Analyzing Carbon Capture Abilities of MOFs Using Graph Neural Networks

Anna Lapre

Primary Advisor: Professor Karianne Bergen

Second Reader: Professor Mara Freilich

Thesis completed in partial fulfillment of honors in Applied Mathematics - Computer Science (under the Brown CS Department requirements).

Acknowledgments

First, I want to thank my primary advisor Professor Karianne Bergen for her steadfast commitment to this project. I am so grateful for our weekly one on ones where I asked many questions (where to pivot when it seemed certain data sources no longer would suffice, when simulations turned out to be infeasible to run, etc.) as well as our more informal conversations about research and academia at large; I felt constantly supported and deeply appreciate the time you put into this project with me. I want to thank Professor Mara Freilich for her support and insights throughout the writing process as well as her passion for climate research. I also want to thank Professor Andrew Peterson for his help in providing broader context for computational research on MOFs. I of course want to acknowledge all of the professors I have had during my time at Brown: your love of learning and teaching is something for which I will forever be grateful. Particular thanks are due to the ENVS and DEEPs professors who have inspired me to work on climate both academically and professionally and to do our best for the one Earth we have. I want to thank my friends for supporting me throughout this process, especially during any of our mutual thesis study sessions; you all are amazing and inspired me with your passion for your own projects. Finally, I owe my family the biggest thanks: with your help, encouragement, and enthusiasm, I was excited to work on this project and tackle any obstacles or barriers when they appeared - thank you for everything.

Abstract

Excess greenhouse gasses are one of the largest contributors to climate change, a crisis which has had a devastating effect on human beings, biodiversity, and the Earth more broadly. The amount of carbon dioxide specifically, which has been emitted in incredibly large quantities since the Industrial Revolution, has served as a catalyst for the Earth's warming; yet, emissions are not decreasing quickly enough to prevent temperatures from crossing irreversible thresholds of damage. A potential solution for this problem is Carbon Capture and Storage (CCS); carbon capture is a method for removing or reducing the amount of greenhouse gasses in the atmosphere via some chemical process. One such process is through the use of crystal materials called Metal Organic Frameworks (MOFs). MOFs have a cage-like structure which can capture carbon by trapping carbon dioxide molecules while letting other atmospheric gasses simply pass through the MOF. The challenge is to find the MOF(s) which are best able to capture carbon, whether that be by having the largest capacity for the amount of CO_2 molecules it can hold or being particularly selective in trapping CO_2 compared to other atmospheric molecules. This, then, leads to the application of machine learning, particularly methods which easily depict chemical structures such as Graph Neural Networks (GNNs), to find the most optimal MOF for capturing carbon. This project uses GNNs to analyze the ability of MOFs to capture carbon holistically, i.e. in which MOFs are able to perform carbon capture such that various metrics of CCS are optimized instead of just one dimension (the literature norm). This project goes on to describe future steps and implications of this work.

Contents

1 Introduction

Climate change is considered one of the largest existential threats to humanity, nature, and the Earth's future. Hans-Otto Pörtner, a co-chair from the Intergovernmental Panel on Climate Change (IPCC), stated that, "the scientific evidence is unequivocal: climate change is a threat to human wellbeing and the health of the planet. Any further delay on concerted global action will miss the brief, rapidly closing window to secure a liveable future" [1]. Climate change has the potential to impact not just nature, but the ability for humans to live on Earth. Another co-chair from the IPCC, Dr. Debra Roberts, stated that "[f]ailure to achieve climate resilient and sustainable development will result in a sub-optimal future for people and nature" [1]. These are just a few statements outlining the necessity of progressing climate sustainable policies and technologies. Some of the impacts of climate change include rising sea level, extinction of various species, degradation of coral reef ecosystems, increased frequency and more severe climate extremes, and more [2].

The Paris Climate Agreement, which is an international treaty adopted by 196 parties, includes the goal of 1) holding average global temperatures "well below" an increased 2°C compared to pre-industrial levels and 2) to limit this global average temperature increase to below 1.5°C [3]. One of climate change's major catalysts is industrial greenhouse gas emissions. In order to meet the Paris goals, emissions need to reach their maximal point before the year of 2025. Prior to the 28th United Nations Climate Change Conference (COP28) near the end of 2023, a new report from UN Climate Change found that current progress towards tackling this goal is insufficient to remain below the 1.5°C threshold [4]. Fossil fuels, when burned, emit climate-damaging gasses such as carbon dioxide (CO_2), reducing the amount of energy that can leave the Earth's atmosphere [2]. Greenhouse gasses' ability to limit heat radiated back into space causes an imbalance in the heat being absorbed by the Earth and the heat sent back out, warming the Earth. In order to prevent this increased warming, carbon emissions need to be curtailed, and this process needs to be done urgently. The United Nations Framework Convention on Climate Change (UNFCCC) outlines the need for technological progress and development in order to address climate change [5]. Renewable energy, such as solar and wind, offer a means of replacing fossil fuel energy sources in the long-term - however, many of these resources are not yet able to be implemented at a global scale due to their cost and/or

inefficiencies. Due to the pressing need to act, scientists are in search of tools which work to both reduce the amount of carbon being emitted as well as remove existing greenhouse gasses in the atmosphere today. One such technology with the potential of reversing the immense amount of greenhouse gasses in the atmosphere is known as carbon capture.

This thesis is structured as follows. Section 2 discusses the necessary background on carbon capture and storage and gives an introduction to Metal Organic Frameworks as a method for capturing carbon; Section 2 also introduces how machine learning and various datasets can be used in this field. Section 3 provides rational for which datasets are used in this project. Section 4 articulates the methods used, including the model architecture, labels, and other hyperparameters. Section 5 communicates the results, including a brief comparison with existing literature. Finally, Section 6 discusses future work, limitations, and concludes the project as a whole.

2 Background

2.1 Carbon Capture and Storage

Carbon Capture and Storage (CCS) is the process of removing carbon dioxide and other prominent greenhouse gasses which exacerbate global warming from the atmosphere. Using this technology at the source of the production of tons of carbon dioxide, such as power plants and heavy transportation-using locations, is deemed to be an opportune approach for reducing the effects of these emission-heavy sectors [6]. The CCS process is typically broken down into three phases: 1) carbon dioxide must be separated from the atmosphere, 2) the captured carbon dioxide must be transported from the source to a final location where 3) it will be sequestered either in the ground, into some liquid or solid form, or some other final form [7]. Many research studies have focused on optimizing the various aspects of this process, from contrasting adsorption and polymer-based capture techniques to optimizing the transportation process using pipelines, etc. [7, 8]. This thesis will focus on the first step: optimizing techniques for capturing carbon from the atmosphere itself.

Researchers are developing a variety of methods for performing carbon capture. One technique is absorption-based carbon capture: amines, sometimes in the form of a liquid absorbent, are used to absorb carbon dioxide from steam, which is then stripped from the absorbent in the presence of

water vapor [9]. Another technique utilizes polymer-based membranes to separate carbon dioxide from other atmospheric gasses [10]. A third technique is to use Metal Organic Frameworks (MOFs) to capture carbon dioxide, either from industrial sources of carbon emissions or directly from the atmosphere. MOFs constitute a class of porous polymers that can be built up from metal compounds and organic ligands [11]. These linkages between organic and inorganic materials create structures that act as cages, which can be used to trap and store gas molecules such as carbon dioxide; then, a change in temperature or pressure works to release the trapped carbon dioxide from the MOF, as seen in Figure 1 [12]. This ability among others - including the structural stability, high surface area, and designability of these structures - makes MOFs a good candidate for carbon capture research [13, 14].

2.2 MOFs and Machine Learning

Metal Organic Frameworks have been studied for the past few decades as a promising technology to be used in carbon capture, with many of these studies invoking machine learning in order to assess the most optimal MOFs. Machine learning techniques are able to improve over the standard trial and error tools to screen materials for certain desirable features when doing carbon capture due to their ability to handle vast amounts of data. Some studies focus on using existing datasets, such as those which contain information on the crystal structure, its physical attributes, and various metrics for carbon capture as discussed in following sections, in order to screen these MOFs for some optimal feature [15]. Other studies invoke the use of generative artificial intelligence in order to compose potentially new MOFs which could be chemically created and tested [16, 17]. The following is a general literature review for the various research papers completed in the past few years on the use of machine learning and MOFs for carbon capture.

