Pores to Process: The In Silico Study of Metal-Organic Frameworks from Crystal Structure to Industrial Pressure Swing Adsorption for Postcombustion Carbon Capture and Storage

Thomas D. Burns

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctorate in Philosophy degree in Chemistry

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

Table of Contents

Abstract vii
Table of Figures x
Table of Tables xviii
List of Acronyms xx
Acknowledgements xxii
1. Chapter 1: Introduction and Background 1
1.1. Sources of GHG Emissions 1
1.2. Carbon Capture and Storage 2
1.2.1. Post-Combustion Carbon Capture 2
1.2.2. Solvent Based Systems 3
1.2.3. Pressure Swing Adsorption 4
1.3. Materials for Pressure Swing Adsorption 5
1.3.1. Metal-Organic Frameworks 6
1.4. Studying MOF systems 7
1.4.1. Atomistic Level Studies 7
1.4.2. Process Level Studies 8
1.4.3. Disconnect Between Material Chemists and Process Engineers 8
1.5. Research Goals 9
1.6. Common Performance Metrics 10
1.6.1. Material Science Targets 10
1.6.2. Process Level Targets 16
1.7. Overview of Chapters 18
1.8. References 22
2. Chapter 2: Simulation and Modelling Methods 30
2.1. Introduction and Overview 30
2.2. Atomistic Simulations 30
2.2.1. Atomic Charges: The REPEAT Method 31
2.2.2. Grand Canonical Monte Carlo Simulations 31
2.3. Process Simulations 38
2.3.1. 4-Stage Light Product Pressurization Cycle 39
2.3.2. 4-Stage LPP Simulator Input Variables 39
2.3.3. Assumptions 41
2.3.4. The Finite Volume Method 42
2.3.5. Convergence and Outputs 45
2.4. Optimization Methods 46
2.4.1. Gradient Descent and the Adam Optimizer 46
2.4.2. Genetic Algorithm 48
2.4.3. Genetic Algorithm: Gene Selection 48
2.4.4. Genetic Algorithm: Latin Hypercube Sampling 49
2.4.5. Genetic Algorithm: Objective Functions 49
2.4.6. Genetic Algorithm: Elitism and Mating 50
2.4.7. Genetic Algorithm: Convergence 52
2.5. Machine Learning Methods 53
2.5.1. Model Types 53
2.5.2. Supervised Learning 53

2.5.3. Unsupervised Learning 54
2.5.4. Feature Scaling 54
2.5.5. Cross-Validation 55
2.5.6. Linear Discriminant Analysis 55
2.5.7. Principal Component Analysis 57
2.5.8. Kernel-Based Principal Component Analysis 58
2.5.9. Decision Trees and Random Forest 59
2.5.10. Random Forest 60
2.5.11. Gradient Boosted Decision Trees 61
2.5.12. Artificial Neural Networks 64
2.6. References 68
2.7. Appendix 2.1: 4-Stage Light-Particle Pressurization Boundary Conditions 73
3. Chapter 3: Guest Atom Localization Algorithm 74
3.1. Abstract 74
3.2. Introduction 74
3.3. Methodology & Results 80
3.3.1. How GALA Works 80
3.3.2. Optimization of Parameters 85
3.3.3. Binding Site Accuracy 97
3.4. Conclusions 101
3.5. Author Contributions 101
3.6. References 102
4. Chapter 4: Pores to Process 105
4.1. Abstract 105
4.2. Introduction 105
4.3. Methodology 109
4.3.1. CoRE Database & Named Materials 109
4.3.2. DFT Calculations 109
4.3.3. Grand-Canonical Monte Carlo Simulations 110
4.3.4. Isotherm Fittings 110
4.3.5. Geometric Calculations 111
4.3.6. Pressure Swing Adsorption Simulator 111
4.3.7. Compression Energy Calculations 112
4.3.8. Tiered Screening Approach 116
4.4. Results and Discussion 123
4.4.1. Results From Screening 123
4.4.2. Relationship to Common Metrics 133
4.5. Conclusions 136
4.6. References 138
4.7. Author Contributions 142
4.8. Appendix 143
4.8.1. Appendix 4.1: List of Named Materials 143
4.8.2. Appendix 4.2: N2-NIMF Lennard-Jones Parameters 144
4.8.3. Appendix 4.3: Density of Supercritical Fluid Mixture 144
4.8.4. Appendix 4.4: Theoretical Thermodynamic Limit 145
5. Chapter 5: Data Mining of Detailed Process Simulations 146
5.1. Abstract 146
5.2. Introduction 146

5.3. Methodology 148
5.3.1. Data Set Classification 148
5.3.2. Generating Isotherms 149
5.3.3. Conventional Performance Metrics 149
5.3.4. Preparation of Univariate Plots 151
5.3.5. Machine Learning Techniques 151
5.4. Results and Discussion 154
5.4.1. Data Set Preparation 154
5.4.2. Isotherm Analysis 154
5.4.3. Univariate Analysis 156
5.4.4. Linear Discriminant Analysis 159
5.4.5. Principal Component Analysis 162
5.4.6. Decision Trees & Random Forest Modelling 164
5.5. Conclusions 166
5.6. References 168
5.7. Appendix 5.1: Equations of Composite Adsorption Metrics 171
5.8. Appendix 5.2: Isotherm Plots for PE and Prod sets 172
5.9. Appendix 5.3: All Univariate Plots 174
5.10. Appendix 5.4: LDA Heatmaps for PE and Prod sets 177
5.11. Appendix 5.5: Additional PCA Plots 179
5.12. Appendix 5.6: Random Forest Feature Rankings 181
5.13. Appendix 5.7: Neural Network Confusion Matrix 182
6. Chapter 6: N_2 Binding Sites 183
6.1. Abstract 183
6.2. Introduction 183
6.3. Methods 184
6.3.1. Data Sets 184
6.3.2. Grand Canonical Monte Carlo Simulations 185
6.3.3. Guest Atom Localization Algorithm 186
6.3.4. Tanimoto Coefficient 186
6.3.5. Uptake Ratios 187
6.3.6. Bootstrapping 188
6.3.7. Linear Discriminant Analysis (LDA) 188
6.4. Results and Discussion 188
6.4.1. Effect of Tanimoto on Binding Sites 188
6.4.2. Probability Plot Comparisons 189
6.4.3. Uptake Ratios 192
6.4.4. Correlation between Tanimoto and Uptake Ratio 193
6.4.5. LDA Prediction of DoE-PRT 194
6.4.6. Design Principles 194
6.5. Conclusions 195
6.6. Future Work 196
6.7. References 198
6.8. Appendix 6.1: Comparison of N_2 Single and Binary Component Binding Sites 200
7. Chapter 7: Interpolation Model in the PSA Simulator 204
7.1. Abstract 204
7.2. Introduction 204
7.3. Methods 206

7.3.1. Metal-Organic Framework Sets 206
7.3.2. Grand Canonical Monte Carlo Simulations 206
7.3.3. Latin Hypercube Sampling 207
7.3.4. Competitive Dual-Site Langmuir 208
7.3.5. Linear Interpolation 209
7.3.6. Percent Mean Absolute Deviations 209
7.4. Results and Discussion 209
7.4.1. Performance of Test Sets 210
7.4.2. Pressure Swing Adsorption 210
7.4.3. Temperature Swing Adsorption 212
7.5. Conclusions 214
7.6. References 216
8. Chapter 8: FoCAS Surrogate PSA Model 217
8.1. Abstract 217
8.2. Introduction 217
8.3. Methodology 219
8.3.1. Dataset 219
8.3.2. Removing Outliers 219
8.3.3. Fitting Subsets 220
8.3.4. Scaling Features 225
8.3.5. Feature Correlation 225
8.3.6. Gradient Boosted Decision Trees 226
8.3.7. Neural Networks – Multi-Layer Perceptron 226
8.3.8. Iterative Grid-Search 226
8.4. Results and Discussion 227
8.4.1. Model Development 227
8.4.2. Comparisons to Detailed PSA Simulations 231
8.4.3. Screening of CoRE Database 234
8.5. Model Limitations and Domain Considerations 236
8.6. Conclusions 236
8.7. References 238
8.8. Appendix 240
8.8.1. Appendix 8.1: Features used in Neural Networks 240
8.8.2. Appendix 8.2: Neural Network Hyperparameters 241
8.8.3. Appendix 8.3: Domain of the four Neural Network Models 241
9. Chapter 9: Conclusions and Future Work 242
9.1. Conclusions 242
9.1.1. Chapter 3: Guest Atom Localization Algorithm (GALA) 242
9.1.2. Chapter 4: Pores to Process 242
9.1.3. Chapter 5: Datamining PSA Results 243
9.1.4. Chapter 6: N_2 Binding Sites 243
9.1.5. Chapter 7: Linear Interpolation in PSA Simulator 244
9.1.6. Chapter 8: FoCAS 245
9.2. Future Work 245
9.2.1. Chapter 3: Guest Atom Localization Algorithm (GALA) 245
9.2.2. Chapters 4 & 5: Pores to Process 246
9.2.3. Chapter 6: N_2 Binding Sites 246
9.2.4. Chapter 7: Linear Interpolation in PSA Simulator 246

9.2.5. Chapter 8: FoCAS..... 247
9.3. References 248

Abstract

This thesis explores the use of computational chemistry and machine learning techniques to aid in the design of Metal-Organic Frameworks (MOFs) for use in postcombustion carbon capture and storage (PoC-CCS). PoC-CCS is an ongoing field of research which aims to selectively remove carbon dioxide, an important greenhouse gas, from the exhaust of fossil-fuel burning powerplants. By using a suite of advanced simulation techniques, high-throughput screenings were performed on thousands of MOFs to study their behaviour in a pressure swing adsorption (PSA) system. To develop a comprehensive picture of a material's performance, the behaviour of individual gas molecules within the pores of the crystal structures to the material's performance in industrial scale PSA columns was evaluated.

To study the behaviour of individual gas molecules within the pores of a MOF, a new algorithm which can accurately determine the locations of gas binding sites was developed. This algorithm, which relies on probability distributions generated through grand canonical Monte Carlo simulations (GCMC), was optimized for CO_2 with the goal of use in high-throughput screening. By tuning the user-controlled parameters for a desired gas, this algorithm, which was named the Guest Atom Localization Algorithm (GALA), was shown to accurately reproduce experimentally determined binding sites while being run in a high-throughput manner with no user intervention.

Studying MOFs at the pore or crystal scale in this manner provides valuable insights into the behaviour of gases within the materials. A major shortcoming, however, is the lack of direct insight into the material's behaviour in industrial systems. Materials scientists and MOF chemists have historically focused on a set of performance metrics measured at this scale; however, no clear connection can be made between such metrics and the performance of that sorbent material in a PSA column. To bridge this gap between MOF chemists and the process engineers studying the PSA systems, a large-scale screening of MOFs was performed using a sophisticated PSA simulator designed to reproduce the performance of an 80 kg PSA column. By supplying isotherms obtained using GCMC simulations to be used as inputs into the PSA simulator, a multi-scale high-throughput screening of MOFs for PoC-CCS was performed for the first time under coal-fired powerplant conditions.

This multi-scale screening provided the ideal conditions to study the materials science performance metrics and their relationships to industrial PSA performance. To study this relationship, a series of machine learning and artificial intelligence techniques were employed. The primary goal was to extract important relationships between the materials science and industrial PSA performance metrics,

with a secondary goal of developing a predictive model which could be used to accelerate the pace of materials discovery. Through the use of machine learning, several metrics were identified which could be used to predict whether a material could meet the minimum target of 95 % purity of captured CO_2, and 90 % removal (or recovery) of CO_2 from the flue gas stream. Among them was the isotherm parameters for N_2, the most abundant species in the flue gas. This finding was significant as to date the focus among MOF chemists studying the PoC-CCS system was placed primarily on the CO_2 metrics, with N_2 only implicitly considered when calculating the CO_2/N_2 selectivity. Although several metrics were identified which could predict the purity and recovery targets, none of the conventional metrics tested could be used to estimate the energetic cost of capture or the size of the capture plant, both important considerations in evaluating the cost of capture.

The relationship between N_2 binding within the pores of the MOF and its ability to meet the purity-recovery targets was explored using GALA. Using a Tanimoto similarity metric and the ratio of single component and competitive loadings, the CO_2 and N_2 binding environments were studied. It was determined that when the N_2 binding environment was significantly altered by the presence of CO_2, the material was more likely to meet the purity-recovery targets. Further analysis found that this change in binding environments was correlated to a reduced N_2 uptake in the presence of CO_2, implying that the competition for binding sites within the pores of the MOF is an important indicator for the material's ability to meet the purity-recovery target. For the first time, a direct relationship between the behaviour of individual gas molecules to industrial PSA performance can be reported.

