of adsorbates more than 60% of the time. Despite these large energy errors, the geometries are more accurate. The average position difference is 1.04 Å, and 71% of the 328 DFT-converged geometries have the same SMILES notation with the previous DFTB pre-optimized geometry. This suggests that while DFTB is a reasonable tool for generating adsorbate geometries, the energy predictions are not sufficiently accurate to yield chemical insight or even correctly order the stability of various adsorbate geometries.



(a) distribution of corrected errors

(b) parity plot of corrected errors

Figure 5.4: Accuracy of DFTB energies after reference alignment (a) error distribution and (b) parity plot of DFTB vs. DFT energy

5.3.2 ML prediction of adsorption energies

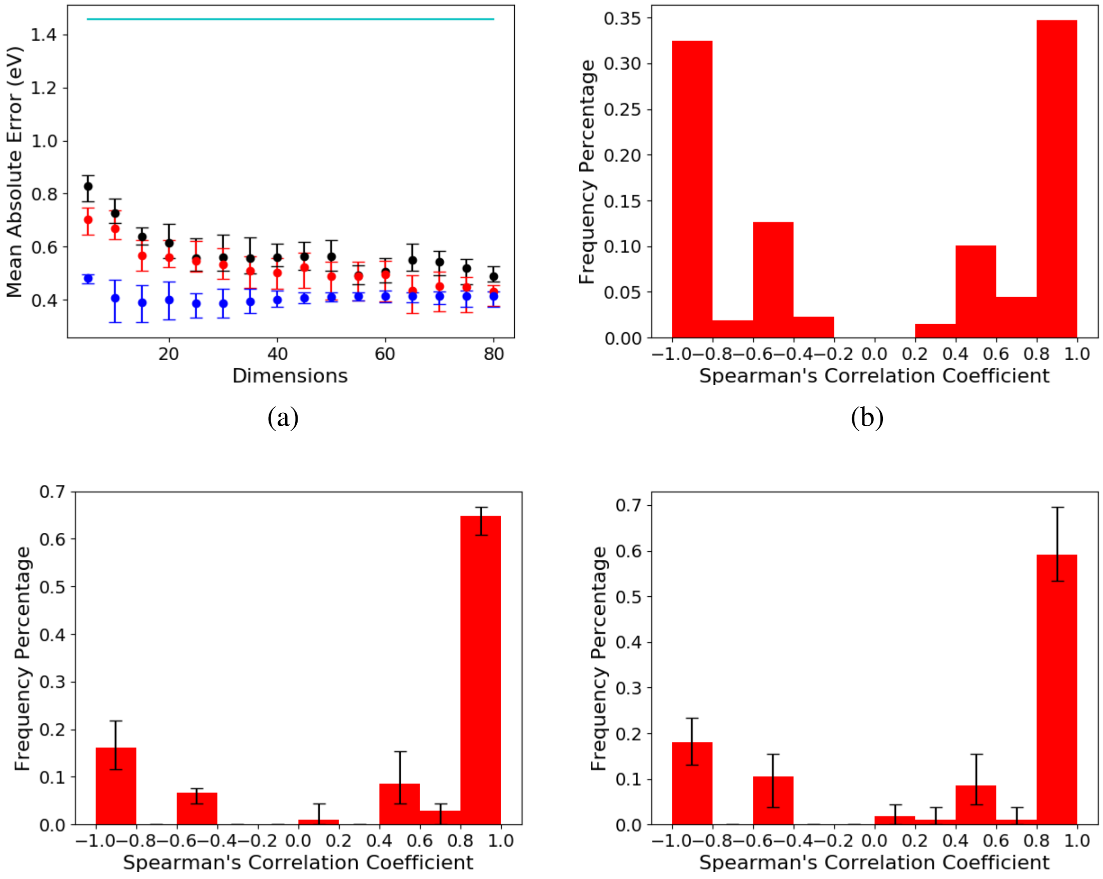
To improve the accuracy of energy predictions we turn to a supervised ML approach. The workflow utilizes the Mol2Vec vectors for each adsorbate as inputs, similar to our previously-developed approach for gas-phase energies [32]. We compare three different linear models for predicting adsorption energies from Mol2Vec vectors: In each case we evaluate the mean absolute error of the test set as a function of the dimension of the input vector, with a maximum dimension of 82 (the maximum dimension of the LDA vectors).

The results of the ML predictions, shown in Figure 5.5a show that the lowest MAE of OLS, PLS and LDA are $\sim$0.51 eV, $\sim$0.39 eV and $\sim$0.43 eV at $\sim$55-dim, $\sim$30-dim and $\sim$65-dim respectively. The MAE of OLS, PLS and LDA decrease at first and reach a plateau or increase slowly after 55-dim, 30-dim and 65-dim. The lowest MAE of the LDA and PLS models are very similar, and we expect that the LDA model will be more transferrable to other adsorbates since the adsorption energies are not used to generate the feature vector. By contrast, we expect that the PLS performance is more specific to this set of adsorbates since the adsorption energies are used to generate the inputs. Comparing the results of the machine-learning models to DFTB, we see that the prediction errors are much lower ($\sim$0.4 eV for ML vs. $\sim$1.5 eV for DFTB). While the error of $\sim$0.4 eV is still somewhat larger than the typical DFT error ($\sim$0.2 eV), the ordering of energies of different geometries with the ML model is relatively good, with a correct ordering at least 59% of the time (Figure 5.5c and Figure 5.5d). However, the ML models requires geometries as inputs, which must be generated by DFTB. This suggests a synergistic approach between the models is required, where DFTB is used to identify geometries and ML is used to predict energies.

5.4 Workflow for identifying stable geometries

5.4.1 Geometry energy order prediction

The average error of the best machine-learning models ($\sim$0.4 eV) is relatively large, indicating that DFT will need to be used when accurate energies are required. However, the number of possible geometries and active sites for biomass molecules makes brute force DFT calculations impractical for large numbers of adsorbates. The combination of geometries from DFTB along with predictions from ML can alleviate this issue by identifying the geometries that are most likely to be stable, thus reducing the number of DFT calculations required. Spearman's correlation coefficient is used to quantitatively assess the model's ability to predict the energy order for different geometries of the same adsorbate. We utilize a 75/25 train/test split with 4 random repeats with the adsorbates for PLS and LDA (approximately 1-3 geometries per adsorbate). Figure 5.5c and Figure 5.5d show the distribution of Spearman's correlation coefficient of PLS and LDA test sets with error bars. PLS and LDA obtain more than 65% and 59% correct energy orders on average with no more than 25% totally inverted (88% of the totally inverted geometries include only 2 geometries). Figure 5.5b shows the distribution of Spearman's correlation coefficient of DFTB with only 35% correct, while more than 30% are totally inverted.



(a)

(b)

(c)

(d)

Figure 5.5: DFTB and ML results of (a) cross validation error of OLS (black), LDA (red), PLS (blue) and DFTB MAE (cyan line), spearman's correlation coefficient of (b) DFTB calculations, (c) 30-dim PLS regression and (d) 65-dim LDA projections with error bars

5.4.2 Criterion for workflow

This demonstrates that both PLS and LDA have similar performance, but both are better than DFTB in predicting energy orders. However, there are still some mis-ordered energies based on the ML predictions, suggesting that the lowest energy structure may not be correctly identified in all cases. To increase robustness, we utilize the average standard deviation of the model errors (0.45 eV) as a tolerance factor, meaning any structure within 0.45 eV of the lowest energy is considered as a possible global minimum. Using simple estimates from probability theory this corresponds to 75% confidence that the true global minimum will be included (see Equation 5.10). The confidence interval is calculated by (assuming $\hat{E}_1 - \hat{E}_2 > 0.45$ and $\sigma = 0.45eV$):

$$E - \hat{E} = Z \sim Normal(0, \sigma^2) \tag{5.3}$$

$$E_1 = \hat{E}_1 + Z \sim Normal(\hat{E}_1, \sigma^2) \tag{5.4}$$

$$E_2 = \hat{E}_2 + Z \sim Normal(\hat{E}_2, \sigma^2) \tag{5.5}$$

$$E_1 E_2 \Rightarrow E_1 - E_2 \sim Normal(\hat{E}_1 - \hat{E}_2, 2\sigma^2) \tag{5.6}$$

$$P(E_1 - E_2 < 0) = \Phi(\frac{\hat{E}_2 - \hat{E}_1}{2\sigma^2}) = 1 - \Phi(\frac{\hat{E}_1 - \hat{E}_2}{2\sigma^2}) \tag{5.7}$$

$$\frac{\hat{E}_1 - \hat{E}_2}{\sqrt{2}\sigma} > \frac{1}{\sqrt{2}} \Rightarrow P(E_1 - E_2 < 0) < 1 - \Phi(\frac{1}{\sqrt{2}}) = 0.25 \tag{5.8}$$

$$P(E_1 < E_2 | \hat{E}_1 - \hat{E}_2 > 0.45) < 0.25 \tag{5.9}$$

$$\Rightarrow P(E_1 > E_2 | \hat{E}_1 - \hat{E}_2 > 0.45) > 0.75 \tag{5.10}$$

This cutoff can be adjusted to improve confidence at the expense of more DFT calculations. Both PLS and LDA models are built based on the 328 DFT calculations and are used for predicting all 857 geometries generated from DFTB and minima hopping. The minimum energy of each adsorbate and the geometries within 0.45 eV of the minimum energy are extracted, and any geometries that have not already been computed are calculated with DFT. Both of the ML models are used, and the geometries are calculated with DFT if the energy predicted by either model is within the threshold.

