Carbon Capture Capstone

Designing a MOF to clean power plant exhaust

By Zachary Brown

Abstract

Power generation is one of the leading sources of greenhouse gas emissions and a key driver of climate change. Metal-organic frameworks (MOFs) are nanoporous materials that are great candidates for carbon capture and sequestration processes due to their highly tunable chemical and structural properties. My goal in this project was to use the ARC-MOF database of calculated MOF properties to generate a predictive model that can help steer chemists in the lab towards developing the best possible carbon capture material. After screening a range of regression models Light GBM was identified as the best performing model for this project. The final Light GBM regressor was trained on 80% of the available data and had root mean squared error of 22.83 when tested on the remaining 20% of the dataset. The four most important features to maximize volumetric working capacity were to target a range of 0-10 angstoms for the pore limiting diameter and largest cavity diameter, minimize unit cell volume, and to target a probe accessible volume fraction of 0.4 to 0.5. In the future I would like to collaborate with a research group working on this problem to dig deeper into this dataset and attempt to synthesize new MOFs targeting these feature values.

Introduction

Power generation is one of the most pressing challenges in dealing with climate change. While there is a widespread push to transition to clean energy, natural gas and coal fired power plants are still being built and operated around the world. One strategy to reduce the ecological impact of these power plants is to capture carbon dioxide as it's emitted and then sequester it underground. This carbon capture and sequestration (CCS) strategy requires a material with both a large CO_2 capacity, selectivity for CO_2 over water, and binding strength that is tuned to hold on to CO_2 at atmospheric pressure, but release it under vacuum so the material can be reused.

Metal-organic frameworks (MOFs) are nanoporous materials that are great candidates for this type of challenge because they can be designed with a near-endless variety of organic linkers that bridge metal oxide nodes in various geometries (Figure 1). When these MOFs are assembled there are a range of resulting geometric properties that play a role in the capacity of the MOF to function as a CCS material such as pore diameters, cavity diameters, accessible and inaccessible surface area, metal charge and exposure, and density.

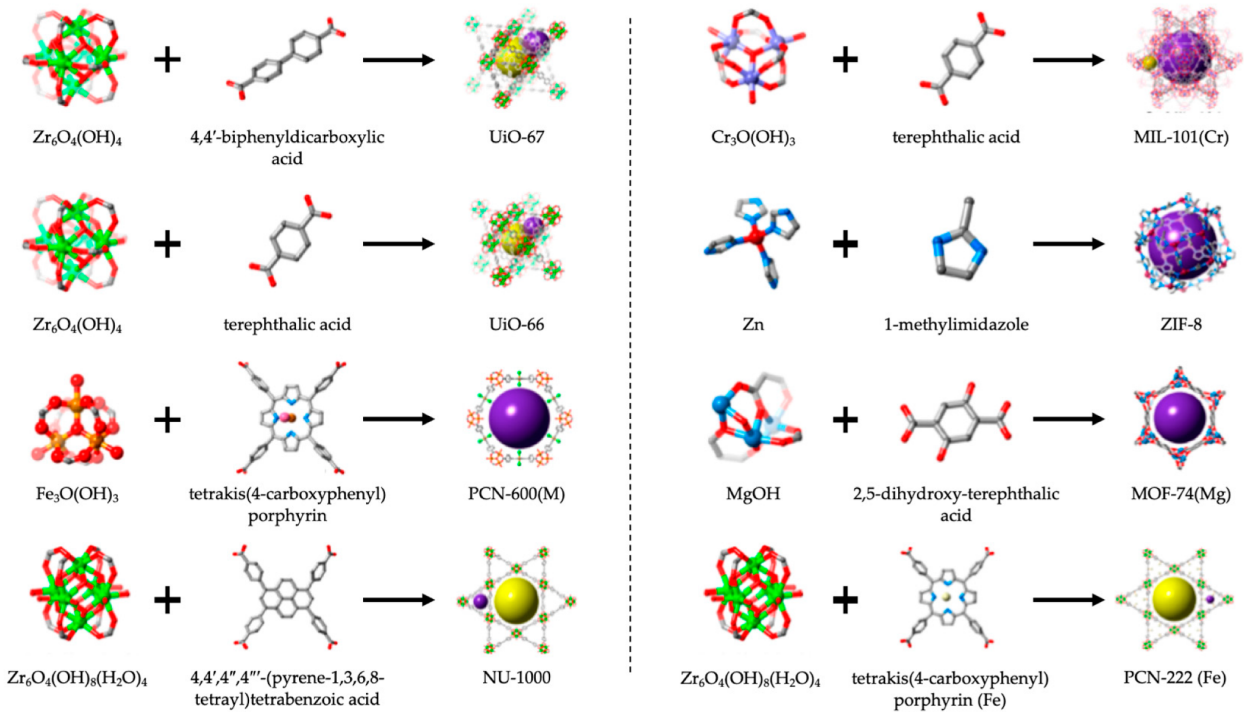

Figure 1. A wide range of linker and metal combinations are shown along with the resulting MOF structure.
Source: www.mdpi.com

Computational chemists have developed multiple large databases, each comprising thousands of experimental and theoretical MOFs along with wide ranges of properties that may be helpful in predicting their efficacy as a CCS material. The computed properties of these MOFs can then be used to predict the amount of a gas adsorbed by that MOF under a given situation. In the case of CCS the MOF is exposed to gas composed of anywhere from 4 - 17% CO_2, some amount of water, and the rest is primarily N_2. A major challenge in this process is maximizing selectivity so that the MOF absorbs CO_2 rather than water, while ensuring that the binding strength to CO_2 is low enough that it can be removed under vacuum under a process called vacuum swing adsorption (VSA). Designing the perfect MOF which meets all of these requirements has proven elusive, so my goal in this project is to use the ARC-MOF database prepared by Burner et. al. to determine how best to design the perfect CCS MOF.