Although the PSA simulator used throughout this work has proven to be a powerful tool for materials discovery, several shortcomings still exist. The first is the method used by the simulator to predict the loadings at various points within the column. This method relies on single component isotherm data despite the ability of GCMC to simulate multi-component isotherms. An alternative method to using single component isotherms was proposed which relies on multi-component isotherm data and a linear interpolation model. The existing method was compared to the new proposed interpolation method, and it was found that the loadings predicted using the interpolation method were more accurate. The second shortcoming of the PSA simulator is the computational expense associated with the optimizations. Using the PSA simulator, a single material may take up to a week to be fully optimized on a high-performance computing cluster. To increase the pace of materials discovery, a surrogate model was developed using the data accumulated over the course of the work presented in this thesis. Using artificial neural networks, a suite of models was developed which reproduces the

outputs of the PSA simulator and is able to optimize a single MOF in a matter of minutes. This suite of models, known as the **Fo**ssil Fuel **C**ombustion for Carbon C**a**pture and **S**torage (FoCAS) was used to perform a screening of over 4,000 materials.

Table of Figures

Figure 1.1 Schematic diagram of a fossil fuel burning powerplant equipped with a PoC-CCS system. 3

Figure 1.2 Schematic diagram of a simple pressure swing adsorption (PSA) system for a CO_2/N_2 separation process. The columns shown in this figure are packed with a solid sorbent material with the red regions representing areas of the column containing CO_2 rich gas. 6

Figure 1.3 Diagram demonstrating the formation of MOFs from organic and inorganic linker which assemble into different network topologies to form the final MOF structure..... 7

Figure 1.4 Three isotherms with the same saturation uptake of 2.0 mmol/g (black dotted line) and different values for the Langmuir constant, b, of 5.0 bar^{-1} (blue line), 1.0 bar^{-1} (green line), and 0.2 bar^{-1} (orange line). 12

Figure 1.5 Demonstration of the working capacity calculation on an isotherm with θsat = 2.0 mmol/g and b = 5.0 bar^{-1}, showing the uptake and pressure at adsorption conditions (green dashed lines) and at the desorption conditions (red dashed lines). 13

Figure 2.1 Plot showing the Lennard-Jones potential for a function with a σij of 1.0 Å and an ϵij of 2.0 kcal/mol. The region representing the steric repulsion is shown in red, and the attractive region resulting from dispersion interactions is shown in green..... 36

Figure 2.2 Comparison of the decay rates of the $1/r$, $erfr/r$, and $erfcr/r$ functions used to calculate the coulombic contribution to the potential calculated by GCMC plotted as a function of the interatomic distance rij between atoms j and j. 38

Figure 2.3 Diagram of the 4-stage light product pressurization cycle used in the pressure swing adsorption simulator, depicting a single column over the course of a single cycle. The columns, packed with sorbent material, are pressurized with flue gas, with the CO_2 rich gas depicted in red and the N_2 rich gas depicted in gray. 40

Figure 2.4 Diagram depicting the column of length L packed with a sorbent material divided into N equally sized segments. The segments, which are assumed to have constant gas pressure and composition, and are positioned along the column's axial axis, z. 42

Figure 2.5 a) Example of 5 randomly selected data points and b) 5 datapoints selected using the Latin hypercube sampling technique..... 49

Figure 2.6 Visual demonstration of the roulette wheel used in the GA mating protocol, where v¬i is the volumetric slice occupied by an individual in the generation and represents the probability of selection. 51

Figure 2.7 a) Generic dataset with two classifications defined by the colour of the ball defined by two features A and B. b) The same generic dataset as in a) with a dashed line representing the line fit through linear discriminant analysis and the arrows demonstrating the projection of the 2-dimensional data onto the line. c) The representation of the dataset shown in a) and b) projected onto the dashed line in figure b), where the variable s is the scatter of the class and μ is the mean of the class along the line's frame of reference..... 56

Figure 2.8 General structure of a decision tree classifier with a maximum depth of 2, dividing an arbitrary dataset into two classifications: Class 1 and Class 2. This decision tree is using the feature space, x, where x0¸ x1, and x2 are the features selected at each node, and the variables a, b, and c represent the values used to split the data. 60

Figure 2.9 General structure of a decision tree estimator from a gradient boosted decision tree with a maximum depth of 3, with each branch terminating in a prediction "γ" _(i,m) where i is the individual in the set, and m is the estimator. This decision tree is using the feature space, x, where xj is the feature selected in node j, and the variables a, b, c, d, e, f, and g represent the values used to split the data..... 63

Figure 2.10 Diagram of an example multi-layered perceptron neural network including 4 hidden layers with 50 neurons / hidden layer. All lines connecting the nodes in the network represent the linear transformation of the data entering a neuron (or node) represented by a circle. 65

Figure 2.11 Plot of the Linear Rectifier (ReLU) function for values of x ranging from -5 to 5. 66

Figure 3.1 a) 3D-isosurfaces of the CO2 probability distributions (cyan – carbon; red – oxygen) in CALF-15 (Zn2(3-amino-1,2,4-triazole)2(oxalate)), determined from a GCMC simulation.17 Also shown in tube representation are the experimental CO2 binding sites determined from X-ray analysis. b) Centre of mass probability density plots of CO2 molecules in CALF-16 (Zn3(3-amino-1,2,4-triazole)3(PO4)). In both a) and b), the framework of the MOF is shown in line representation with H and Zn atoms removed for clarity. 75

Figure 3.2 a) Unrefined CO_2 isosurfaces with the oxygen (red) surfaces and carbon (black) surfaces, generated from a simulation of hypothetical MOF str_m19_o180_ubt with 10 million GCMC cycles. b) The same isosurfaces as in (a) but with a reduced isosurface value revealing areas of high probability corresponding to CO_2 binding sites (circled). c) The same MOF with CO_2 molecules placed by GALA (circled). 77

Figure 3.3 1D probability distribution of the carbon atom derived from GCMC simulation of CO_2 gas adsorption in MOF CALF-15 (red). The probability is plotted along a line which passes through one of the binding sites. The red arrows point to local maxima in the raw probability distribution. The blue line is the result of a 'smoothing' of raw probability distribution with a frequency filter (this work). 77

Figure 3.4 Select CO_2 binding sites determined from maxima in the calculated probability distributions in MOF CALF-16. CO_2 molecules are shown in tube representation while the MOF framework is shown with line representation with the H and Zn atoms removed for clarity. 78

Figure 3.5 A 2D representation of (a) normal binning procedure and (b) equitable binning. The brown circle with the central white dot depicts the position of the atom. The numbers in each grid area show the amount that the atom contributes to the probability distribution in each grid area for the two binning procedures. 81

Figure 3.6 Adsorption and geometric properties of the Small Set (yellow squares) and the Large Set (blue circles). Scatter plots of a) the calculated volumetric versus the gravimetric CO2 uptake (at 1 bar and 298K) of the MOFs and b) the surface area versus the maximum pore diameter. 87

Figure 3.7 Schematic diagram of the built-in sigma and occupancy cut-off parameter optimization scheme which relies on a user defined target value to select the best sigma and occupancy cut-off value for individual MOFs. 91

Figure 3.8 CO_2 binding sites calculated in GALA for CALF-16 with an emphasis on the material's competitive binding sites (circled). Binding sites were calculated using the GALA parameters found in Table 3.1. 92

Figure 3.9 3-Dimensional surface plot of the fitness values for every point in the 2-dimensional grid-search of the *Tanimoto* and the *exclusion radius.* 95

Figure 3.10 Histograms of the optimized parameter for the Large Set of 298 hypothetical MOFs for a) sigma, and b) the occupancy cut-off. For both plots, the most common value is denoted by a dashed line with a sigma value of 2.0 Å and occupancy cut-off of 10%. 97

Figure 3.11 Plot of the number of GALA binding sites vs the number of binding sites determine by hand for all MOFs in the Large Set, including a 1:1 line represented by the black dashed line. 98

Figure 3.12 Binding site location comparison between experimental binding sites (green) and GALA binding sites (orange) in six of the 8 experimental MOFs. Hydrogen atoms have been omitted from the MAF-2 structures to increase visual clarity. Atoms in this figure: Carbon (gray), Oxygen (red), Nitrogen (blue), Sulfur (yellow), Hydrogen (white), and Metals (pink). 99

Figure 3.13 Number of GALA binding sites generated plotted against the GCMC uptake for a) Argon, b) Methane, and c) molecular Nitrogen. The dashed line represents the 1:1 line between the x and y axes. 100

Figure 4.1 Plot of the average density of a supercritical fluid mixture of CO_2 and N_2 as a function of mole fraction of CO2. The calculated densities are presented as blue circles, and the fitted line based on equation 4.9 is shown as a blue dashed line with a Pearson R2 of 0.9998. 115

Figure 4.2 Non-linearity plot denoting the non-linearity values, b, for all MOFs relative to Zeolite-13X. All MOFs in the set are plotted as a heatmap and divided into regions denoting the best purity/recovery obtainable for those materials. The green region corresponds to MOFs which are able to achieve a purity

of 95% with a recovery of 90% with same point. The orange region corresponds to MOFs which are able to meet a more relaxed 90% purity and 90% recovery, whereas the regions in red indicate materials which are unable to meet those targets. 117

Figure 4.3 The parasitic energy plotted against the productivity for the lowest parasitic energy process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line). 125

Figure 4.4 The parasitic energy plotted against the productivity for the highest productivity process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line). 127

Figure 4.5 The parasitic energy plotted against the productivity for the best fitness process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line). 129

Figure 4.6 Pareto fronts showing the performance boundary between the best productivity and parasitic energy points sampled for IISERP-MOF-2 (blue circles), Mg-MOF-74 (yellow diamonds), Zeolite NaA (red squares), QEJYIP (green triangles), UTSA-16 (black circles), and Zeolite-13x (purple diamonds). The DoE target for parasitic energy of 258 kWh/tonne CO_2 is shown as a solid orange line. 131

Figure 4.7 Plots of the energy penalty vs the number of columns required to run a continuous capture in a 100 MW coal-fired powerplant for (a) the best parasitic energy process points, (b) the best productivity process points, and (c) the best overall fitness process points for all 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the energy penalty (green line), the DoE target for the energy penalty (orange line), and the energy penalty from a retrofitted liquid amine capture system (red line). 133

Figure 4.8 (a) The plot of the best parasitic energy for all 482 materials that meet the DoE-PRT plotted against the CO_2/N_2 selectivity, (b) the plot of the best parasitic energy for all 482 materials that meet the DoE-PRT plotted against the CO_2 working capacity, (c) the best productivity for all 482 materials that meet the DoE-PRT plotted against the CO_2/N_2 selectivity, and (d) the best productivity for all 482 materials that meet the DoE-PRT plotted against the CO_2 working capacity. Included in the figures (a) and (b) is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line). 134

Figure 5.1 Histograms of the (a) best parasitic energy and (b) best productivity for all 1,022 fully optimized MOFs which meet the DoE-PRT. 155

Figure 5.2 Plots showing the range of the isotherms, denoted using a shaded area, and the average isotherm shown in a solid line for (a) CO_2 isotherms of the top 150 MOFs ranked by parasitic energy shown in black, (b) CO_2 isotherms of the MOFs which meet the DoE-PRT shown in blue, (c) the CO_2 isotherms of the MOFs which fail the DoE-PRT shown in red, (d) N_2 isotherms of the top 150 MOFs ranked by parasitic energy shown in black, (e) N_2 isotherms of the MOFs which meet the DoE-PRT shown in blue, and (f) the N_2 isotherms of the MOFs which fail the DoE-PRT shown in red..... 156

Figure 5.3 Univariate plots showing smoothed distributions of (a) competitive CO_2 uptake, (b) CO_2 working capacity, (c) competitive N_2 uptake, (d) N_2 working capacity, (e) CO_2/N_2 selectivity, (f) percent Regenerability, (g) Huck's parasitic energy, and (h) the separation potential. In this figure, the top 150 MOFs ranked by parasitic energy are shown in black, the MOFs which meet the DoE-PRT are shown in blue, and the MOFs which fail the DoE-PRT are shown in red. All histograms are normalized so that the area under the curve is equal to 1. 157

Figure 5.4 A heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional linear discriminant analysis models fit to classify MOFs in the DoE-PRT set according to their subsets, with the diagonal representing the balanced accuracy for the 1-dimensional linear discriminant analysis of each feature. The feature IDs on the axes correspond to the features in Table 5.1 and can be separated into four distinct groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties. 161

Figure 5.5 Scatter plots of first two principal components from principal component analysis (a-c) and kernel principal component analysis (d-f) with a gamma of 0.1 using the rbf kernel. The PCAs were run on several sets: (a & d) MOFs which meet the DoE-PRT (blue) and do not meet the DoE-PRT (red), (b & e) MOFs with low PE (blue) and MOFs with high PE (red) divided by the median PE, and (c & f) MOFs with low productivity (red) and high productivity (blue) divided by the median productivity value..... 163