5.5 ML and new DFT calculations

5.5.1 Summary of new global minimas found by ML

The ML model identifies multiple possible new global minima for 65 of the adsorbates, corresponding to 154 additional DFT calculations. The MAE between the model predictions and the DFT energies of the new structures are 0.60 eV and 0.62 eV for LDA and PLS models, about 50% higher than the MAE of the test set. The results of these calculations reveal that the energies of many of these structures are lower than the previously-determined lowest energy structure, as shown in Fig. Figure 5.6a (see SI for details). For 20 of the 65 adsorbates the configuration of the global minima was sufficiently different to lead to a new SMILES representation. For example, for the the adsorbate $[O]CC(O)[CH]$ the ML model identifies a $[CH]$ bidentate-binding geometry with 0.44 eV lower energy than the monodentate geometry that was previously identified as the lowest energy structure, as shown in Fig. Figure 5.6b. For 13 other adsorbates, the difference in geometry was less drastic so that the SMILES string of the new structure was the same as the old structure, but the energy was slightly different by <0.23 eV (see Fig. Figure 5.6b). The original energy was lower than all newly-computed energies in 32 adsorbates. Overall, the original workflow for identifying global minima failed to identify the global minimum for at least 17% of adsorbates, and had qualitative differences in binding geometries (different SMILES string or energy difference $>$ 0.05 eV) for 14% of adsorbates.

5.5.2 Workflow

The results of this work suggest that a combination of physical approximations and ML models is a promising route toward identifying global minima of complex adsorbates. A

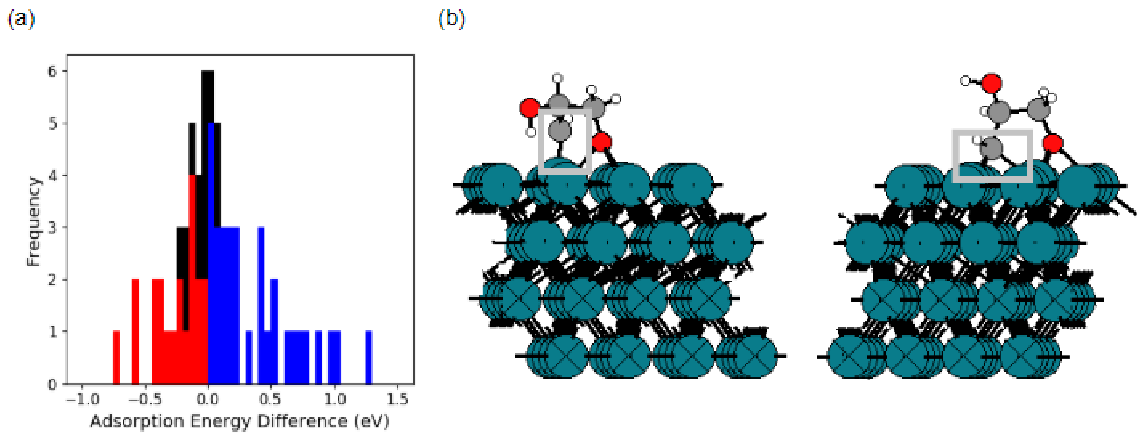


Figure 5.6: Results of DFT calculations for structures predicted to be low-energy by ML model. (a) Stacked bar plot of adsorption energy difference for structures with the same (black) or different (red/blue) SMILES strings from prior global minimum, with lower-energy structures in red and higher-energy structures in blue. (b) A representative example of an adsorbate ($[O]CC(O)[CH]$) where the new lower-energy geometry binds with a qualitatively different structure, where $[CH]$ binds directly to a single Rh atom (left, gray box) instead of two Rh atoms in the previous structure (right, gray box).

general workflow (see Figure 5.8 and Figure 5.8) involves a first step that uses an approximate physical method (DFTB in this case) to rapidly generate many candidate geometries. The second step involves using DFT to calculate the energies of the most stable structures (structures within 1 eV of the minimum in this case, leading to 328 geometries of 171 adsorbates). Third, these DFT energies are used to train ML models, here based on Mol2Vec and linear regression, and the ML models are then used to predict the energies of all candidate structures. Finally, the predictions of the ML model are used to identify new structures that will be computed with DFT, in this case structures within 0.45 eV (the standard error of the ML models on the test set) of the predicted minimum. The results of this work show that this process yields 20 global energy minima that would have been incorrect without the use of the ML model. We note that a similar process could be applied without an ML model, by simply increasing the threshold of DFT calculations from the original DFTB energies. However, increasing this threshold to 1.45 eV (the standard error of DFTB energies) would result in 149 additional DFT calculations, and would still only identify two of the new low-energy structures. This indicates the utility of the ML model in the workflow.

Nevertheless, there is room for improvement, and the workflow can be made more efficient with improved physical approximations and more accurate ML models, which may be necessary to tackle larger and more complex biomass molecules.

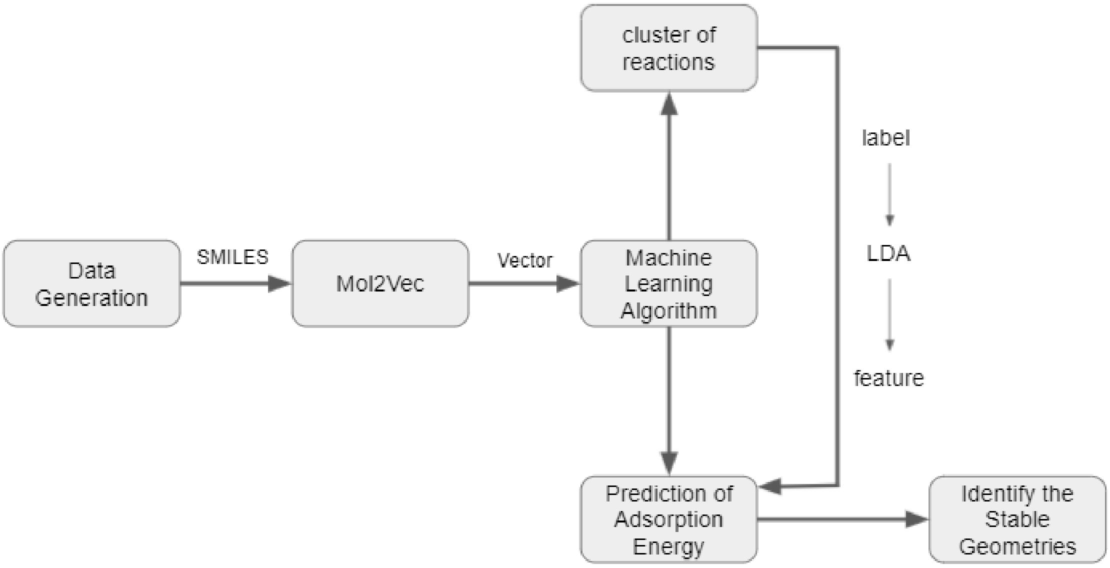

cluster of reactions
label
LDA
feature
Data Generation
SMILES
Mol2Vec
Vector
Machine Learning Algorithm
Prediction of Adsorption Energy
Identify the Stable Geometries

Figure 5.7: Workflow from data generation to model construction.

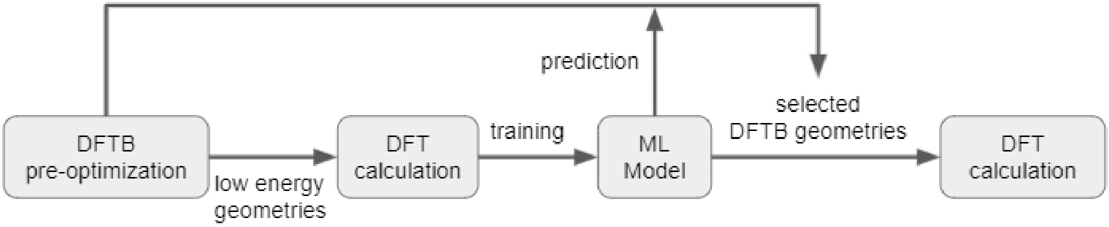


Figure 5.8: Workflow for identification of stable geometries on the Rh (111) surface.

5.6 Conclusion and future work

The size and complexity of biomass molecules leads to a major challenge in predicting the global minimum adsorption geometry, and this challenge is compounded by the number of possible intermediates that appear in biomass reaction networks. It is clear that new techniques are needed to accelerate the study of these systems since direct calculation with

DFT or other quantum chemical techniques is impractical. Here, we show that ML models based on Mol2Vec descriptors can achieve an MAE of 0.39 eV (PLS with 30-dim) and 0.41 eV (LDA with 65-dim) when applied to 171 intermediates derived from erythrose, glyceraldehyde, glycerol and propionic acid. These models provide more accurate estimates of adsorption energies than DFTB (1.46 eV), but the lowest MAE of ML methods are still not comparable to DFT. Spearman's correlation is used here to show that ML methods are much more reliable for assessing the relative stability of different geometries than DFTB. However, ML methods still need the structure input from DFTB. We combine the strengths of both to provide a more robust route to identifying low-energy structures.

The best aspects of ML and DFTB techniques are combined leading to a new workflow of DFTB+minima hopping → DFT → ML → DFT. DFTB is used first to generate up to multiple possible local minima for each of the adsorbates. Low-energy structures are calculated with DFT. The DFT energies together with the vector descriptors are used to build ML model and applied to all geometries from DFTB. Finally, geometries with predicted low-energies are calculated by DFT. This approach allows us to discover 20 new global minima for the 171 adsorbates studied here, which would be missed if only DFTB were used to evaluate candidate structures. Nonetheless, the workflow also has the limitation that there is still uncertainty about whether or not the true global minima is found, since an exhaustive search is not feasible for these complex adsorbates. However, the results indicate that combining physical models and ML predictions is a promising path toward solving this challenging problem.

