Characterization of a Metal Organic Framework Database

Adam Mirmiran

Thesis submitted in
partial fulfillment of the requirements for the
Master's degree in Chemistry

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa

Table of Contents

Abstract

Metal organic frameworks (MOFs) are nanoporous materials composed of inorganic and organic structural building units (SBUs). Over the last several decades, interest in MOFs has grown considerably partially due to their promising capabilities for carbon capture and storage (CCS) technologies. This is mostly due to their tunable pore chemistry, high internal surface area and unique structural diversity. This thesis focuses on computational methods that were used to analyze and organize a database of hypothetical structures to facilitate MOF discovery. The work done is detailed in two main parts.

In the first part of the thesis, a topologically diverse hypothetical MOF database, containing over 300,000 structures, is screened using simplified molecular-input line-entry system (SMILES) strings to identify SBUs in each structure. The structures in the database are then renamed according to the SBUs identified by the SMILES strings algorithm. The renaming of the structures allows users to have a good idea of the geometrical and topological distribution of the database. Furthermore, a quick and reliable test is developed to identify structures with incorrect bonding patterns/missing hydrogen.

In the second part of the thesis, density functional theory (DFT) - derived charges are generated for each structure in the hypothetical MOF database. Using these charges, the CO_2/N_2 selectivity is calculated and compared with the selectivity values obtained from another charge generating method, split-charge equilibration (SQE), and it is determined that there is good agreement, $r = 0.96$, between the two methods. A machine learning model is then developed to identify relationships between geometrical features and CO_2/N_2 selectivity.

Table of Figures

List of Tables

List of Acronyms

- AP-RDF Atomic Property-weighted Radial Distribution Function
- ASA Accessible Surface Area
- ASR All Solvent Removed
- AV Accessible Volume
- CCS Carbon Capture and Storage
- CoRE Computation Ready Experimental
- CSD Cambridge Structural Database
- DFT Density Functional Theory
- ESP Electrostatic Potential
- FSR Free Solvent Removed
- GCMC Grand Canonical Monte Carlo
- GHG greenhouse gas
- HoA Heat of Adsoprtion
- IAST Ideal Adsorption Solution Theory
- MAE Monoethanolamine
- ML Machine Learning
- MOFs Metal organic frameworks
- PAW Projector Augmented Wave
- PBC Periodic Boundary Condition
- PLD Pore Limiting Diameter
- PSA Pressure Swing Adsorption
- PXRD Powered X-Ray Diffraction
- QMOF Quantum Metal Organic Framework
- QSPR Quantitative Structure-Property Relationship
- RCSR Reticular Chemistry Structure Resource
- RDF Radial Distribution Function
- REPEAT Repeating Electrostatic Potential Extracted Atomic (charge)
- SBUs Structural Building Units
- SMILES Simplified Molecular-Input Line-Entry System
- ToBasCCo Topology-Based Crystal Constructor
- TPSA Temperature-Pressure Swing Adsorption
- TSA Temperature swing adsorption
- VASP Vienna Ab Initio Simulation Package
- XRD X-Ray Diffraction

Acknowledgements

I would like to thank the current and past members of the Woo lab, who have been a tremendous help during my journey as a graduate student. I would also like to give special thanks to Tom Burns, Jake Burner, and Jun Luo, for providing valuable insight during our time spent working together. Most of all, I would like to thank Dr. Woo, who has encouraged me and tested my abilities almost as much as I have tested his patience.

I would also like to give thanks to my friends and family for all their support during these years.

1 Introduction

1.1 The Importance of Carbon Capture

Over the last decades, much attention has been brought to scientists and the public on the potentially devastating effects of climate change. With the looming threat of rising temperatures and sea levels, as well as mass extinction of land and aquatic life, scientists have significantly increased their efforts in finding innovative ways to counteract or slow down the effects of global warming. While the trend between increased greenhouse gas (GHG) emissions, which are made up of mostly CO_2, CH_4, and NO_x ('x' meaning either NO or NO_2) and rising global temperatures has become a common talking point for politicians and the public alike, finding ways to reduce gas emissions remains a difficult problem to tackle. According to the United States Environmental Protection Agency, CO_2 accounted for 80% of the total GHG emissions in 2019, approximately half of which came from stationary energy production sources, such as power plants.[1] Several promising methods for reducing CO_2 emissions from power plants have been proposed including converting to nuclear energy and/or renewable energy, as well as increasing the efficiency of existing power plants. Another method that is frequently mentioned for GHG emission reduction is carbon capture and storage (CCS), which refers to the process of capturing, transporting, and permanently storing CO_2. With the current technology, CCS is most cost effective when performed at stationary sources which produce high quantities of concentrated CO_2 gas emissions, such as power plants, rather than from vehicles. Since power plants account for roughly 30% of all CO_2 emissions, they are ideal for large scale CCS implementation.[2] In most coal power plants, CO_2 is produced as a bi-product during the combustion of fossil fuels (coal, oil, natural gas) to produce water and heat, the latter of which is used to generate electricity. As such, a promising strategy is to introduce CCS technology in

power plants, either during construction or by retro-fitting existing ones, in order to capture GHGs before they are released into the atmosphere.

There exist three main methods for carbon capture at point sources: pre-combustion, post-combustion, and oxyfuel combustion.[3] Fig 1.1 shows basic schemes of these processes.

In oxyfuel combustion, nitrogen is removed from the air before combustion, yielding a stream that is roughly 95% oxygen. This method has several benefits. Since nitrogen is removed before combustion, NO_x production is greatly reduced. Also, the volume of the flue gas (gas produced from burning fossil fuel) is 75% less than gas from air-fired combustion, which allows for easier removal of pollutants such as SO_x, mercury particulates, as well as any remaining NO_x. Lastly, the resulting biproducts of this reaction, CO_2 and water vapor, can be easily separated by dehydration and low temperature purification processes.[4] The main drawback of oxyfuel systems is the energy required to produce a highly pure oxygen stream.[5]

Pre-combustion carbon capture involves gasifying coal in steam and oxygen/air at high temperature and pressure to form carbon monoxide synthesis gas, or syngas, as well as H_2. A water shift reaction then converts carbon monoxide and water into H_2 and CO_2, the latter of which can be captured and stored while H_2 can be combusted for additional electricity production.[6] Theoretically, this method is quite promising, however, current power plants cannot be retro-fitted to include it, thus, only new powerplants can use the technology and there is growing pressure to not continue building fossil fuel burning powerplants.

Lastly, there is post-combustion carbon capture.[7] This method captures CO_2 from flue gases of fossil fuels burned in air. One of the main benefits of the post-combustion method is that most power plants can be retrofitted to include it with relatively little modifications. As such, out of

the three methods mentioned, post-combustion carbon capture is the main CCS technology used in industry.

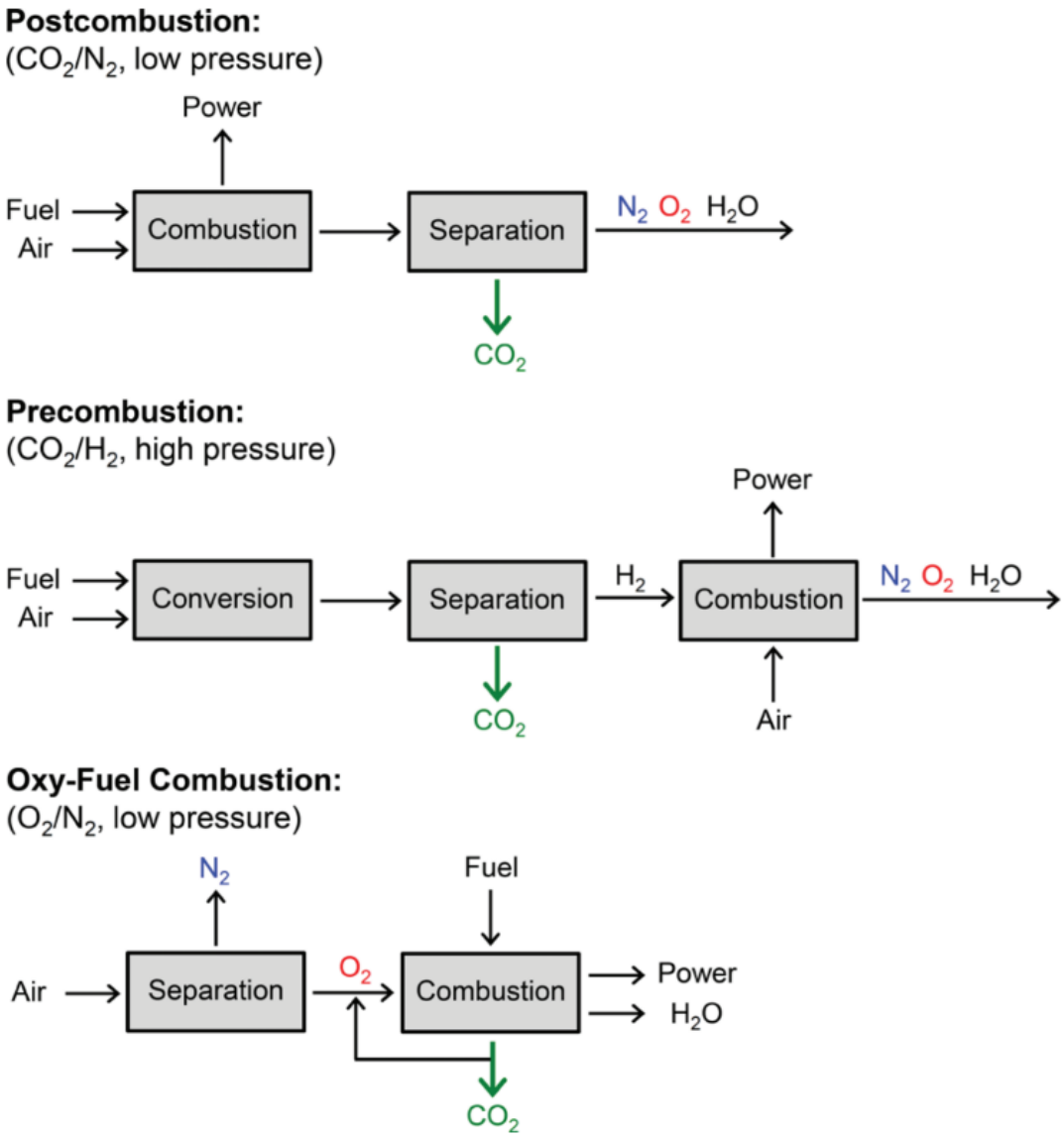


Figure 1.1 Basic schemes for oxy-fuel combustion, precombustion and post-combustion

These technologies are still relatively new and only a select few power plants have been able to implement CCS technology. For example, the Boundary Dam Power Station in

Saskatchewan, Canada, became the first power station in the world the successfully use CCS technology.

One of the main hurdles of promoting CCS is that although reducing CO_2 emission is appealing, from a financial and practical point it can have several drawbacks. Using the Boundary Dam Power Station as an example, since its initiating CO_2 capture in 2014, the plant has had to be shut down and repaired frequently causing it to rarely reach its daily carbon capture goal of 3,200 metrics tons. While the recent announcement from the plant that 4 million tons of CO_2 had been captured since its launch is admirable, it is overshadowed by this milestone being reached almost 3 years later than expected. Furthermore, while the original expectation was to capture 90% of all CO_2 produced by the plant, it has only been able to capture 46% on average, mostly due power outages and maintenance issues. However, these delays are reasonable considering CCS in industry is still an emerging technology.

Post-combustion carbon capture is a broad term that encompasses various methods that remove CO_2 and other gases from burning fuel resources via adsorption[8], absorption[9], membrane separation[10] and chemical reactions.[11] While each method is significant, this work will primarily discuss CO_2 carbon capture via adsorption. Figure 1.2 shows a general scheme of a post-combustion CCS via adsorption. In this scheme, flue gas, composed roughly of 70% N_2, 15% CO_2 and 10% H_2O, interacts with a material, **M**, which will selectively trap CO_2 allowing other the gases to pass through. Then, the CO_2 is released from the material **M-CO2**, using either high temperatures or low pressure to regenerate the material **M**.

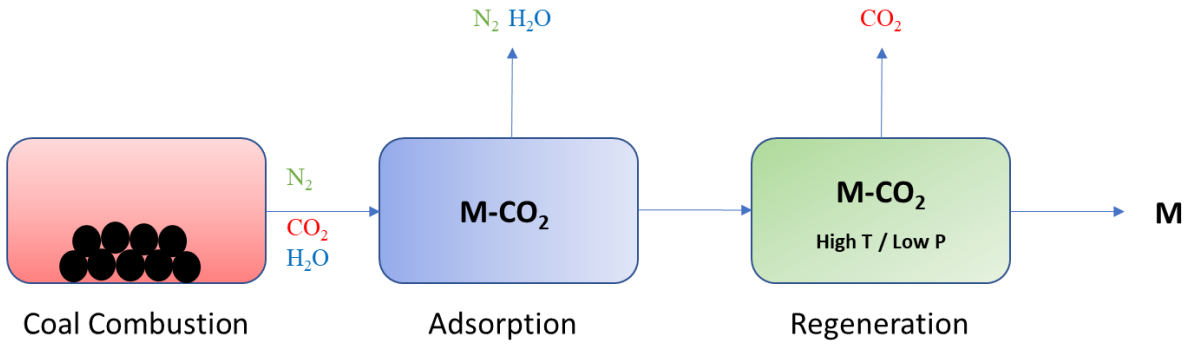


Figure 1.2 Graphic of CCS process from coal combustion. CO_2 is selectivity bound in a material M while other gases pass through. The material is then regenerated, and the CO_2 is compressed and stored.

This process is generally done using cold solutions of various amines, such as monoethanolamine (MAE), to bind the CO_2 and then use high temperature to release the CO_2 once the other gases have passed through the material, as shown in Fig. 1.3

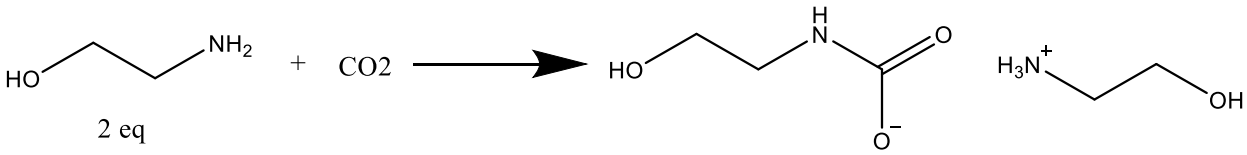


Figure 1.3 Reaction of monoethanolamine and CO_2

The high temperatures (~120 °C) needed for this process require a substantial amount of energy, as such, power plants with CCS technology generally require 10-40% more energy to produce similar outputs as regular power plants, making the process too costly to be financially viable.[12]

To circumvent this issue, many are investigating the benefits of using solid, nano-porous materials in CCS, as doing so could provide many significant advantages. First, the heat capacity of these solids is much lower than water, meaning less energy is used to release the CO_2. Second, the interaction between these materials and CO_2 can be weaker than the amine solutions,

resulting in less energy being required for the regeneration step. Lastly, the interaction between the solids and CO_2 can be reversed by either changing the pressure, called pressure swing adsorption (PSA), or temperature, called temperature swing adsorption (TSA) or a combination of the two, called temperature-pressure swing adsorption (TPSA), after saturation. These techniques have been successfully used for gas separations on various materials including zeolites,[13] activated carbons[14] and most promising, metal organic frameworks (MOFs).[15]

1.2 Metal organic frameworks

Metal Organic Frameworks, shown in Fig 1.4, are highly structured 1,2- or 3-dimensional polymer compounds composed of organic ligands, called linkers, connected by metal nodes. MOFs generally possess low densities, accessible, tunable and often permanent pores ranging from 3 to 100 angstroms, as well as extremely high surface areas, some surpassing 10,000 m^2/g.[16–18]. As such, they have been used for a wide variety of applications, including gas storage and separation, [24–27] biomedicine,[28,29] chemical sensing,[30–32] catalysis,[33–36] as well as energy storage.[37–40] However, the most distinctive characteristic is that they can be combined in a variety of ways by using different organic and inorganic Structural Building Units (SBUs), topologies and functionalization groups. This allows researchers to create/tune MOFs for a certain application by using a specific combination of SBUs, resulting in a near infinite number of possible structures. [19–23]

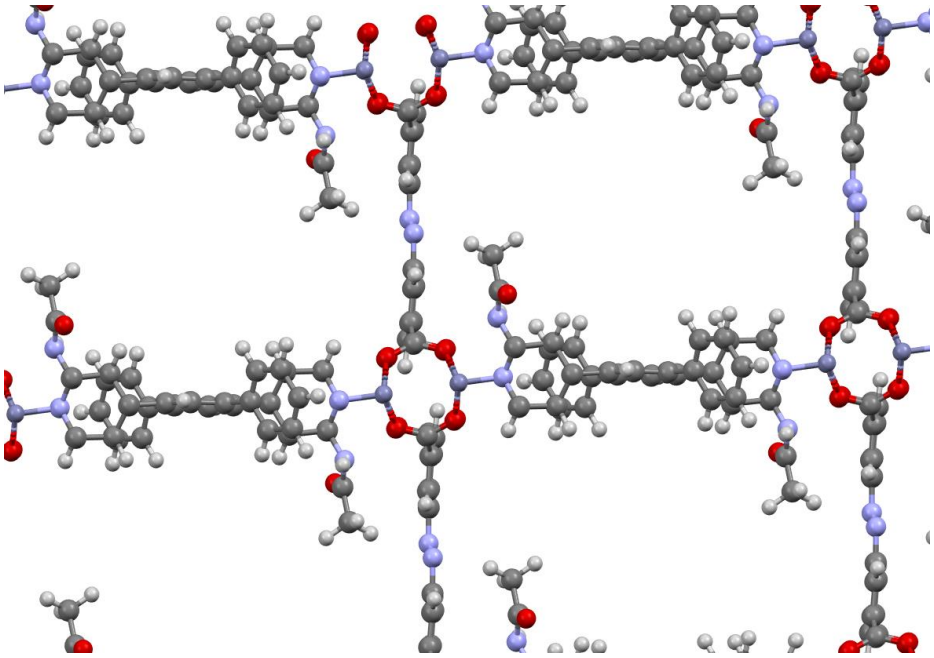


Figure 1.4 Structure of a MOF. Most MOFs are highly ordered and have clearly defined pores.

This, however, is also one of the main challenges MOF researchers face, as finding the best possible structure in an extremely vast chemical space is difficult and time consuming. As of March 2020, there is an estimated 500,000 hypothetical and 100,000 synthesized MOFs.[41]

1.3 Evaluating MOFs for CCS applications

MOFs have been applied to a large variety of applications; water purification,[42] drug storage and delivery,[43,44] optical devices,[45] chemical sensing,[31] catalysis,[46–48] and most notably gas storage and separation.[49,50] When investigating a MOF for gas storage and separation, there are several criteria that govern its effectiveness namely, working capacity, selectivity, and recyclability (regeneration).

1.3.1 Working Capacity

Gas adsorption data is usually generated from a gas adsorption isotherm, shown in Figure 1.5. An isotherm is generated by measuring the amount of gas adsorbed by the material at

various partial pressures and a fixed temperature. Using the isotherm, the working capacity can be calculated by taking the difference in gas adsorbed at the adsorption conditions (low temperature, low purity) and at the desorption conditions (high temperature, high purity).[51]

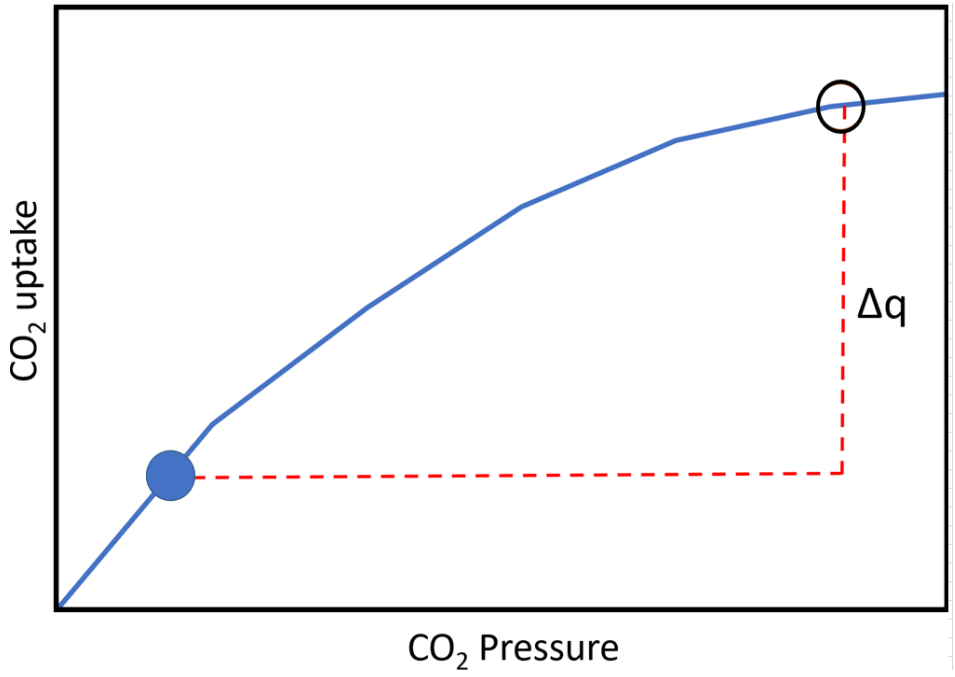


Figure 1.5 Adsorption isotherm of CO_2 in a MOF. Adsorption and desorption pressure values will vary depending on which application/environment the MOF will be used in.

In figure 1.5, the working capacity, Δq, shows how much gas can be removed in one PSA cycle and is measured by taking the difference between the adsorption (empty circle), and desorption (filled circle) conditions. Therefore, the higher the working capacity of a material, the more gas that can be separated in one PSA cycle. It should be noted that experimentally, isotherms are usually only obtainable for one gas at a time. In order to obtain the isotherms when more than one gas is present, one can use single gas isotherms combined with some sort of mixing theory, the most common of which is the Ideal Adsorption Solution Theory (IAST).[52] The method uses two pure isotherms and assumes that the gases mix in an ideal behavior on a homogenous surface. While these assumptions are significant, IAST predicts binary adsorption

well in most cases.[53–55] However, there are some cases where IAST models are poor, such as in materials with flexible frameworks[56,57] and when the competing gases disparate significantly in size.[58]

1.3.2 Selectivity

When testing the adsorption efficacy of a given material, simply testing one guest molecule does not provide enough information. For CCS, materials need to be tested on how well they can discriminate between the various gases in flue gas, such as N_2, O_2, H_2O and CO_2. If, for example, a material adsorbs a lot of CO_2 and N_2, a significant amount of energy will be required to compress and sequester the N_2, drastically increasing the cost of the process. The *selectivity* of a material provides a measurement of how much of a particular gas will adsorb over another. Since nitrogen is a main component of flue gas, and CO_2 is often the desired adsorbent, CO_2/N_2 selectivity calculations are often performed when evaluating MOFs for CCS. As shown in equation 1.1, the selectivity S_{ij} for one gas, i, over another, j, is calculated as a ratio of the mole fractions of i (x_i) over j (x_j) normalized by the ratio of their partial pressures. (p_i / p_j)

$$\boldsymbol{S_{ij} = \frac{x_i p_j}{x_j p_i}} \qquad (\mathbf{1.1})$$

IAST is also used in this calculation to determine the mole fractions of gases in the adsorbed phase. This is because with a single guest, all binding sites in the material are free and there is no competition between the gases and the binding sites but when there are multiple guests, some will have stronger affinities to the material than others. Ignoring the competitive nature of guest binding for binary gases will generally result in low estimations of selectivity.

1.3.3 Stability and Recyclability

Most of the research done on MOFs focuses on validating a material for CCS technology by measuring one or two properties. These measurements are usually done in controlled laboratory environments, thus ignoring some crucial practical limitations.