Figure 5.6 Bar plots demonstrating the percent of nodes showing the top 10 features found in the random forest classifier model, with the percentage of nodes relying on the features shown in the y-axis for (a) the first nodes in the trees, (b) the nodes in the second layer of the decision trees, and (c) the nodes in the third layer of the decision trees. 165

Figure 6.1 Visualization of the binding sites within the pores of MOFs (a) FASQUN for the single component N_2 simulation, (b) FASQUN for the binary component CO_2/N_2 simulation, (c) CUGLTM02 single component N_2 simulation, and (d) CUGLTM02 binary component CO_2/N_2 simulation. In this representation, the framework atoms are shown in a stick representation, the N_2 binding sites are displayed in CPK format in blue, and the CO_2 binding sites are displayed in CPK format with the oxygen atoms in red and carbon in grey. 190

Figure 6.2 Histograms of the Tanimoto coefficients for (a) carbon, (b) oxygen, and (c) nitrogen probability plots, comparing the single component probability distributions to the binary component probability distributions. In these plots, the distributions for MOFs which meet the DoE-PRT are represented by the green lines, whereas MOFs which do not meet the DoE-PRT are represented by the red lines. The 99% confidence intervals of the mean Tanimoto coefficients are shown as green and red bars for the distributions of MOFs which passed and failed the DoE-PRT, respectively. All histograms have been normalized so that the area under the curve is equal to 1. 191

Figure 6.3 Histograms of the uptake ratios for (a) carbon dioxide, (b) molecular nitrogen, comparing the binary component uptakes to single component uptakes. In these plots, the distributions for MOFs which meet the DoE-PRT are represented by the green lines, whereas MOFs which do not meet the DoE-PRT are represented by the red lines. The 99% confidence intervals of the mean uptake ratios are shown as green and red bars for the distributions of MOFs which passed and failed the DoE-PRT, respectively. All histograms have been normalized so that the area under the curve is equal to 1. 192

Figure 6.4 Plot of N_2 uptake ratio against N_2 Tanimoto coefficients for MOFs which meet the DoE-PRT (green circles) and MOFs which do not meet the DoE-PRT (red circles) with a linear tread line fit to the full data set (black dashed line). 193

Figure 6.5 Heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional LDAs including the N_2 Tanimoto coefficient and N_2 uptake ratio parameters combined with 38 metrics tested in Chapters 4 and 5 of this thesis. 195

Figure 7.1 Plots of the productivity vs parasitic energy for MOFs which meet the DoE-PRT in (a) the Pressure Swing Adsorption (PSA) set, and (b) the Temperature Swing Adsorption (TSA) sets. All points displayed are the best parasitic energy point from the optimizations in Chapter 4. 209

Figure 7.2 Plots of the Percent Mean Absolute Deviation (MAD) for (a) CO_2 uptake and (b) N_2 uptake for all 80 MOFs in the PSA set. The black line represents the MAD for the loadings calculated using the competitive DSL model, and the purple, red, green, orange, and blue lines are the percent MADs from the loadings calculated using the linear interpolator using the 11x11x11, 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids, respectively. The percent MAD values are ordered from lowest to highest and are plotted as a function of the number of MOFs. This means that at any given point, the number of MOFs that fall below a certain Percent MAD indicated by the y-axis can be determined by the corresponding point on the x-axis. 210

Figure 7.3 a) Plot of all CO_2 isotherms used in the PSA set, showing the adsorption of CO_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. b) Plot of all N_2 isotherms used in the PSA set, showing the adsorption of N_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. c) The range of CO_2 isotherms tested (blue area) showing the upper and lower bounds of CO_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the PSA set, and the average uptake as a function of pressure (black dotted line). d) The range of N_2 isotherms tested (blue area) showing the upper and lower bounds of N_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the PSA set, and the average N_2 uptake as a function of pressure (black dotted line)..... 212

Figure 7.4 Plots of the Percent Mean Absolute Deviation (MAD) for (a) CO_2 uptake and (b) N_2 uptake for all 101 MOFs in the TSA set. The black line represents the MAD for the loadings calculated using the competitive DSL model, and the purple, red, green, orange, and blue lines are the percent MADs from the loadings calculated using the linear interpolator fit using the 11x11x11, 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids, respectively. The percent MAD values are ordered from lowest to highest and are plotted as a function of the number of MOFs. This means that at any given point, the number of MOFs that fall below a certain Percent MAD indicated by the y-axis can be determined by the corresponding point on the x-axis. 213

Figure 7.5 a) Plot of all CO_2 isotherms used in the TSA set, showing the adsorption of CO_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. b) Plot of all N_2 isotherms used in the TSA set, showing the adsorption of N_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. c) The range of CO_2 isotherms tested (blue area) showing the upper and lower bounds of CO_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the TSA set, and the average uptake as a function of pressure (black dotted line). d) The range of N2 isotherms tested (blue area) showing the upper and lower bounds of N2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the TSA set, and the average N2 uptake as a function of pressure (black dotted line). 214

Figure 8.1 Histograms of the (a) Purity of captured CO_2 and the (b) Recovery of CO_2 from flue gas in the 5.6 million process points used in the initial fittings of the neural network models. 221

Figure 8.2 Histogram of the purity values present in the new purity subset used to train the neural network model for purity. 222

Figure 8.3 Histogram of the recovery values present in the new recovery subset used to train the neural network model for recovery. 223

Figure 8.4 Histogram of the parasitic energy values (excluding compression) present in the new recovery subset used to train the neural network model for parasitic energy. 224

Figure 8.5 Histogram of the productivity values present in the new recovery subset used to train the neural network model for productivity. 225

Figure 8.6 Heatmap of the feature pairs, where each grid entry is the Pearson R^2 correlation coefficient for each feature pair. 227

Figure 8.7 Heatmaps of the test set predicted vs actual values from the ANN models for (a) purity, (b) recovery, (c) parasitic energy, and (d) productivity. The 1:1 line is shown as a black dashed line while the colourmap is shown in the Log_{10} scale. 229

Figure 8.8 Screenshot of the FoCAS A.I. application window. 231

Figure 8.9 Comparison of the predicted vs the actual (a) purity, (b) recovery, (c) parasitic energy (including compression), and (d) productivity. In this figure, the predicted values are calculated using the FoCAS surrogate model and the actual values are calculated using the PSA simulator..... 233

Figure 8.10 Heatmaps of the parasitic energy vs productivity for MOFs found to meet the DoE-PRT showing (a) the best parasitic energy process points, (b) the best productivity process points, and (c) the best overall fitness process points. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line). 235

Table of Tables

Table 2.1 Descriptions of the 7 process variables used in the PSA simulation specific to the 4-stage LPP cycle. 41

Table 2.2 Definitions of variables found in equations 2.17, 2.18, 2.19, 2.23, and 2.24. 44

Table 3.1 GALA control parameters and their optimized values for high throughput screening of CO2 adsorption in nanoporous materials..... 86

Table 3.2 Number of unsmoothed maxima identified in the probability distribution of CO2 carbon at flue gas conditions as a function of the grid size. 89

Table 3.3 The maximum, minimum, and interval values for the parameters being optimized by the built-in GALA optimization scheme. 91

Table 3.4 CO_2 performance data and physical properties for the seven hypothetical MOF structures used in the small set during parameter optimization. 93

Table 3.5 Table of MOFs comparing the number of CO_2 binding sites determined experimentally compared to those determined by GALA. 98

Table 4.1 Definitions and values of variables used in equation 4.4, to calculate energy of compression of the captured gas mixture from 1 to 80 bar. 114

Table 4.2 Fitted from equation 4.9, fit to reproduce the average fluid density of a CO_2/N_2 supercritical fluid over a range of CO_2 mole fractions. 115

Table 4.3 List of all process variables controlling the pressure swing adsorption simulation, including ranges and intervals tested, and the constant values used in the grid-search. 118

Table 4.4 List of all process variables controlling the pressure swing adsorption simulation used as genes in the genetic algorithm, with their minimum and maximum allowable values. 119

Table 4.5 Values used in equation 4.11 to balance the fitness function. 121

Table 4.6 Top 10 materials ranked according to the best parasitic energy points found during the optimization. The productivity values shown are the productivities of the materials at the process point that minimizes the parasitic energy. With the exception of IISERP-MOF-2, zif-36-frl, and UTSA-16, the names provided for each MOF are the designations given in the CoRE database. 126

Table 4.7 Top 10 materials ranked according to the best productivity points found during the optimization. With the exception of NaA and UTSA-16, the names provided for each MOF are the designations given in the CoRE database. 128

Table 4.8 Top 10 materials ranked according to the best overall fitness points found during the optimization. With the exception of IISERP-MOF-2, NaA, and UTSA-16, the names provided for each MOF are the designations given in the CoRE database. 130

Table 4.9 The Pearson and Spearman correlation R^2 values comparing the best parasitic energies and productivities to 22 common adsorption metrics and 7 geometric properties used to evaluate materials for post-combustion carbon capture. 135

Table 5.1 Description of the four main classification sets, and subsets used for data analysis throughout this chapter. 148

Table 5.2 List of all features tested, along with the corresponding ID used in Figure 5.4. These features are separated into four groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties. 150

Table 5.3 Balanced accuracies of random forest classification models fit using alternate sets of features comparing CO_2 and N_2 specific features to the original random forest model for the DoE-PRT set, labeled All Features. Also included are the balanced accuracies from the best PE-set and Prod-set classification models..... 166

Table 7.1 Ranges of conditions testing using GCMC for the PSA system and the TSA system..... 207

Table 8.1 Table of the R^2 coefficients and mean absolute errors (MAE) showing the prediction accuracy of the Gradient Boosted Decision Trees for the training and test sets of the four target variables. 228

Table 8.2 Table of the R^2 coefficients and mean absolute errors (MAE) showing the prediction accuracy of the optimized ANNs for the training test sets of the four target variables. 230

Table 8.3 The number of the top 100 materials based on the rankings from the screening of 1,022 MOFs in Chapter 4 according to parasitic energy, productivity, and overall fitness in the top 100, 200, and 200 rankings of those same MOFs using FoCAS. 233

Table 8.4 The parasitic energy and productivity of a single process point for the top 10 MOFs ranked by (a) parasitic energy, (b) productivity, and (c) overall fitness. 236

List of Acronyms

- ANN Artificial Neural Network
- APS Adsorbent performance score
- CCS Carbon capture and storage
- CCUS Carbon capture, utilization, and storage
- CoRE Computation ready experimental MOF database
- DoE United States Department of Energy
- DSL Dual-site Langmuir
- DT Decision Tree
- EPA United States Environmental Protection Agency
- ESP Electrostatic Potential
- FoCAS Fossil Fuel Combustion Carbon Capture and Storage
- GA Genetic algorithm
- GALA Guest atom localization algorithm
- GBDT Gradient Boosted Decision Tree
- GCMC Grand Canonical Monte Carlo
- GHG Greenhouse gas
- IAST Ideal absorbed solution theory
- IPCC Intergorvernmental panel on climate change
- kPCA Kernel Principal Component Analysis
- LDA Linear discriminant analysis
- LHS Latin Hypercube Sampling
- LJ Lennard-Jones
- LPP Light particle pressurization
- MAD Mean Absolute Deviation
- MEA Monoethanolamine
- MLP Multilayer Perceptron
- MOF Metal-organic framework
- MSE Mean squared error
- N2-NIMF Nitrogen in metal-organic frameworks
- PCA Principal Component Analysis
- PE parasitic energy
- PoC-CCS Post-combustion carbon capture and storage
- Pr/Prod Productivity
- PRT Purity-recovery target
- PSA Pressure Swing Adsorption
- QM-ESP Quantum mechanical electrostatic potential
- REPEAT Repeating Electrostatic Potential Extracted Atomic
- RF Random Forest
- RMSE Root mean squared error
- SSL Single-site Langmuir

- SSP Sorbent selection parameter
- TPD tonnes per day
- TSA Temperature Swing Adsorption
- UFF Universal forcefield
- VASP Vienna Ab-Initio Simulation Package
- VSA Vacuum swing adsorption
- PBCs Periodic Boundary Conditions
- ALR Adaptive learning rate
- RBF Radial basis function

Acknowledgements

Like many things in life, a writing a thesis is something that cannot be accomplished alone. First and foremost, I need to thank my thesis supervisor Tom Woo. Over the years Tom provided me with an environment where I could learn and thrive. He fostered curiosity and ingenuity and was always willing to listen to ideas and provide feedback, even when the ideas were wild and possibly overambitious.

