Unifying Chemistry and Machine Learning for the Study of Noncovalent Interactions

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Jacob Townsend

December 2020

Acknowledgments

I would like to thank my advisor and mentor, Kostas (Prof. Konstantinos Vogiatzis), for his constant encouragement and optimism. His professional and personal commitment has provided me the opportunity to pursue this research and develop as a professional and individual. I am forever grateful to Kostas for the freedom to pursue the research I found interest in and to keep me motivated. Whenever disappointment set in, I could always count on Kostas to keep a lighthearted and cheery attitude to remind me that we don't do "science in a vaccuum". He helped me see the value in my research and made the entire experience possible. I am excited to have contributed to the start of the work of the Vogiatzis lab, even if it just a small impact on the bright future that awaits!

I would also like to thank my committee members, Prof. Brian Long, Prof. Robert Hinde, and Prof. Manolis Doxastakis. I have greatly appreciated your constructive feedback and support throughout this journey. Thanks to my fellow research group members who have helped me through this time. Dr. Rajeshkumar Thayalan was a great friend and help throughout my career. I also have to thank Justin Kirkland, who was was a great friend that I could always count on for a laugh.

Thanks to my girlfriend, Halie Miller, for being a constant source support and happiness. She (somehow) put up with me and never failed to bring a smile to my face. Having moved away from my family to pursue my graduate studies, it has been amazing that I have gotten the opportunity to be a part of yours. Thank you Bruce, Rhonda, and Samantha for accepting me into your family and providing me another home away from home.

Lastly, and most importantly, I owe all my success to my family. They have made me into the person I am today. Words really can't do it justice. I have been blessed with an amazing family. Mom and Dad, all I can really say is "thank you". And I could not mention

my family without thanking my brother, Jamie, who has been there from since day one. There really are not any words that can describe how lucky I am to have my family, but I am forever grateful.

Abstract

Gas separations are in great demand for carbon emission reduction, natural gas purification, oxygen isolation, and much more. Many of these separations rely on cost-prohibitive methods such as cryogenic distillation or strong-binding solvents. As a result, novel materials are being developed to subvert the energetic expense of gas separation processes. These studies focus on improving the performance of alternative materials, including (but not limited to) metal-organic frameworks, covalent organic frameworks, dense polymeric membranes, porous polymers, and ionic liquids.

In this work, the atomistic effects of functional units are explored for gas separations processes using electronic structure theory and machine learning. In particular, the crucial role of noncovalent interactions for these processes is investigated. Specifically, the atomistic effects are studied for the separation of CO_2 and N_2 using vinyl-addition polynorbornenes to interpret how structural modification can contribute to macroscopic permeability properties. An in-depth examination elucidated the role of functionalizing boranes to increase N_2 interaction for the selective separation from O_2. Here, a small-scale screening was performed to append electron-withdrawing moieties with boranes to develop a stronger Lewis acid to interact with N_2. As a next step, electronic structure theory and machine learning were employed to suggest promising functional units for CO_2/N_2 separations. Novel techniques are developed to perform a high-throughput screening of a large molecular database with great accuracy. Lastly, the efficiency and applicability of accurate calculations for noncovalent interactions was improved by reducing computational expense of coupled cluster calculations using machine learning. This work introduces novel methods to study noncovalent interactions crucial for gas separations through the incorporation of accurate electronic structure theory methods and advanced machine learning techniques.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Polymeric Membranes for Gas Separations

Carbon dioxide (CO_2) exists naturally in the atmosphere at small concentrations and is essential in sustaining life.[1] It behaves as a greenhouse gas by emitting and absorbing thermal radiation, and therefore contributes to the global greenhouse effect, a natural process which warms the Earth's surface.[2,3] While this process is necessary, in recent decades atmospheric CO_2 concentration has increased at an unprecedented and alarming rate, contributing to extreme-weather events such as floods, droughts, severe storms, and has led to rising sea levels and increasing ocean acidity.[4] In an effort to reduce anthropogenic contributions to global climate changes, regulatory agencies have set targets to reduce CO_2 emissions over the remainder of the century.[5,6] These initiatives target power generation sectors which are responsible for approximately 40% of global CO_2 emissions,[7] of which approximately two-thirds of U.S.-based emissions derives from the combustion of fossil fuels such as coal and natural gas.[8] In order to meet current climate initiatives, there has been a significant focus on the area of carbon capture and storage (CCS) to meet climate change targets and provide low-carbon energy.[9]

The most industrially relevant gas carbon capture processes are cryogenic separation and absorption processes. Cryogenic distillation for the separation of CO_2 from the primary atmospheric gas, N_2, entails lowering the temperature to condense CO_2 into a liquid, which occurs at extremely low temperatures and yields high energy costs.[10] Sorbents for absorption

of CO_2, typically liquid amines, often have substantial capital, operating, and environmental implications due solvent regeneration after abstraction and secondary pollution.[11] Due to the energy costs, practical application of the previous approaches may reduce efficiency of power production, leading to subsequent increases in prices of electricity, which necessitates the development of more efficient alternatives.[12] One such method is the use of polymeric membrane technology, which has emerged as a primary tool for effective CO_2 separation due to its passive separation mechanism, compact and environmentally-friendly design, and cost-effective solution toward gas separations.[13,14] The ideal membrane is one that is both permeable and selective towards a CO_2. However, attempts to increase selectivity are typically hindered by subsequent reductions in permeability, and vice versa. This trade-off in performance between permeability and selectivity is best observed on a logarithmic plot known as a Robeson plot and is shown for CO_2 and N_2 in Figure 1.1.[15] Further development of polymer membranes that disobey this trade-off is key to the realization of CCS targets.

1.2 Noncovalent Interactions of CO_2 with Functional Groups

One strategy towards enhancing the separation of CO_2 with a material is to introduce functional groups which modify the interaction strength, thereby improving capacity and/or selectivity. Fundamentally, non-interacting carbon dioxide is a linear molecule, and therefore does not have a dipole moment; however, the carbon atom is electron deficient due to the electronegative oxygen atoms. This allows Lewis bases to provide stabilizing electrostatic interactions and can induce a dipole moment by attracting the carbon atom and distorting its linear geometry.[16,17] Typically, the strength of the Lewis base determines the extent of interaction of the functional group and CO_2, where strong interactions are considered as chemisorption (greater than 20 kcal mol^{-1}). Absorption using amine based sorbents such as monoethanolamine (MEA) have strong interactions of approximately 25 kcal mol^{-1}.[18,19] MEA reacts reversibly to produce carbamates, but the strong interactions are responsible for the large energy penalty to regenerate the sorbent.[20,21] To reduce the energetic penalty

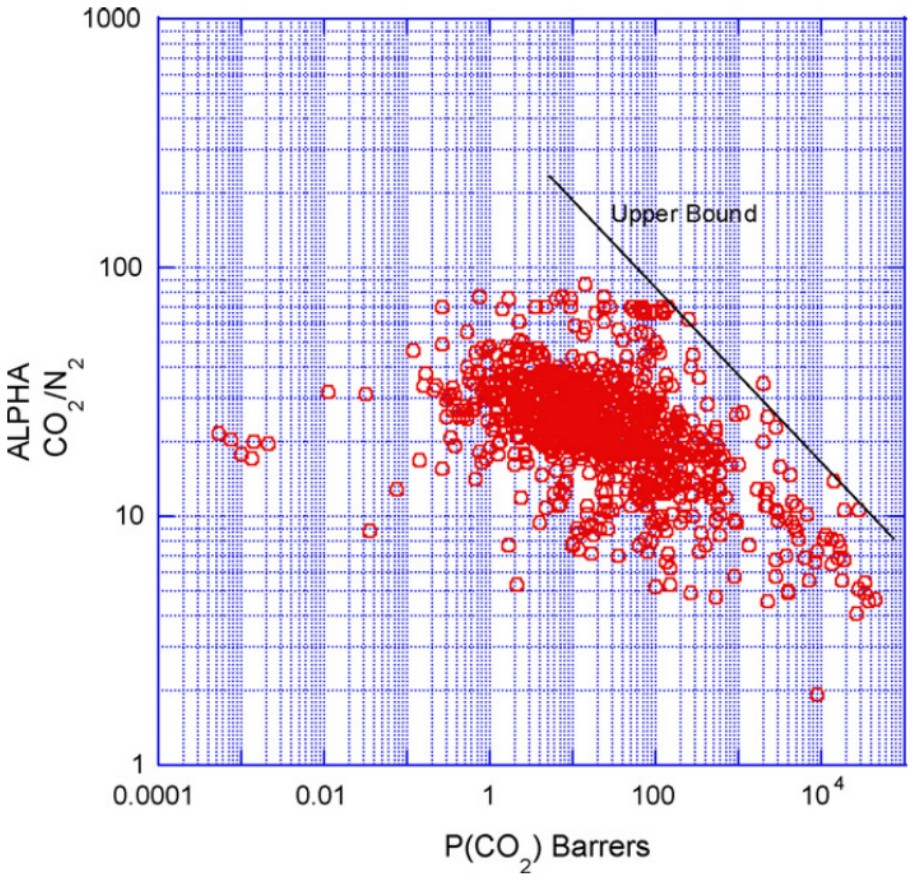


Figure 1.1: 2008 Robeson selectivity versus permeability plot for the CO_2 / N_2 gas pair.[15]

in the use of sorbents for CO_2 separations, it is of interest to investigate methods that utilize weaker binding motifs.

As an alternative to strong-binding chemisorption found in MEA, many separation approaches including polymeric membranes rely on weaker interactions based on the reversible binding of CO_2, leading to easier desorption. These physisorption-based separation methods are characterized by lower binding energies and longer interatomic distances between CO_2 and the binding site, which are comprised of a combination of electrostatic and van der Waals interactions. Materials where physisorption is a primary contributor towards separation includes metal-organic frameworks, covalent organic frameworks, dense polymeric membranes, porous polymers, and ionic liquids.[22] Introducing functional groups to optimize the interaction strength is a strategy towards increasing selectivity, where approximately 11 kcal mol^{-1} has been set as a target for optimal binding.[22,23] However, in-depth studies on functional groups for CO_2 binding have fallen short of this benchmark, with the best performing system featuring imidazopyridamine units with an interaction of approximately 6 kcal mol^{-1}.[16,17,24] These insights gained from theory have been used to enhance understanding and suggest units which have been successfully implemented in materials such as polymer membranes and MOFs.[24,25] So far, the analysis of functional groups has been limited to several dozen cases, but the success of this approach highlights the utility of electronic structure theory calculations for CO_2 separation and material functionalization.

1.3 Theoretical Studies for Gas-Materials Interactions

1.3.1 Electronic Structure Theory Calculations for the Study Atomistic Effects in Polymers

The study of polymeric units with electronic structure theory can shed light on the underlying physical phenomena responsible for bulk characteristics. A method that accurately treats van der Waals interactions is necessary to study CO_2 interactions in polymers where gas-binding is dictated by weak noncovalent interactions. Wave function-based methods

offer a systematic and accurate approach in calculating these noncovalent interactions, where coupled-cluster singles and doubles with perturbative triples [CCSD(T)] is the gold standard. However, the applicability is significantly reduced due to N^7 scaling with respect to the number of basis functions, significantly reducing its applicability. Møller-Plesset perturbation theory (MP2) offers a more computationally practical method with its N^5 scaling, and provides a reasonable treatment of noncovalent interactions by considering two-electron excitations, but tends to overestimate the strength of many of weakly bound systems.[26,27] As an alternative, density functional theory (DFT) methods offer significantly reduced computationally costs (N^3-N^4) by using the electron density to approximate correlation and exchange energies provided by more expensive wave function methods. The mapping of electron density to the correlation and exchange terms is provided by the density functional.[28] Hundreds of functionals have been developed and each utilizes a unique set of parameters and considerations.[29] As a result, the treatment of long-range interactions is not systematic and varies functional to functional. In fact, many functionals neglect crucial dispersion interactions and rely on semi-empirical corrections for quantitative results.[30–34] While DFT and MP2 are routinely applied due to their efficiency, CCSD(T) remains the gold standard for the accurate consideration of noncovalent interactions.

In order to utilize electronic structure calculations to study atomistic effects of polymeric functional units, the materials must be substantially truncated. In practice, a polymeric material is limited to a number of functional repeat units. Even with this truncation, geometry optimizations on polymer membrane units may to converge to local minima on the potential energy surface. As an illustrative example, the many binding sites and their respective binding energies are shown for 7-methyl-1H-imidazo[4,5-b]pyridine in Figure 1.2. This shows multiple calculations are necessary to accurately determine the optimal binding of a gas to a molecular unit, further increasing the complexity and computational expense of these applications.

1.3.2 Data-Driven Exploration of Functional Groups for Gas Separations

Although it is well-accepted that the thermophysical and kinetic properties of the polymer membranes dictate the performance of the material for gas separation, it is often unclear how chemical modification will alter that performance.[9] Traditionally, membrane discovery has been guided by chemical intuition and experiment-driven validation centered on functionalizing materials with moieties that interact selectively with certain gases.[14,22,35–37] This approach is resource intensive, and thus, materials discovery remains a challenge as gas separations are governed by complex structure-property relationships[38–42] and the vastness of chemical space.[43] In light of these challenges, materials discovery is undergoing a paradigm shift, and it is increasingly prevalent to have experiment and theory work together to guide predictions and discovery.[44–49] Although computationally driven-studies to suggest moieties have been performed for the study CO_2 interactions with functional groups,[16,24,50,51] so far they have been limited in scope due to the computational burden of performing rigorous electronic structure theory calculations necessary for the accurate calculation of the noncovalent interactions responsible for binding and separating CO_2.

Modern machine learning techniques are revolutionizing all facets of computational sciences including chemistry. Recent studies have led to significant advances in drug and catalyst design.[49,52–58] However, data-driven approaches to predict favorable noncovalent interactions for gas separations have not seen significant development. The scarcity of such methods is likely due to a) level of theory required for the accurate calculation of such interactions and b) the challenges in generating training sets for such methods. Although highly accurate calculations such as coupled-cluster theory represent the standard for such noncovalent interactions, this is typically not feasible for large-scale implementations due to its significant computational costs. Additionally, the determination of binding sites and strengths requires many calculations for the evaluation of complex potential energy surfaces (see Figure 1.2). Even with accurate reference data, the consideration of how to describe the molecules for machine learning process poses a significant challenge. Therefore, the development of ML-based methods molecular screenings to omit the time-consuming

electronic structure calculations and conformation searches is necessary to accelerate the discovery of higher performing CO_2-binding functionalities.

1.3.3 Machine-Learning for Efficient Quantum Chemical Methods for Noncovalent Interactions

While machine learning has made a significant impact on many areas of chemistry, its role in *ab initio* methods is a relatively new area of study only gaining traction in recent years.[59] Among these studies, significant efforts have been made on the development of DFT functionals and semi-empirical dispersion corrections.[46,60–67] Despite these advances, few studies focus on wave function methods that are necessary to accurately calculate electronic correlation energy crucial for noncovalent interactions. Wave function (post-Hartree-Fock) methods are constructed such that they are systematically improvable as the number of electronic configurations considered and one electron-basis (as the size of the basis set is extended towards the complete basis set [CBS] limit) are extended, which is shown in Figure 1.3. Within a given basis set, solution of the electronic, nonrelativistic, time-independent Born-Oppenheimer Schrödinger equation yields the exact electronic energy by considering all electronic configurations as provided by full configuration interaction (FCI):

$$|\Psi_{\mathrm{FCI}}\rangle = c_0|\Psi_0\rangle + \sum_{i,a} c_i^a|\Psi_i^a\rangle + \sum_{i<j,a<b} c_{ij}^{ab}|\Psi_{ij}^{ab}\rangle + \sum_{i<j<k,a<b<c} c_{ijk}^{abc}|\Psi_{ijk}^{abc}\rangle + \ldots \tag{1.1}$$

where i, j, k represent occupied molecular orbitals, a, b, c represent virtual orbitals, Ψ_0 represents the reference wave function, often a Hartree-Fock determinant, and Ψ_i^a represents the Slater determinant formed by replacing electron in orbital i with orbital a and Ψ_{ij}^{ab} corresponds to the electronic configuration of replacing electrons from orbitals i, j to a, b, and so on. However, this expansion is intractable for all but small systems with truncated basis sets; thus, most wave function methods offer approximations by truncations to the excitation space of FCI (see Figure 1.3). While it is outside the scope of this discussion to derive the full formalism herein, coupled-cluster determines CI coefficients by products of clusters amplitudes, which is exact without truncation of the cluster operator, but is typically truncated for computational considerations.[68] Even with truncation, CC methods

remain cost-prohibitive for many systems of interest. Therefore, the development of ML methods based on the underlying electronic structure can subvert the computational costs and extend the applicability of the method to systems of chemical relevance including gas interactions with functional polymer units without compromising its accuracy.

In recent years, there have been several studies to extend the applicability of FCI calculations using neural networks, but the high cost still prohibits their use in applications exceeding a few atoms.[69–72] As a result, methods to estimate CC energies have been set as a target of many ML studies to accurately and rapidly evaluate noncovalent interactions. In particular, HF properties have been used to predict CC energy at a fraction of the computational expense.[73–77] While these ML-CC methods have reduced cost, they have not shown reliable chemical transferability with respect to system size and composition and rely on expensive machine-learning techniques. As an intermediate step, methods that use ML representations based on correlated methods such as MP2 may offer more suitable performance for predicted CC properties while maintaining a significant cost-savings. While a representation based on MP2 has been introduced, it did not demonstrate chemical transferability between systems.[78] While it has been known for decades that localized electronic correlation methods offer transferable correlation contributions,[79–82] it has not been well-explored for machine-learning CC energies and is an exciting avenue for predicting accurate and transferable electronic correlation energy predictions.

1.4 Dissertation Overview

The objective of work presented herein is to contribute to the development and understanding of gas separation processes through the study noncovalent interactions via the use of electronic structure theory calculations and machine learning. It is our goal to extend beyond validation; we would like to perform rigorous high-throughput screening to identify key moieties for rational design of the next generation of CCS materials. Additionally, we would like to develop methods that can extend the applicability of accurate quantum chemical calculations by reducing their computational expense through insightful machine learning approaches. In the following work, these ideas will be carried through by identifying

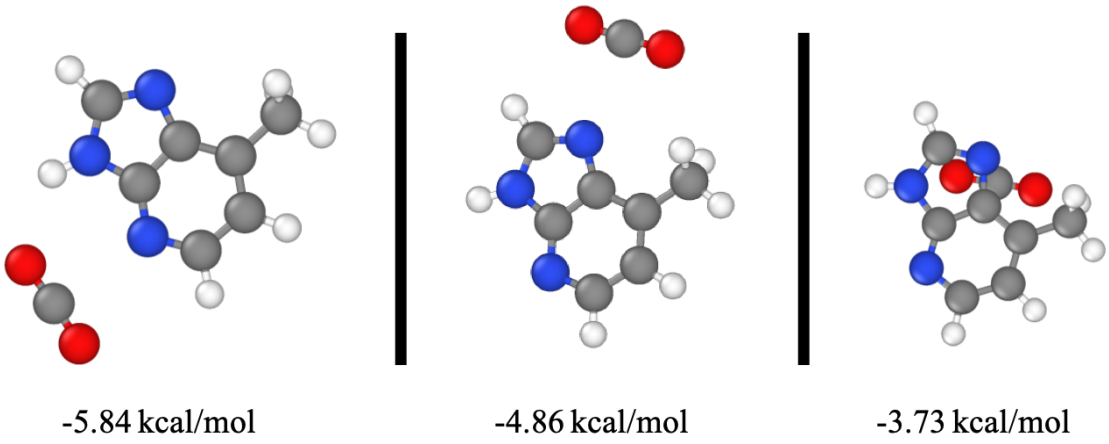


Figure 1.2: Binding Energies of CO_2 with 7-methyl-1H-imidazo[4,5-b]pyridine at several binding sites at the PBE0-D3(BJ)/def2-TZVPP level of theory

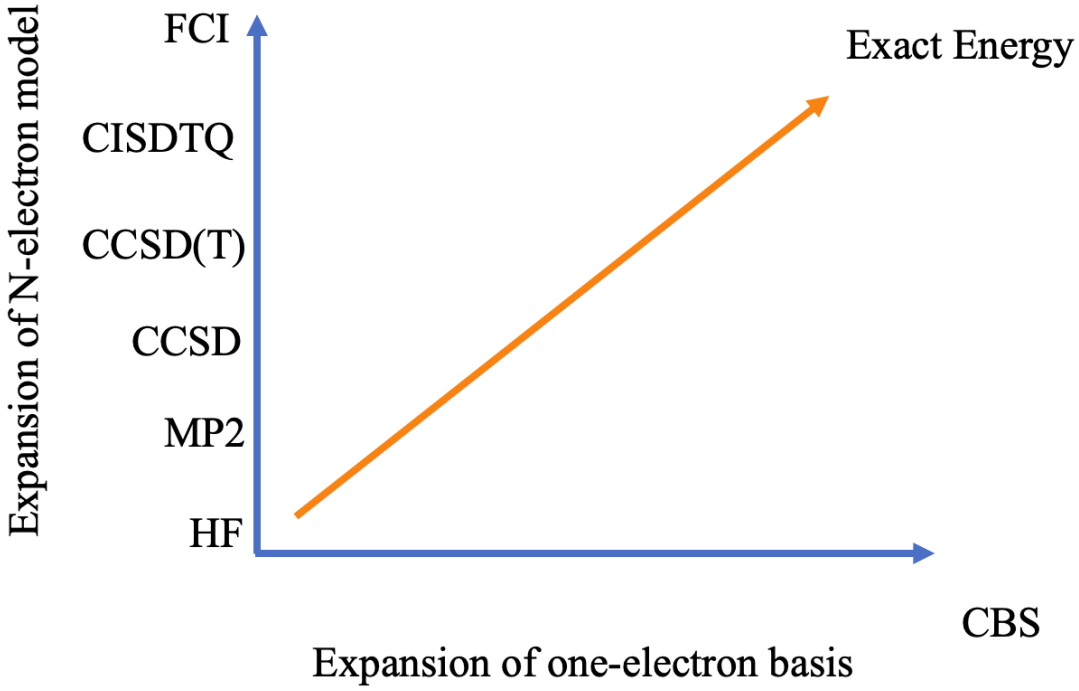


Figure 1.3: Systematic expansion to the exact solution of the electronic, nonrelativistic, time-independent Born-Oppenheimer Schrödinger Equation. The expansion of the N-electron models is located on the y-axis and one electron basis is located on the x-axis.

novel ways to screen databases for CO_2 interactions and the development of techniques to calculate highly accurate coupled cluster theory energetics at reduced cost.

As a first step, we investigated the role of atomistic effects to known polymer systems in Chapter 2. In particular, we examined gas interactions with polymeric repeat units and the role of competing intermolecular interactions between repeat units on observed gas separation properties. In this section, we addressed the challenge of the possibility of sampling bias and converging to geometric local minima in geometry optimizations by implementing an automated conformation search that is highlighted in Figure 1.4. Molecular dynamics simulations can be performed and optimized to generate many conformations for starting points for DFT calculations. DFT geometry optimizations are performed on different conformations, and the most energetically stable geometry is considered as the "global" minimum on the potential energy surface. This approach allowed for efficient study on numerous polymers for the understanding of atomistic effects contributing to material properties. This technique was used in determining the role of functionalization of vinyl-addition polynorbornenes on gas separation performance. Next, the role of hydrogen bonding in a self-healing urea-functionalized poly(dimethyl siloxane) membrane is studied along with the impact of the urea units on gas separation. Lastly, a study was performed on the polyvinyl alcohol-boronate polymer for the separation of aqueous anions to elucidate anion binding and experimentally observed spectroscopic peaks.

In Chapter 3, electronic structure theory calculations were used in a small-scale study to suggest borane-functionalized moieities for the difficult separation of N_2 and O_2. In-depth analysis of the role of Lewis-bases on boranes are explored for increasing the interaction with nitrogen. An extensive benchmark of DFT methods was performed to determine an accurate level of theory for these interactions. Afterwards, the effects of utilizing different electron-withdrawing functional groups was explored. This chapter highlights a small-scale study to suggest borane-functionalized moieities for the difficult separation and serves as a precursor to the data-driven approaches that follow.

Chapter 4 discusses the development of a method to perform high-throughput screening using a novel representation based on an applied branch of mathematics known as persistent homology for the discovery of effective CO_2 binding functional groups. The

novel representation was used to perform a large-scale screening of a database of organic molecules to discover functional groups for enhanced binding and was found to produce accurate results with very small training set sizes. Additionally, a ML-based screening was performed without the need of conformational searches by including only the representation of the bare functional unit at the time of testing (see Figure 1.5). Based on this screening new functional units featuring amino- and hydroxyl- units are suggested.

Chapters 5 and 6 introduce methods that can extend the applicability of accurate quantum chemical calculations by reducing their computational expense through insightful machine learning approaches. Our goal was to introduce methods that can predict CC energies based on MP2, a prerequisite to beginning a CC equation. To understand our approach for this, let us first examine the coupled cluster energy expression which is obtained by projection against the Hartree-Fock state

$$E_{\mathrm{CC}} = \langle \Psi_{\mathrm{HF}} | \hat{H} | \Psi_{\mathrm{CC}} \rangle \tag{1.2}$$

where E_{CC} is the coupled cluster energy, Ψ_{CC} is the CC wave function, and Ψ_{HF} is the Hartree-Fock wave function. Because the Hamiltonian is only a two-particle operator, only the only two particle promotion coefficients c_{ij}^{ab} (see Equation 1.1) directly contribute to E_{CC} with a HF reference. At any level of CC truncation, the CI c_{ij}^{ab} are composed of one electron amplitudes t_i^a and two-electron amplitudes t_{ij}^{ab}. Therefore, if one can map from a c_{ij}^{ab} from a computationally inexpensive level of theory to the c_{ij}^{ab} of a higher level of theory, one can capture the higher accuracy at reduced cost. More concretely, our goal is to use MP2 as the starting guess

$$c_{ij[\mathrm{CC}]}^{ab} = c_{ij[\mathrm{MP2}]}^{ab} + \Delta_{\mathrm{ML}} \tag{1.3}$$

where $c_{ij[\mathrm{CC}]}^{ab}$ are the selected level of CC coefficients, $c_{ij[\mathrm{MP2}]}^{ab}$ are the MP2 coefficients, and Δ_{ML} are the terms for ML to learn. An inherent assumption of this model is that there exists a connection between levels of theory that is systematic and can be learned. Equation 1.3 allows one to map between levels of theory, but does not suggest how to most efficiently learn Δ_{ML} terms. The exploration of how to perform this mapping is the topics of Chapters 5 and 6, where Chapter 5 introduces this formalism to learn each $c_{ij[\mathrm{CC}]}^{ab}$ independently, and

Figure 1.4: A flowchart examining a strategy to remove expert bias from conformational searches. Molecular dynamics simulations can be used to generate conformations which are optimized by DFT to find global minima on the potential energy surface.

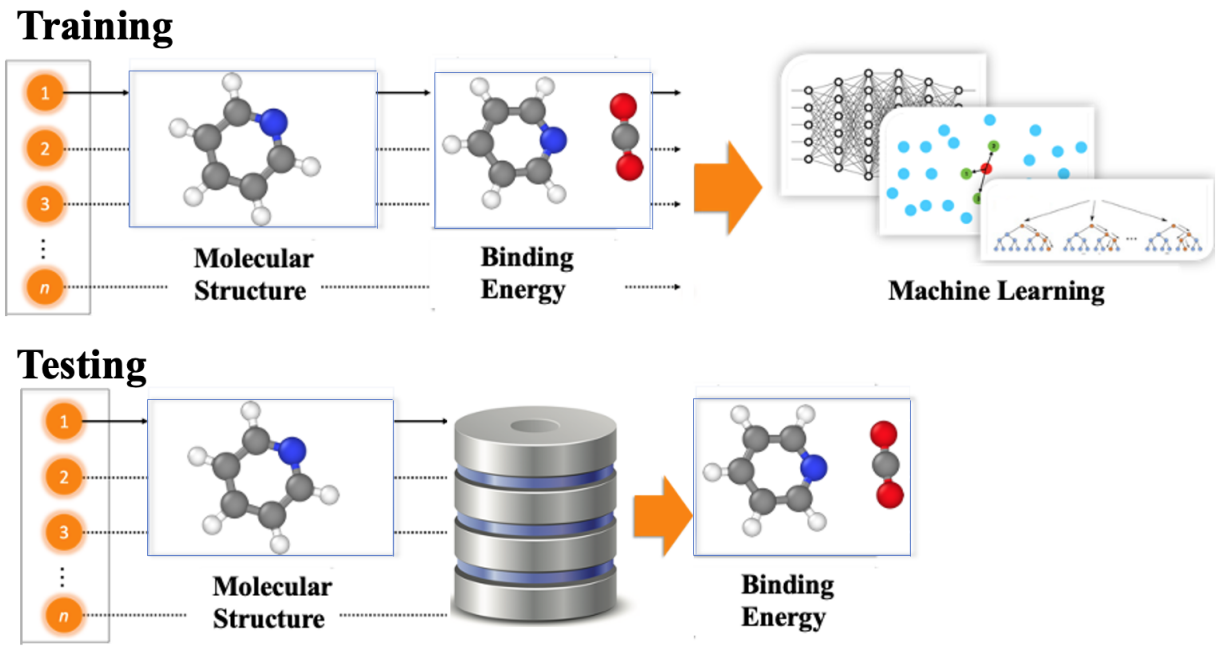


Figure 1.5: Diagram depicting the flow of training and testing for ML-based CO_2 interactions. For training, both the structure of the molecule and the interaction energy are provided. For testing, only the structure of the molecule is provided to predict the interaction energy, which allows the circumvention of costly conformational searches.

Chapter 6 explores how to make this expression robust by exploring locality and learning large collections of $c^{ab}_{ij[\mathrm{CC}]}$ based on occupied pair-orbitals. In this section, the electronic structure properties of the ML-based feature set are explored. Further, the transferability between chemical systems based on varying functional units and the expansion of system size is reported. Ultimately, the data-driven approach towards improved electronic structure calculations and high-throughput screening presented in this dissertation offer an approach towards accurate chemical discovery of CO_2-philic functionalizations for the next-generation of separation media. Finally, Chapter 7 is used to summarize the work presented in the dissertation and draw some insight on the future directions for role of data-driven chemistry in the role of gas separations.

Chapter 2

Gas Separations with Polymeric Membranes

The following chapter examines applications where electronic structure theory calculations have been used to better understand experimental observation of polymeric systems used for separations. Within each section, the focus will be in explaining the context for the material, followed by the experimental properties reported. Each of the following sections corresponds to a published paper, where the author's (Jacob Townsend) contribution was to perform electronic structure calculations. Each system will be described and the atomistic effects as determined by the electronic structure calculations to the understanding of macroscopic properties is explored.

2.1 Polynorbornenes

Disclosure

The following section contains a modified version of work from an article published in *Macromolecules* in collaboration with the group of Professor Brian Long and Aramco Services Company.[14] Two graduate students have contributed to the work that is summarized in section 2.1.1, which includes Christopher R. Maroon and Kevin R. Gmernicki. Jacob

Townsend performed the DFT calculations used in the analysis of sections 2.1.2 and 2.1.4 under the guidance of Professor Konstantinos D. Vogiatzis.

2.1.1 Introduction

Polymer membranes development is one of the areas being pursued for the realization of CCS performance targets, and offer many practical advantages for separation over amine-based sorbents such as a passive separation mechanism, reduced environmental concerns from toxic amine precursors, and reduced energy and operational expense.[83] However, polymeric membrane performance is typically hindered by an observed trade-off where increases in selectivity are usually accompanied by decreases in permeability, and vice versa. Therefore, further developments are needed to find polymers that disobey this relationship and can surpass "upper bound" limits of performance (see Figure 1.1).

Vinyl-addition polynorbornenes (VAPNBs) offer an attractive route as they have many of the traits of highly permeable membranes, such as bulky nonpolar substitutents and rigid backbones.[84] In line with this, VAPNBs with silane functionality were developed and exhibited very high permeability but overall low selectivity for CO_2/N_2.[85] Consistent with the strategies discussed in section 1.2, Lewis base functionalities were incorporated into the polymers to enhance the interaction, and thus selectivity for CO_2.[36,86,87] Having successfully developed alkoxysilyl VAPNBs to greatly improve selectivity towards CO_2, the Long group turned their interest to incorporating tris(2-methoxyethoxy)silane functionalities into VAPNBs given the successful use of glycol-like functionalities in other polymers such as poly(ethylene oxide).[88]

2.1.2 Tris(2-methoxyethoxy)silane Functionality

To investigate the effects of the addition of the alkoxysilanes moieities, 6 unique polymer membranes were synthesized using monomers A, B, and C shown in Figure 2.1. Polymers A.P1, A.P5, and A.P6 were synthesized as homopolymers of monomers A,B, and C, respectively. A.P2-A.P4 are were synthesized as copolymers with ratios of monomers of A:B of 79:21, 59:41, and 43:57, respectively, and the permeability of these polymers is shown

in Table 2.1. The overall performance of the synthesized membranes for the separation of CO_2 from N_2 is illustrated in Figure 2.2 relative to the 2008 upper bound. It is evident that increasing the tris(2-methoxyethoxy)silane functionalities (Monomer B) within the polymers yields a systematic approach towards the Robeson upper bound.

Figure 2.2 shows that the performance enhancements associated with incorporation of Monomer B disobeys the permeability/selectivity trade-off by increasing selectivity with only small decreases in permeability, and thus nears the Robeson upper bound. While the CO_2 permeability of polymer A.P5 is reduced approximately 20% relative to A.P1, the selectivity increases by 123%, which is responsible for the deviation from the usual trade-off. Examining the individual CO_2 and N_2 permeability values more closely shows although there was only a 20% decrease in CO_2 permeability, there was nearly a 275% decrease in N_2 permeability. The understanding of how the incorporation of Monomer B into the polymer matrix could introduce what could be described as an N_2-phobic effect is particularly interesting, especially considering N_2 is an inert gas. To better understand this unusual phenomenon, we performed electronic structure theory computations to aid in an explanation of polymer A.P5's enhanced CO_2/N_2 selectivity in comparison to A.P1. Because experimental results showed that gas solubility within polymers A.P1-A.P5 is primarily dominated by Henry's law dissolution, both polymer-gas and polymer-polymer interactions are known to play key roles in gas diffusion.[89] To study these interactions, saturated monomers A and B were used as small molecule surrogates for polymers A.P1 and A.P5, respectively, and monomer-gas and monomer-monomer interactions were computed.[90–92]

2.1.3 Computational Details

Quantum chemical calculations were performed to elucidate the atomistic effects between individual repeat units of polymers A.P1, A.P5, and A.P6 and the respective gases (CO_2 and N_2). To evaluate the strength and behavior of the interactions, monomer-monomer and monomer-gas interactions were calculated using density functional theory (DFT). Because of the high flexibility of the side chains of monomers A and B, many competitive conformations were considered since conventional DFT geometry optimizations may converge to local minima on the potential energy surface (PES). Therefore, molecular dynamics

Table 2.1: Permeabilities for Polymers A.P1-P6, given in Barrer along with selectivities (CO_2/N_2)

Polymer	CO_2 Permeability	N_2 Permeability	Selectivity α
A.P1	936.6 (± 14.9)	57.3 (± 2.3)	16.4 (± 0.4)
A.P2	868.8 (± 126)	40.5 (± 8.2)	21.6 (± 1.3)
A.P3	733.3 (± 66.7)	27.5 (± 4.2)	26.8 (± 1.6)
A.P4	700.0 (± 79.6)	19.6 (± 2.0)	33.7 (± 2.9)
A.P5	754.8 (± 40.5)	20.7 (± 2.6)	36.7 (± 2.8)
A.P6	4371 (± 855)	326.8 (± 7.7)	13.4 (± 2.4)

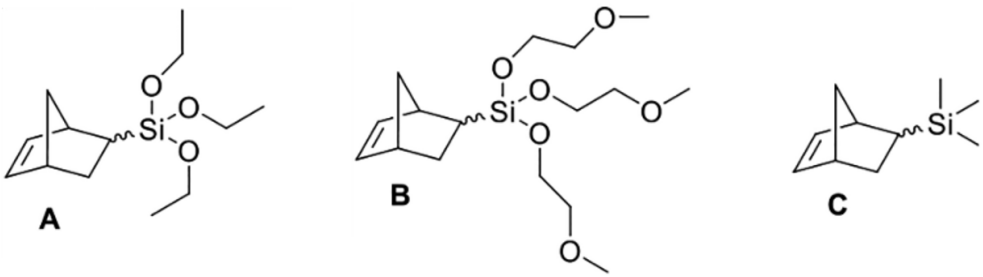


Figure 2.1: Monomer A (5-triethoxysilyl-2-norbornene), Monomer B (5-tris(2-methoxyethoxy)-2-norbornene), and Monomer C (5- trimethylsilyl-2-norbornene)

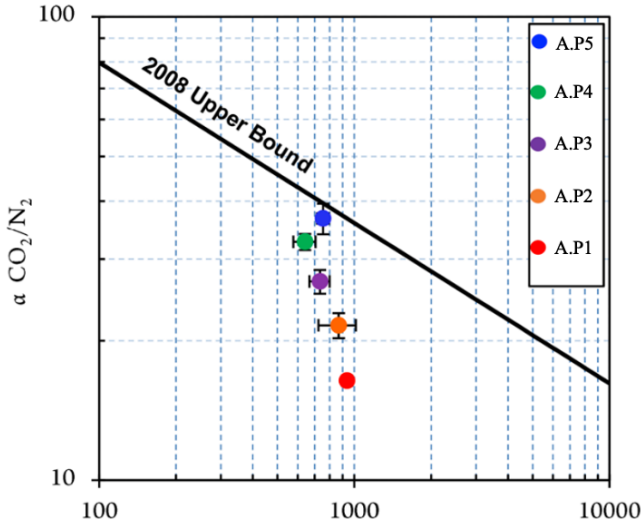


Figure 2.2: Robeson plot showing gas separation performance of polymers A.P1 through A.P5 with error bars.

(MD) simulations were performed to scan the conformational space, providing sufficiently stable starting conformations which were further optimized with DFT. Initial structures were generated with the OPLS-AA force field[93] in a periodic box at 200 K in the NVT ensemble with a Nosè-Hoover thermostat using a 1.0 ps time constant and a 1.0 fs time step. LigParGen[94] was used for the generation of the force field input, and all simulations were performed with the LAMMPS[95] software package. 10000 time steps were evaluated followed by a geometry optimization with respect to the force field to produce initial structures for DFT (Figure 1.4). Thirty conformations were considered for every monomer, monomer-CO_2, monomer-N_2, and monomer-monomer structures. All DFT calculations were performed using the TURBOMOLE[96,97] 7.2 program package with the PBE0[98] functional, def2-TZVPP[99] basis set, and Grimme's "D3" dispersion correction[31] with the Becke-Johnson damping function.[100] An ultrafine grid was used for integral evaluation, and geometry optimizations were performed with tight convergence criteria. The resolution of identity was used in the computation of two electron integrals. Frequency calculations were performed on the most stable conformations to ensure they are minima on the PES. Interaction energies (ΔE_{AB}) between two structures A and B were calculated using

$$\Delta E_{\mathrm{AB}} = E_{\mathrm{AB}} - E_{\mathrm{A}} - E_{\mathrm{B}} \tag{2.1}$$

where E_{AB} represents the energy of the interacting supersystem, and E_{A} and E_{B} represent the energies of the isolated systems.

2.1.4 Results and Discussion

In regards to monomer-gas interactions, monomers A and B each contain silyl ether groups that interact favorably with CO_2 via both dispersion and induction forces. More specifically, the electron-rich silyl ether oxygen atoms were shown to interact with the electron-deficient carbon of CO_2, while the two oxygen atoms of CO_2 form weak hydrogen bonds with the adjacent hydrogen atoms within the monomers. To understand the relative stability of the gas interactions in the polymer, we truncate the polymers A.P1, A.P2, and A.P3 to a saturated repeat units A, B, and C, respectively. The computed interaction energies for

monomer A-CO_2 (ΔE_{1-CO_2}) and monomer B-CO_2 (ΔE_{2-CO_2}) are -4.1 and -6.1 kcal mol^{-1}, respectively, whereas the interaction energy for monomer C that lacks any silyl ether groups is roughly half (-2.3 kcal mol^{-1}). The most stable conformation of the B-CO_2 supersystem is shown in Figure 2.3 in which both the silyl ether and ethereal oxygen atoms were found to interact with CO_2, presumably leading to its more favorable ΔE_{2-CO_2} interaction. Monomers A, B, and C each showed only weak interactions with N_2 of -2.3, -1.9, and -1.6 kcal mol^{-1}, respectively.

To understand the relative stability of interacting polymer units in the polymer, we calculate the interaction energy between two monomeric units of A, B, and C, respectively. The monomer−monomer interaction energies (ΔE_{1-1}, ΔE_{5-5}, and ΔE_{6-6}) which could then be compared to the monomer-gas energies to provide insight into the relative energy barrier for perturbing the monomer-monomer intermolecular interactions by a particular gas. It is expected that a lower energy difference ($\Delta\Delta E$) between monomer- monomer and monomer-gas interactions correlates to easier Henry's law dissolution of the gas molecule within the polymer matrix and therefore higher solubility values. The computed $\Delta\Delta E$ barriers are reported in Figure 2.4 in which monomer C has the lowest values of 4.6 and 5.2 kcal mol^{-1} for CO_2 and N_2, respectively. This agrees well with experimental findings that show that A.P6 has the greatest solubility for both gases, but because both the CO_2 and N_2 barriers are almost identical, A.P6 provides the lowest solubility selectivity.