As such, when evaluating MOFs, stability, either chemical, thermal, and/or mechanical, is a major concern. Chemical stability generally refers to how resistant a structure is to chemicals in the environment such as moisture, solvents, acids, bases, etc. Often the chemical stability can be assessed by comparing powdered X-ray diffraction (PXRD) patterns of a MOF before and after exposure to water or other chemicals.[59] Mechanical stability refers to the preservation of structural integrity after exposure to varying pressures.[60] In many nanoporous materials, a common source of mechanical instability can come from the removal of pore-filling solvent molecules, which often support and maintain the porosity of the structure.[61] It is not uncommon for a MOF to undergo partial or total collapse after solvent removal. Lastly, thermal stability refers to the stability of the MOF under various temperatures, which is especially relevant for carbon capture in TSA. In TSA, once the CO_2 has been captured, it can be indirectly heated to a specific temperature, forcing the high purity stream CO_2 out of the column.[62] Since during this heating process temperatures can reach anywhere between 100-250 °C, a structure with strong thermal stability is required.[63]

Historically MOFs have had issues with stability, especially when compared to zeolites, however this has changed drastically in the recent years with many MOFs showing high thermal, chemical and mechanical stability.[64–66] Furthermore, many groups have started employing computational tools to predict and generate robust structures for specific applications.[67,68]

1.3.4 Topology

The increased interest in crystalline materials came with the realization that these structures could be generated using combinations of clusters of atoms, also called Secondary Building Units (SBUs), that are either organic or metal-containing. The idea would then be to develop theoretical and synthetic methods for which these SBUs could be combined into periodic structures, as shown in Fig 1.6.[69]

This became known as the field of *reticular chemistry*, which had the purpose of understanding and organizing crystalline materials into their underlying topology, also called *net*. For example, in Fig 1.6, the metal SBUs ZnO_4 are displayed as the vertices of the net, and the linking organic SBUs, benzene dicarboxylate, are displayed as edges. By generalizing the connectivity, complex periodic structures can thus be simplified into common geometrical shapes.

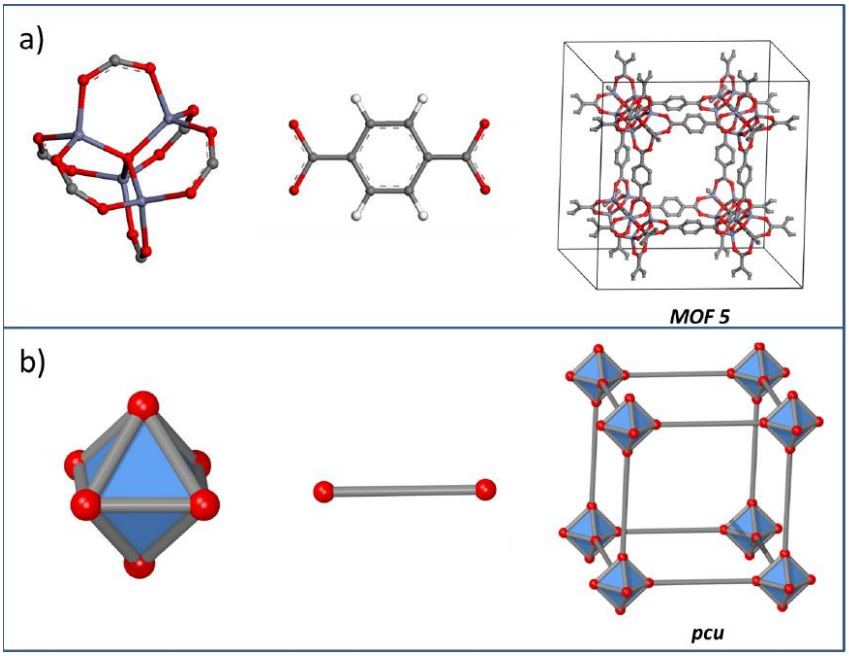


Figure 1.6 a) Metal and organic SBUs which make up MOF-5 b) representation of SBUs as simple geometrical shapes.

The classification, identification and retrieval of these shapes gave rise to the Reticular Chemistry Structure Resource (RCSR)[70] which currently contains about 3000 nets. The database allows for the search of materials using keywords, names, attributes as well as topological three-letter symbols, called RCSR symbols. These symbols can be abbreviations of the structural type or mineral name such as **pcu** for primitive cubic, **sql** for square lattice, **dia** for diamond and **qtz** for quartz. Others are short forms of the chemical formula that make up the material such as **cds** for $CdSO_4$, **srs** for $SrSi_2$ and **pts** for PtS. However, these are the most common ones, and the majority symbols are just randomly assigned.

The common symbols, such as **dia**, gave rise to numerous similar nets which created extensions to be basic three-letter naming scheme, as shown in Fig 1.7. Some common extensions of **dia** include **dia-a,** in which the vertices of **dia** are replaced with tetrahedron of

vertices, or **dia-e** where tetrahedron are placed in the middle of the edges, as well as many others.

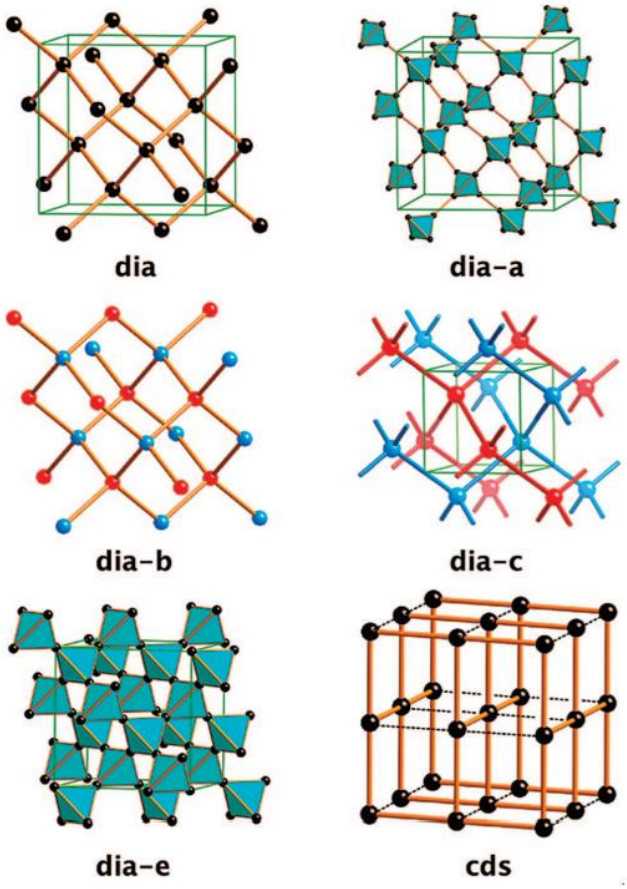


Figure 1.7 Common dervivaties of the popular 'dia' topology.

To assist with this rapidly increasing numbers of nets, several tools have been developed to perform comprehensive topological analysis of crystalline structures.[71–75]

The classification and understanding of nets allow for a highly specific type of MOF discovery in which materials with promising properties can be slightly modified while keeping the same underlying topology, to ideally produce promising candidates. For example, Yan et al.[76] synthesized **NOTT-112**[76] (named after the University of Nottingham), an **rht**-type MOF, which demonstrated exceptionally high H_2 storage capacities and later they also synthesized two

expanded versions namely, **NOTT-116**[77] and **NOTT-119**[78] which are isoreticular with **NOTT-112** but had increase ligand sizes. As expected, the group found that by extending the appropriate ligands and therefore increasing the surface area, the H_2 adsorption capacity increased as well. There is still a fine balance required as other properties may suffer as a result of extending edge lengths, such as the **rht**-based MOF **NU-100**[79] which has longer ligands than **NOTT-119** but undergoes total loss of porosity upon thermal desolvation and requires advanced drying methods to retain its crystallinity.[80] However, the chemical diversity of structures developed this way is in question. For example, highly successful experimental groups will incentivize other groups to investigate the systematic synthesis of similar MOF, resulting in a lot of resources spent on a specific subset of structures.

1.4 Metal Organic Framework Databases

As mentioned previously, one of the main advantages of MOFs is their tunability; by changing the organic or metal SBUs, one can drastically alter the properties of a given structure. As such, when researching how to improve CCS technology, scientists have undergone the momentous task of identifying which MOFs would be the best candidates for CO_2 adsorption. Over the last several decades, MOF-related publications have continuously increased and while this exposure has been a tremendous asset to MOF research, the vast chemical and topological space of these structures creates a problem when trying to identify the best candidates. The majority these structures will yield unsatisfactory results for CCS technology and the traditional method of synthesizing and analyzing CO_2 adsorption performance is simply too slow and costly.[81]

As such, computational methods have been developed to efficiently navigate this space and several groups have made significant contributions in trying to answer the following question:

given millions of potential structures, how does one identify the best for a specific application? Part of the answer may lie in the advancement of MOF repositories, or databases.

1.4.1 The Cambridge Structural Database

The Cambridge Structural Database (CSD)[82] is a popular repository for organic, metal-organic and organometallic structures that are generally characterized using X-ray diffraction (XRD). The CSD, which contains over 100,000 synthesized structures in the 'MOF' subset, has allowed many research groups to perform large-scale screening studies.[83] In 2013, Goldmith and al.[84] screened 38,000 MOFs from the CSD for their H_2 storage capacities and in 2012, Watanabe et al.[85] looked at the pore characteristics and selectivity of 1163 MOFs. Both studies showed the feasibility of large-scale computational screening. However, the structures in the CSD are not computation-ready, meaning that they cannot be used directly for molecular simulations of gas adsorption. There are generally two main hurdles a computational scientist must face before being able to use structures from the CSD database: solvent/counter ions in the pores and structural disorder/missing hydrogen. Solvent in pores, which almost 90% of MOFs in the CSD contain, must be manually removed from each structure before gas simulations can be performed to mimic the effects of experimental *activation*. Activation is the process where heat and/or vacuum is applied to remove residual solvents form the pores of the MOF, thereby allocating space for gas molecules to adsorb. This solvent-removal step is highly time consuming and requires a lot of manual effort.

The second obstacle is the presence of incomplete structures (missing bonds/atoms) and structures that exhibit disorder, which are artifacts of XRD. When using XRD, it is difficult to define atoms that only weakly scatter X-rays, such as hydrogen, which often leads to the

resulting structures missing hydrogen entirely. Furthermore, some ligands that can adopt multiple configurations will appear disordered as shown in Fig 1.8

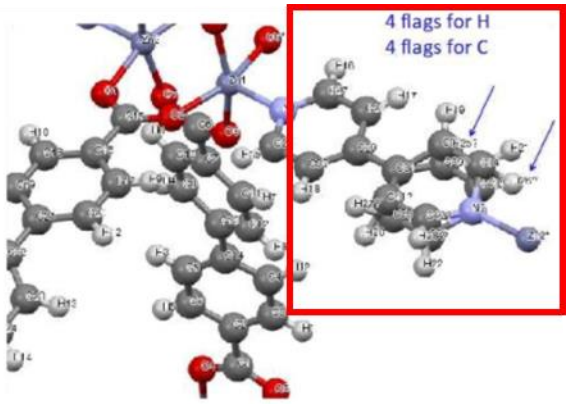


Figure 1.8 Multiple configurations of organic ligand showing 8 carbon and hydrogen atoms in a 6 membered ring.

This issue is well documented, for example, Watanabe et al.[85] extracted 30,000 structures from the CSD and had to perform thorough examinations to ensure structural integrity of the materials. A considerable effort was put into removing solvent molecules in pores and removing disordered atoms after which the resulting structures had to be re-examined to ensure that the cleanup was successful. Some structures also had unphysical disorder (overlapping atoms/extremely short bond distances) and were excluded. After the cleanup process, only 1163 of the 30 0000 structures remained. Similarly, Goldsmith excluded about half of the original structures for similar reasons.

To minimize this issue, a database was designed specifically to facilitate high-throughput screening of a large number of structures.[86]

1.4.2 Computation-Ready Experimental Database

To address the issues mentioned previously, the Computation-Ready Experimental (CoRE) MOF database uses several automated, in-house scripts to make framework structures

computation-ready.[87] Although it only boasts a mere ~14,000 MOFs, the CoRE MOF 2019 database contains structures derived from synthesized materials with the synthesis protocols usually available, which can assist in the synthesis of promising MOFs after high-throughput screening studies. From the CoRE database 2019, two sets of databases were generated: a free solvent removed (FSR) and an all solvent removed (ASR) database. Figure 1.9 shows the distinction between (left) original structure; (middle) free solvent removed; (right) all solvent removed.[87] Bound solvents generally refers to molecules that are part of the framework as determined by the van der wall radii of atoms plus 0.4 Angs.[88] Free solvent simply refers to solvent that is not bound to the framework but are present in the cavity of the MOF.

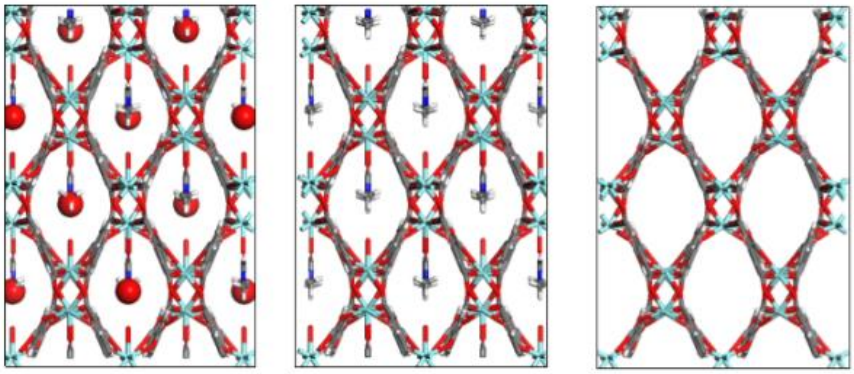


Figure 1.9 Cross section of a MOF where (left) No solvent is removed, (middle) Lightly bound solvent is removed and (right) all atoms except framework atoms are removed. Reproduced from [87]

Having solvent, bound or free, in a MOF can drastically hinder the adsorbates (CO_2 / CH_4) access to the pores, skewing absorbance measurements. Thus by having no solvent it becomes much more efficient to screen many structures as researchers do not need to spend time cleaning the structures.[89]

While the CoRE 2019 database has proven itself to be an extremely powerful tool in the discovery of novel material for CCS,[90,91] a significant amount of manual labor is required to

introduce a structure to the database. Even with this intensive care, several inconsistencies have been reported with both the CSD and CoRE databases. For example, to generate the CoRE ASR database, solvent molecules were simply deleted from crystal structure files, which does not always accurately reflect what the structure will look like upon activation (evacuation of solvent).[88] Some MOFs may be destabilized by the removal of solvents, causing the framework to collapse.[92]

This was demonstrated by Nazarian et al.[93] who used high level computational methods to optimize more than 800 MOFs and found significant structural changes, specifically in materials that were crystallographically resolved with solvent present. They also found that the optimization has a large impact on the simulated gas adsorption properties and noticed a change of more than 25% in over 25% of the structures.

Altintas et al.[94] found that between the ~3500 common MOFs shared between the CSD and CoRE databases, 387 would yield different gas uptakes depending on which database the structure came from.

1.4.3 Hypothetical MOF Databases

Several research groups have developed computational methods to generate hypothetical MOFs, or HMOFs, which are structures that are iteratively assembled using common building blocks found in existing MOFs.[95–97] Since there can be many different combinations of a given set of building blocks, researchers are able to quickly generate many theoretical structures. From there, general trends between MOF properties and various CCS applications can be investigated.

Taking into consideration the vast chemical space created by the various combinations of SBUs with varying topologies, one of the most effective tools for porous crystalline material discovery are hypothetical MOF databases. Wilmer et al.[97] were the first to generate a

hypothetical database, consisting of 137,000 MOFs. This was done using “snapping” algorithm which combined a set a 97 organic SBUs and 5 metal SBUs, partially shown in fig 1.10, derived from crystallographic data of already synthesized MOFs.

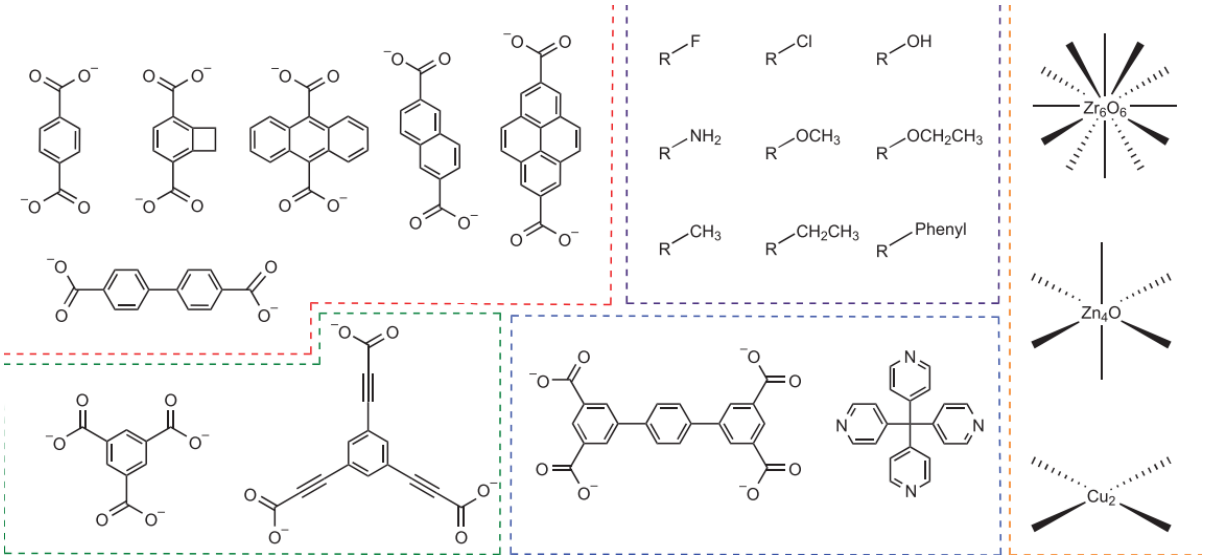


Figure 1.10 Subset of organic and inorganic SBUs used by Wilmer to generate the hypothetical database. Reproduced from [97]

Various structural properties were computed of every structure, including surface area, powder X ray diffraction pattern, and porosity. Using this data, they were able to identify over 300 promising structures with ground-breaking methane storage capacity. Furthermore, they were able to establish important structure-property relationships, such as the relationship between volumetric surface area, gravimetric surface area, and void fraction, with methane adsorption. This work pioneered the use of computational methods for rational MOF design. However, the database did have a few significant drawbacks. Firstly, the 137,000 structures shared a common ~3000 unfunctionalized framework, resulting in a significant lack of diversity, limiting the chemical and structural space sampled. Furthermore, many unphysical structures were later identified, where atoms were too close to each other. Lastly, no bonding information was included, restricting the possibility for further computational work to correct the structures.

As stated by the authors, chemical intuition is absent and low-cost computation can miss important species or lead unrealistic results.

Boyd et al.[20] used a combination of a 'snapping' algorithm, similar to the one used by Wilmer, and a topological builder to generate a database containing over 325,000 topologically diverse hypothetical crystalline structures. They then screened the database to identify and synthesize MOFs containing hydrophobic CO_2 binding sites for post-combustion carbon capture.

1.5 Thesis goals and Outline

The thesis is organized in the following way.

Chapter 2 explores important derivations of commonly used equations for gas adsorption calculations. Various popular approximative models are also discussed as well as emerging methods in MOF discovery.

Chapter 3 describes working with a topologically diverse database of hypothetical MOFs containing over 325,000 structures. In this project, the structures are screened for metal and organic SBUs as well as functional groups using Simplified Molecular Input Line-Entry System (SMILES) strings to rename the structures according to their underlying topologies as well as SBUs. Common geometrical features are also analyzed as well as the topological and chemical diversity of the database. The goal of this chapter is to facilitate usage of the database by developing a clear and logical naming convention of each structure.

Chapter 4 demonstrates the generation of DFT-derived electrostatic potential (ESP) charges using the Vienna Ab initio Simulation Package (VASP) package and the REPEAT method for each structure in the database mentioned previously. The CO_2/N_2 selectivity values obtained from DFT-derived charges are compared to selectivity values of another popular method, SQE. The goal of this chapter is to generate high-level DFT-derived charges on a large and diverse set

of MOFs to facilitate high-level screening and machine learning development studies.

A simple machine learning model is developed to provide an example of how this database could be used.

Chapter 5 will go over the various conclusions of each chapter as well as any potential future work.

1.6 References

1. Inventory of U.S Greenhouse Gas Emissions and Sinks. Published online 2019.

2. Smit B. Carbon Capture and Storage: Introductory lecture. *Faraday Discuss*. 2016;192:9-25. doi:10.1039/C6FD00148C

3. Kanniche M, Gros-Bonnivard R, Jaud P, Valle-Marcos J, Amann JM, Bouallou C. Pre-combustion, post-combustion and oxy-combustion in thermal power plant for CO2 capture. *Appl Therm Eng*. 2010;30(1):53-62. doi:10.1016/j.applthermaleng.2009.05.005

4. Stanger R, Wall T, Spörl R, et al. Oxyfuel combustion for CO2 capture in power plants. *Int J Greenh Gas Control*. Published online 2015. doi:10.1016/j.ijggc.2015.06.010

5. Senior CL, Morris W, Lewandowski TA. Emissions and risks associated with oxyfuel combustion: State of the science and critical data gaps. *J Air Waste Manag Assoc*. 2013;63(7):832-843. doi:10.1080/10962247.2013.791892

6. Pardemann R, Meyer B. Pre-Combustion Carbon Capture. *Handb Clean Energy Syst*. Published online 2015:1-28. doi:10.1002/9781118991978.hces061

7. Basile A, Gugliuzza A, Iulianelli A, Morrone P. Membrane technology for carbon dioxide (CO2) capture in power plants. *Adv Membr Sci Technol Sustain Energy Environ Appl*. 2011;(2008):113-159. doi:10.1533/9780857093790.2.113

8. Tlili N, Grévillot G, Vallières C. Carbon dioxide capture and recovery by means of TSA and/or VSA. *Int J Greenh Gas Control*. 2009;3(5):519-527. doi:10.1016/j.ijggc.2009.04.005

9. Yu CH, Huang CH, Tan CS. A review of CO2 capture by absorption and adsorption. *Aerosol Air Qual Res*. 2012;12(5):745-769. doi:10.4209/aaqr.2012.05.0132

10. Ahmad F, Lau KK, Shariff AM, Murshid G. Process simulation and optimal design of membrane separation system for CO 2 capture from natural gas. *Comput Chem Eng*. 2012;36(1):119-128. doi:10.1016/j.compchemeng.2011.08.002

11. Wu X, Yu Y, Qin Z, Zhang Z. The advances of post-combustion CO2 capture with chemical solvents: Review and guidelines. *Energy Procedia*. 2014;63:1339-1346. doi:10.1016/j.egypro.2014.11.143

12. Eldardiry H, Habib E. Carbon capture and sequestration in power generation: review of impacts and opportunities for water sustainability. *Energy Sustain Soc*. 2018;8(1). doi:10.1186/s13705-018-0146-3

13. Mehdipour M, Fatemi S. Modeling of a PSA-TSA Process for Separation of CH 4 from C 2 Products of OCM Reaction. *Sep Sci Technol*. 2012;47(8):1199-1212. doi:10.1080/01496395.2011.644019

14. Plaza MG, Pevida C, Pis JJ, Rubiera F. Evaluation of the cyclic capacity of low-cost carbon adsorbents for post-combustion CO2 capture. *Energy Procedia*. 2011;4:1228-1234. doi:10.1016/j.egypro.2011.01.178

15. Harlick PJE, Tezel FH. Equilibrium analysis of cyclic adsorption processes: CO2 working capacities with NaY. *Sep Sci Technol*. 2005;40(13):2569-2591. doi:10.1080/01496390500283233

16. Xuan W, Zhu C, Liu Y, Cui Y. Mesoporous metal–organic framework materials. *Chem Soc Rev*. 2012;41(5):1677-1695. doi:10.1039/c1cs15196g

17. Rocío-Bautista P, Taima-Mancera I, Pasán J, Pino V. Metal-organic frameworks in green analytical chemistry. *Separations*. 2019;6(3):1-21. doi:10.3390/separations6030033

18. Martin RL, Haranczyk M. Exploring frontiers of high surface area metal–organic frameworks. *Chem Sci*. 2013;4(4):1781-1785. doi:10.1039/c3sc00033h