Methodology

Loading the data

The ARC-MOF database consists of multiple datasets in .csv format, each with a different set of properties for the MOFs included. For this study I imported and combined the topology, geometry, process, radial distribution function (RDF), and revised autocorrelation (RAC) datasets, joining each on the MOF name. The topology set contained only MOF topologies, which describe the structure of the MOF using three letter abbreviations. The geometry dataset

contains geometric properties of the MOFs such as pore diameter, cavity diameter, surface area, etc. The process dataset lists gas separation properties for each MOF such as gravimetric and volumetric capacity for the target, selectivity for the target over other blended gases, and working capacity. The RDF dataset contained RDFs calculated for electronegativity, atomic hardness, van der Waals, mass, and none at various distances from the atom centers. A radial distribution function describes the property of interest at a given distance from the atom center. Electronegativity is a measure of how strongly the nucleus of an atom pulls electrons towards itself. Atomic hardness describes the reactivity and stability of an atom, and van der Waals forces are described in the Insights and Performance section of this report. The final dataset included RACs which are calculated properties based on features such as metal identity, linker identity, etc. After evaluating the adsorption properties available in the combined dataset I decided to use volumetric CO_2 capacity as the target metric for the project. Volumetric working capacity is defined as the volume of CO_2 that can be adsorbed and then desorbed repeatedly from some volume of MOF, which fits this situation since space in an exhaust stack is limited and it would be ideal to pack as much CO_2 into as small a volume as possible. The distribution for this metric is shown below in Figure 2.

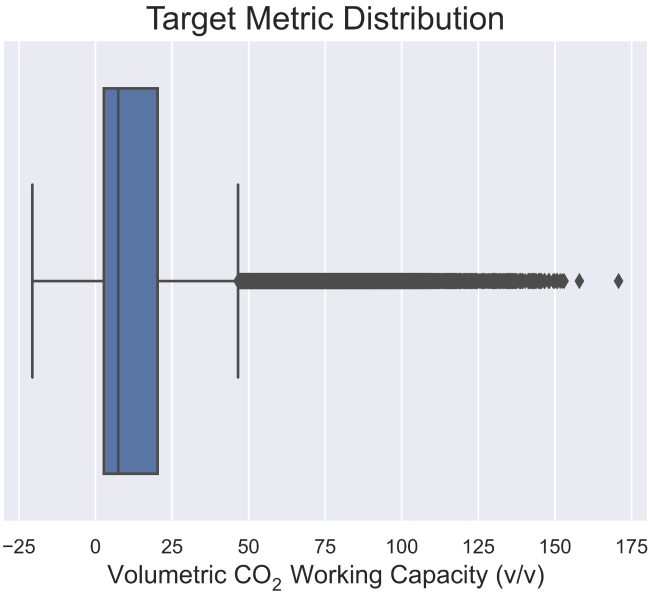

Figure 2. The distribution of volumetric CO_2 working capacity is shown including outliers to show the absolute highest predicted capacity.

To conclude my data wrangling I removed any rows missing this value, dropped features that were constant, and dropped others that were unrelated to this project.

Exploratory Data Analysis

To begin my analysis I focused on a range of geometric properties including density, pore limiting diameter, largest cavity diameter, gravimetric surface area, and volumetric surface area. A scatterplot was prepared for each with an ordinary least squares (OLS) overlay to quickly identify any positive or negative correlations with CO_2 working capacity (Figure 3).