I also need to thank everyone I worked with in the Woo lab, Dr. Sean Collins, Dr. Peter Boyd, Dr. Thomas Daff, Dr. Mykhalo Krykunov, Chris Demone, Hana Durekova, Jun Luo, Jake Burner, Dr. Ohmin Kwon, Adam Mirmiran, Phil De Luna, and Dr. Mo Zein Aghaji. Between them they provided me with support in the form of a sounding board to bounce ideas off, give feedback on presentations or paper drafts, or just be there to when we needed to get lunch or blow off some steam. I wouldn't have been able to get through this degree without them.

Outside the lab I also had a tremendous amount of support. My friends and roommates were always supportive and took an interest in my progress, with a specific shout out to Josh Schram, and David Mayer for taking the time to proofread some chapters for me during my final edits. I need to thank my parents, all four of them, for their constant support and always believing that I would get to this point. Finally, I need to thank my partner Camélia, you held me together during the hardest moments of this process, especially when my comprehensive exam didn't go as well as it could have. Her support during the time of writing this thesis, either by being there to distract me when I got annoyed at word and its terrible formatting options, proofreading chapters, or just being there to distract me when I got stressed. I couldn't have done this without her.

1. Chapter 1: Introduction and Background

The increase in global temperatures as a result of rising CO_2 concentrations in the atmosphere poses the biggest existential threat of our generation. The rising concentrations of CO_2 and other harmful greenhouse gases (GHGs) have been directly linked to human activity around the world. In the 2021 report by the Intergovernmental Panel on Climate Change (IPCC),[1] the authors stated that "It is unequivocal that human influence has warmed the atmosphere, ocean and land" and go on to state that "Observed increases in well-mixed greenhouse gas (GHG) concentrations since around 1750 are unequivocally caused by human activities".[1] As more of the world rapidly industrializes, GHG emissions are expected to rise to unprecedented levels,[1] risking irreversible harm to the planet and the millions of species that inhabit it.

1.1. Sources of GHG Emissions

An important first step in reducing anthropogenic GHG emissions is identifying the sources of those emissions. Although inventory data is not readily available from every country, massive international efforts have been made in recent years to identify causes of GHG emissions and develop models and projections for a variety of future scenarios.[1]

The United States Environmental Protection Agency (US EPA) regularly publishes GHG inventory data which helps identify sources of major emitters. In their 2021 report, CO_2 was identified as the most prominent GHG, accounting for 80% of the country's emissions.[2] When broken down by sector, the same report identified transportation as the biggest source responsible for 29% of GHG emissions, with electricity generation close behind it at 25%.[2] The transportation sector, which includes oceanic shipping, long-haul trucking, and everyday car usage is currently undergoing a major shift with the introduction of electric vehicles. Although electrification of oceanic shipping will be challenging, electric vehicles will have an impact on reducing the sector's overall GHG emissions. This reduction, however, will cause increased demand on the electricity sector which could result in electricity generation overtaking transportation as the largest source of GHG emissions in the United States.[2] Similarly, in a 2017 article published by Jiang and coworkers, power generation from coal-fired powerplants accounted for 30% of China's CO_2 emissions,[3] while a separate report stated that India intends on increasing their coal-fired electricity generation by 123% over the next 10 years.[4]

The decarbonization of the world's electrical grids has become a priority to many countries around the planet, with Canada committing to moving towards net-zero by 2050. Although promising advancements in renewable energy sources such as solar and wind and been made in recent years,[5–7] many challenges still need to be overcome before these technologies can completely replace traditional fossil fuels.[8] This means that the world will continue to rely on power generated from burning fossil fuels for many years to come, and with the ever increasing demand around the world for electrical power, an increase in carbon emissions from fossil fuel burning powerplants is almost guaranteed.

1.2. Carbon Capture and Storage

Reducing emissions from stationary GHG sources such as fossil fuel burning powerplants is of paramount importance. Here, Carbon Capture and Storage (CCS) or Carbon Capture, Utilization and Storage (CCUS) will play a principal role. CCS encompasses a wide array of technologies aimed at selectively removing carbon dioxide or other harmful GHGs from gas mixtures. The applications of CCS technologies range from direct air capture,[9–11] which pulls CO_2 directly out of the atmosphere, pre-combustion capture,[12–15] where the fuel is first gasified and the CO_2 is separated from the fuel thereby producing cleaner by-products, oxy-fuel separation,[16–18] in which oxygen is purified prior to combustion to produce fewer by-products, and post-combustion carbon capture,[19–21] which selectively removes CO_2 from combustion flue gas.

1.2.1. Post-Combustion Carbon Capture

Post-combustion carbon capture and storage (PoC-CCS) is the most promising technology for addressing the challenge posed by the continued use of fossil fuel burning powerplants. Unlike the other proposed CCS technologies, such as pre-combustion and oxy-fuel separation, PoC-CCS provides a framework which allows existing powerplants to be retrofit with a capture unit.[22,23] This system is demonstrated using the schematic in Figure 1.1, where the combustion product flue gas composed of N_2 and CO_2 is passed into a carbon capture unit. The purified CO_2 stream is then compressed to transport conditions of 150 bar,[24] and the N_2 rich product gas is vented to the atmosphere. Importantly, since water is a well-known product of combustion, any PoC-CCS technology would need to operate efficiently in the presence of water. Several different systems have been proposed to perform PoC-CCS, including solvent-based capture,[23,25–27] solid sorbent capture,[19,28–31] and calcium looping,[32,33] however to date only solvent-based systems have been piloted in real coal-fired powerplants on a large scale.

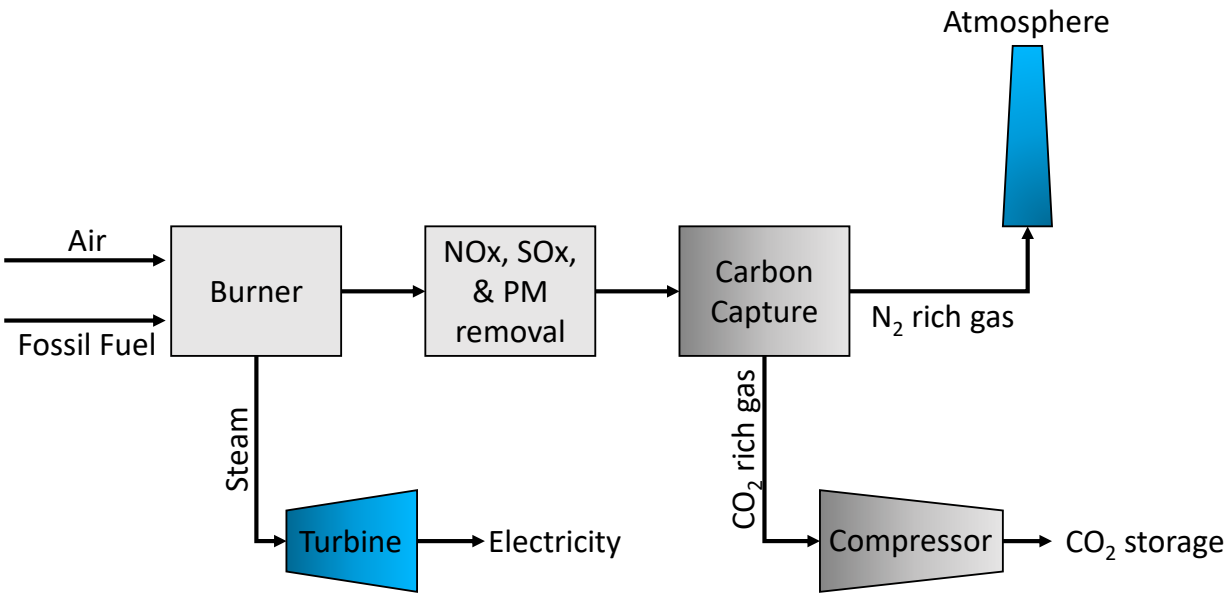
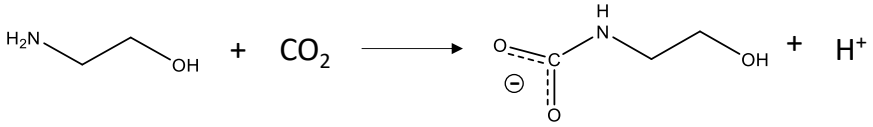


Figure 1.1 Schematic diagram of a fossil fuel burning powerplant equipped with a PoC-CCS system.

1.2.2. Solvent Based Systems

The only system currently in use at an industrial scale power plant are solvent based CO_2 capture systems. Also called liquid amines systems, the process involves bubbling the flue gas through an aqueous medium containing dissolved amines. The CO_2 then undergoes a chemical reaction with the liquid amines which removes the CO_2 from the flue gas, while the purified N_2 gas is allowed to flow through the medium unchanged. A common amine used in solvent-based systems, and the one currently in use at the Boundary Dam powerplant in Saskatchewan, is monoethanolamine (MEA).[34–36] The reaction between MEA and CO_2 is shown in equation 1.1, in which MEA reacts with CO_2 to form a carbamate product.



(1.1)

According to a recent report written in 2021 by Giannaris and coworkers,[37] as of October 2020 the CCUS unit at the Boundary Dam plant has captured over 3.6 million tonnes of CO_2 since 2014, where CO_2 is captured from only one out of the four boilers. This plant uses of a technology which relies on the formation of covalent bonds to capture the CO_2, meaning that recovery of the medium through the removal of CO_2 for storage and utilization requires a tremendous amount of energy. Furthermore, the

chemical bonds are broken through the application of heat to the separation column, filled with an aqueous solution with a high heat capacity. This energetic cost associated with running the capture unit is typically referred to as the *parasitic energy*, or the energy the plant needs to divert to run the capture unit. The need to heat the solution and break covalent bonds to recover the captured CO_2 means that liquid amine CO_2 capture systems have a high parasitic energy. In the case of the Boundary Dam power plant, the PoC-CCS unit is estimated to triple the cost of electricity generation.

Additional challenges also exist in implementing solvent-based capture systems, including the formation of amine foams described in the same report 2021 by Giannaris and coworkers.[37] This foam forms within the carbon capture units and causes degradation of the amine, reducing the efficacy of the capture unit. This means that the amine needs to be replaced continually, further increasing the operational costs of running the solvent-based capture system.

In a separate paper written by Giannaris and coworkers, the authors outline a feasibility study to implement a new CCUS system at another Saskpower coal-fired powerplant – the Shand power station in Saskatchewan. Building on the lessons learned from the Boundary Dam plant, they estimate that in a best-case scenario, the cost of capture could be reduced to 45 USD / tonne of CO_2 captured. Considering that such a powerplant generates millions of tonnes of CO_2 emissions annually and that the reported peak capture rate of the Boundary Dam plant was 3200 tonnes CO_2/day,[38] this cost is still exceedingly high. For comparison, the 2021 carbon pricing in Canada is listed at 32 USD / tonne CO_2 emitted,[37] meaning that the idealized capture unit at the Shand power station would still operate at a significantly higher cost to Saskpower than simply paying the tax. A more energy efficient and lower cost alternative would therefore be necessary to encourage widespread use of CCS and CCUS technologies.

1.2.3. Pressure Swing Adsorption

Pressure swing adsorption (PSA) is an alternative to solvent-based capture systems, that relies on the adsorption of gases onto the surface of porous materials, such as a zeolite. Although this technology is still in development for CCS/CCUS applications, it has been in use since the 1950s for a variety of other applications, such as separating H_2 from oil refinery off-gases,[39] and purifying O_2 for medical applications.[40] Unlike in solvent-based capture, PSA systems are packed with a porous solid material where the CO_2 can be recovered with the application of a vacuum, or if heat is used for the recovery, the energy requirements are lower than solvent-based systems because porous solids have relatively low heat capacities. Additionally, the guest molecules can be physically adsorbed rather than chemically

adsorbed, such that strong covalent bonds do not need to be broken to recover the CO_2 (Note that some high-profile porous solids developed for the PSA process, notably those developed by Long and coworkers,[41–48] utilize chemical adsorption). These factors combine to significantly reduce the energy requirements of PSA based CO_2 capture compared to that of a solvent-based systems.