19. Pulido A, Chen L, Kaczorowski T, et al. Functional materials discovery using energy-structure-function maps. *Nature*. 2017;543(7647):657-664. doi:10.1038/nature21419

20. Boyd PG, Chidambaram A, García-Díez E, et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*. Published online 2019. doi:10.1038/s41586-019-1798-7

21. Hendon CH, Rieth AJ, Korzyński MD, Dincă M. Grand Challenges and Future Opportunities for Metal-Organic Frameworks. *ACS Cent Sci*. 2017;3(6):554-563. doi:10.1021/acscentsci.7b00197

22. Wang Q, Astruc D. State of the Art and Prospects in Metal-Organic Framework (MOF)-Based and MOF-Derived Nanocatalysis. *Chem Rev*. 2020;120(2):1438-1511. doi:10.1021/acs.chemrev.9b00223

23. Li B, Wen HM, Cui Y, Zhou W, Qian G, Chen B. Emerging Multifunctional Metal–Organic Framework Materials. *Adv Mater*. 2016;28(40):8819-8860. doi:10.1002/adma.201601133

24. Yang H, Wang Y, Krishna R, et al. Pore-Space-Partition-Enabled Exceptional Ethane Uptake and Ethane-Selective Ethane-Ethylene Separation. *J Am Chem Soc*. 2020;142(5):2222-2227. doi:10.1021/jacs.9b12924

25. Zhao X, Wang Y, Li DS, Bu X, Feng P. Metal–Organic Frameworks for Separation. *Adv Mater*. 2018;30(37). doi:10.1002/adma.201705189

26. Lin RB, Xiang S, Xing H, Zhou W, Chen B. Exploration of porous metal–organic frameworks for gas separation and purification. *Coord Chem Rev*. 2019;378:87-103. doi:10.1016/j.ccr.2017.09.027

27. Li H, Li L, Lin R-B, et al. Porous metal-organic frameworks for gas storage and separation: Status and challenges. *EnergyChem*. 2019;1(1):100006. doi:10.1016/j.enchem.2019.100006

28. Doonan C, Riccò R, Liang K, Bradshaw D, Falcaro P. Metal-Organic Frameworks at the Biointerface: Synthetic Strategies and Applications. *Acc Chem Res*. 2017;50(6):1423-1432. doi:10.1021/acs.accounts.7b00090

29. Lu K, Aung T, Guo N, Weichselbaum R, Lin W. Nanoscale Metal–Organic Frameworks for Therapeutic, Imaging, and Sensing Applications. *Adv Mater*. 2018;30(37). doi:10.1002/adma.201707634

30. Koo W-T, Jang J-S, Kim I-D. Metal-Organic Frameworks for Chemiresistive Sensors. *Chem*. 2019;5(8):1938-1963. doi:10.1016/j.chempr.2019.04.013

31. Kreno LE, Leong K, Farha OK, Allendorf M, Van Duyne RP, Hupp JT. Metal-organic framework materials as chemical sensors. *Chem Rev*. 2012;112(2):1105-1125. doi:10.1021/cr200324t

32. Wang L, Xu H, Gao J, Yao J, Zhang Q. Recent progress in metal-organic frameworks-based hydrogels and aerogels and their applications. *Coord Chem Rev*. 2019;398:213016. doi:10.1016/j.ccr.2019.213016

33. Wang L, Agnew DW, Yu X, Figueroa JS, Cohen SM. A Metal–Organic Framework with Exceptional Activity for C−H Bond Amination. *Angew Chemie - Int Ed*. 2018;57(2):511-515. doi:10.1002/anie.201709420

34. Liu L, Zhou TY, Telfer SG. Modulating the performance of an asymmetric organocatalyst by tuning its spatial environment in a metal-organic framework. *J Am Chem Soc*. 2017;139(39):13936-13943. doi:10.1021/jacs.7b07921

35. Genna DT, Pfund LY, Samblanet DC, Wong-Foy AG, Matzger AJ, Sanford MS. Rhodium Hydrogenation Catalysts Supported in Metal Organic Frameworks: Influence of the Framework on Catalytic Activity and Selectivity. *ACS Catal*. 2016;6(6):3569-3574. doi:10.1021/acscatal.6b00404

36. Li Z, Schweitzer NM, League AB, et al. Sintering-Resistant Single-Site Nickel Catalyst Supported by Metal-Organic Framework. *J Am Chem Soc*. 2016;138(6):1977-1982. doi:10.1021/jacs.5b12515

37. Baumann AE, Burns DA, Liu B, Thoi VS. Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. *Commun Chem*. 2019;2(1):1-14. doi:10.1038/s42004-019-0184-6

38. Feng D, Lei T, Lukatskaya MR, et al. Robust and conductive two-dimensional metal-organic frameworks with exceptionally high volumetric and areal capacitance. *Nat Energy*. 2018;3(1):30-36. doi:10.1038/s41560-017-0044-5

39. Song Dang, Qi-Long Zhu QX. Nanomaterials derived from metal–organic frameworks. *Nat Rev Mater*. 2018;3(17075).

40. Mehtab T, Yasin G, Arif M, et al. Metal-organic frameworks for energy storage devices: Batteries and supercapacitors. *J Energy Storage*. 2019;21(December 2018):632-646. doi:10.1016/j.est.2018.12.025

41. Moosavi SM, Nandy A, Jablonka KM, et al. Understanding the diversity of the metal-organic framework ecosystem. *Nat Commun*. 2020;11(1):1-10. doi:10.1038/s41467-020-17755-8

42. Wang C, Cheng P, Yao Y, et al. In-situ fabrication of nanoarchitectured MOF filter for water purification. *J Hazard Mater*. 2020;392(December 2019):1-7. doi:10.1016/j.jhazmat.2020.122164

43. Gulcay E, Erucar I. *Metal-Organic Frameworks for Biomedical Applications*. Elsevier B.V.; 2019. doi:10.1016/B978-0-12-817650-4.00006-1

44. Molavi H, Moghimi H, Taheri RA. Zr-Based MOFs with High Drug Loading for Adsorption Removal of Anti-Cancer Drugs: A Potential Drug Storage. *Appl Organomet Chem*. 2020;34(4):4-9. doi:10.1002/aoc.5549

45. Cui Y, Yue Y, Qian G, Chen B. Luminescent functional metal-organic frameworks. *Chem Rev*. 2012;112(2):1126-1162. doi:10.1021/cr200101d

46. Zhao C, Dai X, Yao T, et al. Ionic Exchange of Metal-Organic Frameworks to Access Single Nickel Sites for Efficient Electroreduction of CO2. *J Am Chem Soc*. 2017;139(24):8078-8081. doi:10.1021/jacs.7b02736

47. Yang Q, Xu Q, Jiang HL. Metal-organic frameworks meet metal nanoparticles: Synergistic effect for enhanced catalysis. *Chem Soc Rev*. 2017;46(15):4774-4808. doi:10.1039/c6cs00724d

48. Yang D, Gates BC. Catalysis by Metal Organic Frameworks: Perspective and Suggestions for Future Research. *ACS Catal*. 2019;9(3):1779-1798. doi:10.1021/acscatal.8b04515

49. Rodenas T, Luz I, Prieto G, et al. Metal-organic framework nanosheets in polymer composite materials for gas separation. *Nat Mater*. 2015;14(1):48-55. doi:10.1038/nmat4113

50. Bae TH, Lee JS, Qiu W, Koros WJ, Jones CW, Nair S. A high-performance gas-separation membrane containing submicrometer-sized metal-organic framework crystals. *Angew Chemie - Int Ed*. 2010;49(51):9863-9866. doi:10.1002/anie.201006141

51. Berger AH, Bhown AS. Comparing physisorption and chemisorption solid sorbents for use separating CO2 from flue gas using temperature swing adsorption. *Energy Procedia*. 2011;4:562-567. doi:10.1016/j.egypro.2011.01.089

52. Myers AL, Prausnitz JM. Thermodynamics of Mixed-Gas Adsorption. 11(1):121-127.

53. Bae YS, Farha OK, Hupp JT, Snurr RQ. Enhancement of CO2/N2 selectivity in a metal-organic framework by cavity modification. *J Mater Chem*. 2009;19(15):2131-2134. doi:10.1039/b900390h

54. Simmons JM, Wu H, Zhou W, Yildirim T. Carbon capture in metal-organic frameworks - A comparative study. *Energy Environ Sci*. 2011;4(6):2177-2185. doi:10.1039/c0ee00700e

55. Zheng B, Bai J, Duan J, Wojtas L, Zaworotko MJ. Enhanced CO 2 Binding Affinity of a High-Uptake rht -Type. *J Am Chem Soc*. 2011;133:748-751.

56. Coudert FX. The osmotic framework adsorbed solution theory: Predicting mixture coadsorption in flexible nanoporous materials. *Phys Chem Chem Phys*. 2010;12(36):10904-10913. doi:10.1039/c003434g

57. Fraux G, Boutin A, Fuchs AH, Coudert FX. On the use of the IAST method for gas separation studies in porous materials with gate-opening behavior. *Adsorption*. 2018;24(3):233-241. doi:10.1007/s10450-018-9942-5

58. Cessford NF, Seaton NA, Düren T. Evaluation of ideal adsorbed solution theory as a tool for the design of metal-organic framework materials. *Ind Eng Chem Res*. 2012;51(13):4911-4921. doi:10.1021/ie202219w

59. Howarth AJ, Liu Y, Li P, et al. Chemical, thermal and mechanical stabilities of metal-organic frameworks. *Nat Rev Mater*. 2016;1(15018):1-16. doi:10.1038/natrevmats.2015.18

60. Ding M, Cai X, Jiang HL. Improving MOF stability: Approaches and applications. *Chem Sci*. 2019;10(44):10209-10230. doi:10.1039/c9sc03916c

61. Bennett TD, Sotelo J, Tan JC, Moggach SA. Mechanical properties of zeolitic metal-organic frameworks: Mechanically flexible topologies and stabilization against structural collapse. *CrystEngComm*. 2015;17(2):286-289. doi:10.1039/c4ce02145b

62. Raganati F, Chirone R, Ammendola P. CO2 Capture by Temperature Swing Adsorption: Working Capacity As Affected by Temperature and CO2 Partial Pressure. *Ind Eng Chem Res*. 2020;59(8):3593-3605. doi:10.1021/acs.iecr.9b04901

63. Bo Peh S, Farooq S, Zhao D. A Metal-Organic Framework (MOF)-Based Temperature Swing Adsorption Cycle for Postcombustion CO2 Capture from Wet Flue Gas. *Chem Eng Sci*. 2021;250:117399. doi:10.1016/j.ces.2021.117399

64. Bosch M, Zhang M, Zhou H-C. Increasing the Stability of Metal-Organic Frameworks. *Adv Chem*. 2014;2014:1-8. doi:10.1155/2014/182327

65. Burtch NC, Jasuja H, Walton KS. Water stability and adsorption in metal-organic frameworks. *Chem Rev*. 2014;114(20):10575-10612. doi:10.1021/cr5002589

66. Canivet J, Fateeva A, Guo Y, Coasne B, Farrusseng D. Water adsorption in MOFs: Fundamentals and applications. *Chem Soc Rev*. 2014;43(16):5594-5617. doi:10.1039/c4cs00078a

67. Lee K, Howe JD, Lin LC, Smit B, Neaton JB. Small-molecule adsorption in open-site metal-organic frameworks: A systematic density functional theory study for rational design. *Chem Mater*. 2015;27(3):668-678. doi:10.1021/cm502760q

68. Colón YJ, Snurr RQ. High-throughput computational screening of metal-organic frameworks. *Chem Soc Rev*. 2014;43(16):5735-5749. doi:10.1039/c4cs00070f

69. Li M, Li D, O'Keeffe M, Yaghi OM. Topological analysis of metal-organic frameworks with polytopic linkers and/or multiple building units and the minimal transitivity principle. *Chem Rev*. 2014;114(2):1343-1370. doi:10.1021/cr400392k

70. Keeffe MO, Peskov M a, Ramsden SJ, Yaghi OM. (RCSR) Database of , and Symbols for , Crystal Nets. *Acc Chem Res*. 2008;41(12):1782-1789. http://dx.doi.org/10.1021/ar800124u

71. V. A. Blatov APS and VNS. The program package TOPOS is intended for the automation of all stages of crystal-chemical analysis. *J Appl Crystallogr*. 2000;33(1193).

72. Motherwell JAC and S. COMPACK: a program for identifying crystal structure similarity using distances. *J Appl Crystallogr*. Published online 2005.

73. J. Rohlícek, E. Skorepová MB and JC. CrystalCMP: an easy-to-use tool for fast comparison of molecular packing. *J Appl Crystallogr*. Published online 2016.

74. G. de la Flor, D. Orobengoa, E. Tasci JMP-M and MIA. Comparison of structures applying the tools available at the Bilbao Crystallographic Server. *J Appl Cryst*. Published online 2016.

75. Blatov VA, Shevchenko AP, Proserpio DM. Applied topological analysis of crystal structures with the program package topospro. *Cryst Growth Des*. 2014;14(7):3576-3586. doi:10.1021/cg500498k

76. Yan Y, Lin X, Yang S, et al. Exceptionally high H2 storage by a metal-organic polyhedral framework. *Chem Commun*. 2009;(9):1025-1027. doi:10.1039/b900013e

77. Yan Y, Telepeni I, Yang S, et al. Metal-organic polyhedral frameworks: High H2 adsorption capacities and neutron powder diffraction studies. *J Am Chem Soc*. 2010;132(12):4092-4094. doi:10.1021/ja1001407

78. Yan Y, Yang S, Blake AJ, et al. A mesoporous metal-organic framework constructed from a nanosized C 3-symmetric linker and [Cu24(isophthalate)24] cuboctahedra. *Chem Commun*. 2011;47(36):9995-9997. doi:10.1039/c1cc13170b

79. Farha OK, Yazaydin AÖ, Eryazici I, et al. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat Chem*. 2010;2(11):944-948. doi:10.1038/nchem.834

80. He Y, Li B, O'Keeffe M, Chen B. Multifunctional metal-organic frameworks constructed from meta-benzenedicarboxylate units. *Chem Soc Rev*. 2014;43(16):5618-5656. doi:10.1039/c4cs00041b

81. Zhang X, Chen Z, Liu X, et al. A historical overview of the activation and porosity of metal-organic frameworks. *Chem Soc Rev*. 2020;49(20):7406-7427. doi:10.1039/d0cs00997k

82. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. 2002;58.

83. Li A, Bueno-Perez R, Wiggin S, Fairen-Jimenez D. Enabling efficient exploration of metal-organic frameworks in the Cambridge Structural Database. *CrystEngComm*. 2020;22(43):7152-7161. doi:10.1039/d0ce00299b

84. Goldsmith J, Wong-Foy AG, Cafarella MJ, Siegel DJ. Theoretical limits of hydrogen storage in metal-organic frameworks: Opportunities and trade-offs. *Chem Mater*. 2013;25(16):3373-3382. doi:10.1021/cm401978e

85. Watanabe T, Sholl DS. Accelerating applications of metal-organic frameworks for gas adsorption and separation by computational screening of materials. *Langmuir*. 2012;28(40):14114-14128. doi:10.1021/la301915s

86. Chung YG, Haldoupis E, Bucior BJ, et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. *J Chem Eng Data*. Published online 2019. doi:10.1021/acs.jced.9b00835

87. Chung YG, Camp J, Haranczyk M, et al. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chem Mater*. 2014;26(21):6185-6192. doi:10.1021/cm502594j

88. Sturluson A, Huynh MT, Kaija AR, et al. The role of molecular modelling and simulation in the discovery and deployment of metal-organic frameworks for gas storage and separation*. *Mol Simul*. 2019;45(14-15):1082-1121. doi:10.1080/08927022.2019.1648809

89. Moghadam PZ, Li A, Wiggin SB, et al. Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. *Chem Mater*. 2017;29(7):2618-2625. doi:10.1021/acs.chemmater.7b00441

90. Moghadam PZ, Islamoglu T, Goswami S, et al. Computer-aided discovery of a metal-organic framework with superior oxygen uptake. *Nat Commun*. 2018;9(1):1-8. doi:10.1038/s41467-018-03892-8

91. Ahmed A, Liu Y, Purewal J, et al. Balancing gravimetric and volumetric hydrogen density in MOFs. *Energy Environ Sci*. 2017;10(11):2459-2471. doi:10.1039/c7ee02477k

92. Coupry DE, Addicoat MA, Heine T. Extension of the Universal Force Field for Metal-Organic Frameworks. *J Chem Theory Comput*. 2016;12(10):5215-5225. doi:10.1021/acs.jctc.6b00664

93. Nazarian D, Camp JS, Chung YG, Snurr RQ, Sholl DS. Large-Scale Refinement of Metal-Organic Framework Structures Using Density Functional Theory. *Chem Mater*. Published online 2017. doi:10.1021/acs.chemmater.6b04226

94. Altintas C, Avci G, Daglar H, et al. An extensive comparative analysis of two MOF databases: high-throughput screening of computation-ready MOFs for CH4 and H2 adsorption. *J Mater Chem A*. 2019;7(16):9593-9608. doi:10.1039/c9ta01378d

95. Lin LC, Berger AH, Martin RL, et al. In silico screening of carbon-capture materials. *Nat Mater*. 2012;11(7):633-641. doi:10.1038/nmat3336

96. Martin RL, Haranczyk M. Construction and characterization of structure models of crystalline porous polymers. *Cryst Growth Des*. 2014;14(5):2431-2440. doi:10.1021/cg500158c

97. Wilmer CE, Leaf M, Lee CY, et al. Large-scale screening of hypothetical metal-organic frameworks. *Nat Chem*. 2012;4(2):83-89. doi:10.1038/nchem.1192

98. Majumdar S, Moosavi SM, Jablonka KM, Ongari D, Smit B. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl Mater Interfaces*. Published online 2021. doi:10.1021/acsami.1c16220

2 Methods

This chapter will describe the various computational tools and methods that are frequently used when generating gas adsorption data. These include Grand Canonical Monte Carlo (GCMC) simulations, Density Functional Theory (DFT), the REPEAT method, as well as basic machine learning (ML) theory.

2.1 Monte Carlo simulations

One of the most commonly used tools when evaluating gas adsorption properties of a material are Monte Carlo (MC) simulations.[1–3] Briefly, MC calculations for gas adsorption relate microscopic energy interactions between guest molecules (CO_2, N_2, etc.) and the material (MOF) to macroscopic observed values such as gas uptake. The idea is to observe a guest particle interacting with a framework where the temperature and volume remain constant. By changing the position and momentum of the guest particle, a new configuration of the system is described and repeating this process many times will sample what is known as an *ensemble*. A *canonical ensemble* is therefore a collection of configurations where the volume and temperature stay constant.[4]

To obtain a reasonably statistically accurate ensemble, the energy of many configurations would need to be calculated, resulting in significant computation time. Instead, it is more practical to introduce the Monte Carlo algorithm, which will randomly perturb (rotate or translate) the guest particle from one position to another and either accept or reject a perturbation based on a set of rules proposed by Metropolis et al.[5] The rules are as follows:

1. Accept a perturbation if the resulting energy state is lower than the previous one ($E_{new} < E_{old}$).

2. If a perturbation results in a higher energy configuration, the move can still be accepted if it falls under a probability proportional given by $exp\ (-\frac{\Delta E}{k_b T})$, which describes a Boltzmann weighted change in energy of a random move, where k_bT is the thermal energy of the system and ΔE is the difference in energy between the configurations.
3. If rules 1 or 2 are not met, reject the perturbation, and count the previous configuration again.

The reason for including rule 2 is to allow certain perturbations, that may seem unfavorable, to be explored. As such, if an energy configuration is slightly higher than the previous one, the difference in energy will be low resulting in a higher probability of being accepted. Rule 3 will count the old configuration multiple times if the new perturbation is rejected, thus promoting a favorable lower state energy which will contribute more heavily to the statistical average. Therefore, this algorithm generates configurations that are Boltzmann distributed in terms of their energies, which is what is observed in nature for a system of fixed volume, temperature, and molecules.

2.2 Grand Canonical Monte Carlo Simulations

To determine how much gas a certain material will adsorb, more appropriate perturbations are required. In the Grand Canonical Monte Carlo (GCMC) simulation, two new types of perturbations are introduced: creation and deletion. Their acceptance criteria, are shown in equations 2.1 and 2.2 respectively

$$A_{crt} = \frac{V_p}{kT(N+1)} e^{-\Delta U/kT} \tag{2.1}$$

$$A_{del} = \frac{N}{kTV_p} e^{-\Delta U/kT} \tag{2.2}$$

Where volume is *V*, pressure *p*, Boltzmann constant *k*, temperature (*T*), number of guests (*N*) and ΔU is the potential energy of the system. These two types of perturbations are useful in modeling the physical process of gas entering or leaving the adsorbent. Using these rules and perturbations, a typical GCMC simulation still requires 10^7 moves to produce statistically sound thermodynamic averages. This means that generally, high level *ab initio* methods are too computationally demanding and that empirical methods (discussed later) are usually preferred. To further speed up GCMC calculations, two approximations are commonly used 1) the MOF framework is fixed and 2) intermolecular potentials between the framework and guest are simplified. The first approximation is introduced by fixing the MOF atoms and preventing them from moving while the guest molecules can move but have fixed bonds. While fixing the structure can cause issues for flexible "breathing" MOFs[7,8], the results generated from this method are generally accurate.[9,10] The second approximation simplifies the guest-host interactions into 2-body steric, dispersion and non-polarizable electrostatic potentials.

2.2.1 Periodic Boundary Conditions

When computing adsorption properties on crystalline materials, periodic boundary conditions (PBCs) are commonly used to approximate the infinite periodic nature of these materials. This is done by describing the system as a transitionally invariant set of atoms, called *unit cells* or *primary cells*, shown in figure 2.1 a). These unit cells are then copied and surrounded by *images* of itself in all directions, as shown in figure 2.1b), which describe all surrounding interactions of the atoms in the original unit cell. Applying an appropriate unit cell size is important, if the cell is too small it is unlikely that all pairwise interactions will be included and if the cell is too large, the computational cost will increase exponentially. Inside the cells, the set of atoms are described as transitionally invariant such that a particle can exit on one

side of the cell and re-appear on the opposite side, as shown in figure 2.1 b), which preserves any properties (forces, velocities, etc.) of the particle. While this makes the computation of an infinite system feasible, it does introduce the issue of particles interacting with multiple images of other particles. To deal with this problem, a cutoff, shown in figure 2.1 c), is introduced, where only the interactions between atoms inside the area are computed. The cutoff is only applied to short-ranged interactions (steric and dispersion) and not long-ranged electrostatic interactions.

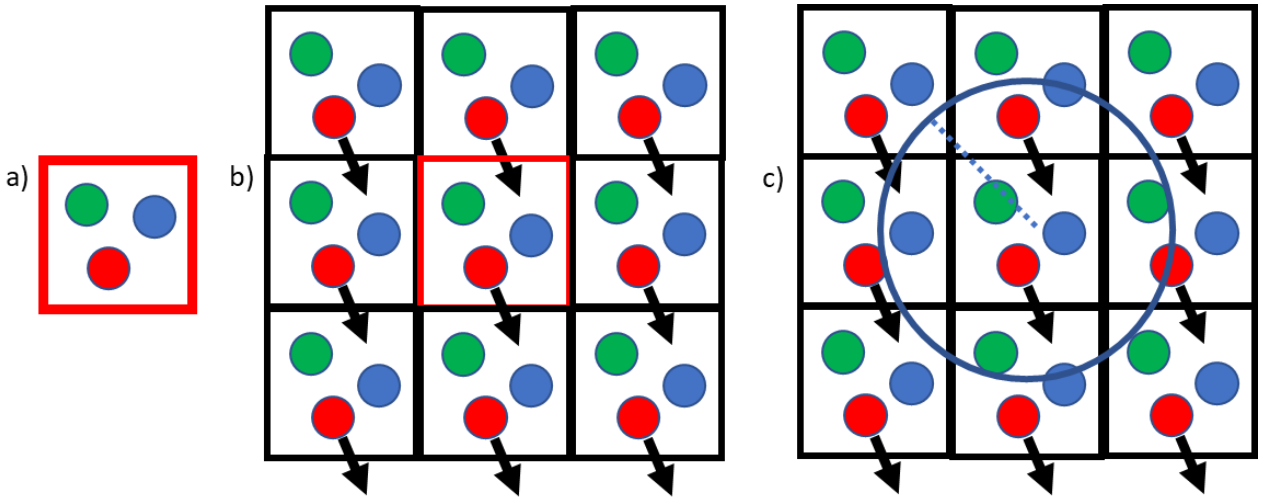


Figure 2.1 Periodic Boundary Conditions. a) a primary cell b) primary cells and copies with transitionally invariant atoms c) radius cutoff showing area of computation

In a GCMC simulation, short range steric and dispersion interactions are calculated separately from long-range Coulombic interactions. Many suitable potentials can be used to represent short-range interactions, such as the Buckingham potential[11] or the Born-Hugging-Meyer potential[12]. Here, the Lennard-Jones, also called 12-6 or LJ, potential will be described as it is one of the most commonly used models in GCMC simulations[13]. For long-range interactions, special considerations are required, and they will be discussed in the next sections.