The monomer-monomer interaction energies for two monomer A units (ΔE_{1-1}) and two monomer B units (ΔE_{2-2}) were calculated to be -10.7 and -12.6 kcal mol^{-1}, respectively. Though the interaction energy between two monomer B units is clearly stronger, monomer B also interacts with CO_2 more strongly. As a result, both monomers A and B have very similar $\Delta\Delta E$ barriers for CO_2 of 6.6 and 6.4 kcal mol^{-1}, respectively, which is consistent with the experimental observation that Henry's law dissolution of CO_2 into both polymers A.P5 and A.P1 are essentially equal. In contrast, the $\Delta\Delta E$ values for monomers A/B and N_2 showed a more substantial difference of 8.4 and 10.7 kcal mol^{-1}, respectively. This result suggests that the Henry's law dissolution of N_2 into polymer A.P5 (signified by monomer B) is far more difficult than dissolution into polymer P1 (signified by monomer A) and that A.P5 should have a significantly decreased N_2 solubility coefficient relative to A.P1. These DFT results

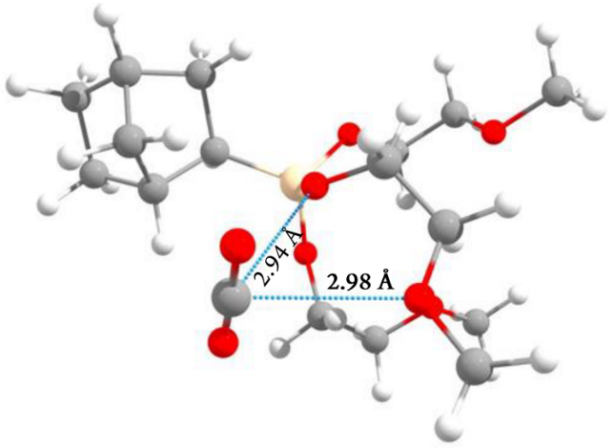


Figure 2.3: Most stable conformation for B-CO_2 contains favorable interactions between the two oxygen atoms of monomer B and the electron-deficient carbon of CO_2. C_{CO_2}-O_2 atom distances below 3.0 Å are shown.

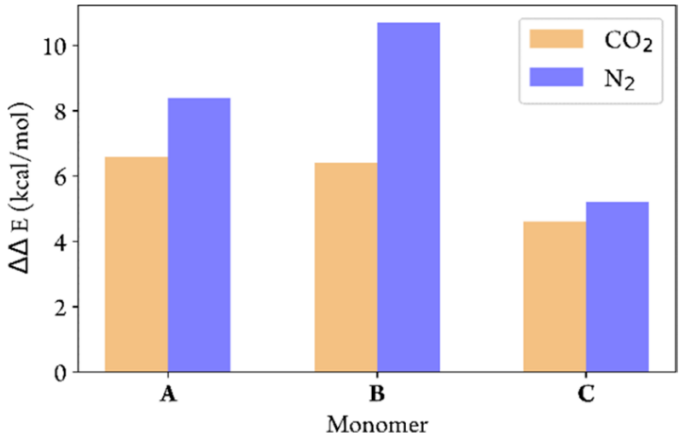


Figure 2.4: $\Delta\Delta E$ values for monomers A-C with CO_2 and N_2.

are in qualitative agreement with their experimentally determined solubility coefficients for A.P1 and A.P5 and therefore corroborate the greatly increased CO_2 selectivity observed for polymer A.P5, which occurs without a substantial loss in overall CO_2 permeability.

2.2 Alkoxysilyl VAPNB Structure-Function Relations

Disclosure

The following section contains a modified version of work from an article published in *Journal of Membrane Science* in collaboration with the group of Professor Brian K. Long and Aramco Services Company.[101] Several graduate students have contributed to the results summarized in 2.2.1, which includes Christopher R. Maroon, Morgan A. Higgins, and Dylan O'Neal. Jacob Townsend performed the DFT calculations used in the analysis of section 2.2.2 under the guidance of Professor Konstantinos D. Vogiatzis.

2.2.1 Overview

Having demonstrated that the alkoxysilane functional modifications that maximize polymer-CO_2 interactions relative to polymer-polymer and polymer-N_2 interactions can significantly enhance CO_2/N_2 selectivity, the Long group turned to understanding the relationship of additional modifications on alkoxysilyl-substituted VAPNBs to their gas separation performance. More specifically, three sets of polymers were examined (see Figure 2.5) to elucidate the effects of alkoxysilane branching, chain length of the alkoxysilane, and the length of the hydrocarbon linker between norbornene ring and siloxane unit. Note that the labels for polymers for this section follow notation B.P, where the only polymer shared between this section and section 2.1.2 is polymer B.P3 which corresponds to A.P1 in section 2.1.2. The experimental results are summarized in Table 2.2 and Figure 2.6. To further explain the observed behaviors, electronic structure theory calculations were performed to elucidate the underlying atomistic interactions contributing to the observed permeability. Computational protocol follows that contained within section 2.1.3.

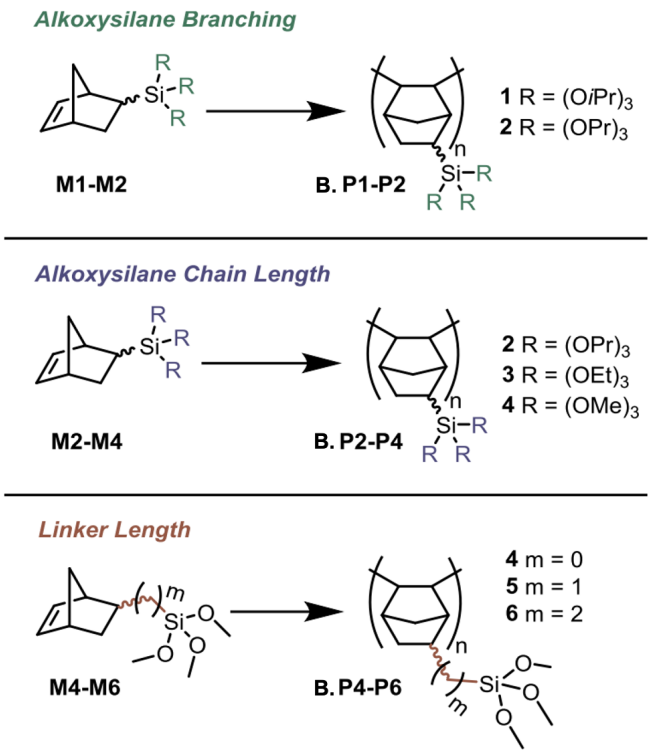


Figure 2.5: Series of alkoxysilyl-substituted monomers and polymers designed to elucidate the structure-property relationship between substituent identity and gas transport behavior.

Table 2.2: Permeability and selectivity of polymers B.P1-B.P6 for CO_2/N_2 gas pair.

Polymer	CO_2 Permeability	N_2 Permeability	Selectivity α
B.P1	1120 ($\pm$ 90)	90 ($\pm$ 10)	12 ($\pm$ 0.7)
B.P2	650 ($\pm$ 30)	47 ($\pm$ 1)	14 ($\pm$ 0.5)
B.P3	940 ($\pm$ 20)	57 ($\pm$ 2)	16 ($\pm$ 0.4)
B.P4	1600 ($\pm$ 200)	56 ($\pm$ 8)	29 ($\pm$ 0.8)
B.P5	770 ($\pm$ 20)	23 ($\pm$ 1)	33 ($\pm$ 1)
B.P6	560 ($\pm$ 60)	16 ($\pm$ 2)	35 ($\pm$ 0.5)

2.2.2 Results and Discussion

With respect to CO_2 interactions, all studied polymers contain a favorable interaction with the alkoxysilyl groups via stabilizing induction and dispersion interactions. The electron deficient carbon atom of CO_2 is able to interact favorably with the electron-rich ether oxygen atoms. Additionally, the oxygen atoms of CO_2 participate in weak hydrogen bond-type interactions with hydrogen atoms located on the monomeric unit, leading to interaction energies (ΔE) ranging from -4.07 to -5.26 kcal mol^{-1} (Table 2.2). In comparison, significantly weaker interaction energies (-2.15 to -2.45 kcal mol^{-1}) were obtained for nitrogen, which interacts primarily via weaker dispersion forces.

To better understand the stability of the gas interactions relative to the intermolecular polymer-polymer interactions, the monomer-monomer binding energies for each of the polymers were also evaluated. By comparing the strength of polymer-gas interactions, we can shed light on the relative energy penalty for perturbing the favorable monomer-monomer binding via a particular gas. The relative energy barrier, denoted here as $\Delta\Delta E$, is considered as a metric of the permeation, where lower barriers correspond to easier dissolution into the polymer matrix. The CO_2, N_2, and monomer-monomer ΔE values and relative energy barriers ($\Delta\Delta E$) are located in Table 2.3 and in appendix Fig. 1.

The effect of branching was investigated by comparing B.P1 and B.P2, which contain the isopropoxy groups and the linear propoxy chains respectively. The branched substituents of polymer B.P1 exhibit substantially lower monomer-monomer interactions, leading to a significantly lower $\Delta\Delta E$ values for both gases, which is in agreement with the experimentally observed increase in permeability values for B.P1 over B.P2. Furthermore, the $\Delta\Delta E$ difference between CO_2 and N_2 is approximately equivalent (2.8 and 2.7 kcal mol^{-1}) for B.P1 and B.P2, respectively. Therefore, the DFT results corroborate the higher permeability for P1 and that its performance should be in line with the permeability-selectivity trade-off.

Comparing polymers B.P2–P4, the longer alkyl chain length substituents exhibit stronger monomer-monomer interactions. By considering the CO_2 $\Delta\Delta E$ values, the DFT results suggest that B.P4 would have the lowest barrier, and thus the highest permeability, which is in line with experimental results. B.P2 and B.P3 have $\Delta\Delta E$ values within 0.3 kcal mol^{-1} for

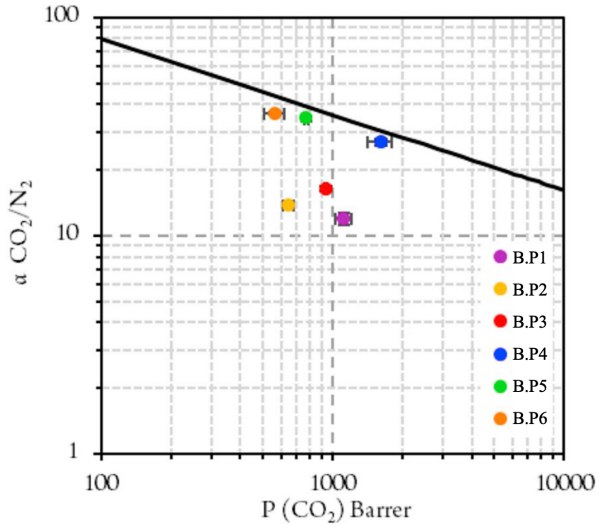


Figure 2.6: Robeson plot for polymers B.P1–B.P6 relative to the 2008 upper bound for CO_2/N_2 separations.

Table 2.3: Interaction energies (ΔE) of CO_2, N_2 and monomers (in kcal mol^{-1}), as well as their relative energy barriers for gas permeation [$\Delta\Delta E$ (in kcal mol^{-1})]

Polymer	$\Delta E_{\text{mon-CO}_2}$	$\Delta E_{\text{mon-N}_2}$	$\Delta E_{\text{mon-mon}}$	$\Delta\Delta E_{\text{CO}_2}$	$\Delta\Delta E_{\text{N}_2}$
B.P1	-5.26	-2.43	-8.37	3.11	5.95
B.P2	-5.01	-2.28	-11.32	6.32	9.04
B.P3	-4.07	-2.28	-10.66	6.59	8.38
B.P4	-4.88	-2.15	-8.50	3.95	6.35
B.P5	-4.76	-2.39	-9.37	4.61	6.98
B.P6	-4.88	-2.45	-10.19	5.30	7.74

CO_2 (-6.32 vs -6.59 kcal mol^{-1}), but the 1 kcal mol^{-1} stronger interaction between B.P2 and CO_2 suggests a slightly higher permeability, though this was not observed experimentally. Similar behavior was observed for N_2, where a $\Delta\Delta E$ of 6.35 kcal mol^{-1} for B.P4 suggests that this material would have the highest permeability.

The same trend is observed when examining how the number of saturated methylene carbons connecting the norbornene ring to the alkoxysilyl moiety affect the computational results (i.e. in polymers B.P4–P6 which contain 0, 1, and 2 saturated carbons as links, respectively). Increasing the length of the tether also increases monomer- monomer interactions. For both gases, going from B.P4 to B.P6, the $\Delta\Delta E$ values increase for each respective gas, suggesting that longer tether links should correspond to lower permeabilities, which agrees well with the observed permeabilities in Table 2.3. The difference in $\Delta\Delta E$ for CO_2 and N_2 is approximately 2.4 kcal mol^{-1} for these three polymers, which supports their similar performance with respect to the permeability-selectivity trade-off.

2.3 Impact on Functional Groups on Membrane-Mediated CO_2/N_2 Separations on a Common Polymeric Backbone.

Disclosure

The following section contains a modified version of work published in the *Journal of Polymer Science* in collaboration with the group of Professor Brian K Long.[102] Several graduate students have contributed to the results summarized in 2.3.1, which includes Morgan A. Higgins, Christopher R. Maroon, and Xinyi Wang. Jacob Townsend performed the DFT and SAPT calculations used in the analysis of section 2.3.3 under the guidance of Professor Konstantinos D. Vogiatzis.

2.3.1 Overview

Having developed many VAPNBs encroaching upon the Robeson upper bound, the Long group turned towards examining the effects of different functional units on a common polymer backbone. While the addition of CO_2-philic units such as imidazoles, amidoximes, and ether units are utilized for their preferential sorption of CO_2, few studies have performed a thorough study on a single polymer backbone, complicating the examination of structure-property relationships. Therefore, many polymers were synthesized (copolymers) for the examination of functional unit incorporation on gas separation performance, and are shown in Figure 2.7 with their permeation and copolymerization properties shown in Table 2.4.

2.3.2 Computational Details

DFT interaction energies follow protocol of Section 2.1.3. The strength of the interactions was calculated using density functional theory (DFT) and further characterized with symmetry adapted perturbation theory (SAPT) to decompose the interactions into contributions of electrostatics, dispersion, and induction. SAPT0[103]/jun-cc-pVDZ[104], as implemented in the Psi4[105] program package with density fitting,[106] was used to further characterize the nature of the interactions of the most stable conformations found with DFT. DFT interactions are located in the appendix in Table 1 and SAPT results are located in the appendices in Tables 2, 3, and 4.

2.3.3 Results and Discussion

Examination of Table 2.4 reveals that C.P1 exhibits enhanced CO_2 permeability and selectivity as compared to B.P2 (79.6 vs. 23.0 Barrer). C.P1 displays a three-fold increase in CO_2 permeability as well as a slight increase in N_2 permeability, resulting in an increase in selectivity, which is surprising given their similar functional composition. DFT calculations for C.P1 and C.P2 reveal similar CO_2 interaction energies (-5.03 and -5.79 kcal mol^{-1}, respectively), and N_2 interaction energies of -1.94 and -2.29 kcal mol^{-1}, which suggest a slight advantage of C.P2 over C.P1, which is in contrast to the experimental observation. The decomposition of the gas interactions reveals a similar mechanism for C.P1 and C.P2

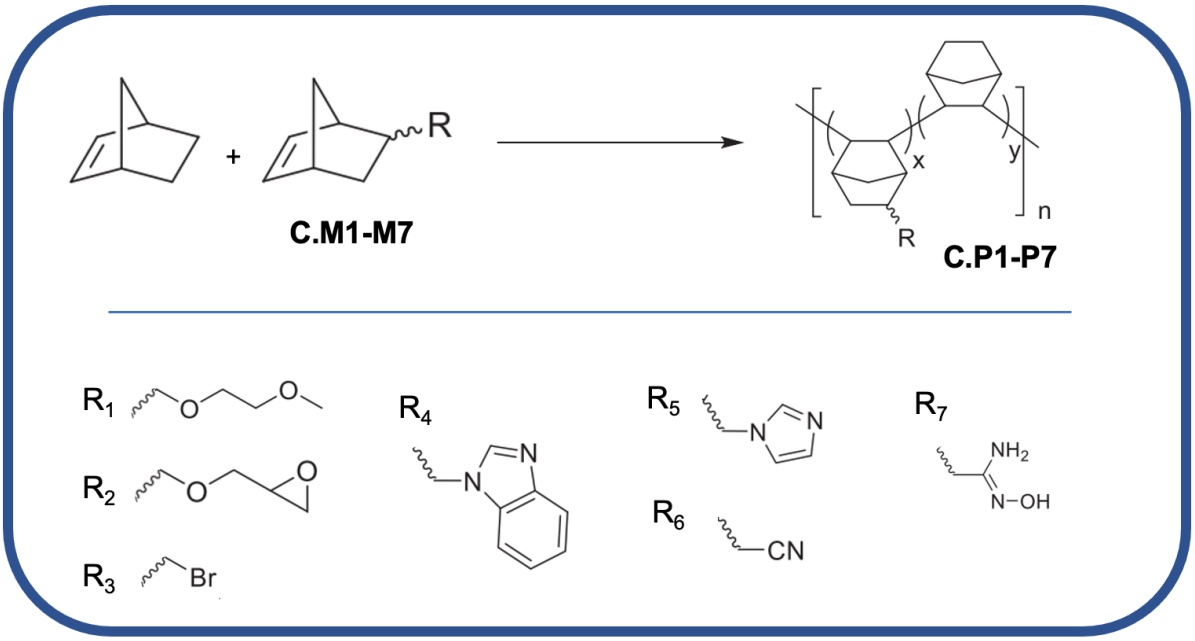


Figure 2.7: Copolymerization of unsubstituted norbornene with the functionalized norbornene (monomers C.M1-M7) with respective units R_1-R_7 to yield copolymers C.P1-P7.

Table 2.4: Permeability and selectivity of polymers C.P1-P7 for CO_2/N_2 gas pair and the ratio of incorporation of the functionalized monomer [shown in Figure 2.7].

Polymer	CO_2 Permeability	N_2 Permeability	Selectivity α	Fuctionalized comonomer %
C.P1	79.6 ($\pm$ 6.0)	2.29 ($\pm$ 0.09)	34.8 ($\pm$ 2.0)	51
C.P2	23.0 ($\pm$ 3.5)	0.82 ($\pm$ 0.16)	28.4 ($\pm$ 2.4)	58
C.P3	25.4 ($\pm$ 2.5)	0.88 ($\pm$ 0.07)	29.0 ($\pm$ 1.7)	53
C.P4	21.7 ($\pm$ 2.6)	0.64 ($\pm$ 0.10)	34.0 ($\pm$ 1.5)	24
C.P5	22.8 ($\pm$ 4.5)	0.71 ($\pm$ 0.13)	32.1 ($\pm$ 2.2)	53
C.P6	23.5 ($\pm$ 2.8)	0.75 ($\pm$ 0.13)	31.6 ($\pm$ 2.1)	53
C.P7	15.3 ($\pm$ 1.1)	0.24 ($\pm$ 0.08)	74.3 ($\pm$ 27.2)	-

in their interactions with the gases: CO_2 has a strong electrostatic attraction (approx. 55% of the interaction) toward C.P1 and C.P2 and N_2 is largely dispersion bound (approx. 70% of the interaction). Additionally, the monomer–monomer interactions are comparable in strength (-9.91 and -9.39 kcal mol^{-1}) and composition (30% electrostatic, 10% induction, and 60% dispersion). The quantum chemical calculations do not reveal the source of this unique increase in performance which disobeys the permeability/selectivity trade-off.

Comparison of copolymers C.P4 and C.P5, which bear benzimidazole and imidazole substituents, respectively, reveals similar CO_2/N_2 permeability and selectivity values (CO_2 permeability = 21.7 vs. 22.8 Barrer; α = 34.0 vs. 32.1), with the results falling largely within the error of one another. Imidazole and benzimidazole were not expected to exhibit drastic differences in CO_2-philicity, which is consistent with the nearly equivalent interaction energies between the benzimidazole (C.P4) and imidazole (C.P5) substituted units with CO_2 (-4.51 and -4.79 kcal mol^{-1}). However, the DFT calculations reveal that monomer–monomer interactions for the bulkier C.P4 are substantially larger than C.P5 (-15.1 and -10.6 kcal mol^{-1}, respectively). Though it may be reasoned that the stronger monomer–monomer interactions of C.P4 may contribute decreased gas permeability, these were not observed experimentally, perhaps due to the lower amount of functionalized monomers in the copolymer in C.P4 (see Table 2.4).

The amidoxime-functionalized copolymer C.P7 had a significant increase in CO_2/N_2 selectivity (α = 74.3) as compared to C.P3 (α = 29.0) and C.P6 (α = 31.6). C.P7 also exhibited the lowest CO_2 permeability of all the tested membranes (15.3 Barrer), for which hydrogen bonding within the polymer matrix was hypothesized as a possible explanation. DFT calculations verify the CO_2-philicity of amidoxime groups with C.P7 displaying the strongest CO_2 interaction energy (-6.18 kcal mol^{-1}), among these polymeric units. Additionally, C.P7 has the largest difference between its respective CO_2 and N_2 interaction energies (3.93 kcal mol^{-1}), which likely contributes to its enhanced CO_2/N_2 selectivity, albeit accompanied by a significant drop in CO_2 permeability. Because amidoxime functional groups have multiple opportunities for hydrogen bonding between polymer chains within the film, it is unsurprising that their calculated monomer–monomer interaction energy was very strong (-13.4 kcal mol^{-1}). The SAPT energy decomposition reveals that

while the C.P7–CO_2 interaction is very similar in nature to that of the C.P4–CO_2 and C.P5–CO_2 (approx. 65% electrostatic, 10% induction, and 25% dispersion). However, the C.P7 is uniquely dominated by electrostatic monomer-monomer interactions primarily due to hydrogen bonding, which may contribute to the reduced gas permeability exhibited experimentally.

2.4 Urea-Functionalized Poly(dimethyl siloxane) Membranes

Disclosure

The following section contains a modified version of work from an article published in Advanced Functional Materials in collaboration with the groups of Dr. Tomonori Saito and Professor Alexei P. Sokolov within the Chemical Sciences Division at Oak Ridge and Dr. Zhe Qiang from Northwestern University.[107] In the following section, Jacob Townsend performed the DFT calculations used in the analysis of section 2.3.3 uner the supervision of Prof. Konstantinos D. Vogiatzis.

2.4.1 Overview

In an effort to mimic biological systems which have built-in repair mechanisms, materials have been developed to repair themselves to maintain functionality. While development based on extrinsic healing mechanisms with limited healing cycles has occurred, recent research has focused on intrinsic healing mechanisms for multiple cycles of self-repair.[108] Self-repair can be realized through the incorporation of covalent or free-radical based chain ends, or through supramolecular dynamic bonds. Supramolecular chemical methods rely on weaker interactions compared to covalent self-healing mechanisms, dictated by hydrogen bonding, metal-ligand coordination, π bonding, and/or van der Waals interactions.[109] These relatively weaker supramolecular interactions result in noncovalent interactions that can be reversibly formed. In particular, hydrogen bonding is an attractive mechanism because it is amongst the strongest noncovalent interactions, and can therefore boost mechanical

strength and is responsible for the self-healing properties.[110,111] With these aims in mind, a self-healing polymer was investigated by incorporating urea-functionality end units to a polydimethylsiloxane (PDMS) polymer. The developed material with a PDMS repeat unit with four possible end units within the material is depicted in Figure 2.8. The resulting polymer was found to be a stretchable material that could undergo physical stress and recover its mechanical form and even recover gas-separation performance after a scratch. Therefore, the material was studied to examine the self-healing interaction strength and the possible implications of introducing CO_2 on the healing mechanism.

2.4.2 Results and Discussion

DFT calculations were performed to evaluate the supramolecular interactions of the polymeric elastomer end units and their binding energy with CO_2 molecules. The computational protocol follows that of section 2.1.3. The interaction of saturated monomer units I-IV, depicted in Figure 2.8, forms a set of ten unique dimer structures: I-I, I-II, I-III, I-IV, II-II, II-III, II-IV, III-III, III- IV, and IV-IV due to the presence of all four associating units in the polymer elastomer system. For each dimer, the energetically most stable structure was determined after evaluating different possible conformers. Each dimer system was studied with CO_2 to produce a set of conformers representing stable dimer-CO_2 supersystems, henceforth referred to as I-I-CO_2, I-II-CO_2, etc. The binding energies for the dimer systems and the average hydrogen bond distances are shown in Appendix Table 5, and ranged from approximately -16.5 to -18.2 kcal mol^{-1}. For all systems, the most stable structures contain two strong hydrogen bond interactions between the carbonyl oxygens and the amine hydrogens in a single plane, shown in Figure 2.9. In monomers with multiple amines bonded to carbonyl groups, the most stable geometry was found not to involve the terminal amine in hydrogen bonding with the adjacent monomer.

The interaction of monomers I-IV The electron deficient carbon of CO_2 (C_{CO_2}) interacts with the oxygen of the C-O group of the monomers and a hydrogen bond is formed between the oxygen of CO_2 (O_{CO_2}) and a hydrogen of a N-H group. A similar binding motif was found for the dimer pairs: the oxygen atoms of the dimer that interact with C_{CO_2}, and the hydrogens of the dimer can form weak hydrogen bonds with O_{CO_2}. Calculations were

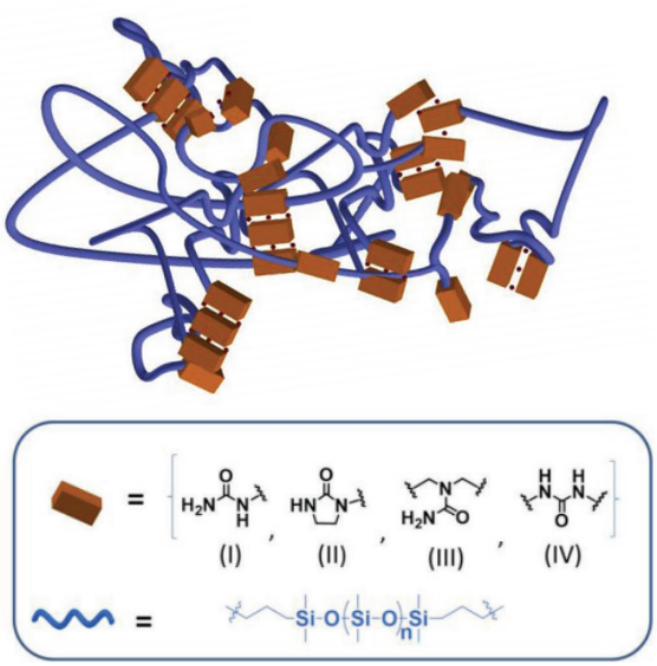


Figure 2.8: Depiction of the developed polymer with urea functional units (I-IV) on a PDMS backbone

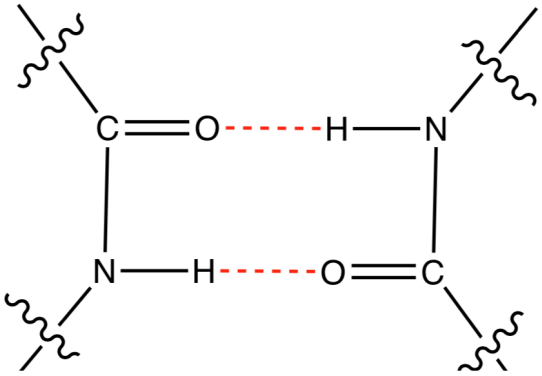


Figure 2.9: Two-dimensional representation of hydrogen bonding between monomer units I-IV

performed to find the most stable conformation of dimer systems and CO_2. The CO_2 binding energy for all systems ranges between -5.0 and -5.8 kcal mol^{-1} (see Appendix Table 6). The O-C_{CO_2} interaction distance falls between 2.66-2.78 Å for all monomer-CO_2 and six dimer-CO_2 supersystems. These supersystems have the CO_2 located adjacent to the plane of hydrogen bonding. The four systems with greater O-C_{CO_2} bond distances (II-IV-CO_2, III-III-CO_2, III-IV-CO_2,and IV-IV-CO_2) have sterically bulky side chains, therefore the CO_2 is more stable above the plane of hydrogen bonding. Introduction of CO_2 to the dimers distorts the plane of hydrogen bonding. When CO_2 is adjacent or diagonal to the hydrogen bond plane, the plane is only slightly shifted. However, when CO_2 is located directly above or below the plane of H-bonding dimers, the plane undergoes significant deformation.

Each urea functional group (I-IV) shows similar dimer interaction strengths, and all possible combinations are within 2 kcal mol^{-1}. The CO_2 interactions for monomers for associating units was even more similar (within 1 kcal mol^{-1}). The interaction strength for all urea functional units is on the same order of magnitude. However, when comparing the dimer interaction and CO_2 interaction strength, the dimer interaction strength is nearly three times stronger than monomer and dimer CO_2 interactions. Therefore, the study supports the interaction of CO_2 in the gas separation process should not impede the supramolecular dimer interactions responsible for self-healing.

2.5 Ion Specific Fluorescence Modulation of Polyvinyl Alcohol-Boronate Matrices

Disclosure

The following section contains a modified version of work from an article published in *Polymer Chemistry* in collaboration with Professor Johnathan N. Brantley and postdoctoral researcher Dr. Brian P. Jacobs.[112] In the following section, Jacob Townsend performed the DFT calculations used in the analysis of section 2.5.3 under the guidance of Professor Konstantinos D. Vogiatzis.

2.5.1 Overview

Aqueous anions such as OH^- and F^- play a critical role in many essential chemical and biological processes, but may pose significant health risks at increasing concentrations.[113,114] Therefore, there has been a substantial effort to develop chemosensors for the selective detection of such anions by spectroscopic methods.[115–118] Boron-containing polymers have been particularly promising as fluorine sensors and extractors. With this in mind, the Brantley group turned to the development of synthetically accessible aromatic boronate polymers for anion binding, extraction, and optical detection which are shown in Figure 2.10. While a series of polymers of polymers for the detection of various anions were investigated, the discussion herein will be on the polymer labeled 4_{CF_3} in Figure 2.10a . Upon examining the UV spectrum of 4_{CF_3}, a broad absorbance around 317 nm was found upon the addition of tetrabutylammonium fluoride (TBAF), with subsequent increases in the absorption at increased loading of TBAF (see Figure 2.10b).

2.5.2 Computational Details

Quantum chemical calculations were performed on saturated repeating units of fluorine to elucidate the atomistic effects of anion binding in PVA-aryl boronate polymers. All calculations were performed using the Orca 4.1 program package.[119] Density functional theory (DFT) geometry optimizations were performed at the B3LYP[120]/def2-SVP[99] level of theory with Grimme's "D3" dispersion correction with the Becke-Johnson damping function.[31,100] Single point calculations were performed with the def2-TZVPPD[121] basis set in order to obtain higher accuracy. An ultrafine grid was used for integral evaluation (Grid6 setting). The resolution of identity was used in the computation of two electron integrals,[122] and the "chain of spheres" approximation was applied for the exchange integrals.[123,124] Frequency calculations were performed to ensure the optimized structures were minima on the potential energy surface. For reducing the computational complexity, TBAF was modeled as tetramethylammonium fluoride (TMAF). Time-dependent density functional theory (TD-DFT) calculations were performed on the optimized structures to improve the interpretation of the observed UV-vis spectra. TD-DFT calculations were performed with

the range-separated exact exchange CAM-B3LYP[125] functional and def2-TZVPPD[121] basis set to generate UV-vis spectra with 150 optimized roots, which has shown to be an accurate method in the calculation of excitation energies.[126] The orca mapspc program was used to generate UV-vis spectra with 2000 cm^{-1} of peak broadening.

2.5.3 Results and Discussion

The interaction energy between the repeating unit, henceforth referred to as a monomer (Mon), and the first TMAF was computed by using the expression in equation 2.1 and is referred to as $\Delta E_{\text{Mon-TMAF}}$. The isolated monomer is planar with an intramolecular hydrogen bond as shown in Figure 2.11a. Upon introducing the cation-anion pair TMAF, the fluorine binds to the boron, which yields a tetrahedral boron. The cation is found to interact favorably with the ether oxygen atoms, as it is shown in 2.11b. The interaction energy, $\Delta E_{\text{Mon-TMAF}}$, is considerable at -33.4 kcal mol^{-1}.

Upon addition of a second cation-anion pair (TMAF), the fluorine hydrogen bonds with the hydroxyl group of the monomer, granting an additional -23.7 kcal mol^{-1} over the $\Delta E_{\text{Mon-Pair}}$, calculated as:

$$\Delta E_{\text{Mon-2TMAF}} = E_{\text{Mon-2TMAF}} - E_{\text{Mon-TMAF}} - E_{\text{TMAF}} \tag{2.2}$$

The third equivalent of TMAF results in an additional -16.5 kcal mol^{-1}. This value was calculated by subtracting the energy of the monomeric unit that has two bound TMAF molecules:

$$\Delta E_{\text{Mon-3TMAF}} = E_{\text{Mon-3TMAF}} - E_{\text{Mon-2TMAF}} - E_{\text{TMAF}} \tag{2.3}$$

The binding of the second and third fluorine equivalent occurs is of similar nature, i.e. at hydroxo sites of the polymeric material. For that reason, we have also considered an average interaction between the Mon-TMAF supersystem (first fluorine equivalent bound on boron) and two cation-anion pairs as:

$$\Delta E'_{\text{Mon-3TMAF}} = E_{\text{Mon-3TMAF}} - E_{\text{Mon-TMAF}} - 2E_{\text{TMAF}} \tag{2.4}$$

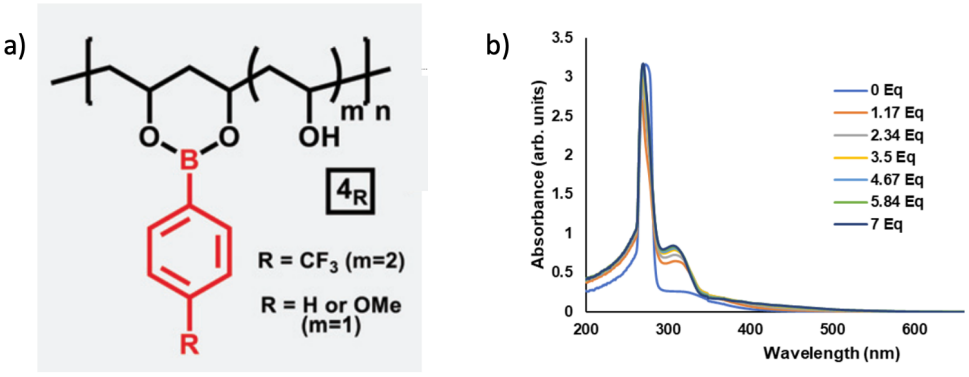


Figure 2.10: Graphical depiction of synthesized polyvinyl alcohol-aryl boranate polymers synthesized.

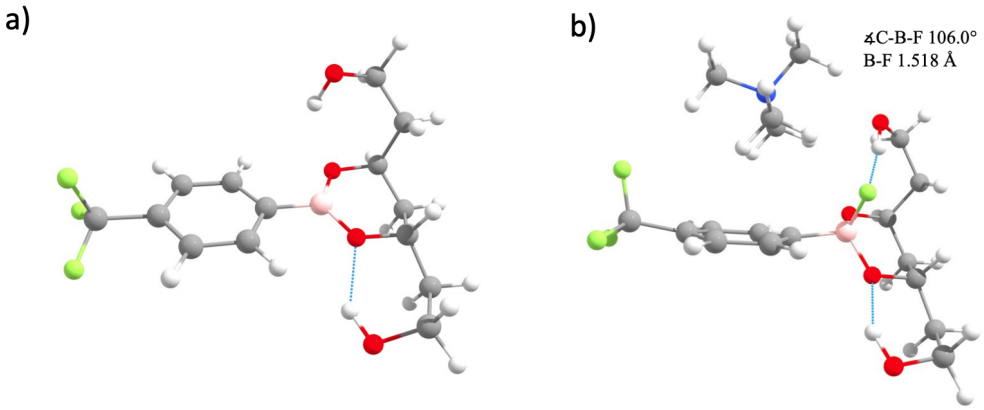


Figure 2.11: (a) Optimized repeating unit and (b) Optimized repeating unit + TMAF. The B-F bond length is 1.518 Å and is further stabilized by intramolecular hydrogen bonding from the neighboring hydroxy unit. Color code: B (pink), O (red), C (grey), F (green), H (white)

The $\Delta E'_{\text{Mon-3TMAF}}$ term is -40.2 kcal mol^{-1}, which leads to an average binding of approximately 20 kcal mol^{-1} for the second and third addition of TMAF. No additional binding sites were for fluorine were located, and therefore we conclude that the monomer can effectively capture multiple fluoride units at the boron sites and hydroxy units, which is consistent with experimental observations.

TD-DFT calculations were performed on a repeating unit of 4_{CF_3} The scope of these calculations is not to reproduce the experimental spectra, since they have been performed on a gas-phase, isolated molecular model and they do not capture the properties of the bulk, amorphous material. For example, effects from $\pi - \pi$ stacking between two repeating units are neglected. Such effects are expected to red-shift the TD- DFT transitions with respect to the experimental results.[127] However, these calculations offer a description on the electronic structure of the ground and the first, low-lying states, which can explain qualitatively the changes of the experimental UV-Vis spectra. The calculated UV-vis spectra for the monomer and transition electron density differences are shown in Figure 2.12. Table 2.5 shows the molecular orbital contribution to the first two excitations. Overall, the leading excitations for both peaks correspond to excitations from the conjugated π ring to the electron deficient atomic p orbital of boron, which is part of the conjugated π system of the molecular unit. Herein, the focus is on the peak labeled **1**, because the peak in this region experimentally exhibits increased ratiometric absorbance upon addition of tetrabutylammonium fluoride, and is the targeted excitation for the observation of fluorescence.

The spectra of the monomeric unit were calculated with the presence of one and two TMAF units and are shown in Figure 2.13. Both spectra contain a strong peak characteristically equivalent to the unsaturated monomer (Figure 2.12) at approximately 6.9 eV. Although the excitations seem equivalent to those found for the unsaturated monomer, the nature of the excitations are not synonymous. For 1-TMAF and 2-TMAF, the $\pi - \pi^*$ excitation is mainly localized to the phenyl group (see Table 2.6)

When comparing the calculated to the experimental spectra in Figure 2.10b, the bare polymer does not show a strong characteristic peak corresponding to Peak 1 in Figure 2.12. Additionally, the presence of solvent and other polymeric units may donate electron density to the boron atom, reducing the prevalence of this peak. However, the presence of peaks

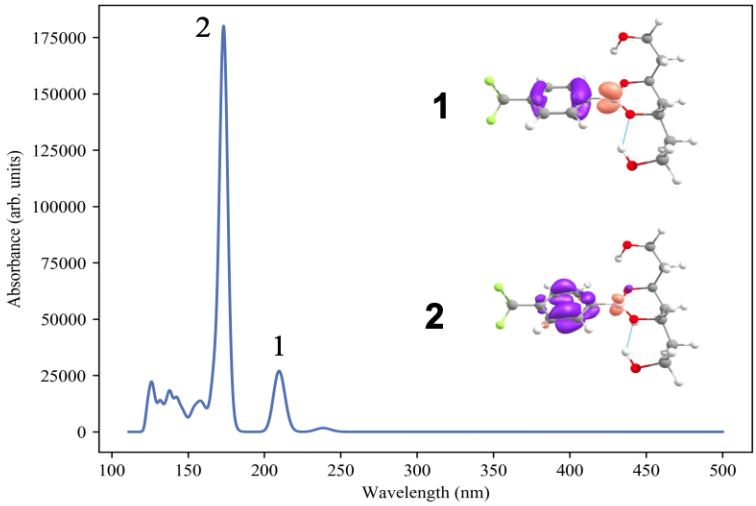
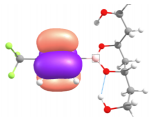
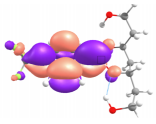
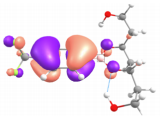
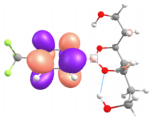
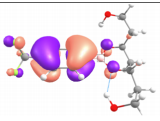
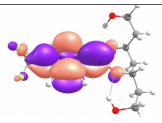
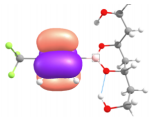
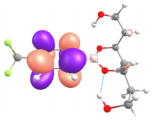


Figure 2.12: Calculated UV-vis spectra for repeating unit at the CAM-B3LYP/def2-TZVPPD level of theory. Inset: Electron density differences of the electronic transition between the ground state and the two excited states. Purple isosurface corresponds to electron deficiency of the molecule related to the ground state, pink isosurface to electron excess. Both 1 and 2 correspond to excitations from the aromatic π orbitals to the empty 2p orbital of boron. Color code: B (pink), O (red), C (grey), F (green), H (white)

Table 2.5: Transition energies (in eV) and the corresponding molecular orbital contribution to the first and second peaks in Figure 2.12

State	Energy	Transition		Weight
1	5.92	 HOMO	 LUMO	65
		 HOMO-1	 LUMO-1	22
2	6.86	 HOMO-1	 LUMO	76
		 HOMO	 LUMO-1	14

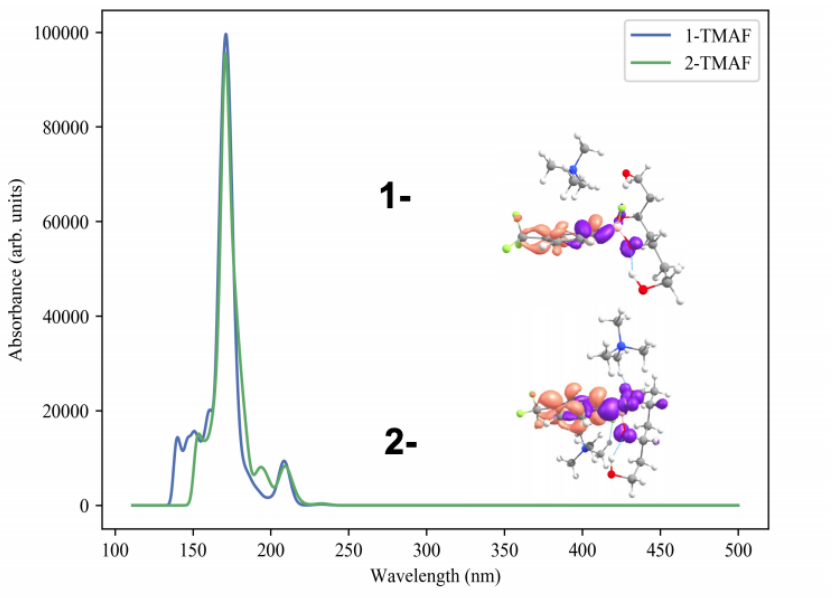
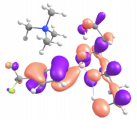
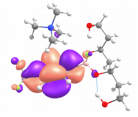
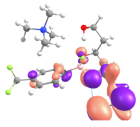
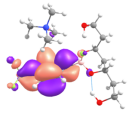
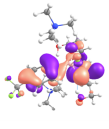
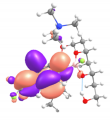
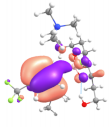
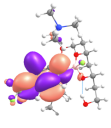


Figure 2.13: Calculated UV-vis absorbance spectra with one (blue) and two (green) TMAFs added to the repeating unit at the CAM-B3LYP/def2-TZVPPD level of theory. Both spectra show a characteristic peak at approximately 215 nm.