A diagram of a simple PSA system is shown in Figure 1.2, which minimally includes two key phases: *adsorption* and *desorption*. During the adsorption phase, the flue gas is allowed to flow through a column packed with a porous material. As the gas flows through the column, CO_2 selectively adsorbs onto the surface and pores of the material, while the N_2 flows freely out of the top valve. Once the porous material is saturated with CO_2, the system switches to the *desorption* phase where the top valve is closed, and a vacuum is applied to the bottom (inlet) valve. This lowers the pressure inside the column and removes the adsorbed CO_2. When the pressure inside the column is lowered below atmospheric pressure, this PSA system is commonly referred to as a Vacuum Swing Adsorption (VSA) system. For the sake of simplicity, the terms PSA and VSA will be used interchangeably throughout this thesis. The cycle shown in Figure 1.2 is a simplified version of a typical PSA system, and in practice the process requires more complex cycle configurations. The configuration discussed in this thesis is a 4-stage light-particle pressurization (LPP) cycle and is described in detail in Chapter 4. In a technoeconomic study performed in 2018 by Bui and coworkers, it was concluded that the use of a PSA system for PoC-CCS would be economically viable and reduce the cost of carbon capture, provided the right sorbent material is found.[49]

1.3. Materials for Pressure Swing Adsorption

The biggest challenge in the development of a PSA system for PoC-CCS is the selection of an appropriate sorbent material.[49] Although PSA technologies are relatively mature, they have historically relied on a class of porous materials called zeolites.[50,51] Zeolites are a form of naturally occurring crystalline material composed of tetrahedral silicon or aluminum atoms, bridged by oxygen atoms. Due to the presence of tetrahedral aluminum atoms within the zeolites, the frameworks carry a negative charge which is balanced by a counter-ion, often Na^+. The presence of this framework charge and counter ions can cause zeolites to be unstable in the presence of water [52–54] or may result in water favourably binding to the zeolites preventing CO_2 adsorption,[55] making them poor candidates for PoC-CCS applications. As such, alternative materials are being explored for use in a PoC-CCS PSA system.

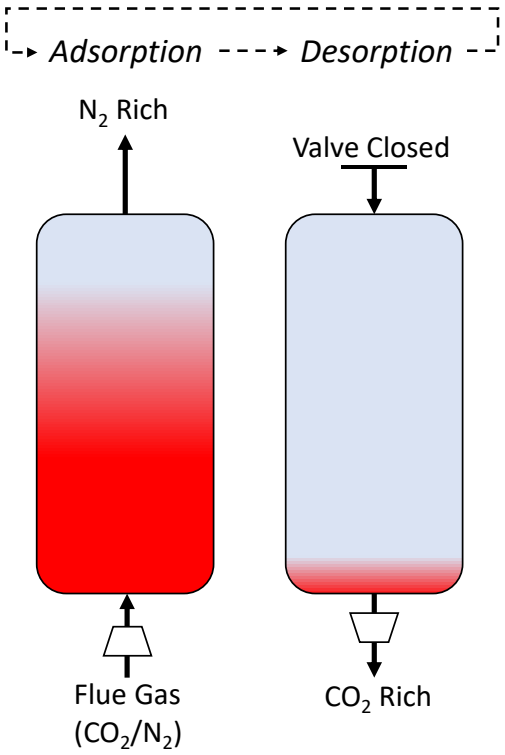


Figure 1.2 Schematic diagram of a simple pressure swing adsorption (PSA) system for a CO_2/N_2 separation process. The columns shown in this figure are packed with a solid sorbent material with the red regions representing areas of the column containing CO_2 rich gas.

1.3.1. Metal-Organic Frameworks

Amongst the most promising class of materials for use in a PSA system for PoC-CCS are metal-organic frameworks (MOFs).[56,57] MOFs are a relatively new class of materials composed of organic and inorganic building units, often called linkers, which self assemble to form 3-dimensional crystal lattice frameworks as depicted in Figure 1.3. These materials boast record breaking internal surface areas[58] and can be constructed out of thousands of known organic and inorganic linkers. Additionally, the same set of linkers can be assembled to form different MOFs with different topologies. The topologies are the underlying network structure of the MOF depicted in Figure 1.3. This diversity means that there is a seemingly infinite number of possible MOF structures, making them highly tunable for any gas separation application.

Although MOFs have a reputation for being unstable in the presence of water, there are now many MOFs which have been found to be water-stable or at least stable towards humid gas streams[59,60] and that retain their gas adsorption properties after prolonged exposure to humid conditions.[59] This combination of tunability and water stability makes MOFs ideal candidates for PoC-CCS applications and

the subject of numerous high-profile publications,[21,29,61–63] including a recent publication detailing the use of the MOF CALF-20 in carbon capture pilot plant at a cement production facility.[64]

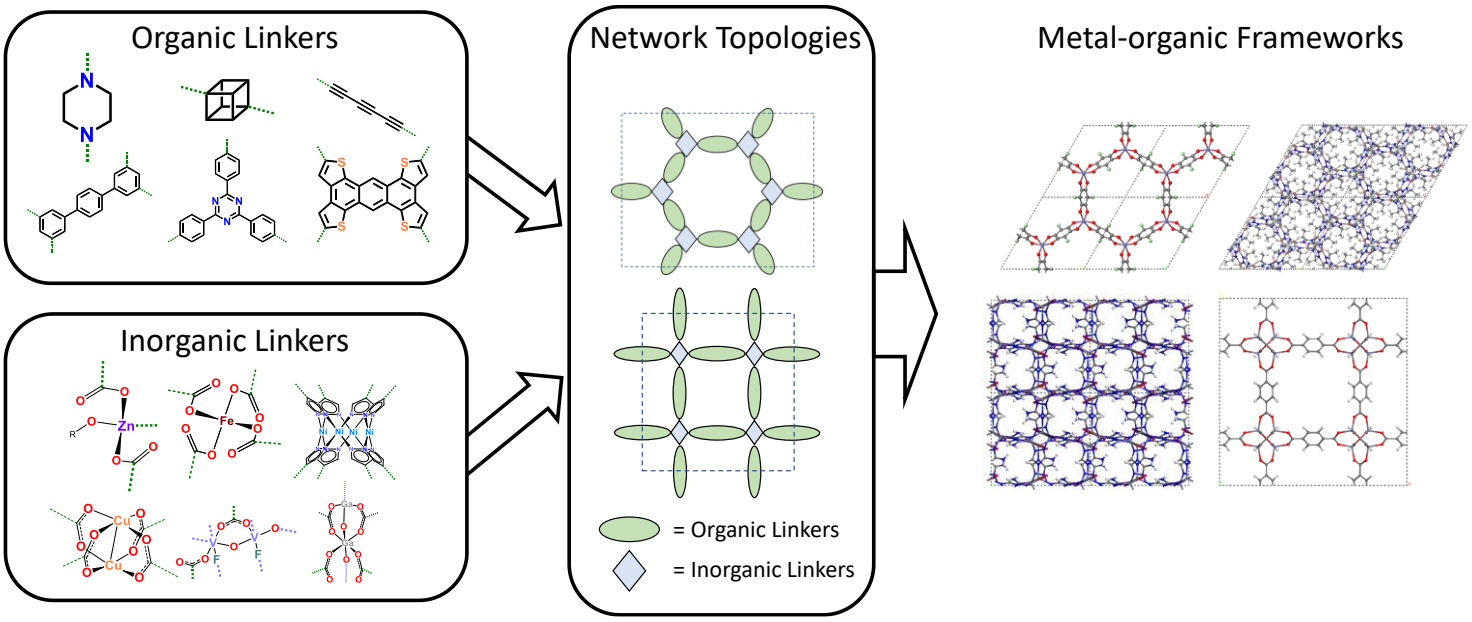


Figure 1.3 Diagram demonstrating the formation of MOFs from organic and inorganic linker, which assemble into different network topologies to form the final MOF structure.

1.4. Studying MOF systems

Although the discovery of MOFs is relatively recent,[65] they have garnered significant interest for a wide variety of applications, including PoC-CCS.[19,20,59,66,67] The methods used to study and test these materials vary significantly according to the expertise of the researchers performing those studies, ranging from atomistic level studies carried out by materials scientists and chemists, to large scale process level studies carried out by process and chemical engineers.

1.4.1. Atomistic Level Studies

The study of gas separations commonly performed by chemists and materials scientists are referred to in this thesis as atomistic level studies. These often involve bench scale experiments to test the physical adsorption properties with only small quantities of the material available. Adsorption experiments commonly involve exposing a material with a known quantity and pressure of a pure gas, for example CO_2, to determine how much of that gas can adsorb onto the surface of the material at a given set of conditions. By performing this experiment at a constant temperature while varying the pressure of the gas, calculating the amount adsorbed at each pressure point, the researcher can

estimate the strength of binding of the guest to the material. This plot of describing the amount absorbed as a function of pressure at a constant temperature is known as an isotherm.

To increase the pace and reduce the cost of MOF research, simulation techniques which reproduce such experimental adsorption metrics are often employed.[68–70] Using high-throughput screening techniques, thousands of MOFs can be tested for a given application. This technique was demonstrated by Snurr and coworkers[29,63] when 130,000 MOFs were evaluated for their ability to separate CO_2 from methane using grand canonical Monte Carlo simulations (see section 2.2.2). These studies rely on a series of metrics describing the adsorption behaviour of the gasses within the pores of the material at equilibrium. These metrics are then used to evaluate the material's ability to perform the desired gas separation. For full details on common atomistic performance metrics, see section 1.6.1.

1.4.2. Process Level Studies

The study of gas separation commonly performed by process and chemical engineers are referred to in this thesis as process level studies. These studies typically involve packing separation columns with the sorbent material and allowing the gas mixture to flow through. This form of experiment allows the researcher to directly observe the sorbent's ability to separate the gases in the mixture.

Although these studies provide crucial insight into a material's ability to perform the gas separation, they require significantly larger quantities of the sorbent material. As a result, Haghpanah and coworkers[71] developed a technique for simulating this process using the isotherm, heats of adsorption, and density of the sorbent material. At the time that the work in this thesis was performed, only a handful of studies had been completed using this technique.[72–74] For full details on the process level performance metrics, see section 1.6.2.

1.4.3. Disconnect Between Material Chemists and Process Engineers

There are many important considerations when vetting materials for PoC-CCS, however, researchers often rely on different criteria to perform their evaluations. A significant amount of research is being conducted at the atomistic scale using gas adsorption experiments and simulations,[29,62,63,75,76] while other research has been focused on the materials' performance in bench-scale separation columns.[77,78] Although both areas of focus are working towards the same goal of identifying the best possible material to separate CO_2 from N_2 at flue gas conditions in PSA systems, these experiments are typically performed separately and as a result have different perceptions of what defines a high

performance material. This disconnect between the atomistic and process level design of PSA systems for PoC-CCS was addressed in the 2018 *Mission Innovation Report:* "Accelerating Breakthrough Innovation in Carbon Capture, Utilization and Storage", highlighting the need to bridge the gap between materials science and process level design.[79]

The challenges in bridging this gap were discussed by Farmahini and coworkers,[80] where they outlined the challenge in consistently implementing an in silico multi-scale workflow, and the potential issues associated with performing molecular simulations on structured materials, where the characteristics of the structuring agents are not known. To date, a limited number of studies exist which attempt to bridge the gap between materials science and process engineers. Prior to the work performed in this thesis, only one such study existed published by Hasan and coworkers.[81] This study performed a hierarchical screening on a small set of completely silicious zeolite materials (zeolites containing only Si and O atoms) involving molecular simulations and a PSA simulator that used a simple capital expenditure model to estimate the cost of capture / tonne of CO_2. Although this study was the first to bridge the gap between materials scientists and process engineers, it relied on a relatively simplistic model to study a small set of materials with little diversity. To date, no large multi-scale screenings had been performed on a large diverse set of materials for the application of PoC-CCS in a PSA system.

1.5. Research Goals

A major shortcoming of MOF research in the field of PoC-CCS has been the disconnect between the materials chemists performing studies at the lab scale, and the process engineers designing the separation units themselves. The goals of the research presented in this thesis is to bridge that gap between the atomistic and process level, using sophisticated simulation techniques. In this way, the first large-scale screening was performed on a database of MOFs, studying the materials from their crystal structure to their industrial process performance.

The secondary goal of the research presented here-in was to study the results of this large-scale screening using sophisticated statistical and machine learning techniques in an attempt to relate the materials science targets discussed in section 1.6.1 to the process level performance metrics discussed in section 1.6.2. By relating these metrics, we hoped to develop insights allowing for more rapid vetting of MOF materials at the bench scale, to streamline the materials discovery process.

Finally, we aimed to develop predictive models which would be able to rapidly determine the process level performance of MOFs in an industrial PSA system, allowing for the rapid screening of thousands of materials without the need to high level process simulations or experiments, ultimately increasing the pace of materials discovery to find a viable solution to reducing GHG emissions from fossil-fuel burning powerplants.

1.6. Common Performance Metrics

In this section, supplementary information describing the common performance metrics used by both material scientists and process engineers is provided.