Following the extension from gas-phase model to Rh surface species in this chapter, a potential future computational work would be the estimation of the energies for more of the intermediates and predict the probability for each of the recursions. Also, as a model and workflow for Rh surface is already built, we are also interested to see how the experiment results will compare to our ML models. Also, these experimental results from reaction pathway, intermediates formation and spectrum can help with further improvement of the model.

CHAPTER 6
DEVELOPMENT AND APPLICATION OF NEURAL NETWORK FORCE FIELDS FOR BIOMASS MOLECULES

6.1 Introduction

As discussed in chapter 5, DFTB provides a fast physics-based method to pre-optimize the geometry and estimate the energy of adsorption systems. This is an important step in exploring the different possible adsorbate geometries and identifying the global minimum adsorption energy. However, DFTB needs significant parameterization and stable parameters are often difficult to obtain. The DFTB method also requires a lot of constraints on the geometry optimization (e.g. fixed bond lengths), and the energies are wildly inaccurate as well (see chapter 5). It is clear that an improved method for rapidly estimating geometries and energies of adsorbates on surfaces is needed to better identify global energy minima for large adsorbates. Neural network force fields (NNFF) [118] could be a potential alternative, offering a more transferable data-driven alternative to DFTB that can simultaneously optimize geometries and offer more accurate predictions of energies.

Behler and co-workers constructed high-dimensional neural network potentials back in 2010s [118], they used the structural descriptors as well as tested the previous BP and SOAP descriptors [119]. The feed forward neural network was training on energies and the forces were calculated by taking the derivative of the predicted energies. They tested the neural network potentials on Cu and Zinc oxide systems (single atomic species), both achieves RMSE $\sim$ 4.5 meV/atom [120, 121]. After Behler's work and with the development of deep learning tools recently, there are a few other descriptors with neural network force fields implementation in different areas, for example, constructions of the translational and rotational descriptors by Wang and co-workers in neural network and found that

1000 MD steps is sufficient to train with satisfied accuracy [122] and similar descriptors based on electronic structure is used in Hu and co-workers work, they are able to apply it into predicting thermal transport of Si [123]. In our study, we are using the GMP descriptor [124]. GMP is proved to be a universal electronic structure based descriptor on atomic structures.

6.2 Methods

For the work in this section, the dataset is changed to 61 smaller molecules/intermediates that contain no more than 4 C and 4 O on 5 transition metals (211) surfaces: Ag, Au, Pd, Pt, Rh. We obtain more training data but compromise on molecule size. Quantum ESPRESSO [88] with (4, 4, 1) k-points, 500 eV planewave cutoff and PBE exchange correlation function are used for all DFT calculations in this chapter. AMPtorch, combining the AMP package and the Pytorch package, is the tool used for training the NNFF with GMP descriptor. A total of $\sim$18,000 data points are taken from the converged and non-converged structures of the 61 adsorbates on 5 different metal (211) surfaces. We start from Rh since the rest of the thesis is based on Rh surface. The train/test is split as 80/20 based on the adsorbates. There are $\sim$ 30 trajectory structures for each adsorbate both converged and non-converged. The train/test is split in a way that all the structures of a specific adsorbate is either in train or test datasets.

6.3 Results on Rh surfaces

We start by evaluating the performance of NNFF models on the Rh surface. The middle steps between the initial and the converged structures of adsorbates on Rh (211) surface are taken randomly and a total of $\sim$15,000 structures are used with a 80/20 train/test split based on adsorbates. The hyperparameters for the NNFF are determined by an orthogonal experimental design method. The hyperparameter combinations are shown in Table 6.1. Finally, descriptor order of 4, nodes of 25, layers of 4 and epochs of 2400 are used based on the test error of energy training. We also checked the errors for different train/test splits under these hyper-parameters to evaluate the influence of more training data (see Figure 6.1).

Table 6.1: Orthogonal design experiments of hyperparameters

Descriptor order	Nodes	Layers	Epochs
3	10	3	1600
3	15	4	2000
3	20	5	2400
3	25	6	2800
4	10	4	2400
4	15	3	2800
4	20	6	1600
4	25	5	2000
5	10	5	2800
5	15	6	2400
5	20	3	2000
5	25	4	1600
6	10	6	2000
6	15	5	1600
6	20	4	2800
6	25	3	2400

The energy error provides an indication of how accurately the NNFF will be able to predict adsorption energies, and errors approaching 0.3 eV are promising since this approaches the accuracy of typical exchange-correlation error ($\sim$0.2 eV). However, for predicting geometries it is also necessary to predict forces. AMPtorch allows the hyperparameter “force coefficient” to add the trained force error to the loss function, and the default value of 0 leads to the results in Figure 6.1. A total of 4 additional force coefficients of 0.01, 0.1, 1, 10 are implemented here with the same hyperparameters obtained from the previous orthogonal experiment design. The results, shown in Table Table 6.2, reveal a trade-off between the force error and energy error. The test energy error is only 0.32 eV when the forces are not trained, and the force error decreases while the energy error increases with the increasing force coefficient. The force error estimated at force coefficient 10 is 0.066 eV/$\AA$, which is very close to the DFT self-consistent setting 0.05 eV/$\AA$. However, the energy error at 10

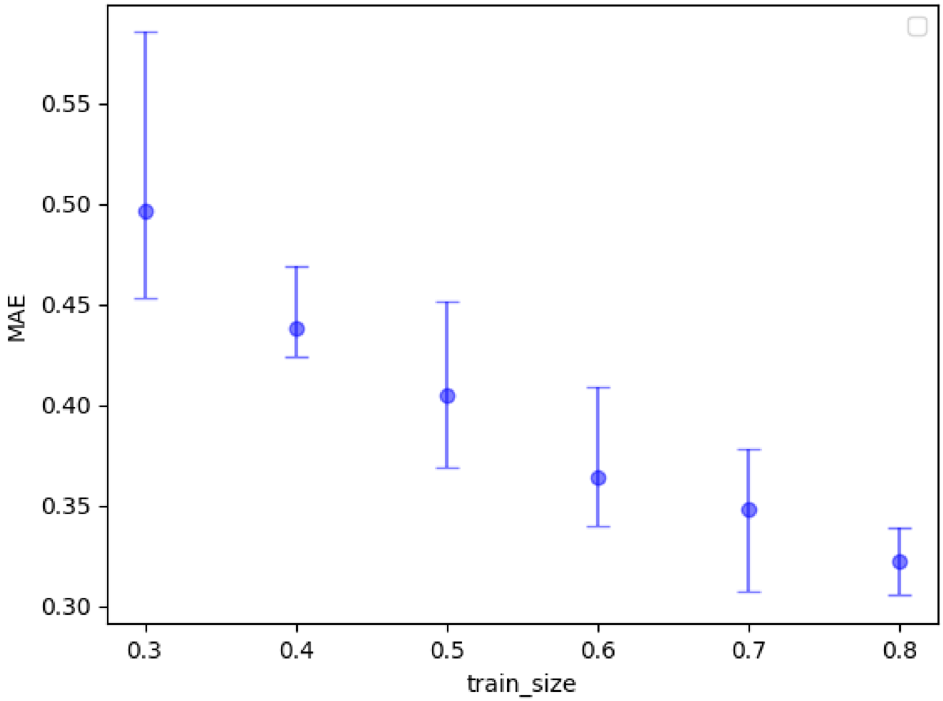


Figure 6.1: Validation errors (energy) under descriptor order of 5, nodes of 25, layers of 4 and epochs of 2400 with different train/test split

is as high as 0.94 eV, which is far more than energy error at force coefficient 1. Finally, a force coefficient 1 is used for the following analysis since it provides a good compromise between force and energy errors. In the Rh test case, we achieve a validation force error of 0.083 eV/$\AA$, which is more accurate than most reactive force fields for catalysts [125]. The energy error increases to 0.53 eV in this case, but this is still comparable to the accuracy of the more specialized machine-learning models presented in Chapter chapter 5.

Table 6.2: MAE for energy and force with different force coefficient errors

Force coefficient	Energy MAE	Force MAE
0	0.32 eV	N/A
0.01	0.46 eV	0.26 eV/$\AA$
0.1	0.49 eV	0.12 eV/$\AA$
1	0.53 eV	0.083 eV/$\AA$
10	0.94 eV	0.066 eV/$\AA$

6.4 Transfer learning

In the subsequent sections we explore two strategies for improving the accuracy and utility of the NNFFs using "transfer learning", where models are trained on one type of data and then transferred to another. In the first section we seek to reduce the amount of DFT training data needed by leveraging the large amounts of DFTB data. In the subsequent section we show that models can be transferred between different metals, opening the door to more general purpose neural network force fields.

6.4.1 Transfer learning from DFTB to DFT

As DFTB parameterization is already trained with Rh surface and a large amount of Rh data could be obtained by this approach, the DFTB data from different binding geometries of the 61 small molecules on Rh (211) surface are pre-trained to obtain an initial NNFF with $\sim$15k data points and validation MAE $\sim$0.8-0.9 eV. Then, the parameters from this initial NNFF are taken as the starting point for the DFT results (see Figure 6.2).

However, the pre-trained NNFF parameters with DFTB does not either show a faster convergence or a better final accuracy. The reason might be that the energies and forces calculated by DFTB is of low accuracy, and the pre-trained NNFF is not approaching a correct description of the system (see chapter 5 for corrected DFTB errors compared to DFT) . Thus there is not much difference compared to starting from the random weights and bias, and both could reach a minimum error of 0.53 eV for energy and 0.083 eV/$\AA$ for force. So, we conclude that starting from models pre-trained with DFTB data is not likely to be a a good strategy for transfer-learning.