2.3 Lennard-Jones potential

The Lennard-Jones (LJ) potential approximates the short-range interactions according to equation 2.3

$$V_{LJ}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.3}$$

where the distance scalar, σ_{ij}, and the potential well, ε_{ij}, are parameters assigned for interactions between atoms *i* and *j* and r_{ij} is the distance between atoms *i* and *j*. This function decays rapidly at a rate of r_{ij}^{-6} as the atoms move away from each other, which allows for a cutoff to be used without introducing significant error. In contrast, the potential which describes the long-range, or Coulombic, interactions decay much slower, at a rate of r_{ij}^{-1},as shown in equation 2.4, where q represents the charge of the atoms.

$$V_C = \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \tag{2.4}$$

Equations 2.3 and 2.4 are shown graphically in figure 2.2.

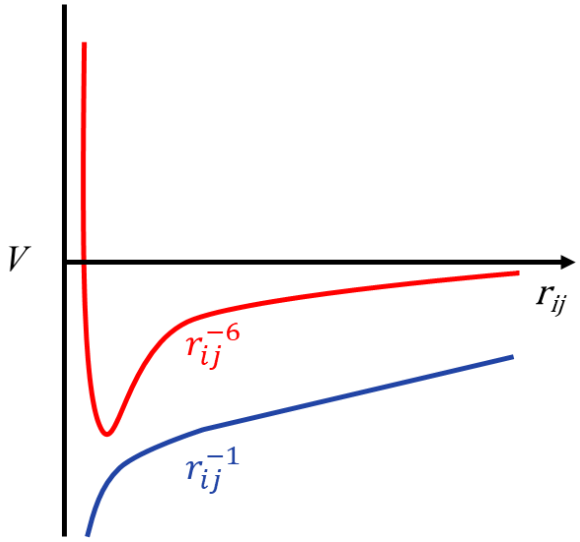


Figure 2.2 Energy potential of (top) LJ interactions and (bottom) Coulombic interactions between two atoms.

For the interactions between the atoms in the primary cell and all the other cells, the Coulomb energy is given by equation 2.5, where *N* is the number of atoms, the $'$ on the first sum indicates that the terms $i = j$ are omitted when n = 0 (i.e., the original unit cell), and the ½ factor accounts for double counting.

$$\mathrm{V} = \frac{1}{2}\sum_{n}{}' \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{|r_i - r_j| + n} \qquad (2.5)$$

Equation 2.5 converges conditionally and slowly, and therefore computing all relevant Coulombic interactions would require introducing many periodic image cells (*n*), thus drastically increasing the computational cost. To circumvent this issue, it is more appropriate to use the Ewald Summation to compute long-range interactions in periodic systems.[14] However, electrostatic interactions between the MOF and guest are approximated using the fixed partial charge approximation, which generates partial charges that are independent of conformation and environment.

2.4 Atomic Point Charges

The performance of MOFs can be assessed using simulations, which require a set of point charges, $\{q\}$, as one of its inputs. These charges would ideally be obtained from high level quantum calculations, e.g., Density Functional Theory (DFT) which yield highly accurate point charges. However for MOFs, which generally have hundreds of atoms per unit cells, high level quantum calculations can take hours to compute, resulting in enormous computational time when screening thousands of candidates.[15] As such, when conducting large-scale screenings, it is generally more suitable to use cheaper, approximative methods to obtain point charges, allowing for fast and reasonably accurate computation for a variety of structures. One of the most common of these methods is charge equilibrium (Qeq), which performs quite well in predicting partial charges in MOFs, with the original Qeq method being developed in 1991 by Rappé and Goddard.[16]

2.5 Charge Equilibration Methods

While the full explanation of the charge equilibration method is beyond the scope of this work, a summary is still discussed. The Qeq method defines the energy of an atom as a function of its charge as the first two terms of a Taylor series around its neutral state.

$$E_A = E_{A0} + Q_A \left(\frac{\partial E}{\partial Q}\right)_{A0} + \frac{1}{2} Q_A^2 \left(\frac{\partial^2 E}{\partial Q^2}\right)_{A0} \tag{2.6}$$

This equation can also be written as:

$$E_A = E_{A0} + \chi_A^0 Q_A + \frac{1}{2} J_{AA}^0 Q_A^2 \tag{2.7}$$

Where the electronegativity, χ_A^0, and the idempotential, J_{AA}^0, are the adjustable parameters of the model. By summing the energy of each atom as well as the contribution of the interatomic interactions, the charge-dependent energy of a set of atoms is obtained:

$$E_Q(Q_1 \dots Q_N) = \sum_A (E_{A0} + \chi_A^0 Q_A) + \frac{1}{2}\sum_{A,B} Q_A Q_B J_{AB} \qquad (2.8)$$

The next step is to find a solution where the first derivative of this expression with respect to the charge Q_A is equal for all atoms at equilibrium, hence the name *charge equilibration* method.

$$\frac{\partial E}{\partial Q_A} = \chi_A^0 + J_{AA}^0 + \sum_{B \neq A} Q_B J_{AB} \qquad (2.9)$$

By setting this equal for the *N* atoms in a system, $N - 1$ equations are obtained and when combined with the total charge equation, $Q_{\text{tot}} = \sum_{i=1}^{N} Q_i$, a system with *N* equations and *N* unknowns is obtained, which can be solved.

Charge equilibration methods can calculate partial atomic charges for large structures such as MOFs in seconds but the high parameterization from the electronegativity and idempotential makes them unfeasible in certain chemical environments, thus limiting its applicability.[17] For example, Q_{eq} would generate nonphysical partial charges of +3 for Li atoms in Ag_5Li_5 clusters or give imprecise energy for small interatomic distances.[18,19]

Since then, several modifications and improvements have been made to the Q_{eq} method to increase its accuracy in various environments and for a wide variety of applications. [17,20,21] These methods will be discussed in the next sections.

2.5.1 Periodic Charge Equilibration

Periodic Qeq (PQeq) [22] is a variation of Qeq that is applicable to periodic structures by accounting for the effects of the crystal lattice energy on charge distribution, which is done, in part, by using an Ewald summation. The effect of the lattice is incorporated in the expression for the chemical potential, equation 2.15. This gives

$$\mu_{A,periodic}(Q_1 \dots Q_N) = \chi_A^0 + \chi_{c,A}^0 + J_{AA}^0 Q_A + \sum_{B \neq A} Q_B J_{AB} \qquad (2.10)$$

Where $\chi_{c,A}^0$, is the lattice electronegativity given by

$$\chi_{c,A}^0 = \sum_{L \neq 0} \sum_{j}^{n} J'_{ij,L} + \sum_{L \neq 0} \sum_{j}^{n} \frac{Q_A}{R_{ij,L}} \qquad (2.11)$$

Where the cloud penetration, $J'_{ij,L} = J_{ij,L} - \frac{1}{R}$ converges rapidly with respect to *R*. The index *L* refers to the periodic image cell, where *L* = 0 is the primary cell. The $Q_A / R_{ij,L}$ term is a direct Coulombic term which is calculated using an Ewald summation.

One of the main advantages of PQeq is the user only needs lattice parameters and atomic coordinates of a single cell. Furthermore, *a priori* guess of the equilibrium charges as well as choosing appropriate cutoffs and boundaries for each calculation was not needed, unlike many methods at the time. [24] PQeq was first used in large scale screening by Haldoupis to assign charges in 500 MOFs for CO_2 and N_2 adsorption simulation.[23] The results were reasonable when compared to DFT values but more importantly very little computational cost was required, making this a promising method for rapid screening of many structures.

2.5.2 Extended Charge Equilibration

Wilmer et al. [20] developed an extended charge equilibration method, coined EQeq, which would require fewer fitting parameters than the Qeq and PQeq methods and would therefore converge much faster while yielding similar accuracy. To demonstrate this, Wilmer et al.[25] used their EQeq method to screen a database of 137,000 hypothetical MOFs. While their findings emphasized several important connections between physical (surface area, pore size and pore volume) characteristics and adsorption capabilities in MOFs, there were considerable issues with the method and the study. The method would treat metal cations differently than other atoms as the original Qeq method was known to yield poor results in ionic bonded materials. Therefore, EQeq ensured accurate representations of metal cations but this also resulted in a background positive charge in the unit cell. The background charge was then minimized through an empirically determined dielectric constant. As mentioned by the authors in their work on screening the 137,000 of hypothetical MOFs, the group chose a large dielectric constant of 2, which resulted in an overall underestimation of CO_2 adsorption. [25,26] Furthermore, while considerable efforts were put in place to minimize the effects of cations, the halogen parameters were not changed, which resulted in significant overestimation in fluorine and chlorine - functionalized MOFs.[21]

Lastly, and most importantly, the structures did not have diverse enough underlying network topologies, meaning that trends they observed could only be confirmed for a specific subset of MOFs that shared this topology.

2.5.3 MEPO-Qeq

The MOF electrostatic-potential-optimized-charge (MEPO-Qeq)[21] method is based on the PQeq method implemented in the GULP[27] package. In the MEPO scheme, the atomic values χ_A^0

and J_{AA}^{0} are fitted to reproduce DFT/PBE[28] electrostatic potentials (ESP). This was done by taking a training set of 543 hypothetical MOFs and performing DFT calculations to produce ESPs. The standard Qeq parameters, shown in table 2.1, were then adjusted to reproduce these DFT-derived ESPs in a least square manner. These adjusted parameters, referred to as MEPO-Qeq parameters, shown in table 2.1, as well as QEq and EQeq parameters, were then used to generate CO_2 uptakes and heats of adsorption (HoA) in the 543 MOFs. The uptakes and HoAs were compared with results from DFT-derived charges, and it was found that while MEPO-Qeq underestimated CO_2 uptakes, it yielded closer results to DFT values than the other methods.

Table 2.1 Electronegativities and electrostatic potential parameters used for Qeq and MEPO-Qeq

Element	Qeq paramters		MEPO-Qeq	
	χ (eV)	1/2J (eV)	χ (eV)	1/2J (eV)
H	4.528	6.945	4.528	6.945
V	3.650	3.410	4.093	4.217
Cu	4.200	4.220	5.429	3.468
Zn	5.106	4.285	3.701	4.468
C	5.343	5.063	5.431	5.857
N	6.899	5.880	6.688	6.622
O	8.741	6.682	8.714	8.568

A validation set of 693 MOFs was used to investigate the robustness of MEPO-Qeq. Using the same process as for the training set, CO_2 uptakes and HOAs were generated using Qeq, EQeq, and MEPO-Qeq parameters and compared to DFT derived results. The MEPO-Qeq performed well and better than the other methods for this set as well. To further test the transferability of the method, the CO_2 uptakes of 11 Zn-zeolitic imidozolate (ZIFs) and 16 Vanadium MOFs, neither of which were part of the training set, were calculated. Compared to

DFT data, the ZIF CO_2 uptakes were reasonably close, with a Pearson correlation of 0.96 and an RMS error of 0.16. As for the vanadium MOF set, MEPO did not perform well with a Pearson correlation of 0.92 and an RMS error of 0.86. The is most likely due the presence of open-metal site vanadium centers, which were not included in the training set. However, given a set of structures with reasonably similar environments as the training set, the MEPO parameters have been successfully used in various studies to generate gas adsorption data in crystalline structures.[29,30] Most notably, they were used by Boyd et al. to generate partial atomic charges in over 300,000 hypothetical MOFs to then compute CO_2 working capacity and CO_2/N_2 selectivity.[31]

2.5.4 Split-Charge Equilibration

The split-charge method (SQE), developed by Mueser and coworkers,[32] is a relatively simple and robust technique that has been widely used for determining partial charges in small nonmetallic molecules[33] and simple periodic systems.[34] In the SQE method, an energy expression is defined in terms of split-charges, q_{ij}, that are associated with each covalent bond to then generate partial atomic charges, Q_i.

$$Q_i = \sum_j q_{ij} \tag{2.12}$$

The calculation for Qeq charges is obtained by minimizing equation 2.17 by adjusting each partial atomic charge ,Q_i, in the set of all partial atomic charges $\boldsymbol{Q}$.

$$E_{QEq}(\boldsymbol{Q}) \sum_i \left(\chi_i Q_i \frac{1}{2}\kappa_i Q_i^2\right) + \sum_{i,j>1} Q_i Q_j J_{ij}(r_{ij}) \tag{2.13}$$

Where χ is the atomic electronegativity, κ is the atomic hardness and J is the distance-dependent electrostatic potential.

The SQE energy expression introduces bond hardness, κ_{ij}^{b}, and bond electronegativity, χ_{ij}^{b} for each bond type. SQE charges are then obtained by minimizing equation 2.20 over the set of all split-charges, $\boldsymbol{q}$.

$$E_{SQE}(\boldsymbol{q}) = \sum_{i,j>1} \left(\chi_{ij}^{b} q_{ij} + \frac{1}{2} \kappa_{ij}^{b} q_{ij}^{2} \right) + E_{Qeq}(\boldsymbol{Q}) \qquad (2.14)$$

The additional bonding information obtained from the split-charges allows for accurate prediction of bond dissociation limits, particularly in covalent systems. This is not possible with the main Qeq model. Also, the two additional bond type parameters result in greater number of degrees of freedom during the charges fitting procedure, making this model more flexible. Lastly, when compared to Qeq, the SQE model allows for better modeling of the dielectric properties and polarizability of the system.[33,35] This is mainly due to SQE being based on a more accurate physical description of the charge distribution. The main drawback of this model is that it is only applicable to charge neutral systems, due to its *charge conservation* property, where $q_{ij} = -q_{ji}$. However, SQE was later reformulated by Krykunov and coworkers to treat non-neutral systems and was found to perform better in charged systems than basic Qeq models.[36] Collins et al.[37] reparametrized the values for SQE analogously to MEPO-Qeq and used a diverse training set of 559 and a validation set of 585 hypothetical MOFs. Due to the addition of bonding information, 91 parameters were fitted, where only 34 parameters would be sufficient for a Qeq method.

2.6 Density Function Theory

Density Functional Theory (DFT)[38,39] is a quantum mechanical (QM) method that has been widely used in chemistry and physics to calculate the electronic structure of atoms, molecules, and solids since the 1970's and is still one of the most popular electronic structure method.[40] Briefly, the goal for QM calculations is to find the ground state energy for a collection of atoms by solving the Schrödinger equation:

$$\hat{\mathrm{H}}\psi(\{r_i\},\{R_I\}) = E\psi(\{r_i\},\{R_I\}) \tag{2.15}$$

Where Ψ is a wavefunction which describes a system of nuclei, R_I, and electrons, r_i, with Ĥ being an energy operator that is applied to the wavefunction. Naturally, given any system with more than a few particles, this equation becomes impossible to solve. This is known as the many particle problem. As such, the *Born-Oppenheimer* approximation is used, which separates the nucleic motion from the electron kinetic energy. In most cases this approximation is reliable as the nuclei are essentially static when observed at an electronic time scale. This allows the decoupling of the nuclear and electronic wave function:

$$\psi(\{r_i\},\{R_I\}) \rightarrow \psi_N(\{R_I\}) * \psi_e(\{r_i\}) \tag{2.16}$$

To solve the Schrödinger equation for electrons:

$$\hat{\mathrm{H}}\psi_e(r_1, r_2 \ldots r_N) = E\psi_e(r_1, r_2 \ldots r_N) \tag{2.17}$$

However, even with the Born-Oppenheimer approximation, this equation is still a 3N dimensional problem, where N is the number of electrons. As such, solving the Schrödinger for any practical material is still unfeasible. DFT offers an accurate and computationally feasible way to solve the equivalent of the Schrödinger equation in terms of the *electron density*:

$$n(\boldsymbol{r}) = \psi^*(r_1, r_2 \ldots r_N)\psi(r_1, r_2 \ldots r_N) \tag{2.18}$$

Where ψ^* denotes the complex conjugate of the wavefunction. This reduces the dimensions of the problem from 3N to 3. The next step is to treat each electron as a point charge in the field of all other electrons, which simplifies the many-electron problem into many *one* electron problems:

$$\psi(r_1, r_2 \ldots r_N) = \psi_1(r_1) * \psi_2(r_2) * \ldots * \psi_N(r_N) \tag{2.19}$$

The electron density can then be defined as a sum of the individual electron wavefunctions:

$$n(\boldsymbol{r}) = 2\sum_i \psi^*(r)\psi(r) \tag{2.20}$$

The two fundamental theorems of DFT, proposed by Hohenberg and Kohn[41] , are:

1) The ground state energy *E* is a unique functional of the electron density.

$$E = E[n(r)] \tag{2.21}$$

2) The electron density that minimizes the energy of the overall functional is the true ground state of electron density.

$$E[n(r)] > E_0[n_0(r)] \tag{2.22}$$

The first theorem introduces a *functional* that only requires the electron density to compute the ground state energy of a system. The second theorem states that by minimizing the energy functional, the ground state energy density is obtained. The exact form of the energy functional, generally referred to as the *exchange-correlation* (XC) functional, is not known approximations must be made. The most basic approximation is the local density approximation (LDA), which depends only on the value of the electron density at each point in space. There exist more complex functionals specifically designed to obtain better description of a desired property, such as nuclear magnetic resonance chemical shift[42], formation energy[43], ionization

potential[44] and many others. As such, there is an estimated 500 functionals, however only a handful of them are commonly used, such as Perdew, Burke, and Ernzerhof [45] (PBE)[28] and Becke, 3-parameter, Lee-Yang-Parr (B3LYP) functionals [46,47]

Practically, the ground state electron density can be calculated using approximative methods developed by Kohn and Sham.[39] Briefly, the Kohn-Sham DFT (KS-DFT) scheme proposes a fictitious system of *non-interacting* electrons with the same density as that of one with interacting electrons. This drastically simplifies the calculation as the kinetic energy of non-interacting electrons is known. The electron density is then computed using a *self-consistency* (iteratirative) scheme based on the single, non-interacting electron wavefunctions. The KS-DFT scheme is by far the most successful way to implement DFT and is often simply called DFT.

Often, a system can be further simplified by *freezing* ion cores, which means that calculations are done under the assumption that inner shell electrons are not involved in chemical or physical interactions in solids, this is known as the *frozen-core* approximation. [49] The complex effects of the motion of the core electrons on a valance system can therefore be replaced by a *pseudopotential* (also called an effect potential), which greatly reduces the numerical effort required to describe the interaction

In this thesis, variation of DFT for periodic systems, called *periodic* DFT, is used to optimize the framework of the crystalline structures. Periodic DFT facilitates the computation of large and periodic structures by introducing periodic boundary conditions (PBCs)[48], which are discussed in section 2.1.1.. In this work, periodic DFT is used primarily to generate electrostatic potentials used by the REPEAT method (section 2.7) to generate partial atomic charges. Periodic DFT will also be used to optimize the 325,000 theoretical MOFs using the Vienna Ab Initio Simulation (VASP)[50–53] package.

2.7 REPEAT

To compute the electrostatic interactions between the MOF and the guest molecules, the atomic charges must be calculated, as shown in Eq. 2.13. The Repeating Electrostatic Potential Extracted Atomic (REPEAT) method, generated by Campañá et al.[54] , generates electrostatic potential (ESP) fitted partial atomic charges from quantum mechanical (QM) calculations in crystalline solids. While the full explanation is beyond the scope of this work, a summary is described. To calculate the REPEAT charges, a periodic DFT calculation is performed to generate a QM ESP. REPEAT then employs a least-squares fitting method to fit the partial atomic charges to best represent the QM ESP. The main distinction between REPEAT and other methods that generate QM derived partial atomic charges is its ability to be used for crystalline systems, such as MOFs. In an infinite periodic system, the reference point of the ESP is ill-defined, meaning that the resulting potentials will be shifted by a constant offset amount. REPEAT addresses this issue by using an error functional which minimizes the relative difference between the QM ESP and the ESP from point charges.

While REPEAT has been successfully used in various studies,[26,55–57] the computational cost of performing QM calculations to obtain ESP for each material, especially crystalline structures, is expensive.

2.8 Machine Learning in MOF Discovery

Machine learning (ML) methods have yielded ground-breaking results in various fields in recent years and materials science is no exception. While the theory of ML is too broad for this work, a general overview of how it can be used in MOF discovery is discussed. Recently, ML algorithms have been developed to relate chemical and/or structural features , also known as *descriptors*, to various gas adsorption and storage properties with astonishing results. [58–60] A

generic ML algorithm learns a function, *f*, that predicts a desired property, such as working capacity or selectivity, by using existing data. The goal is for the function, or *model*, to accurately predict the desired property for a new structure, never seen before, represented by the easy to calculate descriptors.

2.8.1 Descriptors

When constructing a ML model one of the most important and challenging part of the process is choosing the appropriate descriptors for a particular property. For example, Anderson et al.[63] described the relative importance of commonly used descriptors in ML MOF working capacities (WC) and selectivity (S) and adsorption loading (N) for H_2, CO_2 and N_2 mixtures. Some descriptors like SE of FG (sum of epsilons of functional group, where the epsilons are obtained from LJ values) and MPC of FG (most positive charge of functional group) are relatively unimportant, while LPD (lowest pore diameter) and topology are highly important in most cases.

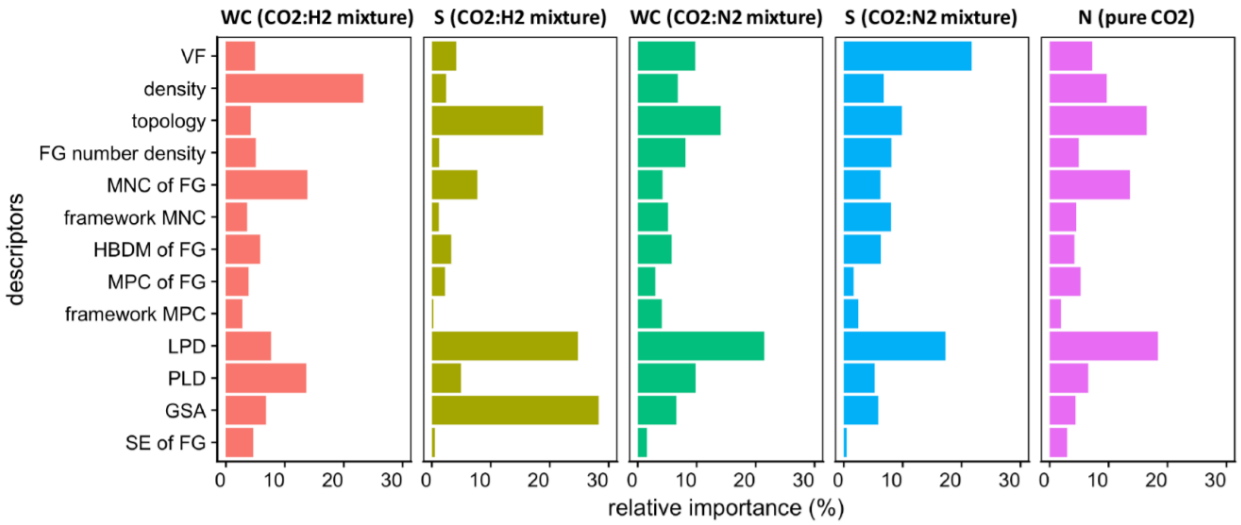


Figure 2.3 Relative importance of various descriptors in gas adsorption calculations. Reproduced from [63]

This was also demonstrated by Ward et al. [61] who identified 145 potential descriptor for materials research. The selection of descriptors is especially complicated for MOFs due to their vast chemical and structural diversity. [62] Some of the more commonly used descriptors for MOFs and their respective property are shown in figure 2.4.