Table 2.6: Transition energies (in eV) and the corresponding molecular orbital contribution to the first peaks shown in Figure 2.13

System	Energy	Transition		Weight
1-TMAF	5.95	 HOMO	 LUMO	48
		 HOMO-1	 LUMO	14
2-TMAF	5.96	 HOMO-1	 LUMO	33
		 HOMO	 LUMO	14

after the addition of the fluorine correspond to polymer backbone to π^* excitations in the material. Increasing the equivalents of fluorine did not increase the intensity of these peaks, suggesting that the ratiometric increase in absorbance is due to increased fluorine occupancy of the boron sites, since there are many competitive binding sites with strong interactions (cations and hydroxyl units)

Chapter 3

Understanding the Nature of Weak Interactions Between Functionalized Boranes and N_2/O_2, Promising Functional Groups for Gas Separations

Disclosure

The following section contains an adapted reproduction of the published manuscript in the *Journal of Physical Chemistry A* with permission.[27] Jacob Townsend was responsible for the coupled-cluster calculations, DFT functional screening, SAPT calculations and analyses, and manuscript preparation. Nicole Braunscheidel assisted in the DFT functional and basis set benchmark under the supervision of Jacob Townsend. All work was performed under the guidance of Professor Konstantinos D. Vogiatzis.

Abstract

The separation of nitrogen and oxygen gases is considered as a very challenging processes, since both O_2 and N_2 are nonpolar molecules with similar kinetic diameters. Electronic structure theory can provide a fundamental understanding on effects that can lead to selective binding of oxygen or nitrogen gas for the development of novel separation processes. Boranes can bind dinitrogen through a dative bond, where the boron acts as a σ acceptor and back-donates through π orbitals. To better understand these interactions, we have performed highly accurate CCSD(F12)(T) and CCSDT(Q) computations for the BH_3-N_2 and BH_3-O_2 complexes. The coupled-cluster binding energies were used as reference for benchmarking different density functionals, and larger functionalized boranes were examined at the M05/def2-TZVPPD level. Symmetry adapted perturbation theory (SAPT) calculations were performed for the elucidation of the nature of the interaction between nitrogen and substituted boranes, and how direct or distal functionalizations affect the strength of the weak dative bonds. Using these methods, several boranes were found to bind N_2 over O_2. These complexes may be promising for incorporation into the next generation of materials in the effort towards more efficient N_2/O_2 separations.

3.1 Introduction

Oxygen-rich gas is demanded for various industrial applications including medical use, steel production, chemical manufacturing, and oxy-fuel combustion.[128–130] In particular, oxy-fuel combustion has been suggested as an efficient approach towards environmental carbon capture, which uses pure oxygen as a reagent. Thus, removal of N_2 from the feed gas provides CO_2 and water vapors as products of combustion, which simultaneously reduces or eliminates NO_x pollutants.[131–134] Currently, these processes are performed via cryogenic distillation or pressure swing adsorption to produce highly pure oxygen, but they prove to be very energy intensive. Cryogenic distillation is the most prominent method due to excellent oxygen purity of up to 99% and high gas production volume. However, this process requires extensive cooling, which results in the high-energy cost of approximately 200kWh per ton

of O_2.[135,136] Therefore, it is essential to produce alternative methods to perform high purity N_2/O_2 separations with lower energy costs.

Metal organic frameworks (MOFs), a family of porous hybrid organic/inorganic materials with metal-containing subunits connected by organic linkers, have been considered a promising alternative for numerous gas separations.[135,137] The high surface area of MOFs makes them attractive for gas separation and storage,[135] including the separation of O_2 over N_2 at higher temperatures than cryogenic conditions and, consequently, at lower energy cost.[138] Many computational[138–140] and experimental[141–145] studies have demonstrated applicability through selectivity which occurs at a coordinatively unsaturated metal center. Functionalization of the organic linkers offers a different, attractive approach to increase MOF selectivity,[146–154] but this approach has not yet matured for N_2/O_2 gas separations.

An additional alternative to distillation is membrane-based separation, which is driven by a gas pressure gradient between upstream and downstream membrane interfaces. Membrane gas separations are energetically efficient and can be performed at ambient conditions.[155,156] However, polymeric membranes have an observed tradeoff between permeability and selectivity,[15,38] which is especially challenging in the separation of oxygen and nitrogen gas. To date, membranes have not shown to be highly selective for oxygen purification, resulting in oxygen gas purities of only 25-40%.[157] To increase the purity of the oxygen stream, membranes must be developed with improved gas selectivity.

To improve the gas separation performance of materials, one approach is to incorporate functional groups that interact selectively towards one of the gases. Driven by the recent progress in nitrogen fixation by borylenes,[158,159] we made the hypothesis that functionalization of materials with boranes can enhance N_2/O_2 selectivity. Indeed, many boranes have already been incorporated into many materials for gas separation. For example, boron nitride nanosheets have been developed into membranes and show promise in separating CO_2/N_2 and CH_4/H_2.[160–162] A computational study on MOF-5 showed a 3-fold increase in H_2 interaction energies in replacing organic linker carbons with boron-substituted groups.[163] Additionally, MOFs have shown improved H_2 storage ability after the incorporation of amine boranes.[164,165] Polynuclear boranes have also been incorporated into polymer matrices in a mixed matrix membrane (MMM) for enhancing gas separation.[166]

However, functionalized materials with boranes have not yet been studied for N_2/O_2 separation.

Many nitrile-boron complexes have been studied computationally[167–169], and have shown especially strong binding energies with BH_3.[170,171] In particular, the BH_3-N_2 complex has been examined in an early study with the uncorrelated Hartree-Fock (HF)/4-31G scheme, which predicted a B-N distance of 2.644 Å.[172] Later, the complex was studied by means of density functional theory (DFT) at the B3LYP/6-311+G** level, which gave a remarkably shorter bond distance of 1.568 Å.[173] In 2010, Smith *et al.* performed a thorough examination of the BH_3-N_2 complex by performing higher-level coupled-cluster singles-and-doubles with perturbative triples (CCSD(T)) calculations with the 6-311+G(2df,2pd) basis set to determine a bond length of 1.641 Å.[171] However, a substantial difference in binding energies and geometries was found between the CCSD and CCSD(T) levels. The addition of the perturbative triples (T) shortens the B-N distances from 1.666 Å (CCSD) to 1.641 Å (CCSD(T)) and increased the interaction energy from -3.8 kcal mol^{-1} to -5.7 kcal mol^{-1}, respectively. Significantly weaker binding energies were observed for coupled-cluster (-5.7 kcal mol^{-1}) compared to DFT, which varied from -7.4 to -14.1 kcal mol^{-1}, depending on the choice of the functional.[171] However, the lack of counterpoise correction on the CCSD(T) energies (particularly with the 6-311+G(2df,2pd) basis) may not provide the expected accuracy for the interaction strength.

To elucidate the nature of the interaction in this complex, we have performed high level post-Hartree-Fock computations on the BH_3-N_2 and BH_3-O_2 complexes to understand their electronic structure and energetics. A full potential energy curve was computed for the BH_3-X complexes (X = N_2 or O_2) and studied with post-Hartree-Fock methods, including the Møller-Plesset second order perturbation theory (MP2) and explicitly correlated F12 coupled-cluster (CC) methods. Additionally, the sufficiency of the CCSD(T) level of theory for the description of the BH_3-N_2 complex was examined by obtaining reference binding energies beyond the "gold standard" CCSD(T) scheme. Full triples and non-iterative quadruples excitations were included (CCSDT(Q) level) for highly accurate treatment of electronic correlation. Based on these results, many popular DFT functionals were examined to determine the best level of theory for these molecular supersystems. O_2 and N_2 binding

energies with BR_3 complexes were examined, where R corresponds to functional groups appended to the boron complex, to suggest functionalized boranes for the separation of O_2 and N_2. Finally, the symmetry-adapted perturbation theory (SAPT) decomposition scheme was applied in order to elucidate the individual contributions to the interaction energies from a subset of the functionalized boranes.

3.2 Computational Methods

Calculations were performed using TURBOMOLE 7.2 (unless otherwise noted). BH_3-N_2/O_2 dissociation curves were computed by means of DFT with the B3LYP[120,174] functional, Grimme's "D3" dispersion correction[31] with the Becke-Johnson (BJ) damping function,[100] and the def2-TZVPP[99] basis set. Dispersion corrections significantly increase the accuracy of DFT calculations, including open shell cases such as molecular oxygen.[31] The resolution of identity (RI) was used for the computation of the two-electron integrals.[175] Single point calculations were performed with second-order Møller-Plesset (MP2) perturbation theory and coupled-cluster with singles-and-doubles and perturbative triples excitations (CCSD(T)) on the B3LYP-D3(BJ)/def2-TZVPP geometries, where the B-X distance (X = N or O) was kept fixed. The counterpoise correction (CP) was applied in all MP2 and CCSD(T) calculations to correct the error from the truncation of the one-electron basis set:[176]

$$E_{\mathrm{CP}} = E_{\mathrm{AB}}^{\mathrm{A+B}} - E_{\mathrm{A}}^{\mathrm{A+B}} - E_{\mathrm{B}}^{\mathrm{A+B}} \tag{3.1}$$

The subscripts denote the atoms present from the geometry of the supersystem AB and the superscripts denote that basis functions of the structures A and B that were included. Two relative energy terms are used to describe the interaction of complexes: binding energy and interaction energy. Binding energy is denoted as the energy difference between the supersystem and the relaxed monomers, and interaction energy as the difference between the supersystem and the monomers in the geometry of the supersystem. Therefore, the binding energy will be described in the conformation of the complex as:

$$\Delta E_{\mathrm{binding}} = \Delta E_{\mathrm{interaction}} + E_{\mathrm{deformation}} \tag{3.2}$$

where $\Delta E_{\text{interaction}}$ represents the interaction energy in the conformation of the complex and $\Delta E_{\text{deformation}}$ the (positive) energy difference between the monomers in the supersystem geometry and the relaxed, isolated monomers.

Two approaches were used for the estimation of the complete-basis set (CBS) limit in the post-Hartree-Fock methods. The first was by applying explicitly correlated (F12) methods, which include functions that depend on interelectronic coordinates to efficiently converge toward the CBS limit.[177,178] These were utilized in TURBOMOLE[96] 7.2 using the frozen core approximation with the cc-pVQZ-F12[179] basis set and corresponding complementary auxiliary basis set (CABS). The cc-pVQZ-F12 is a large basis set with noncovalent binding errors typically less than 0.01 kcal mol^{-1}. However, they have not demonstrated improvement over the standard aug-cc-pVQZ basis for noncovalent interactions.[180] The aug-cc-pwCV5Z (aug-cc-pV5Z for "H")[181] was used for fitting the F12 and electron-repulsion integrals (CBAS), while the aug-cc-pV5Z[182] was used for two-electron contributions to the Fock matrix (JKBAS). The 2B ansatz was used in all F12 calculations. The second strategy involves the two-point Helgaker extrapolation scheme:[183]

$$E^{\text{corr}}_{X(X+1)} = \frac{X^3 E^{\text{corr}}_X - Y^3 E^{\text{corr}}_Y}{X^3 - Y^3} \tag{3.3}$$

where X and Y are the cardinal numbers to the basis sets used. This extrapolation was performed on the CCSD(F12)(T) calculations with the cc-pVTZ-F12 and cc-pVQZ-F12 basis sets, corresponding to X = 3 and Y = 4, respectively. Herein, the results labeled CBS will be the those with the explicitly correlated F12 functions, using the triple and quadruple-zeta scheme (Eq. 3.3).

Additionally, coupled-cluster calculations at a higher correlation level were performed with full single, double, and triple excitations and perturbative quadruple excitations (CCSDT(Q)). Due to the inherently large number of excitations, the smaller cc-pVTZ[184] basis set was used. These calculations were counterpoise corrected for the complexes, and the CCSDT(Q)/cc-pVTZ energies were used to estimate a higher-level electron correlation effects via the following model:

$$E_{\text{CCSDT(Q)}} \approx E^{\text{CBS}}_{\text{CCSD(F12)(T)}} + \delta T(Q) \tag{3.4}$$

$$\delta T(Q) = E^{\text{cc-pVTZ}}_{\text{CCSDT(Q)}} - E^{\text{cc-pVTZ}}_{\text{CCSD(T)}} \tag{3.5}$$

where the $\delta T(Q)$ term provides the correction to the energy by adding higher-level correlation terms at the CCSD(T)-F12 level. CCSDT(Q)/cc-pVTZ and CCSD(T)/cc-pVTZ calculations were performed with the NWChem 6.6 program package.[185]

An alternative method to supermolecular energies, which relies on the indirect calculation of interactions based on a supersystem and isolated fragments, is the symmetry-adapted perturbation theory (SAPT). SAPT decomposes the interactions between two molecular fragments into the following meaningful terms: electrostatics, exchange-repulsion, polarization/induction, and dispersion forces.[186] All SAPT calculation methods were performed with density fitting as implemented in the Psi4 program package.[105,187–189] The density fitted SAPT0[106] calculations were performed with the jun-cc-pVDZ[104] basis and a corresponding auxiliary basis of jun-cc-pVDZ-JK[190] for the self-consistent field (SCF) step and jun-cc-pVDZ-RI[122] for the open shell oxygen systems. For a highly-accurate analysis of functionalized boranes and their interaction with nitrogen, the higher correlated SAPT2+3[191] method with coupled-cluster doubled (CCD) treatment of dispersion was used.[192,193] For these calculations, the aug-cc-pVTZ basis set along with density fitting were used.[105,187–189] SCF was performed the aug-cc-pVTZ-JK[190] auxiliary basis set, and the higher order terms are calculated with the aug-cc-pVTZ-RI.[194] Lastly, the charge transfer stabilization was calculated as described by Stone and Misquitta.[195]

3.3 Results and Discussion

3.3.1 Reference Data from CCSD(F12)(T) and CCSDT(Q)

Two highly-accurate coupled-cluster schemes were applied for the assessment of the interactions between boranes and the N_2/O_2 gas molecules. Figure 3.1 includes the dissociation curves for the BH_3-N_2 and BH_3-O_2 complexes (not in scale) as they were obtained from different explicitly-correlated methods. Results from the B3LYP-D3(BJ) are shown for comparison.

The corresponding energies, counterpoise corrections, and equilibrium bond distances between BH_3 and N_2/O_2 for each level of theory and basis sets are shown in Table 3.1. The explicitly correlated calculations are sufficiently converged with the quadruple-zeta basis, since CP corrections are less than 0.1 kcal mol^{-1} for all complexes and methods. The degree of electron correlation greatly impacts the interaction energies for both BH_3-N_2 and O_2 complexes, as expected. MP2(F12)/cc-pVQZ-F12 predicts an BH_3-N_2 interaction strength of -7.50 kcal mol^{-1} with a B-N distance of 1.62 Å, whereas CCSD(F12)/cc-pVQZ-F12 provides an interaction strength of -3.89 kcal mol^{-1} at 1.66 Å. MP2 predicts overbinding of N_2 whereas CCSD underbinds, which is consistent with previous observations,[196,197] while the CCSD(F12)(T) with the same basis set provides a bond distance at 1.64 Å with a strength of -5.96 kcal mol^{-1}. For oxygen, both CCSD(F12) and MP2(F12) predict interaction distances of 2.68 Å, which is 0.06 Å longer than CCSD(F12)(T). However, due to the relatively flat potential energy profile of BH_3-O_2, the MP2(F12) and CCSD(F12) are less than 0.3 kcal mol^{-1} from to the CCSD(F12)(T) results.

It is also evident from the reference CCSDT(Q) results that the choice of an appropriate basis set is necessary for increased accuracy. Using the two-point extrapolation scheme of Eq. 3.3, the BH_3-N_2 interaction energy is calculated to be -6.17 kcal mol^{-1} at the CCSD(F12)(T)/CBS level of theory. The δT(Q) correction (Eq. 3.5) is -0.08 kcal mol^{-1} for the BH_3-N_2 complex. The correction from including explicit triples (δT) reduces the interaction energy by 0.04 kcal mol^{-1}, which is consistent with the results of Smith *et al.* which found positive δT in many polar noncovalent interactions and the N_2 dimer.[170] Addition of the perturbative quadruples provides a larger correction to the interaction energy than the explicit triples since the δ(Q) contribution is -0.13 kcal mol^{-1}. Summing these two terms, the δT(Q) correction is -0.08 kcal mol^{-1}, in line with a previous study finding that the δT(Q) term should always be negative (stabilizing) for noncovalent interaction systems at the basis set limit, where it should capture nonconverged dispersion energy.[198] Thus, the best estimate to the BH_3-N_2 interaction energy provided by CCSD(F12)(T)/CBS+δT(Q) is -6.26 kcal mol^{-1}. For the BH_3-O_2 system, the explicit triples correction is almost negligible (-0.003 kcal mol^{-1}); however, the perturbative quadruple correction is -0.015 kcal mol^{-1}, giving a total δT(Q) correction of -0.018 kcal mol^{-1}. This results in a -1.22 kcal mol^{-1}

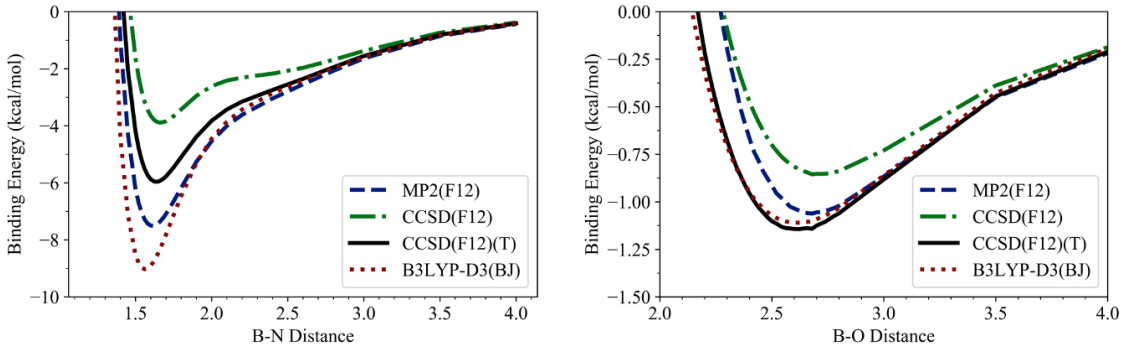


Figure 3.1: Dissociation curves for BH_3-N_2 (left) and BH_3-O_2 (right) obtained at the CP corrected MP2(F12), CCSD(F12), and CCSD(F12)(T) levels of theory (cc-pVQZ-F12 basis set) using B3LYP-D3(BJ)/def2-TZVPP geometries. Note that the binding energies for the N_2/O_2 dissociation binding energies are not to scale.

Table 3.1: Binding energies (kcal mol^{-1}), Counterpoise Correction (kcal mol^{-1}), and B-X (X= N or O) bond distances (Å) for the BH_3 complexes using explicitly correlated MP2 and coupled-cluster levels of theory.

Method	ΔE	CP Correction	$\Delta E(CP)$	R_{B-X}
$BH_3 - N_2$				
MP2(F12)/cc-pVQZ-F12	-7.55	0.04	-7.50	1.62
CCSD(F12)/cc-pVQZ-F12	-3.93	0.04	-3.89	1.66
CCSD(F12)(T)/cc-pVTZ-F12	-5.88	0.08	-5.80	1.64
CCSD(F12)(T)/cc-pVQZ-F12	-6.00	0.04	-5.96	1.64
CCSD(F12)(T)/CBS	-6.17	-0.01	-6.17	1.64
CCSD(F12)(T)/CBS+δT(Q)			**-6.26**	**1.64**
$BH_3 - O_2$				
MP2(F12)/cc-pVQZ-F12	-1.10	0.04	-1.06	2.68
CCSD(F12)/cc-pVQZ-F12	-0.90	0.05	-0.86	2.68
CCSD(F12)(T)/cc-pVTZ-F12	-1.14	0.05	-1.10	2.68
CCSD(F12)(T)/cc-pVQZ-F12	-1.19	0.05	-1.14	2.62
CCSD(F12)(T)/CBS	-1.25	0.05	-1.20	2.62
CCSD(F12)(T)/CBS+δT(Q)			**-1.22**	**2.62**

reference binding energy at the CCSD(F12)(T)/CBS+δT(Q) level for BH_3-O_2. This energy value is in agreement with the B3LYP-D3(BJ)/def2-TZVPP results for the oxygen complex, which predicts a -1.11 kcal mol^{-1} binding energy. On the contrary, B3LYP-D3(BJ) fails to describe the BH_3-N_2 interaction accurately by overbinding nitrogen by 2.7 kcal mol^{-1}.

3.3.2 DFT Benchmark Study

Smith *et al.* have shown that the choice of exchange-correlation functional drastically affects the interaction strength of the dative bond between BH_3 and N_2, which ranged from -7.4 to -14.1 kcal mol^{-1}.[171] To further investigate this behavior, a systematic study on the choice of density functionals that adequately describe the BH_3-N_2 interaction was performed. All functionals were evaluated based on their strength of gas interactions and the distance between boron and the closest nitrogen or oxygen atom, respectively (Figure 3.2). The extent of HF exchange greatly impacts the accuracy of the BH_3-N_2 interaction. Pure DFT functionals performed poorly, since they predict overbinding of N_2, with interaction strengths two to three times greater than the reference value. M06-L[199] was found to be the most accurate non-hybrid functional, which provided an interaction energy of -13.7 kcal mol^{-1}. Across the hybrid M06 functionals, altering exchange from low to high M06, M06-2X[200]and M06-HF[201] lowered the relative errors to 40%, 23%, and 21% (Figure 3.2). Similar dependence of the HF exchange on the B-N distance was also observed. For the M06 series (M06-L, M06, M06-2X, M06-HF), the computed distances are 1.53, 1.57, 1.59, and 1.69 Å, respectively, where the interaction energy and distance decreases with increasing exchange. M05-2X[202] outperformed all other DFT functionals for the geometry of the BH_3-N_2 complex, with a B-N distance of 1.62 Å. This finding is in agreement with a recent study showing M05-2X amongst the best functionals for weaker noncovalent interaction geometries.[203] In terms of binding energies, the most accurate functionals were M05[204] and M05-2X[202], with errors of -0.25 and 0.82 kcal mol^{-1} from the coupled-cluster reference values, respectively. Ultimately, the M05 was selected since it provides energies within 0.3 kcal mol^{-1} of the reference in both the BH_3-N_2 and BH_3-O_2 systems. The def2-TZVPPD[121] basis set was found to provide the most accurate results among the basis sets tested (see Appendix Table

10). Thus, the M05/def2-TZVPPD level of theory was used for the examination of larger, functionalized boranes.

3.3.3 SAPT Analysis for BH_3-N_2 and BH_3-O_2

The previous analysis showed that there is a substantial interaction between BH_3 and N_2, described as a weak dative bond. The individual energy contributions of such weak bond were studied with SAPT for understanding how functionalized boranes contribute to intermolecular interactions. The BH_3-N_2 interaction has been studied with absolutely localized molecular orbitals (ALMO) energy decomposition analysis (EDA)[205] at the B3LYP/6-31(+,+)G(d,p) level of theory, a methodology that is in agreement with DFT formulations of SAPT.[206] At 1.56 Å the interaction energy was -17.1 kcal mol^{-1}, but after including monomer relaxation by accounting for the deformation energy, the resulting binding energy is -7.2 kcal mol^{-1}. The ALMO EDA showed a nearly equivalent interaction between the donation from N_2 and π back-donation from BH_3 to N_2.[205] To further this discussion, the nature of this type of interactions has been evaluated by means of SAPT, which is the wave function-based equivalent towards understanding energy decomposition, and provides the interaction in terms of electrostatics, induction, dispersion, and exchange repulsion. Various levels of SAPT were investigated at three interatomic distances, 1.2 Å, 1.56 Å, and 1.64 Å. These results are summarized in Appendix Table 7, which shows that the SAPT2+3+δ_{MP2} nicely agrees with the coupled-cluster reference (without considering deformation energy). The SAPT2+3+δ_{MP2} predicts an interaction energy of -15.1 kcal mol^{-1}, which is within one kcal mol^{-1} of the coupled-cluster reference of -15.9 kcal mol^{-1}. The charge-transfer (CT) as suggested by the Stone and Misquitta[195] (difference in induction energy in the dimer-centered and monomer-centered basis sets), was calculated to be -22.3 kcal mol^{-1}. It should be noted that the SAPT CT scheme includes the stabilization from charge delocalization by considering the dimer basis functions. This results in basis set and intermolecular distance dependence in the charge-transfer term,[207] and for noncovalent interactions at close distances, the distinction between the polarization induction and charge-transfer is somewhat arbitrary,[208] but greater than -20 kcal mol^{-1} charge transfer shows there is significant polarization in the electron density about the intermolecular distance.

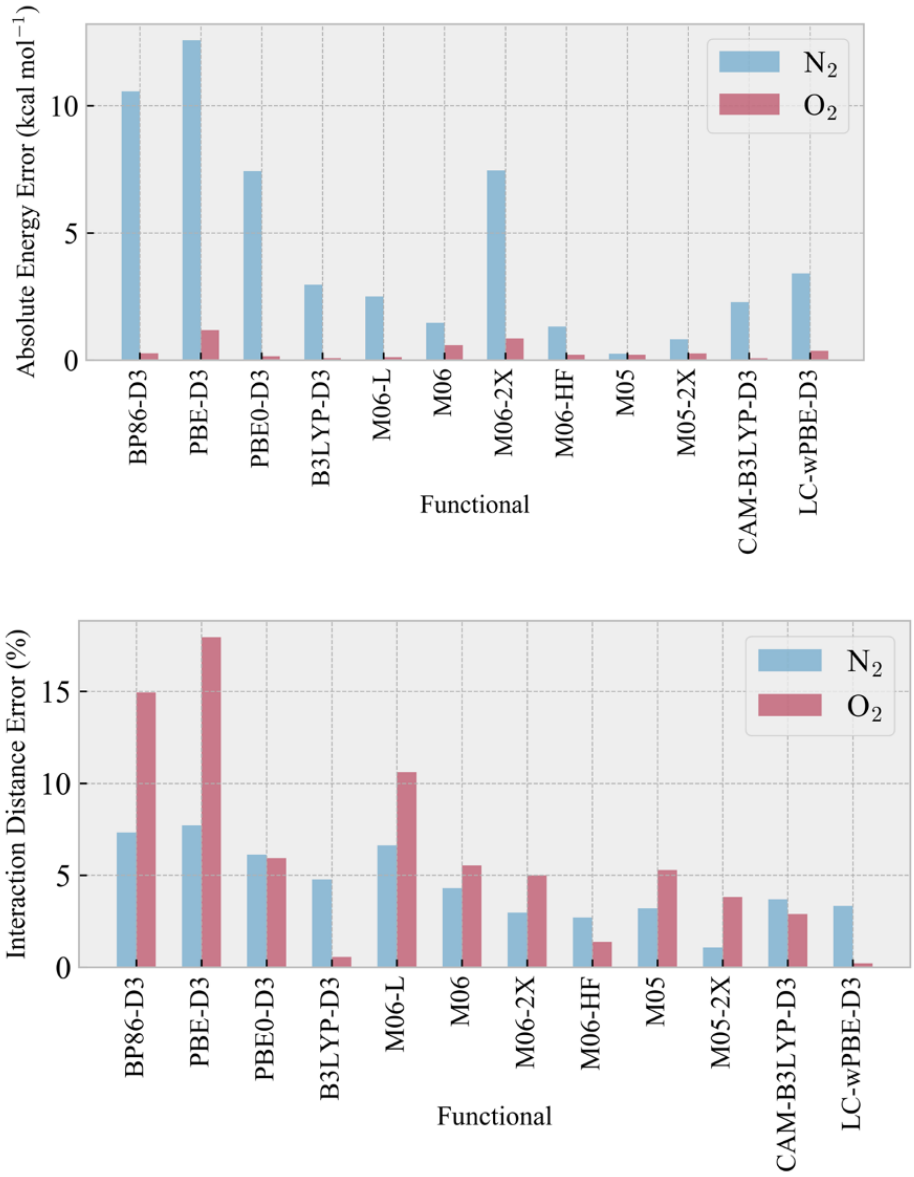


Figure 3.2: Absolute binding energy deviations from CCSDT(Q)/CBS for the BH_3-N_2/O_2 systems are reported for all functionals in the def2-TZVPPD basis (top). The percent error in the B-N and B-O distance at the equilibrium geometry is provided as a percent error (bottom).

A full dissociation curve was calculated at the SAPT2+3+δ_{MP2} level using the B3LYP-D3(BJ)/def2-TZVPP geometries also used in the CCSD(F12) reference calculations and is shown in Appendix Figure 2. The strength of the individual contributions heavily dependent on the distance. However, the total binding energy is much less affected. For example, at 1.50 Å the exchange repulsion is 126 kcal mol^{-1} and lowers to 87.6 kcal mol^{-1} at 1.64 Å. In those distances, the total binding energy drops from -15.5 to -13.4 kcal mol^{-1}. Interestingly, the stronger binding energy at 1.50 Å does not result in a minimum interaction distance due to the larger deformation energy at small intermolecular distances. On the other hand, Oxygen is almost noninteracting with the borane, providing less than -1.0 kcal mol^{-1} stability from induction and less than -1.5 kcal mol^{-1} from dispersion. On the other hand, the borane-nitrogen complex receives more than -40 and -20 kcal mol^{-1} stabilization from induction and dispersion, respectively. Due to the different levels of SAPT, these values are not directly comparable, but it does however demonstrate that the BH_3-O_2 interaction is significantly weaker compared to N_2 with BH_3.

3.3.4 Functionalized Boranes

Monosubstituted Boranes. The interaction between functionalized boranes and N_2/O_2 was evaluated at the M05/def2-TZVPPD level of theory. 33 unique functional groups were examined, while the five most promising molecules for N_2/O_2 separation are discussed here in more detail (Figure 3.3). The M05/def2-TZVPPD binding energies for both BH_3-N_2 and BH_3-O_2 and the corresponding B-N distance and N-B-H angle are reported in Appendix Table 8. The strongest interactions with N_2 were obtained from structures **2** and **5** (-7.49 and -7.39 kcal mol^{-1}, respectively). Structure **5** has the closest B-N distance and largest N-B-H angles (1.586 Å and 104.7°, respectively) but interestingly, structure **2** has a much larger interaction distance of 1.609 Å while also having the strongest binding energy with N_2. Upon applying SAPT2+3(CCD)+δ_{MP2}, the energies of the five functionalized boranes interacting with nitrogen were decomposed into the exchange, electrostatic, induction, and dispersion terms (see Appendix Figure 3). Fragments **1** and **5** show the strongest $E_{exchange}$ due to their closer B-N distances. In all the complexes, the interaction is primarily composed of electrostatics and induction. Within the five complexes, all showed similar stability except

for **5**, which had a much more favorable interaction at -21.19 kcal mol^{-1}, due to a larger $E_{\text{induction}}$ (-50.73 kcal mol^{-1}).

Effect of Multiple Modifications to BH_3. To assess the sensitivity of the binding energies of boranes to functionalizing multiple sites, the strongest binding functional group (-CF_3 group from 2) was chosen for further studies with mono-, di-, and trisubstituted boranes. The M05/def2-TZVPPD binding energies and SAPT interaction energies are presented in the Appendix Table 9. The incorporation of more -CF_3 functional groups increases the B-N distance with each subsequent addition from 1.609 to 1.629 and 1.651 Å, respectively. Despite the larger distance with multiple functional groups, the interaction and binding strength increases. The total interaction energy increases more than the binding energies due to the increase in deformation energy. As the distance increases, the exchange repulsion and electrostatic contribution drop, while the presence of the additional functional groups contributes to the increased dispersion. These effects increase the total interaction energy from -17.90 kcal mol^{-1} to -23.91 kcal mol^{-1}. However, this does not translate to considerably larger binding energy, as the deformation energies offset much of this increase. Ultimately, appending multiple functional groups does increase the nitrogen affinity of the borane; moving from mono-substituted to tri substituted –CF_3 grants a 17% increase in the total binding.

Effects of Functionalization at Distal Positions. The varying interaction energies found in Appendix Table 8 demonstrate how the affinity of the borane to nitrogen can be modified by using various electron withdrawing functional groups. In the examples provided, the functional groups varied at the first or second atom from the boron. It is also interesting to study whether modifications beyond the first coordination sphere will significantly impact the electronic interaction of nitrogen with the borane. To demonstrate this, we have examined variants of **4** due to its intermediate binding energy of -4.82 kcal mol^{-1} and its distal hydrogen which is available for further functionalization. Three substituted structures considered in this study (**4a-4c**) are shown in Figure 3.4. In Table 3.2, the equilibrium B-N bond distance, DFT binding and deformation energy, and the relative weight of the attractive SAPT interactions are shown for **4** and **4a-4c**, and differences in the SAPT interactions between **4** and **4a-4c** are presented in Figure 3.5. Interestingly, the bond

distance stays relatively static across all complexes (within 0.003 Å). However, the binding energy differs by more than 3.0 kcal mol^{-1}, which demonstrates distal functional groups alter the electronic structure of the borane. We found that the methyl substitution in **4a** has the weakest binding at -4.38 kcal mol^{-1}, whereas the cyano-functionalized **4c** increases the interaction energy to -7.63 kcal mol^{-1}.

The SAPT2+3(CCD)+δ_{MP2} results demonstrate how the interactions change as different electron-withdrawing functional groups are considered (Table 3.2). The stronger binding is tied to a higher percentage of the stability due to induction forces. Analyzing **4a** in Figure 3.5, we can see electrostatics and dispersion interactions are more favorable than **4**, whereas the exchange and induction effects are less favorable. Overall, these sum to give weaker binding than **4**. For **4b** and **4c**, the opposite is observed; exchange and induction are more favorable while dispersion and electrostatics terms are less favorable. This suggests that **4b** and **4c** are effectively withdrawing more electron density from boron since the electrostatic and dispersion effects are smaller due to less electron density interacting with the nitrogen. To visualize this concretely, the difference in electron density between the boron and nitrogen was plotted in Figure 3.6, which confirms that 4b and 4c have less electron density near the nucleus of boron. From approximately 0.20 Å to 1.0 Å, there is an increase in electron density, which is a result of the electron density from nitrogen polarizing toward the more electron deficient boron atom. In the range 1.0 to 1.5 Å, **4b** and **4c** show lower electron density, which corresponds to the valence electrons about the nitrogen. In contrast, **4a** has more electron density at the boron atom shows the opposite behavior: more electron density about the nitrogen and less polarization toward boron. This finding is in agreement with the SAPT induction ordering of **4c** > **4b** > **4** > **4a**, which mirrors the ordering of interaction and binding strength. Therefore, even distal functional groups can effectively change the electronic structure and the binding of N_2. Those groups that can pull electron density away from boron allow for more polarization and induction, leading to stronger interactions with nitrogen. In this case, distal functionalization gave a 60% increase in binding energies, showing that such substitutions strongly influence nitrogen binding.

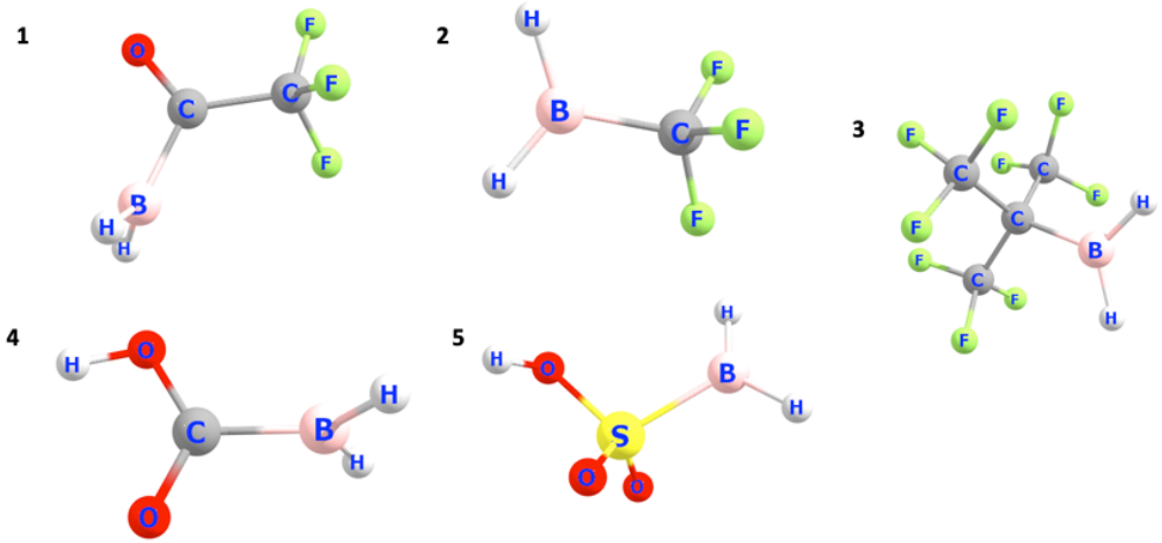


Figure 3.3: Five most promising functionalized boranes with electron withdrawing groups for N_2 binding.

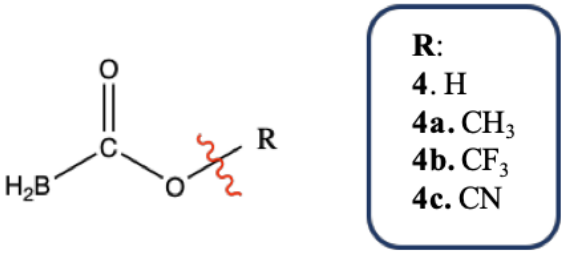


Figure 3.4: Structure **4** and its functionalized variants **4a**-**4c**.

Table 3.2: The DFT interaction energy (kcal mol^{-1}) and distances as provided by DFT for complexes **4** and **4a-4c**. The individual attractive SAPT2+3(CCD)+δ_{MP2} components are also reported as a percentage of the total attractive terms, as well as total interaction energies are also reported.