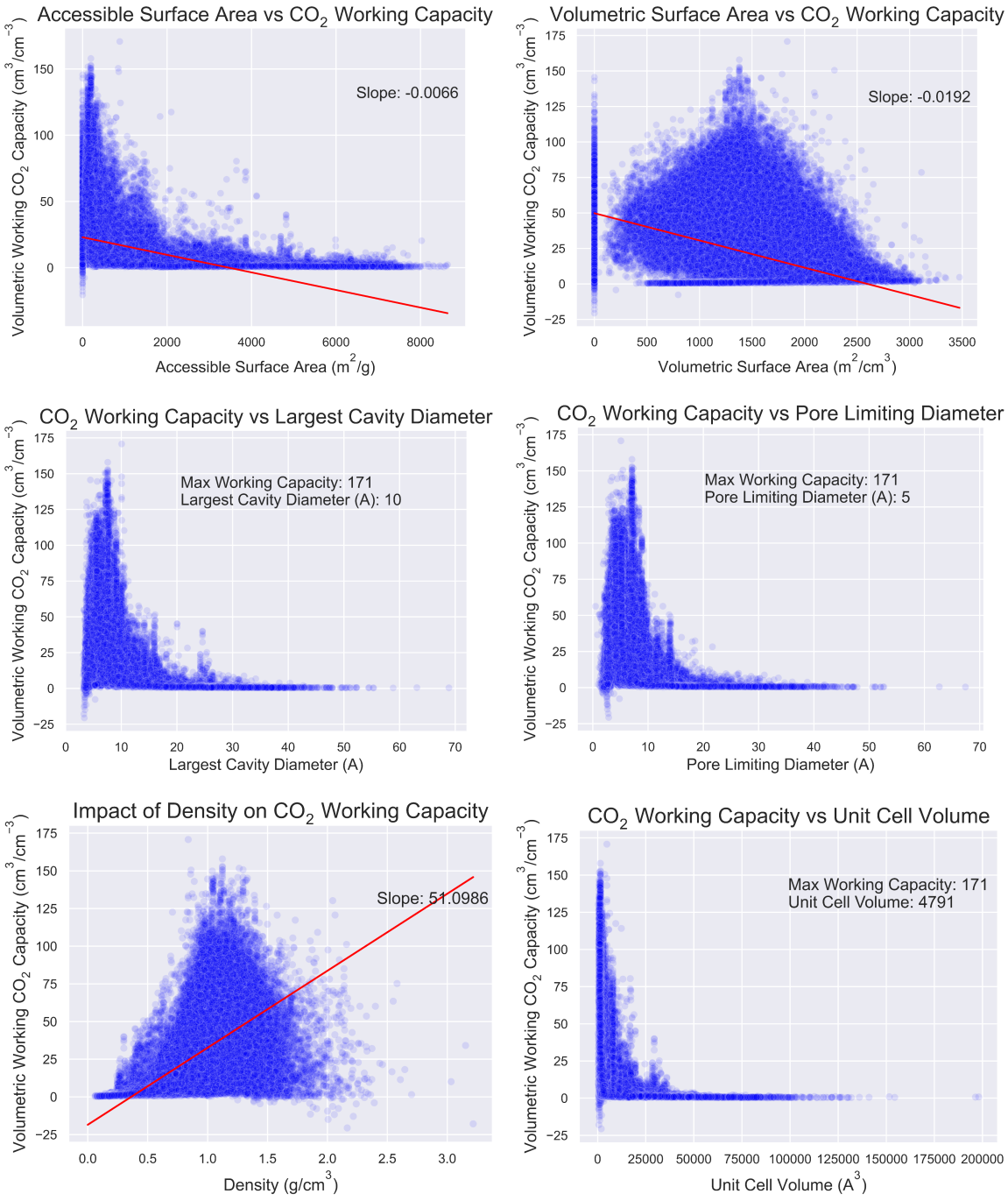

Figure 3. Correlations between volumetric CO_2 working capacity and accessible surface area, volumetric surface area, largest cavity diameter, pore limiting diameter, density, and unit cell volume are plotted with trend lines overlaid.

Some of these features needed to be transformed to achieve linear relationships, so in some cases the log of volumetric working capacity was taken, and in one case the inverse of unit cell volume was plotted against working capacity (Figure 4).

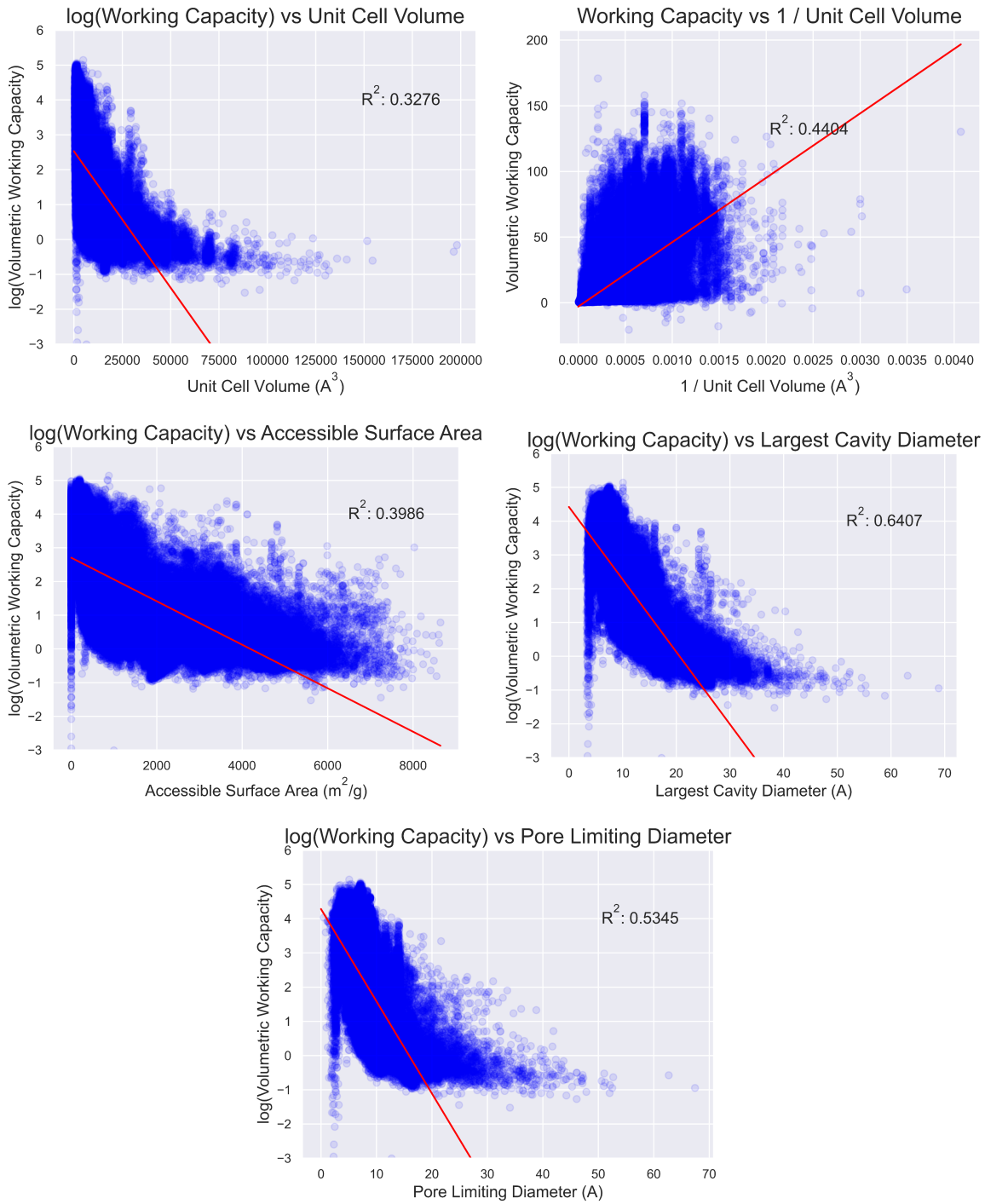

Figure 4. Correlations are plotted between transformed data in effort to demonstrate linear relationships between independent and dependent features.

For each RDF I identified whether the data was normal by performing a Shapiro-Wilk test. If the test result was above 0.05, suggesting a normal distribution, I performed a Pearson correlation test to calculate the correlation coefficient. If the p-value of that coefficient was less than 0.05 then I added it to a plot of the RDF correlation with volumetric CO_2 working capacity at each

distance. The plots for each RDF feature were overlaid below in Figure 5 showing which distances had the strongest positive and negative correlations.