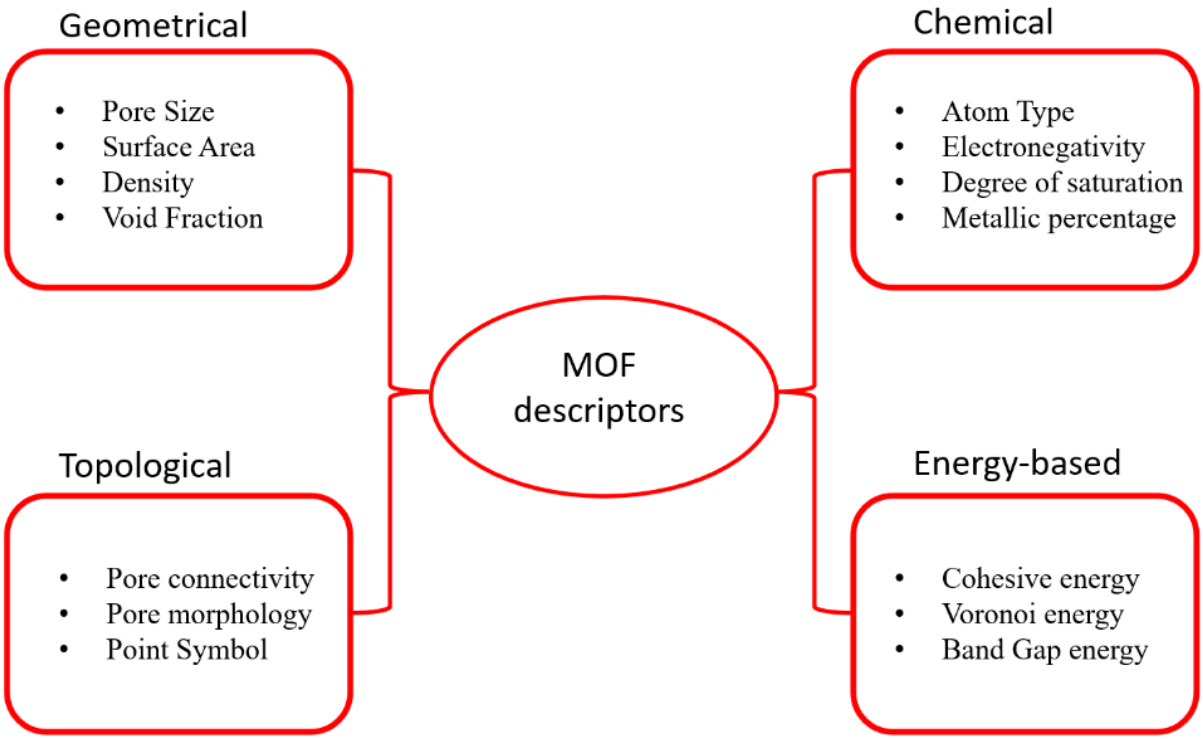


Figure 2.4 Common descriptors used in ML models for various applications.

The importance of descriptor choice is further highlighted by the work done by Fernandez et al.[2], who were the first to use ML models for prediction of methane storage performance in MOFs at high pressure (100 bar). They did this by generating a quantitative structure-property relationship (QSPR)[64,65] model using geometrical features only, such as pore size, surface area, and void fraction. While their model was accurate ($r^2 = 0.93$) for predicting methane uptake capacity at high pressure, prediction for low pressure (1bar) was relatively poor ($R^2 = 0.72$). The group later introduced an atomic property weighted radial distribution function

(AP-RDF)[66] descriptor into their model that could capture both chemical and geometric features and found noticeable improvement ($r^2 = 0.83$) in prediction accuracy for low pressure gas adsorption. The AP-RDF is based on the radial distribution function (RDF) which was first proposed by Gasteiger et al. [67,68] to encode the 3D structure of molecules. The AP-RDF, shown in equation 2.23, can be described as the weighted probability distribution of finding a pair of atoms in a volume of radius R

$$RDF^{P}(R) = f \sum_{i,j} P_i P_j e^{-B\left(r_{ij}-R\right)^2} \qquad (2.23)$$

Where B is a smoothing parameter, r_{ij} is the minimum image convention distance of atom pairs and f is a scaling factor. P_i and P_j are atomic properties, namely electronegativity, polarizability and van de Waals volume. As such, the AP-RDF describes both geometrical and chemical features in a structure. Several variations of the descriptor are still being used years later in MOF ML research.[58,69,70]

2.8.2 Input vs output data

For property prediction there are two 'types' of data, *input*, and *output*. The input data refers to the chemical, geometrical and/or energy-based descriptors which are easily accessible or computationally inexpensive. [71,72]. Output data refers to the property that is being investigated, which is often computationally expensive to generate, especially when aiming for QM levels for accuracy. There are therefore two main strategies for obtaining output data, either by large-scale screening or by obtaining output data from the literature. Several groups used large-scale screening to generate their own data but this is often too computationally expensive, especially for DFT calculations.[73,74] Generally in ML, having a larger and more diverse dataset will result in a more robust algorithm as it has been trained on a wider variety of data. As such,

many researchers have used results from large-scale screening studies to use as data points and have also made use of various data mining techniques to obtain information from different publications.[75–77]. Recently, Rosen et al. [78] proposed a database, called Quantum MOF (QMOF), which consists of quantum mechanical properties (energy charge density, band gap and density of states) containing over 14,000 experimentally synthesized crystalline structures. This "ML-ready" database could provide researchers the tools needed to develop their own ML models and accelerate MOF discovery. The Boyd-Woo database, which contains over 300,000 structures is another tool which could be used for developing machine learning models, although the structures are computationally generated.

2.9 References

1. Kaija AR, Wilmer CE. High-Pressure Methane Adsorption in Porous Lennard-Jones Crystals. *J Phys Chem Lett*. 2018;9(15):4275-4281. doi:10.1021/acs.jpclett.8b01421

2. Fernandez M, Woo TK, Wilmer CE, Snurr RQ. Large-scale quantitative structure-property relationship (QSPR) analysis of methane storage in metal-organic frameworks. *J Phys Chem C*. 2013;117(15):7681-7689. doi:10.1021/jp4006422

3. Snurr RQ, Bell AT, Theodorou DN. Prediction of Adsorption of Aromatic Hydrocarbons in Silicalite from Grand Canonical Monte Carlo Simulations with Biased Insertions. *J Phys Chem*. 1993;97(51):13742-13752.

4. Hansen J-P, McDonald IR. *Theory of Simple Liquids, Third Edition*.; 2006.

5. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21(6):1087-1092. doi:10.1063/1.1699114

6. Alavi S. Statistical Mechanics: Theory and Molecular Simulation. By Mark E. Tuckerman. *Angew Chemie Int Ed*. 2011;50(51):12138-12139. doi:10.1002/anie.201105752

7. Rogge SMJ, Goeminne R, Demuynck R, et al. Modeling Gas Adsorption in Flexible Metal–Organic Frameworks via Hybrid Monte Carlo/Molecular Dynamics Schemes. *Adv Theory Simulations*. 2019;2(4):1-15. doi:10.1002/adts.201800177

8. Chen L, Morrison CA, Düren T. Improving predictions of gas adsorption in metal-organic frameworks with coordinatively unsaturated metal sites: Model potentials, ab initio parameterization, and gcmc simulations. *J Phys Chem C*. 2012;116(35):18899-18909. doi:10.1021/jp3062527

9. Basdogan Y, Keskin S. Simulation and modelling of MOFs for hydrogen storage. *CrystEngComm*. 2015;17(2):261-275. doi:10.1039/c4ce01711k

10. Daglar H, Keskin S. Computational Screening of Metal-Organic Frameworks for Membrane-Based CO2/N2/H2O Separations: Best Materials for Flue Gas Separation. *J Phys Chem C*. 2018;122(30):17347-17357. doi:10.1021/acs.jpcc.8b05416

11. Buckingham RA. The classical equation of state of gaseous helium , neon and argon. *Proc R Soc London Ser A Math Phys Sci*. 1938;168(933):264-283.

12. Huggins ML, Mayer JE. Interatomic distances in crystals of the alkali halides. *J Chem Phys*. 1933;1(9):643-646. doi:10.1063/1.1749344

13. Jones JE, Ingham AE. On the Calculation o f certain Crystal Potential the Cubic Crystal o f Least Potential Energy. *Proc R Soc A*. 1925;107(744):636-653.

14. Ewald P. Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. *Ann Phys*. 1921;369(3):253-287.

15. Ongari D, Boyd PG, Kadioglu O, MacE AK, Keskin S, Smit B. Evaluating Charge Equilibration Methods to Generate Electrostatic Fields in Nanoporous Materials. *J Chem Theory Comput*. Published online 2019. doi:10.1021/acs.jctc.8b00669

16. Rappe AK, Iii WAG, Qa E. Charge Equilibration for Molecular Dynamics Simulations. *J Phys Chem*. 1991;(8340):3358-3363.

17. Ogawa T, Kurita N, Sekino H, Kitao O, Tanaka S. Consistent charge equilibration (CQEq) method: Application to amino acids and crambin protein. *Chem Phys Lett*. 2004;397(4-6):382-387. doi:10.1016/j.cplett.2004.09.003

18. Zhang M, Fournier R. Self-Consistent charge equilibration method and its application to Au 13Na n (n = 1,10) Clusters. *J Phys Chem A*. 2009;113(13):3162-3170. doi:10.1021/jp8063273

19. Oda A, Hirono S. Geometry-dependent atomic charge calculations using charge equilibration method with empirical two-center Coulombic terms. *J Mol Struct THEOCHEM*. 2003;634(1-3):159-170. doi:10.1016/S0166-1280(03)00338-5

20. Wilmer CE, Kim KC, Snurr RQ. An extended charge equilibration method. *J Phys Chem Lett*. 2012;3(17):2506-2511. doi:10.1021/jz3008485

21. Kadantsev ES, Boyd PG, Daff TD, Woo TK. Fast and accurate electrostatics in metal organic frameworks with a robust charge equilibration parameterization for high-throughput virtual screening of gas adsorption. *J Phys Chem Lett*. Published online 2013. doi:10.1021/jz401479k

22. Ramachandran S, Lenz TG, Skiff WM, Rappé AK. Toward an understanding of zeolite Y as a cracking catalyst with the use of periodic charge equilibration. *J Phys Chem*. Published online 1996. doi:10.1021/jp952864q

23. Haldoupis E, Nair S, Sholl DS. Finding MOFs for highly selective CO 2/N 2 adsorption using materials screening based on efficient assignment of atomic point charges. *J Am Chem Soc*. 2012;134(9):4313-4323. doi:10.1021/ja2108239

24. Wilmer CE, Snurr RQ. Towards rapid computational screening of metal-organic frameworks for carbon dioxide capture: Calculation of framework charges via charge equilibration. *Chem Eng J*. 2011;171(3):775-781. doi:10.1016/j.cej.2010.10.035

25. Wilmer CE, Farha OK, Bae YS, Hupp JT, Snurr RQ. Structure-property relationships of porous materials for carbon dioxide separation and capture. *Energy Environ Sci*. 2012;5(12):9849-9856. doi:10.1039/c2ee23201d

26. Wells BA, De Bruin-Dickason C, Chaffee AL. Charge equilibration based on atomic ionization in metal-organic frameworks. *J Phys Chem C*. 2015;119(1):456-466. doi:10.1021/jp510415h

27. Gale JD. GULP: A computer program for the symmetry-adapted simulation of solids. *J Chem Soc*. 1997;93(4):629-637.

28. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett*. 1996;77(18):3865-3868. doi:10.1103/PhysRevLett.77.3865

29. Aghaji MZ, Fernandez M, Boyd PG, Daff TD, Woo TK. Quantitative Structure–Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2Working Capacity and CO2/CH4Selectivity for Methane Purification. *Eur J Inorg Chem*. 2016;2016(27):4505-4511. doi:10.1002/ejic.201600365

30. Qiao Z, Zhang K, Jiang J. In silico screening of 4764 computation-ready, experimental metal-organic frameworks for CO2 separation. *J Mater Chem A*. 2016;4(6):2105-2114. doi:10.1039/c5ta08984k

31. Boyd PG, Chidambaram A, García-Díez E, et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*. Published online 2019. doi:10.1038/s41586-019-1798-7

32. Nistor RA, Polihronov JG, Müser MH, Mosey NJ. A generalization of the charge equilibration method for nonmetallic materials. *J Chem Phys*. 2006;125(9). doi:10.1063/1.2346671

33. Verstraelen T, Bultinck P, Van Speybroeck V, Ayers PW, Van Neck D, Waroquier M. The significance of parameters in charge equilibration models. *J Chem Theory Comput*. 2011;7(6):1750-1764. doi:10.1021/ct200006e

34. Verstraelen T, Sukhomlinov S V., Van Speybroeck V, Waroquier M, Smirnov KS. Computation of charge distribution and electrostatic potential in silicates with the use of chemical potential equalization models. *J Phys Chem C*. 2012;116(1):490-504. doi:10.1021/jp210129r

35. Verstraelen T, Ayers PW, Van Speybroeck V, Waroquier M. ACKS2: Atom-condensed Kohn-Sham DFT approximated to second order. *J Chem Phys*. 2013;138(7). doi:10.1063/1.4791569

36. Krykunov M, Demone C, Lo JWH, Woo TK. A New Split Charge Equilibration Model and REPEAT Electrostatic Potential Fitted Charges for Periodic Frameworks with a Net Charge. *J Chem Theory Comput*. 2017;13(6):2858-2869. doi:10.1021/acs.jctc.6b00998

37. Collins SP, Woo TK. Split-charge equilibration parameters for generating rapid partial atomic charges in metal-organic frameworks and porous polymer networks for high-throughput screening. *J Phys Chem C*. 2017;121(1):903-910. doi:10.1021/acs.jpcc.6b10804

38. Kohn W. Nobel lecture: Electronic structure of matter - Wave functions and density functional. *Rev Mod Phys*. 1999;71(5):1253-1266. doi:10.1103/revmodphys.71.1253

39. Kohn W, Sham LJ. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys Rev*. 1965;140.

40. Van Mourik T, Bühl M, Gaigeot MP. Density functional theory across chemistry, physics and biology. *Philos Trans R Soc A Math Phys Eng Sci*. 2014;372(2011). doi:10.1098/rsta.2012.0488

41. Hohenberg P, Kohn W. Inhomogeneous Electron Gas. *Phys Rev*. 1964;136:B864. doi:10.1007/BF01198136

42. Wiitala KW, Hoye TR, Cramer CJ. Hybrid density functional methods empirically optimized for the computation of 13C and 1H chemical shifts in chloroform solution. *J Chem Theory Comput*. 2006;2(4):1085-1092. doi:10.1021/ct6001016

43. Sarmiento-Pérez R, Botti S, Marques MAL. Optimized Exchange and Correlation Semilocal Functional for the Calculation of Energies of Formation. *J Chem Theory Comput*. 2015;11(8):3844-3850. doi:10.1021/acs.jctc.5b00529

44. Jin Y, Bartlett RJ. Accurate computation of X-ray absorption spectra with ionization potential optimized global hybrid functional. *J Chem Phys*. 2018;149(6). doi:10.1063/1.5038434

45. Lehtola S, Steigemann C, Oliveira MJT, Marques MAL. Recent developments in LIBXC — A comprehensive library of functionals for density functional theory. *SoftwareX*. 2018;7:1-5. doi:10.1016/j.softx.2017.11.002

46. Stephen PJ, Devlin FJ, Chabalowski CF, Frisch MJ. Ab Initio Calculation of Vibrational Absorption. *J Phys Chem*. 1994;98(45):11623-11627.

47. Heyd J, Scuseria GE, Ernzerhof M. Hybrid functionals based on a screened Coulomb potential. *J Chem Phys*. 2003;118(18):8207-8215. doi:10.1063/1.1564060

48. Makov G, Payne MC. Periodic boundary conditions in ab initio calculations. *Phys Rev B*. 1995;51(7):4014-4022. doi:10.1103/PhysRevB.51.4014

49. Mazurek AH, Szeleszczuk Ł, Pisklak DM. Periodic DFT calculations—Review of applications in the pharmaceutical sciences. *Pharmaceutics*. 2020;12(5). doi:10.3390/pharmaceutics12050415

50. Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys Rev B - Condens Matter Mater Phys*. Published online 1996. doi:10.1103/PhysRevB.54.11169

51. Kresse G, Hafner J. Ab initio molecular dynamics for liquid metals. *Phys Rev B*. 1993;47(1):558-561. doi:10.1103/PhysRevB.47.558

52. Kresse G, Hafner J. Ab initio molecular-dynamics simulation of the liquid-metalamorphous-semiconductor transition in germanium. *Phys Rev B*. 1994;49(20):14251-14269. doi:10.1103/PhysRevB.49.14251

53. Kresse G, Furthmüller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput Mater Sci*. 1996;6(1):15-50. doi:10.1016/0927-0256(96)00008-0

54. Campañá C, Mussard B, Woo TK. Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional. *J Chem Theory Comput*. 2009;5(10):2866-2878. doi:10.1021/ct9003405

55. Cho HS, Tanaka H, Lee Y, et al. Physicochemical Understanding of the Impact of Pore Environment and Species of Adsorbates on Adsorption Behaviour. *Angew Chemie*. 2021;133(37):20667-20673. doi:10.1002/ange.202107897

56. Vaidhyanathan R, Iremonger SS, Shimizu GKH, Boyd PG, Alavi S, Woo TK. Competition and Cooperativity in Carbon Dioxide Sorption by Amine-Functionalized Metal-Organic Frameworks. *Angew Chemie*. 2012;124(8):1862-1865. doi:10.1002/ange.201105109

57. Fischer M, Bell RG. Modeling CO2 adsorption in zeolites using DFT-derived charges: Comparing system-specific and generic models. *J Phys Chem C*. 2013;117(46):24446-24454. doi:10.1021/jp4086969

58. Burner J, Schwiedrzik L, Krykunov M, Luo J, Boyd PG, Woo TK. High-performing deep learning regression models for predicting low-pressure CO2 adsorption properties of metal−organic frameworks. *J Phys Chem C*. 2020;124(51):27996-28005. doi:10.1021/acs.jpcc.0c06334

59. Raza A, Sturluson A, Simon CM, Fern X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. *J Phys Chem C*. 2020;124(35):19070-19082. doi:10.1021/acs.jpcc.0c04903

60. Borboudakis G, Stergiannakos T, Frysali M, Klontzas E, Tsamardinos I, Froudakis GE. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Comput Mater*. 2017;3(1):1-6. doi:10.1038/s41524-017-0045-8

61. Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput Mater*. 2016;2(June):1-7. doi:10.1038/npjcompumats.2016.28

62. Altintas C, Altundal OF, Keskin S, Yildirim R. Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation. *J Chem Inf Model*. 2021;61(5):2131-2146. doi:10.1021/acs.jcim.1c00191

63. Anderson R, Rodgers J, Argueta E, Biong A, Gómez-Gualdrón DA. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem Mater*. 2018;30(18):6325-6337. doi:10.1021/acs.chemmater.8b02257

64. Cramer RD, Poss MA, Hermsmeier MA, Caulfield TJ, Kowala MC, Valentine MT. Prospective identification of biologically active structures by topomer shape similarity searching. *J Med Chem*. 1999;42(19):3919-3933. doi:10.1021/jm990159q

65. Sheridan RP, Rusinko A, Nilakantan R, Venkataraghavan R. Searching for pharmacophores in large coordinate data bases and its use in drug design. *Proc Natl Acad Sci U S A*. 1989;86(20):8165-8169. doi:10.1073/pnas.86.20.8165

66. Fernandez M, Tre NR, Woo TK. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal − Organic Frameworks for the Prediction of Gas Uptake Capacity. *J Phys Chem c*. 2013;117:14095-12105.

67. Hemmer MC, Gasteiger J. Prediction of three-dimensional molecular structures using information from infrared spectra. *Anal Chim Acta*. 2000;420(2):145-154. doi:10.1016/S0003-2670(00)00876-X

68. Hemmer MC, Steinhauer V, Gasteiger J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib Spectrosc*. 1999;19(1):151-164. doi:10.1016/s0924-2031(99)00014-4

69. Dureckova H, Krykunov M, Aghaji MZ, Woo TK. Robust machine learning models for predicting high co 2 working capacity and co 2 /h 2 selectivity of gas adsorption in metal organic frameworks for precombustion carbon capture. *J Phys Chem C*. 2019;123(7):4133-4139. doi:10.1021/acs.jpcc.8b10644

70. Krykunov M, Woo TK. Bond Type Restricted Property Weighted Radial Distribution Functions for Accurate Machine Learning Prediction of Atomization Energies. *J Chem Theory Comput*. 2018;14(10):5229-5237. doi:10.1021/acs.jctc.8b00788

71. Pinheiro M, Martin RL, Rycroft CH, Haranczyk M. High accuracy geometric analysis of crystalline porous materials. *CrystEngComm*. 2013;15(37):7531-7538. doi:10.1039/c3ce41057a

72. Willems TF, Rycroft CH, Kazi M, Meza JC, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater*. Published online 2012. doi:10.1016/j.micromeso.2011.08.020

73. He Y, Cubuk ED, Allendorf MD, Reed EJ. Metallic Metal-Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J Phys Chem Lett*. 2018;9(16):4562-4569. doi:10.1021/acs.jpclett.8b01707

74. Kancharlapalli S, Gopalan A, Haranczyk M, Snurr RQ. Fast and Accurate Machine Learning Strategy for Calculating Partial Atomic Charges in Metal-Organic Frameworks. *J Chem Theory Comput*. 2021;17(5):3052-3064. doi:10.1021/acs.jctc.0c01229

75. Iacomi P, Llewellyn PL. Data Mining for Binary Separation Materials in Published Adsorption Isotherms. *Chem Mater*. 2020;32(3):982-991. doi:10.1021/acs.chemmater.9b03376

76. Park S, Kim B, Choi S, Boyd PG, Smit B, Kim J. Text Mining Metal-Organic Framework Papers. *J Chem Inf Model*. 2018;58(2):244-251. doi:10.1021/acs.jcim.7b00608

77. Nandy A, Duan C, Kulik HJ. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. *J Am Chem Soc*. 2021;143(42):17535-17547. doi:10.1021/jacs.1c07217

78. Rosen AS, Iyer SM, Ray D, et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*. 2021;4(5):1578-1597. doi:10.1016/j.matt.2021.02.015

3 Characterizing a MOF database

3.1 Abstract

With the advancement of computational tools, databases of hypothetical materials have become important tools when searching for promising candidates for various applications. In the case of the highly diverse and tunable Metal-Organic Frameworks (MOFs), this search can be extremely daunting. To accurately sample the vast chemical and structural space of these materials, databases should contain structures with a variety of features such as surface area, chemical components, porosity, and topology. In this study, a database of 325, 000 hypothetical MOFs generated from 28 metal structural building units (SBUs), 164 organic SBUs and over 1000 underlying net topologies is characterized using Simplified molecular-input line-entry system (SMILES) strings.