Many studies make use of existing datasets to investigate some measure of optimality for capturing carbon in order to discover well-suited MOFs. The various metrics used for measuring an MOF's ability to capture carbon are outlined in Table 1 below. One study used a large-scale screening of MOF structures and monte carlo simulations to assess the amount of CO_2 captured by MOFs in the presence of water [18]. This served as an interesting context for carbon capture as water molecules could compete with CO_2 to bind with (or be "captured" by) the MOF. This study made use of an

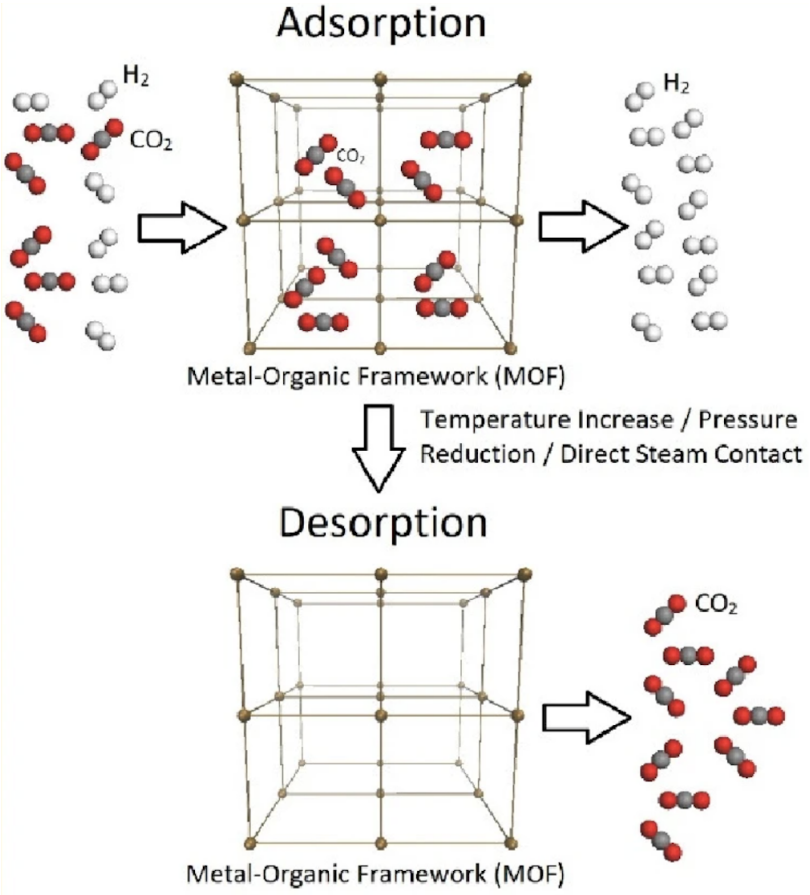


Figure 1: Depicts process of MOFs capturing carbon dioxide during adsorption, then releasing the gas during desorption. Image source: [12].

existing experimental MOF database and the authors performed their initial screening of MOFs by looking at Henry's law constants for carbon dioxide and water, rather than any form of machine learning. Another study used a similar approach of screening large datasets, instead using Grand Canonical Monte Carlo simulations, but used CH_4/H_2 selectivity as its threshold for optimal MOFs [19].

A challenge when attempting to screen MOF databases is the immense variability in these crystalline structures and the substantial computational power needed to effectively do so [15]. ML-assisted screening, however, offers a computationally cheaper and potentially farther-reaching tool to assess vast amounts of structures. One such study used a gradient boosted trees regression method

to assess a diverse set of network topologies of MOFs when looking at CO_2 capacity and CO_2/H_2 selectivity [20]. A different study looked at methane (CH_4), a very potent greenhouse gas, adsorption using a random forest model [21]. Like several ML assisted screenings of MOFs, this paper trained its model on a small portion of the dataset (in this case 8%) and tested the model on the remaining data so as to improve efficiency in model training and the need for relatively few labels. Until fairly recently, many of these studies, did not utilize the benefits of deep learning architectures, as well as physics-informed ML models via terms introduced in the loss functions, etc., which could serve as a catalyst for discovery by screening many more MOFs than those which currently have been studied. Now, many screening-based studies use some form of neural networks to identify MOFs with high carbon capture potential [22, 23, 24].

Other studies attempt to *generate* structures instead of screening and testing existing datasets in order to achieve optimal carbon capture. One example of this type of work is a study which used generative AI to create optimal MOF structures; the researchers focused on CO_2 adsorption capacity and synthesizable linkers. The authors identified six MOF structures surpassing CO_2 capacities greater than 2m mol/g; this benchmark demonstrates high carbon capture ability [16]. Another study trained a variational autoencoder (a type of generative deep learning architecture) to design MOFs, resulting in structures competitive with some of the best performing MOFs reported at the time of publishing [17]. This study produced $\sim$ 2 million MOFs. Due to computational resource constraints, only 45,000 randomly selected MOFs were used in further analysis. The authors also used a physics-informed loss function which encourages their models to produce MOFs which adhere to desired physical properties. They then performed Grand Canonical Monte Carlo Simulations on an even smaller select pool of MOFs to rank and produce the most competitive structures. Comparatively, there is less existing research on the utility of generative methods for producing MOFs used in carbon capture applications than that of high-throughput screening of these structures.

2.3 Commonly Used Data

Datasets used for analyzing MOFs with machine learning typically come from two sources: 1) data from experimental sources and published papers or 2) publicly available data [15]. Commonly used public data sets include MOFX-DB, CoRE, MOFomics, and RCSR. MOFX-DB is an online database

built from researchers at Northwestern University which tracks computational adsorption data for molecules including CH_4 and CO_2 for over 160,000 MOFs [25]. Computation-Ready, Experimental (CoRE) MOF 2019 database contains more than 14,000 three-dimensional MOF structures which includes physical properties, structural data, and pore analytics [26]. MOFomics is a database built from researchers at Princeton University which includes a variety of structural data, such as pore size distribution, surface area, various diameters, etc. for over 800 experimental MOFs and 1600 hypothetical MOFs [27]. Reticular Chemistry Structure Resources (RCSR) is another comprehensive database with roughly 1600 periodic nets which characterize the structure of MOFs [28]. Lastly, the Simplified Molecular Input Line Entry System (SMILES) is commonly used to uniquely and uniformly identify each MOF which is necessary for pre-processing of the above datasets [29].

Variables which appeared to be commonly used in the literature when looking at porous materials and their carbon capture abilities can be drawn from looking at studies of porous carbon materials, another type of porous materials similarly investigated for carbon capture, as well as MOFs. One study used features such as framework density, surface area, and pore size in combination with several Monte Carlo methods to both construct hypothesized porous carbon materials and compute their CO_2 capacities. [30]. Another study used various chemical descriptors including the "atomic property-weighted radial distribution function, the bag-of-atoms, and the chemical motif density" as well as geometric descriptors of MOFs in order to predict carbon dioxide capacity via neural networks [22]. A different study found that the moderate diameter, which measures the largest sphere fitting within the MOF structure, and the accessible surface area were important in constructing a MOF with maximal carbon dioxide capacity [31]. One study made use of textural properties of porous carbon materials to predict their carbon dioxide adsorption capacity; they trained a deep neural network using features including surface area, micropore and mesopore volume, pressure, and temperature [23]. These authors furthered their study and found that as pressure increased, mesopore volume contributed more to the overall prediction. When comparing an MOF's selectivity of trapping CO_2 over N_2, the authors found that high CO_2 uptake could not be achieved when selectivity of CO_2 was optimized [32]. This notion of tradeoffs between different carbon capture metrics is an important issue to tackle in determining the highest performing MOFs and will be discussed further in following sections. Some studies, however, use the entire structure itself, including atom

placement and bond measurements, in order to train their model; this is the approach taken in this work [24].

3 Data

3.1 Data Needed for Machine Learning with MOFs

In order to train a supervised machine learning algorithm on MOFs, researchers require a dataset containing a large amount of MOFs and their associated labels encompassing some metric for carbon capture. Due to the complexity of MOFs, including the wide range of potential atoms, their positions, and the bonds between them, the crystal structures cannot easily be recorded as standard numerical datasets with various features. Instead, researchers encode them as Crystallographic Information Files (.cif). These files contain substantial amounts of text, encoding a dictionary linking atoms to their positions, which can then be read by various programs to construct bond mapping and overall crystallographic structure [33].

These files also need associated labels; for a carbon capture application, this could be a measure of carbon uptake by the MOF, the MOF's selectivity for taking up carbon over nitrogen, its working capacity in an industry setting, or several others. Screening MOFs for their carbon capture abilities typically takes form as 1) experimental, 2) computational, or 3) as a combination of the formers [34]. The difficulty in experimentally derived labels stems from the time and financial costs of such experiments [35]. Rather than experimentally constructing and testing MOFs for various carbon capture attributes, which would be especially infeasible for large datasets of MOFs needed for deep learning models, researchers utilize algorithms to simulate the exposure of MOFs to atmospheric gas and record these results for their datasets. In particular, many researchers have turned to high throughput screening which relies on a computational approach in order to analyze thousands of MOFs and quickly identify top performers [34].