1.6.1. Material Science Targets

Materials science experiments, or lab-scale studies of materials, are performed on small quantities of MOF crystals for a given carbon capture and storage application. These studies typically involve performing measurements of gas adsorption uptakes for pure gases at a given set of temperature and pressure points, when the system is at equilibrium.[82,83] This allows researchers to determine the maximum amount of each gas that will be adsorbed onto the pores of a material at PoC-CCS conditions and provide an estimate of its ability to separate CO_2 from N_2 in a PSA system.[84,85] Typically, materials scientists use these equilibrium experiments because they are relatively easy and inexpensive to perform,[86] however, materials scientists are increasingly reporting breakthrough experiments which can directly measure the MOFs ability to separate gas mixtures.[87,88] These breakthrough experiments are significantly more expensive and require more technical expertise, and as a result the simple metrics described in this section are still commonly reported.[89,90] These equilibrium experiments are typically performed on a single gas (single component) and only require small quantities of the sorbent material. Competitive models are then applied to single component data to determine the material's preference to bind CO_2 over N_2, however this can be overcome using simulations, which are able to predict gas adsorption behaviour of gas mixtures.[29] In this section, I describe the metrics commonly used by material scientists to predict the PSA performance of a solid sorbent material at PoC-CCS conditions.

1.6.1.1. Isotherms

Understanding the gas adsorption behaviour of a MOF at a range of pressures is critical in determining its performance in a PSA system. This behaviour is typically studied using gas adsorption isotherms, a plot of the gas adsorption uptake, or the amount of a gas adsorbed onto the pores of the material, as a function of pressure at a constant temperature. The most commonly observed isotherm behaviour is known as a Type I or a Langmuir isotherm and can be modelled using equation 1.2, where θ is the gas uptake, θ_{sat} is the saturation uptake, b is the fitted Langmuir constant, and P is the partial pressure of the gas.

$$\theta = \frac{\theta_{sat} b P}{1 + bP} \tag{1.2}$$

The behaviour of this function is demonstrated in Figure 1.4, which shows three isotherms with the same saturation uptake of 2.0 mmol of captured gas/g of material (mmol/g). In this function, the saturation uptake is a fitted parameter and is the value the function will reach when P approaches infinity. The Langmuir constant b can be thought of as the rate at which the isotherm approaches the saturation uptake, demonstrated by the three isotherms in Figure 1.4 for values of 5.0 bar^{-1} (blue line), 1.0 bar^{-1} (green line), and 0.2 bar^{-1} (orange line). The Langmuir constant b can also be defined as the strength of the binding between the guest molecules and the material since materials with stronger interactions between the guest and the framework will reach the saturation uptake at lower pressures than those with lower values of b.

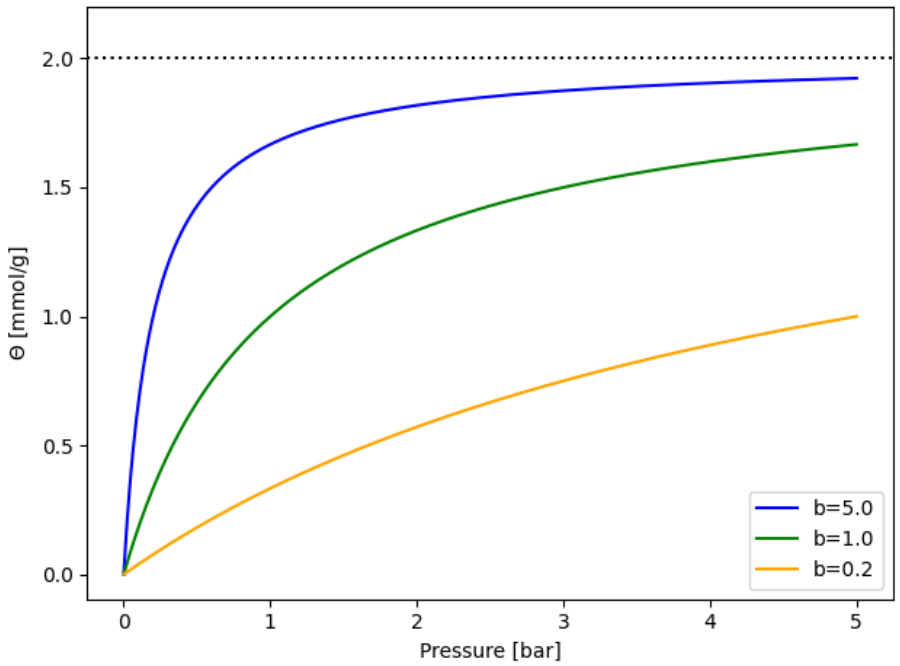


Figure 1.4 Three isotherms with the same saturation uptake of 2.0 mmol/g (black dotted line) and different values for the Langmuir constant, $\boldsymbol{b}$, of 5.0 bar^{-1} (blue line), 1.0 bar^{-1} (green line), and 0.2 bar^{-1} (orange line).

The key assumption made by the Langmuir isotherm is that the binding within the pores of the material is uniform, meaning all guests are occupying identical binding sites. In practice this is not always the case, with some MOFs containing many binding motifs within their pores.[91,92] An expansion on the Langmuir isotherm attempts to address this assumption by adding additional parameters for a second binding site. This is known as the dual-site Langmuir (DSL), defined by equation 1.3, and describes the uptake θ as the sum of the loadings for a strong and weak binding site. In the DSL formula, b and d are the fitted Langmuir parameters for the strong and weak binding sites, respectively. The order that the binding sites appear in equation 1.3 becomes important when dealing with competitive Langmuir models (see section 1.6.1.5).

$$\theta = \frac{\theta_{sat,strong} bP}{1 + bP} + \frac{\theta_{sat,weak} dP}{1 + dP} \tag{1.3}$$

1.6.1.2. Working Capacity

Although the uptake capacity at adsorption conditions tells a researcher the maximum amount of gas the material will be able to adsorb at a given set of conditions, a PSA system like the one presented in Figure 1.2 will need to be cycled hundreds of times for it to be economically viable. When the capture

medium is recovered through the application of a vacuum, the lowest pressure achieved will need to be selected to minimize the cost of capture. Since the vacuum pumps and compressors are the largest contributors to the parasitic energy during a PSA cycle, completely evacuating the column would be energetically expensive and not industrially feasible. This means that some of the captured gas will remain in the column after the desorption step is complete and the sorbent material will not be fully recovered. The working capacity (Δw) is used to calculate an effective adsorption capacity over the course of the PSA cycle, by considering the difference in uptake at adsorption and desorption conditions (equation 1.4). By fitting isotherm parameters to experimental or simulated adsorption data, the working capacity can be estimated for any pair of adsorption-desorption conditions, as demonstrated in Figure 1.5.

$$\Delta w = \theta_{adsorption} - \theta_{desorption} \tag{1.4}$$

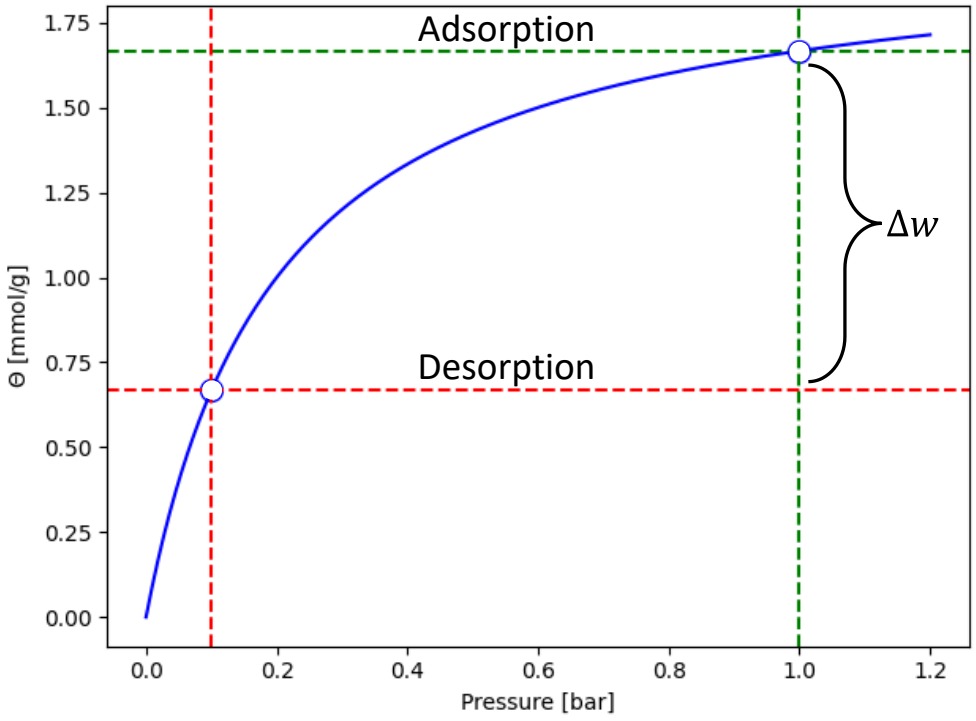


Figure 1.5 Demonstration of the working capacity calculation on an isotherm with $\boldsymbol{\theta_{sat}}$ = 2.0 mmol/g and $\boldsymbol{b}$ = 5.0 bar^{-1}, showing the uptake and pressure at adsorption conditions (green dashed lines) and at the desorption conditions (red dashed lines).

1.6.1.3. Isosteric Heat of Adsorption

When the adsorption of a gas onto the surface of a material takes place, it causes a release of heat. This heat is called the isosteric heat of adsorption and is the change in enthalpy resulting from the gas transitioning from the bulk to the adsorbed phase. This property is specific to each material and gas pair and is often considered when performing estimates of parasitic energy.[93] The assumption is that the isosteric heat of adsorption, Q_{st}, is the energy penalty associated with desorption, or removing adsorbed molecules from the surface of the material. This implies that the higher the Q_{st}, the greater the parasitic energy. In contrast, an exceedingly low Q_{st} may indicate weak adsorption of the gas onto the framework or poor selectivity as other gases may also weakly adsorb to the material's surface.

Measuring Q_{st} experimentally requires adsorption isotherms at different temperatures. The isotherms are then compared using the pressure points, p, that yield the same loading θ which are then fit to the Clausius-Clapeyron equation given by equation 1.5. In this equation, R is the gas constant, T is the temperature, and c is a fitted constant. For a Q_{st} more representative of the pure interaction energy between a single guest and the material with no influence from other guests in the system, the zero-coverage isosteric heat of adsorption (Q_{st}^0) can be estimated for a point where $\theta = 0$.

$$\ln (p)_\theta = -\left(\frac{Q_{st}}{R}\right)\frac{1}{T} + c \tag{1.5}$$

1.6.1.4. Selectivity

The aim of a pressure swing adsorption system is to separate one or more gases from a mixture. This means that consideration of the equilibrium loadings and heats of adsorption of the individual gases in isolation does not provide a complete picture of a material's performance. A commonly used metric to estimate a material's ability to separate the component gases is known as the selectivity of the material, S_{ij}, calculated using equation 1.6. This metric considers the ratio of the loadings for the two gases weighted by their partial pressures at the adsorption conditions. This provides insight into a material's preference to adsorb the target gas, for example CO_2, over another gas, example N_2. Importantly, although the selectivity should be measured for gas mixtures to accurately represent the conditions in a PSA column, it is often measured using single component experiments.[94–96] This metric is considered to be of great importance to material scientists when selecting materials for PoC-CCS in a PSA system.[97–99]

$$S_{ij} = \frac{\theta_i / p_i}{\theta_j / p_j} \tag{1.6}$$

1.6.1.5. Competitive Models

As mentioned previously, equilibrium properties determined experimentally often rely on single component gases instead of the gas mixtures relevant to the process, since single component experiments are less complex and can be performed with smaller quantities of sorbent. This means that the loadings and working capacities do not consider the competition between the two gases present in a binary mixture, such as a flue gas. Competitive isotherm models have been developed to calculate the loadings of gas mixtures based on single component adsorption data, with the competitive model most relevant to the work in this thesis being the competitive DSL model. This model, defined in equation 1.7, is the IAST[100] solution for mixing DSL isotherms for multi-component gases. The equation weights the loading of component i, θ_i, according to the strength of the corresponding binding sites for n components in the gas mixture. Recall that in section 1.6.1.1 it was discussed that the order of the binding sites in the DSL equation is significant. This is because the key assumption made by the competitive DSL model is that the strong sites only compete with the other strong sites, and the weak sites only compete with the other weak sites.

$$\theta_i = \frac{\theta_{sat,strong,i} b_i p_i}{1 + \sum_{j=1}^{n} b_j p_j} + \frac{\theta_{sat,weak,i} d_i p_i}{1 + \sum_{j=1}^{n} d_j p_j} \tag{1.7}$$

1.6.1.6. Ideal Absorbed Solution Theory (IAST)

The Ideal Absorbed Solution Theory (IAST)[100] is a method that models the competitive adsorption characteristics of a material exposed to a gas mixture based on single component adsorption data. It relies on fitted isotherm parameters, like those described in section 1.6.1.1, to calculate the spreading pressure of each gas type. The spreading pressure for gas species i, Π_i, is defined as the pressure exerted by a gas on an adsorbent surface and is calculated using equation 1.8,[100] where c_i is the multi-component concentration of species i, c_i^0 is the single component concentration, and $\theta_i(c_i)$ is the single component isotherm function.

$$\Pi_i = \int_0^{c_i^0} \frac{\theta_i(c_i)}{c_i} \delta p \tag{1.8}$$

The aim of the IAST method is to equate the spreading pressure of all component gases in the mixture. This is done by minimizing the difference between the spreading pressures by performing iterative adjustments of the c_i^0 value for each gas. The value of c_i^0 can then be expressed using equation 1.9, where z_i is the fraction of species i in the adsorbed phase, calculated using equation 1.10. Once the value of c_i^0 is optimized for each gas species, the final loadings can then be determined using equation 1.11. Due to the difficulty associated with measuring the loadings of individual gases in gas mixtures, IAST is often used to calculate working capacities and selectivities based on single component adsorption experiments. This is vital as these properties require consideration of competing gas molecules to better represent the conditions in a PSA column.

$$c_i^0 = \frac{c_i}{z_i} \tag{1.9}$$

$$z_i = \frac{\theta_i}{\sum_j^n \theta_j} \tag{1.10}$$

$$\frac{1}{\sum_i^n \theta_i} = \sum_i^n \frac{z_i}{\theta_i(c_i)} \tag{1.11}$$

1.6.2. Process Level Targets

As mentioned in section 1.6, lab-scale equilibrium and process level studies have historically been performed separately, with little to no consensus as to what constitutes a high performing material. The lack of collaboration between these two important facets of the materials discovery process has resulted in a disconnect between materials chemists and process engineers. This disconnect has resulted in many MOF chemists relying on a wide range of performance metrics to boast about a material's performance with no concrete understanding as to how they relate to actual process level performance.