3.2 Introduction

MOFs have been investigated for various applications including, drug storage and delivery,[1,2] optical devices,[3] chemical sensing,[4] catalysis,[5–7] and most notably, gas storage and separation.[8,9] At the experimental level, material discovery for CO_2 performance is extremely difficult for MOFs since synthesis is slow and manpower hungry, thus causing a bottleneck. Computational methods have provided alternative methods that have shown a lot of promise for materials discovery as several studies have used computational tools to establish important properties for gas adsorption. For example, Farha et al.[10] used computational modeling to design and predictively characterize a MOF with a particular high BET surface area, a property that is linked to high storage capacities. Based on simulated N_2 isotherm, they were able to calculate an ultrahigh BET surface area of 6,605 $m^2\ g^{-1}$ which motivated them to proceed with synthesis. The experimental and simulated BET surface areas of the MOF, which they named NU-100, were in

excellent agreement with one another, only differing by ~5%, demonstrating the efficacy of computational tools. The design of NU-100 was accomplished by replacing an organic ligand in a structure synthesized by Yan et al.[11] while using the same metal SBU and cubic space group to ensure the same topology. While these studies are important, the strategy of modifying existing materials fails to include an adequate sample of the nearly infinite chemical and physical space spanned by MOFs. Another good example of this is MOF UiO-66, which was synthesized by Cavka et al.[12] over a decade ago. UiO-66 is an archetypal Zr-MOF with a high surface area and high thermal stability as well as remarkable aqueous, acidic, mechanical, and water vapour stability due to its strong Zr-O bonds. The Zr-O node in the MOF has also shown powerful catalytic properties[13] while the aqueous stability and porosity have made the structure suitable for dye adsorption. Close to 400 papers have listed UiO-66 as the topic of paper, with experiments ranging from improving synthesis protocol by using modulators[14] and deprotonation,[15] to promoting structural defects to increase catalysis, hydrophilicity, and heats of adsorption.[16,17] A more practical way of approaching material discovery comes from high throughput screening studies. For example, Wilmer et al. used their 137,000 hypothetical MOF database to identify common geometrical characteristics shared amongst MOFs with optimal CH_4 uptake. While their database lacked sufficient topological diversity, they were able to conclude that surface areas of 1500-2000 $m^2\ g^{-1}$ and pore diameters of 5-6 Å were commonly found in highly performing MOFs. [18] More recently, Boyd et al. screened 300,000 hypothetical MOFs to identify water stable structures with strong CO_2 adsorption from wet flue gases. [19] This was done by first screening the database for MOFs with both high CO_2/N_2 selectivity and CO_2 working capacity and then analyzing the binding sites of these structures. It was found that materials with binding sites in the form of two parallel aromatic rings had the lowest water

affinity, and the CO_2/N_2 selectivity of the materials in dry and wet conditions were similar. Two of the structures that retained their capture performance in humid conditions were synthesized and their CCS properties were comparable to commercially available MOFs, which perform considerably worse in wet environments.

While the increasing computational power has given rise to numerous high throughput screening studies, gathering, and organizing data remains a significant roadblock.
In this work, the hypothetical MOF database generated by Boyd. et al.[19,20] is screened using Simplified molecular-input line-entry system (SMILES)[21] strings and the Cambridge Structural Database (CSD) Python API[22] to rename structures using a more appropriate naming convention. The renamed structures are then analysed for various geometrical features including volumetric and accessible surface area, cell volume, as well as various pore properties. Furthermore, structures with common errors and/or missing hydrogen were identified using SMILES strings and the CSD Python API.

3.3 Methods

MOF structures were generated using Topology Based Crystal Constructor (ToBasCCo) code developed in house by Dr. Boyd.[20] The construction algorithm is described next. First, the underlying topology is chosen. Based on the topology, metal and organic linkers with matching coordination linkers are chosen, i.e., if the topology calls for 3 coordinate linkers, then the SBU chosen must also be 3 coordinate. The list of metal and organic SBUs are shown in Figure 3.1 and 3.2, respectively.

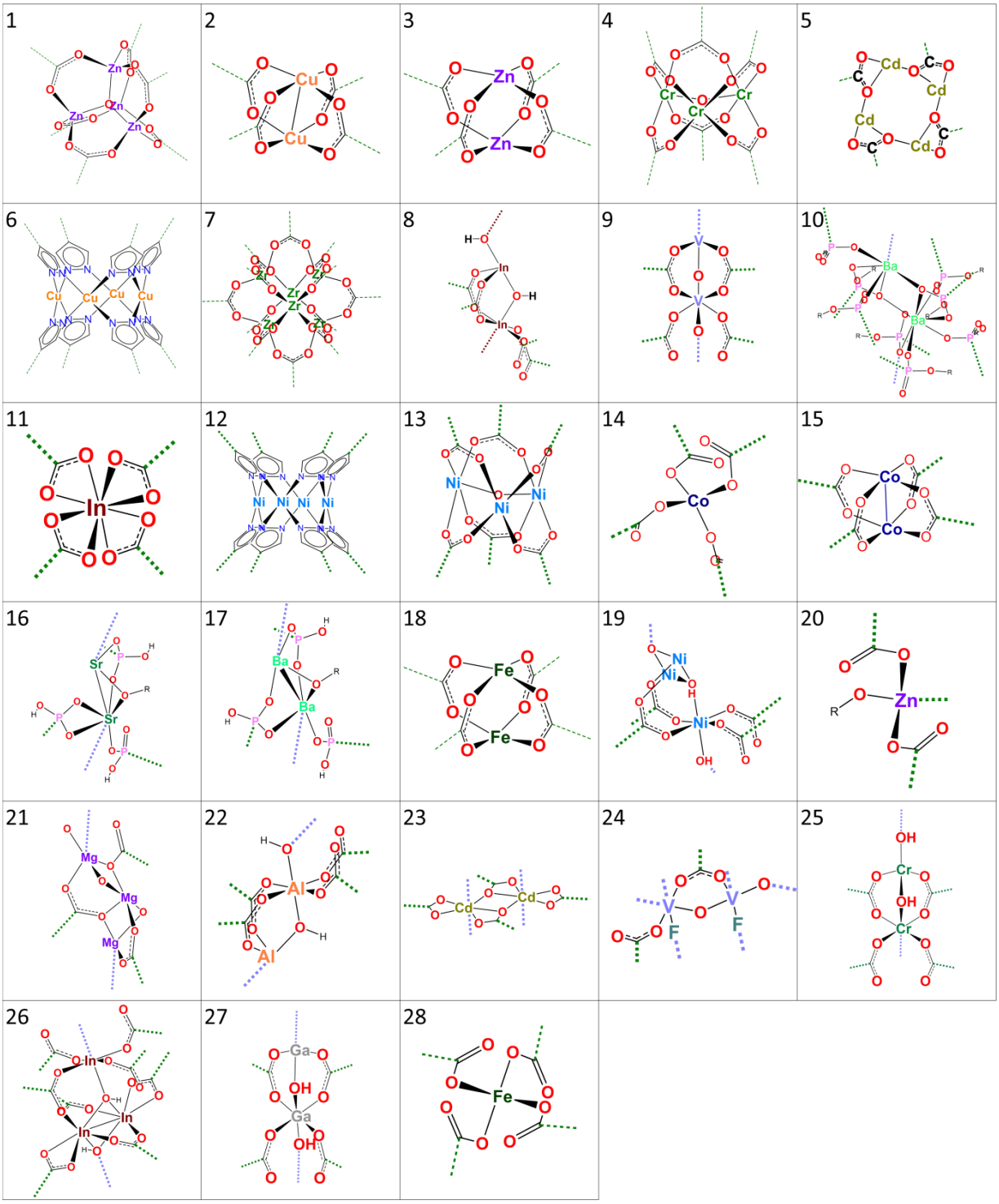
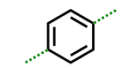
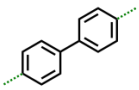
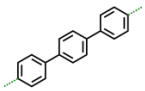
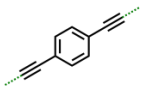
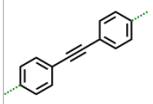
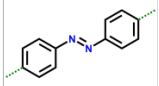
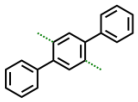
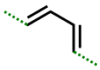
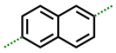
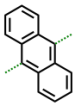
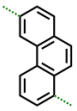
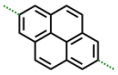
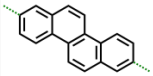
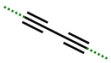
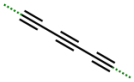
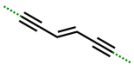
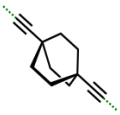
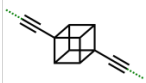
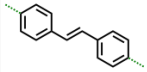
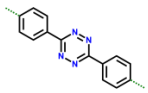
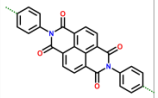
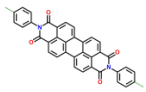
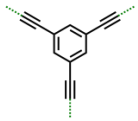
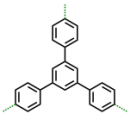
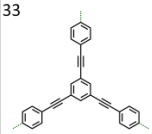
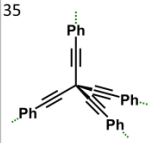
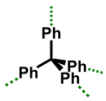
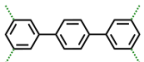
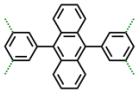
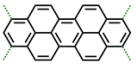
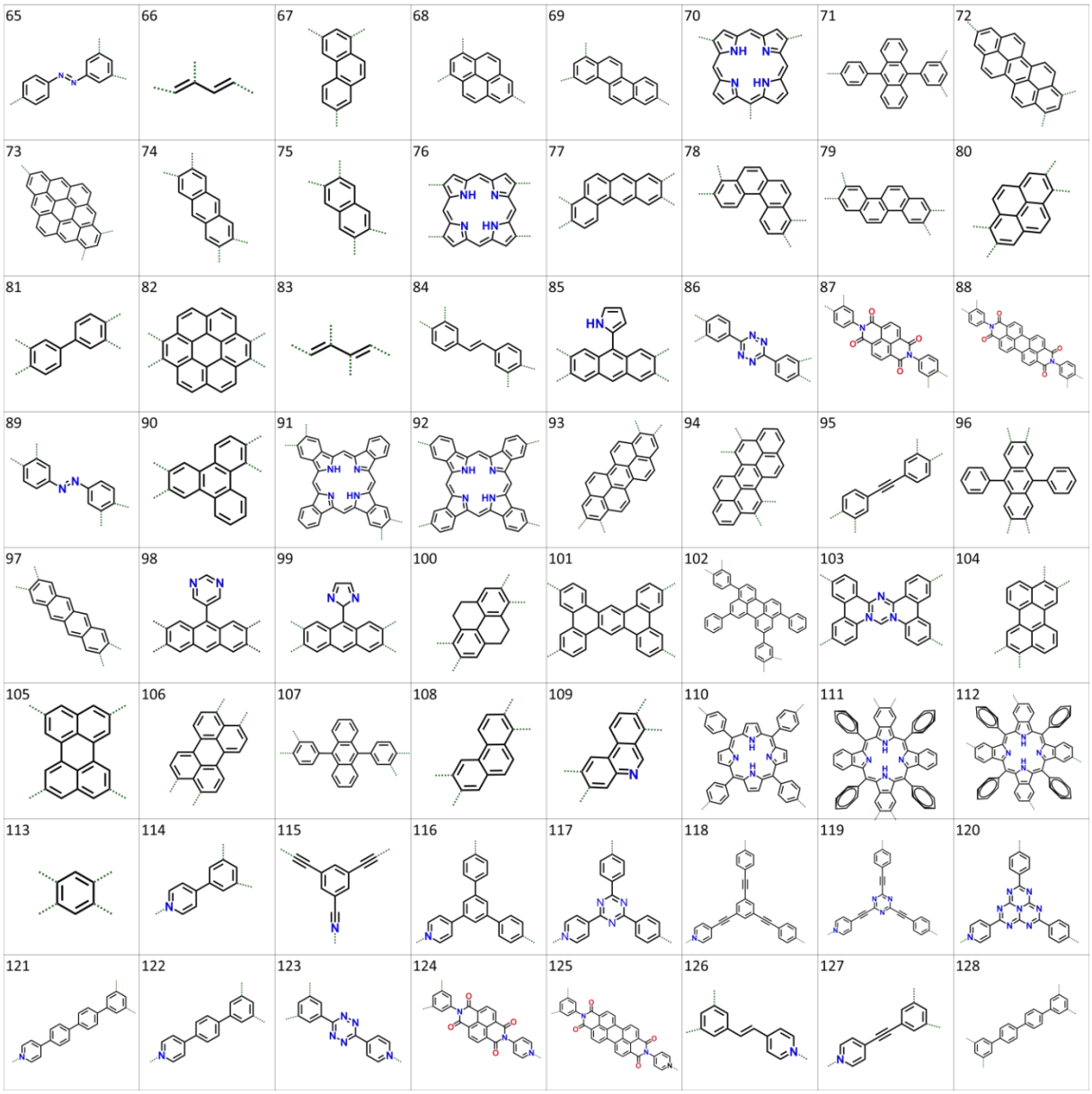


Figure 3.1 Metal SBUs used to generate the hypothetical MOF database.

1 	2 	3 	4 	5 	6 	7 	8 
9 	10 	11 	12	13 	14 	15 	16 
17 	18 	19 	20 	21 	22 	23 	24 
25 	26 	27 	28 	29 	30 	31 	32 
33 	34 	35 	36 	37 	38 	39 	40 
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64


65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128

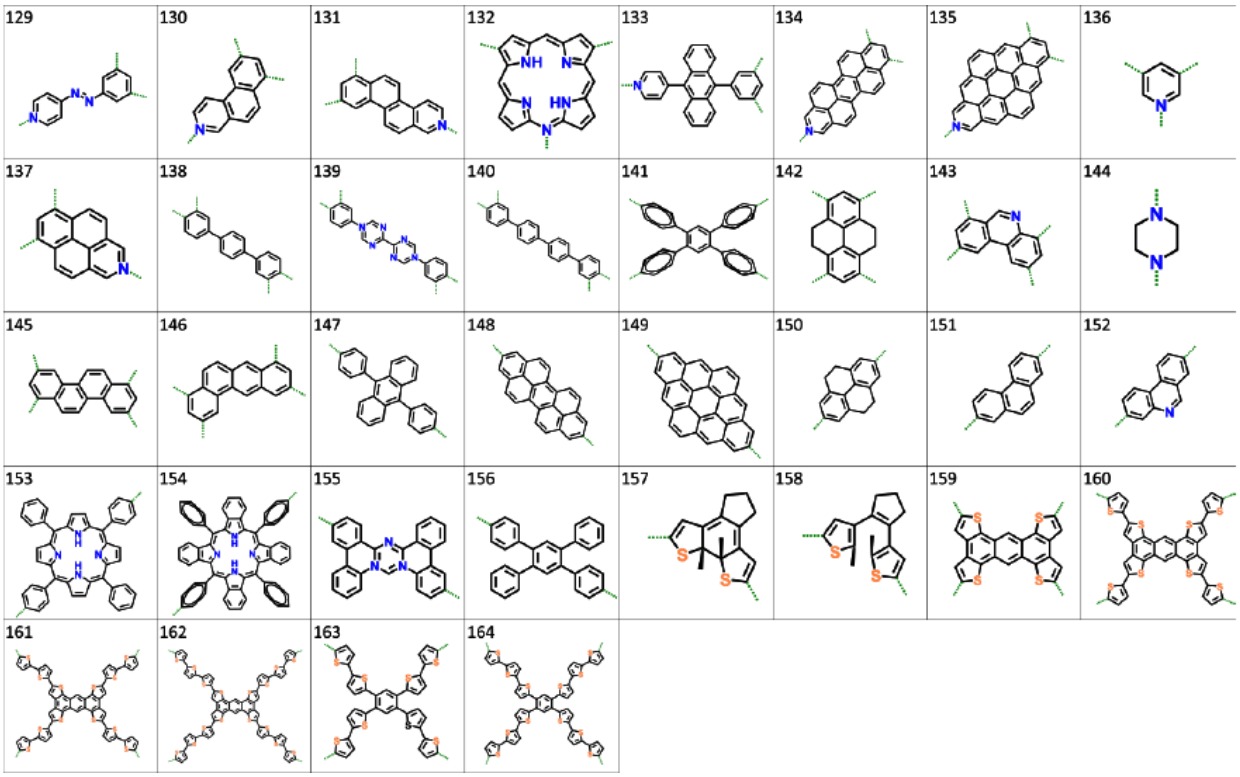


Figure 3.2 Organic SBUs used to generate the hypothetical MOF database.

The dotted green bonds represent carboxylate- or N capped moiety that binds to the metal SBUs.

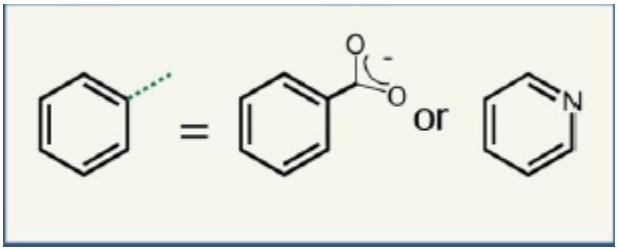


Figure 3.3 Carboxylate and nitrogen capping organic moieties.

Once chosen, pre-defined site locations will be used to connect the metal and organic SBUs together, as shown in Figure 3.4.

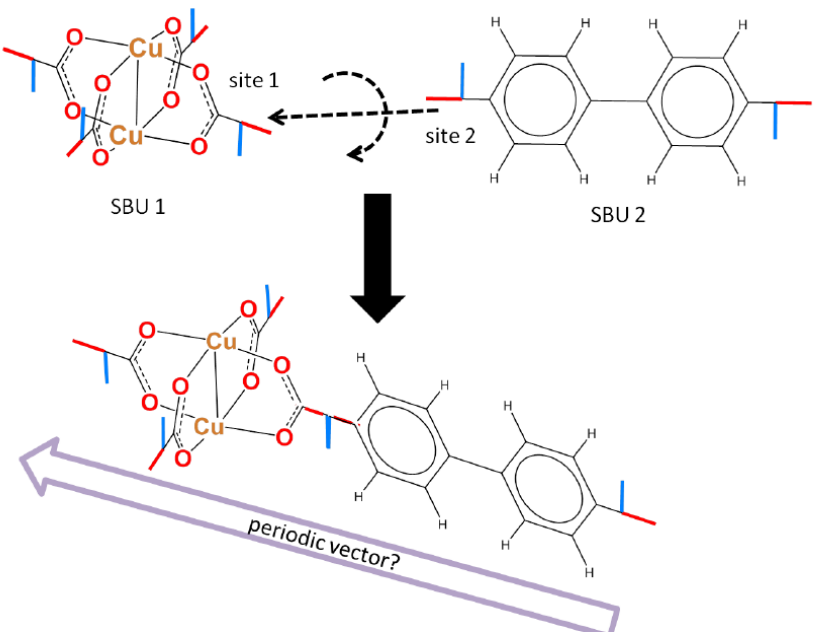


Figure 3.4 Binding sites used to connect metal and organic SBUs. The SBUs can connect in various orientations, as shown by the red and blue dashes.

When the SBUs are bonded, the algorithm will screen the resulting structure for other potential bond locations. The process can be repeated many times according to the desired underlying topology, as shown in Fig. 3.5.

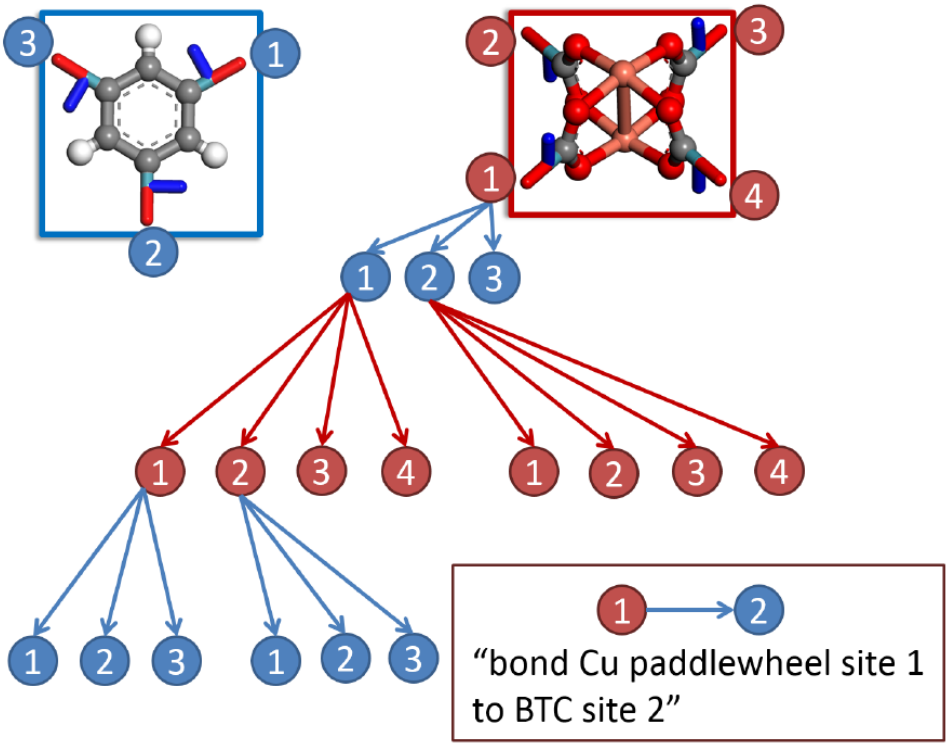


Figure 3.5 Recursive binding of metal and organic SBUs. Chemically similar SBUs (e.g., Organic #1 and #30) can have different numbers of binding sites, resulting in different topologies.

The method used to construct the structures is similar to the one described earlier by Wilmer et al.[23] where a set of metal and organic SBUs as well as functional groups are attached together iteratively like building blocks. Once built, the structures could be functionalized according to the chemical motifs shown in Fig 3.6.

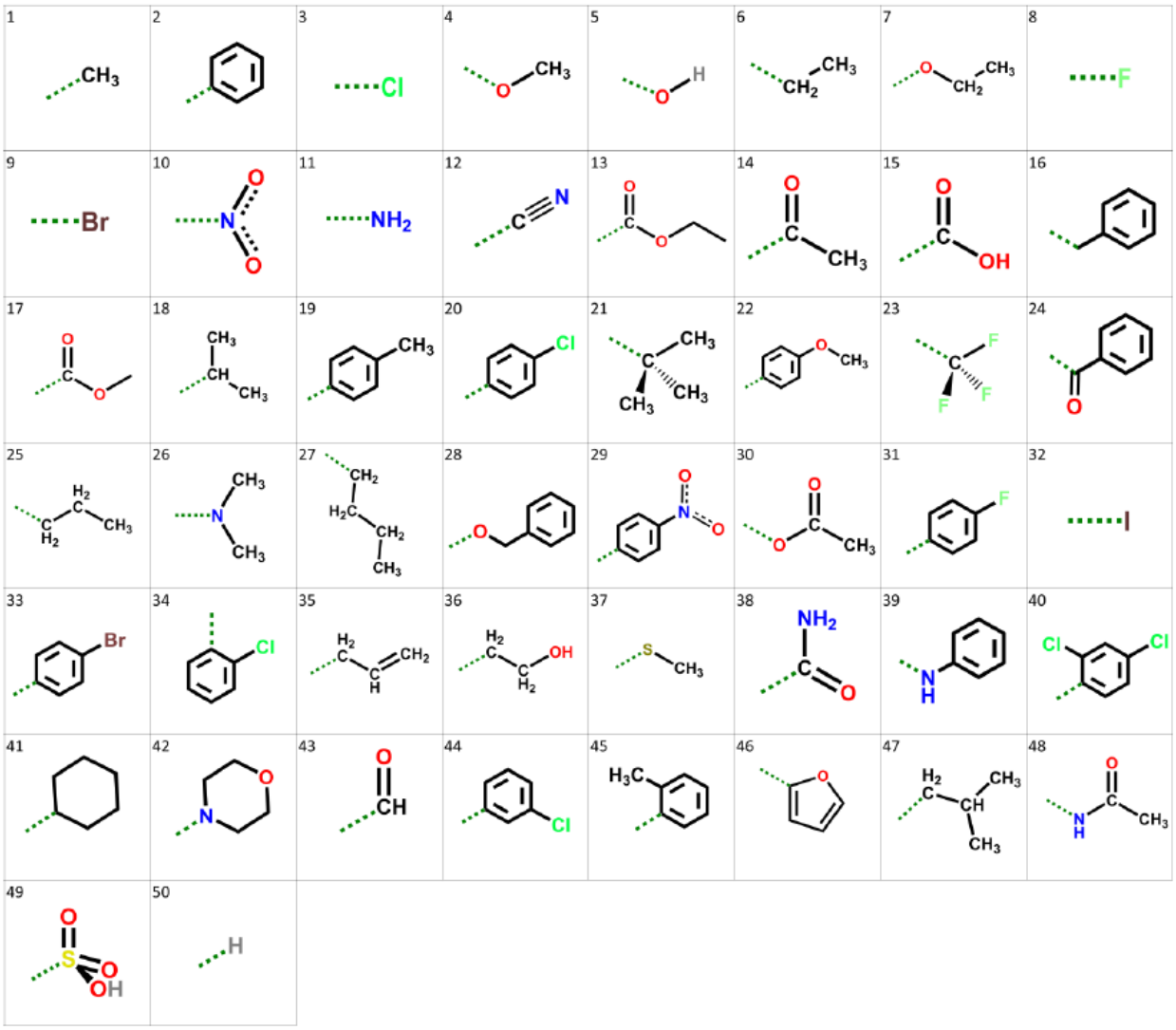


Figure 3.6 Functional groups used to further diversify the database.