	4	**4a**	**4b**	**4c**
$\Delta E_{\text{Binding}}$ (DFT)	-4.82	-4.38	-6.29	-7.63
$\Delta E_{\text{Deformation}}$ (DFT)	11.48	11.37	11.10	11.05
$\Delta E_{\text{Interaction}}$ (DFT)	-16.30	-15.74	-17.39	-18.68
B-N distance (Å)	1.614	1.613	1.616	1.616
$E_{\text{Electrostatic}}$	39.2%	39.4%	38.7%	38.2%
$E_{\text{Induction}}$	40.2%	39.8%	40.8%	41.5%
$E_{\text{Dispersion}}$	20.7%	20.8%	20.5%	20.3%
$\Delta E_{\text{Interaction}}$ (SAPT)	-17.33	-16.67	-18.48	-19.68

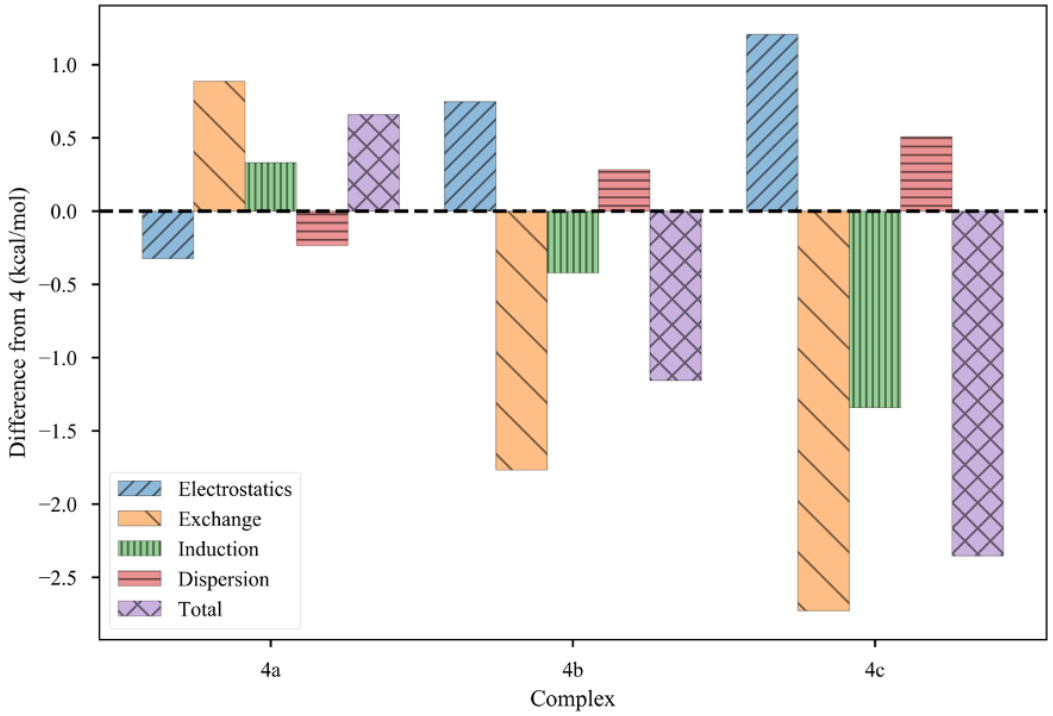


Figure 3.5: The DFT interaction energy (kcal mol^{-1}) and distances as provided by DFT for complexes **4** and **4a-4c**. The individual attractive SAPT2+3(CCD)+δ_{MP2} components are also reported as a percentage of the total attractive terms, as well as total interaction energies are also reported.

3.4 Conclusions and Outlook

We have performed an in-depth study on the BH_3 complex with N_2 and O_2 using a highly accurate CCSDT(Q)/CBS scheme for the calculation of reference binding energies. For the BH_3-N_2 complex, a value -6.26 kcal mol^{-1} with an equilibrium distance of 1.64 Å was obtained, while for the BH_3-O_2 complex a significantly weaker binding energy of -1.22 kcal mol^{-1} at 2.62 Å was computed. Many DFT functionals fail to accurately describe the BH_3-N_2 interaction, providing exceedingly strong interactions with truncated bond distances. However, M05/def2-TZVPPD was shown to accurately model the binding energy, with less than 0.3 kcal mol^{-1} error for both N_2 and O_2. The highly-accurate SAPT2+3(CCD)+MP2 level of theory was used to quantify the stabilizing interactions for many boranes, which were shown to be bound largely due to electrostatics and induction (roughly 40% of the interaction energies, respectively). Next, an examination of many electron withdrawing moieties was performed using the M05/def2-TZVPPD method to determine effective functionalization for strong interactions with N_2. Additionally, the effect of adding multiple functional groups to boranes revealed only modest (20%) increases in binding due to larger deformation energies for the di- and trisubstituted boranes. Lastly, distal functionalization was shown to provide a very effective for increasing N_2 binding energies via modifying induction strength of the borane-nitrogen interaction. Based on our results, perfluorinated (-CF_3), bisulfite (-HSO_3), and functionalized formate groups (-CO_2X, where X is an electron withdrawing moiety) are suggested as promising molecular units that can enhance N_2/O_2 gas separations. Such groups can be introduced in polymeric membranes or in the framework of porous materials (eg. MOFs) for the targeted synthesis of the next generation of functional materials. The potential oxidation of substituted boranes from molecular O_2 is currently examined from our group. These studies will provide further evidence on the utilization of such molecular groups for efficient gas separations.

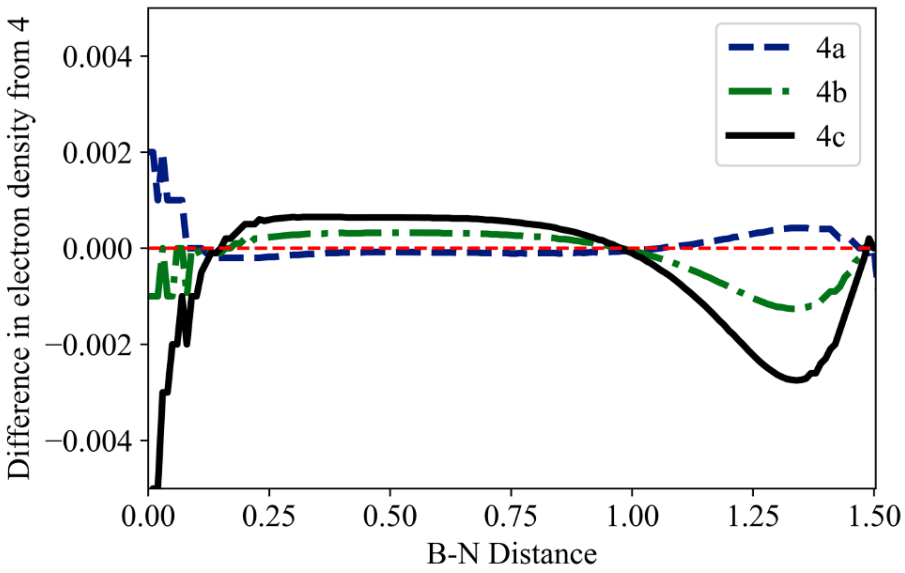


Figure 3.6: The electron density difference between **4** and complexes **4a-4c** along the boron-nitrogen bond. The boron nucleus is located at 0.0 Å

Chapter 4

Representation of Molecular Structures with Persistent Homology for Machine Learning Applications in Chemistry

Disclosure

The following section contains an adapted reproduction of the published manuscript in *Nature Communcations* taken with permission.[209] Jacob Townsend and John H. Hymel and performed the DFT calculations for this study. Jacob Townsend and Cassie Putman Micucci worked on machine learning, method development, and manuscript preparation under the guidance of Professors Konstantinos D. Vogiatzis and Vasileios Maroulas.

Abstract

Machine learning and high-throughput computational screening have been valuable tools in accelerated first-principles screening for the discovery of the next generation of functionalized molecules and materials. The application of machine learning for chemical applications requires the conversion of molecular structures to a machine-readable format known as a

molecular representation. The choice of such representations impacts the performance and outcomes of chemical machine learning methods. Herein, we present a new concise molecular representation derived from persistent homology, an applied branch of mathematics. We have demonstrated its applicability in a high-throughput computational screening of a large molecular database (GDB-9) with more than 133,000 organic molecules. Our target is to identify novel molecules that selectively interact with CO_2. The methodology and performance of the novel molecular fingerprinting method is presented and the new chemically-driven persistence image representation is used to screen the GDB-9 database to suggest molecules and/or functional groups with enhanced properties.

4.1 Introduction

The increasing concentration of greenhouse gases has been identified as a primary factor of many facets of environmental degradation such as higher global temperature, rising sea levels, increased ocean acidity, and more extreme weather-related events. CO_2 is the most prominent greenhouse gas, and its atmospheric concentration has exceeded 400 ppm, which is more than a 40% increase from pre-industrial conditions, potentially leading to a rise in global temperatures of more than 2^o C by the year 2100.[4] Lowering CO_2 emissions is therefore mandatory to meet ambitions to limit temperature increases to 1.5^oC.[5] Advancements in carbon capture and storage technology are desired for meeting these goals of lowered atmospheric greenhouse gas emissions and reduced global temperature increases year to year. At an industrial level, liquid amine-based solvents are used for separation and capture of CO_2 via chemisorption, but the solvent regeneration step is an energy intensive process. Membrane-based technologies offer an alternative, cost-effective process for CO_2. Unlike solvents, where chemisorption involves a reaction with binding strengths exceeding 20 kcal mol^{-1} through the creation of chemical bonds between CO_2 and solvent, membranes utilize much weaker noncovalent interactions. Different types of materials have been suggested for the fabrication of permeable membranes including amorphous, non-porous polymeric membranes,[210–213] or crystalline materials with permanent porosity such as metal-organic frameworks (MOFs) or zeolites.[135]

The understanding of how the atomistic structure of materials affects the gas selectivies is a crucial process for the development of more efficient carbon capture technologies. Most often, this involves the separation of CO_2 from N_2, which are two atmospheric gases with similar kinetic diameters, making size-sieving a challenging task. The introduction of functional groups which selectively interact with CO_2 has been a successful approach for increasing membrane performance.[22,213] Such CO_2-philic functional groups (usually Lewis bases) can be either introduced into the framework of porous crystalline materials (e.g., MOFs) or functionalized into the repeat units of non-porous polymeric membranes. Electronic structure theory calculations between molecular units and the respective gases provide a quantification of these noncovalent interactions, as well as elucidate their nature and properties.[16,24,27,51,214–216] However, the number of potential CO_2-philic groups is intractably large, which leads to an excessive study of such systems with accurate *ab initio* methods. In addition, the determination of gas interaction energies may require multiple calculations to evaluate competitive gas binding sites for every structure, which further increases the computational cost and expert intervention. However, high-throughput computational screening can accelerate the discovery of new, functional materials for rational synthesis through the circumvention of the expensive and time-demanding synthesis and testing process.[217,218]

High-throughput computational design has shown great success in identifying new molecules[47,58] and materials[219–222] with enhanced properties and advanced functionality. For many applications, first-principles studies are essential to virtual screening, but the high computational cost of these methods makes the search of large parts of the chemical space cost-prohibitive. In recent years, machine learning (ML) has become a valuable tool in reducing the cost of a systematic chemical space exploration by enhancing the search for structure-property relationships,[223–225], guiding molecular design,[48,49,226,227] and predicting electronic structure properties.[228–231] ML algorithms are used for their ability to learn complicated relationships in data with high computational efficiency that can be systematically improved through additional training data, but may require extensive training set sizes before predicting out-of-sample properties accurately. The efficiency of ML depends on how these data are passed to the algorithm. For chemical applications, this occurs through

molecular representations, which are the featurization of molecular compounds from their molecular structure into a vector of values. The ML algorithm then infers the relation between the structure and the property of interest. Recent developments in the formulation of molecular representations, particularly in the realm of quantum properties and structure-function relations, have increased the efficiency of ML for chemical applications.[232–238] In addition, such representations generalize to more complex instances such as reaction barriers.[239] Despite the inherent ability of ML to extract important features, ML-model accuracy is dependent on the molecular representation.

A molecular representation reduces the dimensionality of a molecular structure into a chemically meaningful format that relays important chemical information. For example, a chemical formula conveys a three-dimensional molecule as a string of characters but it is an ambiguous input for ML. Atomistic and molecular structure should be converted into a machine-readable format that can be parsed efficiently to ML algorithms as input features. One of the most prominent representations is the Coulomb matrix (CM) introduced by Rupp *et al.*, which is a square atom-by-atom matrix containing an approximate potential energy of the free atom along the diagonal and pair Coulombic potentials on the off-diagonal terms.[230] An improvement over CM is typically observed using the Bag-of-Bonds (BoB) representation, where each atomic pair is placed in specific vectors (bags) based off the elemental pairs and sorted by value.[228] Faber *et al.* have developed FCHL, a representation based on Gaussian distribution functions for the universal kernel ridge regression-based quantum machine-learning models.[240] In addition, the smooth overlap of atomic positions (SOAP) representation[241] calculates the local density of atoms around all atoms in a given chemical environment but suffers from an increased computational cost over the pairwise CM and BoB representations.

Herein, we present a new molecular representation scheme based on persistent homology, a branch of computational topology. Application of persistent homology on molecules encodes three-dimensional structural data into two-dimensional persistence images. Since persistence images hold topological features of chemical structures, we are suggesting them as alternative molecular fingerprints which are transposed into ML input and used to identify relationships in the data. A molecular representation is introduced for encoding chemical

structures, which is applied in the prediction of interaction energies of organic molecules with gas molecules. Persistence images offer a similar-size ML vectorization regardless of system size, and a numerical example is given as a proof of this concept. Whereas CM and BoB do not provide a constant-size representation by this definition, it is effectively achieved through padding empty cells. However, the input vector space takes the dimensions of the largest molecule in the data and requires significant padding for smaller molecules, whereas the introduced method has a predefined vector size regardless of number of atoms. It has been suggested that using feature vectors with sizes independent of system size may result in improved generalization between small and large systems.[242] The new method is used for screening a large database of organic molecules for the discovery of CO_2-philic functional groups.

4.2 Results and Discussion

4.2.1 Persistence Diagrams.

The mapping of molecular structure to a chemically driven persistence image entails several steps. First, the molecular homological features, which measure the connectedness, proximity, and the empty space among the atoms, are computed and stored. These homological features are summarized in a persistence diagram (PD).[243–255] A PD encodes molecular features such as bonds and rings. The PDs can then be vectorized into a persistence image[256] (PI) for use as a molecular representation. However, employing persistent homology and its derivative, PI, purely focuses on detecting topological attributes but lacks explicit incorporation of key chemical information such as element identity, leading to limited applicability in molecular systems. Here, we describe the application of persistent homology with domain-specific knowledge for the generation of persistence images based on atomic properties. The basic steps are presented in the following paragraphs. Anisole was selected as a representative example because it contains two distinct functional units (phenyl and methoxy groups).

To construct a persistence diagram for a given molecule, spheres of a given radius centered at each atom are considered and, as the radius increases, the spheres intersect and lead to the evolution of homological features, called connected components and holes. The connected components encode interatomic distances, while holes describe molecular attributes such as rings and functional groups.

The PDs hold information about the generation or birth and the lifetime length or persistence of connected components and holes. The placement of a birth is denoted by its location along the x-axis of the PD, whereas the persistence is denoted by its location on the y-axis. Birth of connected components occurs at 0, since every atom is given an initial sphere with radius 0 at the start of the algorithm (see anisole as an example on Fig. 4.1). The spheres are then systematically expanded (Fig. 4.1a, d, g, j, and m) until spherical intersections occur, which effectively generate a new connected component by merging older ones. The persistence of connected components is then recorded on the associated PD. In some sense, a PD records the time in terms of spheres' radii for the atoms to form a single cluster. For anisole that is used here as an example, four different types of connected components are generated. The first two appear at 1.1 and correspond to the C–H bonds of the methoxyl and phenyl groups, respectively. The other two appear at 1.4 and correspond to the C–O and C–C bonds, respectively. This means that the units of the two axes are given in angstroms (Å). When sphere intersections lead to the formation of connected atoms (connected component) on a ring, for example when all the six spheres of the phenyl carbons have met, a hole is generated (Fig. 4.1l). The death of a hole occurs when all spheres that form a given hole intersect, and its persistence is recorded on the persistence diagram (Fig. 4.1o).

It becomes now evident that the connected components depend on the distances of neighboring atoms and the holes correspond to topological features of the functional groups. Two holes are now formed for the anisole example of Fig. 4.1, which correspond to the phenyl and methoxy groups. These holes are unique for each respective unit and differentiate between different conformations of the same molecule and subtle differences in geometry.[257–259]

4.2.2 Chemically Driven Persistence Images as Molecular Representations.

The PD is vectorized into a pixelized image, called persistence image (PI), which is a stable, computationally tractable representation.[256] PIs are constructed by placing a Gaussian kernel centered at each point on the PD as it is highlighted for anisole in Fig. 4.2, where the pixel intensity corresponds to the multiplicity. For example, five C–H pairs in the phenyl group, three C–H pairs in the methyl group etc. Next, the surface is transferred into pixel values (Fig. 4.2b, c). The resulting image effectively encodes the molecular geometry.

The transformation of a PD to a PI may lead to inconsistencies if all atoms are treated identically, especially when molecular structures with the same geometries but different atom types are encoded into a PI. For example, the diatomic molecules HBr and F_2 have approximately the same bond distance (1.41 Å), and therefore generate the same PI (Fig. 4.2f, g, respectively). To ameliorate this shortcoming, we introduce atomistic information in the variance of the Gaussian kernels that yielded a PI. The variance determines the spread of the kernel, or how "smeared" each point on the PD is when placed onto the PI. This variance is chosen based on the atom type that created the point in the PD. Specifically, we define the variance in the persistence images by the difference in electronegativity for connected components. Electronegativity differences are chosen because they provide a general description between the nature of different bonds. For the example of HBr and F_2, HBr has a very polar chemical bond, whereas molecular fluorine is nonpolar. Large variance is provided to atom pairs with large electronegativity differences, which ultimately generates unique PIs (Fig. 4.2h, i, respectively). Our new chemically driven persistence image differentiates between molecules which have similar geometric configurations but different atomic compositions.

Another important feature is related to the dimensionality of the PI molecular representation which remains of the same order with respect to the molecular size as we show empirically in Appendix Fig. 4. This is what we call herein a similar-size representation. For example, for a small molecule like anisole that is composed of 16 atoms, a 3 Å by 3 Å PD was generated. The equivalent PD of a medium-size molecule such as the tert-butylcalixarene

(105 atoms) is of comparable size (4Å by 4Å). Similarly, the PD of a large structure, the main protease of the new coronavirus identified as COVID-19[260] (shown in Appendix Fig. 5)in complex with an inhibitor N3 (2500 non-hydrogen atoms) has size of 6 Å by 6 Å. As it was mentioned in the earlier, a similar-size representation is desirable for many chemical applications.[242]

4.2.3 Performance of Persistence Images as Molecular Representations.

Here, we demonstrate the performance of the chemically driven PIs on an application relevant to green chemistry. Our aim is to screen a large molecular database in order to discover molecular groups that show a stronger affinity for CO_2 interaction over N_2. Such molecular groups can be introduced in polymeric materials for the development of the next generation of functional gas separation membranes. Since it is desirable to avoid any density functional theory (DFT)-optimized geometries as input for ML models, which introduce a significant computational bottleneck for the screening of large molecular databases, we resort to structures generated by the OpenBabel[261] software package (gen3d function). Our target is to train a ML model that maps low-cost geometries with accurate quantum chemical data, so it can provide reliable interaction energies for molecular species with geometries generated on-the-fly. We tested the performance of PI as alternative molecular representations that effectively encode chemical structures. The initial subset of 100 organic molecules was used and their interaction energies of each of these structures with CO_2 and N_2 were computed by means of DFT. We also wanted to compare PI with the widely utilized Coulomb matrices (CM), Bag-of-Bonds (BoB), FCHL, and Smooth Overlap of Atomic Positions (SOAP) representations, since each of these are produced with little computational burden and widely implemented in a number of programming libraries.[262–264] PI, CMs, BoBs, FCHL, and SOAP representations were generated for each structure and the performance of each representation scheme was evaluated for the prediction of gas interaction energies. For each scheme, a variety of machine-learning algorithms were tested, including random forest, Gaussian process regression, and kernel ridge regression. Overall, two machine-learning

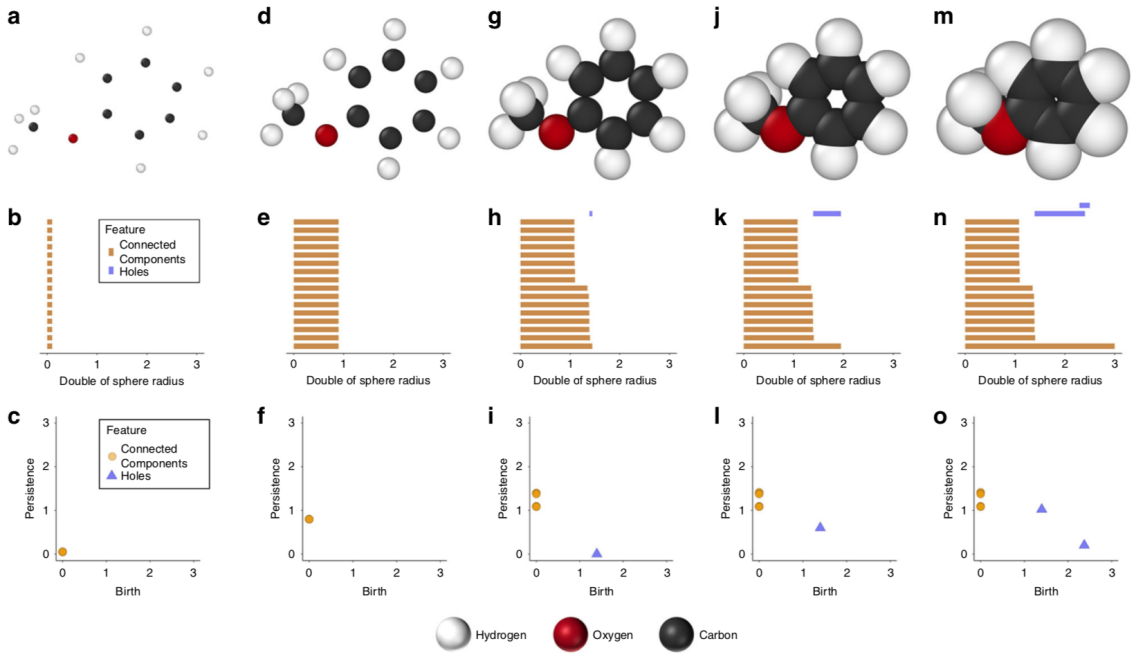


Figure 4.1: Graphical representation of the evolution of the persistent diagram (PD) for anisole. **a** Each atom of anisole is represented by a point. Each connected component is born at time birth = 0 in the evolution of **b** homological features plot and **c** PD. Each bar at the evolution plot tracks the connected components or holes. Note that there are 16 overlapping points in the PD associated with the number of atoms. **d-f** The radius of a sphere centered at each atom is increased, and the connected components continue to persist so the persistence of a connected component is not in its final location. **g-i** When the radius increase leads to sphere overlap between atom pairs, the connected components on the PD are finalized. A hole is created by the six carbon atoms of the phenyl group, and a blue triangle appears in the persistence diagram to indicate its birth. **j-l** The radius is again expanded, and the hole (phenyl ring) still persists. **m-o** A second hole is formed between the methoxyl group and the two carbons of the phenyl group. The death of both holes occurs when their corresponding spheres intersect. The final PD is obtained when the death of all but one connected component and all holes are observed. Note that the bottom bar never terminates and it is always excluded from the PD.

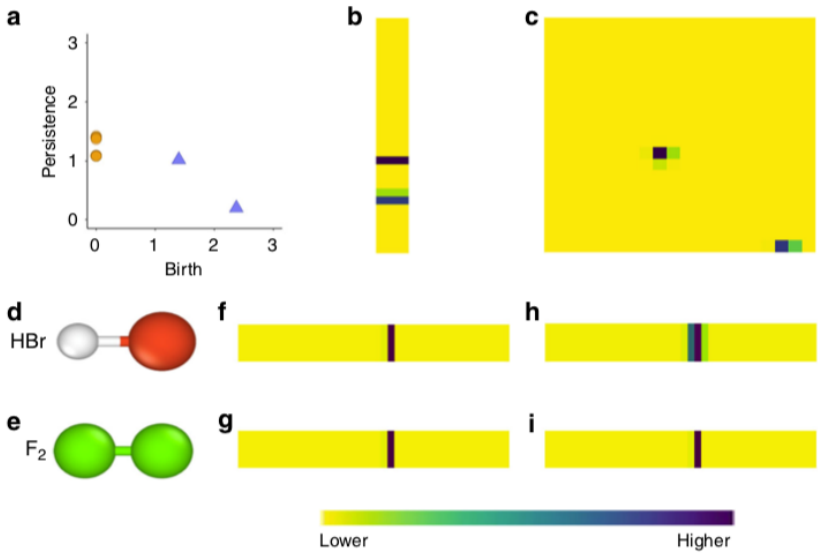


Figure 4.2: Generation of a persistence image (PI) from a persistence diagram (PD).**a** PD of anisole, as shown in Figure. 4.1. The PI for **b** connected components and **c** holes. The molecules **d** HBr and **e** F_2 have the same bond length, resulting in equivalent features on the persistence diagram. Without incorporating the chemically driven PI, the representations in **f** and **g** are equivalent. By incorporating the pairwise electronegativity difference into the variance of the Gaussian kernel, the input vectors **h** and **i** for the two molecules are distinguished by the variance of the vector for HBr.

models were trained per molecular representation scheme, one for CO_2 and one for N_2 interaction energies. The 10-fold cross validation root-mean-squared error (RMSE) for the five trained models on the CO_2 energies are shown in Fig. 4.3. The error bars represent the standard deviation of the RMSE. Similar results were obtained for N_2 interaction energies (see Appendix Figure 6). Comparing the best learners for each representation, CM showed the highest deviation (RMSE of 0.63 kcal mol^{-1}), followed by BoB and FCHL (RMSE of 0.52 and 0.50 kcal mol^{-1}, respectively). The most accurate models were PI with kernel ridge regression (Laplacian kernel, 0.44 kcal mol^{-1}) and SOAP with kernel ridge regression (linear kernel, 0.41 kcal mol^{-1}), where PI showed a tighter variance, yielding a higher confidence in the predictions.

4.2.4 Screening the GDB-9 Database.

High-throughput computational screening using ML is an efficient method to survey molecules for numerous chemical applications. Here, we are applying the PI method for identifying molecules and functional groups that enhance CO_2 interactions with little computational cost. ML models trained on DFT-quality data can estimate DFT-quality results for hundreds of thousands of systems within seconds, while the explicit computation at the DFT level is a cost-prohibitive process. The GDB-9 database[265] was screened, which includes 133,885 organic molecules containing no more than nine non-hydrogen atoms to determine the most promising molecules for CO_2 binding.

The data from the initial 100 organic molecules discussed in the previous section do not adequately capture the properties of the chemical space spanned by the GDB-9 database. The initial training set can be considered as biased since it contains largely N-containing heterocycles with small functionalizations on aromatic carbons. For surpassing this limitation and reliably screening the full GDB-9 space, we applied a methodology known as active learning. In active learning, the training set is systematically expanded to capture the necessary missing physics to accurately predict for the targeted space. The top 40 molecules were selected with respect to predicted CO_2 interaction strength and further investigated by the MD/DFT scheme described in the section 2.1.3. Therefore, the training set was expanded to better infer the relationship between the molecular representation and

the chemical space spanning the GDB-9 database. We have repeated this processes four times by considering different yet optimized molecular representation methods (CM, BoB, SOAP, and PI). The individual steps that were followed are shown schematically in Fig. 4.4 and analyzed in the next paragraphs.

Three iterations were performed with each representation scheme together with the optimized machine-learning algorithm, as it is discussed in the previous section. Thus, the kernel ridge regression (Laplacian) was used for CM and PI, Gaussian process regression for BoB, and kernel ridge regression (linear) for SOAP. The active-learning process resulted a total of 220 data points, i.e., 220 molecular structures with their corresponding CO_2 and N_2 interaction energies computed by DFT per method. The distributions of the interaction energies of these molecules for each method are visualized in Fig. 4.5. We set a mark at 6.0 kcal mol[1] for molecules with significantly strong CO_2 interaction energy. The first iteration contains only the original 100-molecule training set. By expanding the training set with the 40 molecules from the first iteration, the next 40 best-predicted molecules from the model that utilizes the PI representation have significantly improved. On the contrary, no significant changes were observed from the other three models. By the third iteration, the CO_2 distribution from CM, BoB, and PI remained almost unchanged, while a small shift toward stronger interaction energies was found for the SOAP model. Overall, PI showed the greatest performance since each respective iteration increased the number of promising structures, from 10 (first iteration) to 43 (second iteration), and ultimately to 75 out of 120 molecules. Active learning with the PI molecular representation has systematically expanded the training set to better represent the chemical space of the dataset, yielding more reliable predictions every round. Since promising structures are rare within the dataset, this strategy allows the model to account for these rare instances within the training set in a way that would be impossible with a randomly chosen training set. In addition, the three top candidates for CO_2 separations were found to demonstrate stronger interaction energies than 6.50 kcal mol^{-1}, which are shown in Fig. 4.6 (bottom, right). Our computational procedure allowed us to discover new molecules with higher CO_2 affinity that combine previously unknown binding motifs. In particular, we found that cooperative effects between N-containing heterocycles with amino or hydroxo groups at ortho position increases the CO_2

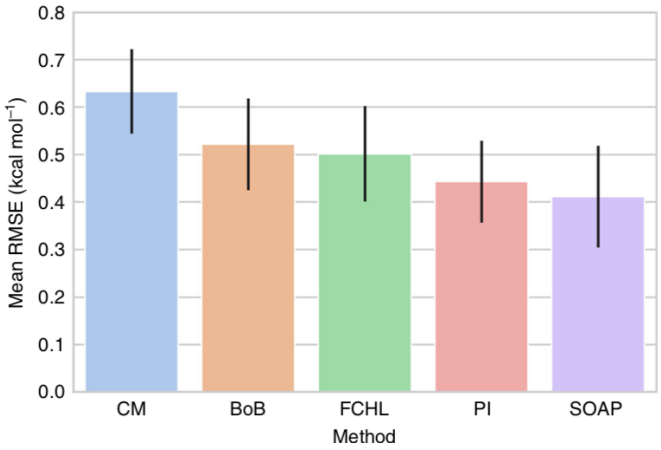


Figure 4.3: Root-mean-square error (RMSE, in kcal mol^{-1}) of CO_2 interaction energies.The automated OpenBabel-generated structures were used for each molecular representation model. The error bars represent the standard deviation of the RMSE.

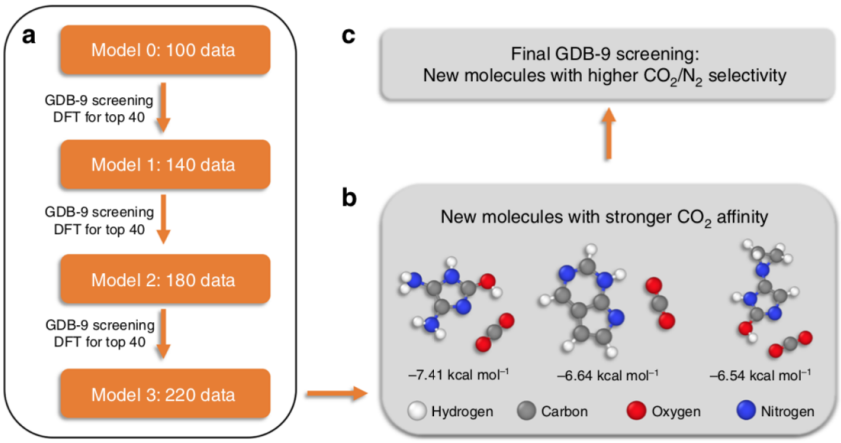


Figure 4.4: Flowchart of the intermediate steps followed for the high-throughput computational screening of the GDB-9 database. **a** The three active- learning steps starting from a model with 100 molecules (Model 0), **b** the results obtained from DFT, and **c** the screening of the GDB-9 database.

strength. The lone electron pair of nitrogen induces a dipole moment on CO_2 that allows stronger interactions with hydrogen atoms of the NH_2 and/or OH functional groups.

After completing three iterations of active learning, the full dataset (220 molecules) is used to create a ML model for predictions on the whole GDB-9 database. For comparison, the database was screened with the four different models, where each of them uses a different molecular representation method (CM, BoB, SOAP, and PI) and data generated from the corresponding active-learning steps. The optimum learner for each method was used, as it was discussed in the previous section, except for BoB, where the kernel ridge regression (linear) was applied. Figure 4.6 includes a plot for each model, where all predicted N_2 and CO_2 interaction energies are on the x- and y-axis, respectively. Only the method that utilized the PI molecular representation was able to identify 4,5-diamino-1H imidazol-2-ol as one of the molecules with the strongest CO_2 affinity as indicated with an orange dot on each plot of Fig. 6. 4,5-diamino- 1H-imidazol-2-ol was part of the training set introduced in the second step of the active-learning process (Fig. 4), and has a DFT CO_2 interaction energy of 7.41 kcal mol^{-1}. All methods agree that the majority of the molecular entries of the GDB-9 dataset have a mean CO_2 interaction energy centered between 3.0 and 4.0 kcal mol^{-1} and a N_2 interaction energy at 2.0 kcal mol^{-1}. Most of the molecules have predicted CO_2 interaction energies between 3.0 and 5.0 kcal mol^{-1}, which emphasize the difficulty in determining new molecules with high CO_2 affinity. However, CM and BoB were less effective in identifying rare instances. On the contrary, results from SOAP were significantly scattered, and predicted many cases with false CO_2 interaction energies close to 0 kcal mol^{-1} (in a few cases, SOAP even predicted repulsive interaction energies). Interestingly, training of machine-learning models with the CM, BoB or the SOAP representations that utilize the molecules identified by active learning with the PI representation yielded more concise distributions, while all models identified the molecular species with the stronger CO_2 interaction energy among the top candidates. In other words, PI provided higher quality data for all methods. Therefore, the model that utilizes PIs for molecular representations provided the most consistent distributions for both CO_2 and N_2 interaction energies. The PI screening revealed a total of 44 of the 133,885 molecules with CO_2 interaction energies exceeding 6.5 kcal mol^{-1}. DFT calculations were performed for verification of these results. It

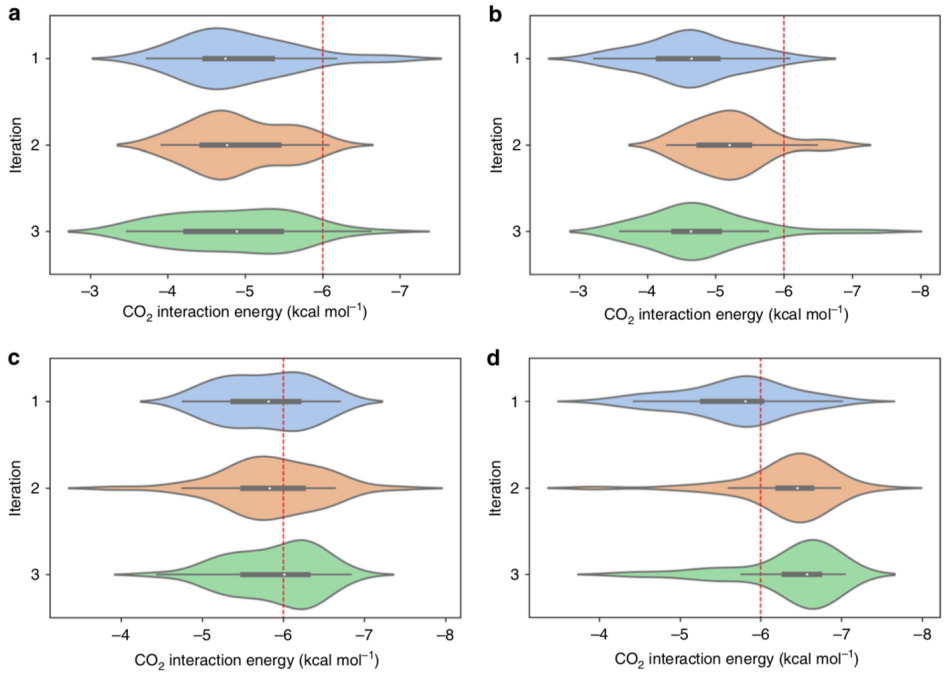


Figure 4.5: CO_2 interaction energy distribution shown as horizontal violin plots for the first, second, and third active-learning steps. (**a**) CM, (**b**) BoB, (**c**) SOAP, and (**d**) PI. The height of the shape shows the frequency of occurrences, the white dot along the axis iteration depicts the median, and the bold line the middle 50% of the data. The thinner line corresponds to the range of all the data. The dashed line at 6.00 kcal mol^{-1} marks cases with stronger CO_2 interaction.

should also be mentioned that SOAP needed 88,287 s for screening the full GDB-9 database, while the screening with the novel PI representation was almost 40 times faster (only 2219 s). All screenings were performed on an Intel i5-4278U processor.

4.3 Conclusions

A novel molecular representation utilizing persistence images with embedded chemical bonding information has been introduced for predicting DFT-quality CO_2/N_2 interaction energies. From our investigation, this new chemically driven persistence image is a concise, computationally efficient, and effective representation that generally outperforms other representations for prediction of CO_2 interaction energies since the computational cost is low and does not suffer from dimensionality problems. This representation accounts for underlying topological structure in the molecule, providing a method to control uncertainty due to differing geometric configurations. The new methodology has been applied for the screening of the GDB-9 database to suggest new CO_2-philic moieties. By using an active-learning approach, our ML-based screening was able to identify many promising molecules in the GDB-9 database despite a very small training set (220 molecules). Specifically, 44 molecules were identified that exceed -6.5 kcal mol^{-1} CO_2 interaction energy. In addition, candidates that may exhibit strong CO_2 interactions while maintaining weak N_2 interactions were examined, yielding a strategy for identifying species with potentially strong gas separation capabilities. Ultimately, chemically driven persistence images are promising molecular representations for larger supermolecular systems due to compact vectorization. Therefore, we believe that the chemically driven PI molecular representations can be applied in a plethora of chemical problems. The PI method described herein relies on a topological representation of a molecular compound that allows a flexible summary of the diversity of the atomic geometries. Due to this flexibility in terms of topological equivalence, the PI is robust and provides accurate predictions in contrast to other methods that need to learn rigid geometric representations. We are currently expanding the applicability of the novel molecular fingerprinting method to high-throughput screening of molecular databases for catalysis and ligand-based lanthanide/actinide separations. For this type of chemical

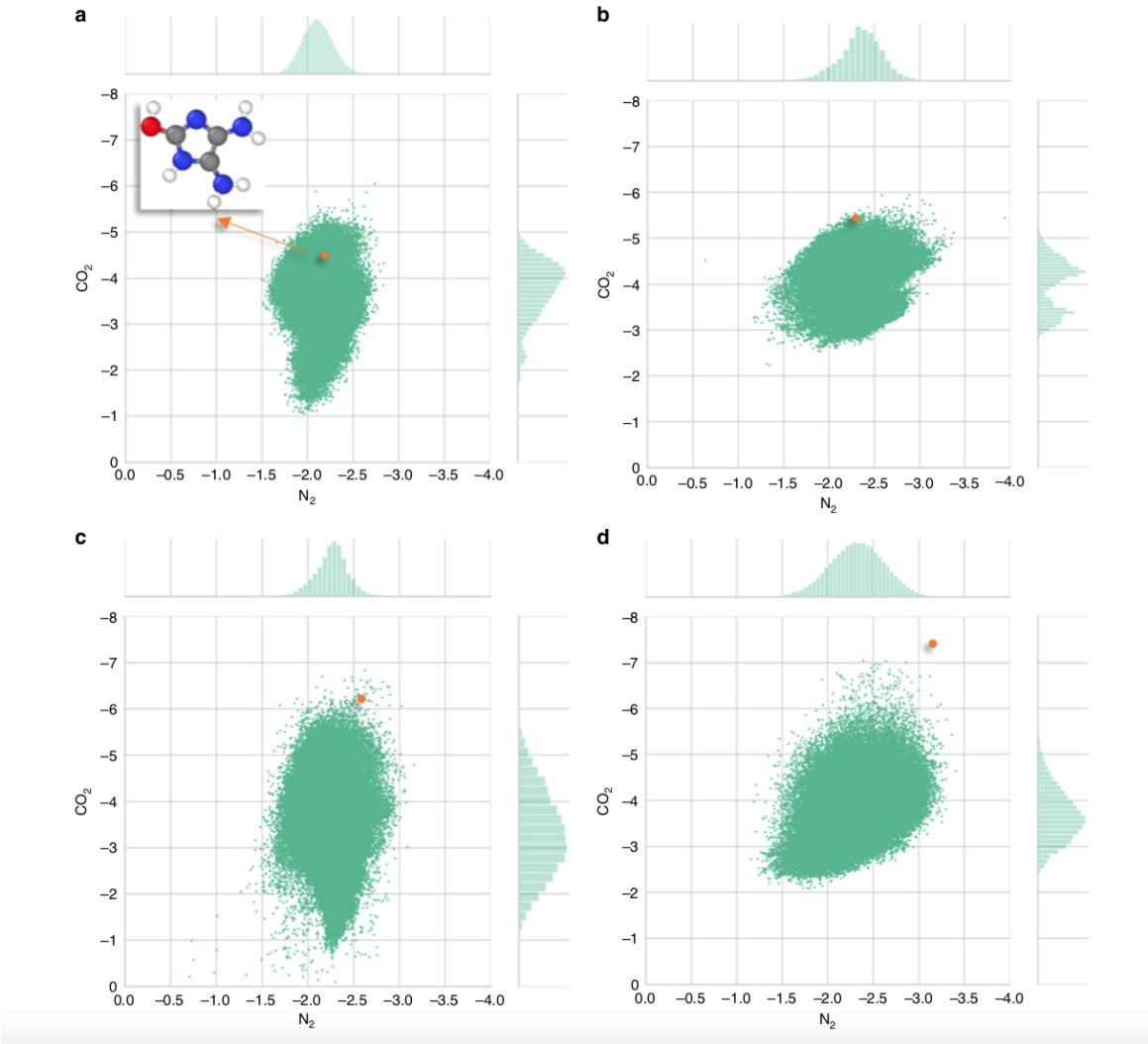


Figure 4.6: Predicted CO_2 and N_2 interaction energies (in kcal mol^{-1}) for all molecules in the GDB-9 database using four molecular representation models. (**a**) CM, (**b**) BoB, (**c**) SOAP, and (**d**) PI. Only the model that utilized the PI molecular representation was able to identify 4,5-diamino-1H-imidazol-2-ol (inset) as the strongest CO_2-philic groups (shown with an orange dot on each plot).

applications, additional features are taken into consideration, such as intensity normalization when a PI is generated from the corresponding PD and predictability of properties of larger molecules from data generated from smaller ones.