One such computational technique to gather carbon capture results is to use Grand Canonical Monte Carlo simulations. In the existing literature, Grand Canonical Monte Carlo (GCMC) simulations have been used to accurately predict carbon dioxide or methane absorptivity and selectivity as compared to nitrogen [36]. This is a standard simulation used for modeling adsorption abilities

of porous materials such as MOFs [37]. I spent some time looking at various packages for running GCMC simulations, such as RASPA, in order to produce an applicable label [38]. However, these simulations can be quite computationally expensive and require substantial amounts of time to run, on the order of hours to days for just one MOF. The technical understanding of porous material chemistry required to accurately run and gain new insights through these simulations also seemed too high for a senior thesis project. After deciding against simulating the carbon capture attributes myself, I looked for publicly available datasets which contained common MOFs in conjunction with some sort of applicable labels. One such dataset had been created by the COSMOS group which contained various labels for $\sim$ 4000 MOFs [39]. As with most machine learning models, more data with labels is typically more promising, so I continued looking for other datasets.

3.2 Computation-Ready Experimental database (CoRE)

Computation-Ready Experimental (CoRE) MOF 2019 database contains more than 14,000 MOF structures and is a common benchmark dataset for ML work on MOFs [26]. These MOF structures are stored as Crystallographic Information Files. Each MOF is stored in two separate datasets: 1) Free Solvents Removed (FSR) in which researchers removed free solvents contained in the crystal structure and 2) All Solvents Removed (ASR) in which researchers also removed any bound solvents attached to the structure (see Figure 2) The structures used in this project are the MOFs with all solvents removed, leaving just the crystal framework atoms for analysis. The CoRE dataset is complemented with structure-level attributes, including largest cavity diameter, pore limiting diameter, and volumetric surface area, which can be used in conjunction with the structures themselves to predict carbon capture-related attributes. The dataset does not include carbon capture-related labels for the MOFs.

In the existing literature GCMC simulations have been used for analyzing the CoRE dataset to accurately predict carbon dioxide or methane absorptivity and selectivity as compared to nitrogen [40, 41, 42, 43]. As previously mentioned, though, these simulations are not feasible for this project. Instead of applying GCMC simulations to every MOF in a dataset, several papers have used this dataset to test their models and apply GCMC simulations on their model's predicted highest performing MOFs. This allows researchers to verify their model's accuracy while reducing

the computational needs of their project.

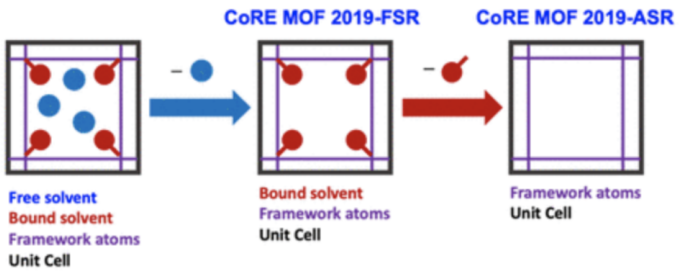


Figure 2: CoRE 2019-FSR and 2019-ASR datasets depicted visually. Image source: [26].

3.3 Boyd et al. Materials Cloud (BMC)

In order to have a dataset with a large amount of MOFs and have appropriate carbon capture labels, we needed to look beyond CoRE and the small datasets mentioned above. Fortunately, a study published in 2019 compiled a database of 324,426 hypothetical MOFs [44]. Each MOF was generated (meaning they were not physically constructed in laboratory experiments) using a graph theory method from Peter Boyd and Tom Woo [45]. This method combines a labeled quotient graph, representing a desired net topology, with structural building units (SBUs) present in nanoporous materials in order to construct MOFs. They repeat this process by altering the net topology to match the desired SBUs, creating various MOFs. They then ensured the MOFs were physically realizable with the use of the MEPO-QEq charge generation method, meaning that the electrostatic potential of each MOF is accurate. Each MOF is encoded as a crystallographic information file, accompanied by a multitude of carbon capture related labels. The authors obtained these labels by simulating each MOF and its interaction with post-combustion flue gas, the gas emitted at power plants and other high carbon-emissive centers. This narrowed the scope of their application of MOFs as a carbon removal technique at particularly emissive locations, optimizing their impact. Thus, the rest of this project focuses on this application as well, rather than simple direct air capture taking place at any location in the word. The authors also state that, in ensuring electrostatic potential was physically accurate, the carbon capture metrics generated will also be accurate with few incorrect

results. Among the over 300k MOFs, the authors chose roughly 8,000 as "top" MOFs for carbon capture and recorded a further study of their abilities. The various carbon capture adjacent metrics include the following:

- Adsorption of carbon dioxide as simulated at various temperature and pressure combinations, including at 298K and 313K at 0.15 bar CO_2.
- Selectivity of CO_2 adsorption compared to N_2 adsorption (Nitrogen being a dominant gas in the atmosphere with which carbon dioxide would compete to bind with the MOF).
- Working capacities, which combine information about the adsorption of CO_2 in different physical states.

The benefit of having such a rich dataset with various features, beyond the structure of each MOF, is that it allows for more nuanced analysis. It gives researchers the ability to predict various dimensions of an MOF in carbon capture rather than solely one label. This depth is the crux on which the rest of this project lies.

4 Methods

4.1 Creating a Composite Score

The aforementioned literature primarily relies on computing single-dimension statistics for measuring how well an MOF will perform as a carbon capture tool. The Boyd et al. dataset provides a vast array of these labels, as mentioned in the previous section. This allows for a variety of research questions, such as how selective is the MOF of bonding with, and capturing, CO_2 as opposed to N_2, what the working capacity of the MOF is, or what is its carbon capacity. All of these are important metrics from which industry users can decide which MOFs to potentially use in carbon capture applications when attempting large scale removal. Demir and Keskin noted the importance of three specific metrics in which MOFs should excel in order to be considered optimal carbon capture tools: adsorption selectivity, working capacity, and regenerability [46]. Raganati et al. also noted in their review paper, *Adsorption of Carbon Dioxide for Post-combustion Capture* the need for a carbon

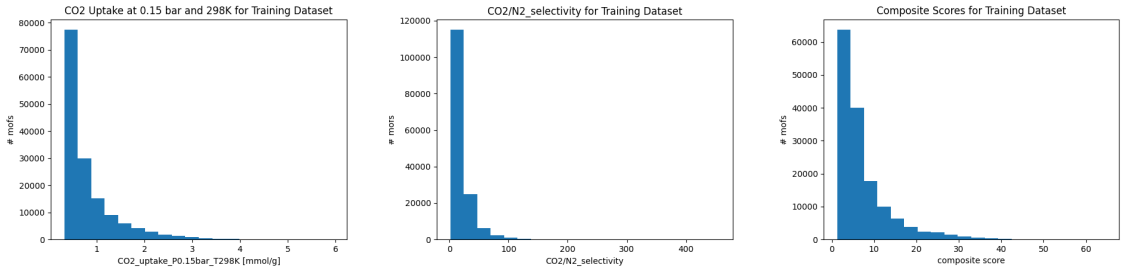


Figure 3: Plot A shows the distribution of MOFs for carbon dioxide uptake at 0.15 Bar and 298K, Plot B shows the distribution of MOFs for CO_2/N_2 selectivity, and Plot C shows the distribution of MOFs by the composite score defined in this project. Each plot has a similar right skewed-ness, preserving the distribution of MOFs in carbon capture abilities in the composite score calculation.

removal agent such as MOFs to satisfy a range of metrics in order to be economically and physically strong enough to be deployed in the real world [47]. Of the metrics they mention as being important, they include adsorption capacity/uptake and CO_2/N_2 selectivity, among others, some of which are more economically focused such as ease of regeneration and costs to create a hypothetical MOF. The Boyd et al. dataset has labels concerning adsorption uptake, selectivity, and working capacity; it does not have a regenerability measure, which could lead to potential further work if this data can be acquired or simulated. Importantly, each of these metrics plays an important role in understanding how well an MOF can capture carbon, and choosing just one metric to focus on limits the researcher in understanding the overall effectiveness of an MOF.

This project creates a "composite score," as has been done in public health research, in order to address this concern [48]. This composite score is a combination of different, relevant metrics wrapped into a single target label. One valid concern of this approach could be that the composite score itself lacks any physical meaning. However, this score still preserves the relative distributions of MOFs along these various metrics. In Figure 3, the distributions of working capacity and carbon adsorption are shown alongside the distribution for the composite score. We see similar skewness, meaning the MOFs along the right side of the plot excel in those dimensions, while taking into account these various metrics in a single distribution. The goal of the following model is to be able to predict this composite score, and in doing so predict which MOFs would generally be good, holistic candidates for industrial carbon capture applications.

The score takes the following form:

$$\begin{aligned}\text{composite score} = (1 + 0.1 * log(CO_2/N_2 \text{ selectivity})) \\ * (\mathbb{1}[\text{ excess } CO_2 \text{ uptake} \geq 2] \\ + 4(CO_2 \text{ adsorption uptake}) \\ + 2(\text{working capacity vacuum swing}) \\ + 2(\text{working capacity temperature swing}))\end{aligned}$$

Other variations of composite scores were tried, such as including CO_2 uptake at different pressures or altering the weighting of variables in the above mentioned loss. These often had worse results in training or did not create the correct distribution from which to train and predict; this misaligned the distribution in a way that prevented the model from being able to predict and isolated the best MOFs from mediocre MOFs.