Process engineers measure the performance of industrial PSA systems through direct observation, reporting metrics that explicitly describe the separation performance and cost of the system. This is in contrast to the more abstract metrics to estimate PSA performance such as those used by material scientists, discussed in section 1.6.1. Instead, the process level performance metrics are lifted directly from the separation apparatus and include the purity of the captured CO_2, the percentage of CO_2 recovered from the flue gas stream, the energy required to run the separation, and overall productivity of the material.

1.6.2.1. Purity and Recovery Targets

The most important question to answer when vetting a material at the process level is whether the process will adequately separate a gas mixture. Without an answer to this question, the energetic cost of running a separation column and the overall productivity of a material is moot. The process's ability to separate the gas is measured by two metrics: the purity of captured CO_2 and the recovery percentage of CO_2 captured from the flue gas.

The purity of the captured gas is an important consideration since the captured gas needs to be compressed to supercritical levels for energy efficient, long-distance transport to a storage site.[24] The lower the purity of CO_2 in the captured gas, the more energy is required to compress the same quantity of CO_2. Moreover, the more impurities in the CO_2, the more difficult it is to make the fluid supercritical. To make sure that sorbent-based PoC-CCS technologies are economically viable, the United States Department of Energy (DoE) has set a series of benchmarks that these separation processes need to meet to be considered for industrial use. The first of those requirements is a 95% purity of the captured gas,[101] meaning the gas being compressed needs to have a mole fraction greater than or equal to 0.95. This value was chosen as it is the minimum purity required for injection into a pipeline.[101]

The recovery of CO_2 from the flue gas stream is defined as the percentage of the CO_2 that enters the column that is captured for storage. The US DoE has also set a minimum target for sorbent-based PoC-CCS systems of 90% recovery of CO_2 from the flue gas,[101] meaning that, at most, 10% of the CO_2 that enters the column is emitted to the atmosphere while the remaining 90% is captured and stored. While the 95% purity target is required for technical reasons, the 90% recovery target is more arbitrary.

Together, these purity and recovery targets set by the US DoE are commonly referred to as the DoE purity-recovery targets (DoE-PRT),[101] and are considered the absolute minimum requirement for a sorbent-based PoC-CCS systems.

1.6.2.2. Parasitic Energy

Once a material is found that can meet the DoE-PRT, the parasitic energy (PE) is considered. The parasitic energy includes the energy required to run the separation process and compress the captured gas to transport conditions (150 bar).[24] This performance metric is considered to be one of the most crucial as it determines a significant portion of the operational cost of running a PoC-CCS capture system. Since costing estimates are complex and will vary significantly according to geographic location

and political climate surrounding the powerplant, the parasitic energy acts as a more holistic metric defining the cost of capture.[102] The US DoE has also provided a parasitic energy target that sorbent-based PoC-CCS systems need to meet to be considered for industrial use. This target states that a sorbent-based capture systems require a 30% improvement in PE over existing solvent-based capture systems.[101] Although this target may seem arbitrary, solvent-based capture systems for PoC-CCS are a comparatively mature technology, whereas sorbent-based systems require significant research investment before they are industrially viable. As such, this target reflects the performance required to justify the additional expense of developing this technology.

1.6.2.3. Productivity

The final consideration for process performance is the productivity of the material. The productivity for a PoC-CCS system can be defined as the amount of CO_2 captured by the PSA unit in a day per cubic meter of sorbent. Although no US DOE targets have been defined for the productivity of a sorbent, ideally this parameter would need to be maximized to reduce the amount of material needed and ultimately the size of the capture plant. Since productivity dictates the size of the plant required to continually capture CO_2 from a powerplant's flue gas, it is also inherently linked to the initial capital cost of building the capture plant. Smaller productivity would mean larger capture plants, likely including additional columns. Inclusion of additional columns increases the complexity of the plant resulting in more possible points of failure. As a result, to effectively reduce the cost of capture, a material would need to display low parasitic energy while simultaneously having exceptionally high productivity. This further complicates the optimization of a PSA system for PoC-CCS, as the productivity and parasitic energy are competing properties, meaning that improvements in parasitic energy are often associated with losses in productivity.[19]

1.7. Overview of Chapters

The contents of this thesis are briefly described and summarized herein. In Chapter 2, I outline in detail the computational methods used throughout this thesis to calculate gas adsorption properties of MOFs using Grand Canonical Monte Carlo (GCMC) simulations. I also describe a sophisticated pressure swing adsorption simulator, used in Chapter 4 to perform a large-scale screening of MOFs, as well as a description of the optimization and machine learning techniques used throughout this thesis.

In Chapter 3, I present my work to develop a Guest Atom Localization Algorithm (GALA) to locate binding sites of guest molecules within the pores of nanoporous materials based on GCMC simulations. Using a set of experimentally determined binding sites, and a series of hypothetical materials with binding sites located by hand, the optimization of the GALA code to calculate the binding sites of CO_2 for use in high-throughput screening is described in detail. The GALA code was successfully optimized for CO_2 and the resulting binding sites were found to be in good agreement to those located experimentally. The work in this chapter is the subject of a manuscript currently in progress.

In Chapter 4, I present a large-scale screening of MOFs from the Computation Ready Experimental (CoRE) MOF database[103] from pores to process. Over the course of this screening, MOFs were studied from simple crystal structures to process level performance, calculating charges using the REPEAT method, CO_2 and N_2 isotherms, and performing full process optimizations in a 4-stage light particle pressurization (LPP) cycle using a custom genetic algorithm (GA). From this study, 482 MOFs were identified that could meet the DoE-PRT, with 223 of those MOFs outperforming the DoE's PE target. The top 10 materials according to PE, productivity, and a performance metric which balances PE and productivity are presented, with the best overall material identified as being IISERP-MOF-2, which was able to achieve a parasitic energy of 221.36 kWh/tonne CO_2 captured while maintaining a productivity of 3.73 TPD CO_2 / m^3 of adsorbent. The work presented in this chapter formed the basis of a publication in the journal *Environmental Science and Technology.*

In Chapter 5, I present the in-depth analysis of the resulting data from the screening in Chapter 4. The overall goal of this work was to relate a series of equilibrium level performance metrics commonly used describe MOFs to their process level performance. Using a wide range of advanced machine learning techniques including Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Decision Trees, and the Random Forest, the equilibrium metrics failed to predict the PE or productivity of the MOFs studied in Chapter 4. This demonstrated that studying these materials at the lab scale does not provide a complete picture of a material's performance. Further it was found that some metrics were able to predict whether or not a material could meet the DoE-PRT with some success, relating equilibrium metrics to process level performance for the first time. It was concluded that the N_2 adsorption characteristics play a more significant role in determining a material's ability to meet the DoE-PRT than the CO_2 adsorption characteristics, a conclusion that was contrary to the belief in the MOF community at the time this work was published in the journal *Environmental Science and Technology.*

In Chapter 6, the CO_2 and N_2 binding sites were located using GALA in a series of MOFs with process performance determined in the screening described in Chapter 4. The conclusions of the work done in Chapter 5, that N_2 and not CO_2 drove the material's ability to meet the DoE-PRT, led to the hypothesis that the N_2 binding site locations played an important role in that behaviour. A comparison was performed between single component and binary component probability distributions used by GALA to calculate the bindings sites in 704 MOFs whose ability to meet the DoE-PRT was known. It was found that MOFs which showed low similarity between the single component and binary component N_2 probability distributions had a higher tendency of meeting the DoE-PRT. The implication was that the N_2 binding sites in the presence of CO_2 were displaced, reducing the uptake of N_2 in the presence of CO_2 and improving the MOF's ability to meet the DoE-PRT. This secondary hypothesis was tested by comparing the N_2 uptake ratios and performing visual inspections of the CO_2 and N_2 binding sites over a range of Tanimoto similarity values. The conclusion of this chapter is that the displacement of N_2 binding sites by CO_2 plays an important role in a material's ability to meet the DoE-PRT. The work performed in this chapter is the subject of a manuscript currently in progress.

In Chapter 7, the competitive dual-site Langmuir (DSL) model used by the PSA simulation code is evaluated against binary GCMC simulations and competitive loadings calculated using a linear interpolation model. This linear interpolation model was built using a grid of multi-component GCMC data, allowing for a more direct prediction of competitive loadings in the PSA column. The goal of this work was to determine whether the existing PSA model, which relies on single component adsorption data, could be improved using multi-component simulations. To test the two methods, a randomized set of 1000 points at different temperatures, pressures, and mole fractions relevant to the conditions within a PSA column were generated using Latin Hypercube Sampling. The loadings for 85 MOFs at every point were calculated using GCMC as a benchmark and compared to the competitive DSL and interpolated loadings using a percent mean absolute deviation function. It was concluded that the predicted loadings using the linear interpolation model were more accurate than the competitive DSL for both CO_2 and N_2 loadings. This work was then duplicated for 101 MOFs at temperature swing adsorption (TSA) conditions due to the processes' relevance to future work being performed within the Woo lab. The analysis performed on the TSA system was in agreement with the results seen for the PSA system. The chapter concludes that the implementation of a linear interpolation model into the PSA simulator could be beneficial, however more work would be required to compare the PSA results with both systems to determine the magnitude of this change.

In Chapter 8, I describe the work done in developing a surrogate model for the PSA simulator using a suite of artificial neural networks (ANNs). This suite of ANNs contains 4 unique regression models which predict purity, recovery, parasitic energy, and productivity of a material given the material's isotherm parameters, heat of adsorption, crystal density, and the conditions at the defined process point. Construction of the training sets for each individual model is described in detail, along with an initial proof of concept test involving the fitting of 4 Gradient Boosted Decision Tree (GBDT) models. This initial proof of concept showed that unoptimized GBDT models were able to predict all four metrics with high accuracy, indicating that the development of more computationally intensive ANN models was possible. Using an iterative grid-search algorithm, the structure of the ANNs were optimized for all four target variables and the model suite was constructed. Once the application was complete, a full screening of the CoRE database was performed, and the result of that screening are presented.

In Chapter 9, I present general conclusions for each chapter and discuss possible directions for future projects.

1.8. References

1. *Climate Change 2021 Working Group I contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Summary for Policymakers*.

2. U.S. E.P.A. Climate Change Division. *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2019*. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks (2021).

3. Jiang, S., Chen, Z., Shan, L., Chen, X. & Wang, H. Committed CO_2 emissions of China's coal-fired power generators from 1993 to 2013. *Energy Policy* **104**, 295–302 (2017).

4. Shearer, C., Fofrich, R. & Davis, S. J. Future CO_2 emissions and electricity generation from proposed coal-fired power plants in India. Earth's Future 5: 408–416. (2017).

5. Ahmadi, M. H., Ghazvini, M., Sadeghzadeh, M., Alhuyi Nazari, M., Kumar, R., Naeimi, A. & Ming, T. Solar power technology for electricity generation: A critical review. *Energy Science and Engineering* vol. 6 340–361 (2018).

6. He, Y. L., Wang, K., Qiu, Y., Du, B. C., Liang, Q. & Du, S. Review of the solar flux distribution in concentrated solar power: Non-uniform features, challenges, and solutions. *Applied Thermal Engineering* vol. 149 448–474 (2019).