Chapter 5

Data-driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver

Disclosure

The following section contains an adapted reproduction of the published manuscript in *Physical Chemistry Letters* taken with permission.[266] Jacob Townsend performed all work in the manuscript under the supervision of Professor Konstantinos D. Vogiatzis.

Abstract

Solving the coupled-cluster (CC) equations is a cost-prohibitive process which exhibits poor scaling with system size. These equations are solved by determining the set of amplitudes (t) that minimize the system energy with respect to the coupled-cluster equations at the selected level of truncation. Here, a novel approach to predict the converged coupled-cluster singles-and-doubles amplitudes, thus the coupled-cluster wave function, is explored by using machine learning and electronic structure properties inherent to the MP2 calculation. Features are collected from quantum chemical data, such as orbital energies, one-electron Hamiltonian, Coulomb and exchange terms. The data-driven CCSD (DDCCSD) is not an

alchemical method since the actual iterative coupled-cluster equations are solved. However accurate energetics can also be obtained by bypassing solving the CC equations entirely. Our preliminary data shows it is possible to achieve remarkable speedups in solving the CCSD equations, especially when the correct physics are encoded and used for training of a machine learning models.

5.1 Introduction

The coupled-cluster singles and doubles with perturbative triples (CCSD(T)) has become the "gold standard" in computational chemistry since it can provide accurate electronic energies and properties for molecules with a single-reference ground state.[267,268] The CCSD(T) scheme has been applied in a plethora of computational studies since it provides accurate atomization energies,[269,270] reaction barriers,[271–273] noncovalent interactions,[214,274] and more. However, conventional CCSD(T) is a computationally intensive method in terms of CPU time and disc space, which limits its applicability to small and medium-size molecular systems; a large basis close to the complete basis set (CBS) limit are needed for achieving the expected accuracy. Solutions to the slow convergence with respect to the basis set include extrapolation schemes[183,269,275–277] and addition of two-electron functions that explicitly depend of the interelectronic r_{12} distance (explicitly correlated methods or R12/F12 methods).[177,178,278,279] Alternative methodologies that speed-up CC schemes are the exploration of the locality of electron excitations,[80] fragmentation schemes,[17,280] pair-natural orbital (PNO) expansions,[17,280–283] linear scaling methods,[284] and high-performance computing.[285] Herein, we present a novel approach towards implementing low-cost, highly accurate CC calculations through machine learning (ML). The data-driven coupled-cluster (DDCC) scheme uses ML for the accurate prediction of the double excitation amplitudes, t_{ij}^{ab}. A new wavefunction is generated which aims to replicate the converged CCSD wavefunction. The machine-learned wavefunction can either be used directly for accurate CCSD-level observables, such as energy, at MP2-level cost, or as an improved guess for the t_{ij}^{ab} amplitudes, which reduces the number of iterations to converge the CCSD equations.

In recent years, machine learning has had an increased role in chemistry as a promising technique since is fast and scalable for a broad range of chemical applications. ML has been widely applied in determining potential energy surfaces[286–297] and has been recently applied as an efficient method for screening the chemical space for quantum-level accuracy prediction and materials discovery.[55,239,298–301] Achieving quantum-level accuracy at low computational cost has received much attention due to developments in machine-learnable molecular representations such as Coulomb matrices,[230] Bag-of-Bonds[302] and many other fingerprinting methods.[228,238,303–308] However, these methods ultimately rely on mapping molecular geometries to properties, which can require extensive amounts of training data and may have limited transferability between systems. Additionally, this may require the optimization of chemical structures in order to properly relate molecular structure to function.

In contrast to mapping properties from molecular structure for the study of materials, ML has been used to improve the electronic structure description directly. For density functional theory (DFT)-based approaches, ML has been applied to directly learn the electron density and for development of exchange-correlation functionals.[60,309–311] Recently, there have also been developments in applying ML to wave function-based methods. For instance, Coe has developed a technique for choosing important electronic configurations in a selected configuration interaction (CI) expansion.[69] Additionally, there have been several implementations to predict coupled-cluster energies using electronic structure techniques. McGibbon *et al.* have developed spin-network-scaled-MP2 (SNS-MP2), which uses Hartree-Fock, MP2 (second-order Møller-Plesset perturbation theory), and symmetry adapted perturbation theory (SAPT0) calculations coupled with an artificial neural network (ANN) to predict noncovalent interaction energies at the CCSD(T) level.[312] Margraf and Reuter developed a method to predict CC correlation energy on systems by implementing a transformation on a subset of the largest MP2 amplitudes modified using a transformed tensor representation.[78] Their technique uses features independent of the number of atoms and includes only electronic structure features but does not demonstrate transferability across different molecular systems. Lastly, Welborn *et al.* have presented a ML method for predicting CCSD energies based on Hartree-Fock localized molecular orbital features,

yielding an accurate method with promising chemical transferability.[73] These previous ML-CC methods have used MP2 or HF wave functions to predict CC energies. Building on these ideas, we present a data-driven approach that uses the MP2 wave function and its implicit properties to predict the CCSD wave function. This new DDCCSD wave function can be used as a starting guess for CCSD iterations for faster convergence of the CC equations or implemented as a low-cost improvement to the MP2 wave function for more accurate properties and energetics.

The CC ansatz uses an exponential form, where the cluster operator typically acts on a canonical Hartree-Fock (HF) reference:

$$|\Psi_{\mathrm{CC}}\rangle = e^{\widehat{T}}|\Psi_0\rangle \tag{5.1}$$

$\widehat{\mathrm{T}}$ is the cluster operator and Ψ_0 represents the HF reference wave function. The cluster operator is truncated for practical purposes, and the most common implementation uses only explicit singles-and-doubles excitations (CCSD). The correlation energy is given by:

$$E_{\mathrm{corr}}^{\mathrm{CCSD}} = \sum_{a<b,i<j} \langle ij||ab\rangle t_{ij}^{ab} + \sum_{a<b,i<j} \langle ij||ab\rangle t_i^a t_j^b \tag{5.2}$$

where ij and ab refer to occupied and virtual orbitals, respectively, and t_i^a and t_j^b are the one- and two-excitation amplitudes that are determined through the iterative solution of the CC equations. Conventional CCSD uses as an initial guess the MP2 amplitudes:

$$t_{ij[\mathrm{MP2}]}^{ab} = \frac{\langle ab||ij\rangle}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{5.3}$$

where ε_k corresponds to the energy of the respective orbital, k. These amplitudes and properties of such orbitals will ultimately comprise the feature set used to design a ML-based wavefunction.

To produce a model that is general enough to be transferable to molecular systems of varying size, orientation and atom composition, the features are restricted to electronic structure properties and, therefore, the features do not explicitly incorporate the identity or position of any of the atoms. Additionally, all the features consist of values already obtained

in the HF or MP2 calculation, adding no additional cost for the generation of the feature set. An obstacle in applying ML to quantum chemical applications is dimensionality: i.e. how do we frame an input feature matrix of amplitudes when the number of amplitudes changes for all systems? Molecular representations of molecules such as the Coulomb Matrix have used padding, where all empty cells are filled with zeros.[230] Here, the size of the input and output matrices will vary, and for the case of amplitudes, the size of the matrix of amplitudes is $(M_{occ})^2(M_{vir})^2$, where M_{occ} and M_{vir} corresponds to the number of occupied and virtual molecular orbitals, respectively. However, this is overcome as the implemented model predicts each amplitude independently. Therefore, only the properties of one instance of a double excitation will be considered by the ML model at a time.

For a given double excitation, many elements from the Fock, Coulomb, and exchange matrices are collected, which can be considered as an extension of the approach taken by Welborn and coworkers.[73] For each double excitation, the following are collected:

- Orbital Energies (as well as the broken-down contributions to these energies from the one-electron Hamiltonian matrix (h), Coulombic matrix (J) and exchange matrix (K), which come from the diagonalization of their respective matrices)
- Coulombic and exchange integrals
- Whether two promoted electrons go into same virtual orbital or not (value of 1 or 0)
- Two electron integral
- MP2 denominator
- Initial MP2 amplitude (see Eq. 5.3)

These features and some mathematical transformations comprise the 30 features for each doubles excitation. Such selection of features provides a representation that contains properties analogous to the HF method used to generate the features; the predictions are invariant to rotation and translation. However, the predictions may not be invariant towards different localization schemes, and therefore, it necessitates the use of the same scheme for the generation of the MOs. It may be beneficial to use localized molecular orbitals, as

many of the amplitudes would tend toward zero values at increasing distances between atoms, which is encoded in the feature set by proxy through the Coulombic integrals.[73] The standard molecular orbitals are delocalized, and this may lead to more non-vanishing amplitudes However, the performance change for such an implementation will be examined in a future study. The k-Nearest Neighbors (k-NN) algorithm was used for the ML-prediction of the double excitation amplitudes. k-NN's have several distinct advantages for a proof-of-concept. First, they are conceptually simple: they measure the distance of a test sample in the feature space from the training samples, take the k closest neighbors and predicts with the average value of those neighbors. In this way, it does not make assumptions about the data and is strictly instance-based learning. Most ML methods fit with a mean squared error loss function, and this would lead to fitting the largest amplitudes more accurately and ultimately neglecting the smaller amplitudes. This was shown to be problematic when attempting to use the ML-predicted wave function as a starting point for iterations to calculate the converged CCSD wave function. However, k-NN's avoid the need for scaling amplitudes as they are not trained via a loss function. Accurate predictions with a simpler method such as k-NN are promising for future implementations using more sophisticated ML algorithms. By default, k-NN algorithms weigh the distance in feature space for all features equivalently. This involves calculation of the Euclidean distance:

$$d(p,q) = \sqrt{((q_1 - p_1))^2 + (q_2 - p_2))^2 + ... + (q_n - p_n))^2} \tag{5.4}$$

$d(p,q)$ represents the distance measured by the k-NN, p and q represent a given test and training sample respectively, and the subscripts denote the feature number up to n features. In practice, this means all features distances contribute equally in their importance toward the calculated total distance used to decide the nearest neighbor, which is undesirable for this implementation, as all the features may not contribute equally or perhaps a small error in a certain feature is more significant than another. This can be visualized by the modified Euclidean distance:

$$d(p,q) = \sqrt{[(x_1(q_1 - p_1))]^2 + [x_2(q_2 - p_2))]^2 + ... + [x_n(q_n - p_n))]^2} \tag{5.5}$$

where the feature weights, x_n, were determined via a grid search on the scaled data to optimize their weights. The most significant features include the corresponding two-electron integral, the MP2-predicted double excitation amplitude, the orbital energy differences, and whether or not the two promoted electrons are being excited into the same virtual orbital. Optimized weights are listed in the Supporting Information. The features were scaled using a MinMaxScaler in the Sci-Kit Learn[313] module to accentuate and normalize the distances between points in the feature space to allow for a truer recognition of neighbors. All k-NN models were also implemented in Sci-Kit learn[313] with the one nearest neighbor (k = 1), which was shown to have the best performance. All HF calculations were performed in the Psi4[105] program. The MP2 and coupled-cluster calculations were performed using the Psi4Numpy[314] spin-factored CCSD implementation.[315] The implementation of the DDCC code described herein uses the CCSD code modified to generate the ML features and predict the CC wave function using k-NNs. All ML results included use the ML predicted double excitation amplitudes (t_{ij}^{ab}) as initial guess, and both t_i^a and t_{ij}^{ab} are further converged as in the conventional CCSD scheme. All structures were generated from simulations at 200K using the OPLS-AA force field,[93] as generated by LigParGen[94] and implemented in the LAMMPS[95] software.

5.2 Results and Discussion

5.2.1 Single Molecule Prediction: H_2O as a Test Case.

There are two metrics that will be used to evaluate the effectiveness: (1) energy differences between the converged CCSD energy and the starting energy before performing iterations and (2) the number of iterations until convergence. Errors in energy will be reported as the mean absolute error in mHartree (mE_h). Another metric to determine the accuracy of the method will be the percent improvement over MP2:

$$\text{Percent Improvement} = (1 - \frac{\Delta E_\mathrm{ML}}{\Delta E_\mathrm{MP2}}) * 100 \tag{5.6}$$

ΔE corresponds to the energy difference between the converged CCSD energy and the initial energy calculated before iterations, and the subscripts ML and MP2 correspond to the respective method's energies. The CCSD iterations include also the relaxation of the single excitation amplitudes, but since their number is significantly smaller than the double excitation amplitudes, we have followed their standard initialization (set to zero and solved separately). The projected equations for the t_i^a amplitudes were solved for all schemes discussed in the next paragraphs. The method was investigated on a set of water molecules at different geometries generated with OPLS-AA.[93] The performance of the method is evaluated on a sample of 1, 5, 10, 20, 40, and 100 water molecules in the training set, and the trained models were tested on a different set of 50 water molecules. The results for both energy differences and iterations for the STO-3G,[316] cc-pVDZ,[184] aug-cc-pVDZ,[317] and aug-cc-pVTZ[317] basis sets are presented in Figures 5.1 and 5.2. Even training on one molecule, the error in energy is significantly lower than the MP2 error, which is depicted for the aug-cc-pVTZ basis set (Figure 5.1, inset). By systematically increasing the number of molecules in the training set, the average deviation from exact CCSD significantly decreases. The larger basis sets require more training examples to reduce error due to the increased number of amplitudes, however they all show systematic improvement in the expansion of training molecules, and with 10 training structures, all the average deviation is from CCSD is sub-mE_h. With 40 training examples, the average deviation for all basis sets is below 0.2 mE_h, or approximately 98% of the difference between MP2 and CCSD. By using 40 molecules for training, the energy difference from the true CCSD correlation energy has nearly converged (Figure 5.1) and thus, 40 training examples will be used in the following applications, unless otherwise noted.

Another important consideration is the number of steps (iterations) taken to solve the CCSD equations. By building the wavefunction with the DDCC amplitudes, the number of steps to convergence is reduced significantly (Figure 5.2). Conventional CCSD converges in eight iterations for all basis sets, except cc-pVDZ which converges in seven steps. The 100-water k-NN models result in up to a 40% reduction in iterations in the test set.

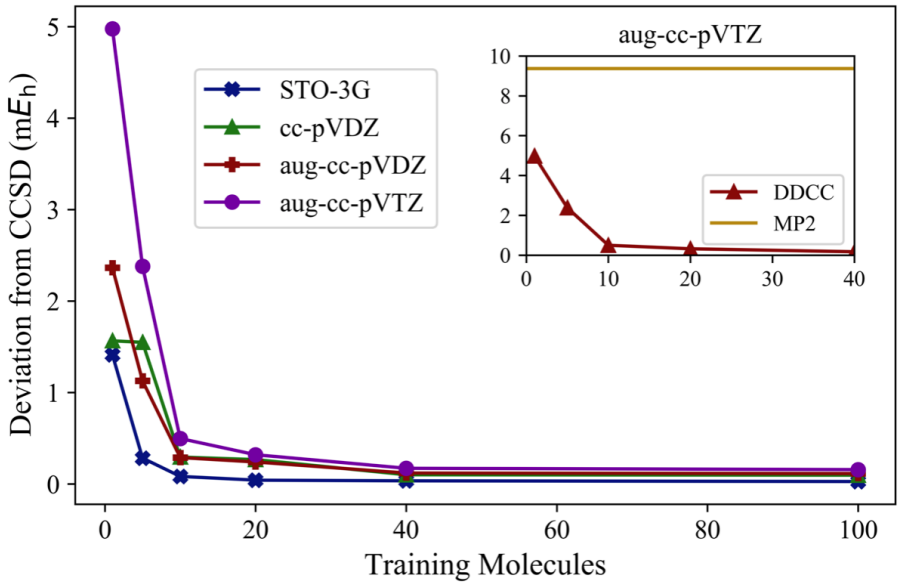


Figure 5.1: Average absolute deviation from the DDCC energy with the converged CCSD energies for the 50 waters in the test set by varying the size (number of waters) in the training set for four basis sets. For comparison, the MP2 deviation from the CCSD results in included for the aug-cc-pVTZ basis set.

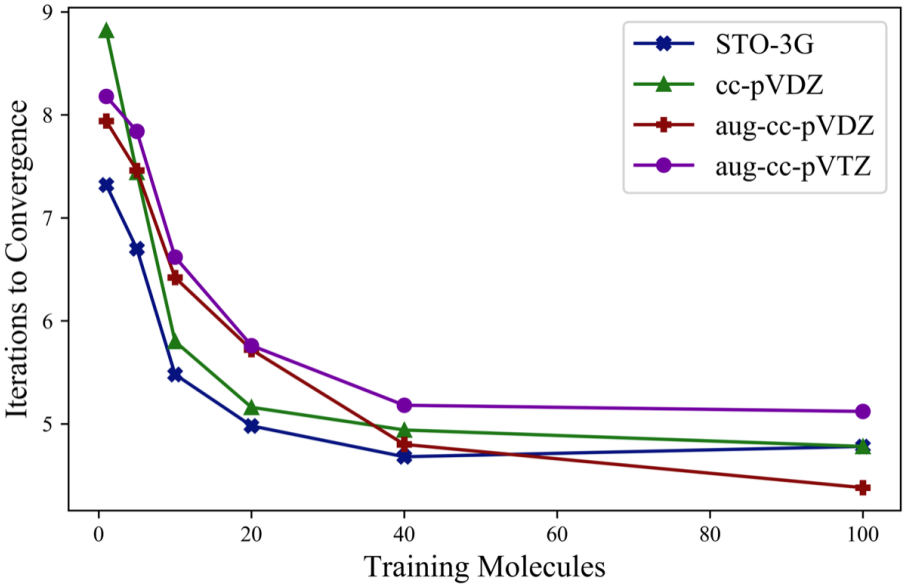


Figure 5.2: Number of steps to converge the CC equations based on the number of training molecules in the training set.

5.2.2 Transferability: Data from Seven Small Molecules.

DDCC is able to predict better starting amplitudes for the coupled-cluster wave function than MP2. For demonstrating transferability and model performance, we have trained ML models where electronic structure data from other molecules are used which are tested on a set of seven small molecules. Eight different models were created using electronic structure data from seven different molecules obtained with the 6-311G basis set[318]: water, methane, ethylene, ethane, methanol, ammonia and formaldehyde. Each model was trained on one molecule type and tested on all seven (ML-molecule name of Figure 5.3). Additionally, an eighth model was trained using electronic structure data from all of the molecules combined (ML-All of Figure 5.3). Figure 5.3 shows the summary of the results from these calculations on sets of 50 test molecules in the form of a heat map which plots the percent improvement.

A remarkable percent improvement is achieved when each molecule uses a model trained from data of the same type of molecule (diagonal of Figure 5.3). Off-diagonal cases show varying degrees of accuracy. For instance, formaldehyde test cases see a large percent improvement from the model trained on water whereas, it does not perform better with the model trained on ethylene. Such performance enhancements are not intuitive, and it is unclear when a performance enhancement will occur, but it is probably related to lack of data that capture the physics (vide infra) of a specific atom type. For example, the ML-water model predicts well (80% improvement or greater) for five of the seven molecules, but its energy is not well predicted by the models generated from other molecules. However, the ML model trained on the data with all seven molecules shows impressive improvement for all molecules with at least 98% improvement, which corresponds to less than a 0.5 mE_h error for all molecules. Additionally, a 15-60% reduction in iterations to convergence was observed.

5.2.3 Molecular Clusters: Extrapolating from Few Water Molecules to Larger Clusters.

While it is useful to be able to predict the energies of isolated molecules, as in the seven molecule heatmap (Figure 5.3), it is also useful to generate accurate energies for systems

containing several interacting molecules. A successful model must be able to capture the physics of a larger interacting system based on those provided by smaller subsystems. To demonstrate this, models have been trained using conformations containing between one and three waters to predict the amplitudes of four different clusters containing six (book, prism, cage, and ring conformations), eight (D_{2d} and S_4) and 16-water clusters (Figure 5.4) from Miliordos *et al.*[319] An optimized monomer, 11 dimers generated through a relaxed dissociation, and 24 trimer systems were used for training. The goal is to demonstrate that these smaller systems are capable of capturing the physics of the larger clusters. Due to the larger system sizes, the STO-3G basis was used for testing purposes.

For demonstration purposes, models that use data from each supersystem structure will be added on top of the previous model's data - for example, the trimer model will contain data from the monomer, dimer, and trimer systems. Therefore, we can examine how interactions of the larger supersystem are captured given the increasing system complexity through the monomer, dimer, and trimer systems. The deviations and improvements for each of the models on the four 6-molecule supersystems are shown in Figure 5.5. The model based on only isolated water failed to predict hexamer test cases, and none of the ML systems showed an improvement on the MP2 energy. This finding is in line with intuition, that the physics of the intermolecular interactions between the water molecules are not captured from an isolated water. However, by adding data from dimer conformations in the training set, the errors drop drastically (less than 9.0 mE_h for all systems) which results in a 90% improvement over MP2. By including trimer systems in the training set, the DDCC deviations from the CCSD energy drop to below 1.0 mE_h for all systems, which results percent improvements above 99% for all of the 6-molecule systems. To further demonstrate the extrapolation, the 3-water k-NN model was used to predict the CCSD energy on two different eight water clusters (D_{2d} and S_4) and a 16-water system. The absolute errors were 9.65, 15.92, and 8.33 mE_h, respectively. These errors correspond to a 90.5%, 81.3%, and 95.7% improvement over the MP2 energies, which demonstrates the model of 3-waters can be extrapolated to much larger systems and the DDCC scheme demonstrates a significant improvement to the MP2 wavefunction.

5.3 Conclusions

This study herein demonstrates the prediction of the CC amplitudes with a great deal of efficiency. In the first test case (convergence of CCSD for one water molecule), the ML-predicted amplitudes resulted in an energy two orders of magnitude closer to the true CCSD correlation energy in 40% the number of iterations on average for all basis sets tested. To demonstrate the transferability of the proposed methodology, seven ML models were trained on data from different molecules, plus a model which incorporated data from all seven molecules. Across the board, all molecules showed to have improvement over the MP2 predicted amplitudes using the model trained on all seven molecules, meaning more data, in general, should lead to more accurate models using k-NNs. This shows that the model can be extrapolated to new molecules with good accuracy. Lastly, subsystems containing few waters (1-3) were used to predict larger systems of water (6,8,16). It was found that the models including one, two, and three water could predict 6-water systems with a 99% improvement over MP2, since the chemical interactions between water molecules were encoded successfully. The 8- and 16-water systems also showed significant improvements. In conclusion, we have shown that machine-learned coupled-cluster amplitudes can provide a better starting approximation to the CC wave function than MP2 amplitudes and are capable of extrapolating to larger systems. The suggested methodology is not alchemical since the exact CCSD energies can be computed by solving the exact CC equations, and can be used in a plethora of chemical applications (such as the calculation of full potential surfaces from only a few structures, efficient energy and properties of molecular clusters from smaller subclusters). Our methodology is transferable to other methods that use iterative solvers, such as the complete-active space second-order perturbation theory (CASPT2) methods, which are currently being examined in our group. Finally, a data-driven scheme for the accurate computation of the (T) term is also underway.

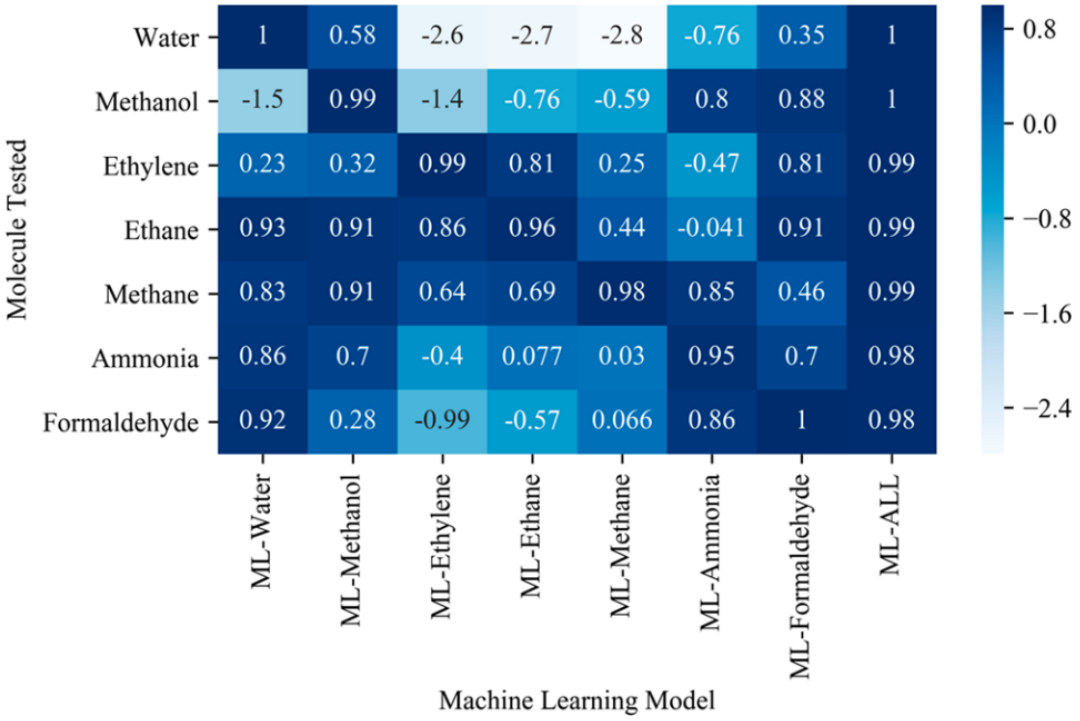


Figure 5.3: Heatmap showing the performance of different k-NNs on predicting different molecules where the value being plotted is the percent improvement (Eq. 5.6).

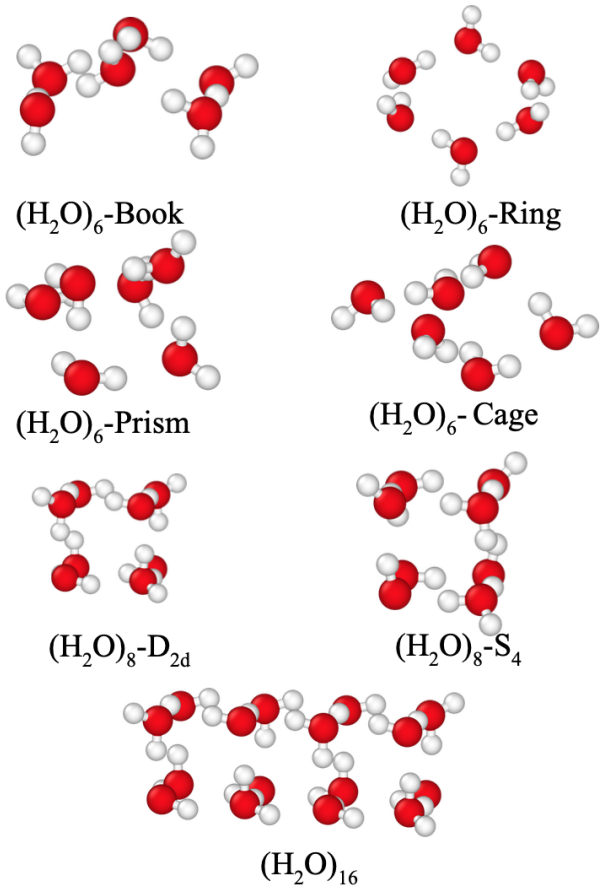


Figure 5.4: Molecular structures of water hexamers in the book, ring, prism, and cage conformation, water octamers in D_{2d} and S_4 conformation, and of ordered water decahexamer conformation.

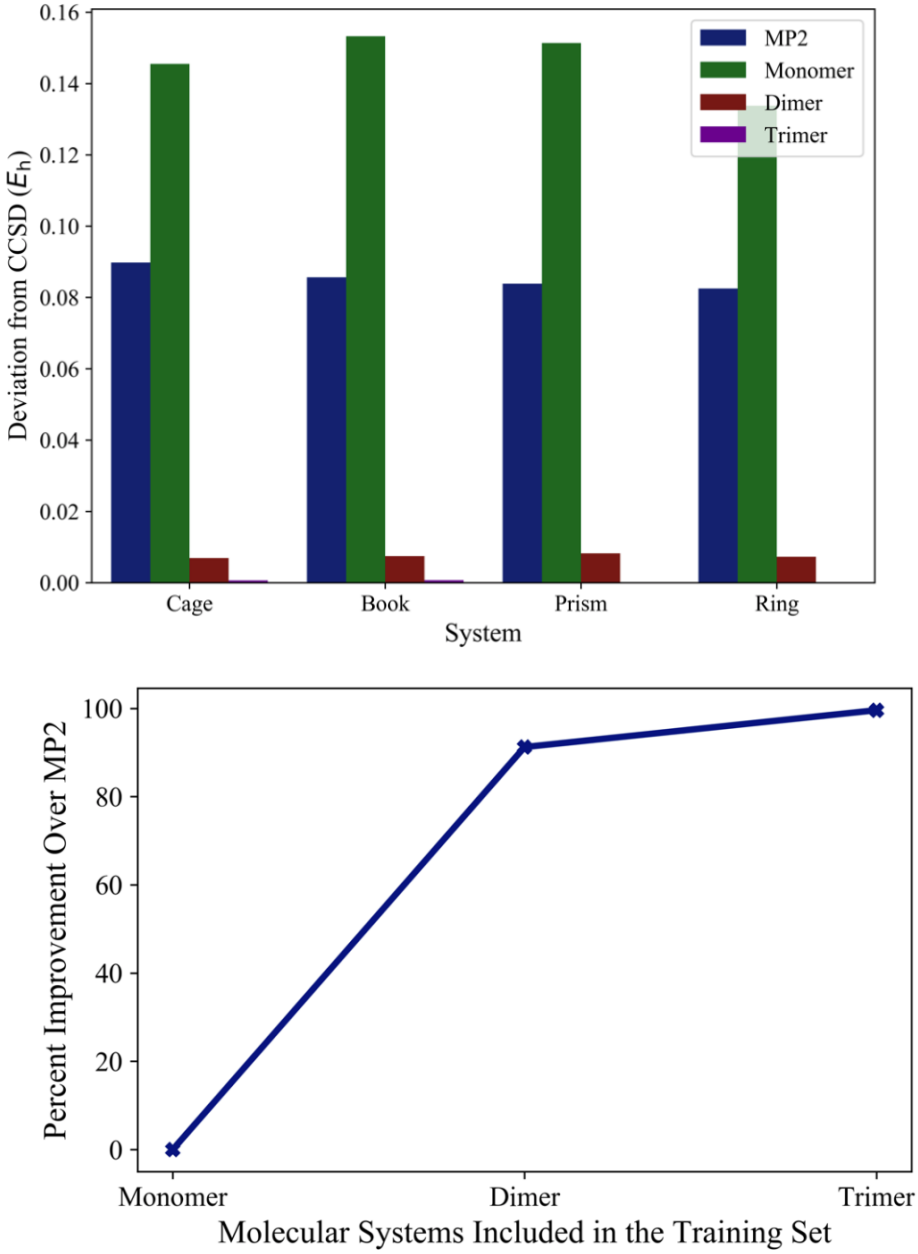


Figure 5.5: The DDCC absolute deviation (E_h) from the converged CCSD is plotted for the four systems containing data with the monomer, dimer, and trimer systems, as well as from CCSD using from MP2 amplitudes (top). The average percent improvement for the four systems at each level of physics is included (below).

Chapter 6

Transferable MP2-based Machine Learning for Accurate Coupled-Cluster Energies

Disclosure

Jacob Townsend performed all work in the following section under the supervision of Professor Konstantinos D. Vogiatzis. This work has been accepted in the *Journal of Chemical Theory and Computation.*[320]

Abstract

Machine learning methods have enabled the low-cost evaluation of molecular properties such as energy at an unprecedented scale. While many of such applications have focused on molecular input based on geometry, few studies consider representations based on the underlying electronic structure. Directing the attention to the electronic structure offers a unique challenge that allows for a more detailed representation of the underlying physics and how they affect molecular properties. The target of this work is to efficiently encode a lower-cost correlated wave function derived from MP2 to predict a higher-cost CCSD wave function based on correlation-pair energies and the contributing electronic promotions and

integrals. The new molecular representation explores the short-range behavior of electron correlation and utilizes distinct models that differentiate between two-electron promotions from the same molecular orbital or from two different orbitals. We present a re-engineered set of input features that provide an intuitive description of the orbital properties involved in electron correlation. The overall models are found to be highly transferable and size-extensive, necessitating very few training instances to approach chemical accuracy of a broad spectrum of organic molecules. The efficiency and transferability of the novel representation is demonstrated on a series of linear hydrocarbons, the potential energy surface of the water dimer, and on the GDB-9 database. For the GDB-9 database, we found that data from only 140 molecules are adequate to achieve chemical accuracy for more than 133,000 organic molecules.

6.1 Introduction

Machine learning (ML) methods continue to have increasingly prominent role in many scientific fields including chemistry. Such techniques have been widely applied for evaluating potential energy surfaces,[286–297,302,321] drug discovery,[46,322–327] and material and molecular design.[44,55,239,298–301,328–331] The application of ML to the prediction of quantum-mechanical energies or properties provides massive computational savings by encoding a molecular structure via a fingerprint or molecular representation,[209,230,241,242,302–304,332,333] where a popular strategy is to predict energetics, typically at the density functional theory (DFT) level.[209,239]

While most ML applications in chemistry have been directed towards connecting physical structure to desired properties, a significant effort towards developing ML methods that directly improve the electronic structure description of molecules and materials has been explored.[59] For example, ML has been adopted to improve density functionals or bypass the Kohn-Sham equations altogether.[60,309–311,334–336] Lei and Medford have used convolutional neural networks to map Maxwell-Cartesian spherical harmonic kernels for functional construction.[335] Chandrasekaran *et al.* have introduced a representation to map atomic environments of grid points to generate electron charge densities, and therefore bypassing

the Kohn-Sham equations altogether.[334] Similarly, wave function-based ML approaches have been developed which largely focus on the reduction of their substantial computational cost. Coe has developed a ML method to learn a selective configuration interaction (CI) expansion more efficiently than the conventional perturbative or Monte Carlo-based sampling,[69] a method that has been applied to potential energy curves.[70] Additionally, Peng-Jian and coworkers have developed an artificial neural network (ANN)-based ansatz to solve complete active space (CAS)-CI wave functions.[337] Schütt and coworkers have also shown a remarkable technique to predict quantum mechanical wave functions using a localized basis of atomic orbitals based on atoms and atom-pairs, which can be used to accelerate convergence of the self-consistent field procedure.[338]

In particular, efforts have been made to model the coupled-cluster methods, as the coupled-cluster singles-and-doubles with perturbative triples [CCSD(T)] has been denoted the "gold standard" in computational chemistry for its accuracy in systems well-described by a single reference state.[267] Nudejima and coworkers performed an ML-based energy density analysis and were able to predict complete-basis-limit-extrapolated CCSD(T) energetics with small-basis Hartree-Fock (HF) densities obtained from small basis sets.[339] McGibbon *et al.* predicted CCSD(T) interaction energies using MøllerPlesset perturbation theory (MP2) and symmetry adapted perturbation theory (SAPT0) as features for an artificial neural network.[312] Margraf and Reuter developed a method to predict CCSD correlation energies utilizing a representation based on the MP2 amplitudes.[78] Miller *et al.* have developed a transferable molecular-orbital-based approach for the calculation of CCSD(T) energetics using Fock, Coulomb, and exchange matrices obtained from HF theory using Gaussian Process Regression (GPR).[73,74] More recently, they have demonstrated that regression clustering can be used to significantly reduce the training times of GPR by producing an ensemble of smaller models.[75] Lastly, there have been efforts to extend the approach of using HF orbitals to predict correlated energies using density matrices.[76,77]

In our previous work, we introduced the prediction of CCSD wave function parameters, the double excitation amplitudes, based on properties of the MP2 wave function.[266] In this work, we expand on the previous study to provide a new method based on the MP2 wave function and predict highly accurate CCSD energies, with particular focus is the role of

localization of the electron correlation and its performance on machine-learned energetics. With traditional post-Hartree-Fock methods, the effective correlation space grows with molecular size, which is cost-prohibitive for methods that incur harsh computational scaling such as coupled-cluster.[268] However, the short-range locality of electron correlation has been targeted to reduce computational expense since its introduction to correlated methods in the 1980s,[340] in realization that the electron promotions associated with correlation energy should not grow with molecular size. Localized correlation methods omit unimportant electron configurations from the correlated domain or approximate them with a lower level of theory to capture correlation at a fraction of the computational expense.[341–345] However, a perhaps underutilized property of localized orbitals is the transferability of correlation contributions,[79–82] which has been utilized within machine-learning of correlation[73–75,77] but its role in the success of such methods has not been well-explored. Our goal is to introduce a novel representation to accurately predict correlation energies within chemical accuracy of the respective method (1 kcal mol^{-1}) that are chemically transferable with respect to changes in molecular structure and system size based on the underlying MP2 wave function. In Section 6.2, the construction of the representation is explained, and its properties are explored in Section 6.4.1. The transferability with respect to size and system are examined in Section 6.4.2 and 6.4.5, respectively. The examination of the performance of the model when the underlying MP2 performs poorly is demonstrated on the carbon dimer in Section 6.4.3. The potential energy surface of water dimer computed by the new methodology is discussed in Section 6.4.4, while basis set effects are presented in Section 6.4.4. Finally, the conclusions from this study are summarized in Section 6.5.

6.2 Theory and Implementation

In CCSD theory, the projected CC equations are solved to determine the converged cluster amplitudes which subsequently determine the energy as shown in Equation 5.2, and an initial guess is usually provided by the MP2 amplitudes (see Equation 5.3) Therefore, to circumvent the solution of the CC equations, one needs to generalize a representation that can effectively map between the two-electron promotions of MP2 to the converged cluster amplitudes. One

way to accomplish this is to rewrite Eq. 5.2 in terms of Nesbet's theorem as:

$$E_{\text{corr}} = \sum_{i>j} \varepsilon_{ij} \tag{6.1}$$

Where each ε_{ij} is the total pair-correlation energy corresponding to the occupied orbital-pair ij. Each ε_{ij} is composed from a square matrix containing elements from Equation 5.2, or more explicitly

$$\varepsilon_{ij} = \sum_{a>b} e_{ij}^{ab} \tag{6.2}$$

$$e_{ij}^{ab} = \langle ij||ab\rangle (t_i^a t_j^b - t_i^b t_j^a + t_{ij}^{ab}) \tag{6.3}$$

where ab run over the indices of the number of virtual orbitals. In principle, Equation 6.1 would correspond to the CCSD correlation energy if the exact coefficients comprising the two-electron coefficients were known $(t_i^a, t_j^b, t_{ij}^{ab})$. While the basis of our approach to predict pair energies was inspired by the work of Miller *et al.*,[73] our approach is to focus on the using the properties of the underlying MP2 wavefunction. By exploiting the locality and transferability of electron correlation, we aim to find a systematic, learnable connection between MP2 and more accurate correlated methods. We hypothesize that it possible to learn the connection between the MP2 and CC wave functions and provide a consistent methodology for the accurate prediction of CC energies by circumventing the solution of the projected CC equations.

In this paragraph, we turn our attention on the development of the new representation for correlation-pair energies ε_{ij}. Based on the notion that electron correlated is local (short-range) and transferable, we introduce a Δ-ML implementation[229] where the MP2 pair-correlation energies, $\varepsilon_{ij}^{\text{MP2}}$, and their associated promotions and integrals were used to predict the CCSD pair-correlation energies that are summed to produce the total CCSD electron energy. Therefore, the goal is to learn the systematic connection between MP2 and CCSD correlation. With these aims in mind, the resulting representation aims to capture as much of the full MP2 correlation energy as possible while keeping a compact feature matrix for computational tractability. In our implementation, diagonal and off-diagonal ε_{ij} elements

were predicted on two independent models, an approach that enhances model flexibility and increased accuracy. The rationale behind this selection is that the correlation energy term ε_{ii} of two electrons promoted from the same molecular orbital i has higher contribution to the total CCSD energy than the off-diagonal terms ε_{ij}. Overall, the CC correlation-pair energies, ε_{ij}, were predicted from the following features

- $\varepsilon_{ij}^{\mathrm{MP2}}$
- $\langle ii||jj\rangle$
- $e_{ij[\mathrm{MP2}]}^{ab}$ matrix
- $\langle ij||ab\rangle$ matrix
- $\langle ii||aa\rangle$ matrix
- $\langle jj||bb\rangle$ matrix
- $\langle aa||bb\rangle$ matrix
- MP2 t_2 amplitude matrix
- Missing $\varepsilon_{ij}^{\mathrm{MP2}}$ correlation in representation *(vida infra)*

where all the matrices have been sorted with respect to the energy contributions of $e_{ij[\mathrm{MP2}]}^{ab}$. The two-electron integrals were added to indirectly provide additional information about the energy contributions, $e_{ij[\mathrm{MP2}]}^{ab}$, and their corresponding orbitals. For the diagonal elements, ε_{ii}, the $\langle ii||jj\rangle$ were excluded (since $i = j$) and thus, their corresponding model has one less feature. In the spirit of localized correlation methods, each of these matrices were truncated to a fixed number of elements. Specifically, the most positive and most negative contributions were included in the representation to ensure the representation does not contain correlation exceeding the MP2 value. This truncation parameter is chosen based on the chemical application (vide infra) but can be considered a hyperparameter; however, a general consideration is that larger basis sets will require more elements from these matrices as the electron correlation becomes more delocalized over a greater number of virtual orbitals. This study uses 20 two-electron promotions, since larger feature spaces did

not lead to increased accuracy and utilized more computational resources in model training. The final feature, the missing $\varepsilon_{ij}^{\mathrm{MP2}}$ correlation, contains the truncated MP2 energy not contained in the 20 $e_{ij[\mathrm{MP2}]}^{ab}$ values. In the construction of a model based solely on electronic structure, the produced model is invariant with respect to molecular rotations, translations, and permutations.