4.2 Crystal Graph Convolutional Neural Network (CGCNN)

The underlying model for this project largely is based off of the model created by Tian Xie and Jeffrey C. Grossman in their paper published in 2018, *Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties* [24]. In this paper, Xie and Grossman create a deep learning framework which takes in crystal materials (such as MOFs) and learns material properties. They focus specifically on density functional theory (DFT) properties of these structures: this includes attributes like formation energy, absolute energy, and Poisson ratio. Although these properties are not directly linked to carbon capture, the ability of the model to learn structure-level properties is crucial for a similar application to the carbon metrics listed in Table 1. Xie and Grossman designed an architecture which closely resembles a traditional convolutional neural network, containing a series of convolutional layers followed by linear hidden layers and outputting a prediction label. The process of converting crystal files into tensors over which to be convolved is detailed in the next subsection, followed by a detailing of the architecture, the loss function, and training parameters and statistics for this project.

Metric	Description	Studies Using this Metric
CO_2 adsorption uptake *	Measures the uptake of CO_2 by an MOF via adsorption. In [44], this is measured in mmol/g and occurs in two different thermodynamic combinations: the measure used in the composite score if uptake simulated at 298K and 0.15 bar CO_2.	[44, 25, 32, 20, 16, 23, 30, 31]
CO_2/N_2 selectivity *	Measures the ratio of CO_2 adsorption uptake compared to N_2 adsorption uptake by the same MOF.	[44, 32]
excess CO_2 uptake *	Measures the "excess uptake" of CO_2 - this is not well defined in [44], but the purpose it serves in the composite score is as another measure for ability to absorb a high amount of CO_2.	[44]
working capacity vacuum swing *	Working capacity measures the difference of adsorption between thermodynamical points. For vacuum swing, adsorption was conducted as stated above and desorption was conducted at 363K and 0.1 bar CO_2.	[44]
working capacity temperature swing *	Working capacity is defined above. For temperature swing, adsorption was conducted as stated above and desorption was conducted at 413K and 0.7 bar CO_2.	[44]
CH_4/H_2 selectivity	Analogous to CO_2/N_2 selectivity, now comparing Methane and Hydrogen.	[19]
CO_2/H_2 selectivity	Analogous to CO_2/N_2 selectivity, now comparing CO_2 and Hydrogen.	[20]
CO_2/H_2O selectivity	Analogous to CO_2/N_2 selectivity, now comparing CO_2 and water.	[18]
CH_4 adsorption uptake	Analogous to CO_2 adsorption uptake, now focusing on Methane.	[21]

Table 1: Carbon Capture Metrics. * denotes a metric used in the model's composite score.

4.2.1 Graphical Representation of MOFs

A graph used in machine learning is typically composed of vertices and edges which are then encoded as tensors (matrices used in deep learning algorithms). In chemical structures such as an MOF, each atom signifies a node and a bond signifies an undirected edge. Crystal structures in particular are often represented as undirected multigraphs, meaning that two nodes may share more than one edge between them, representing different chemical bonds. Each MOF is represented as

three tensors, which are visualized in Figure 4a:

1. Atom Features - tensor of shape: (# of atoms in the MOF, # of features of an atom (atom embedding))
2. Neighbor Features - tensor of shape: (# of atoms in the MOF, (max) # of neighbors for each atom, # of features of a bond (bond embedding))
3. Neighbor Feature Indices - tensor of shape: (# of atoms in the MOF, (max) # of neighbors for each atom)

Atom Features maps each atom in the MOF to its atom embedding: initially, the atom embedding is simply a pre-set encoding used in Xie and Grossman's work, with each element (Hydrogen, Carbon, etc.) equated to a vector of length 92. This is then passed through a fully connected (embedding) layer with an outputted vector of dimension 64. So, the Atom Features tensor is of dimension (N, atom_embed_length) where N is the number of atoms in the structure and atom_embed_length is a specified embedding (64). Neighbor Features maps each atom to its neighbors and the bond embedding that signifies the edge between the atom and each of its neighbors. Due to the fact that each atom could have a different number of neighbors, this dimension of the tensor is set to be some max neighbor cutoff limit, chosen to be 12 by Xie and Grossman. This means that if there were to be an atom with more than 12 neighbors, the 12 atoms closest to the atom in question would be recorded here and all others discarded. This also means if an atom has fewer than 12 neighbors, all of its neighbors would be recorded and the remaining values in the tensor would be filled in with zeros. Xie and Grossman found 12 neighbors to be sufficient, as including more neighbors would introduce fairly weak bonds which in practice do not have a substantial effect. The bond embedding length is 41 due to how Xie and Grossman process the crystal structures from their initial .cif files. So, the Neighbor Features tensor is of dimension (N, M, nbr_embed_length) where N is the number of atoms in the structure, M is the number of neighbors (12), and nbr_embed_length is the embedding of bond features. Finally, Neighbor Feature Indices maps each atom to each of its neighbors in the Neighbor Features tensor. This is useful during the convolution in order to index the Atom Features tensor for each atom's appropriate neighbors.

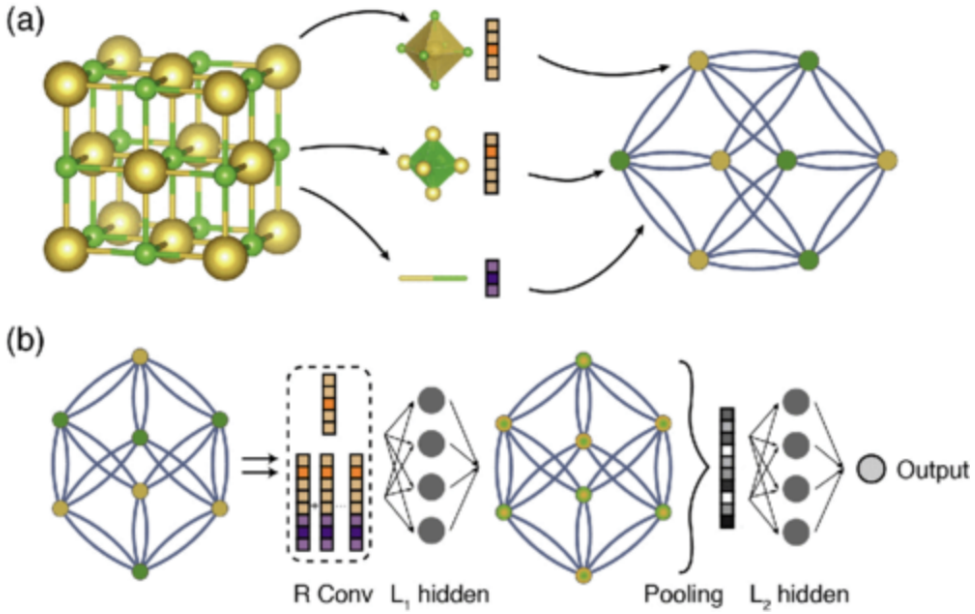


Figure 4: A: Crystal Representation and Embedding, B: CGCNN Model Architecture. Image source: [24].

4.2.2 Architecture

The general architecture is depicted in Figure 4b. First, each crystal, modeled by the three tensors outlined above, is passed into R convolutional layers. During one of these convolutional layers, each node in the Atom Features tensor (i.e. an atom in the crystal structure) is transformed by combining information about its neighbors, being other atoms with which this node shares a bond (these neighbors are indexed using the Neighbor Feature Indices tensor), and the bonds connecting that atom to its neighbors (information found in the Neighbor Features tensor). This is then wrapped in a nonlinear function in order to add more complexity and more opportunity for learning within each convolution. So, each of these layers updates each node represented in the Atom Features tensor using information about itself and its neighboring nodes. The other tensors remain unchanged, but are necessary for encoding the neighborhood context for performing each convolution and thus are repeatedly used but not changed. In this project, 6 convolutional layers are used to best characterize and understand the relationships between atoms in the crystal structure. Figure 4b shows that the

convolutional layers are followed by L_1 fully connected hidden layers; in this project, however, no hidden layers were used directly after convolution. Instead, the pooling layer is used immediately after the convolutional layers in order to summarize the information learned in the first segment of this architecture. Xie and Grossman choose to use a normalized summation over each node in the Atom Features tensor as their pooling layer, which is also used here. Then, this normalized and summed representation is passed through L_2 fully connected hidden layers; in this project, 3 of these fully connected layers were used. Finally, a linear output layer produces a single value as the predicted composite score for the inputted crystal structure.