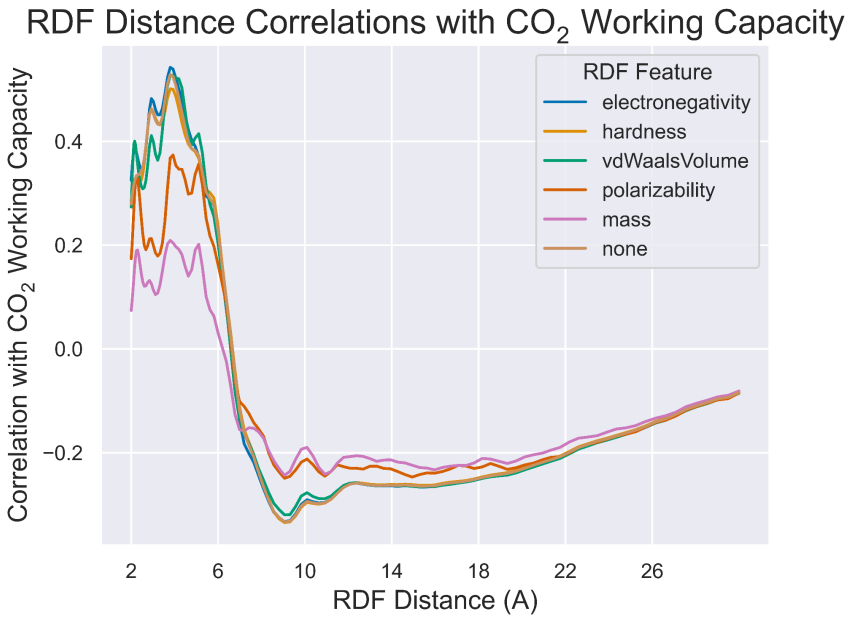

Figure 5. The correlations between each RDF feature and volumetric CO_2 working capacity are plotted against the RDF distance to identify the distances with the strongest positive and negative correlations.

After evaluating the RDFs, I calculated the average CO_2 working capacity for each MOF topology. I then reduced the list of topologies to only those with at least 20 MOFs, sorted by working capacity, and plotted the distributions of the top 10 (Figure 6).

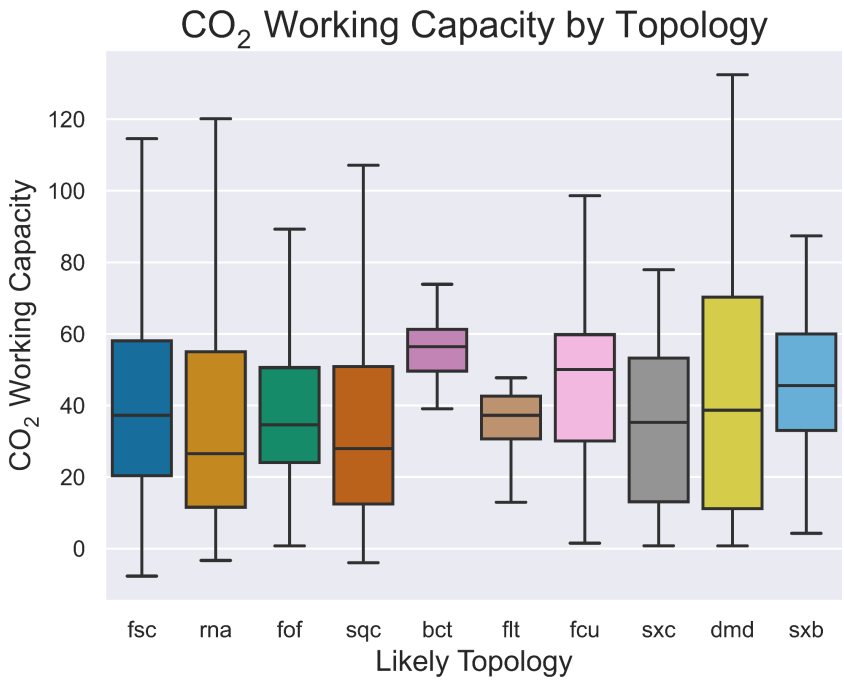

Figure 6. Distributions of the 10 topologies with highest average volumetric CO_2 working capacity and at least 20 entries in the dataset are plotted.

Although most had distributions covering almost the entire range of working capacity, bct had a much tighter distribution with a higher average than the others, indicating that it might be a topology worth targeting in MOF design.

For my final step in exploratory analysis I calculated Pearson correlation coefficients for all of the remaining features in the dataset that hadn't been analyzed to this point. I noticed gaps around 0.5 and -0.5, so I created a list of all features with correlation coefficients below -0.5 or above 0.5 to see which looked most likely to be significant. What I found was that most of the features with high correlations were features that were either a metric of how the MOF performs in this situation (adsorption_fig_of_merit) or features that are also a consequence of the MOF design, rather than something the scientist could control up front, such as gas selectivity or various measures of uptake. These were dropped from the dataset to ensure that any model could only use features that a scientist could design into their MOF.

Feature Engineering

To prepare the data for modeling I began by dropping unnecessary or unexplained columns (unfortunately this ARC-MOF database doesn't have great metadata explaining all of the acronyms or feature meanings). At this point I chose to work with just the 'likely topology' column so I dropped the other topology columns and then created dummies, dropping the first to avoid collinearity. I then followed the procedure shown below in Figure 7.

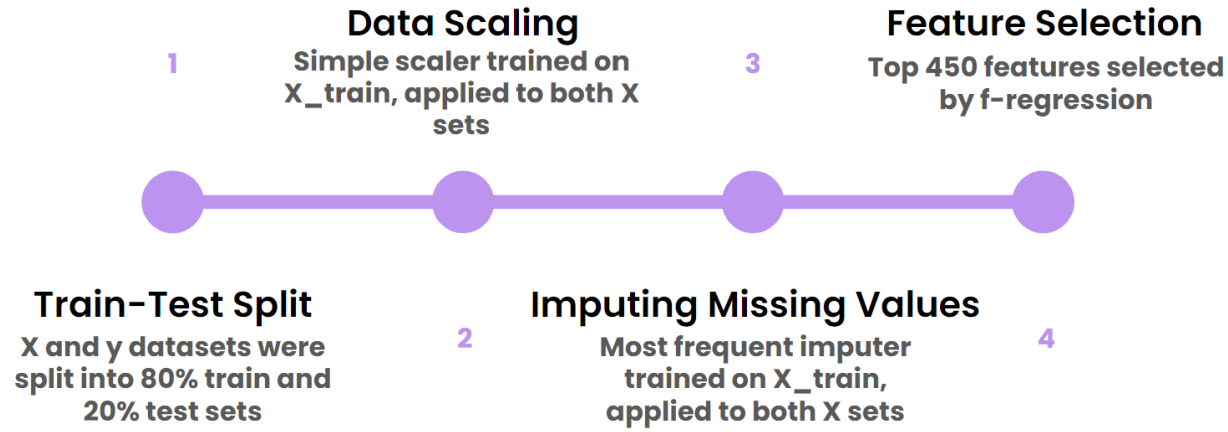

Figure 7. The feature engineering process is shown from left to right.