6.4.2 Transfer learning between different metals

As the NNFF is already trained for the adsorbates on Rh (211) surface, we take the previously-trained parameters as the starting point for the extension to small biomass molecules on

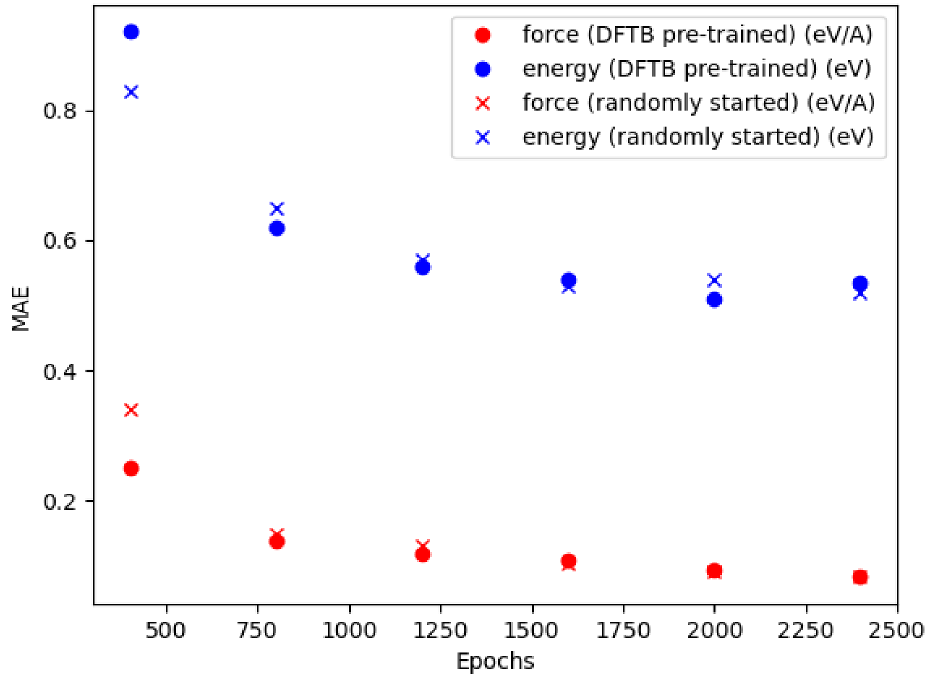


Figure 6.2: Energy error and force error for starting from pre-trained DFTB NNFF weights and random-generated weights with different epochs for Rh DFT results

multiple metal surfaces. As a comparison, we also tried the random starting point of the NNFF with the same hyperparameters. The pre-trained parameters show faster convergence at the beginning of the training process, while the validation error is similar after 1600 epochs and the final lowest MAE for both pre-trained and random-started parameters are similar to each other (see Figure 6.3). After training NNFF for 2400 epochs, both could reach a force error as 0.06 eV/$\AA$ and energy error 0.11 eV. This implies that transfer learning from the Rh neural network would probably save time and computation force with lower epochs to train.

We also noticed that the validation error for multiple metals (0.11 eV for energy and 0.06 eV/$\AA$ for force) are much lower than the Rh metal only (0.53 eV for energy and 0.08 eV/$\AA$ for force). To find out the reason for this, NNFF for each of the rest 4 single metals are trained and corresponding validation errors are calculated. The validation errors for Rh,

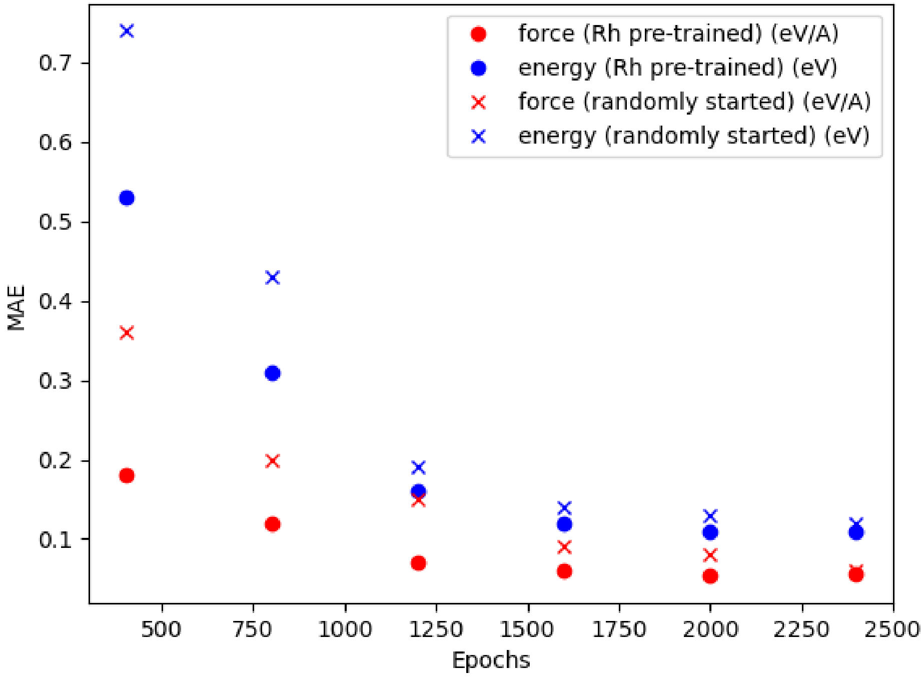


Figure 6.3: Energy error and force error for starting from pre-trained Rh DFT results weights and random-generated weights with different epochs for multiple metal DFT results

Pd, Pt, Au and Ag are shown in Table 6.3 and all 5 type of metals have an energy MAE $\sim$ 0.5 eV and force MAE of $\sim$ 0.1 eV/$\AA$. Thus, the decrease in validation error when training on multiple metal data indicates that the overall model is considerably improved due to the increase in total number of data points ($\sim$4K points per metal). This is a promising result that suggests that the NNFF for multiple metals are consistent with each other, meaning there is not a tradeoff between training with different metals. This indicates that the model trained on all metals is likely to be a good starting point for a general purpose model for optimization of molecules on metal surfaces.

One exciting capability of the GMP features is their ability to extrapolate between elements. The decrease in energy and force error as more elements are added is a promising indication that the model is learning to interpolate between elemental species. To further test this ability, extrapolation to new metals with the NNFF is explored. To achieve this, we

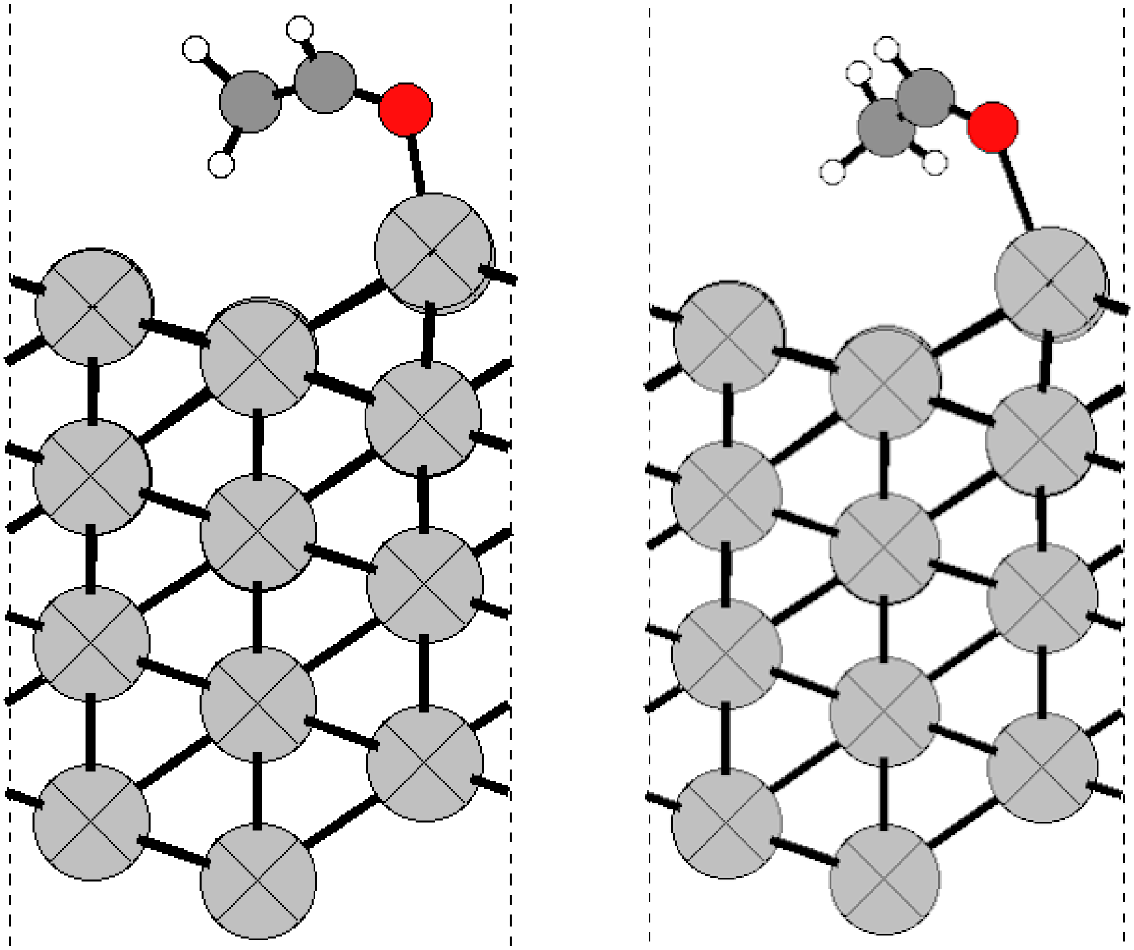
Table 6.3: Single metal NNFF validation errors

Metal	Energy MAE	Force MAE
Rh	0.53 eV	0.083 eV/$\AA$
Pt	0.48 eV	0.094 eV/$\AA$
Pd	0.56 eV	0.100 eV/$\AA$
Au	0.49 eV	0.108 eV/$\AA$
Ag	0.62 eV	0.104 eV/$\AA$

train on all metals except one, and test the model on the metal that was omitted from training. We do include a single example of the test metal to enable reasonable feature scaling, but this single point represents a negligible fraction of the total training data (one point out of ∼4K). This test is performed with two metals, Ag and Pd. The validation errors are 0.42 eV and 0.39 eV, respectively. While these errors are higher than the errors of the metals included in the training set (∼0.1 eV), it is remarkable that they are lower than the errors for models trained on ∼4K examples of Ag and Pd (Table Table 6.3). This further confirms the ability of the GMP-based NNFF models to predict energies and forces across multiple metals.