4.2.3 Custom Loss Function

The goal of this research project is to find MOFs with the best carbon capture ability or, quantitatively, the highest composite score. Due to the skewness displayed in Figure 3, using a uniform loss function such as Mean Average Error would result in the model's not understanding these skewed data points. It would most likely always predict values in the lower end of the distribution, as that is where the bulk of the data lies and would thus minimize loss. In order to combat this, one could use mean squared error to emphasize points on the ends of the distribution. However, due to the quantity of data in the center of the distribution as opposed to the right extreme, this loss can be improved upon by explicitly weighting the loss by where the MOFs fall on this distribution. This was achieved by using loss defined by:

$$\text{Loss(prediction, target)} = (w * \text{(target-prediction)})^2 \text{ where:}$$

$$w = \begin{cases} 0.5 & \text{target} < 40 \\ 10 & \text{target} \geq 40 \end{cases}$$

This weighted mean squared error emphasizes these extreme cases with high carbon capture potential while ensuring the model still is learning from the bulk of the database. Other varieties of this loss function included different thresholds (40 in the final loss above) between which to differ weights, multiple different thresholds with differing weights (staggered weighting), etc., but the above loss

performed the best, as determined by looking at precision and recall for the model's validation predictions (see Results section).

4.2.4 Hyperparameters and Training Statistics

The model was trained using Brown's super computer, OSCAR, using 4 CPU cores and 1 GPU. The end to end runtime for training the model took several hours; it varied by training run, but the final training run took 7 hours and 4 minutes. This was primarily due to the time needed to load and compile the data from the .cif files into the three tensors specified above (the graphical format to be loaded) which was done on the CPUs as opposed to the GPU. In future work, more or all of the data could be used if the code were also better optimized to load in the data. As a result, 150,000 MOFs were used in training rather than the entire available dataset (around 300k MOFs), which still provided a breadth of MOFs on which the model was able to train. Furthermore, as we wanted to particularly focus our model on the best MOFs, the top 150,000 MOFs according to carbon uptake were chosen rather than a random subset of the 300k dataset; this mitigated some of the skewness present in the data as mentioned in the previous section. A potential downside to this approach is the model will not have been exposed to many MOFs with low carbon uptake; however, due to the nature of this project being about identifying MOFs with particularly high carbon capture abilities, this is not too big of an issue. Of the 150k MOFs, 80% were used directly for training, 10% were used as a validation set, and 10% were used as a test set. Then, the CoRE dataset was used as a completely separate test set only used for the final model. Other hyperparameters include: 30 epochs, optimizer is Stochastic Gradient Descent, learning rate is 0.1.

5 Results

5.1 Model Results on Boyd et al.

Multiple models were trained with varying hyperparameters. In order to decide which model best predicted carbon capture ability, the prediction and target values were plotted followed by a precision and recall analysis. First, prediction vs target values for the initial validation set were plotted. This reduced the pool of potential models down to two; their differing hyperparameters are

shown in Table 1.

Parameter	Model 1	Model 2
# convolutional layers	3	6
# hidden linear layers	2	3
custom loss weight threshold	25	40
weight value above threshold	2	10

Table 2: Hyperparameter Differences between Model 1 and Model 2.

Figure 5 plots the models' predictions as compared to the target values for the BMC test set. Each sample is colored by whether or not it was classified as a top MOF in the BMC dataset. BMC identified top MOFs based on their carbon uptake; although only one metric is used for this distinction, it still serves as a useful guide for labeling MOFs as good at carbon capture when looking at model results. As depicted, both have relatively good prediction capabilities in terms of their ability to roughly match the target values regardless of where along the target distribution the sample is. Due to the skewness of the data, this was a concern if the model would just predict lower values, but both models overcame this challenge. The question then became what do we want to prioritize in order to evaluate what characteristics determine a "good" model. To do so, we look at the precision and recall rates of our model. As this is a regression task, we can establish thresholds for which MOF scores above this threshold are "positive" cases (MOFs that are good at carbon capture) and MOF scores below are "negative" cases (MOFs that are bad at carbon capture). We can then adjust this threshold and see how both precision and recall vary by threshold for both models, as shown in Figure 6.

In this project, we seek to discover MOFs that will be exceptionally good at carbon capture. So, we will take the MOFs with the highest predicted scores and proceed with next steps for experimenting in the real world. As a result, we want our model to only predict high scores for MOFs with actual high scores. This means we care about optimizing precision, which is the rate of correctly predicting positive cases out of all predicted positive cases. Looking at Figure 6, we see that precision is on average higher for Model 2 than for Model 1 with comparable recall. We also see Model 1's decline in both precision and recall with very high thresholds (> 45) as well as a steep decline

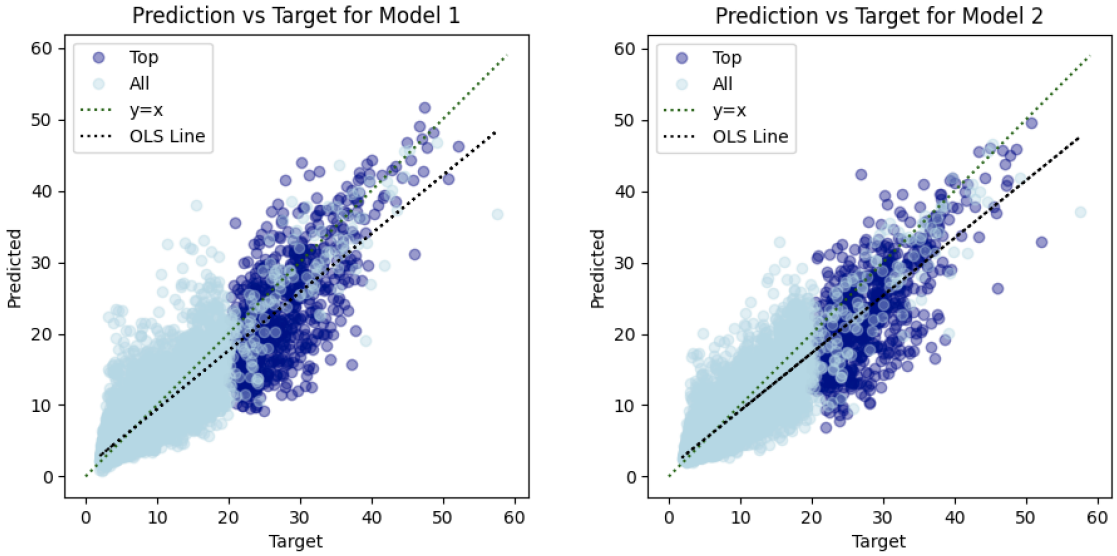


Figure 5: Model 1 (left) and Model 2 (right) predictions compared to target composite scores. The green line is the identity function (signals perfect prediction) and the black line is the OLS line of prediction by target value.

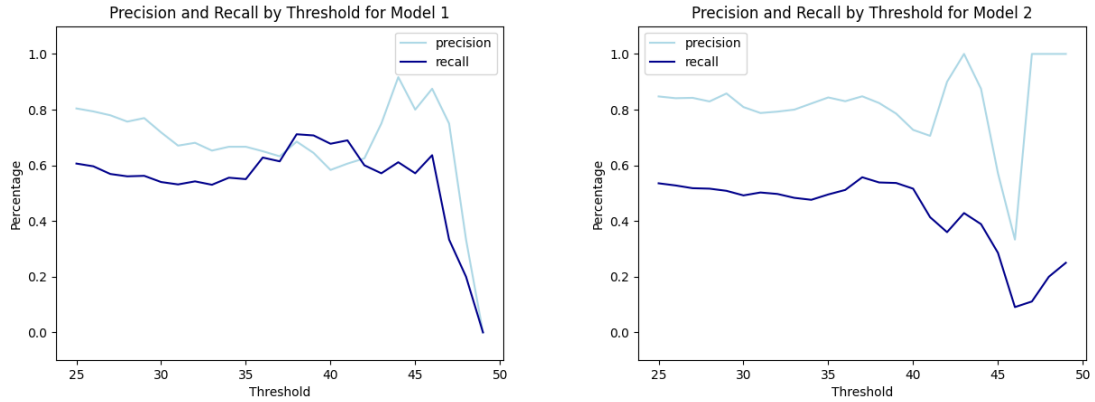


Figure 6: Model 1 (left) and Model 2 (right) precision and recall values as a function of thresholds determining positive (good) and negative (poor) cases of MOFs.

followed by an incline for Model 2; due to the small sample size of values above 45, these results are most likely not the most reliable, so we focus on the values below 45, particularly between 40 and 45 (this range has a higher sample size while still signaling high carbon capture ability). Due to the higher precision values in this range, we choose Model 2 as our final model.

The mean average error for the training and validation sets for Model 2 are shown in Figure 7. We see a potential overfit in the final epochs, so we choose the best model (lowest validation loss) as the final model. This was done for both Model 1 and Model 2, but we will only focus on Model 2 from now on.