6.3 Computational Details

All Hartree-Fock calculations were performed in the Psi4 program.[346] MP2 and CCSD calculations for training used a modified Psi4Numpy[314] spin-factored CCSD implementation.[315] The Psi4 program package was used for the computation of CCSD energies for evaluating the accuracy of the trained models. MP2 and CCSD calculations utilize the frozen-core approximation. Orbital localization was performed with a threshold of 10^{-12} on both the occupied and virtual orbital sub-blocks. In this study Boys[347] and Pipek-Mezey[348] (PM) orbital localization schemes were used. All calculations were performed with the 6-31G[349] basis set except where the basis set effects were explored, where STO-3G[316] and cc-pVDZ[184] were used as representative smaller and larger basis sets than the 6-31G, respectively. All features were scaled using MinMaxScaler in the Sci-Kit Learn[313] module, and XGBoostRegressor in the xgboost package was used for regression due its scalability, efficiency, and accuracy.[350] While kernel-based methods such as kernel ridge regression and Gaussian process regression were also tested and often provided superior performance for smaller training sets, their N^3 scaling with respect to training instances was impractical for larger training sets. For consistency we have chosen to use xgboost throughout this study, which follows $Nlog(N)$ scaling. Neural networks were also tested but did not produce greater accuracy for this application. Hyperparameters were selected via a 3-fold training cross-validation amongst training examples to optimize the depth and child weight, and L2 regularization for the models.

6.4 Results and Discussion

6.4.1 Properties of the New Representation

As previously discussed, in the generation of the representation for each pair energy, ε_{ij}, the corresponding orbital space comprising e_{ij}^{ab} were truncated to a fixed number of electron promotions. Thus, the model only considers a fraction of the promotions containing the full $\varepsilon_{ij}^{\mathrm{MP2}}$. In principle, localizing the orbitals should concentrate the e_{ij}^{ab} terms such that the electron correlation can be captured with significantly fewer terms. We have chosen to examine the behavior of the representation on hexane because its size is large enough so that the benefits of localization can be assessed, while it sets the stage for exploration of transferability towards other larger hydrocarbons (Section 6.4.2). To evaluate the behavior of orbital localization on the aforementioned feature space, the fraction of the full MP2 energy was plotted with respect to the number of promotions included in the representation of each pair energy, which is shown for hexane in Figure 6.1a. Canonical orbitals include the least amount of correlation within the representation at approximately 20% of the total MP2 correlation. Both Boys and PM orbitals significantly increased the total MP2 energy contained in the representation, capturing more than 70% of the total MP2 correlation energy using only 20 promotions for each ε_{ij}. PM orbitals outperform Boys with respect to capturing MP2 correlation in fewer promotions.

All data shown on Figure 6.1a were obtained with the 6-31G basis set, but as the basis set is expanded, the e_{ij}^{ab} matrix grows substantially, and the correlation contained within the representation is subsequently diminished. This is shown for hexane with PM orbitals for STO-3G, 6-31G, and cc-pVDZ, corresponding to 44, 82, and 154 basis functions, respectively. STO-3G quickly captures the complete MP2 correlation, due to the small number of virtual orbitals. Subsequently, 6-31G and cc-pVDZ representations contain substantially less of the overall MP2 correlation, as the increasing virtual orbital space creates a sparser e_{ij}^{ab} matrix. The same behavior can be seen for the canonical and Boys orbitals. Therefore, this suggests larger basis sets may be increasingly difficult as much of the correlation is not contained in the representation. However, as the basis set size increases, $\varepsilon_{ij}^{\mathrm{MP2}}$ is a better approximation to ε_{ij}^{CCSD}, which may offset this challenge. The results found in Figure 6.1 provide intuition

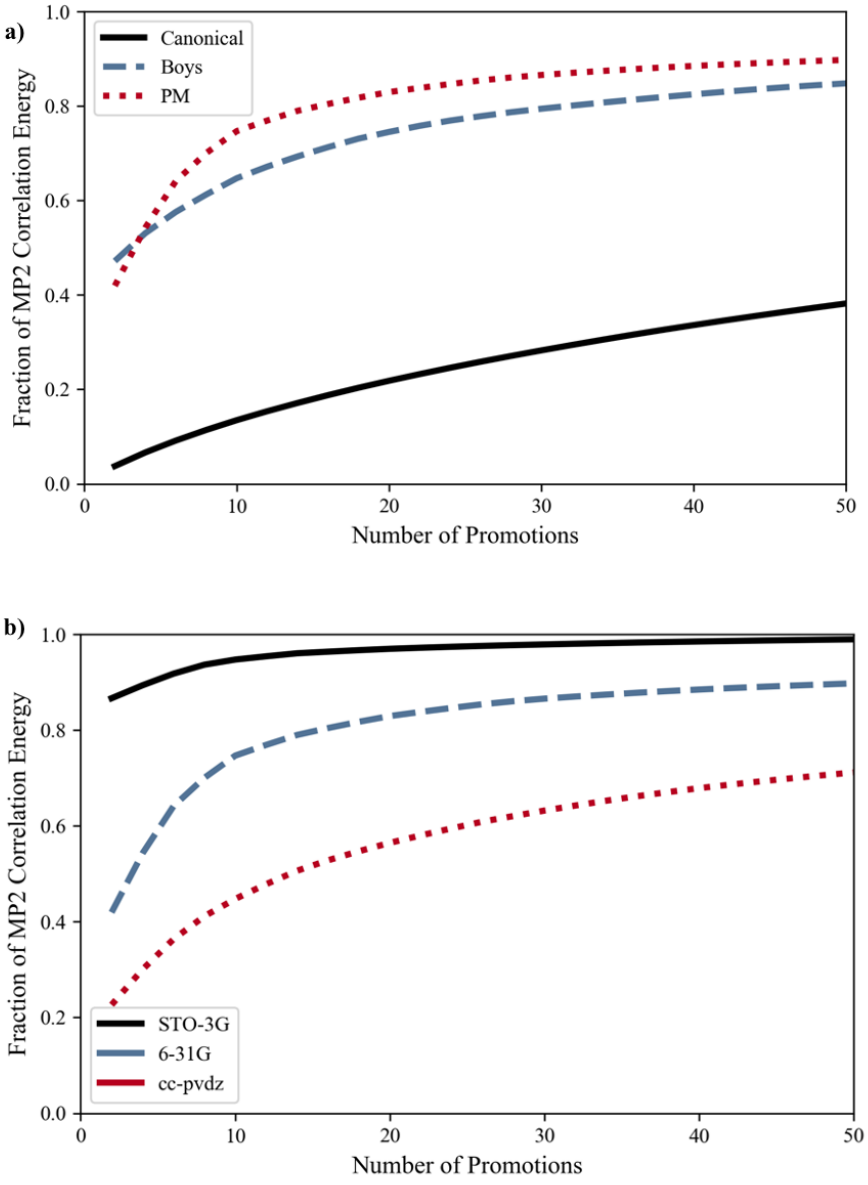


Figure 6.1: a) The fraction of the total MP2 correlation energy included in the representation of hexane versus the number of promotions from e_{ij}^{ab} (Eq. 6.3) in the 6-31G basis with canonical, Boys, and Pipek-Mezey orbitals. b) Fraction of MP2 correlation energy included in the representation of hexane with PM-localized orbitals versus the number of promotions with the STO-3G, 6-31G, and cc-pVDZ basis.

about the representation and should be considered as a guide for a reasonable selection for the number of promotions to be included. The accuracy of the method is not significantly affected if a reasonable number is selected, as it is demonstrated in the next sections. The performance impact of increasing the basis set size is further investigated in Section 6.4.6.

6.4.2 Hydrocarbon Series and Size-extensivity

An important consideration for a useful application of *ab initio*-based ML applications is the extrapolation of system size, i.e., where systems can be trained on smaller systems and have transferable accuracy to larger systems. Such extrapolations are essential due to the poor scaling with respect to system size of electronic structure theory methods, and therefore benefit from training on small systems to predict the properties of large systems and has shown promising results for the ML-based force fields and previous ML-based *ab initio* studies. In order for system size extrapolation to be effective, the representation must show systematic treatment for both small and large molecules. By considering that larger molecules have more pair-correlation terms, the method should be extensive in its treatment towards molecule size. Lack of size-extensivity has been a highlighted as a weakness for molecular representations that feature global feature sets for a molecule.[76,351] However, given that the representation is centered on the MP2 pair-correlation energies and sorted contributions, one must determine whether the representation will have a learnable behavior as the correlation about the molecule becomes more delocalized. Therefore, the properties and performance of the new representation with increasing system size was evaluated on saturated linear chain hydrocarbons.

The fraction of total MP2 correlation captured within the truncated promotion space for ethane to decane with 20 promotions that were used throughout this study is shown in Figure 6.2. With canonical orbitals, the amount of correlation contained within the representation decreases with system size. Ethane contains approximately 66% of the total MP2 correlation energy, whereas, decane contains only 13%. Using either Boys or PM localization schemes, the correlation contained within the representations remains relatively static. For both sets of localized orbitals, moving from pentane to decane results in approximately 1% reduction in the total MP2 correlation within the representation. PM orbitals showed the highest

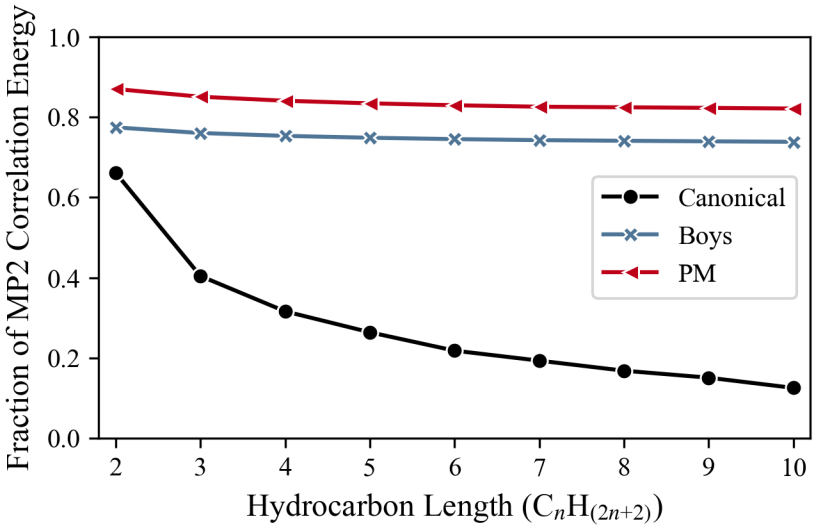


Figure 6.2: The fraction of MP2 correlation energy contained within the representation of hydrocarbons ethane through decane considering 20 pair promotions.

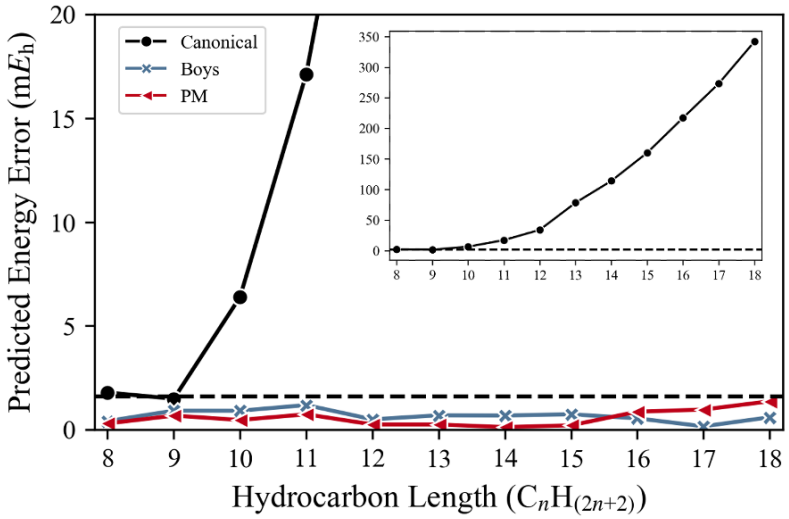


Figure 6.3: Predicted CCSD Energy Error for hydrocarbons octane through octadecane (n = 8-18) based off the model trained on single configurations of propane through heptane (n = 3-7). In the inset, the canonical energy error is shown in greater detail.

efficiency, with 82% of the correlation captured for decane at 20 promotions, followed by Boys with 74%, and lastly canonical with only 13%. Using a single conformation of propane through heptane hydrocarbons ($C_nH_{(2n+2)}$, n = 3-7), a model was made to predict the CCSD energies of the lengthier octane through octadecane (n = 8-18) for canonical, Boys localized, and PM localized orbitals, respectively (Figure 6.3). For the canonical orbitals, the error was near chemical accuracy relative to the reference for octane and nonane. However, for the larger molecules, the errors for the larger molecules increases significantly, as the larger molecules contain decreasing correlation in the representation, leading to substantial prediction errors. For both Boys and PM localized orbitals, the models achieved accuracy below the 1 kcal mol^{-1} threshold for the entire hydrocarbon series, with mean absolute errors of 0.67 and 0.57 mE_h, respectively. This result suggests that the PM orbital localization scheme is the best choice when the aim is molecular size extrapolation.

6.4.3 C_2 Dimer Dissociation

Another consideration is the behavior of the model when performing calculations on challenging electronic structures. As an example, we investigate the dissociation of the carbon dimer (C_2) which remains a challenge for theory as it contains many electronic near-degeneracies even near equilibrium geometry.[352] C_2 presents an unusual ground-state bonding configuration, contains two π bonds with little or no σ bonding.[353] While C_2 has been extensively studied using full CI and a plethora of multireference methods,[352–358] it has been selected herein because MP2 fails to describe its dissociation while CCSD performs reasonably well.[355] By selecting a case where MP2 is a poor description for CCSD, it evaluates whether the model is robust even when the initial representation is a poor approximation of CCSD. The C_2 dimer dissociation was modeled by uniform sampling 20 points from 0.8 to 3.7 Å for training and to reproduce the full dissociation which is shown in Figure 6.4. CCSD and MP2 are in relative agreement until approximately 2.5 Å, when MP2 begins to fail considerably and the correlation energy becomes unphysical. Despite this failure, the ML model is still able to qualitatively regenerate the CCSD dissociation. From the region of 3.0-3.7 Å, MP2 shows an absolute average deviation from CCSD of 108 mE_h, whereas the ML model was 6.7 mE_h. While the results are not quantitatively accurate, they demonstrate

that the model is able to qualitatively reproduce the CCSD result on a challenging electronic structure when the underlying MP2 representation is a poor description of CCSD.

6.4.4 Potential Energy Surface of Water 2510

As an additional example, we have examined the performance of the method in predicting energies on a complex dataset of interacting systems to evaluate a full potential energy surface (PES). In order to demonstrate this capability, the Water 2510 dataset[359–361] was utilized, which contains 2510 water dimers at a vast array of distances and conformations. Although the database focuses only on water dimers, it contains a diverse set of noncovalent interactions, including highly-repulsive short-range electrostatic interactions (up to +150 kcal mol^{-1} interaction energy).[362] For this application, 10% of the dimers were used for training and 90% for testing. The results for Boys and PM localized orbitals are found in Figure 6.5, which show a mean absolute error (MAE) of 9.7 x 10-5 E_h and 1.1 x 10^{-4} E_h, respectively, whereas canonical orbitals had a substantial MAE of 7.67 x 10^{-4} E_h. The mean absolute percentage error (MAPE) for Boys and PM is 0.0392% and 0.0457%, respectively. For the Boys-localized orbitals, of the 2259 molecular dimers in the testing set, only 7 had an error exceeding 1.0 kcal mol^{-1}. Figure 6.5 shows the systems predicted most poorly were those that at the edges of the distribution of correlation energy. This error can be ameliorated with a more sophisticated sampling strategy such as active learning.[209,363,364]

6.4.5 Examination of Transferability

With the previous models having demonstrated promising applicability with respect to accurate CCSD energies for PES mapping that is scalable to larger systems, we considered the transferability between different chemical systems. For this, we utilize the GDB-9 database[265] which contains approximately 134,000 small organic molecules optimized at the DFT level. The performance of CCSD energy predictions was evaluated using 0.01%, 0.1%, 1%, and 10% of the database respectively to test the remaining approximately 121,000 molecules (90%) and is shown for each orbital localization scheme in Figure 6.6.

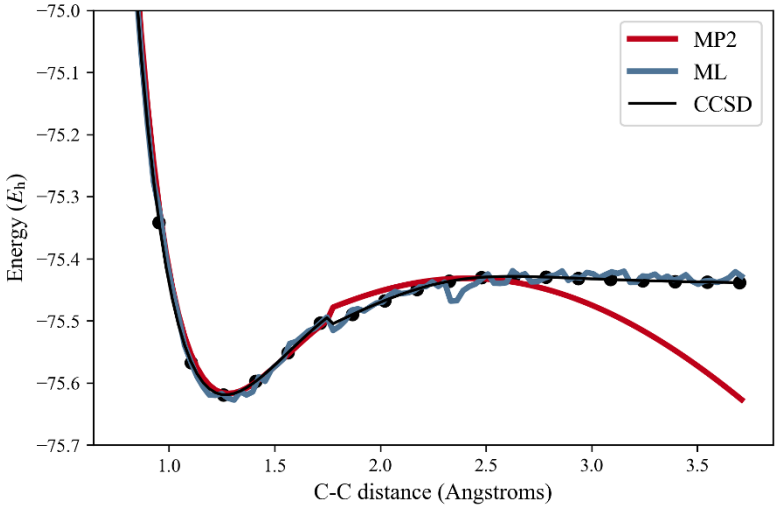


Figure 6.4: C_2 dimer dissociation using MP2 (red), CCSD (black), and a Boys-localized ML model (blue) utilizing the 6-31G basis set. The black points represent training points for the ML model.

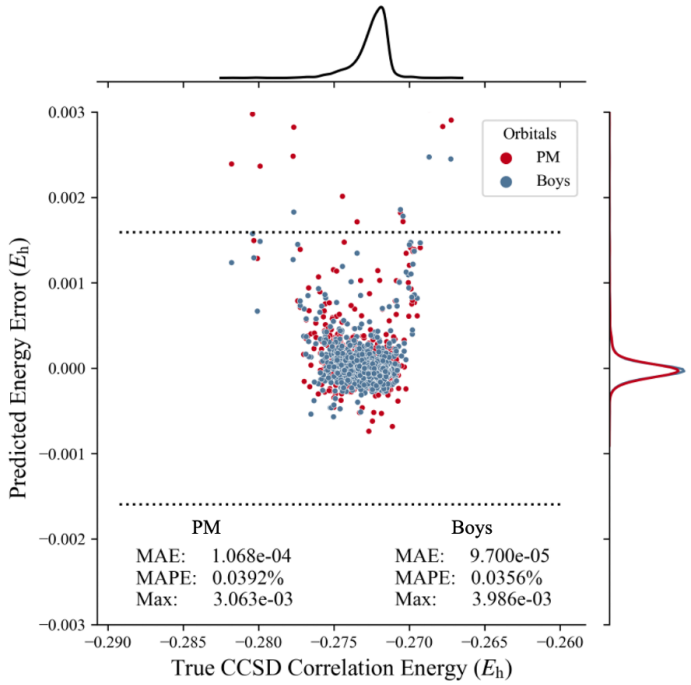


Figure 6.5: Predicted CCSD energy error versus the true correlation energy with their respective distributions for Boys and PM-localized orbitals. The dotted black lines represent the 1 kcal mol^{-1} limit.

In agreement with the previous applications, canonical orbital model accuracy is poorer than those trained with localized orbitals. In the GDB-9 screening, the Boys localization has lower error across all training set sizes. Both Boys and PM model mean absolute errors are below 1 kcal mol^{-1} using only 140 molecules for training. The Boys localization model had MAEs of 0.78 and 0.49 mE_{h} for 1% and 10% of the database as training, respectively, while the MAEs from PM are 0.86 mE_{h} and 0.53 mE_{h}, respectively. The predicted and actual correlation energies for the model using Boys-localized orbitals of 10% of the GDB-9 (i.e. 13,388 molecules used as input data) for training is shown in Figure 6.7. The correlation between predicted and exact data is perfectly linear with a slope of 1.0012 and a Pearson's correlation (R^2) of 0.9998.

A PM7-based geometry-featurized method to compute coupled-cluster energetics needed approximately 20k training instances from the GDB-9 dataset to reduce the MAE to 1 kcal mol^{-1}, which is over 100 times the number of training instances of our method to achieve such accuracy (0.1% or 140 molecules for our model). However, using DFT energies and a geometry-based representation, sub 1 kcal mol^{-1} error was achieved in just 500 training points. An electronic structure-theory based representation using HF features, named the MOB-ML model,[74] predicts coupled-cluster energies on a 350K thermally sampled QM7b dataset with mean absolute errors of approximately 1.25 kcal mol^{-1} with 140 training structures, which is impressive given the small training set size and small basis set (cc-pVDZ), as well as the addition of perturbative triples $E_{(T)}$ correction. However, extrapolating to the larger molecules of the GDB-13 dataset, the errors were increased to 3.88 mE_{h} (2.4 kcal mol^{-1}).

6.4.6 Basis Set Effects

Lastly, the model was examined to determine the performance deprecation in the expansion of the basis set. As shown in Figure 6.1b, the electron correlation becomes more delocalized over a larger number of orbital promotions as the virtual space grows, leading to a sparser representation. Therefore, less of the total MP2 energy is contained within the truncated promotion space which comprises our models. To examine the implications of expanding the basis set, we have generated models for predictions on a subset of the GDB-9 dataset. Each

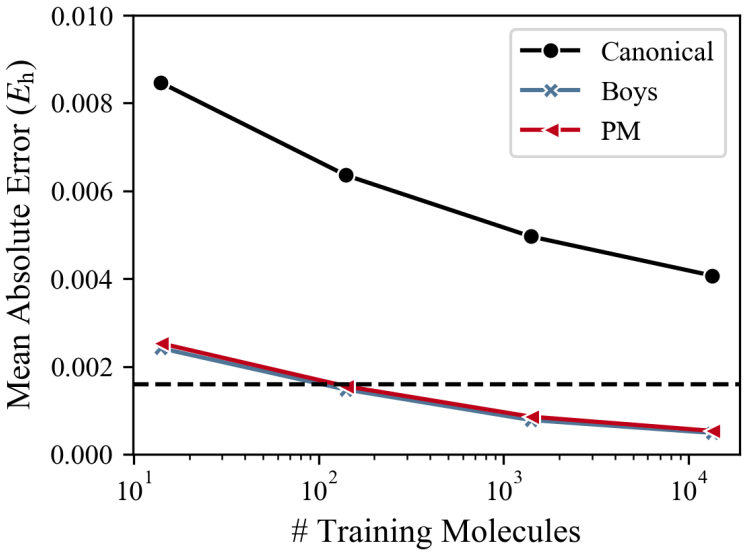


Figure 6.6: Learning curves for the models based on canonical, Boys localized, and Pipek-Mezey localized orbitals in the evaluation of CCSD energetics on the GDB-9 database. The horizontal black dashed line represents the 1 kcal mol^{-1} limit.

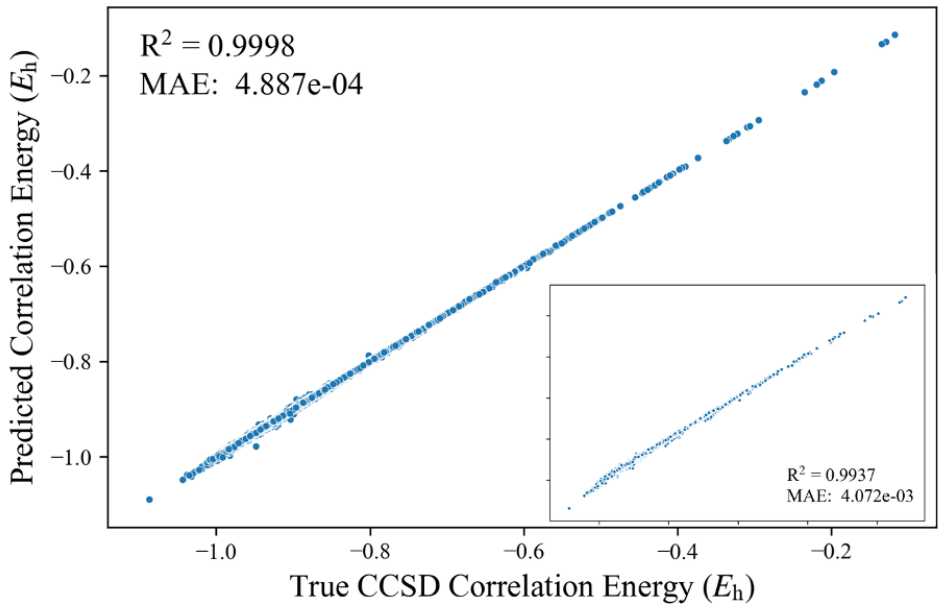


Figure 6.7: Predicted CCSD correlation energy versus the true CCSD correlation energy for 120,493 molecules in the GDB-9 database using the Boys-localized orbital model with the canonical model in the inset for comparison. The R^2 and MAE (in E_h) values are provided for both models.

model was trained on the same set of molecules that spanned 1% of the full dataset and used to predict the energies of 6,100 randomly selected molecules using models trained and tested on the STO-3G, 6-31G, and cc-pVDZ basis sets, respectively, and are shown in Figure 6.8. For each subset of orbital types, there is a systematic decrease in model performance upon expansion of the basis set. For the Boys-localized models in increasing basis set size from STO-3G to 6-31G the MAE goes from 0.69 to 0.77 mE_h. Further increasing the basis set size to the substantially larger cc-pVDZ increases the MAE to 1.09 mE_h, which is still below the 1 kcal mol^{-1} threshold. Similar relative increases in error are shown for the canonical and PM-localized orbitals. While the performance remains reasonable for these basis sets, it is clear further model development may be necessary to utilize routinely applied basis sets for coupled-cluster methods. These improvements may be found by further investigated the locality of the electron correlation or providing a more global representation of each orbital through the use of persistence images[209,250,256] and is a target for future studies.

6.5 Conclusions

This study introduced a new representation of the MP2 wave function for ML that allowed for accurate prediction of CCSD energies using localized orbitals. The generated representations require no additional calculations, and since they are based on the electronic structure, are invariant with respect to translation, permutation, and rotation. A re-engineered set of input features that provide an intuitive description of the orbital properties involved in electron correlation were introduced. The pair-energy representation contains a subset of the largest contributions to the MP2 energies and properties of those respective promotions. Combination of this approach together with orbital localization showed that the representation essentially converges its contained correlation with respect to size, and therefore was able to predict CCSD energies within chemical accuracy of larger linear chain hydrocarbons based on smaller ones. The new method also shows promising results on evaluating potential energy surfaces and was able to map a complex dataset containing many unique conformations of water dimers, yielding a MAE of 0.06 kcal mol^{-1}. The chemical transferability was also examined through evaluation of the GDB-9 database. The

localized models were able to accurately model electronic CCSD correlation energies with mean absolute errors below 1.0 kcal mol^{-1} using as few as 140 molecules for training. The Boys-localized model using approximately 10% of the GDB-9 database for training provided a mean absolute error of 0.31 kcal mol^{-1}. We believe this work is a step towards building a general-purpose and transferable model to accurately predict coupled-cluster energies across the periodic table. While methods that circumvent the MP2 step have a significantly reduced cost at the time of testing, ultimately, timings will also depend on the amount and size of molecules included for training. Finally, we envision that the new representation, together with recent advances in localized equation-of-motion coupled cluster,[365] can be used for accelerating the computation of excited states with machine learning.[322,366,367] Future works will expand the applicability with the inclusion of the perturbative triples term (T) and an unrestricted wave function implementation.

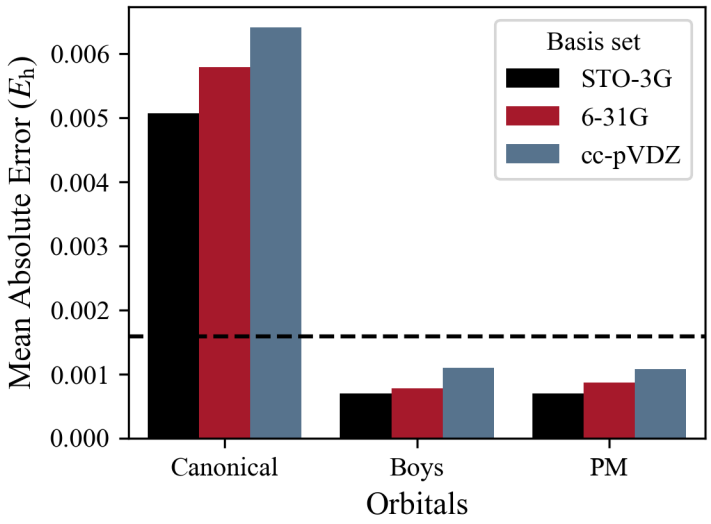


Figure 6.8: Mean absolute errors on a subset of the GDB-9 Dataset utilizing different basis sets sizes for canonical and the Boys- and PM-localized orbitals. The red dotted line represents an error of 1.0 kcal mol^{-1}.

Chapter 7

Summary and Outlook

In this work, we have performed in-depth calculations on the electronic structure of functional units of polymers. We have developed protocols to determine the global minima of complex potential energy surfaces between polymeric units and separation targets. From the results of our calculations, we were able to better interpret physical phenomena in polymeric materials in Chapter 2. Next, we performed an electronic-structure-theory driven screening of functional groups for the suggestion of boranes for the separation of N_2/O_2, where functionalized boranes were found to interact favorable with N_2, and thus suggest a promising pathway for binding nitrogen in materials.

Next, we utilized the conformational protocol previously mentioned to do a large-scale screening of organic molecules for CO_2 binding. In doing so, we examined the suitability of many molecular representations and developed a new representation based on persistent homology which encodes the topology of molecules with chemical information to represent molecules for machine learning. By combining the conformational search procedure with new representation, we performed active learning to selectively expand the training set to effectively capture rare events in the training set where CO_2 interaction energies exceeded -6.0 kcal mol^{-1}. With respect to other molecular representations, the chemically driven persistence images were able to accurately identify new functional groups for CO_2 binding with lower computational overhead and smaller training sets. The performance of this representation is promising and may perform well in other chemical applications.We believe our approach a productive first step towards suggesting functional groups for effective CO_2

separations. While there has been a recent surge in high-throughput screenings for metal-organic frameworks for gas separations,[368–371] similar advances for polymeric membranes have not been reported. There are also significant barriers towards practical realization of accurate suggestion of bulk materials for gas separation. Although it is possible to produce permeability values from theory, it requires both Monte Carlo and molecular dynamics methods which are complex and time-consuming, particularly for amorphous materials and rely on having accurate force fields for the material under consideration.[369] Therefore, we believe the development of representations such as PIs used synergetically with machine learning alorithms can connect atomistic effects to the bulk permeability properties, and offers a promising avenue for future suggestion of membrane technology for CO_2 separations.

While our DFT-based screenings offer productive insights, we ultimately would like to use more accurate wave function-based methods. These methods, although more accurate, have substantial costs that prohibit their use. To combat this cost, in Chapters 5 and 6, we have introduced methods that can accelerate the accurate prediction of coupled-cluster energy calculations based on MP2. As a proof of concept, we used ML to build CCSD wavefunctions by using MP2 wavefunctions and machine learning. The resulting method was shown to learn the electronic promotions of CCSD where the electronic structure could interpolate from training sets. A highlight of this study was the ability to accurately predict the energies of water hexamers based on water dimers and trimers. This promising method was also able to generate approximate CCSD wavefunctions that converge to the exact CCSD solution in fewer iterations than conventional intialization.

While Chapter 5 showed promising ability to interpolate within a chemical system, the transferability was largely limited to chemical species in the training set. In Chapter 6 we introduced a method that utilizes locality to exploit the local nature of electron correlation to produce orbitals that are transferable between chemical systems. A re-engineered set of input features that provide an intuitive description of the orbital properties involved in electron correlation were introduced. This method was able to predict energy of linear hydrocarbons (octane-octadecane) within chemical accuracy using a single configuration of smaller linear hydrocarbons (propane-heptane). The method also showed remarkable

chemical transferability with a training set of 140 molecules yielded mean absolute errors below 1 kcal mol^{-1} across a set of 120,000 unique organic molecules.

Locality in machine learning quantum-chemical methods offer an enormous opportunity for highly accurate calculations at unprecedented scales. The applications herein predicted CCSD energy, but as introduced in Chapter 1, our approach has a logical extension to higher levels of CC all the way to FCI. Additionally, these representations could perform well with pair-domain localized methods, which should further extend the transferabiity and applicabiity.[341–345] This representation could also be used to better extrapolate correlated methods to the complete basis set limit or determine missing correlation from approximate localized correlation methods, allowing one to get closer to exact solutions of a respective method. We believe that machine learning representations based on the electronic structure similar to those we have introduced offer an exciting new approach towards accurately predicting properties of systems at a substantially reduced computational expense. The synergistic cooperativity between lower cost, highly-accurate training data at lower computational expense and effective molecular representations for high-throughput screening offers unprecedented opportunity in calculating and predicting noncovalent interactions for gas separation applications.

Bibliography

[1] P. Hunter. The impact of CO_2. *EMBO reports*, 8(12):1104–1106, 2007.

[2] M. Hertzberg and H. Schreuder. Role of atmospheric carbon dioxide in climate change. *Energy and Environment*, 27(6-7):785–797, 2016.

[3] R. Cassia, M. Nocioni, N. Correa-Aragunde, and L. Lamattina. Climate change and the impact of greenhouse gasses: CO_2 and NO, friends and foes of plant oxidative stress. *Frontiers in Plant Science*, 9:1–11, 2018.

[4] I. Capellán-Pérez, I. Arto, J. M. Polanco-Martínez, M. González-Eguino, and M. B. Neumann. Likelihood of climate change pathways under uncertainty on fossil fuel resource availability. *Energy and Environmental Science*, 9(8):2482–2496, 2016.

[5] M. Hulme. 1.5 °C and climate research after the Paris Agreement. *Nature Climate Change*, 6(3):222–224, 2016.

[6] T. Kuramochi, N. Höhne, M. Schaeffer, J. Cantzler, B. Hare, Y. Deng, S. Sterl, M. Hagemann, M. Rocha, P. A. Yanguas-Parra, G. U. R. Mir, L. Wong, T. El-Laboudy, K. Wouters, D. Deryng, and K. Blok. Ten key short-term sectoral benchmarks to limit warming to 1.5° C. *Climate Policy*, 18(3):287–305, 2018.

[7] L. Abdallah and T. El-Shennawy. Reducing carbon dioxide emissions from electricity sector using smart electric grid applications. *Journal of Engineering (United Kingdom)*, 2013.

[8] Electricity in the United States. *US Energy Information and Administration*, 2020.

[9] M. Bui, C. S. Adjiman, A. Bardow, E. J. Anthony, A. Boston, S. Brown, P. S. Fennell, S. Fuss, A. Galindo, L. A. Hackett, J. P. Hallett, H. J. Herzog, G. Jackson, J. Kemper, S. Krevor, G. C. Maitland, M. Matuszewski, I. S. Metcalfe, C. Petit, G. Puxty, J. Reimer, D. M. Reiner, E. S. Rubin, S. A. Scott, N. Shah, B. Smit, J. P. Trusler, P. Webley, J. Wilcox, and N. Mac Dowell. Carbon capture and storage (CCS): The way forward. *Energy and Environmental Science*, 11(5):1062–1176, 2018.

[10] G. Xu, F. Liang, Y. Yang, Y. Hu, K. Zhang, and W. Liu. An improved CO_2 separation and purification system based on cryogenic separation and distillation theory. *Energies*, 7(5):3484–3502, 2014.

[11] C. Nwaoha, T. Supap, R. Idem, C. Saiwan, P. Tontiwachwuthikul, M. J. AL-Marri, and A. Benamor. Advancement and new perspectives of using formulated reactive amine blends for post-combustion carbon dioxide (CO_2) capture technologies. *Petroleum*, 3 (1):10–36, 2017.

[12] D. Cebrucean, V. Cebrucean, and I. Ionel. CO_2 capture and storage from fossil fuel power plants. *Energy Procedia*, 63:18–26, 2014.

[13] M. Chawla, H. Saulat, M. Masood Khan, M. Mahmood Khan, S. Rafiq, L. Cheng, T. Iqbal, M. I. Rasheed, M. Z. Farooq, M. Saeed, N. M. Ahmad, M. B. Khan Niazi, S. Saqib, F. Jamil, A. Mukhtar, and N. Muhammad. Membranes for CO_2 /CH_4 and CO_2/N_2 Gas Separation. *Chemical Engineering and Technology*, 43(2):184–199, 2020.

[14] C. R. Maroon, J. Townsend, K. R. Gmernicki, D. J. Harrigan, B. J. Sundell, J. A. Lawrence, S. M. Mahurin, K. D. Vogiatzis, and B. K. Long. Elimination of CO_2 /N_2 Langmuir Sorption and Promotion of "N_2-Phobicity" within High- T_g Glassy Membranes. *Macromolecules*, 52(4):1589–1600, 2019.

[15] L. M. Robeson. The upper bound revisited. *Journal of Membrane Science*, 320(1): 390–400, 2008.

[16] K. D. Vogiatzis, A. Mavrandonakis, W. Klopper, and G. E. Froudakis. Ab initio Study of the Interactions between CO_2 and N-Containing Organic Heterocycles. *ChemPhysChem*, 10(2):374–383, 2009.

[17] K. D. Vogiatzis, W. Klopper, and J. Friedrich. Non-covalent interactions of CO_2 with functional groups of metal-organic frameworks from a CCSD(T) scheme applicable to large systems. *Journal of Chemical Theory and Computation*, 11(4):1574–1584, 2015.

[18] Y. Matsuzaki, H. Yamada, F. A. Chowdhury, T. Higashii, S. Kazama, and M. Onoda. Ab initio study of CO_2 capture mechanisms in monoethanolamine aqueous solution: Reaction pathways from carbamate to bicarbonate. *Energy Procedia*, 37:400–406, 2013.

[19] B. Arstad, R. Blom, and O. Swang. CO_2 absorption in aqueous solutions of alkanolamines: Mechanistic insight from quantum chemical calculations. *Journal of Physical Chemistry A*, 111(7):1222–1228, 2007.

[20] E. F. Da Silva and H. F. Svendsen. Ab initio study of the reaction of carbamate formation from CO_2 and alkanolamines. *Industrial and Engineering Chemistry Research*, 43(13):3413–3418, 2004.

[21] H. B. Xie, N. He, Z. Song, J. Chen, and X. Li. Theoretical investigation on the different reaction mechanisms of aqueous 2-amino-2-methyl-1-propanol and monoethanolamine with CO_2. *Industrial and Engineering Chemistry Research*, 53(8):3363–3372, 2014.

[22] Z. Tian, S. Dai, and D.-e. Jiang. What can molecular simulation do for global warming? *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 6(2): 173–197, 2016.

[23] S. Chu and A. Majumdar. Opportunities and challenges for a sustainable energy future. *Nature*, 488(7411):294–303, 2012.

[24] Z. Tian, T. Saito, and D.-e. Jiang. Ab Initio Screening of CO_2-philic Groups. *The Journal of Physical Chemistry A*, 119(16):3848–3852, apr 2015.

[25] S. Kazemi and V. Safarifard. Carbon dioxide capture in MOFs: The effect of ligand functionalization. *Polyhedron*, 154(2018):236–251, 2018.

[26] G. Chalasinski and M. M. Szczesniak. Origins of Structure and Energetics of van der Waals Clusters from ab Initio Calculations. *Chemical Reviews*, 94(7):1723–1765, 1994.

[27] J. Townsend, N. M. Braunscheidel, and K. D. Vogiatzis. Understanding the Nature of Weak Interactions Between Functionalized Boranes and N_2 / O_2 , Promising Functional

Groups for Gas Separations. *The Journal of Physical Chemistry A*, 123:3315–3325, 2019.

[28] R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of Modern Physics*, 87(3):897–923, 2015.

[29] N. Mardirossian and M. Head-Gordon. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017.

[30] Y. Zhao and D. G. Truhlar. Comparative DFT study of van der Waals complexes: Rare-gas dimers, alkaline-earth dimers, zinc dimer and zinc-rare-gas dimers. *Journal of Physical Chemistry A*, 110(15):5121–5129, 2006.

[31] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, apr 2010.

[32] G. A. Dilabio and A. Otero-de-la Roza. Noncovalent Interactions in Density Functional Theory. *Reviews in Computational Chemistry*, 29:1–97, 2016.

[33] A. Kovács, J. Cz. Dobrowolski, S. Ostrowski, and J. E. Rode. Benchmarking density functionals in conjunction with Grimme's dispersion correction for noble gas dimers (Ne_2, Ar_2, Kr_2, Xe_2, Rn_2). *International Journal of Quantum Chemistry*, 117(9), 2017.

[34] A. Fraccarollo, L. Canti, L. Marchese, and M. Cossi. Accurate Evaluation of the Dispersion Energy in the Simulation of Gas Adsorption into Porous Zeolites. *Journal of Chemical Theory and Computation*, 13(4):1756–1768, 2017.

[35] S. Mazinani, R. Ramezani, S. Darvishmanesh, G. F. Molelekwa, R. Di Felice, and B. Van Der Bruggen. A ground breaking polymer blend for CO_2/N_2 separation. *Journal of CO_2 Utilization*, 27(March):536–546, 2018.