I began by splitting the target feature, volumetric CO_2 working capacity, from the rest of the dataset. I then divided the datasets into train and test sets, reserving 20% of the data for final testing. Next I used a simple scaler to scale the entire dataset so that each feature was weighted equally in the model development process. I then used a simple imputer to fill any missing values with the most common value for that feature. Scaling and imputing were both trained on the X_train set, and were then applied to the X_train and X_test sets to avoid data

leakage. Finally, feature selection was performed using f-regression to reduce the dataset to only the 450 most useful features.

Modeling

To begin modeling I split the X_train and y_train sets into 10% training and 90% testing subsets. I trained mean and median dummy models along with OLS, ridge, lasso, random forest, decision tree, XGBoost regressor, XGBoost linear regressor, XGBoost linear coordinate descent, and light GBM models, all using default hyperparameter settings and only limiting max depth for tree-based models to a depth of 10. I had each predict the 90% test set and calculated the root mean squared error (RMSE) of the predictions against the actual data. These RMSE values were then plotted to create a quick visualization of which models performed the best (Figure 8).

Figure 8. The RMSE values of the initial round of models using default hyperparameters and trained on 10% of the train data are plotted.

After scoring models that had been trained on 10% of the data, the same process was repeated, this time using 80% of the training data and reserving only 20% for testing. The RMSE scores for these were then plotted alongside the original 10% train models (Figure 9), confirming that

the best models for this project were random forest, XGBoost, Light GBM, KNN, and decision tree.

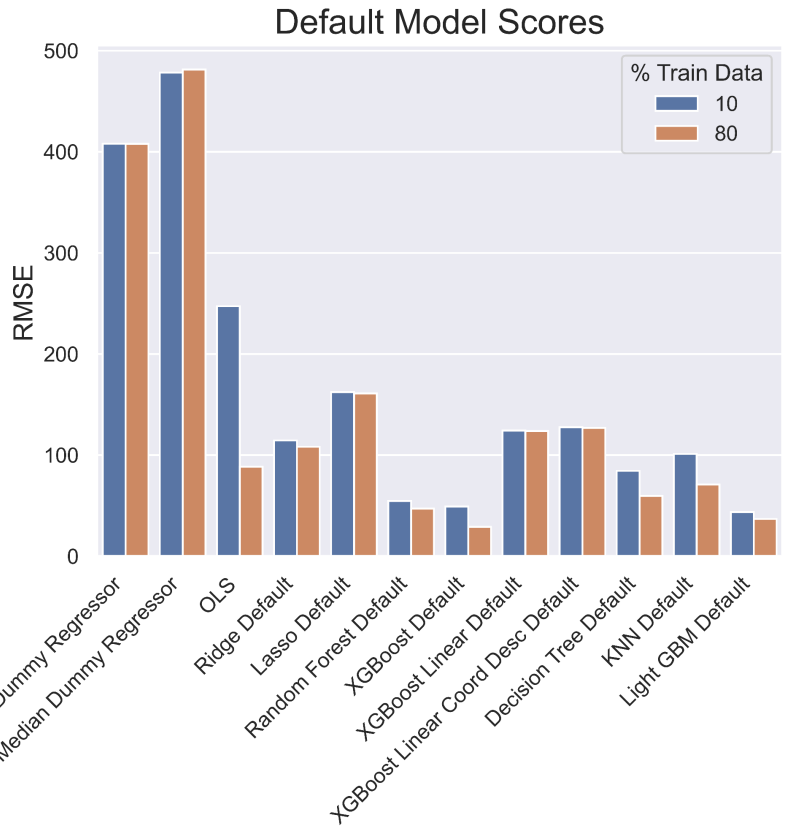

Figure 9. The RMSE values of the first two rounds of models using default hyperparameters are plotted after being trained on both 10 and 80 percent of the training data..

Next I began hyperparameter tuning the aforementioned top performing models using random search CV with 10% train data, five folds, and ten iterations each, starting with the hyperparameters shown in Table 1.

Table 1. The hyperparameters that were tuned in this first stage of model refinement are listed for each model still in consideration.

Model:	**Random Forest**	**XGBoost**	**Decision Tree**	**KNN**	**Light GBM**
Hyperparameters:	-Max features, -Max depth, -Min samples per leaf, -Number of estimators	-% of columns used per tree, -ETA, -Max depth, -Number of estimators	-Max depth, -Min samples per leaf	-Number of neighbors	-Number of leaves, -Max depth, -Learning rate, -Number of estimators

The RMSE for these tuned models were plotted below in Figure 10.