6.5 Conclusions and future work

A NNFF based on GMP features and AMPtorch is being developed for biomass molecules and transition metal catalytic systems. We have already done a preliminary optimization of hyperparameters and the training of the force field based on these hyperparameters. The force error of 0.06 eV/$\AA$ and energy error of 0.11 eV could be achieved in the independent test set. The errors here are already low enough since DFT calculations have an estimated error of $\sim$ 0.2 eV. We also checked the NNFF predicted error of our previous biomass adsorbates on Rh(111) surface. The energy error is 0.44 eV and the force error is 0.15 eV/$\AA$. We would need include more training data from the biomass-Rh(111) system to get a more accurate result. However, the predicted error is already a lot better than the DFTB method. The preliminary results suggest that the NNFFs can be combined with ASE as a



(a) AMPtorch converged structure (convergence takes ∼5 mins)

(b) DFT converged structure (convergence takes ∼20 hrs)

Figure 6.4: Converged optimization of structures by (a) AMPtorch (within 5 mins) and (b) DFT (∼ 20 hrs)

potential pre-optimization tool instead of DFTB. Figure 6.4 shows an example starting from the same initial guess for a small C2 molecule (CH_3CHO). DFT takes more than 20 hrs to converge with the final energy as 2.97 eV while AMPtorch only takes 5 mins to converge with the final energy as 2.45 eV (only ∼0.5 eV in difference). The advantage of AMPtorch as compared to Hotbit is that the resulting models are transferrable to other elements and more accurate in terms of both energies and forces. Also, as we already have the previous workflow for combining the physics-based method (DFTB) and ML models to identify the stable binding geometries, we could also compare the pre-optimization workflow between physics-based DFTB method and AMPtorch NNFF in future work.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusion

In summary, we first validated the DFT level of theory by comparing the calculation results and experimental results for small oxygenates on the MoO_3 surface. Similar DFT settings are used for the remainder of the calculations. Then, a data generation algorithm is developed to generate all possible intermediates for gas-phase biomass molecules. Next, Mol2Vec is used for embedding the generated intermediates. DFT calculations of reaction energies and the vector descriptors are combined for predicting reaction energies and classification of reaction types through machine learning methods. The framework is then adapted to study biomass intermediates adsorbed to a transition-metal surface, where DFTB, DFT and machine learning methods are combined for identifying stable geometries and their adsorption energies on the Rh (111) surface. Finally, a neural network force field model is developed as a route to simultaneously solve the problem of geometry optimization and energy prediction on various transition-metal surfaces.

First, the adsorption energy of 6 small biomass molecules (ethanol, methanol, acetaldehyde, formaldehyde, glycolaldehyde and crotonaldehyde) are calculated with BEEF-vdW. The results show that biomass molecules are physisorbed on the pristine MoO_3 surface and chemisorbed on MoO_3 surfaces with an oxygen vacancy. This is found to be in qualitative agreement with experimental results. Also, the trends in vibrational frequency calculations of ethanol, crotonaldehyde and acetaldehyde are in semi-quantitative agreement with the experimental DRIFT spectrum for most vibrational modes. The agreement between experiment and theory verifies the DFT approach as a valid method for calculation of energies and vibrational frequencies.

Subsequently, the problem of complexity in biomass intermediates was tackled using gas-phase structures. Energies of intermediates are needed for studying the pathways and mechanisms of biomass molecules conversion on transition metal surfaces. A recursive chemical bond breaking algorithm is developed for the generation of all possible intermediates of several key biomass molecules. The bond breaking reactions are also kept while generating the intermediates. A total of 91,098 intermediates (represented by SMILES notations) are used as the input into Mol2Vec for an unsupervised embedding and 13,422 reactions are used for classification. Euclidean distance in radius 0 and cut-off 0.05 are used for clustering the reactions according to the reaction type and atomic environment. LDA is used for classification on 90 classes given by radius 0 clustering. General OLS and LDA projected vector descriptors are implemented for reaction energy prediction and a lowest MAE of 0.59 eV (compared to group additivity MAE 1.59 eV). The model compound assessment is also based on vector descriptors. The length of the vector represents the size of the structure and the direction of the vector represents the chemical properties of the structure. So the similar-sized model compound is assessed according to the distance between its vector and the vector of the compound studied.

Besides gas-phase molecules, intermediates on metal surfaces are also important. An algorithm for adding metals recursively to the SMILES notations is developed to get surface species. Following the previous gas-phase workflow, we are able to predict adsorption energies of intermediates taken from the first two recursions of erythrose, glyceraldehyde, glycerol and propionic acid with MAE as 0.39 eV and 0.41 eV with PLS and LDA respectively. Furthermore, the physics-based fast DFTB method is combined with ML for identifying the stable binding geometries on the Rh (111) surface. A new workflow of DFTB+minima hopping $\rightarrow$ DFT $\rightarrow$ ML $\rightarrow$ DFT is proposed and is able to find 14% new energy-minima structures of the 171 studied adsorbates. Though the workflow has the limitation of uncertainty about whether the true global minima is found, and suffers from inaccuracies in DFTB, it is still a promising general strategy to solve the challenging problem of global energy minimization for biomass molecules on surfaces.

Finally, we aim to use NNFFs based on the new GMP framework to create machine-learning models that can predict both geometries and energies. The results show that the approach is able to reach comparable accuracies to other ML approaches for Rh surfaces, and also indicate that accuracies can be improved further by using training data from multiple metals. This exciting preliminary result suggests that it should be possible to develop a general-purpose NNFF model for predicting the geometries and energies of biomass intermediates on any metal surfaces.

DFT is a useful tool in predicting the energies of biomass molecules on catalytic surfaces, but the computational cost is too high for it to be applied directly to all intermediates and transition-states. Our results show that machine-learning techniques provide a promising route to tackling the complexity of biomass reaction networks, although the current accuracy suggests that improved methods are required. Moreover, we show that physics-based methods like DFTB and DFT can be combined with machine learning to develop workflows that accelerate searches for intermediate geometries and energies, and that this general strategy is a promising way to capitalize on the speed of ML methods and the accuracy of physics-based approaches. However, considerable room remains for the development of more efficient and accurate strategies of analyzing biomass reaction pathways with computational methods.

7.2 Future work

7.2.1 Experimental validation of adsorption stability

In chapter 5, a ML model for the prediction of adsorption energies on Rh (111) surface is built and a new workflow for identification of stable binding geometries of biomass intermediates are proposed. Experimental validation should start from the 4 biomass molecules studied (erythrose, glyceraldehyde, glycerol, propionic acid). Surface science studies using single crystals would be ideal, although results from Rh nanoparticles on an inert support would also yield insight. A combination of reactivity studies and spectroscopic investigations would be necessary to validate the results. Infrared spectra can be simulated for the different geometries and validated against experimental results, where the spectra would correspond to the most stable species and geometries. NMR spectroscopy would also yield insight into the details of the adsorption configurations. Reactivity studies would yield insight into which products were most favorable from decomposition of these 4 molecules, and could be a starting point for development and validation of microkinetic models.

7.2.2 Application to fragmentation patterns

Mass spectroscopy was found to be able to study the fragmentation patterns of linear chemicals in 1970s [126], and is a widely used technique in quantifying the concentrations of biomass molecules. However, the complex structure of biomass molecules leads to thousands of gas-phase species with similar spectroscopic signals, which is a significant challenge of mass spectroscopy of biomass molecules. Our algorithm for recursive intermediate generation could be used to generate gas-phase fragmentation possibilities of biomass compounds. The algorithm could be used to determine the possible molecular weight and structure of possible species, and the results could aid in the interpretation of mass spectra. The algorithm could be extended by constructing a library based on the bond breaking intermediates results, which would be helpful for interpreting the experimental results. Finally, the ML model could be used to estimate the energy of each intermediate and, by coupling with experimental results, the probability of existence of each intermediate could be determined.

7.2.3 NNFF development and pre-optimization tool

Low-cost physical approximations such as DFTB exhibit great speed, but require specific parameters, and suffer from relatively low accuracy. NNFFs are a potential alternate data-driven substitution of the DFTB method. NNFFs that we use is based on the python package AMPtorch and GMP descriptor and takes multiple binding geometries with forces and energies to train. NNFFs, combined with geometry optimization, can pre-optimize the structure and calculate the energy for each step just like other physics-based methods. The neural network force field is being developed for the biomass molecule and transition metal catalytic systems. The best hyper-parameter tuning for the force field and the convergence of the optimization step with the interface of ASE package still remains a challenge. The future work of the development of force field will be testing different hyper-parameters as well as applying transfer learning from simple systems to complex systems. The trained force field could be used as a potential pre-optimization tool if all the problems are solved and it gives a reasonable estimation of the geometries with local minima energy.