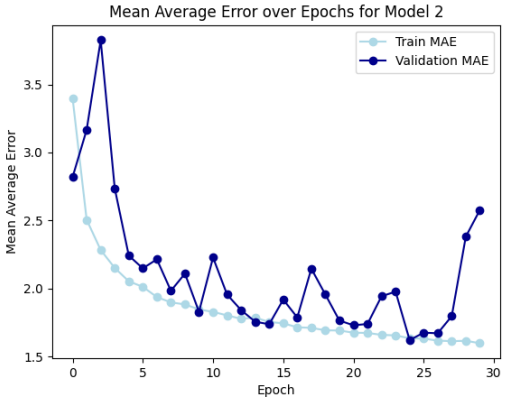
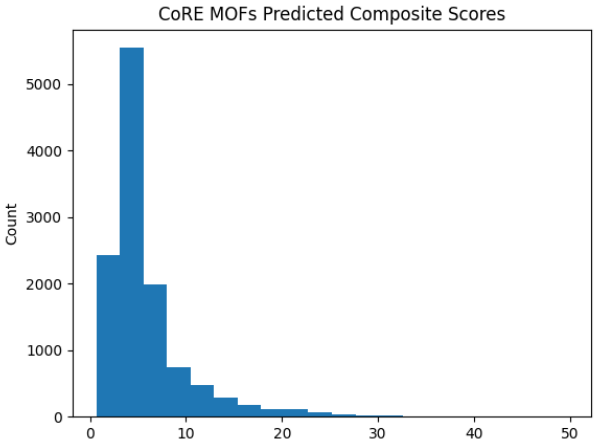


Figure 7: Model 2 Mean Average Error for both Training and Validation datasets over the training process.

5.2 CoRE Testing

The model's distribution of predicted composite scores for the CoRE dataset are shown in Figure 8. We see that this distribution follows the overall composite score distribution for the Boyd et al. dataset. This is a good check that ensures the CoRE dataset is not unexpectedly different than the BMC dataset in a way that would affect the model's ability to predict carbon capture capability. The MOFs with the highest predicted composite score are shown in Table 3 along with their predicted score. The top seven MOFs were identified, arbitrarily using a threshold of 40. Other thresholds could have been chosen, but based off of the precision and recall plots above, 40 seemed the most reasonable while also limiting the number of MOFs identified for further testing for financial and



MOF Name	Predicted Score
BIWSOQ	49.770844
PETXAP	48.537819
HETDUG	48.193043
PETWUI	45.564739
EDUTUS	45.353573
MOKVOZ	44.317356
FUYCAF	40.143955

Figure 8: CoRE Prediction Distribution. Table 3: Top CoRE MOFs based on their predicted score.

practical reasons discussed in the following section.

None of these MOFs were previously identified in other studies as being promising carbon capture candidates. This is partially the nature of the project; the goal is to identify MOFs based on their holistic ability to carbon capture, which could mean identifying novel MOFs as candidates for further experimentation.

5.3 Literature Comparison

In order to validate the model as being able to accurately predict an MOFs ability to capture carbon, we turn to the existing literature for MOF candidates already identified as promising. We can then input these candidates into the model to predict their composite score. Many studies have identified Mg-MOF-74 as having high carbon capture qualities [42, 49]. The model in this project predicted a composite score of 32.3203. This is fairly high and does increase our confidence that the model can correctly identify "good" MOFs. Other MOFs were passed through the model but did not achieve a high composite score and, similar to the literature surrounding these structures, did not appear to be as good of candidates for carbon capture and removal as Mg-MOF-74. This also reinforces the models' capability of rejecting structures which are not as holistically high performing as the top MOFs in the field.

6 Discussion

The work done here exemplifies the use of graph neural networks for high throughput screening of MOF candidates for holistic carbon capture; however, due to the limiting factors of this work being a one year project, there are several other directions this project could have gone and could be continued. This section will elaborate on 1) future work that can be done as a continuation of the work done here, 2) use cases for the existing model and contextualizing it in broader deep learning work, and 3) limitations of this work and what could be improved.

6.1 Future Work

The MOFs with the highest composite score predictions in the CoRE dataset were identified in Table 3. The next steps for determining if these MOFs would be exceptional candidates for carbon capture would be to run some type of simulation, e.g. the Grand Canonical Monte Carlo Simulations mentioned in Section 3. If these simulations gave promising results, such as high carbon adsorption, high CO_2/N_2 selectivity, high working capacity, etc., then the next step would be to run a physical experiment measuring carbon adsorption. These are clearly more expensive to actually facilitate due to the materials and financial resources needed but would provide chemical evidence for employing these MOFs in an industry setting. If proven to be better candidates for carbon capture than existing technologies, the hope would be that these MOFs could be incorporated into carbon reduction processes as soon as possible. The original goal of this project was to accelerate carbon removal in order to combat anthropogenic emissions and modern carbon concentration levels in the atmosphere; identifying and creating new MOFs that are particularly efficient in carbon capture could be one step to doing just that.

6.2 Model Use Cases

Now that this model is created, it can be used to screen potential MOF candidates for carbon capture applications. This is two-fold: it can be used for high-throughput screening given a large dataset of MOFs or it can be used to assess a small pool of candidates to validate existing or future research. The first is beneficial if given a very large dataset that is too computationally expensive

to use. By narrowing down this data to those with high predicted composite scores, a large dataset can become more manageable and allow researchers to focus on high performing MOFs. The second is also helpful if researchers need to validate their own results with another model. If a different project focuses on a specific attribute of carbon capture, e.g. selectivity, then this model can be used to validate that their results also are predictive of holistically capable MOFs (or alternatively demonstrate that their model only centers on one use case and is not meant for general application).

A related work, though much larger in scope, breadth, and expertise, is Google DeepMind's AlphaFold project [50]. This project aimed to predict the folding structure of proteins based solely on the sequence of amino acids it contains. It does so using a neural network based architecture in which it is able to generate the 3D structure of a protein, even for novel cases for which there is not pre-existing knowledge. This drastically decreases the time needed to determine protein structures and can have wide reaching effects for bioinformatics and modern biology more broadly. While this thesis project is much smaller, it aims to similarly harness the ability of deep learning architectures to quickly classify structures in a way that is unrealistic to do in physical experiments.

6.3 Limitations

Due to the limited time frame of a senior thesis, there were several directions this project could have evolved towards and was restricted in our exploration. Beyond future steps of running simulations and experiments as describe in a previous section, some primary limitations included data availability, skewness in the chosen data's distribution of carbon capture labels, and the limited manipulation of the chosen composite score. Each of these concerns are addressed in this section.

This project required 1) a fairly large dataset of MOFs with 2) carbon capture adjacent labels. The Boyd et al. dataset accounted for both of these necessities; however, other choices were available as described in the Background section and, if given a longer time frame, could have been explored. The disadvantages of using the Boyd et al. dataset was that all of the MOFs were hypothetical. They have not been physically manufactured and the labels were based on simulations rather than chemical experiments. This inherently poses a risk in that the training data for this model has not been physically realized and therefore the model could learn hypothetical relations about MOF structure that fail to translate to the real world. On the other hand, there are some datasets on

real MOFs along with labels concerning carbon capture, but they lack the quantity the hypothetical datasets are able to offer in terms of breadth of MOFs [39]. These trade offs are important to acknowledge and can help drive the end goal of a research project depending on which dataset is used. In the future, if new technology allows, researchers could consider creating an MOF database which contains a large amount of MOFs that have been experimentally realized and tested for various carbon capture attributes; this may be expensive and somewhat unrealistic, particularly in the short term, but such a database could accelerate MOF discovery for carbon capture.

The second limitation is the skewness present in MOFs as they relate to carbon capture. The goal of much of deep learning research on MOFs is to discover what makes an MOF particularly well able to capture carbon and which MOFs will be the most efficient and cost effective for industrial application. This is such a difficult yet interesting problem because most MOFs are not well adapted to capture carbon efficiently. This leads to the issue of skewness in the overall distribution of MOF carbon capture abilities. We see this in our data in the aforementioned figures: most of the carbon capture labels and the composite score distribution are relatively low with a right skewed tail containing exceptional candidates. This project tackled this skewed data by introducing a weighted loss function, putting higher emphasis on outliers above a threshold, but there are other techniques which could have been invoked. These include reducing the number of low scoring MOFs in the training set or artificially duplicating high scoring MOFs.

The third is the limited exploration of formulations for the carbon capture composite score. The one chosen in the project is a combination of various carbon-related attributes which prioritized carbon uptake, selectivity, and working capacity which together created a decent distribution for a training set for a deep learning model. Other composite scores were created, including a weighted average, different variables included, or different weights on each variable; however, the chosen composite score had the smoothest distribution which would best function in a machine learning context. The other issue with this score is the lack of physical meaning behind it; a higher score signifies better carbon capture abilities but it does not signify any physicality that could be validated by an experiment. In potential future work, this composite score could be created in tandem with chemical experts on crystal structures in order to best identify what should be included in the score and which factors should be weighted the highest in order to have both a holistic view of an MOFs

carbon capture ability as well as some physical significance to the prediction. Other work could also focus on creating a composite score that aligns with the priorities of those in industry, which may differ from academic experts. This could take into account the cost required to produce an MOF and its ability to be recycled over long periods of time.