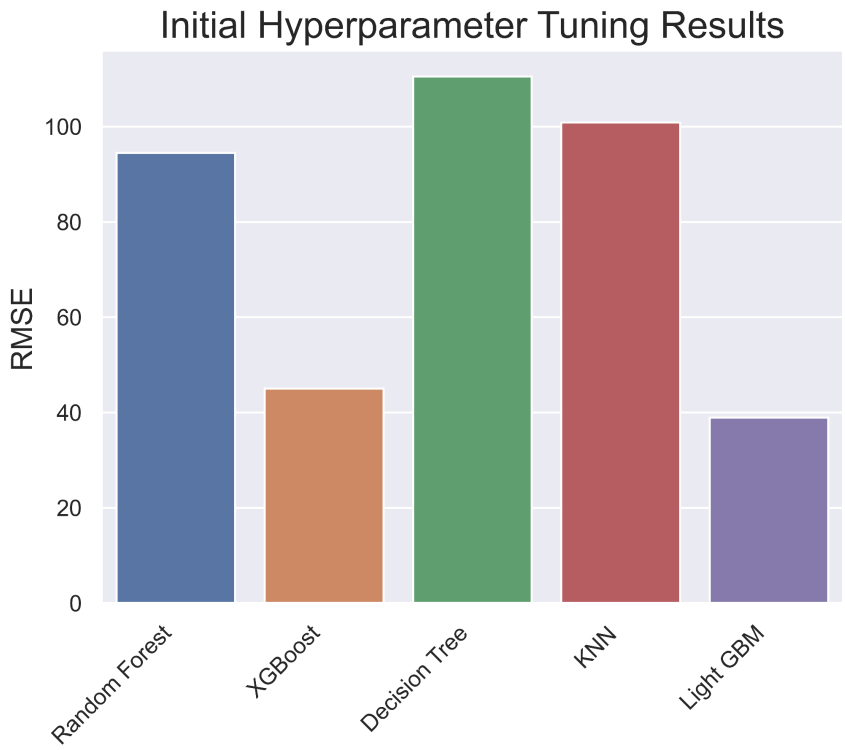

Figure 10. The RMSE of the five hypertuned models are plotted after being trained on 10% of the training data..

As Figure 10 illustrates, XGBoost and Light GBM continue to outperform all of the other models with only a small gap between them. I trained both on 80% of the data and recalculated the RMSE of each against the 20% test data and Light GBM beat out XGBoost by a small margin, 23.76 vs 32.70. To complete the Light GBM optimization, I began by plotting the RMSE against each hyperparameter tested in the first round of hyperparameter tuning (Figure 11).

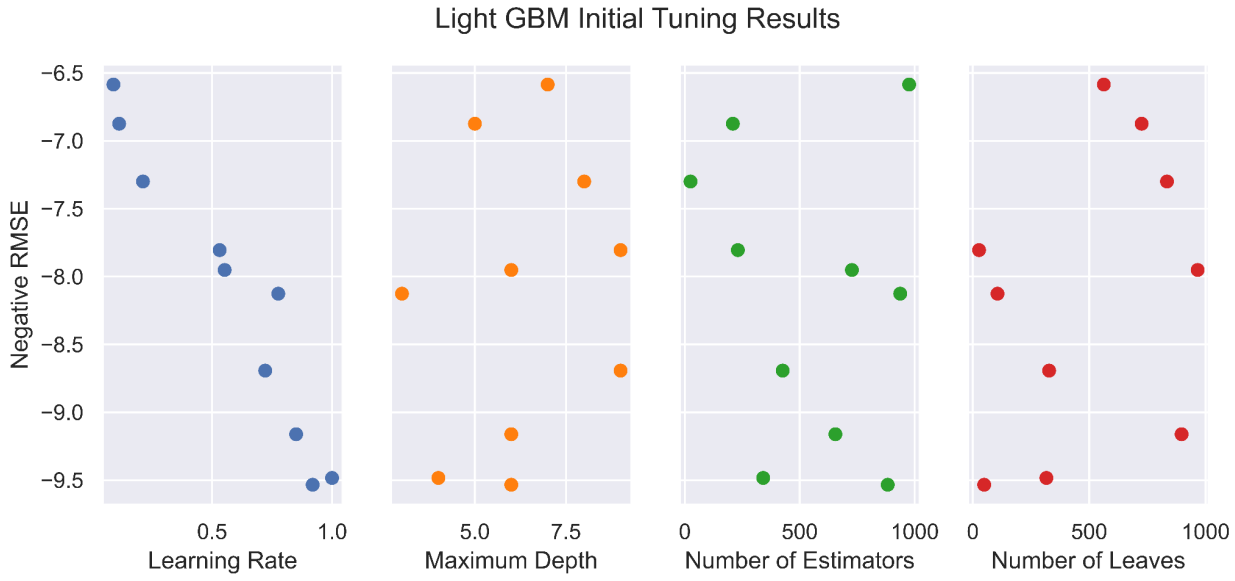

Figure 11. The initial RMSE results are plotted for each hyperparameter that was tuned for the Light GBM model in the first round of hyperparameter tuning.

These plots show that maximum depth and number of estimators had relatively little impact on RMSE, however, learning rate and number of leaves could potentially be improved. A second round of hyperparameter tuning was carried out narrowing the range for learning rate to 0.00001 through 0.1, and narrowing the number of leaves to 450 through 800. Random search CV was used again, training on 80% of the data, testing 10 iterations and five folds. The resulting fully optimized model had 23.65 RMSE, which was a very minor improvement from the prior model. The hyperparameters from the final tuning were then locked for the remainder of the project since this was the best model created. Figure 12 shows the progression of the RMSE as the Light GBM model was improved throughout the project.

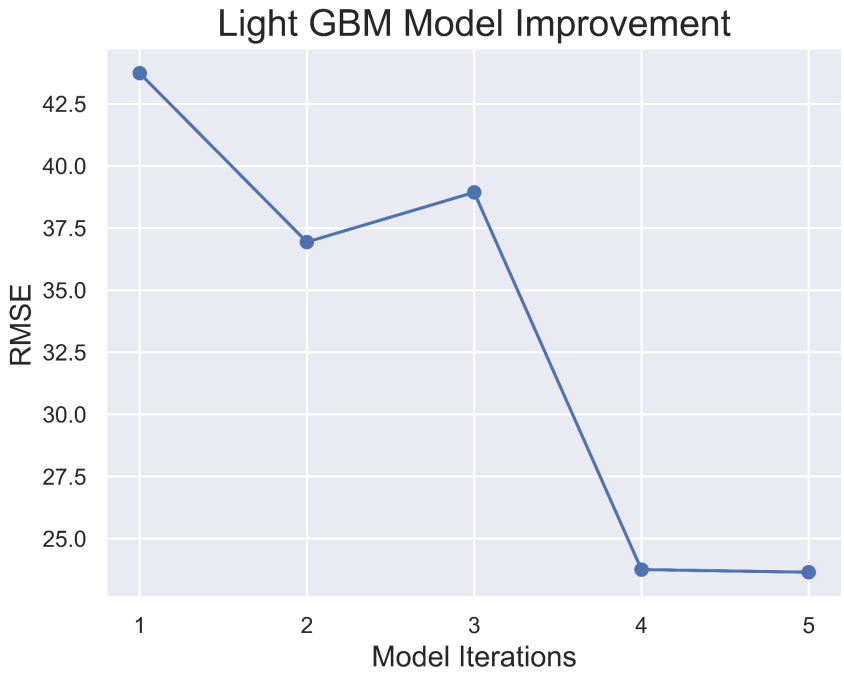

Figure 12. The Light GBM model RMSE is plotted as the model was iterated throughout the project showing that the fourth iteration, after initial hyperparameter tuning, was the best model.