The estimated energy could potentially be sufficient accurate to construct a micro-kinetic model directly without requiring additional DFT calculations as the current trained force field has a validation energy error of 0.11 eV and force error of 0.06 eV/$\mathring{A}$. The error is impressive since the accuracy of DFT itself is generally only considered to be about 0.2 eV. A neural-network force field capable of predicting energies of arbitrary biomass molecules on arbitrary transition-metal surfaces would greatly accelerate screening efforts for biomass catalysis. For example, descriptor-based microkinetic screening studies may require thousands of DFT calculations for complex reaction mechanisms [127]. The ability to rapidly predict these energies to within DFT accuracy would make it practical to study the complex reaction networks for biomass conversion reactions, and provide an improved route to rationally engineering active and selective catalysts for interconversion of biomass compounds. The results of this thesis indicate that machine-learning techniques are a promising strategy to achieving this goal.

Appendices

APPENDIX A

EXPERIMENTAL EQUIPMENT

A telescope and a spectrometer were used to analyze the sun. Many other instruments were used.

APPENDIX B

DATA PROCESSING

Data was processed before being added to this document.

REFERENCES

[1] T. Werpy, G. Petersen, A. Aden, J. Bozell, J. Holladay, J. White, A. Manheim, D. Eliot, L. Lasure, and S. Jones, “Top value added chemicals from biomass. volume 1-results of screening for potential candidates from sugars and synthesis gas,” Department of Energy Washington DC, Tech. Rep., 2004.

[2] J. J. Bozell and G. R. Petersen, “Technology development for the production of biobased products from biorefinery carbohydrates—the us department of energy’s “top 10” revisited,” *Green Chemistry*, vol. 12, no. 4, pp. 539–554, 2010.

[3] H. Yan, H. Qin, X. Feng, X. Jin, W. Liang, N. Sheng, C. Zhu, H. Wang, B. Yin, Y. Liu, *et al.*, “Synergistic pt/mgo/sba-15 nanocatalysts for glycerol oxidation in base-free medium: Catalyst design and mechanistic study,” *Journal of Catalysis*, vol. 370, pp. 434–446, 2019.

[4] T. Jedsukontorn, N. Saito, and M. Hunsom, “Photoinduced glycerol oxidation over plasmonic au and aum (m= pt, pd and bi) nanoparticle-decorated tio2 photocatalysts,” *Nanomaterials*, vol. 8, no. 4, p. 269, 2018.

[5] L. S. Ribeiro, E. G. Rodrigues, J. J. Delgado, X. Chen, M. F. R. Pereira, and J. J. Orfao, “Pd, pt, and pt–cu catalysts supported on carbon nanotube (cnt) for the selective oxidation of glycerol in alkaline and base-free conditions,” *Industrial & Engineering Chemistry Research*, vol. 55, no. 31, pp. 8548–8556, 2016.

[6] E. de Jong, A. Higson, P. Walsh, and M. Wellisch, “Product developments in the bio-based chemicals arena,” *Biofuels, Bioproducts and Biorefining*, vol. 6, no. 6, pp. 606–624, 2012.

[7] E. Lam and J. H. Luong, “Carbon materials as catalyst supports and catalysts in the transformation of biomass to fuels and chemicals,” *ACS catalysis*, vol. 4, no. 10, pp. 3393–3410, 2014.

[8] D. S. Pisal and G. D. Yadav, “Single-step hydrogenolysis of furfural to 1, 2-pentanediol using a bifunctional rh/oms-2 catalyst,” *ACS omega*, vol. 4, no. 1, pp. 1201–1214, 2019.

[9] J. Zhang, Z. Zhong, X.-M. Cao, P. Hu, M. B. Sullivan, and L. Chen, “Ethanol steam reforming on rh catalysts: Theoretical and experimental understanding,” *ACS Catalysis*, vol. 4, no. 2, pp. 448–456, 2014.

[10] S. Cavallaro, “Ethanol steam reforming on rh/al2o3 catalysts,” *Energy & Fuels*, vol. 14, no. 6, pp. 1195–1199, 2000.

[11] D. Mei, V. Lebarbier Dagle, R. Xing, K. O. Albrecht, and R. A. Dagle, "Steam reforming of ethylene glycol over mgal2o4 supported rh, ni, and co catalysts," *ACS Catalysis*, vol. 6, no. 1, pp. 315–325, 2016.

[12] K. Abdelfatah, W. Yang, R. Vijay Solomon, B. Rajbanshi, A. Chowdhury, M. Zare, S. K. Kundu, A. Yonge, A. Heyden, and G. Terejanu, "Prediction of transition-state energies of hydrodeoxygenation reactions on transition-metal surfaces based on machine learning," *The Journal of Physical Chemistry C*, vol. 123, no. 49, pp. 29 804–29 810, 2019.

[13] F. Auneau, C. Michel, F. Delbecq, C. Pinel, and P. Sautet, "Unravelling the mechanism of glycerol hydrogenolysis over rhodium catalyst through combined experimental–theoretical investigations," *Chemistry–A European Journal*, vol. 17, no. 50, pp. 14 288–14 299, 2011.

[14] W. Yang, R. V. Solomon, O. Mamun, J. Q. Bond, and A. Heyden, "Investigation of the reaction mechanism of the hydrodeoxygenation of propionic acid over a rh (1 1 1) surface: A first principles study," *Journal of Catalysis*, vol. 391, pp. 98–110, 2020.

[15] S. Wang, K. Yin, Y. Zhang, and H. Liu, "Glycerol hydrogenolysis to propylene glycol and ethylene glycol on zirconia supported noble metal catalysts," *ACS Catalysis*, vol. 3, no. 9, pp. 2112–2121, 2013.

[16] B. R. Goldsmith, J. Esterhuizen, J.-X. Liu, C. J. Bartel, and C. Sutton, "Machine learning for heterogeneous catalyst design and discovery," 2018.

[17] G. H. Gu, P. Plechac, and D. G. Vlachos, "Thermochemistry of gas-phase and surface species via lasso-assisted subgraph selection," *Reaction Chemistry & Engineering*, vol. 3, no. 4, pp. 454–466, 2018.

[18] V. Vorotnikov, S. Wang, and D. G. Vlachos, "Group additivity for estimating thermochemical properties of furanic compounds on pd (111)," *Industrial & Engineering Chemistry Research*, vol. 53, no. 30, pp. 11 929–11 938, 2014.

[19] Y. Chen and D. G. Vlachos, "Hydrogenation of ethylene and dehydrogenation and hydrogenolysis of ethane on pt (111) and pt (211): A density functional theory study," *The Journal of Physical Chemistry C*, vol. 114, no. 11, pp. 4973–4982, 2010.

[20] M. Salciccioli, S. Edie, and D. Vlachos, "Adsorption of acid, ester, and ether functional groups on pt: Fast prediction of thermochemical properties of adsorbed oxygenates via dft-based group additivity methods," *The Journal of Physical Chemistry C*, vol. 116, no. 2, pp. 1873–1886, 2012.

[21] B. Schweitzer, S. N. Steinmann, and C. Michel, "Group additivity for adsorbed polyols at a pt (111) surface under aqueous conditions: Dft precision on a spreadsheet," in *CatBior*, 2017.

[22] J. Lu, M. Wang, X. Zhang, A. Heyden, and F. Wang, "β-o-4 bond cleavage mechanism for lignin model compounds over pd catalysts identified by combination of first-principles calculations and experiments," *ACS Catalysis*, vol. 6, no. 8, pp. 5589–5598, 2016.

[23] O. Mamun, M. Saleheen, J. Q. Bond, and A. Heyden, "Investigation of solvent effects in the hydrodeoxygenation of levulinic acid to γ-valerolactone over ru catalysts," *Journal of Catalysis*, vol. 379, pp. 164–179, 2019.

[24] C. Minero, A. Bedini, and V. Maurino, "Glycerol as a probe molecule to uncover oxidation mechanism in photocatalysis," *Applied Catalysis B: Environmental*, vol. 128, pp. 135–143, 2012.

[25] S. Wang, Y. Zhang, and H. Liu, "Selective hydrogenolysis of glycerol to propylene glycol on cu–zno composite catalysts: Structural requirements and reaction mechanism," *Chemistry–An Asian Journal*, vol. 5, no. 5, pp. 1100–1111, 2010.

[26] J. Sun and H. Liu, "Selective hydrogenolysis of biomass-derived xylitol to ethylene glycol and propylene glycol on supported ru catalysts," *Green Chemistry*, vol. 13, no. 1, pp. 135–142, 2011.

[27] T. Prasomsri, M. Shetty, K. Murugappan, and Y. Román-Leshkov, "Insights into the catalytic activity and surface modification of moo 3 during the hydrodeoxygenation of lignin-derived model compounds into aromatic hydrocarbons under low hydrogen pressures," *Energy & Environmental Science*, vol. 7, no. 8, pp. 2660–2669, 2014.

[28] S. Kim, E. E. Kwon, Y. T. Kim, S. Jung, H. J. Kim, G. W. Huber, and J. Lee, "Recent advances in hydrodeoxygenation of biomass-derived oxygenates over heterogeneous catalysts," *Green Chemistry*, vol. 21, no. 14, pp. 3715–3743, 2019.