6.4 Conclusion

Excess greenhouse gasses contribute heavily to climate change due to their ability to trap solar radiation, warming the Earth. A potential solution for this problem is Carbon Capture and Storage, the process of removing greenhouse gasses such as carbon dioxide from the atmosphere. One method for performing carbon capture is to use the cage-like structure of Metal Organic Frameworks to trap carbon dioxide molecules while letting other atmospheric gasses simply pass through. Machine learning, particularly methods which easily depict molecules such as Graph Neural Networks, have been used in recent years to find the most optimal MOFs for capturing carbon. This thesis utilizes a crystal graph neural network architecture developed by Xie and Grossman (2018) in tandem with a holistic composite score which measures an MOF's overall capability to perform carbon capture. The model was trained using Boyd et al.'s (2019) hypothetical MOF database and then used to search for the most effective MOFs in the Computation-Ready, Experimental (CoRE) MOF 2019 database. The results produced seven MOFs with high composite score predictions signaling strong carbon capture abilities. These candidates were then compared with the existing literature, including the widely recognized Mg-MOF-74 crystal structure; this validated the model's ability to identify high performing MOFs and reject low performing MOFs. The hope is that these candidates or this model can be used to identify MOFs for industrial applications of carbon capture, thereby removing CO_2 efficiently and helping reduce anthropogenic forcing on climate change.

References

[1] IPCC. *Climate change: a threat to human wellbeing and health of the planet. Taking action now can secure our future*. Last accessed 10 March 2025. Feb. 2022. URL: https://www.ipcc.ch/2022/02/28/pr-wgii-ar6/.

[2] James Hansen et al. "Assessing "Dangerous Climate Change": Required Reduction of Carbon Emissions to Protect Young People, Future Generations and Nature". In: *PLOS ONE* 8.12 (Dec. 2013), pp. 1–26. DOI: 10.1371/journal.pone.0081648. URL: https://doi.org/10.1371/journal.pone.0081648.

[3] United Nations. *Paris Agreement*. Last accessed 13 March 2025. 2015. URL: https://unfccc.int/sites/default/files/english_paris_agreement.pdf.

[4] United Nations. *New Analysis of National Climate Plans: Insufficient Progress Made, COP28 Must Set Stage for Immediate Action*. Last accessed 13 March 2025. Nov. 2023. URL: https://unfccc.int/news/new-analysis-of-national-climate-plans-insufficient-progress-made-cop28-must-set-stage-for-immediate.

[5] United Nations. *Technology and the UNFCCC*. Last accessed 13 March 2025. 2016. URL: https://unfccc.int/news/technology-and-the-unfccc.

[6] Wai Lip Theo et al. "Review of pre-combustion capture and ionic liquid in carbon capture and storage". In: *Applied Energy* 183 (2016), pp. 1633–1663. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2016.09.103. URL: https://www.sciencedirect.com/science/article/pii/S0306261916314052.

[7] Tabbi Wilberforce et al. "Progress in carbon capture technologies". In: *Science of The Total Environment* 761 (2021), p. 143203. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2020.143203. URL: https://www.sciencedirect.com/science/article/pii/S0048969720367346.

[8] "CO2 Transport Systems Development: Status of Three Large European CCS Demonstration Projects with EEPR Funding". In: *Energy Procedia* 63 (2014). 12th International Conference on Greenhouse Gas Control Technologies, GHGT-12, pp. 2458–2466. ISSN: 1876-6102. DOI:

https://doi.org/10.1016/j.egypro.2014.11.268. URL: https://www.sciencedirect.com/science/article/pii/S1876610214020839.

[9] "Recent progress and new developments in post-combustion carbon-capture technology with amine based solvents". In: *International Journal of Greenhouse Gas Control* 40 (2015). Special Issue commemorating the 10th year anniversary of the publication of the Intergovernmental Panel on Climate Change Special Report on CO2 Capture and Storage, pp. 26–54. ISSN: 1750-5836. DOI: https://doi.org/10.1016/j.ijggc.2015.06.017. URL: https://www.sciencedirect.com/science/article/pii/S1750583615002704.

[10] "CO2 capture from pre-combustion processes—Strategies for membrane gas separation". In: *International Journal of Greenhouse Gas Control* 4.5 (2010), pp. 739–755. ISSN: 1750-5836. DOI: https://doi.org/10.1016/j.ijggc.2010.04.001. URL: https://www.sciencedirect.com/science/article/pii/S1750583610000472.

[11] Catherine P. Raptopoulou. "Metal-Organic Frameworks: Synthetic Methods and Potential Applications". In: *Materials* 14.2 (2021). ISSN: 1996-1944. DOI: 10.3390/ma14020310. URL: https://www.mdpi.com/1996-1944/14/2/310.

[12] L. Li S. Gupta. "The Potential of Machine Learning for Enhancing CO2 Sequestration, Storage, Transportation, and Utilization-based Processes: A Brief Perspective." In: *JOM* 74 (Feb. 2022). DOI: https://doi.org/10.1007/s11837-021-05079-x. URL: https://link.springer.com/article/10.1007/s11837-021-05079-x#citeas.

[13] C. Trickett et al. "The chemistry of metal–organic frameworks for CO2 capture, regeneration and conversion." In: *Nat Rev Mater* (2017). DOI: https://doi.org/10.1038/natrevmats.2017.45. URL: https://www.nature.com/articles/natrevmats201745.

[14] Zikri Bayraktar, Shahnawaz Molla, and Sharath Mahavadi. "Graph Neural Network Generated Metal-Organic Frameworks for Carbon Capture". In: *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*. 2023. URL: https://www.climatechange.ai/papers/iclr2023/12.

[15] C et al. Zhang. "Accelerated discovery of porous materials for carbon capture by machine learning: A review." In: *MRS Bulletin* 47 (2022). DOI: https://doi.org/10.1557/s43577-022-00317-2.

[16] H. et al. Park. "A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture." In: *Commun Chem* 7 (2024). DOI: https://doi.org/10.1038/s42004-023-01090-2.

[17] Z. et al. Yao. "Inverse design of nanoporous crystalline reticular materials with deep generative models." In: *Nat Mach Intell* 3 (2021). DOI: https://doi.org/10.1038/s42256-020-00271-1.

[18] Song Li, Yongchul G. Chung, and Randall Q. Snurr. "High-Throughput Screening of Metal Organic Frameworks for CO2 Capture in the Presence of Water". In: *Langmuir* 32.40 (2016). PMID: 27627635, pp. 10368–10376. DOI: 10.1021/acs.langmuir.6b02803. URL: https://doi.org/10.1021/acs.langmuir.6b02803.

[19] Cigdem Altintas et al. "An extensive comparative analysis of two MOF databases: high-throughput screening of computation-ready MOFs for CH4 and H2 adsorption". In: *J. Mater. Chem. A* 7 (16 2019), pp. 9593–9608. DOI: 10.1039/C9TA01378D. URL: http://dx.doi.org/10.1039/C9TA01378D.

[20] Hana Dureckova et al. "Robust Machine Learning Models for Predicting High CO2 Working Capacity and CO2/H2 Selectivity of Gas Adsorption in Metal Organic Frameworks for Precombustion Carbon Capture". In: *The Journal of Physical Chemistry C* 123.7 (2019), pp. 4133–4139. DOI: 10.1021/acs.jpcc.8b10644. URL: https://doi.org/10.1021/acs.jpcc.8b10644.

[21] Maryam Pardakhti et al. "Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs)". In: *ACS Combinatorial Science* 19.10 (2017). PMID: 28800219, pp. 640–645. DOI: 10.1021/acscombsci.7b00056. eprint: https://doi.org/10.1021/acscombsci.7b00056. URL: https://doi.org/10.1021/acscombsci.7b00056.

[22] Jake Burner et al. “High-Performing Deep Learning Regression Models for Predicting Low-Pressure CO2 Adsorption Properties of Metal–Organic Frameworks”. In: *The Journal of Physical Chemistry C* 124.51 (2020), pp. 27996–28005. DOI: 10.1021/acs.jpcc.0c06334. eprint: https://doi.org/10.1021/acs.jpcc.0c06334. URL: https://doi.org/10.1021/acs.jpcc.0c06334.

[23] Zihao Zhang et al. “Prediction of Carbon Dioxide Adsorption via Deep Learning”. In: *Angewandte Chemie International Edition* 58.1 (2019), pp. 259–263. DOI: https://doi.org/10.1002/anie.201812363. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201812363. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201812363.