The naming convention for the structures in this database is relatively straightforward, each structure is named based on the metal and organic SBUs used to create it as well as the functional groups and underlying topology of the structure. For example, if a structure was built using metal SBU 1, organic SBUs 5 and 6, functional group 8 with a **pcu** topology, the MOF would be named m1_o5_o6_f8.pcu.sym.100. The symmetry of the structure was obtained using an algorithm made to search for crystal symmetries called Spglib, here 100 is simply used as an example.[24]

After building the 300,000 hypothetical MOFs, a highly parameterized charge equilibration method was used to generate partial atomic charges. Using these charges as input, CO_2/N_2 selectivity and CO_2 working capacity simulations were performed for every structure to generate adsorption data.[19] Unfortunately, while the charge equilibration methods (Qeq, SQE, MEPO-Qeq, EQeq) can generally provide fast and reliable partial charges, due to their high parameterization they also suffer from generating inaccurate results in uncommon environments. These errors would consistently occur in certain metal or organic SBUs, and functional groups. It was therefore decided that the chemical motifs that yielded unsatisfactory results would simply be removed. As such, the remaining structures were re-numbered for presentability and continuity in the published work, as having a numbering system while skipping some numbers would most likely result in significant confusion. This resulted in many structure having filenames that did not correspond with the chemical motifs of the structure. The task of evaluating every structure for their corresponding SBUs, functional groups and topology would be a momentous one and it was therefore decided to keep the straightforward but incorrect convention. While the database enabled many groups to perform high level screening and machine learning work, the inconsistent labelling remained an issue for those interested in performing geometrical screening on MOFs with chemical motifs and/or topologies.[25–27] As such, Simplified molecular-input line-entry system (SMILES) strings were used to screen the structures for every organic SBU, metal SBU and functional groups mentioned previously. A brief explanation is provided in this work, but the original paper by David Weininger provides a more thorough explanation of SMILES methodology.[21]

3.4 Smiles Strings

One of the best tools for practicing chemists is the ability to search various databases (e.g., PubChem, SciFinder, ChemSpider). SMILES strings provide a reliable way to convert molecular structures into a computer-readable format. While there are many specific rules for certain scenarios, the most relevant to this work are shown below.

1. All atoms are represented by their atomic symbols (oxygen is ‘O’, carbon is ‘C’, etc.).
2. Double bonds are represented by ‘=’ while triple bonds are represented by a ‘#’, single bonds are generally omitted.
3. Parentheses are used when atoms branch off the parent chain, for example 2-methylhexane is represented by CCCCC(C)C.
4. For cyclic molecules, numbers are used to bond atoms that appear disconnected in writing (eg. cyclohexane is C1CCCCC1). Here the two ‘1’’s represent a bond between the carbon atoms.
5. Aromatic atoms are represented using lower case atomic symbols. (e.g., Toluene is Cc1ccccc1, where ‘C’ represents the methyl group, and c1ccccc1 represents the ring.

While these rules are relatively straightforward for small molecules, the labelling grows exponentially as larger chemical motifs are investigated. For example, organic SBU #52 (shown below), can be described using a SMILES string as:

O=c2c4ccc5c(=O)n(c1ccccc1)c(=O)c6ccc(c(=O)n2c3ccccc3)c4c56

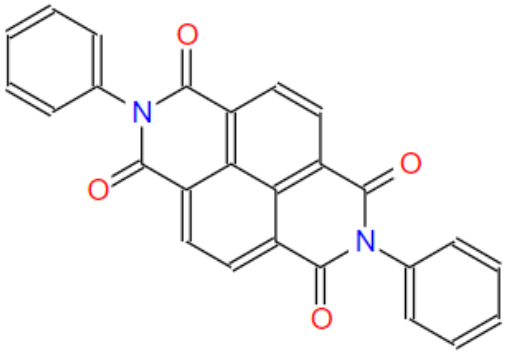


Figure 3.7 Organic SBU #52 and associated SMILES string.

While it may be difficult for humans to process this notation, computers are able to generate the matching chemical motif quickly and accurately from SMILES string. With the advancement and accessibility of powerful computers, many research groups have started including SMILES strings to facilitate characterization of various features in databases.[28,29] The CSD Python API Crystal Reader module was implemented into an inhouse code that could quickly screen crystalline structures for large numbers of chemical motifs. It is important to note that SMILES strings only work for molecules, not periodic structures like MOFs, however, the CSD Python API can convert the atomic positions of a given crystal into a molecule, allowing for SMILES strings substructure search.

In writing the code, the SMILES string for each organic and metal SBUs, as well as functional groups were generated manually. One of the main challenges with using SMILES strings is the inevitability to double count similar chemical motifs. For example, if organic SBU #140 is present in a structure, organic SBU #2 will also be found every time. This is because the SMILES string for organic SBU #140, c4ccc(c3ccc(**c2ccc(c1ccccc1)cc2**)cc3)cc4, **contains** the SMILES string for organic SBU #2 c2ccc(c1ccccc1)cc2 (shown in bold). Similarly for functional groups, if a MOF contained an ethyl group, it would also contain a methyl group. Overlooking

this issue would undermine the goal is the project, which is to rename each structure accurately according to the corresponding chemical motifs. Furthermore, the double counting would drastically inflate the number of some SBUs in the database, thus reducing the reliability of the naming convention.

There are two reasonable strategies for dealing with this double counting.

1. Explicitly defining every hydrogen atom when possible. Hydrogen are generally not included in SMILES string; it is assumed that the charges are neutral unless specified otherwise, however, they can be explicitly added. For example, without adding hydrogen labels, naphthalene's (c2ccc1ccccc1c2) SMILES string contains a benzene (c1ccccc1) SMILES string By adding every hydrogen where applicable, the SMILES strings for benzene becomes [H]c1c([H])c([H])c([H])c([H])c1[H], which is not present in the naphthalene SMILES strings [H]c2c([H])c([H])c1c([H])c([H])c([H])c([H])c1c2[H]
2. There are some organic groups that do not have convenient hydrogen that can be used to differentiate between them and other groups. The most common example for this is alkyne SBUs, like #5, C#C, and #6, C#Cc1ccc(C#C)cc1. There is no way to distinguish between them by specifying hydrogen bonds as there are no suitable locations, thus the best method would be to ensure that when SBU #6 is detected, SBU #5 is not counted. However, this can be quite challenging as every SBU must be mapped against every other similar SBUs to ensure no double counting is occurring. For example, the SMILES string for SBU #5 is present in 16 other organic SBUs.

Once finished, the structures could be renamed using the same naming convention used by Boyd, as described previously. With the structures renamed, geometrical analysis of the database was done.

3.5 Identification of chemically incorrect structures

Another benefit of using SMILES strings was the ability to identify structures with chemically ambiguous configurations, as shown in Fig 3.8. The benzene rings in this case have lost their aromaticity and are significantly distorted resulting in strange bonding patterns, which can happen upon optimizing the structure. Thus, when searching for organic SBU 26 with the SMILES string 'C(=C\c1ccccc1)/c2ccccc2' no matching strings would be identified as the rings are no longer aromatic (and would therefore have upper case 'C' labelling).

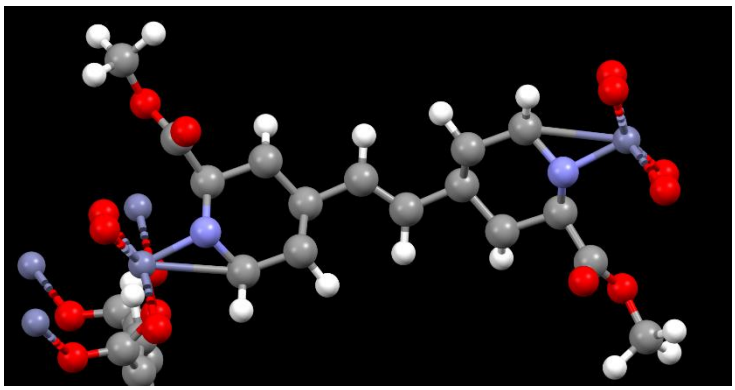


Figure 3.8 Example of highly deformed rings in a MOF.

However, this is not enough to categorize a structure as incorrect since the only information obtained is the lack of SBU 26. Furthermore, unless the MOF was built using only SBU 26 as its organic SBUs (which is possible), the SMILES string algorithm would still detect other organic SBUs in the structure. For example, if a structure contained SBU 26 (distorted) and any other SBU, say SBU 105, the algorithm would detect SBU 105, and no mistake would be identified. Fortunately, there are no other organic SBUs with the SMILES string 'C1CCNCC1'

(corresponding to a non-aromatic carbon ring with one nitrogen) and so, to identify all structures with this deformity, the code was modified to search for this chemical motif. If the SMILES string was found, the structure could then be removed from the database.

Furthermore, structures with chemically invalid bonding/missing hydrogen were identified using the CSD Python API's 'add_hydrogen' function, which generates hydrogen in locations deemed necessary based on bond connectivity. This function, in conjunction with the molecular weight function, can easily identify structures that are missing hydrogen. By taking the molecular weight of the structure, then using the add_hydrogen function and taking the molecular weight again, the chemically invalid structures can be identified if the weights are different. This test is extremely fast and reliable and was able to identify over 30,000 structures with chemically invalid bonds/missing hydrogen.

3.6 Geometric characterization of a MOF database

Various geometrical features were generated from using Zeo++,[30,31] an open source software for performing geometry-based analysis of porous materials. The calculated properties include cell volume, accessible surface area (ASA), accessible volume (AV), volumetric surface area, as well as the diameters for the largest included sphere (Di) also known as largest cavity diameter (LCD), the largest free sphere (Df), also known as pore limiting diameter (PLD), and the largest included sphere along the free path (Dif). The accessible surface area (ASA), as firstly described by Lee and Richards,[32] represents the area along the atomic surface a spherical probe with radius R can navigate without penetrating any other atoms. Similarly, the accessible volume (AV) is the volume that is accessible by the center of a probe.[33] The size of the probe can be changed depending on what application is being investigated. For example, for CO_2 adsorption properties a probe size of 1.65 Å is often used as this is the kinetic diameter (effective size and

based on the mean free path in the gas phase) of CO_2. ASA and AV often play a crucial role in CCS as various studies have demonstrated that high pressure gravimetric gas storage capacities are directly proportional pore volumes and/or surface areas.[34,35]

Gravimetric and Volumetric Surface area are measures of the surface area per weight and volume, respectively. These metrics offer a practical evaluation of a given structure as finding MOFs that are smaller and lighter but still have a high surface area are much more convenient and cheaper to implement than larger and heavier structures. It is therefore often best to find the optimal trade-off between adsorption properties and volume/size.[36]

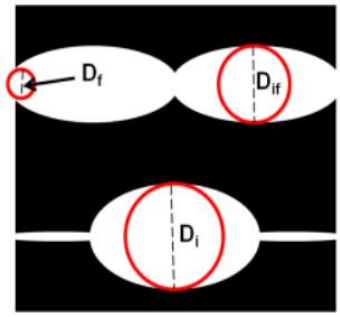


Figure 3.9 Difference between largest probe size (D_f), largest accessible cavity size (D_{if}) and largest cavity (D_i).

As shown in Fig 3.9, the largest included sphere, D_i, measures the size of the largest cavity in a porous material. The largest free sphere, D_f, represents the largest probe that can diffuse through the structure and the largest included sphere along the free path, D_{if}, is the largest cavity accessible along the path described by D_f. Pore size is another important feature for gas adsorption properties as having too large pores will cause guests to pass through and too small will be too restricting.

The importance of optimizing the properties mentioned previously is excellently described by Liu et al.[37]where they analyzed 8 isoreticular MOFs and showed a 94% CO_2 adsorption increase by simply varying the pore size.

3.7 Results

In this work, all properties were calculated with a probe size of 1.2 Å. These properties include void fraction, accessible volume per weight, accessible volume per unit cell, volumetric and gravimetric surface area. The probe radius of 1.2 Å is frequently used in literature as it is the smallest chemically relevant length, making it important when describing porosity in a large dataset. Alternatively, probe radii of 1.82 Å, corresponding to N_2, and 1.65 Å, corresponding to CO2, can also be used when characterizing MOF porosity for specific applications. The results of geometrical properties of the database are shown in Fig 3.10.

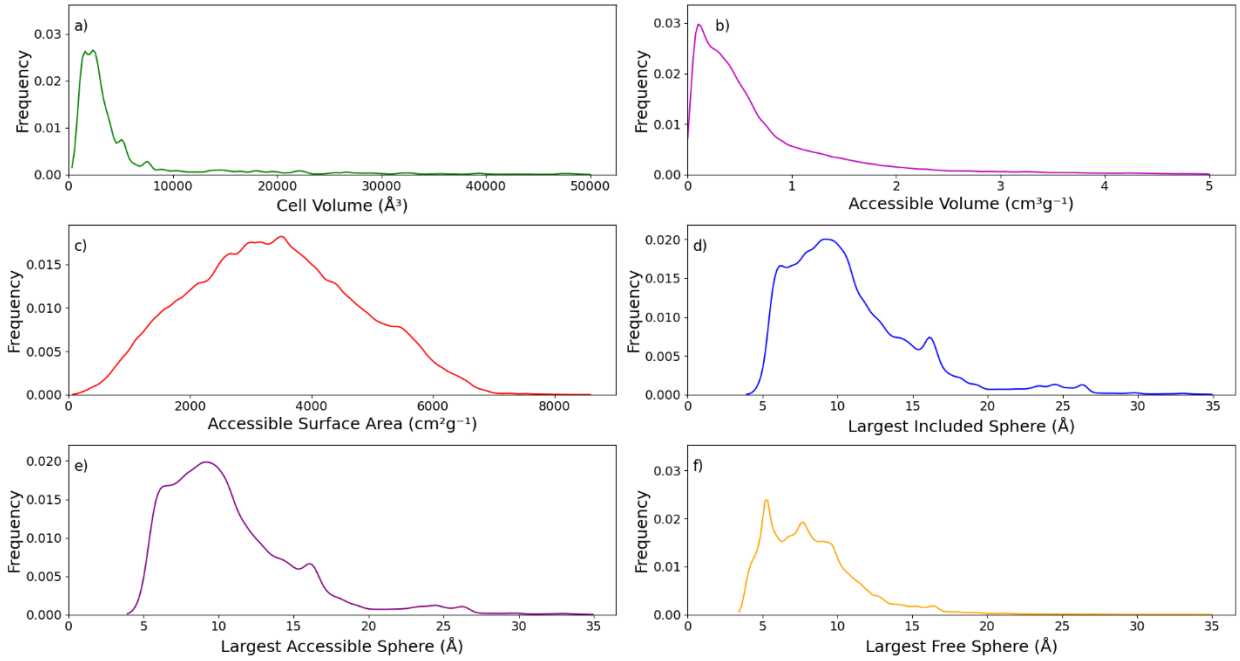


Figure 3.10 Geometrical distributions in the BW MOF database, N=300,000 a) accessible surface area, b) accessible volume c) largest accessible sphere, d) cell volume e) largest free sphere, f) largest included sphere.

Furthermore, using the cleaned-up filenames, the database can also be screened for topological and SBU diversity, as shown in Fig 3.11

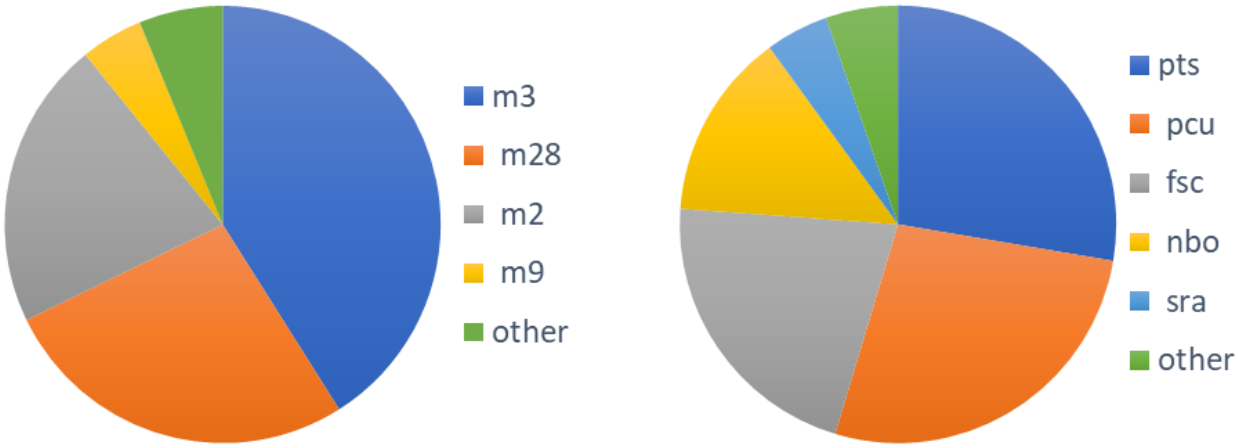


Figure 3.11 Distribution of the most common (left) metal SBUs and (right) topologies in the BW database

3.8 Conclusions

As the interest in metal-organic frameworks continues to grow, MOF databases become larger and more varied. As such, being able to organise and classify the structures in the databases in a simple yet effective manner is crucial. In this chapter, the CSD Python API and SMILES strings were used to characterize the BW MOF database, containing over 300,000 computationally generated MOFs. The analysis of the resulting SMILES strings yielded valuable information regarding the geometrical distribution of various properties in the BW MOF database. More importantly, the CSD Python API and SMILES strings were used to identify over 30,000 structures which contained either deformities and/or missing hydrogen. This method provides a good test for quickly filtering out chemically invalid structures.

3.9 References

1. Gulcay E, Erucar I. *Metal-Organic Frameworks for Biomedical Applications*. Elsevier B.V.; 2019. doi:10.1016/B978-0-12-817650-4.00006-1

2. Molavi H, Moghimi H, Taheri RA. Zr-Based MOFs with High Drug Loading for Adsorption Removal of Anti-Cancer Drugs: A Potential Drug Storage. *Appl Organomet Chem*. 2020;34(4):4-9. doi:10.1002/aoc.5549

3. Cui Y, Yue Y, Qian G, Chen B. Luminescent functional metal-organic frameworks. *Chem Rev*. 2012;112(2):1126-1162. doi:10.1021/cr200101d

4. Kreno LE, Leong K, Farha OK, Allendorf M, Van Duyne RP, Hupp JT. Metal-organic framework materials as chemical sensors. *Chem Rev*. 2012;112(2):1105-1125. doi:10.1021/cr200324t

5. Zhao C, Dai X, Yao T, et al. Ionic Exchange of Metal-Organic Frameworks to Access Single Nickel Sites for Efficient Electroreduction of CO2. *J Am Chem Soc*. 2017;139(24):8078-8081. doi:10.1021/jacs.7b02736

6. Yang Q, Xu Q, Jiang HL. Metal-organic frameworks meet metal nanoparticles: Synergistic effect for enhanced catalysis. *Chem Soc Rev*. 2017;46(15):4774-4808. doi:10.1039/c6cs00724d

7. Yang D, Gates BC. Catalysis by Metal Organic Frameworks: Perspective and Suggestions for Future Research. *ACS Catal*. 2019;9(3):1779-1798. doi:10.1021/acscatal.8b04515

8. Rodenas T, Luz I, Prieto G, et al. Metal-organic framework nanosheets in polymer composite materials for gas separation. *Nat Mater*. 2015;14(1):48-55. doi:10.1038/nmat4113

9. Bae TH, Lee JS, Qiu W, Koros WJ, Jones CW, Nair S. A high-performance gas-separation membrane containing submicrometer-sized metal-organic framework crystals. *Angew Chemie - Int Ed*. 2010;49(51):9863-9866. doi:10.1002/anie.201006141

10. Farha OK, Yazaydin AÖ, Eryazici I, et al. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat Chem*. 2010;2(11):944-948. doi:10.1038/nchem.834

11. Yan Y, Lin X, Yang S, et al. Exceptionally high H2 storage by a metal-organic polyhedral framework. *Chem Commun*. 2009;(9):1025-1027. doi:10.1039/b900013e

12. Cavka JH, Jakobsen S, Olsbye U, et al. A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. *J Am Chem Soc*. 2008;130(42):13850-13851. doi:10.1021/ja8057953

13. Rimoldi M, Howarth AJ, Destefano MR, et al. Catalytic Zirconium/Hafnium-Based Metal-Organic Frameworks. *ACS Catal*. 2017;7(2):997-1014. doi:10.1021/acscatal.6b02923

14. Schaate A, Roy P, Godt A, et al. Modulated synthesis of Zr-based metal-organic frameworks: From nano to single crystals. *Chem - A Eur J*. 2011;17(24):6643-6651. doi:10.1002/chem.201003211

15. Zhao Y, Zhang Q, Li Y, Zhang R, Lu G. Large-Scale Synthesis of Monodisperse UiO-66 Crystals with Tunable Sizes and Missing Linker Defects via Acid/Base Co-Modulation. *ACS Appl Mater Interfaces*. 2017;9(17):15079-15085. doi:10.1021/acsami.7b02887

16. Shan B, McIntyre SM, Armstrong MR, Shen Y, Mu B. Investigation of Missing-Cluster Defects in UiO-66 and Ferrocene Deposition into Defect-Induced Cavities. *Ind Eng Chem Res*. 2018;57(42):14233-14241. doi:10.1021/acs.iecr.8b03516

17. Winarta J, Shan B, McIntyre SM, et al. A Decade of UiO-66 Research: A Historic Review of Dynamic Structure, Synthesis Mechanisms, and Characterization Techniques of an Archetypal Metal-Organic Framework. *Cryst Growth Des*. 2020;20(2):1347-1362. doi:10.1021/acs.cgd.9b00955

18. Wilmer CE, Farha OK, Bae YS, Hupp JT, Snurr RQ. Structure-property relationships of porous materials for carbon dioxide separation and capture. *Energy Environ Sci*. Published online 2012. doi:10.1039/c2ee23201d

19. Boyd PG, Chidambaram A, García-Díez E, et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*. Published online 2019. doi:10.1038/s41586-019-1798-7

20. Boyd PG, Woo TK. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*. Published online 2016. doi:10.1039/c6ce00407e

21. WEININGER D. SMILES, a Chemical Language and Information System. *Am Chem Soc*. Published online 1987. doi:10.4159/harvard.9780674368446.c10

22. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. *Acta Crystallogr Sect B Struct Sci Cryst Eng Mater*. Published online 2016. doi:10.1107/S2052520616003954

23. Wilmer CE, Leaf M, Lee CY, et al. Large-scale screening of hypothetical metal-organic frameworks. *Nat Chem*. 2012;4(2):83-89. doi:10.1038/nchem.1192

24. Togo A, Tanaka I. Spglib: a software library for crystal symmetry search. Published online 2018:1-11. http://arxiv.org/abs/1808.01590

25. Fernandez M, Boyd PG, Daff TD, Aghaji MZ, Woo TK. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO2 capture. *J Phys Chem Lett*. 2014;5(17):3056-3060. doi:10.1021/jz501331m

26. Aghaji MZ, Fernandez M, Boyd PG, Daff TD, Woo TK. Quantitative Structure–Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2Working Capacity and CO2/CH4Selectivity for Methane Purification. *Eur J Inorg Chem*. 2016;2016(27):4505-4511. doi:10.1002/ejic.201600365

27. Moosavi SM, Nandy A, Jablonka KM, et al. Understanding the diversity of the metal-organic framework ecosystem. *Nat Commun*. 2020;11(1):1-10. doi:10.1038/s41467-020-17755-8

28. Werner JE, Swift JA, Quirós M, et al. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J Cheminform*. 2018;10(43):7290-7297. doi:10.1039/d0ce00273a

29. Werner JE, Swift JA. Data mining the Cambridge Structural Database for hydrate-anhydrate pairs with SMILES strings. *CrystEngComm*. 2020;22(43):7290-7297. doi:10.1039/d0ce00273a

30. Martin RL, Smit B, Haranczyk M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J Chem Inf Model*. 2012;52(2):308-318. doi:10.1021/ci200386x

31. Willems TF, Rycroft CH, Kazi M, Meza JC, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater*. Published online 2012. doi:10.1016/j.micromeso.2011.08.020

32. Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol*. 1971;55(3). doi:10.1016/0022-2836(71)90324-X

33. Do DD, Do HD. Appropriate volumes for adsorption isotherm studies: The absolute void volume, accessible pore volume and enclosing particle volume. *J Colloid Interface Sci*. 2007;316(2):317-330. doi:10.1016/j.jcis.2007.08.020

34. He Y, Zhou W, Yildirim T, Chen B. A series of metal-organic frameworks with high methane uptake and an empirical equation for predicting methane storage capacity. *Energy Environ Sci*. 2013;6(9):2735-2744. doi:10.1039/c3ee41166d

35. Kong GQ, Han Z Da, He Y, et al. Expanded organic building units for the construction of highly porous metal-organic frameworks. *Chem - A Eur J*. 2013;19(44):14886-14894. doi:10.1002/chem.201302515

36. Gómez-Gualdrón DA, Wang TC, García-Holley P, et al. Understanding volumetric and gravimetric hydrogen adsorption trade-off in metal-organic frameworks. *ACS Appl Mater Interfaces*. 2017;9(39):33419-33428. doi:10.1021/acsami.7b01190

37. Liu L, Konstas K, Hill MR, Telfer SG. Programmed pore architectures in modular quaternary metal-organic frameworks. *J Am Chem Soc*. 2013;135(47):17731-17734. doi:10.1021/ja4100244

4 Generating DFT charges for a MOF database

4.1 Abstract

This work discusses the screening of a hypothetical MOF database to obtain DFT-derived partial atomic charges. The DFT calculations are computed using the Vienna Ab initio Simulation Package (VASP)[1–4] and the partial atomic charges were derived from the DFT electrostatic potential using the REPEAT method. The database contains over 300,000 structures with a variety of topologies, as well as several thousand MOFs from the experimentally derived CoRE database. The distributions of the REPEAT charges are compared to the Split Charge Equilibrium (SQE) charges (discussed in section 2.5.4). Lastly, a simple machine learning model is developed to demonstrate relationships between geometrical features and CO_2/N_2 selectivity in MOFs.