[36] K. R. Gmernicki, E. Hong, C. R. Maroon, S. M. Mahurin, A. P. Sokolov, T. Saito, and B. K. Long. Accessing Siloxane Functionalized Polynorbornenes via Vinyl-Addition

Polymerization for CO_2 Separation Membranes. *ACS Macro Letters*, 5(7):879–883, jul 2016.

[37] P. Bernardo, E. Drioli, and G. Golemme. Membrane Gas Separation: A Review/State of the Art. *Ind. Eng. Chem. Res*, 48(10):4638–4663, 2009.

[38] L. M. Robeson, C. D. Smith, and M. Langsam. A group contribution approach to predict permeability and permselectivity of aromatic polymers. *Journal of Membrane Science*, 132(1):33–54, 1997.

[39] D. V. Laciak, L. M. Robeson, and C. D. Smith. *Group Contribution Modeling of Gas Transport in Polymeric Membranes*. 1999.

[40] D. W. van Krevelen. *Properties of Polymers: Their Correlation with Chemical Structure, Their Numerical Estimation and Prediction from Additive Group Contributions*. Elsevier, 1997. ISBN 9780444596123.

[41] Y. Yampolskii, S. Shishatskii, A. Alentiev, and K. Loza. Group contribution method for transport property predictions of glassy polymers: focus on polyimides and polynorbornenes. *Journal of Membrane Science*, 149(2):203–220, 1998.

[42] V. Ryzhikh, D. Tsarev, A. Alentiev, and Y. Yampolskii. A novel method for predictions of the gas permeation parameters of polymers on the basis of their chemical structure. *Journal of Membrane Science*, 487:189–198, 2015.

[43] J. L. Reymond. The Chemical Space Project. *Accounts of Chemical Research*, 48(3): 722–730, 2015.

[44] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, and S. K. Saikin. Autonomous Molecular Design: Then and Now. *ACS Applied Materials and Interfaces*, 11(28):24825–24836, 2019.

[45] H. J. Kulik. Making machine learning a useful tool in the accelerated discovery of transition metal complexes. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(1):1–14, 2020.

[46] G. Schneider. Automating Drug Discovery. *Nature*, 17:97–113, 2018.

[47] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.

[48] B. Sanchez-Lengeling and A. Aspuru-Guzik. Inverse molecular design using machine learning:Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[49] 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*, 4(2):268–276, 2018.

[50] W. Wu, Z. Tian, S. Wang, C. Peng, H. Liu, S. Dai, and D. E. Jiang. Design of Calix-Based Cages for CO_2 Capture. *Industrial and Engineering Chemistry Research*, 56 (15):4502–4507, 2017.

[51] J. H. Hymel, J. Townsend, and K. D. Vogiatzis. CO_2 Capture on Functionalized Calixarenes: A Computational Study. *Journal of Physical Chemistry A*, 123(46): 10116–10122, 2019.

[52] D. Schwalbe-Koda and R. Gómez-Bombarelli. Generative Models for Automatic Chemical Design. *Lecture Notes in Physics*, 968:445–467, 2020.

[53] W. Yang, T. T. Fidelis, and W. H. Sun. Machine Learning in Catalysis, from Proposal to Practicing. *ACS Omega*, 5(1):83–88, 2020.

[54] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan, and R. B. Sunoj. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proceedings of the National Academy of Sciences of the United States of America*, 117(3):1339–1345, 2020.

[55] B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, and C. Sutton. Machine learning for heterogeneous catalyst design and discovery. *AIChE Journal*, 64(7):2311–2323, 2018.

[56] C. Réda, E. Kaufmann, and A. Delahaye-Duriez. Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, 18:241–252, 2020.

[57] 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*, 9(9):8243–8255, 2019.

[58] A. Nandy, C. Duan, J. P. Janet, S. Gugler, and H. J. Kulik. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Industrial and Engineering Chemistry Research*, 57(42):13973–13986, 2018.

[59] P. O. Dral. Quantum Chemistry in the Age of Machine Learning. *Journal of Physical Chemistry Letters*, 11(6):2336–2347, 2020.

[60] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Müller. Bypassing the Kohn-Sham equations with machine learning. *Nature Communications*, 8(1):872, 2017.

[61] K. Yao and J. Parkhill. Kinetic Energy of Hydrocarbons as a Function of Electron Density and Convolutional Neural Networks. *Journal of Chemical Theory and Computation*, 12(3):1139–1147, 2016.

[62] Y. Zhou, J. Wu, S. Chen, and G. H. Chen. Toward the Exact Exchange-Correlation Potential: A Three-Dimensional Convolutional Neural Network Construct. *Journal of Physical Chemistry Letters*, 10(22):7264–7269, 2019.

[63] P. D. Mezei and O. A. Von Lilienfeld. Noncovalent Quantum Machine Learning Corrections to Density Functionals. *Journal of Chemical Theory and Computation*, 16(4):2647–2653, 2020.

[64] T. Gao, H. Li, W. Li, L. Li, C. Fang, H. Li, L. Hu, Y. Lu, and Z. M. Su. A machine learning correction for DFT non-covalent interactions based on the S22, S66 and X40 benchmark databases. *Journal of Cheminformatics*, 8(24):1–17, 2016.

[65] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf. Electron density learning of non-covalent systems. *Chemical Science*, 10(41):9424–9432, 2019.

[66] J. Proppe, S. Gugler, and M. Reiher. Gaussian Process-Based Refinement of Dispersion Corrections. *Journal of Chemical Theory and Computation*, 15(11):6046–6060, 2019.

[67] T. Weymuth, J. Proppe, and M. Reiher. Statistical Analysis of Semiclassical Dispersion Corrections. *Journal of Chemical Theory and Computation*, 14(5):2480–2494, 2018.

[68] R. J. Bartlett, C. E. Dykstra, and J. Paldus. *Coupled Cluster Methods for Molecular Calculations*. Springer, 1984. ISBN 9788578110796.

[69] J. P. Coe. Machine Learning Configuration Interaction. *Journal of Chemical Theory and Computation*, 14(11):5739–5749, 2018.

[70] J. P. Coe. Machine Learning Configuration Interaction for ab Initio Potential Energy Curves. *Journal of Chemical Theory and Computation*, 15(11):6179–6189, 2019.

[71] J. Hermann, Z. Schätzle, and F. Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry*, 2020.

[72] K. Choo, A. Mezzacapo, and G. Carleo. Fermionic neural-network states for ab-initio electronic structure. *Nature Communications*, 11(1):1–7, 2020.

[73] M. Welborn, L. Cheng, and T. F. Miller. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *Journal of Chemical Theory and Computation*, 14(9):4772–4779, 2018.

[74] L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *Journal of Chemical Physics*, 150(13):131103, 2019.

[75] L. Cheng, N. B. Kovachki, M. Welborn, and T. F. Miller. Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning. *Journal of Chemical Theory and Computation*, 15:6668–6677, 2019.

[76] B. G. Peyton, C. Briggs, R. D'Cunha, J. T. Margraf, and T. D. Crawford. Machine-Learning Coupled Cluster Properties through a Density Tensor Representation. *Journal of Physical Chemistry A*, 124(23):4861–4871, 2020.

[77] Y. Chen, L. Zhang, H. Wang, and W. E. Ground State Energy Functional with Hartree–Fock Efficiency and Chemical Accuracy. *The Journal of Physical Chemistry A*, 2020.

[78] J. T. Margraf and K. Reuter. Making the Coupled Cluster Correlation Energy Machine-Learnable. *Journal of Physical Chemistry A*, 122:6343–6348, 2018.

[79] A. F. Silva, M. A. Vincent, J. L. McDonagh, and P. L. Popelier. The Transferability of Topologically Partitioned Electron Correlation Energies in Water Clusters. *ChemPhysChem*, 18(23):3360–3368, 2017.

[80] S. Saebo and P. Pulay. Local treatment of electron correlation. *Annu. Rev. Phys. Chem*, 44:213–236, 1993.

[81] N. Flocke and R. J. Bartlett. Localized correlation treatment using natural bond orbitals. *Chemical Physics Letters*, 367(1-2):80–89, 2003.

[82] J. E. Carpenter and F. Weinhold. Transferability of Natural Bond Orbitals. *Journal of the American Chemical Society*, 110(2):368–372, 1988.

[83] P. Luis. Use of monoethanolamine (MEA) for CO_2 capture in a global scenario: Consequences and alternatives. *Desalination*, 380:93–99, 2016.

[84] Y. Yampolskii, I. Pinnau, and B. Freeman. *Materials Science of Membranes for Gas and Vapor Separation*, volume 598. John Wiley & Sons, Ltd, 2006. ISBN 9780470853450.

[85] E. Sh Finkelshtein, K. L. Makovetskii, M. L. Gringolts, Y. V. Rogan, T. G. Golenko, L. E. Starannikova, Y. P. Yampolskii, V. P. Shantarovich, and T. Suzuki. Addition-Type Polynorbornenes with $Si(CH_3)_3$ Side Groups: Synthesis, Gas Permeability, and Free Volume. *Macromolecules*, 39:7022–7029, 2006.

[86] N. Belov, R. Nikiforov, L. Starannikova, K. R. Gmernicki, C. R. Maroon, B. K. Long, V. Shantarovich, and Y. Yampolskii. A detailed investigation into the gas permeation properties of addition-type poly(5-triethoxysilyl-2-norbornene). *European Polymer Journal*, 93:602–611, 2017.

[87] B. J. Sundell, J. A. Lawrence, D. J. Harrigan, J. T. Vaughn, T. S. Pilyugina, and D. R. Smith. Alkoxysilyl functionalized polynorbornenes with enhanced selectivity for heavy hydrocarbon separations. *RSC Advances*, 6(57):51619–51628, 2016.

[88] H. Lin and B. D. Freeman. Gas solubility, diffusivity and permeability in poly(ethylene oxide). *Journal of Membrane Science*, 239(1):105–117, 2004.

[89] T. A. Barbari, W. J. Koros, and D. R. Paul. Gas sorption in polymers based on bisphenol-A. *Journal of Polymer Science Part B: Polymer Physics*, 26(4):729–744, 1988.

[90] S. Kilic, Y. Wang, J. K. Johnson, E. J. Beckman, and R. M. Enick. Influence of tert-amine groups on the solubility of polymers in CO_2. *Polymer*, 50(11):2436–2444, 2009.

[91] H. Feng, T. Hong, S. M. Mahurin, K. D. Vogiatzis, K. R. Gmernicki, B. K. Long, J. W. Mays, A. P. Sokolov, N. G. Kang, and T. Saito. Gas separation mechanism of CO_2 selective amidoxime-poly(1-trimethylsilyl-1-propyne) membranes. *Polymer Chemistry*, 8(21):3341–3350, 2017.

[92] S. Zulfiqar, F. Karadas, J. Park, E. Deniz, G. D. Stucky, Y. Jung, M. Atilhan, and C. T. Yavuz. Amidoximes: Promising candidates for CO_2 capture. *Energy and Environmental Science*, 4(11):4528–4531, 2011.

[93] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[94] L. S. Dodda, I. C. De Vaca, J. Tirado-Rives, and W. L. Jorgensen. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Research*, 45:331–336, 2017.

[95] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[96] F. Furche, R. Ahlrichs, C. Hättig, W. Klopper, M. Sierka, and F. Weigend. Turbomole. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(2):91–100, 2014.

[97] R. Alrichs, M. Baer, M. Haeser, H. Horn, and C. Koelmel. Electronic structure calculations on workstation computers: the program system TURBOMOLE. *Chem Phys Lett.*, 162(3):165–169, 1989.

[98] C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics*, 110(13): 6158–6170, 1999.

[99] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

[100] S. Grimme, R. Huenerbein, and S. Ehrlich. On the Importance of the Dispersion Energy for the Thermodynamic Stability of Molecules. *ChemPhysChem*, 12(7):1258–1261, 2011.

[101] C. R. Maroon, J. Townsend, M. A. Higgins, D. J. Harrigan, B. J. Sundell, J. A. Lawrence, J. T. O'Brien, D. O'Neal, K. D. Vogiatzis, and B. K. Long. Addition-type alkoxysilyl-substituted polynorbornenes for post-combustion carbon dioxide separations. *Journal of Membrane Science*, page 117532, 2019.

[102] M. A. Higgins, C. R. Maroon, J. Townsend, X. Wang, K. D. Vogiatzis, and B. K. Long. Evaluating the impact of functional groups on membrane-mediated CO_2/N_2 gas separations using a common polymer backbone . *Journal of Polymer Science*, 58 (18):2644–2653, 2020.

[103] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno, and C. D. Sherrill. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *Journal of Chemical Physics*, 140(9), 2014.

[104] E. Papajak, J. Zheng, X. Xu, H. R. Leverentz, and D. G. Truhlar. Perspectives on basis sets beautiful: Seasonal plantings of diffuse basis functions. *Journal of Chemical Theory and Computation*, 7(10):3027–3034, 2011.

[105] R. M. Parrish, L. A. Burns, D. G. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *Journal of Chemical Theory and Computation*, 13(7):3185–3197, 2017.

[106] E. G. Hohenstein and C. D. Sherrill. Density fitting and Cholesky decomposition approximations in symmetry-adapted perturbation theory: Implementation and application to probe the nature of π - π. *Journal of Chemical Physics*, 132(18):184111, 2010.

[107] P. F. Cao, B. Li, T. Hong, J. Townsend, Z. Qiang, K. Xing, K. D. Vogiatzis, Y. Wang, J. W. Mays, A. P. Sokolov, and T. Saito. Superstretchable, Self-Healing Polymeric Elastomers with Tunable Properties. *Advanced Functional Materials*, 28(22):1800741, 2018.

[108] S. R. White, N. R. Sottos, P. H. Geubelle, J. S. Moore, M. R. Kessler, S. R. Sriram, E. N. Brown, and S. Viswanathan. Autonomic healing of polymer composites. *Nature*, 409(6822):794–797, 2001.

[109] S. Wang and M. W. Urban. Self-healing polymers. *Nature Reviews Materials*, 5(8): 562–583, 2020.

[110] L. Brunsveld, B. J. Folmer, E. W. Meijer, and R. P. Sijbesma. Supramolecular polymers. *Chemical Reviews*, 101(12):4071–4097, 2001.

[111] M. C. Fyfe and J. F. Stoddart. Synthetic Supramolecular Chemistry. *Accounts of Chemical Research*, 30(10):393–401, 1997.

[112] B. P. Jacobs, J. Townsend, K. D. Vogiatzis, and J. N. Brantley. Ion specific fluorescence modulation of polyvinyl alcohol-boronate matrices. *Polymer Chemistry*, 11(11):1919–1925, 2020.

[113] O. Lantz, M. H. Jouvin, M. C. De Vernejoul, and P. Druet. Fluoride-Induced Chronic Renal Failure. *American Journal of Kidney Diseases*, 10(2):136–139, 1987.

[114] K. Heard, R. E. Hill, C. B. Cairns, and R. C. Dart. Calcium neutralizes fluoride bioavailability in a lethal model of fluoride poisoning. *Journal of Toxicology - Clinical Toxicology*, 39(4):349–353, 2001.

[115] L. E. Figueroa, M. E. Moragues, E. Climent, A. Agostini, R. Martínez-Máñez, and F. Sancenón. Chromogenic and fluorogenic chemosensors and reagents for anions. A comprehensive review of the years 2010-2011. *Chemical Society Reviews*, 42(8):3489–3613, 2013.

[116] N. Busschaert, C. Caltagirone, W. Van Rossom, and P. A. Gale. Applications of Supramolecular Anion Recognition. *Chemical Reviews*, 115(15):8038–8155, 2015.

[117] T. D. Ashton, K. A. Jolliffe, and F. M. Pfeffer. Luminescent probes for the bioimaging of small anionic species in vitro and in vivo. *Chemical Society Reviews*, 44(14):4547–4595, 2015.

[118] V. Blažek Bregović, N. Basarić, and K. Mlinarić-Majerski. Anion binding with urea and thiourea derivatives. *Coordination Chemistry Reviews*, 295:80–124, 2015.

[119] F. Neese. Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):4–9, 2018.

[120] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993.

[121] D. Rappoport and F. Furche. Property-optimized Gaussian basis sets for molecular response calculations. *Journal of Chemical Physics*, 133(13):134105, 2010.

[122] F. Weigend. Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics*, 8(9):1057–1065, 2006.

[123] F. Neese, F. Wennmohs, A. Hansen, and U. Becker. Efficient, approximate and parallel Hartree-Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree-Fock exchange. *Chemical Physics*, 356(1-3):98–109, 2009.

[124] A. K. Dutta, F. Neese, and R. Izsák. Speeding up equation of motion coupled cluster theory with the chain of spheres approximation. *Journal of Chemical Physics*, 144(3): 034102, 2016.

[125] T. Yanai, D. P. Tew, and N. C. Handy. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters*, 393 (1-3):51–57, 2004.

[126] I. Borges, E. Uhl, L. Modesto-Costa, A. J. Aquino, and H. Lischka. Insight into the Excited State Electronic and Structural Properties of the Organic Photovoltaic Donor Polymer Poly(thieno[3,4-b]thiophene benzodithiophene) by Means of ab Initio and Density Functional Theory. *Journal of Physical Chemistry C*, 120(38):21818–21826, 2016.

[127] M. Kolaski, C. R. Arunkumar, and K. S. Kim. Aromatic Excimers: Ab Initio and TD-DFT Study. *Journal of Chemical Theory and Computation*, 9(1):847–856, 2012.

[128] T. Burdyny and H. Struchtrup. Hybrid membrane/cryogenic separation of oxygen from air for use in the oxy-fuel process. *Energy*, 35(5):1884–1897, 2010.

[129] D. Gielen. CO_2 Removal in the Iron and Steel Industry. *Energy Conversion and Management*, 44(7):1027–1037, 2003.

[130] J. Baukal, editor. *Oxygen-Enhanced Combustion*. CRC Press, New York, 1st ed. edition, 1998.

[131] A. Kather and G. Scheffknecht. The oxycoal process with cryogenic oxygen supply. *Naturwissenschaften*, 96(9):993–1010, 2009.

[132] B. J. Buhre, L. K. Elliott, C. D. Sheng, R. P. Gupta, and T. F. Wall. Oxy-fuel combustion technology for coal-fired power generation. *Progress in Energy and Combustion Science*, 31(4):283–307, 2005.

[133] G. Scheffknecht, L. Al-Makhadmeh, U. Schnell, and J. Maier. Oxy-fuel coal combustion-A review of the current state-of-the-art. *International Journal of Greenhouse Gas Control*, 5(S1):16–35, 2011.

[134] C. R. Shaddix and A. Molina. Fundamental investigation of NO_x formation during oxy-fuel combustion of pulverized coal. *Proceedings of the Combustion Institute*, 33(2): 1723–1730, 2011.

[135] K. Sumida, D. L. Rogow, J. A. Mason, T. M. Mcdonald, E. D. Bloch, Z. R. Herm, T.-H. Bae, and J. R. Long. Carbon Dioxide Capture in Metal Organic Frameworks. *Chemical Reviews*, 112(2):724–781, 2011.

[136] T. Banaszkiewicz, M. Chorowski, and W. Gizicki. Comparative analysis of cryogenic and PTSA technologies for systems of oxygen production. *AIP Conference Proceedings*, 1573(2014):1373–1378, 2014.

[137] Y. J. Colón and R. Q. Snurr. High-throughput computational screening of metal organic frameworks. *Chem. Soc. Rev.*, 43(16):5735–5749, apr 2014.

[138] P. Verma, R. Maurice, and D. G. Truhlar. Identifying the Interactions That Allow Separation of O_2 from N_2 on the Open Iron Sites of Fe_2(dobdc). *Journal of Physical Chemistry C*, 119(51):28499–28511, 2015.

[139] K. Lee, W. C. Isley, A. L. Dzubak, P. Verma, S. J. Stoneburner, L. C. Lin, J. D. Howe, E. D. Bloch, D. A. Reed, M. R. Hudson, C. M. Brown, J. R. Long, J. B. Neaton, B. Smit, C. J. Cramer, D. G. Truhlar, and L. Gagliardi. Design of a metal-organic framework with enhanced back bonding for separation of N_2 and CH_4. *Journal of the American Chemical Society*, 136(2):698–704, 2014.

[140] D. J. Xiao, M. I. Gonzalez, L. E. Darago, K. D. Vogiatzis, E. Haldoupis, L. Gagliardi, and J. R. Long. Selective, Tunable O_2 Binding in Cobalt(II)-Triazolate/Pyrazolate Metal-Organic Frameworks. *Journal of the American Chemical Society*, 138(22):7161–7170, 2016.

[141] G. R. Lorzing, B. A. Trump, C. M. Brown, and E. D. Bloch. Selective Gas Adsorption in Highly Porous Chromium(II)-Based Metal-Organic Polyhedra. *Chemistry of Materials*, 29(20):8583–8587, 2017.

[142] L. J. Murray, M. Dinca, J. Yano, S. Chavan, S. Bordiga, C. M. Brown, and J. R. Long. Highly-selective and reversible O_2 binding in Cr_3(1,3,5-benzenetricarboxylate)$_2$. *Journal of the American Chemical Society*, 132(23):7856–7857, 2010.

[143] M. V. Parkes, D. F. Sava Gallis, J. A. Greathouse, and T. M. Nenoff. Effect of metal in M_3(btc)$_2$ and M_2(dobdc) MOFs for O_2/N_2 separations: A combined density functional theory and experimental study. *Journal of Physical Chemistry C*, 119(12):6556–6567, 2015.

[144] M. V. Parkes, J. A. Greathouse, D. B. Hart, D. F. Sava Gallis, and T. M. Nenoff. Ab initio molecular dynamics determination of competitive O_2 vs. N_2 adsorption at open metal sites of M_2(dobdc). *Phys. Chem. Chem. Phys.*, 18(16):11528–11538, 2016.

[145] Y. Wang, J. Yang, Z. Li, Z. Zhang, J. Li, Q. Yang, and C. Zhong. Computational study of oxygen adsorption in metal–organic frameworks with exposed cation sites: effect of framework metal ions. *RSC Advances*, 5(42):33432–33437, 2015.

[146] A. Demessence, D. M. D'alessandro, L. Foo, and J. R. Long. Strong CO_2 Binding in a Water-Stable, Triazolate-Bridged Metal-Organic Framework Functionalized with Ethylenediamine. *J. Am. Chem. Soc*, 131(25):8784–8786, 2009.

[147] S. S. Chen, M. Chen, S. Takamizawa, P. Wang, G. C. Lv, and W. Y. Sun. Porous cobalt(ii)-imidazolate supramolecular isomeric frameworks with selective gas sorption property. *Chemical Communications*, 47(17):4902–4904, 2011.

[148] J. Lincke, D. Lässig, J. Moellmer, C. Reichenbach, A. Puls, A. Moeller, R. Gläser, G. Kalies, R. Staudt, and H. Krautscheid. A novel copper-based MOF material: Synthesis, characterization and adsorption studies. *Microporous and Mesoporous Materials*, 142(1):62–69, 2011.

[149] H. J. Park and M. P. Suh. Stepwise and hysteretic sorption of N_2, O_2, CO_2, and H_2 gases in a porous metal-organic framework $[Zn_2(BPnDC)_2(bpy)]$. *Chemical Communications*, 46(4):610–612, 2010.

[150] T. K. Kim and M. P. Suh. Selective CO_2 adsorption in a flexible non-interpenetrated metal-organic framework. *Chemical Communications*, 47(14):4258–4260, 2011.

[151] P. Pachfule, R. Das, P. Poddar, and R. Banerjee. Solvothermal synthesis, structure, and properties of metal organic framework isomers derived from a partially fluorinated link. *Crystal Growth and Design*, 11(4):1215–1222, 2011.

[152] C. A. Fernandez, P. K. Thallapally, R. K. Motkuri, S. K. Nune, J. C. Sumrak, J. Tian, and J. Liu. Gas-induced expansion and contraction of a fluorinated metal-organic framework. *Crystal Growth and Design*, 10(3):1037–1039, 2010.

[153] S. Wu, L. Ma, L.-s. Long, L.-s. Zheng, and W. Lin. 3D Metal-Organic Frameworks Based on Functionalized Tetracarboxylate Linkers : Synthesis , Structures , and Gas Sorption Studies. *Inorganic Chemistry*, 48(6):2436–2442, 2009.

[154] T. Gadzikwa, O. K. Farha, K. L. Mulfort, J. T. Hupp, and S. T. Nguyen. A Zn-based, pillared paddlewheel MOF containing free carboxylic acids via covalent post-synthesis elaboration. *Chemical Communications*, 0(25):3720–3722, 2009.

[155] D. Aaron and C. Tsouris. Separation of CO_2 from flue gas: A review. *Separation Science and Technology*, 40(1-3):321–348, 2005.

[156] K. C. Chong, S. O. Lai, H. S. Thiam, H. C. Teoh, and S. L. Heng. Recent Progress of Oxygen / Nitrogen Separation Using Membrane Technology. *Journal of Engineering Science and Technology*, 11(7):1016–1030, 2016.

[157] D. F. Sanders, Z. P. Smith, R. Guo, L. M. Robeson, J. E. McGrath, D. R. Paul, and B. D. Freeman. Energy-efficient polymeric gas separation membranes for a sustainable future: A review. *Polymer (United Kingdom)*, 54(18):4729–4761, 2013.

[158] C. Hering-Junghans. Metal-Free Nitrogen Fixation at Boron. *Angewandte Chemie International Edition*, pages 6738–6740, 2018.

[159] M.-A. Légaré, G. Bélanger-Chabot, R. D. Dewhurst, E. Welz, I. Krummenacher, B. Engels, and H. Braunschweig. Nitrogen fixation and Reduction At Boron. *Science*, 359(6378):896–900, 2018.

[160] Q. Sun, Z. Li, D. J. Searles, Y. Chen, G. Lu, and A. Du. Charge-Controlled Switchable CO_2 Capture on Boron Nitride Nanomaterials. *Journal of the American Chemical Society*, 135(22):8246–8253, 2013.

[161] J. Azamat, A. Khataee, and F. Sadikoglu. Separation of carbon dioxide and nitrogen gases through modified boron nitride nanosheets as a membrane: Insights from molecular dynamics simulations. *RSC Advances*, 6(97):94911–94920, 2016.

[162] Y. Zhang, Q. Shi, Y. Liu, Y. Wang, Z. Meng, C. Xiao, K. Deng, D. Rao, and R. Lu. Hexagonal Boron Nitride with Designed Nanopores as a High-Efficiency Membrane for Separating Gaseous Hydrogen from Methane. *Journal of Physical Chemistry C*, 119 (34):19826–19831, 2015.

[163] R. Lotfi and Y. Saboohi. Effect of metal doping, boron substitution and functional groups on hydrogen adsorption of MOF-5: A DFT-D study. *Computational and Theoretical Chemistry*, 1044:36–43, 2014.

[164] C. W. Liao, P. S. Tseng, B. K. Chang, and C. Y. Wang. Facilitated hydrogen release kinetics from amine borane functionalization on gate-opening metal-organic framework. *Surface and Coatings Technology*, 350:12–19, 2018.

[165] S. Barman, A. Remhof, R. Koitz, M. Iannuzzi, O. Blacque, Y. Yan, T. Fox, J. Hutter, A. Züttel, and H. Berke. Post-Synthesis Amine Borane Functionalization of a Metal-Organic Framework and Its Unusual Chemical Hydrogen Release Phenomenon. *Chemistry - A European Journal*, 23(37):8823–8828, 2017.

[166] M. M. Khan, S. Shishatskiy, and V. Filiz. Mixed matrix membranes of boron icosahedron and polymers of intrinsic microporosity (PIM-1) for gas separation. *Membranes*, 8(1):1–18, 2018.

[167] J. A. Phillips and C. J. Cramer. Quantum Chemical Characterization of the Structural and Energetic Properties of HCN-BF_3. *Journal of Chemical Theory and Computation*, 1(5):827–833, 2005.

[168] J. A. Phillips, D. J. Giesen, N. P. Wells, J. A. Halfen, C. C. Knutson, and J. P. Wrass. Condensed-phase effects on the structural properties of C_6H_5CN-BF_3 and $(CH_3)_3CCN$-BF_3: IR spectra, crystallography, and computations. *Journal of Physical Chemistry A*, 109(36):8199–8208, 2005.

[169] S. J. Grabowski. Boron and other Triel Lewis Acid Centers: From Hypovalency to Hypervalency. *ChemPhysChem*, 15(14):2985–2993, 2014.

[170] D. G. Smith, P. Jankowski, M. Slawik, H. A. Witek, and K. Patkowski. Basis set convergence of the post-CCSD(T) contribution to noncovalent interaction energies. *Journal of Chemical Theory and Computation*, 10(8):3140–3150, 2014.

[171] E. L. Smith, D. Sadowsky, J. A. Phillips, C. J. Cramer, and D. J. Giesen. A short yet very weak dative bond: Structure, bonding, and energetic properties of N_2-BH_3. *Journal of Physical Chemistry A*, 114(7):2628–2636, 2010.

[172] A. Garcia-leigh and J. N. Murreii. On the Stability of BH_3 and BF_3 Adducts with BF , CO and N_2. *Croatica Chemica Acta*, 57(5):879–886, 1984.

[173] Q. Li and Y. Xu. A Quantum Chemistry Study: A new Kind of Boron Nitrides. *Journal of Computational Chemistry*, 28(8):1446–1455., 2007.

[174] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988.

[175] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs. Auxiliary basis sets to approximate Coulomb potentials. *Chemical Physics Letters*, 242(6):652–660, sep 1995.

[176] S. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics*, 19(4):553–566, oct 1970.

[177] C. Hättig, W. Klopper, A. Köhn, and D. P. Tew. Explicitly correlated electrons in molecules. *Chemical Reviews*, 112(1):4–74, jan 2012.

[178] L. Kong, F. A. Bischoff, and E. F. Valeev. Explicitly correlated R12/F12 methods for electronic structure. *Chemical Reviews*, 112(1):75–107, 2012.

[179] K. A. Peterson, T. B. Adler, and H. J. Werner. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al-Ar. *Journal of Chemical Physics*, 128(8):84102, 2008.

[180] D. A. Sirianni, L. A. Burns, and C. D. Sherrill. Comparison of explicitly correlated methods for computing high-accuracy benchmark energies for noncovalent interactions. *Journal of Chemical Theory and Computation*, 13(1):86–99, 2017.

[181] C. Hättig. Optimization of auxiliary basis sets for RI-MP2 and RI-CC2 calculations: Core–valence and quintuple-ζ basis sets for H to Ar and QZVPP basis sets for Li to Kr. *Phys. Chem. Chem. Phys.*, 7(1):59–66, 2005.

[182] F. Weigand. Hartree–Fock Exchange Fitting Basis Sets for H to Rn. *Journal of Computational Chemistry*, 29(2):167–175, 2008.

[183] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson. Basis-set convergence in correlated calculations on Ne, N_2, and H_2O. *Chemical Physics Letters*, 286(3-4):243–252, 1998.

[184] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2): 1007–1023, 1989.

[185] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus, and W. A. De Jong. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477–1489, 2010.

[186] B. Jeziorski, R. Moszynski, and K. Szalewicz. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chemical Reviews*, 94(7):1887–1930, 1994.

[187] O. Vahtras, J. Almlöf, and M. W. Feyereisen. Integral approximations for LCAO-SCF calculations. *Chemical Physics Letters*, 213(5-6):514–518, 1993.

[188] B. I. Dunlap, J. W. Connolly, and J. R. Sabin. On some approximations in applications of Xα theory. *The Journal of Chemical Physics*, 71(8):3396–3402, 1979.

[189] J. L. Whitten. Coulombic potential energy integrals and approximations. *The Journal of Chemical Physics*, 58(10):4496–4501, 1973.

[190] F. Weigend. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Physical Chemistry Chemical Physics*, 4(18):4285–4291, 2002.

[191] E. G. Hohenstein and C. D. Sherrill. Efficient evaluation of triple excitations in symmetry-adapted perturbation theory via second-order Møller-Plesset perturbation theory natural orbitals. *Journal of Chemical Physics*, 133(10):104107, 2010.

[192] R. M. Parrish, E. G. Hohenstein, and C. D. Sherrill. Tractability gains in symmetry-adapted perturbation theory including coupled double excitations: CCD+ST(CCD) dispersion with natural orbital truncations. *Journal of Chemical Physics*, 139(17): 174102, 2013.

[193] E. G. Hohenstein and C. D. Sherrill. Wavefunction methods for noncovalent interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2 (2):304–326, 2012.

[194] F. Weigend, A. Köhn, and C. Hättig. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *Journal of Chemical Physics*, 116 (8):3175–3183, 2002.

[195] A. J. Stone and A. J. Misquitta. Charge-transfer in Symmetry-Adapted Perturbation Theory. *Chemical Physics Letters*, 473(1-3):201–205, 2009.

[196] B. W. Hopkins and G. S. Tschumper. Ab Initio Studies of $\pi\cdots\pi$ Interactions: The Effects of Quadruple Excitations. *The Journal of Physical Chemistry A*, 108(15):2941–2948, 2004.

[197] P. Jurečka, J. Šponer, J. Černý, and P. Hobza. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics*, 8(17): 1985–1993, 2006.

[198] J. Řezáč, M. Dubecký, P. Jurečka, and P. Hobza. Extensions and applications of the A24 data set of accurate interaction energies. *Physical Chemistry Chemical Physics*, 17(29):19268–19277, 2015.

[199] Y. Zhao and D. G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *Journal of Chemical Physics*, 125(19):194101, 2006.

[200] Y. Zhao and D. G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states,

and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(1-3): 215–241, 2008.

[201] Y. Zhao and D. G. Truhlar. Density functional for spectroscopy: No long-range self-interaction error, good performance for Rydberg and charge-transfer states, and better performance on average than B3LYP for ground states. *Journal of Physical Chemistry A*, 110(49):13126–13130, 2006.

[202] Y. Zhao, N. E. Schultz, and D. G. Truhlar. Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *Journal of Chemical Theory and Computation*, 2(2):364–382, 2006.

[203] D. A. Sirianni, A. Alenaizan, D. L. Cheney, and C. D. Sherrill. Assessment of Density Functional Methods for Geometry Optimization of Bimolecular van der Waals Complexes. *Journal of Chemical Theory and Computation*, 14(6):3004–3013, 2018.

[204] Y. Zhao, N. E. Schultz, and D. G. Truhlar. Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions. *Journal of Chemical Physics*, 123(16):161103, 2005.

[205] R. Z. Khaliullin, E. A. Cobar, R. C. Lochan, A. T. Bell, and M. Head-Gordon. Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals. *Journal of Physical Chemistry A*, 111(36):8753–8765, 2007.

[206] O. A. Stasyuk, R. Sedlak, and P. Hobza. Comparison of the DFT-SAPT and Canonical EDA Schemes for the Energy Decomposition of Various Types of Noncovalent Interactions. *Journal of Chemical Theory and Computation*, 14(7):3440–3450, 2018.

[207] A. J. Misquitta. Charge transfer from regularized symmetry-adapted perturbation theory. *Journal of Chemical Theory and Computation*, 9(12):5313–5326, 2013.

[208] T. Clark. Polarization, donor–acceptor interactions, and covalent contributions in weak interactions: a clarification. *Journal of Molecular Modeling*, 23(297):1–10, 2017.

[209] J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas, and K. D. Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature Communications*, 11(1):1–9, 2020.

[210] N. Norahim, P. Yaisanga, K. Faungnawakij, T. Charinpanitkul, and C. Klaysom. Recent Membrane Developments for CO_2 Separation and Capture. *Chemical Engineering and Technology*, 41(2):211–223, 2018.

[211] J. Ahmad, W. U. Rehman, K. Deshmukh, S. K. Basha, B. Ahamed, and K. Chidambaram. Recent Advances in Poly (Amide-B-Ethylene) Based Membranes for Carbon Dioxide (CO_2) Capture: A Review . *Polymer-Plastics Technology and Materials*, 58(4):366–383, 2019.

[212] Y. Wang, X. Ma, B. Ghanem, F. Alghunaimi, I. Pinnau, and Y. Han. Polymers of intrinsic microporosity for energy-intensive membrane-based gas separations. *Materials Today Nano*, 3:69–95, 2018.

[213] T. Hong, S. Chatterjee, S. M. Mahurin, F. Fan, Z. Tian, D.-e. Jiang, B. K. Long, J. W. Mays, A. P. Sokolov, and T. Saito. Impact of tuning CO_2-philicity in polydimethylsiloxane-based membranes for carbon dioxide separation. *Journal of Membrane Science*, 530:213–219, 2017.

[214] H. M. Lee, I. S. Youn, M. Saleh, J. W. Lee, and K. S. Kim. Interactions of CO_2 with various functional molecules. *Phys. Chem. Chem. Phys.*, 17:10925–10933, 2015.

[215] L. Chen, F. Cao, and H. Sun. Ab initio study of the π-π Interactions between CO_2 and benzene, pyridine, and pyrrole. *International Journal of Quantum Chemistry*, 113 (20):2261–2266, 2013.

[216] M. A. Hussain, Y. Soujanya, and G. N. Sastry. Evaluating the efficacy of amino acids as CO_2 capturing agents: A first principles investigation. *Environmental Science and Technology*, 45(19):8582–8588, 2011.

[217] J. Kim, M. Abouelnasr, L. C. Lin, and B. Smit. Large-scale screening of zeolite structures for CO_2 membrane separations. *Journal of the American Chemical Society*, 135(20):7545–7552, 2013.

[218] E. Haldoupis, S. Nair, and D. S. Sholl. Finding MOFs for highly selective CO_2/N_2 adsorption using materials screening based on efficient assignment of atomic point charges. *Journal of the American Chemical Society*, 134(9):4313–4323, 2012.

[219] K. A. Persson, D. Skinner, W. Chen, S. Cholia, G. Hautier, D. Gunter, S. P. Ong, S. Dacek, W. D. Richards, A. Jain, and G. Ceder. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[220] O. Levy, S. Curtarolo, S. Sanvito, M. B. Nardelli, N. Mingo, and G. L. W. Hart. The high-throughput highway to computational materials design. *Nature Materials*, 12(3): 191–201, 2013.

[221] J. K. Nørskov and T. Bligaard. The catalyst genome. *Angewandte Chemie - International Edition*, 52(3):776–777, 2013.

[222] K. D. Collins, T. Gensch, and F. Glorius. Contemporary screening approaches to reaction discovery and development. *Nature Chemistry*, 6(10):859–871, 2014.

[223] X. Ma, Z. Li, L. E. Achenie, and H. Xin. Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening. *Journal of Physical Chemistry Letters*, 6(18):3528–3533, 2015.

[224] Z. Li, X. Ma, and H. Xin. Feature engineering of machine-learning chemisorption models for catalyst design. *Catalysis Today*, 280:232–238, 2017.

[225] J. P. Janet and H. J. Kulik. Predicting electronic structure properties of transition metal complexes with neural networks. *Chemical Science*, 8(7):5137–5152, 2017.

[226] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):1–14, 2017.

[227] A. Sturluson, M. T. Huynh, A. H. York, and C. M. Simon. Eigencages: Learning a Latent Space of Porous Cage Molecules. *ACS Central Science*, 4(12):1663–1676, 2018.

[228] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K. R. Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

[229] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015.

[230] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. Anatole Von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, 108(5):058301, 2012.

[231] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K. R. Müller, and A. Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015.

[232] F. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, 2015.

[233] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. Von Lilienfeld. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation*, 13(11):5255–5264, 2017.

[234] B. Huang and O. A. Von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *Journal of Chemical Physics*, 145(16), 2016.

[235] T. Bereau, D. Andrienko, and O. A. Von Lilienfeld. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *Journal of Chemical Theory and Computation*, 11(7):3225–3233, 2015.

[236] N. J. Browning, R. Ramakrishnan, O. A. Von Lilienfeld, and U. Roethlisberger. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *Journal of Physical Chemistry Letters*, 8(7):1351–1359, 2017.

[237] G. Csányi, J. R. Kermode, S. De, N. Bernstein, M. Ceriotti, A. P. Bartók, and C. Poelking. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12):e1701816, 2017.

[238] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld. Alchemical and structural distribution based representation for improved QML. *Journal of Chemical Physics*, 148:241717, 2018.

[239] B. Meyer, B. Sawatlon, S. Heinen, O. A. Von Lilienfeld, and C. Corminboeuf. Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts. *Chemical Science*, 9(35):7069–7077, 2018.

[240] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *Journal of Chemical Physics*, 148(24):241717, 2018.

[241] A. P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Physical Review B - Condensed Matter and Materials Physics*, 87(18):1–16, 2013.

[242] C. R. Collins, G. J. Gordon, O. A. Von Lilienfeld, and D. J. Yaron. Constant size descriptors for accurate machine learning models of molecular properties. *Journal of Chemical Physics*, 148(24):241718, 2018.

[243] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer. Persistent homology analysis of brain artery trees. *Annals of Applied Statistics*, 10(1):198–218, 2016.

[244] M. Kramar, A. Goullet, L. Kondic, and K. Mischaikow. Persistence of force networks in compressed granular media. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(4):1–8, 2013.

[245] D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter, and P. J. Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, 6, 2015.

[246] A. Takiyama, T. Teramoto, H. Suzuki, K. Yamashiro, and S. Tanaka. Persistent homology index as a robust quantitative measure of immunohistochemical scoring. *Scientific Reports*, 7(1):1–9, 2017.

[247] A. Marchese and V. Maroulas. Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification*, 12 (3):657–682, 2018.

[248] V. Maroulas, J. L. Mike, and C. Oballe. Nonparametric estimation of probability density functions of random persistence diagrams. *Journal of Machine Learning Research*, 20:1–49, 2019.

[249] V. Maroulas, C. P. Micucci, and A. Spannaus. A stable cardinality distance for topological classification. *Advances in Data Analysis and Classification*, 2019.