Insights and Performance

Once the final hyperparameter settings were identified the model was trained on the full X_train/y_train dataset and was tested against the X_test/y_test portion. This final model had an RMSE of 22.83, which was an improvement over every prior test. The most important features of this final model were plotted below in Figure 13 to identify what drives volumetric CO_2 working capacity.

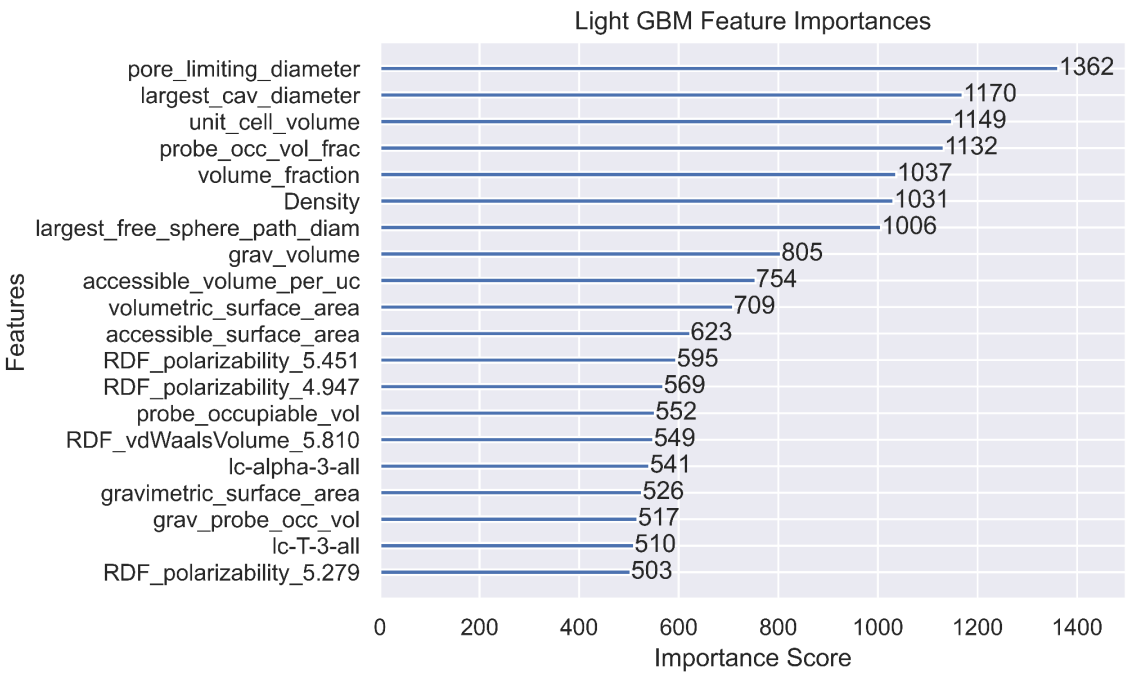

Figure 13. The 20 most important features to the final Light GBM model are listed in order of importance.

Interestingly, the top 7 most important features were all geometry-related features rather than electronic properties such as RDFs or RACs. To understand why these geometric properties are so important, it's helpful to first understand the interactions that drive gas adsorption in MOFs. The most common adsorption interactions in MOFs are dipole-dipole, and dipole-induced dipole. The bonds in a CO_2 molecule are polar, meaning that even though the electrons in the bond are shared between the carbon and oxygen, they are pulled towards the oxygen and reside on average more closely to those atoms. This in combination with the two electron pairs on the oxygens make those atoms partially negatively charged and the central carbon partially positively charged (Figure 14 below).

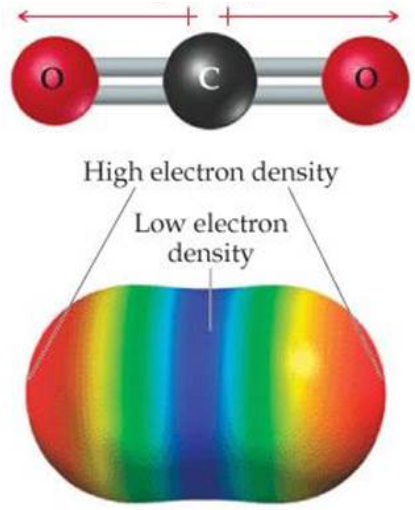

Figure 14. A carbon dioxide molecule is shown with dipoles pointing in the direction of higher electron density. The image below shows high electron density in red and low electron density in blue. **Source:** https://schoolbag.info

As CO_2 molecules come into contact with the surface area of a MOF, one of three effects can cause adsorption:

1. A (partially) negatively charged feature in the MOF will attract and hold the carbon atom at the center
2. A (partially) positively charged feature in the MOF will attract and hold an oxygen atom at one of the ends
3. The CO_2 comes into contact with a nonpolar surface and distorts the electron cloud of the MOF surface, temporarily creating a weak attraction between the surface and the CO_2 molecule

It makes sense then that the most important feature was pore limiting diameter, which has a target range for volumetric working capacity. Pore limiting diameter refers to the largest opening in the MOF that would allow a gas to enter the pores of the MOF. As shown in Figure 3 there is a clear sweet spot between 0 and 10 angstroms. The kinetic diameter of CO_2, the smallest diameter necessary to fit through an opening, is 3.3 angstroms which fits perfectly in that window. As the pore limiting diameter continues to increase beyond that value the CO_2 isn't forced into such close contact with the surfaces of the MOF, leading to fewer interactions with the MOF and less overall adsorption.