4.2 Introduction

With the rapid rise of CO_2 concentrations in the atmosphere, and climate change becoming one of the primary problems faced by our generation, scientists have continuously searched for innovative solutions. The main source of greenhouse gases (GHGs) emissions comes from the burning of fossil fuel for heat and electricity production, transportation, and industrial processes. Several strategies have been proposed to reduce GHG emissions such as renewable energy sources, increased energy system efficiency, and implementation of carbon capture and storage (CCS). [5, 6] Currently, the CCS technology is too costly, as the process requires high temperatures (120-180 °C) thus reducing the electrical output of power plants with CCS technology by more than 30%.[7] As such, significant research has been done to find an alternative material that would make the overall process more feasible. Same of the more promising candidates for CCS applications are known as metal organic frameworks (MOFs), which are solid, nanoporous

materials which have shown several advantageous properties for CCS. Due to their high surface areas and pore volumes, they are ideal for CCS implementation. MOFs span a wide range of structures but can generally be described as porous structures composed of metal nodes connected by organic linkers, also called structural building units (SBUs). Since SBUs describe a wide variety of chemical motifs, which can be combined in different ways, the chemical space spanned by MOFs is enormous.[8,9] Therefore, conventional experimental approaches are considered too slow and many research groups have turned to computational methods to generate hypothetical structures in the hopes of identifying promising structures for CCS.[10,11] With the advancement of computational power, studies have been able to screen hundreds of thousands of hypothetical structures relatively quickly. To access the properties the structures during the screening, computational methods, such as Grand Canonical Monte Carlo (GCMC) simulations, are frequently used. [12] Briefly, MC simulations observe a guest particle, such as CO_2 or CH_4, interacting with a frozen framework (MOF). This particle can then be moved, rotated, or translated and each time it does this, the various forces and interactions between the particle and the framework are calculated. This process is repeated many times (100000-10000000) to create a simulation of the particle interacting with the framework. When combined with a grand canonical (GC) scheme, two other steps are added; insertion, where a guest particle is introduced in the system, and deletion, where a guest particle is removed from the system. Using these five actions, as well as various acceptance criteria discussed in section 2.1, a GCMC can sample a grand canonical ensemble. The only information required to perform a GCMC calculations are atomic positions and partial charges of the MOF framework. The atomic positions of the framework can be obtained from either experiments, or by using a construction algorithm to make hypothetical MOFs. For the atomic partial charges, several methods are detailed in chapter

2 which use pre-defined parameterized values or electrostatic potentials to fit the charges. Historically, due to the restriction of computational power, large-scale screening studies would often rely on lower level, less accurate methods for generating partial charges. It is also a common and reliable strategy to first screen many structures with low-cost methods and then further investigate potential candidates using higher level methods.

4.3 Methods

The database, referred to as the Boyd-Woo (BW) database, screened in this study was developed by Boyd. et al. using the ToBasCCo algorithm[10,13], a detailed explanation is provided in section 3.3. The algorithm used 28 inorganic and 164 organic SBUs, 50 functional groups, as well as 1166 underlying topologies to generate a chemically and topologically diverse set of 325,000 hypothetical structures. Periodic density functional theory (DFT) calculations were performed on most MOFs using the VASP package with the Perdew-Burke-Ernzerhof (PEB)[14] functional and the Project Augmented-Wave (PAW)[15] method. The input files for the VASP calculations, namely the INCAR, POTCAR, KPOINTS, and POSCAR files, were generated using a code developed in-house. The VASP calculation produced a LOCPOT file which contains the electrostatic potential used by the REPEAT method to generate the DFT-derived partial atomic charges, also called REPEAT charges. Due to the high computing demands of DFT calculations, Compute Canada resources were used to accelerate the screening process.

The resulting REPEAT charges were compared to split charge equilibrium (SQE) charges, discussed in section 2.5.4, which were already available in the BW database. For SQE, some of the properties, such as atomic hardness and electronegativity, are parameterized, reducing the computational cost as well as the accuracy. REPEAT uses the electrostatic potential generated from density functional theory (DFT) calculations to fit the partial atomic charges. While

generating DFT-derived ESPs is extremely computationally expensive (especially for large structures like MOFs), the resulting partial atomic charges are significantly more accurate than those generated from parameterized methods. Furthermore, since the partial atomic charges are required as inputs for simulations, the accuracy of the charges has a significant effect on adsorption values.

For both the REPEAT and SQE charges, the CO_2/N_2 selectivity simulations were performed using a temperature of 298K and pressure of 0.15 bar.

A Gradient Boosted Decision Tree (GBDT) model is used to find relationships between the geometric features; cell volume, accessible surface area (ASA), accessible volume (AV), largest included sphere (LIP), largest free sphere (LFP), and largest accessible sphere (LAF) and the target: CO_2/N_2 selectivity. A decision tree (DT) is a machine learning model that is created by making decisions based on the presented features. For example, when trying to figure out the CO_2/N_2 selectivity of a structure, the decision tree will use the values of the geometrical features (LFP, AV, etc.,) to generate a result. Using the training set (contains data for both the geometrical feature and the CO_2/N_2 selectivity) the model will be able to determine which geometrical features are most important when predicting CO_2/N_2 selectivity. The model can be further improved by gradient boosting the decision tree, resulting in a GBDT. Gradient boosting simply means tunning the hyperparameters (the parameters that define the **model**) to generate the best model (usually evaluated using an R test).

The model presented here used a 70:10:20 train-development-test split on 275,000 structures from the BW database. A grid search of the hyperparameters was conducted using the training set (n=192,500) to generate the models. The development set (n=27,500) was used to evaluate

each model generated from the training set. Finally, the best model (i.e. best combination of hyperparameters) is evaluated by the test set (n=55,000).

Table 4.1 Hyperparameters used in the development of the GBDT model and their associated tunning range to generate the search grid.

Hyperparameter	Tuning Range	Description
Loss	None	Function to be optimized
Learning_rate	[0.02,0.04,0.06,0.08,0.1]	Adjust learning speed of model
N_estimator	[100,200,300,400]	Number of trees to be used
Max depth	[3,10,20]	Maximum number of nodes in a tree
Min_sample_split	[2,5]	The minimum samples required to split an internal node

The table shows all the different hyperparameter values used to generate the search grid, which contain 120 grid points to be evaluated.

4.4 Results

The REPEAT charges were compared to SQE charges, as shown in Fig 4.1, for common atoms from the same database of MOFs. The plots are kernel density estimate (KDE) plots, which represent the data using a continuous probability density curve, analogous to a histogram. As expected, the distribution of charges for REPEAT values is more spread out than the parameterized SQE charges. Furthermore, while there is generally good agreement between the two methods in the case of nitrogen, sulfur, hydrogen and carbon, there are significant differences in values for zinc and oxygen values.

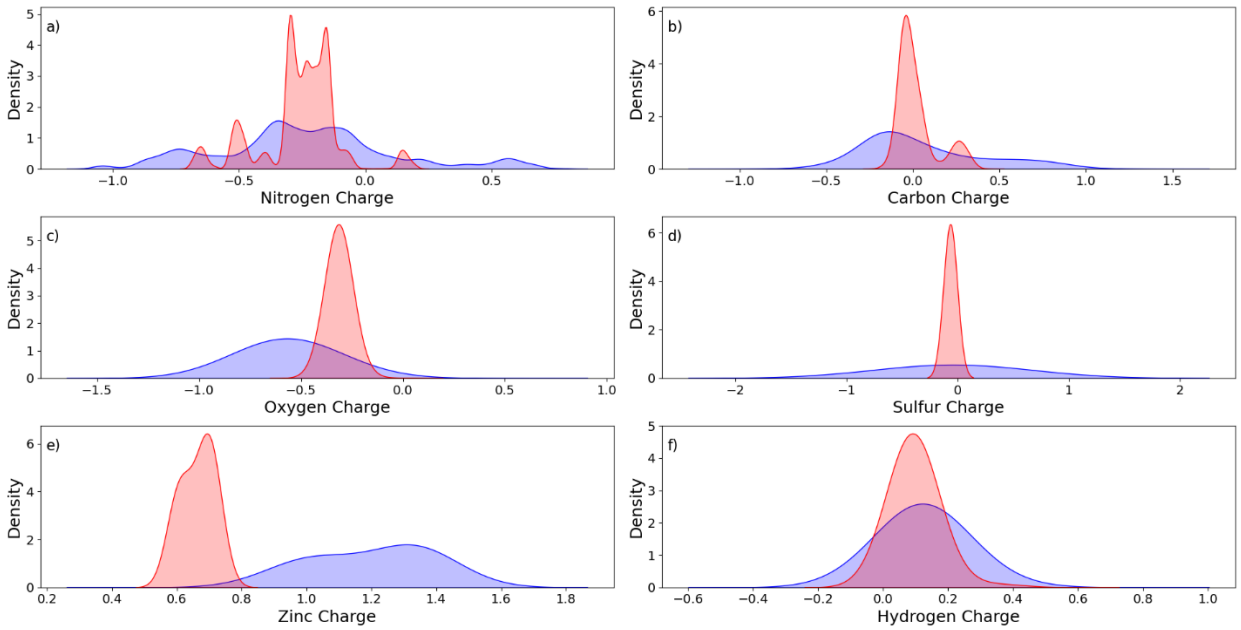


Figure 4.1 DFT (blue) and SQE (red) charge distributions in the BW MOF database for a) hydrogen (n=20,780,592), b) carbon (n=37,741,379), c) sulfur (n=132,885), d) nitrogen (n=2,086,161), e) oxygen (n=1,0789,833), f) zinc (n=713,898)

The CO_2/N_2 selectivity was also calculated for every structure using SQE and DFT charges, as shown in Fig 4.2. There is a good correlation, $r^2 = 0.96$, between SQE (y-axis) and DFT (x-axis) values, although SQE values are more localized due to the high parameterized nature of the method. The colors show the number of MOFs at certain values, where induvial structures are shown in purple while values corresponding to many structures are shown in green to yellow. As shown in the figure, most MOFs have value low selectivity (in the 0-100 range for both methods).

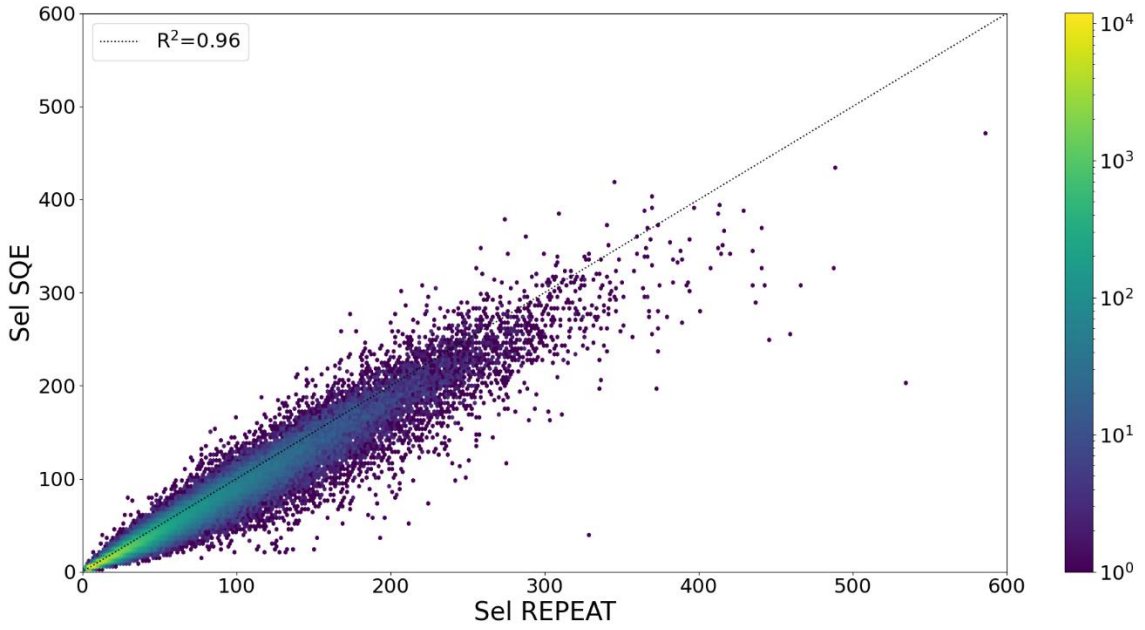


Figure 4.2 CO_2/N_2 selectivity generated from SQE charges (y axis) and REPEAT charges (x axis), n=300,000

The database, now equipped with REPEAT charges, is a useful tool for those interested in performing high-level machine learning as it provides a large and diverse set of structures.

4.4.1 Developing a machine learning model

As proof of concept, a gradient boosted decision tree model is developed to investigate the relationship between the previously discussed geometrical features and the CO_2/N_2 selectivity generated from the REPEAT charges. From the grid search, the optimized values for hyperparameters are a learning rate of 0.08, and max depth of 10, an n_estimator of 300, and a min samples split of 5. The resulting combination is tested on the test set resulting in an r^2 value of 0.95, suggesting a strong correlation between some, or all, the geometric features and CO_2/N_2 selectivity in the MOFs. To obtain a better understanding of which geometric feature most heavily impacts selectivity values in this model, a feature importance plot can be created.

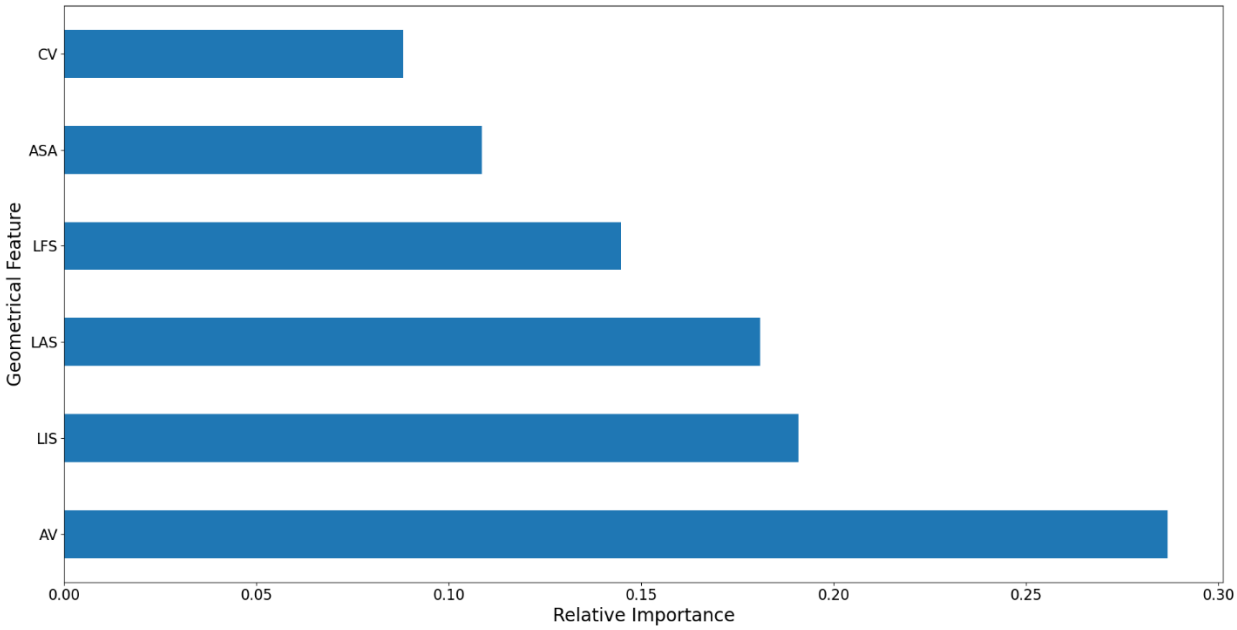


Figure 4.3 Relative importance of geometrical features on the GBDT model for predicting CO_2/N_2 selectivity. CV = Cell Volume, LAS, largest accessible sphere, LFS, largest free sphere, ASA = accessible surface area, LIS = largest included sphere and AV = accessible volume,

From the feature importance plot, the biggest contributor to CO_2/N_2 selectivity is the accessible volume (AV). Therefore, according to the model, when searching for structures with high CO_2/N_2 selectivity, this geometrical feature will most likely significantly affect the values. As mentioned in Chapter 2, the relative importance of these geometrical features can change significantly when predicting adsorption properties for different gases (such as CO_2/H_2).

4.5 Conclusion

In this chapter, DFT-derived partial atomic charges were generated for each structure in the BW Database. The CO_2 selectivity was then computed using the REPEAT charges and was compared to previously generated SQE CO_2 selectivity data. A good agreement, r^2=0.96, was found between the two methods, confirming the robustness of the SQE method in various environments. The atom specific charges were also compared for the two methods. For carbon,

sulfur, hydrogen, and nitrogen there was good agreement. There was a significant difference between the SQE and DFT values for Zinc. The database, now equipped with REPEAT charges, is a useful tool for those interested in performing high-level machine learning as it provides a large and diverse set of structures.

4.6 References

1. Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys Rev B - Condens Matter Mater Phys*. Published online 1996. doi:10.1103/PhysRevB.54.11169

2. Kresse G, Hafner J. Ab initio molecular dynamics for liquid metals. *Phys Rev B*. 1993;47(1):558-561. doi:10.1103/PhysRevB.47.558

3. Kresse G, Hafner J. Ab initio molecular-dynamics simulation of the liquid-metalamorphous- semiconductor transition in germanium. *Phys Rev B*. 1994;49(20):14251-14269. doi:10.1103/PhysRevB.49.14251

4. Kresse G, Furthmüller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput Mater Sci*. 1996;6(1):15-50. doi:10.1016/0927-0256(96)00008-0

5. Fawzy S, Osman AI, Doran J, Rooney DW. Strategies for mitigation of climate change: a review. *Environ Chem Lett*. Published online 2020. doi:10.1007/s10311-020-01059-w

6. Stanger R, Wall T, Spörl R, et al. Oxyfuel combustion for CO2 capture in power plants. *Int J Greenh Gas Control*. Published online 2015. doi:10.1016/j.ijggc.2015.06.010

7. Eldardiry H, Habib E. Carbon capture and sequestration in power generation: review of impacts and opportunities for water sustainability. *Energy Sustain Soc*. 2018;8(1). doi:10.1186/s13705-018-0146-3

8. Li B, Wen HM, Cui Y, Zhou W, Qian G, Chen B. Emerging Multifunctional Metal–Organic Framework Materials. *Adv Mater*. 2016;28(40):8819-8860. doi:10.1002/adma.201601133

9. Li H, Li L, Lin R-B, et al. Porous metal-organic frameworks for gas storage and separation: Status and challenges. *EnergyChem*. 2019;1(1):100006. doi:10.1016/j.enchem.2019.100006

10. Boyd PG, Woo TK. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*. Published online 2016. doi:10.1039/c6ce00407e

11. Wilmer CE, Snurr RQ. Towards rapid computational screening of metal-organic frameworks for carbon dioxide capture: Calculation of framework charges via charge equilibration. *Chem Eng J*. 2011;171(3):775-781. doi:10.1016/j.cej.2010.10.035

12. Kaija AR, Wilmer CE. High-Pressure Methane Adsorption in Porous Lennard-Jones Crystals. *J Phys Chem Lett*. 2018;9(15):4275-4281. doi:10.1021/acs.jpclett.8b01421

13. Boyd PG, Chidambaram A, García-Díez E, et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*. Published online 2019. doi:10.1038/s41586-019-1798-7

14. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett*. 1996;77(18):3865-3868. doi:10.1103/PhysRevLett.77.3865

15. Blöchl PE. Projector augmented-wave method. *Phys Rev B*. 1994;50(24):17953-17979. doi:10.1103/PhysRevB.50.17953

5 Conclusions

5.1 Summary

This thesis discussed two projects in the field of computational MOF research, primarily focused on data management and structure discovery. Chapter 3 focused on renaming the Boyd-Woo (BW) database, which was created by Boyd et al.[1,2] using the structure building ToBasCCO[3] program, resulting in a topologically diverse set of 325,000 hypothetical frameworks built from a wide variety of metal and organic SBUs. The database had previously established itself as a powerful tool generating machine learning MOF models and MOF discovery. However, due to the inconsistency in the naming convention, relating SBUs and topologies to various geometrical characteristics proved to be an issue. The incorrect naming also made it more difficult to have a good representation of the structural and chemical diversity of the dataset. Therefore, the motivation of this project was to use SMILES strings to rename each structure more appropriately, such that each filename would accurately describe the organic and metal SBUs, as well as the topology of the MOF. The database was then screened for many geometrical characteristics including, pore size properties, density, gravimetric surface area, volumetric surface area, and volume fraction.

Chapter 4 discussed the generation of REPEAT charges for the structures in the BW database. The motivation for doing this was to produce highly accurate DFT-derived partial atomic charges which could be used for a wide variety of applications, most notably generating machine learning models and high level GCMC calculations. The accuracy of the SQE CO_2 selectivity data was compared to the 'true' values generated using REPEAT charges and it was found that there was good agreement between the two. Furthermore, a simple machine learning gradient boosted decision tree model was developed to investigate any relationships between

various geometrical features and CO_2/N_2 selectivity. It was found that, according to the model, largest free sphere, accessible volume, and largest included sphere all have a significant impact on selectivity values.

5.2 Future Work

In chapter 3, a strategy for screening structures with missing atoms using SMILES strings was discussed. As more structures are added to the BW database, generating an algorithm to run frequent ‘cleanup’ of the database could be worthwhile to ensure consistent data format.
The REPEAT charges generated for the BW database in chapter 4 can prove useful for developing machine learning algorithm as it provides charges originating form a variety of chemical environments. While the machine learning model presented in this work provided some insight, more complex models could be developed to obtain a better understanding of the relationship between geometrical features and other gas adsorption properties.

5.3 References

1. Boyd PG. Computational High Throughput Screening of Metal Organic Frameworks for Carbon Dioxide Capture and Storage Applications. Published online 2016.

2. Boyd PG, Chidambaram A, García-Díez E, et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*. Published online 2019. doi:10.1038/s41586-019-1798-7

3. Boyd PG, Woo TK. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*. Published online 2016. doi:10.1039/c6ce00407e