[250] V. Maroulas, F. Nasrin, and C. Oballe. A Bayesian Framework for Persistent Homology. *SIAM Journal on Mathematics of Data Science*, 2(1):48–74, 2020.

[251] Z. Cang and G. W. Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, 2017.

[252] Z. Cang, L. Mu, and G. W. Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Computational Biology*, 14(1):1–44, 2018.

[253] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8, 2017.

[254] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit. High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *Journal of Chemical Theory and Computation*, 14(8):4427–4437, 2018.

[255] M. Kimura, I. Obayashi, Y. Takeichi, R. Murao, and Y. Hiraoka. Non-empirical identification of trigger sites in heterogeneous processes using persistent homology. *Scientific Reports*, 8(1):1–9, 2018.

[256] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18:1–35, 2017.

[257] A. Zomorodian and G. Carlsson. Computing persistent homology. *Proceedings of the Annual Symposium on Computational Geometry*, pages 347–356, 2004.

[258] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[259] L. Wasserman. Topological Data Analysis. *Annu. Rev. Stat. App.*, 5:501–532, 2018.

[260] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, and H. Yang. Structure of M^{pro} from COVID-19 virus and discovery of its inhibitors. *Nature*, 582:289–293, 2020.

[261] N. M. O. Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3 (33):1–14, 2011.

[262] L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[263] A. Christensen, F. Faber, L. Bratholm, A. Tkatchenko, K. Muller, and O. von Lilienfeld. Qml: A python toolkit for quantum machine learning. page https://github.com/qmlcode/qml.

[264] C. R. Collins. MolML. page https://github.com/crcollins/molml, 2017.

[265] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:1–7, 2014.

[266] J. Townsend and K. D. Vogiatzis. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *The Journal of Physical Chemistry Letters*, 10 (14):4129–4135, 2019.

[267] T. Helgaker, S. Coriani, P. Jørgensen, K. Kristensen, J. Olsen, and K. Ruud. Recent Advances in Wave Function-Based Methods of Molecular-Property Calculations. *Chemical Reviews*, 112(1):543–631, 2012.

[268] J. Townsend, J. K. Kirkland, and K. D. Vogiatzis. Post-Hartree-Fock methods: configuration interaction, many-body perturbation theory, coupled-cluster theory. In *Mathematical Physics in Theoretical Chemistry*, pages 63–117. 2019.

[269] D. Feller, K. A. Peterson, and J. Grant Hill. On the effectiveness of CCSD(T) complete basis set extrapolations for atomization energies. *Journal of Chemical Physics*, 135: 044102, 2011.

[270] W. Klopper, B. Ruscic, D. P. Tew, F. A. Bischoff, and S. Wolfsegger. Atomization energies from coupled-cluster calculations augmented with explicitly-correlated perturbation theory. *Chemical Physics*, 356(1-3):14–24, 2009.

[271] J. Pu and D. G. Truhlar. Benchmark calculations of reaction energies, barrier heights, and transition-state geometries for hydrogen abstraction from methanol by a hydrogen atom. *Journal of Physical Chemistry A*, 109(5):773–778, 2005.

[272] J. Zhang and E. F. Valeev. Prediction of reaction barriers and thermochemical properties with explicitly correlated coupled-cluster methods: A basis set assessment. *Journal of Chemical Theory and Computation*, 8(9):3175–3186, 2012.

[273] K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu, and E. A. Pidko. Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chemical Reviews*, 119(4):2453–2523, 2019.

[274] D. A. Sirianni, A. Alenaizan, D. L. Cheney, and C. D. Sherrill. Assessment of Density Functional Methods for Geometry Optimization of Bimolecular van der Waals Complexes. *Journal of Chemical Theory and Computation*, 14(6):3004–3013, 2018.

[275] J. M. Martin. Ab initio total atomization energies of small molecules - Towards the basis set limit. *Chemical Physics Letters*, 259(5-6):669–678, 1996.

[276] F. N. Pansini and A. J. Varandas. Toward a unified single-parameter extrapolation scheme for the correlation energy: Systems formed by atoms of hydrogen through neon. *Chemical Physics Letters*, 631-632:70–77, 2015.

[277] T. Helgaker and W. Klopper. Basis-set convergence of correlated calculations on water. *The Journal of Chemical Physics*, 9639:243–252, 1997.

[278] D. P. Tew, W. Klopper, and T. Helgaker. Electron correlation: The many-body problem at the heart of chemistry. *Journal of Computational Chemistry*, 28(8):1307–1320, 2007.

[279] A. Köhn and D. P. Tew. Explicitly correlated coupled-cluster theory using cusp conditions. I. Perturbation analysis of coupled-cluster singles and doubles (CCSD-F12). *Journal of Chemical Physics*, 133(17), 2010.

[280] H. Stoll. The correlation energy of crystalline silicon. *Chemical Physics Letters*, 191 (6):548–552, 1992.

[281] D. G. Liakos, M. Sparta, M. K. Kesharwani, J. M. Martin, and F. Neese. Exploring the accuracy limits of local pair natural orbital coupled-cluster theory. *Journal of Chemical Theory and Computation*, 11(4):1525–1539, 2015.

[282] J. Brabec, J. Lang, M. Saitow, J. Pittner, F. Neese, and O. Demel. Domain-Based Local Pair Natural Orbital Version of Mukherjee's State-Specific Coupled Cluster Method. *Journal of Chemical Theory and Computation*, 14(3):1370–1382, 2018.

[283] K. D. Vogiatzis, E. C. Barnes, and W. Klopper. Interference-corrected explicitly-correlated second-order perturbation theory. *Chemical Physics Letters*, 503(1-3):157–161, feb 2011.

[284] C. Ochsenfeld, J. Kussmann, and D. S. Lambrecht. Linear-Scaling Methods in Quantum Chemistry. In *Reviews in Computational Chemistry*, chapter 1, pages 1–82. John Wiley & Sons, Ltd, 2007. ISBN 9780470116449.

[285] W. A. De Jong, E. Bylaska, N. Govind, C. L. Janssen, K. Kowalski, T. Müller, I. M. Nielsen, H. J. Van Dam, V. Veryazov, and R. Lindh. Utilizing high performance computing for chemistry: Parallel computational chemistry. *Physical Chemistry Chemical Physics*, 12(26):6896–6920, 2010.

[286] S. Lorenz, A. Groß, and M. Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, 395 (4-6):210–215, 2004.

[287] C. M. Handley and P. L. Popelier. Potential energy surfaces fitted by Artificial Neural Networks. *Journal of Physical Chemistry A*, 114(10):3371–3383, 2010.

[288] S. Manzhos and T. Carrington. Using neural networks to represent potential surfaces as sums of products. *Journal of Chemical Physics*, 125(19), 2006.

[289] R. Dawes, D. L. Thompson, A. F. Wagner, and M. Minkoff. Interpolating moving least-squares methods for fitting potential energy surfaces: A strategy for efficient automatic data point placement in high dimensions. *Journal of Chemical Physics*, 128 (8):084107, 2008.

[290] S. Lorenz, M. Scheffler, and A. Gross. Descriptions of surface chemical reactions using a neural network representation of the potential-energy surface. *Physical Review B - Condensed Matter and Materials Physics*, 73(11):1–13, 2006.

[291] Y. Guo, I. Tokmakov, D. L. Thompson, A. F. Wagner, and M. Minkoff. Interpolating moving least-squares methods for fitting potential energy surfaces: Improving efficiency via local approximants. *Journal of Chemical Physics*, 127(21):184108, 2007.

[292] J. Behler, S. Lorenz, and K. Reuter. Representing molecule-surface interactions with symmetry-adapted neural networks. *Journal of Chemical Physics*, 127(1), 2007.

[293] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri. Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. *Journal of Chemical Physics*, 130(13):134101, 2009.

[294] J. B. Witkoskie and D. J. Doren. Neural network models of potential energy surfaces: Prototypical examples. *Journal of Chemical Theory and Computation*, 1(1):14–23, 2005.

[295] T. Morawietz and J. Behler. A density-functional theory-based neural network potential for water clusters including van der waals corrections. *Journal of Physical Chemistry A*, 117(32):7356–7366, 2013.

[296] K. V. Jose, N. Artrith, and J. Behler. Construction of high-dimensional neural network potentials using environment-dependent atom pairs. *Journal of Chemical Physics*, 136 (19), 2012.

[297] V. Botu and R. Ramprasad. Learning scheme to predict atomic forces and accelerate materials simulations. *Physical Review B - Condensed Matter and Materials Physics*, 92(9):094306, 2015.

[298] J. P. Janet and H. J. Kulik. Predicting electronic structure properties of transition metal complexes with neural networks. *Chemical Science*, 8(7):5137–5152, 2017.

[299] J. P. Janet and H. J. Kulik. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *Journal of Physical Chemistry A*, 121(46):8939–8954, 2017.

[300] J. P. Janet, L. Chan, and H. J. Kulik. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *Journal of Physical Chemistry Letters*, 9(5):1064–1071, 2018.

[301] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, and P. Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design and Engineering*, 3(3):442–452, 2018.

[302] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. Anatole von Lilienfeld, K.-R. Mu, and A. Tkatchenko. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *Journal of Physical Chemistry Letters*, 6(12):2326–2331, 2015.

[303] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017.

[304] O. A. Von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16): 1084–1093, 2015.

[305] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev. Machine learning of molecular properties: Locality and active learning. *Journal of Chemical Physics*, 148(24):241727, 2018.

[306] S. Choi, Y. Kim, J. W. Kim, Z. Kim, and W. Y. Kim. Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chemistry - A European Journal*, 24(47):12354–12358, 2018.

[307] W. Pronobis, K. T. Schütt, A. Tkatchenko, and K. R. Müller. Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *European Physical Journal B*, 91(8):178, 2018.

[308] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K. R. Müller, and O. Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15:095003, 2013.

[309] L. Li, T. E. Baker, S. R. White, and K. Burke. Pure density functional for strong correlation and the thermodynamic limit from machine learning. *Physical Review B*, 94(24):245129, 2016.

[310] J. C. Snyder, M. Rupp, K. Hansen, K. R. Müller, and K. Burke. Finding density functionals with machine learning. *Physical Review Letters*, 108(25):253002, 2012.

[311] J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai. Semi-local machine-learned kinetic energy density functional with third-order gradients of electron density. *Journal of Chemical Physics*, 149(7):241705, 2018.

[312] R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Hargus, K.-H. Law, J. L. Klepeis, and D. E. Shaw. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *The Journal of Chemical Physics*, 147(16): 161725, 2017.

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

[314] D. G. Smith, L. A. Burns, D. A. Sirianni, D. R. Nascimento, A. Kumar, A. M. James, J. B. Schriber, T. Zhang, B. Zhang, A. S. Abbott, E. J. Berquist, M. H. Lechner, L. A. Cunha, A. G. Heide, J. M. Waldrop, T. Y. Takeshita, A. Alenaizan, D. Neuhauser, R. A. King, A. C. Simmonett, J. M. Turney, H. F. Schaefer, F. A. Evangelista, A. E. Deprince, T. D. Crawford, K. Patkowski, and C. D. Sherrill. Psi4NumPy: An Interactive Quantum Chemistry Programming Environment for Reference Implementations and Rapid Development. *Journal of Chemical Theory and Computation*, 14(7):3504–3511, 2018.

[315] J. F. Stanton, J. Gauss, J. D. Watts, and R. J. Bartlett. A direct product decomposition approach for symmetry exploitation in many-body methods. I. Energy calculations. *The Journal of Chemical Physics*, 94(6):4334–4345, 1991.

[316] W. J. Hehre, R. Ditchfield, R. F. Stewart, and J. A. Pople. Self-Consistent Molecular Orbital Methods. IV. Use of Gaussian Expansions of Slater-Type Orbitals. Extension to Second-Row Molecules. *The Journal of Chemical Physics*, 52(2769), 1970.

[317] R. A. Kendall, T. H. Dunning, and R. J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics*, 96(9):6796–6806, 1992.

[318] A. D. McLean and G. S. Chandler. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11-18. *The Journal of Chemical Physics*, 72(10): 5639–5648, 1980.

[319] E. Miliordos and S. S. Xantheas. An accurate and efficient computational protocol for obtaining the complete basis set limits of the binding energies of water clusters at the MP2 and CCSD(T) levels of theory: Application to $(H_2O)_m$, m = 2-6, 8, 11, 16, and 17. *Journal of Chemical Physics*, 142(23), 2015.

[320] J. Townsend and K. D. Vogiatzis. Transferable MP2-Based Machine Learning for Accuate Coupled-Cluster Energies. *Journal of Chemical Theory and Computation*, Accepted, 2020.

[321] J. Behler. Perspective: Machine learning potentials for atomistic simulations. *Journal of Chemical Physics*, 145(17):170901, 2016.

[322] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018.

[323] Y. Jing, Y. Bian, Z. Hu, L. Wang, and X. Q. S. Xie. Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS Journal*, 20(3):1–10, 2018.

[324] N. E. Thomford, D. A. Senthebane, A. Rowe, D. Munro, P. Seele, A. Maroyi, and K. Dzobo. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. *International Journal of Molecular Sciences*, 19(6), 2018.

[325] A. Lavecchia. Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discovery Today*, 24(10):2017–2032, 2019.

[326] C. W. Coley, D. A. Thomas, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. John Hart, T. F. Jamison, and K. F. Jensen. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, 365(6453), 2019.

[327] E. Gawehn, J. A. Hiss, J. B. Brown, and G. Schneider. Advancing drug discovery via GPU-based deep learning. *Expert Opinion on Drug Discovery*, 13(7):579–582, 2018.

[328] J. P. Janet, S. Ramesh, C. Duan, and H. J. Kulik. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Central Science*, 6(4):513–524, 2020.

[329] J. P. Janet, T. Z. Gani, A. H. Steeves, E. I. Ioannidis, and H. J. Kulik. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Industrial and Engineering Chemistry Research*, 56(17):4898–4910, 2017.

[330] J. G. Freeze, H. R. Kelly, and V. S. Batista. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chemical Reviews*, 119 (11):6595–6612, 2019.

[331] M. Foscato and V. R. Jensen. Automated in Silico Design of Homogeneous Catalysts. *ACS Catalysis*, 10:2354–2377, 2020.

[332] 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 - Condensed Matter and Materials Physics*, 89(20):1–5, 2014.

[333] O. A. von Lilienfeld. Quantum Machine Learning in Chemical Compound Space. *Angewandte Chemie - International Edition*, 57(16):4164–4169, 2018.

[334] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad. Solving the electronic structure problem with machine learning. *npj Computational Materials*, 5(1), 2019.

[335] X. Lei and A. J. Medford. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Physical Review Materials*, 3(6):1–18, 2019.

[336] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, 2019.

[337] Y. Peng-Jian, M. Sugiyama, K. Tsuda, and T. Yanai. Artificial Neural Networks Applied as Molecular Wave Function Solvers. *Journal of Chemical Theory and Computation*, 2020.

[338] K. T. Schütt, M. Gastegger, A. Tkatchenko, K. R. Müller, and R. J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):1–10, 2019.

[339] T. Nudejima, Y. Ikabata, J. Seino, T. Yoshikawa, and H. Nakai. Machine-learned electron correlation model based on correlation energy density at complete basis set limit. *Journal of Chemical Physics*, 151(2), 2019.

[340] P. Pulay. Localizability of Dynamic Electron Correlation. *Chemical Physics Letters*, 100(2):151–154, 1983.

[341] C. Riplinger and F. Neese. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *Journal of Chemical Physics*, 138(3), 2013.

[342] C. Riplinger, B. Sandhoefer, A. Hansen, and F. Neese. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *Journal of Chemical Physics*, 139(13), 2013.

[343] M. Sparta and F. Neese. Chemical applications carried out by local pair natural orbital based coupled-cluster methods. *Chemical Society Reviews*, 43(14):5032–5041, 2014.

[344] P. R. Nagy and M. Kállay. Approaching the Basis Set Limit of CCSD(T) Energies for Large Molecules with Local Natural Orbital Coupled-Cluster Methods. *Journal of Chemical Theory and Computation*, 2019.

[345] P. R. Nagy, G. Samu, and M. Kállay. Optimization of the Linear-Scaling Local Natural Orbital CCSD(T) Method: Improved Algorithm and Benchmark Applications. *Journal of Chemical Theory and Computation*, 14(8):4193–4215, 2018.

[346] D. G. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of Chemical Physics*, 152(18):184108, 2020.

[347] J. Foster and S. F. Boys. Canonical Configurational Interaction Procedure. *Reviews of Modern Physics*, 32(2):300–302, 1960.

[348] J. Pipek and P. G. Mezey. A fast intrinsic localization procedure applicable for ab initio and semiempirical linear combination of atomic orbital wave functions. *The Journal of Chemical Physics*, 90(9):4916–4926, 1989.

[349] W. J. Hehre, K. Ditchfield, and J. A. Pople. Self-consistent molecular orbital methods. XII. Further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. *The Journal of Chemical Physics*, 56(5):2257–2261, 1972.

[350] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[351] H. Jung, S. Stocker, C. Kunkel, H. Oberhofer, B. Han, K. Reuter, and J. T. Margraf. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem*, 1900052:1–8, 2020.

[352] C. D. Sherrill and P. Piecuch. The X $^1\Sigma_{g'}^+$, B $^1\Delta_{g'}$, and B $^1\Sigma_g^+$ states of C_2: A comparison of renormalized coupled-cluster and multireference methods with full configuration interaction benchmarks. *Journal of Chemical Physics*, 122(12), 2005.

[353] W. Zou and D. Cremer. C_2 in a Box: Determining Its Intrinsic Bond Strength for the X $^1\Sigma_g^+$ Ground State. *Chemistry - A European Journal*, 22(12):4087–4099, 2016.

[354] S. Sharma. A general non-Abelian density matrix renormalization group algorithm with application to the C_2 dimer. *Journal of Chemical Physics*, 142(2), 2015.

[355] M. L. Abrams and C. D. Sherrill. Full configuration interaction potential energy curves for the X $^1\Sigma_g^+$,B $^1\Delta_g$, and B $^1\Sigma_g^+$ states of C_2: A challenge for approximate methods. *Journal of Chemical Physics*, 121(19):9211–9219, 2004.

[356] G. H. Booth, D. Cleland, A. J. Thom, and A. Alavi. Breaking the carbon dimer: The challenges of multiple bond dissociation with full configuration interaction quantum Monte Carlo methods. *Journal of Chemical Physics*, 135(8), 2011.

[357] A. J. Varandas. Extrapolation to the complete-basis-set limit and the implications of avoided crossings: The X $^1\Sigma_g^+$, B $^1\Delta_g$, and B $^1\Sigma_g$ states of C_2. *Journal of Chemical Physics*, 129(23), 2008.

[358] W. Jiang and A. K. Wilson. Multireference composite approaches for the accurate study of ground and excited electronic states: C_2, N_2, and O_2. *Journal of Chemical Physics*, 134(3), 2011.

[359] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. Van Der Avoird. Predictions of the properties of water from first principles. *Science*, 315(5816):1249–1252, 2007.

[360] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. Van Der Avoird. Polarizable interaction potential for water from coupled cluster calculations. I. Analysis of dimer potential energy surface. *Journal of Chemical Physics*, 128(9), 2008.

[361] E. M. Mas, R. Bukowski, K. Szalewicz, G. C. Groenenboom, P. E. Wormer, and A. Van Der Avoird. Water pair potential of near spectroscopic accuracy. I. Analysis of potential surface and virial coefficients. *Journal of Chemical Physics*, 113(16):6687–6701, 2000.

[362] D. G. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *Journal of Physical Chemistry Letters*, 7(12):2197–2203, 2016.

[363] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg. Less is more: Sampling chemical space with active learning. *Journal of Chemical Physics*, 148(24): 1–29, 2018.

[364] M. Gastegger, J. Behler, and P. Marquetand. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science*, 8(10):6924–6935, 2017.

[365] R. Izsák. Single-reference coupled cluster methods for computing excitation energies in large molecules: The efficiency and accuracy of approximations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(3):1–29, 2020.

[366] S. Kiyohara, M. Tsubaki, and T. Mizoguchi. Learning excited states from ground states by using an artificial neural network. *npj Computational Materials*, 6(1):1–6, 2020.

[367] P. O. Dral, M. Barbatti, and W. Thiel. Nonadiabatic Excited-State Dynamics with Machine Learning. *Journal of Physical Chemistry Letters*, 9(19):5660–5663, 2018.

[368] C. Altintas, G. Avci, H. Daglar, A. Nemati Vesali Azar, S. Velioglu, I. Erucar, and S. Keskin. Database for CO_2 Separation Performances of MOFs Based on Computational Materials Screening. *ACS Applied Materials and Interfaces*, 10(20): 17257–17268, 2018.

[369] J. Jiang. Computational screening of metalorganic frameworks for CO_2 separation. *Current Opinion in Green and Sustainable Chemistry*, 16:57–64, 2019.

[370] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, and T. K. Woo. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *Journal of Physical Chemistry Letters*, 5(17):3056–3060, 2014.

[371] Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha, and R. Q. Snurr. In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances*, 2(10), 2016.

Appendix

Supplementary Tables and Figures for Chapter 2: Gas Separations with Polymeric Membranes

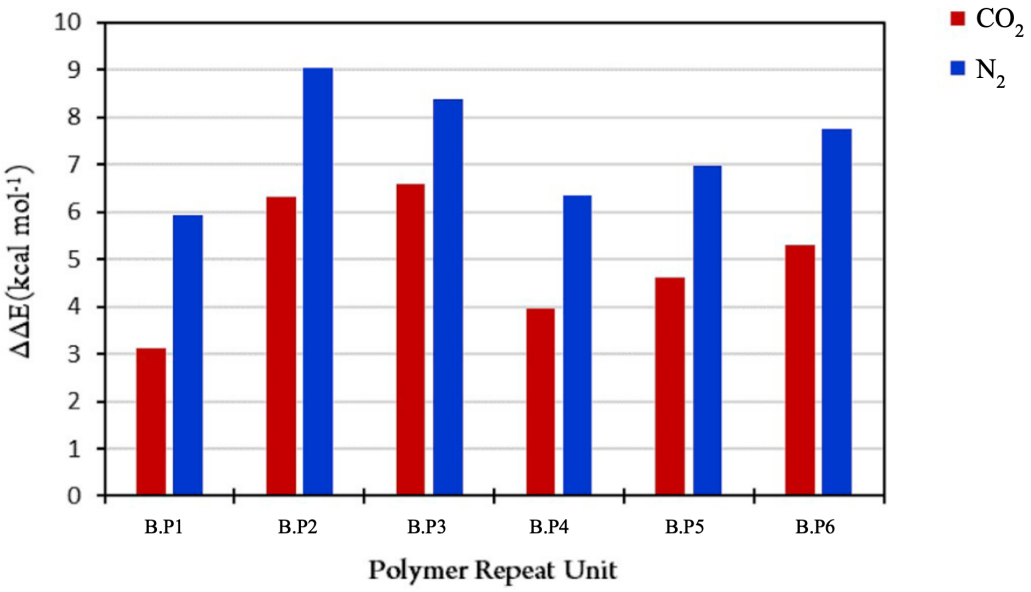


Figure 1: $\Delta\Delta E$ values for saturated repeat units of polymers B.P1–B.P6 for CO_2 and N_2 (kcal mol^{-1}).

Table 1: Interaction energies (ΔE) of CO_2, N_2 and monomers (in kcal mol^{-1}) for polymers C.P1-P7

Polymer	$\Delta E_{\text{mon-CO}_2}$	$\Delta E_{\text{mon-N}_2}$	$\Delta E_{\text{mon-mon}}$
C.P1	-5.03	-1.94	-9.91
C.P2	-5.79	-2.29	-9.39
C.P3	-3.10	-1.68	-7.23
C.P4	-4.51	-2.59	-15.10
C.P5	-4.79	-1.83	-10.57
C.P6	-3.49	-1.88	-8.84
C.P7	-6.18	-2.26	-13.37

Table 2: SAPT0/jun-cc-pVDZ interactions for polymer unit with CO_2 in kcal mol^{-1}

Polymer	Electrostatics	Exchange	Induction	Dispersion	Total
B.P1	-7.81	6.71	-1.29	-4.48	-6.87
B.P2	-7.41	6.74	-1.26	-4.55	-6.49
B.P3	-3.35	3.65	-0.53	-2.43	-2.67
B.P4	-8.97	7.46	-1.48	-2.99	-5.98
B.P5	-9.04	7.47	-1.56	-2.78	-5.91
B.P6	-2.75	3.43	-0.52	-2.95	-2.79
B.P7	-11.61	10.84	-2.20	-4.22	-7.20

Table 3: SAPT0/jun-cc-pVDZ interactions for polymer unit with N_2 in kcal mol^{-1}

Polymer	Electrostatics	Exchange	Induction	Dispersion	Total
B.P1	-0.70	1.73	-0.17	-1.98	-1.12
B.P2	-1.01	1.94	-0.18	-2.35	-1.60
B.P3	-0.84	1.61	-0.19	-1.70	-1.11
B.P4	-1.26	3.24	-0.41	-4.09	-2.52
B.P5	-0.70	1.90	-0.21	-2.13	-1.14
B.P6	-0.88	1.69	-0.20	-1.93	-1.32
B.P7	-1.03	2.07	-0.18	-2.09	-1.24

Table 4: SAPT0/jun-cc-pVDZ interactions for polymer units (monomer-monomer dimer) in kcal mol^{-1}

Polymer	Electrostatics	Exchange	Induction	Dispersion	Total
B.P1	-6.42	13.13	-1.77	-13.67	-8.73
B.P2	-6.82	12.30	-1.70	-13.23	-9.45
B.P3	-4.59	8.44	-1.27	-8.45	-5.87
B.P4	-12.00	18.44	-3.15	-20.27	-16.98
B.P5	-9.38	13.30	-2.58	-13.08	-11.71
B.P6	-8.17	9.37	-2.14	-8.00	-8.93
B.P7	-16.63	21.14	-4.91	-11.52	-11.92

Table 5: PBE0/def2-TZVPP interactions for polymer units (monomer-monomer dimer) in kcal mol^{-1} and the average bond distance between functional units in Å.

Dimer System	Interaction Energy (kcal mol^{-1})	Average H-bond Distance (Å)
I-I	-18.2	1.80
I-II	-17.2	1.80
I-III	-17.5	1.78
I-IV	-17.9	1.80
II-II	-16.4	1.80
II-III	-16.5	1.80
II-IV	-17.0	1.80
III-III	-16.5	1.81
III-IV	-17.0	1.80
IV-IV	-17.1	1.80

Table 6: PBE0/def2-TZVPP interaction energies, polymer oxygen to CO_2 carbon distances, and polymer-CO_2 hydrogen bonding distances for monomer and dimers containing urea based associating units I-IV

System	CO_2 Interaction Energy (kcal mol^{-1})	O-C_{CO_2} (Å)	H-O_{CO_2}
I	-5.2	2.74	2.23
II	-5.0	2.74	2.30
III	-5.7	2.72	2.25
IV	-5.4	2.73	2.24
I-I	-5.4	2.77	2.26
I-II	-5.9	2.66	2.38
I-III	-5.3	2.66	2.34
I-IV	-5.1	2.78	2.26
II-II	-5.2	2.95	2.93
II-III	-5.0	2.75	2.76
II-IV	-5.3	2.95	2.93
III-III	-5.8	3.09	3.10
III-IV	-5.6	3.00	3.03
IV-IV	-5.6	2.96	2.98

Supplementary Tables and Figures for Chapter 3: Understanding the Nature of Weak Interactions Between Functionalized Boranes and N_2/O_2, Promising Functional Groups for Gas Separations

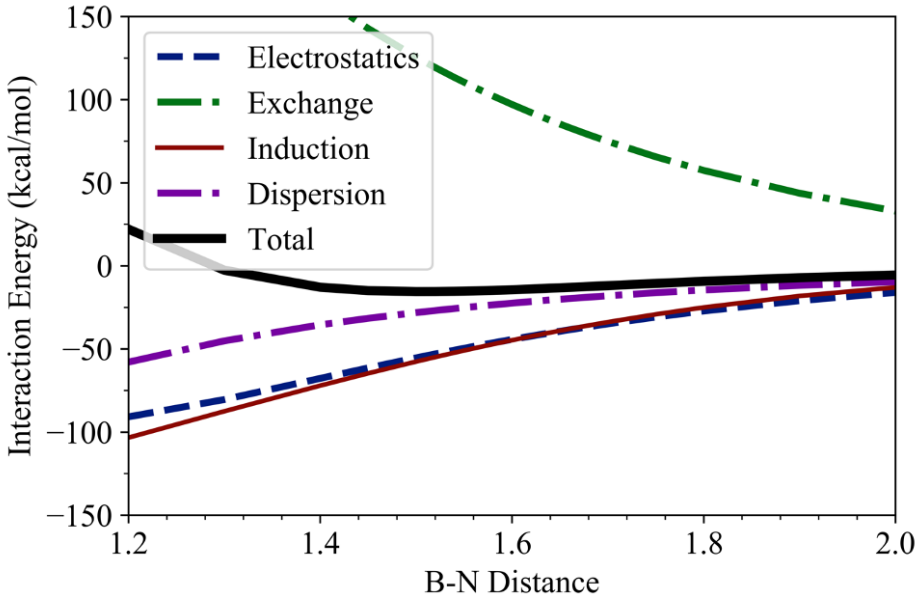


Figure 2: SAPT2+3+δMP2 interaction energies components and total energy for the BH_3-N_2 complex

Table 7: SAPT/aug-cc-pVTZ BH_3-N_2 interaction energies (kcal mol^{-1}) at various levels of the SAPT series and at 1.2, 1.56, and 1.64 Å distances displayed with the reference CCSD(T) method. The respective deformation energy is also reported for completeness.

BH_3-N_2 Distance	1.2 Å	1.56 Å	1.64 Å
SAPT0	3.90	-20.32	-17.19
SAPT2+(3)(CCD)	-2.41	-19.78	-16.30
SAPT2+(3)(CCD)+δMP2	22.59	-14.74	-13.11
SAPT2+3(CCD)	-24.66	-26.73	-21.70
SAPT2+3(CCD)+δMP2	22.06	-15.10	-13.42
CCSD(F12)(T)/cc-pVQZ-F12	20.50	-15.86	-14.14
$\Delta E_{\text{Deformation}}$ CCSD(F12)(T)/cc-pVQZ-F12	23.40	10.35	8.18
$\Delta E_{\text{Binding}}$ CCSD(F12)(T)/cc-pVQZ-F12	43.90	-5.51	-5.96

Table 8: Interaction energies for functionalized boranes B(X)(H_2), where X is one of the functional groups present in structures **1-5** and N_2/O_2 molecules. The optimized B-N distance (Å) and N-B-H angles (taken as an average over the two hydrogens) are provided.

B(X)(H_2)	X	ΔE_{Bind}-N_2	B-N dist.	N-B-H angle	ΔE_{Bind}-O_2	ΔE_{Bind}-N_2/O_2
Reference	H	-6.51	1.588	102.7	-1.00	6.51
1	$COCF_3$	-5.12	1.598	102.5	-1.32	3.88
2	CF_3	-7.49	1.609	102.9	-1.50	5.00
3	$C(CF_3)_3$	-5.00	1.617	101.0	-1.22	4.09
4	CO_2H	-4.82	1.614	102.8	-0.94	5.15
5	SO_3H	-7.39	1.586	104.7	-1.31	5.62

Table 9: DFT (M05/def2-TZVPPD) binding and interaction energies (kcal mol^{-1}) and distances are presented with SAPT2+3(CCD)+δ_{MP2} for the mono-, di-, and triperfluorinated boron interacting with N_2.

	$B(CF_3)H_2$	$B(CF_3)_2H$	$B(CF_3)_3$
$\Delta E_{\text{Binding}}$ (DFT)	-7.49	-8.34	-8.81
$\Delta E_{\text{Deformation}}$ (DFT)	9.78	10.81	11.31
$\Delta E_{\text{Interaction}}$ (DFT)	-17.27	-19.15	-20.13
B-N distance (Å)	1.609	1.629	1.651
$E_{\text{Electrostatic}}$	-43.75	-42.51	-41.94
E_{Exchange}	94.49	89.81	87.10
$E_{\text{Induction}}$	-45.57	-45.41	-44.97
$E_{\text{Dispersion}}$	-23.07	-23.39	-24.11
$\Delta E_{\text{Interaction}}$ (SAPT)	-17.90	-21.49	-23.91

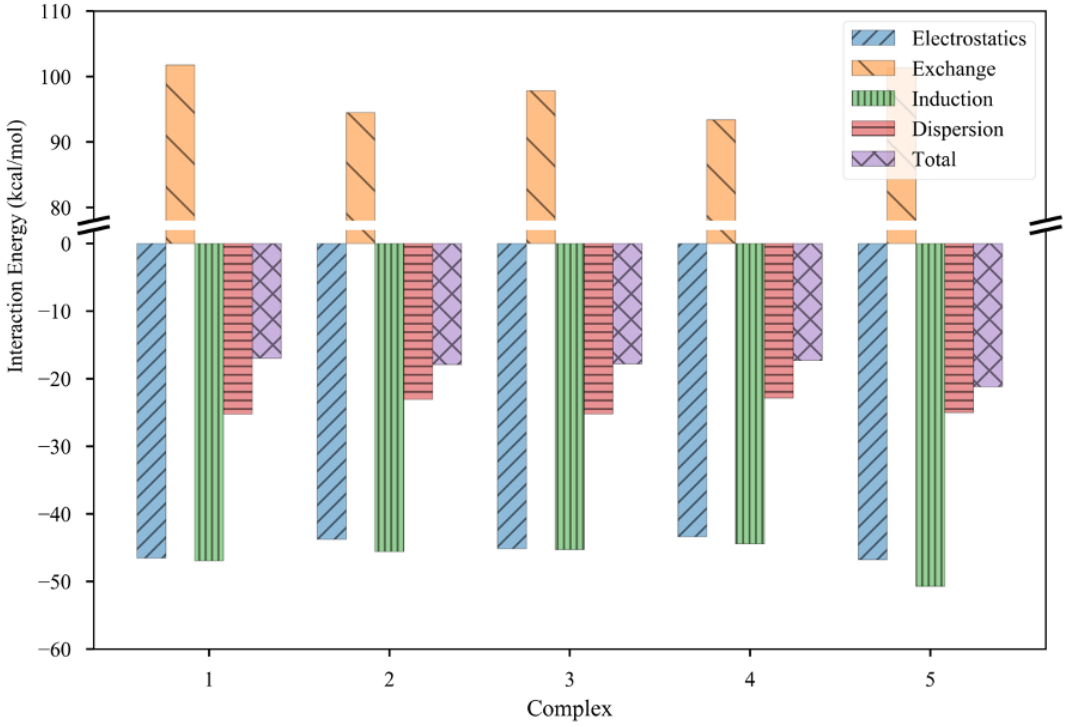


Figure 3: Energy decomposition terms for complexes **1-5** as calculated at the SAPT 2+3+δ_{MP2} level of theory.

Table 10: Binding Energies and their respective counterpoise corrections (CP) in kcal mol^{-1} at the B3LYP-D3(BJ) level of theory for BH_3 and the N_2 and O_2 complexes, respectively

Basis	ΔE	CP Correction	ΔE(CP)	B-(N/O) dist Å
BH_3-N_2				
def2-SVP	-13.00	1.96	-11.04	1.55
def2-TZVPP	-9.41	0.38	-9.03	1.56
def2-TZVPPD	-9.22	0.21	-9.01	1.56
aug-cc-pVTZ	-9.05	0.24	-8.81	1.56
aug-cc-pVQZ	-9.21	0.14	-9.08	1.56
def2-QZVPP	-9.42	0.39	-9.03	1.56
def2-QZVPPD	-9.05	0.02	-9.03	1.56
BH_3-O_2				
def2-SVP	-2.81	1.51	-1.30	2.43
def2-TZVPP	-1.34	0.23	-1.11	2.58
def2-TZVPPD	-1.13	0.05	-1.09	2.61
aug-cc-pVTZ	-1.06	0.09	-0.98	2.60
aug-cc-pVQZ	-1.06	0.03	-1.03	2.60
def2-QZVPP	-1.48	0.33	-1.15	2.56
def2-QZVPPD	-1.14	0.01	-1.13	2.59

Supplementary Tables and Figures for Chapter 4: Representation of Molecular Structures with Persistent Homology for Machine Learning Applications in Chemistry

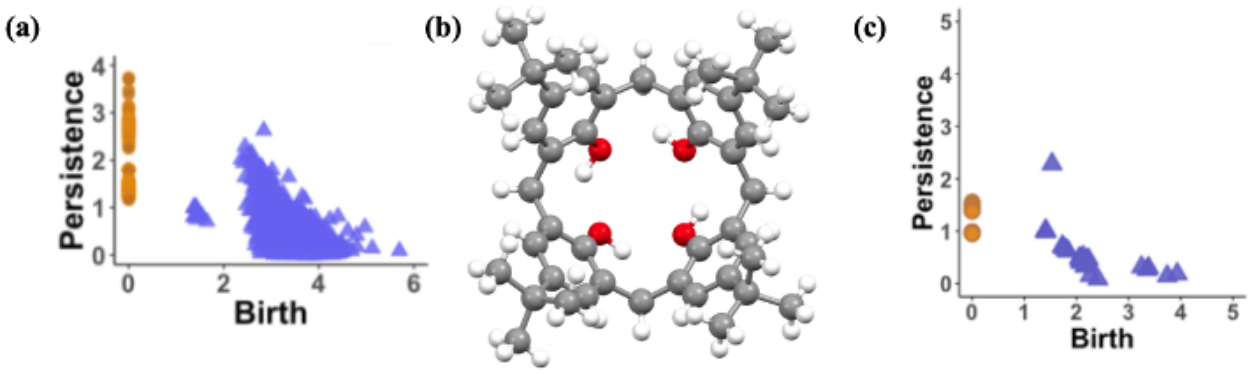


Figure 4: (a) Persistence diagram of 2500 atom main protease in complex with an inhibitor N3 of the new coronavirus (CoV) identified as COVID-19, (b) ball-and-stick model of 4-tertbutylcalix[4]arene (R = tert-butyl groups). (c) The persistent diagram of the 4-tertbutylcalix[4]arene.

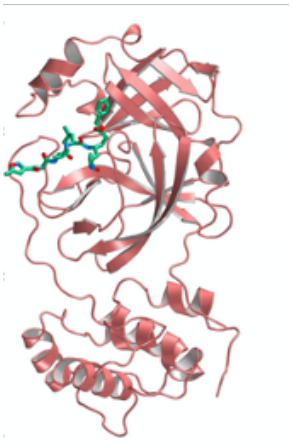


Figure 5: Structure of SARS-CoV-2 main protease

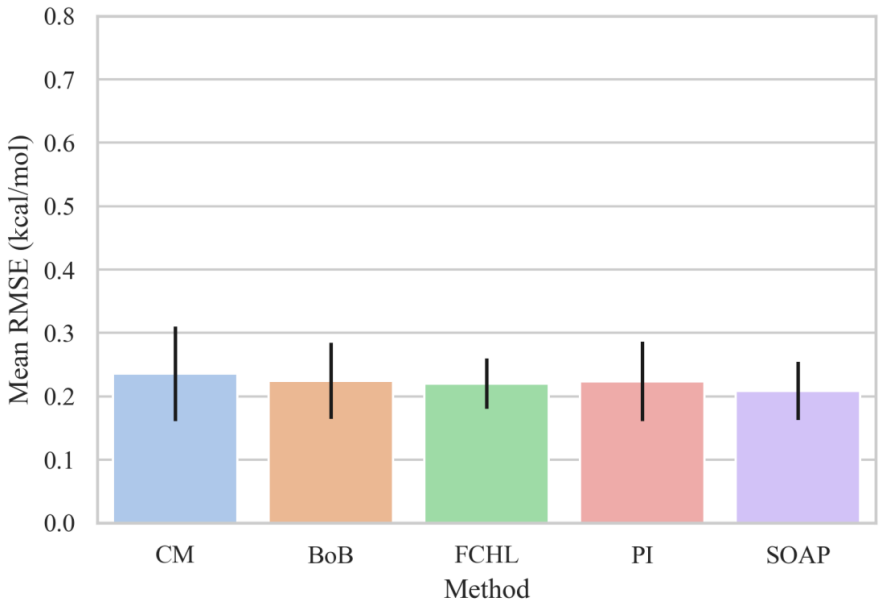


Figure 6: Average RMSE (in kcal mol^{-1}) using 10-fold CV on N_2 interaction energies for Coulomb Matrices (CM), Bag of Bonds (BoB), FCHL representation, Persistence Images (PI), and Smooth Overlap of Atomic Positions (SOAP) representations. The error bars represent the standard deviation of the RMSE. Overall, all models show similar behavior due to the small range of output values since N_2 interaction energies range roughly between -1.6 and - 2.6 kcal mol^{-1}.

Vita

Jacob Andrew Townsend was raised in Kennesaw, GA by parents Brian and Jodi Townsend alongside his younger twin brother, Jamie. In 2011, he graduated from North Cobb High School, and elected to pursue his undergraduate career in chemistry at Kennesaw State University. In 2015, Jacob performed pharmaceutical analytical research with professors Rodolfo Romañach and Rafael Mendez at the University of Puerto Rico Mayaguez. Following his graduation from Kennesaw State University in 2015, Jacob pursued his Ph.D. in Physical Chemistry at the University of Tennessee under Professor Konstantinos D. Vogiatzis. While at the University of Tennessee, Jacob published many manuscripts and was honored as the chemistry department's Glev Mamantov Graduate Research Scholar and won the PsiCon 2018 "Best Lightning Talk" award.