Largest cavity diameter was the second most important feature and also had a sweet spot around 8 angstoms - roughly the diameter of two CO_2 molecules. Much like the pore limiting diameter, as the cavity diameter increases beyond that there is more void space where CO_2 may be present, but not adsorbed to the MOF. As shown in Figure 2, the distribution of volumetric working capacity goes as high as 175 mL of CO_2 per mL of MOF, and the key to fitting so much CO_2 in such a small volume is to maximize interactions with the MOF surface rather than allowing space for gaseous CO_2.

The third most important feature to the Light GBM model was unit cell volume, which was shown in my exploratory analysis to have a negative correlation with volumetric working capacity. To understand this feature let's start by imagining that the volume we have in our exhaust stack is represented by a 4x4x4 cube. The unit cell of a MOF is the smallest cubic volume required to fit one repeat unit of the MOF. If we have one MOF with a unit cell that is 1x1x1 vs another that is 2x2x2, we can fit eight times as many repeat units of the smaller MOF in our exhaust stack than the larger one, so that MOF with the larger unit cell must have at least eight times more surface area per unit cell than the smaller MOF to make up for that difference.

The fourth most important feature to the model was probe occupiable volume fraction. This feature is the fraction of total volume that can be accessed by a probe molecule, meaning that space occupied by the MOF, or spaces too small for the probe to access are factored out. Since this wasn't one of the features explored in the initial analysis I plotted it against volumetric working capacity to examine that relationship (Figure 15).

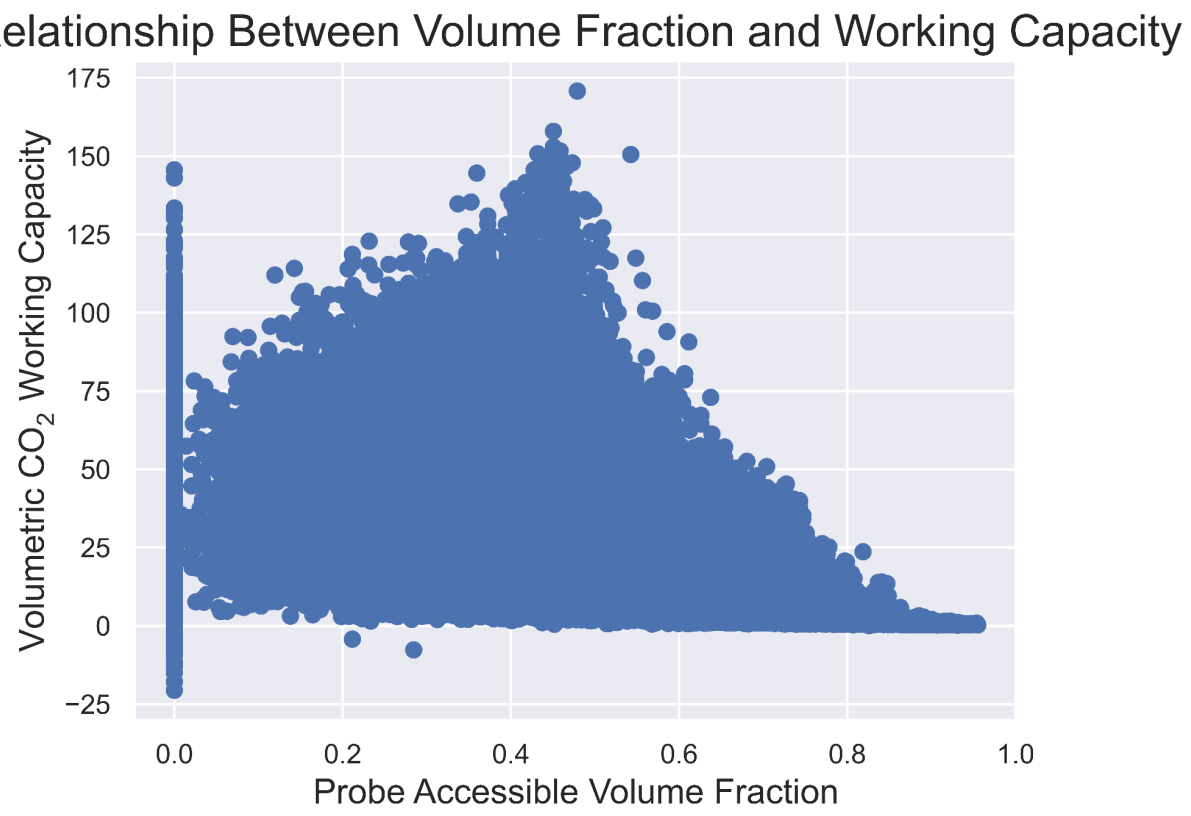

Figure 15. Volumetric working capacity is plotted against probe occupiable volume fraction showing a peak at roughly 0.4-0.5.

As we can see, the best performing MOFs have roughly 40-50% free volume for CO_2 to occupy, the rest ideally being accessible MOF surface area. What these features show is that we really want to pack as much surface area into as small a volume as possible.

Conclusions

After testing a wide range of regression models, Light GBM was identified as the most accurate option for predicting volumetric CO_2 working capacity in a MOF. This model found four features to be particularly important:targeting pore limiting diameter and largest cavity diameter to 0-10 angstroms, minimizing unit cell volume, and targeting a probe accessible volume fraction between 0.4 and 0.5. While many other factors contribute to the working capacity, these paint the picture that the most important strategy in designing a MOF for CCS is to design a dense MOF with lots of surface area shaped into pores that are roughly the diameter of two CO_2 molecules. It would be interesting to work with a research group focused on CCS to dig deeper into this dataset and attempt to synthesize some novel MOFs following this paradigm. I'd also like to find datasets to join into this study specifying metal atom identity and linker identity to help guide the synthesis even further.