[29] M. Valter, E. C. Dos Santos, L. G. Pettersson, and A. Hellman, "Partial electrooxidation of glycerol on close-packed transition metal surfaces: Insights from first-principles calculations," *The Journal of Physical Chemistry C*, vol. 124, no. 33, pp. 17 907–17 915, 2020.

[30] Y. Kwon and M. T. Koper, "Electrocatalytic hydrogenation and deoxygenation of glucose on solid metal electrodes," *ChemSusChem*, vol. 6, no. 3, pp. 455–462, 2013.

[31] R. Kumar, N. Enjamuri, S. Shah, A. S. Al-Fatesh, J. J. Bravo-Suarez, and B. Chowdhury, “Ketonization of oxygenated hydrocarbons on metal oxide based catalysts,” *Catalysis Today*, vol. 302, pp. 16–49, 2018.

[32] C. Chang and A. J. Medford, “Classification of biomass reactions and predictions of reaction energies through machine learning,” *The Journal of Chemical Physics*, vol. 153, no. 4, p. 044 126, 2020.

[33] H. Gelernter, A. Sanders, D. Larsen, K. Agarwal, R. Boivie, G. Spritzer, and J. Searleman, “Empirical explorations of synchem,” *Science*, vol. 197, no. 4308, pp. 1041–1049, 1977.

[34] B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos, and T. Klucznik, “Chematica: A story of computer code that started to think like a chemist,” *Chem*, vol. 4, no. 3, pp. 390–398, 2018.

[35] E. Asgari and M. R. Mofrad, “Continuous distributed representation of biological sequences for deep proteomics and genomics,” *PloS one*, vol. 10, no. 11, e0141287, 2015.

[36] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, “Deepppi: Boosting prediction of protein–protein interactions with deep neural networks,” *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[37] L. Khalatbari, M. R. Kangavari, S. Hosseini, H. Yin, and N.-M. Cheung, “Mcp: A multi-component learning machine to predict protein secondary structure,” *Computers in biology and medicine*, vol. 110, pp. 144–155, 2019.

[38] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, and B. Yu, “Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure,” *Genomics*, vol. 111, no. 6, pp. 1839–1852, 2019.

[39] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, “Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets,” *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4150–4158, 2019.

[40] E. Parhizkar, H. Saeedzadeh, F. Ahmadi, M. Ghazali, and A. Sakhteman, “Partial least squares-least squares-support vector machine modeling of atr-ir as a spectrophotometric method for detection and determination of iron in pharmaceutical formulations,” *Iranian journal of pharmaceutical research: IJPR*, vol. 18, no. 1, p. 72, 2019.

[41] X. Shan, X. Wang, C.-d. Li, Y. Chu, Y. Zhang, Y. Xiong, and D.-Q. Wei, “Prediction of cyp450 enzyme–substrate selectivity based on the network-based label space

division method," *Journal of chemical information and modeling*, vol. 59, no. 11, pp. 4577–4586, 2019.

[42] B. Tang, S. T. Kramer, M. Fang, Y. Qiu, Z. Wu, and D. Xu, "A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility," *Journal of Cheminformatics*, vol. 12, no. 1, pp. 1–9, 2020.

[43] F. Gharagheizi, D. Tang, and D. S. Sholl, "Selecting adsorbents to separate diverse near-azeotropic chemicals," *The Journal of Physical Chemistry C*, vol. 124, no. 6, pp. 3664–3670, 2020.

[44] B. J. Bucior, N. S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O. K. Farha, N. Bagheri, and R. Q. Snurr, "Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks," *Molecular Systems Design & Engineering*, 2019.

[45] Z. Alizadeh and M. Mohammadizadeh, "Predicting electron-phonon coupling constants of superconducting elements by machine learning," *Physica C: Superconductivity and its Applications*, vol. 558, pp. 7–11, 2019.

[46] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, "Machine learning modeling of superconducting critical temperature," *npj Computational Materials*, vol. 4, no. 1, pp. 1–14, 2018.

[47] S. Le Roux and P. Jund, "Ring statistics analysis of topological networks: New approach and application to amorphous ges2 and sio2 systems," *Computational Materials Science*, vol. 49, no. 1, pp. 70–83, 2010.

[48] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058 301, 2012.

[49] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B*, vol. 89, no. 9, p. 094 104, 2014.

[50] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller, and E. K. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," *Physical Review B*, vol. 89, no. 20, p. 205 118, 2014.

[51] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, "Informatics-aided bandgap engineering for solar materials," *Computational Materials Science*, vol. 83, pp. 185–195, 2014.

[52] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, “Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids,” *Physical Review B*, vol. 89, no. 5, p. 054 303, 2014.

[53] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, and K.-i. Shimizu, “Machine learning for catalysis informatics: Recent applications and prospects,” *Acs Catalysis*, vol. 10, no. 3, pp. 2260–2297, 2019.

[54] A. O. Oliynyk and A. Mar, “Discovery of intermetallic compounds from traditional to machine-learning approaches,” *Accounts of chemical research*, vol. 51, no. 1, pp. 59–68, 2018.

[55] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, “Finding nature’s missing ternary oxide compounds using machine learning and density functional theory,” *Chemistry of Materials*, vol. 22, no. 12, pp. 3762–3767, 2010.

[56] A. Vojvodic, A. Hellman, C. Ruberto, and B. I. Lundqvist, “From electronic structure to catalytic activity: A single descriptor for adsorption and reactivity on transition-metal carbides,” *Physical review letters*, vol. 103, no. 14, p. 146 103, 2009.

[57] R. A. Van Santen, M. Neurock, and S. G. Shetty, “Reactivity theory of transition-metal surfaces: A brønsted- evans- polanyi linear activation energy- free-energy analysis,” *Chemical reviews*, vol. 110, no. 4, pp. 2005–2048, 2009.

[58] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, and H. Xin, “High-throughput screening of bimetallic catalysts enabled by machine learning,” *Journal of Materials Chemistry A*, vol. 5, no. 46, pp. 24 131–24 138, 2017.

[59] X. Ma, Z. Li, L. E. Achenie, and H. Xin, “Machine-learning-augmented chemisorption model for co2 electroreduction catalyst screening,” *The journal of physical chemistry letters*, vol. 6, no. 18, pp. 3528–3533, 2015.

[60] I. Takigawa, K.-i. Shimizu, K. Tsuda, and S. Takakusagi, “Machine-learning prediction of the d-band center for metals and bimetals,” *RSC advances*, vol. 6, no. 58, pp. 52 587–52 595, 2016.

[61] ——, “Machine learning predictions of factors affecting the activity of heterogeneous metal catalysts,” in *Nanoinformatics*, Springer, Singapore, 2018, pp. 45–64.

[62] A. J. Medford, C. Shi, M. J. Hoffmann, A. C. Lausche, S. R. Fitzgibbon, T. Bligaard, and J. K. Nørskov, “Catmap: A software package for descriptor-based microkinetic mapping of catalytic trends,” *Catalysis Letters*, vol. 145, no. 3, pp. 794–807, 2015.

[63] A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman, and H. J. Kulik, “Machine learning accelerates the discovery of design rules and exceptions in stable metal–oxo intermediate formation,” *ACS Catalysis*, vol. 9, no. 9, pp. 8243–8255, 2019.

[64] S. Gugler, J. P. Janet, and H. J. Kulik, “Enumeration of de novo inorganic complexes for chemical discovery and machine learning,” *Molecular Systems Design & Engineering*, 2020.

[65] A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman, and A. Aspuru-Guzik, “A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315 000 redox reactions,” *ACS central science*, vol. 5, no. 7, pp. 1199–1210, 2019.

[66] B. Sanchez-Lengeling, L. M. Roch, J. D. Perea, S. Langner, C. J. Brabec, and A. Aspuru-Guzik, “A bayesian approach to predict solubility parameters,” *Advanced Theory and Simulations*, vol. 2, no. 1, p. 1 800 069, 2019.

[67] A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, and G. A. Terejanu, “Prediction of adsorption energies for chemical species on metal catalyst surfaces using machine learning,” *The Journal of Physical Chemistry C*, vol. 122, no. 49, pp. 28 142–28 150, 2018.

[68] M. Reveil and P. Clancy, “Classification of spatially resolved molecular fingerprints for machine learning applications and development of a codebase for their implementation,” *Molecular Systems Design & Engineering*, vol. 3, no. 3, pp. 431–441, 2018.

[69] I. Muegge and P. Mukherjee, “An overview of molecular fingerprint similarity search in virtual screening,” *Expert Opinion on Drug Discovery*, vol. 11, no. 2, pp. 137–148, Dec. 2015.

[70] D. Weininger, “Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules,” *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[71] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[72] S. W. Benson, F. Cruickshank, D. Golden, G. R. Haugen, H. E. O’Neal, A. Rodgers, R. Shaw, and R. Walsh, “Additivity rules for the estimation of thermochemical properties,” *Chemical Reviews*, vol. 69, no. 3, pp. 279–324, 1969.

[73] S. Jaeger, S. Fulle, and S. Turk, “Mol2vec: Unsupervised machine learning approach with chemical intuition,” *Journal of chemical information and modeling*, vol. 58, no. 1, pp. 27–35, 2018.

[74] J. Romero, J. P. Olson, and A. Aspuru-Guzik, “Quantum autoencoders for efficient compression of quantum data,” *Quantum Science and Technology*, vol. 2, no. 4, p. 045 001, 2017.

[75] J. P. Janet and H. J. Kulik, “Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships,” *The Journal of Physical Chemistry A*, vol. 121, no. 46, pp. 8939–8954, 2017.