[24] Tian Xie and Jeffrey C. Grossman. “Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties”. In: *Phys. Rev. Lett.* 120 (14 Apr. 2018), p. 145301. DOI: 10.1103/PhysRevLett.120.145301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

[25] N. Scott Bobbitt et al. “MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials”. In: *Journal of Chemical & Engineering Data* 68.2 (2023), pp. 483–498. DOI: 10.1021/acs.jced.2c00583. eprint: https://doi.org/10.1021/acs.jced.2c00583. URL: https://doi.org/10.1021/acs.jced.2c00583.

[26] Yongchul G. Chung et al. “Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019”. In: *Journal of Chemical & Engineering Data* 64.12 (2019), pp. 5985–5998. DOI: 10.1021/acs.jced.9b00835. eprint: https://doi.org/10.1021/acs.jced.9b00835. URL: https://doi.org/10.1021/acs.jced.9b00835.

[27] Eric L. First and Christodoulos A. Floudas. “MOFomics: Computational pore characterization of metal–organic frameworks”. In: *Microporous and Mesoporous Materials* 165 (2013), pp. 32–39. ISSN: 1387-1811. DOI: https://doi.org/10.1016/j.micromeso.2012.07.049. URL: https://www.sciencedirect.com/science/article/pii/S1387181112004817.

[28] Michael O'Keeffe et al. "The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets". In: *Accounts of Chemical Research* 41.12 (2008). PMID: 18834152, pp. 1782–1789. DOI: 10.1021/ar800124u. eprint: https://doi.org/10.1021/ar800124u. URL: https://doi.org/10.1021/ar800124u.

[29] Benjamin J. Bucior et al. "Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis". In: *Crystal Growth & Design* 19.11 (2019), pp. 6682–6697. DOI: 10.1021/acs.cgd.9b01050. eprint: https://doi.org/10.1021/acs.cgd.9b01050. URL: https://doi.org/10.1021/acs.cgd.9b01050.

[30] Boran Li et al. "Understanding the CO2/CH4/N2 Separation Performance of Nanoporous Amorphous N-Doped Carbon Combined Hybrid Monte Carlo with Machine Learning". In: *Advanced Theory and Simulations* 5.1 (2022), p. 2100378. DOI: https://doi.org/10.1002/adts.202100378. eprint: https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/adts.202100378. URL: https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adts.202100378.

[31] Xiangyu Zhang et al. "Machine Learning-Driven Discovery of Metal–Organic Frameworks for Efficient CO2 Capture in Humid Condition". In: *ACS Sustainable Chemistry & Engineering* 9.7 (2021), pp. 2872–2879. DOI: 10.1021/acssuschemeng.0c08806. eprint: https://doi.org/10.1021/acssuschemeng.0c08806. URL: https://doi.org/10.1021/acssuschemeng.0c08806.

[32] Song Wang et al. "Insights into CO2/N2 Selectivity in Porous Carbons from Deep Learning". In: *ACS Materials Letters* 1.5 (2019), pp. 558–563. DOI: 10.1021/acsmaterialslett.9b00374. eprint: https://doi.org/10.1021/acsmaterialslett.9b00374. URL: https://doi.org/10.1021/acsmaterialslett.9b00374.

[33] I.D. Brown. "CIF (Crystallographic Information File): A Standard for Crystallographic Data Interchange". In: *J Res Natl Inst Stand Technol.* 101.3 (1996), pp. 341–346. DOI: doi:10.6028/jres.101.035. URL: https://doi.org/10.6028/jres.101.035.

[34] R. Aniruddha, I. Sreedhar, and Benjaram M. Reddy. "MOFs in carbon capture-past, present and future". In: *Journal of CO2 Utilization* 42 (2020), p. 101297. ISSN: 2212-9820. DOI: https:

//doi.org/10.1016/j.jcou.2020.101297. URL: https://www.sciencedirect.com/science/article/pii/S2212982020306338.

[35] Sumida et al. "Carbon dioxide capture in metal-organic frameworks." In: *Chem Rev* 112 (2012), pp. 724–81. DOI: 10.1021/cr2003272. URL: https://pubmed.ncbi.nlm.nih.gov/22204561/.

[36] Hong Xu et al. "Machine learning-assisted high-throughput screening of MOFs for efficient adsorption and separation of CF4/N2". In: *Journal of Cleaner Production* 461 (2024), p. 142634. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2024.142634. URL: https://www.sciencedirect.com/science/article/pii/S0959652624020821.

[37] Xiangyu Zhang, Kexin Zhang, and Yongjin Lee. "Machine Learning Enabled Tailor-Made Design of Application-Specific Metal–Organic Frameworks". In: *ACS Applied Materials & Interfaces* 12.1 (2020). PMID: 31820913, pp. 734–743. DOI: 10.1021/acsami.9b17867. eprint: https://doi.org/10.1021/acsami.9b17867. URL: https://doi.org/10.1021/acsami.9b17867.

[38] Donald E. Ellis David Dubbeldam Sofía Calero and Randall Q. Snurr. "RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials". In: *Molecular Simulation* 42.2 (2016), pp. 81–101. DOI: 10.1080/08927022.2015.1010082. eprint: https://doi.org/10.1080/08927022.2015.1010082. URL: https://doi.org/10.1080/08927022.2015.1010082.

[39] Cigdem Altintas et al. "Database for CO2 Separation Performances of MOFs Based on Computational Materials Screening". In: *ACS Applied Materials & Interfaces* 10.20 (2018). PMID: 29722965, pp. 17257–17268. DOI: 10.1021/acsami.8b04600. eprint: https://doi.org/10.1021/acsami.8b04600. URL: https://doi.org/10.1021/acsami.8b04600.

[40] Ramsahye et al. "Adsorption of CO2 in metal organic frameworks of different metal centres: Grand Canonical Monte Carlo simulations compared to experiments." In: *Adsorption* 13 (2007), pp. 461–467. URL: https://doi.org/10.1007/s10450-007-9025-5.

[41] Xin et al. "Functionalized linker to form high-symmetry adsorption sites in micropore COF for CO2 capture and separation: insight from GCMC simulations." In: *J Mater Sci* 57 (2022), pp. 6282–6292. URL: https://doi.org/10.1007/s10853-022-07008-y.

[42] Guanhua Zhang et al. “Grand canonical Monte Carlo simulation of the adsorption and separation of carbon dioxide and methane using functionalized Mg-MOF-74”. In: *Energy Reports* 9 (2023), pp. 2852–2860. ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2023.01.121. URL: https://www.sciencedirect.com/science/article/pii/S2352484723001300.

[43] Orhan et al. “Accelerating the prediction of CO2 capture at low partial pressures in metal-organic frameworks using new machine learning descriptors.” In: *Commun Chem* 6 (214 2023). URL: https://doi.org/10.1038/s42004-023-01009-x.

[44] Boyd et al. “Data-driven design of metal–organic frameworks for wet flue gas CO2 capture.” In: *Nature* 576 (2019), pp. 253–256. URL: https://doi.org/10.1038/s41586-019-1798-7.

[45] Peter G. Boyd and Tom K. Woo. “A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory”. In: *CrystEngComm* 18 (21 2016), pp. 3777–3792. DOI: 10.1039/C6CE00407E. URL: http://dx.doi.org/10.1039/C6CE00407E.

[46] Hakan Demir and Seda Keskin. “Hypothetical yet effective: Computational identification of high-performing MOFs for CO2 capture”. In: *Computers and Chemical Engineering* 160 (2022), p. 107705. ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2022.107705. URL: https://www.sciencedirect.com/science/article/pii/S0098135422000485.

[47] Federica Raganati, Francesco Miccio, and Paola Ammendola. “Adsorption of Carbon Dioxide for Post-combustion Capture: A Review”. In: *Energy & Fuels* 35.16 (2021), pp. 12845–12868. DOI: 10.1021/acs.energyfuels.1c01618. eprint: https://doi.org/10.1021/acs.energyfuels.1c01618. URL: https://doi.org/10.1021/acs.energyfuels.1c01618.

[48] Ann-Kathrin Schalkamp et al. “Machine learning-based personalized composite score dissects risk and protective factors for cognitive and motor function in older participants”. In: *Frontiers in Aging Neuroscience* 16 (2024). ISSN: 1663-4365. DOI: 10.3389/fnagi.2024.1447944. URL: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2024.1447944.

[49] David Britt et al. “Highly efficient separation of carbon dioxide by a metal-organic framework replete with open metal sites”. In: *Proceedings of the National Academy of Sciences* 106.49 (2009), pp. 20637–20640. DOI: 10.1073/pnas.0909718106. eprint: https://www.pnas.org/

doi/pdf/10.1073/pnas.0909718106. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0909718106.

[50] J. et al. Jumper. "Highly accurate protein structure prediction with AlphaFold." In: *Nature* 596 (2021), pp. 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2.

AI acknowledgment: minimal general background research at the beginning of this project included questions to LLMs about graph neural networks, MOF and carbon capture applications, and potential Python libraries/packages for working with .cif files. These outputs were not used in the coding or writing of this project.