[76] M. Karthikeyan and A. Bender, “Encoding and decoding graphical chemical structures as two-dimensional (pdf417) barcodes,” *Journal of chemical information and modeling*, vol. 45, no. 3, pp. 572–580, 2005.

[77] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, “Automatic chemical design using a data-driven continuous representation of molecules,” *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[78] A. Bhattacharya and S. Shivalkar, “Re-tooling benson’s group additivity method for estimation of the enthalpy of formation of free radicals: C/h and c/h/o groups,” *Journal of Chemical & Engineering Data*, vol. 51, no. 4, pp. 1169–1181, 2006.

[79] C. H. Reynolds and R. C. Reynolds, “Group additivity in ligand binding affinity: An alternative approach to ligand efficiency,” *Journal of chemical information and modeling*, vol. 57, no. 12, pp. 3086–3093, 2017.

[80] M. Salciccioli, Y. Chen, and D. G. Vlachos, “Density functional theory-derived group additivity and linear scaling methods for prediction of oxygenate stability on metal catalysts: Adsorption of open-ring alcohol and polyol dehydrogenation intermediates on pt-based metals,” *The Journal of Physical Chemistry C*, vol. 114, no. 47, pp. 20 155–20 166, 2010.

[81] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, and Y. Jung, “Inverse design of solid-state materials via a continuous representation,” *Matter*, vol. 1, no. 5, pp. 1370–1384, 2019.

[82] B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik, and R. Q. Snurr, “Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis,” *Crystal Growth & Design*, vol. 19, no. 11, pp. 6682–6697, 2019.

[83] W. T. B. Kelvin, *Mathematical and physical papers*. University Press, 1890, vol. 3.

[84] D. Sholl and J. A. Steckel, *Density functional theory: a practical introduction*. John Wiley & Sons, 2011.

[85] H. Toffoli, “Principles of density functional theory,” *Lecture Scripts-Middle East Technical University*, 2016.

[86] P. Hohenberg and W. Kohn, “Inhomogeneous electron gas,” *Physical review*, vol. 136, no. 3B, B864, 1964.

[87] Y. Wang and R. G. Parr, “Construction of exact kohn-sham orbitals from a given electron density,” *Physical Review A*, vol. 47, no. 3, R1591, 1993.

[88] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, *et al.*, “Quantum espresso: A modular and open-source software project for quantum simulations of materials,” *Journal of physics: Condensed matter*, vol. 21, no. 39, p. 395 502, 2009.

[89] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. Hansen, *et al.*, “Electronic structure calculations with gpaw: A real-space implementation of the projector augmented-wave method,” *Journal of physics: Condensed matter*, vol. 22, no. 25, p. 253 202, 2010.

[90] S. R. Bahn and K. W. Jacobsen, “An object-oriented scripting interface to a legacy electronic structure code,” *Computing in Science & Engineering*, vol. 4, no. 3, pp. 56–66, 2002.

[91] A. J. Medford, *Computational Insight into catalytic hydrogenation of nitrogen and carbon monoxide*. Stanford University, 2015.

[92] J. Mercereau and R. Feynman, “Physical conditions for ferromagnetic resonance,” *Physical Review*, vol. 104, no. 1, p. 63, 1956.

[93] H. J. Monkhorst and J. D. Pack, “Special points for brillouin-zone integrations,” *Physical review B*, vol. 13, no. 12, p. 5188, 1976.

[94] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, “Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon,” *Physical Review B*, vol. 51, no. 19, p. 12 947, 1995.

[95] G. Seifert, D. Porezag, and T. Frauenheim, “Calculations of molecules, clusters, and solids with a simplified lcao-dft-lda scheme,” *International journal of quantum chemistry*, vol. 58, no. 2, pp. 185–192, 1996.

[96] H. Eschrig and I. Bergert, “An optimized lcao version for band structure calculations application to copper,” *physica status solidi (b)*, vol. 90, no. 2, pp. 621–628, 1978.

[97] G. Seifert, “Tight-binding density functional theory: An approximate kohn- sham dft scheme,” *The Journal of Physical Chemistry A*, vol. 111, no. 26, pp. 5609–5613, 2007.

[98] W. A. Harrison, *Electronic structure and the properties of solids: the physics of the chemical bond*. Courier Corporation, 2012.

[99] V. M. P. Koskinen, “Density-functional tight-binding for beginners,” *Computational Material Science*, vol. 47, p. 237, 2009.

[100] N. Yang, A. J. Medford, X. Liu, F. Studt, T. Bligaard, S. F. Bent, and J. K. Nørskov, “Intrinsic selectivity and structure sensitivity of rhodium catalysts for c2+ oxygenate production,” *Journal of the American Chemical Society*, vol. 138, no. 11, pp. 3705–3714, 2016.

[101] A. A. Peterson, “Global optimization of adsorbate–surface structures while preserving molecular identity,” *Topics in Catalysis*, vol. 57, no. 1-4, pp. 40–53, 2014.

[102] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, “The atomic simulation environment—a python library for working with atoms,” *Journal of Physics: Condensed Matter*, vol. 29, no. 27, p. 273 002, 2017.

[103] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[104] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[105] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis-a brief tutorial,” *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.

[106] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[107] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[108] M. Rellán-Piñeiro and N. López, "The active molybdenum oxide phase in the methanol oxidation to formaldehyde (formox process): A DFT study," *ChemSusChem*, vol. 8, no. 13, pp. 2231–2239, Jun. 2015.

[109] E. Miliordos, S. Caratzoulas, and D. G. Vlachos, "A periodic-dft study of retro-aldol fragmentation of fuctose on moo3," *Applied Catalysis A: General*, vol. 530, no. Supplement C, pp. 75–82, 2017.

[110] M. Orazov and M. E. Davis, "Tandem catalysis for the production of alkyl lactates from ketohexoses at moderate temperatures," *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. 11 777–11 782, 2015.

[111] M. Shetty, B. Buesser, Y. Román-Leshkov, and W. H. Green, "Computational investigation on hydrodeoxygenation (hdo) of acetone to propylene on α-moo3 (010) surface," *The Journal of Physical Chemistry C*, vol. 121, no. 33, pp. 17 848–17 855, 2017.

[112] D. Mei, A. M. Karim, and Y. Wang, "Density functional theory study of acetaldehyde hydrodeoxygenation on moo3," *The Journal of Physical Chemistry C*, vol. 115, no. 16, pp. 8155–8164, 2011.

[113] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, "Density functionals for surface science: Exchange-correlation model development with bayesian error estimation," *Physical Review B*, vol. 85, no. 23, p. 235 149, 2012.

[114] L. Bengtsson, "Dipole correction for surface supercell calculations," *Physical Review B*, vol. 59, no. 19, p. 12 301, 1999.

[115] S. Najmi, M. Rasmussen, G. Innocenti, C. Chang, E. Stavitski, S. R. Bare, A. J. Medford, J. W. Medlin, and C. Sievers, "Pretreatment effects on the surface chemistry of small oxygenates on molybdenum trioxide," *ACS Catalysis*, vol. 10, no. 15, pp. 8187–8200, 2020.

[116] J. Zhang, L. Lin, J. Zhang, and J. Shi, "Efficient conversion of d-glucose into d-sorbitol over mcm-41 supported ru catalyst prepared by a formaldehyde reduction process," *Carbohydrate research*, vol. 346, no. 11, pp. 1327–1332, 2011.

[117] M. Niu, Y. Hou, W. Wu, S. Ren, and R. Yang, "Successive c1–c2 bond cleavage: The mechanism of vanadium (v)-catalyzed aerobic oxidation of d-glucose to formic acid in aqueous solution," *Physical Chemistry Chemical Physics*, vol. 20, no. 26, pp. 17 942–17 951, 2018.

[118] J. Behler, “Constructing high-dimensional neural network potentials: A tutorial review,” *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1032–1050, 2015.

[119] E. Kocer, J. K. Mason, and H. Erturk, “A novel approach to describe chemical environments in high-dimensional neural network potentials,” *The Journal of chemical physics*, vol. 150, no. 15, p. 154 102, 2019.

[120] N. Artrith and J. Behler, “High-dimensional neural network potentials for metal surfaces: A prototype study for copper,” *Physical Review B*, vol. 85, no. 4, p. 045 439, 2012.

[121] N. Artrith, T. Morawietz, and J. Behler, “High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide,” *Physical Review B*, vol. 83, no. 15, p. 153 101, 2011.

[122] Y. Huang, J. Kang, W. A. Goddard III, and L.-W. Wang, “Density functional theory based neural network force fields from energy decompositions,” *Physical Review B*, vol. 99, no. 6, p. 064 103, 2019.

[123] A. Rodriguez, Y. Liu, and M. Hu, “Spatial density neural network force fields with first-principles level accuracy and application to thermal transport,” *Physical Review B*, vol. 102, no. 3, p. 035 203, 2020.

[124] X. Lei and A. J. Medford, “A universal framework for featurization of atomistic systems,” *arXiv preprint arXiv:2102.02390*, 2021.

[125] J. Behler, “Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations,” *Physical Chemistry Chemical Physics*, vol. 13, no. 40, pp. 17 930–17 955, 2011.

[126] Y. Nakadaira, Y. Kobayashi, and H. Sakurai, “Mass spectroscopy of organosilicon compounds: Fragmentation patterns of linear organopolysilanes,” *Journal of Organometallic Chemistry*, vol. 63, pp. 79–83, 1973.

[127] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, “To address surface reaction network complexity using scaling relations machine learning and dft calculations,” *Nature communications*, vol. 8, no. 1, pp. 1–7, 2017.

VITA