7. Chen, H., Zhang, R., Chen, X., Zeng, G., Kobera, L., Abbrent, S., Zhang, B., Chen, W., Xu, G., Oh, J., Kang, S.-H., Chen, S., Yang, C., Brus, J., Hou, J., Gao, F., Li, Y. & Li, Y. A guest-assisted molecular-organization approach for >17% efficiency organic solar cells using environmentally friendly solvents. *Nature Energy* (2021).

8. Cherp, A., Vinichenko, V., Tosun, J., Gordon, J. A. & Jewell, J. National growth dynamics of wind and solar power compared to the growth required for global climate targets. *Nature Energy* **6**, 742–754 (2021).

9. Kumar, A., Madden, D. G., Lusi, M., Chen, K.-J., Daniels, E. A., Curtin, T., Perry, J. J. & Zaworotko, M. J. Direct air capture of CO_2 by physisorbent materials. *Angewandte Chemie* **127**, 14580–14585 (2015).

10. Gambhir, A. & Tavoni, M. Direct air carbon capture and sequestration: How it works and how it could contribute to climate-change mitigation. *One Earth* vol. 1 405–409 (2019).

11. Seipp, C. A., Williams, N. J., Kidder, M. K. & Custelcean, R. CO_2 capture from ambient air by crystallization with a guanidine sorbent. *Angewandte Chemie - International Edition* **56**, 1042–1045 (2017).

12. Zhou, T., Shi, H., Ding, X. & Zhou, Y. Thermodynamic modeling and rational design of ionic liquids for pre-combustion carbon capture. *Chemical Engineering Science* **229**, (2021).

13. Erlach, B., Schmidt, M. & Tsatsaronis, G. Comparison of carbon capture IGCC with pre-combustion decarbonisation and with chemical-looping combustion. *Energy* **36**, 3804–3815 (2011).

14. Pardemann, R. & Meyer, B. Pre-combustion carbon capture. in *Handbook of Clean Energy Systems* 1–28 (John Wiley & Sons, Ltd, 2015).

15. Theo, W. L., Lim, J. S., Hashim, H., Mustaffa, A. A. & Ho, W. S. Review of pre-combustion capture and ionic liquid in carbon capture and storage. *Applied Energy* vol. 183 1633–1663 (2016).

16. Zhu, Y., Chen, M., Yang, Q., Alshwaikh, M. J. M., Zhou, H., Li, J., Liu, Z., Zhao, H., Zheng, C., Bartocci, P. & Fantozzi, F. Life cycle water consumption for oxyfuel combustion power generation with carbon capture and storage. *Journal of Cleaner Production* **281**, (2021).

17. Seddighi, S., Clough, P. T., Anthony, E. J., Hughes, R. W. & Lu, P. Scale-up challenges and opportunities for carbon capture by oxy-fuel circulating fluidized beds. *Applied Energy* vol. 232 527–542 (2018).

18. Kunze, C. & Spliethoff, H. Assessment of oxy-fuel, pre- and post-combustion-based carbon capture for future IGCC plants. *Applied Energy* **94**, 109–116 (2012).

19. Burns, T. D., Pai, K. N., Subraveti, S. G., Collins, S. P., Krykunov, M., Rajendran, A. & Woo, T. K. Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

20. Chao, C., Deng, Y., Dewil, R., Baeyens, J. & Fan, X. Post-combustion carbon capture. *Renewable and Sustainable Energy Reviews* **138**, (2021).

21. Liang, L., Liu, C., Jiang, F., Chen, Q., Zhang, L., Xue, H., Jiang, H. L., Qian, J., Yuan, D. & Hong, M. Carbon dioxide capture and conversion by an acid-base resistant metal-organic framework. *Nature Communications* **8**, (2017).

22. Jockenhoevel, T., Schneider, R. & Rode, H. Development of an economic post-combustion carbon capture process. in *Energy Procedia* vol. 1 1043–1050 (2009).

23. Lucquiaud, M. & Gibbins, J. Retrofitting CO_2 capture ready fossil plants with post-combustion capture. Part 1: Requirements for supercritical pulverized coal plants using solvent-based flue gas scrubbing. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy* **223**, 213–226 (2009).

24. Aspelund, A. & Jordal, K. Gas conditioning-The interface between CO_2 capture and transport. *International Journal of Greenhouse Gas Control* **1**, 343–354 (2007).

25. Liang, Z. (Henry), Rongwong, W., Liu, H., Fu, K., Gao, H., Cao, F., Zhang, R., Sema, T., Henni, A., Sumon, K., Nath, D., Gelowitz, D., Srisang, W., Saiwan, C., Benamor, A., Al-Marri, M., Shi, H., Supap, T., Chan, C., *et al.* Recent progress and new developments in post-combustion carbon-capture technology with amine based solvents. *International Journal of Greenhouse Gas Control* vol. 40 26–54 (2015).

26. Knudsen, J. N., Andersen, J., Jensen, J. N. & Biede, O. Evaluation of process upgrades and novel solvents for the post combustion CO_2 capture process in pilot-scale. *Energy Procedia* **4**, 1558–1565 (2011).

27. Knudsen, J. N., Jensen, J. N., Vilhelmsen, P. J. & Biede, O. Experience with CO_2 capture from coal flue gas in pilot-scale: Testing of different amine solvents. *Energy Procedia* **1**, 783–790 (2009).

28. Torrisi, A., Bell, R. G. & Mellot-Draznieks, C. Functionalized MOFs for enhanced CO_2 capture. *Crystal Growth and Design* **10**, 2839–2841 (2010).

29. Lin, L.-C., Berger, A. H., Martin, R. L., Kim, J., Swisher, J. a., Jariwala, K., Rycroft, C. H., Bhown, A. S., Deem, M. W., Haranczyk, M. & Smit, B. In silico screening of carbon-capture materials. *Nature Materials* **11**, 633–641 (2012).

30. Li, J.-R., Ma, Y., McCarthy, M. C., Sculley, J., Yu, J., Jeong, H.-K., Balbuena, P. B. & Zhou, H.-C. Carbon dioxide capture-related gas adsorption and separation in metal-organic frameworks. *Coordination Chemistry Reviews* **255**, 1791–1823 (2011).

31. D'Alessandro, D. M., Smit, B. & Long, J. R. Carbon dioxide capture: Prospects for new materials. *Angewandte Chemie International Edition* **49**, 6058–6082 (2010).

32. Blamey, J., Anthony, E. J., Wang, J. & Fennell, P. S. The calcium looping cycle for large-scale CO_2 capture. *Progress in Energy and Combustion Science* vol. 36 260–279 (2010).

33. Zhu, L., Jiang, P. & Fan, J. Comparison of carbon capture IGCC with chemical-looping combustion and with calcium-looping process driven by coal for power generation. *Chemical Engineering Research and Design* **104**, 110–124 (2015).

34. Lee, J., Kolawole, T. & Attidekou, P. Carbon capture from a simulated flue gas using a rotating packed bed adsorber and mono ethanol amine (MEA). in *Energy Procedia* vol. 114 1834–1840 (Elsevier Ltd, 2017).

35. Gladich, I., Sinopoli, A., Abotaleb, A. & Pietrucci, F. Stability of a monoethanolamine-co2zwitterion at the vapor/liquid water interface: Implications for low partial pressure carbon capture technologies. *Journal of Physical Chemistry B* **125**, 4890–4897 (2021).

36. Ayittey, F. K., Saptoro, A., Kumar, P. & Wong, M. K. Energy-saving process configurations for monoethanolamine-based CO_2 capture system. *Asia-Pacific Journal of Chemical Engineering* **16**, (2021).

37. Giannaris, S., Janowczyk, D., Ruffini, J., Hill, K., Jacobs, B., Bruce, C., Feng, Y. & Srisang, W. *SaskPower's Boundary Dam Unit 3 Carbon Capture Facility-The Journey to Achieving Reliability*. (2021).

38. Giannaris, S., Bruce, C., Jacobs, B., Srisang, W. & Janowczyk, D. Implementing a second generation CCS facility on a coal fired power station – results of a feasibility study to retrofit SaskPower's Shand power station with CCS. *Greenhouse Gases: Science and Technology* **10**, 506–518 (2020).

39. Wiessner, F. G. Basics and industrial applications of pressure swing adsorption (PSA), the modern way to separate gas. *Gas Separation & Purification* **2**, 115–119 (1988).

40. Jee, J. G., Lee, J. S. & Lee, C. H. Air separation by a small-scale two-bed medical O_2 pressure swing adsorption. *Industrial and Engineering Chemistry Research* **40**, 3647–3658 (2001).

41. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

42. Dinakar, B., Forse, A. C., Jiang, H. Z. H., Zhu, Z., Lee, J. H., Kim, E. J., Parker, S. T., Pollak, C. J., Siegelman, R. L., Milner, P. J., Reimer, J. A. & Long, J. R. Overcoming metastable CO_2 adsorption in a bulky diamine-appended metal-organic framework. *Journal of the American Chemical Society* **143**, 15258–15270 (2021).

43. Milner, P. J., Siegelman, R. L., Forse, A. C., Gonzalez, M. I., Runčevski, T., Martell, J. D., Reimer, J. A. & Long, J. R. A Diaminopropane-appended metal-organic framework enabling efficient CO_2 capture from coal flue gas via a mixed adsorption mechanism. *Journal of the American Chemical Society* **139**, 13541–13553 (2017).

44. Siegelman, R. L., McDonald, T. M., Gonzalez, M. I., Martell, J. D., Milner, P. J., Mason, J. A., Berger, A. H., Bhown, A. S. & Long, J. R. Controlling cooperative CO_2 adsorption in diamine-appended Mg2(dobpdc) metal-organic frameworks. *Journal of the American Chemical Society* **139**, 10526–10538 (2017).

45. Drisdell, W. S., Poloni, R., McDonald, T. M., Pascal, T. A., Wan, L. F., Pemmaraju, C. das, Vlaisavljevich, B., Odoh, S. O., Neaton, J. B., Long, J. R., Prendergast, D. & Kortright, J. B. Probing the mechanism of CO_2 capture in diamine-appended metal-organic frameworks using measured and simulated X-ray spectroscopy. *Physical Chemistry Chemical Physics* **17**, 21448–21457 (2015).

46. McDonald, T. M., Lee, W. R., Mason, J. A., Wiers, B. M., Hong, C. S. & Long, J. R. Capture of carbon dioxide from air and flue gas in the alkylamine-appended metal-organic framework mmen-Mg 2(dobpdc). *Journal of the American Chemical Society* **134**, 7056–7065 (2012).

47. Demessence, A., D'Alessandro, D. M., Foo, M. L. & Long, J. R. Strong CO_2 binding in a water-stable, triazolate-bridged metal-organic framework functionalized with ethylenediamine. *Journal of the American Chemical Society* **131**, 8784–8786 (2009).

48. McDonald, T. M., D'Alessandro, D. M., Krishna, R. & Long, J. R. Enhanced carbon dioxide capture upon incorporation of N,N'- dimethylethylenediamine in the metal-organic framework CuBTTri. *Chemical Science* **2**, 2022–2028 (2011).

49. Bui, M., Adjiman, C. S., Bardow, A., Anthony, E. J., Boston, A., Brown, S., Fennell, P. S., Fuss, S., Galindo, A., Hackett, L. A., Hallett, J. P., Herzog, H. J., Jackson, G., Kemper, J., Krevor, S., Maitland, G. C., Matuszewski, M., Metcalfe, I. S., Petit, C., *et al.* Carbon capture and storage (CCS): The way forward. *Energy and Environmental Science* **11**, 1062–1176 (2018).

50. Grande, C. A. & Rodrigues, A. E. Propane/propylene separation by Pressure Swing Adsorption using zeolite 4A. *Industrial and Engineering Chemistry Research* vol. 44 8815–8829 (2005).

51. Alonso-Vicario, A., Ochoa-Gómez, J. R., Gil-Río, S., Gómez-Jiménez-Aberasturi, O., Ramírez-López, C. A., Torrecilla-Soria, J. & Domínguez, A. Purification and upgrading of biogas by pressure swing adsorption on synthetic and natural zeolites. *Microporous and Mesoporous Materials* **134**, 100–107 (2010).

52. Lutz, W., Toufar, H., Kurzhals, R. & Suckow, M. *Investigation and Modeling of the Hydrothermal Stability of Technically Relevant Zeolites *. Adsorption* vol. 11 (2005).

53. Maag, A. R., Tompsett, G. A., Tam, J., Ang, C. A., Azimi, G., Carl, A. D., Huang, X., Smith, L. J., Grimm, R. L., Bond, J. Q. & Timko, M. T. ZSM-5 decrystallization and dealumination in hot liquid water. *Physical Chemistry Chemical Physics* **21**, 17880–17892 (2019).

54. Wang, Y., Guerra, P., Zaker, A., Maag, A. R., Tompsett, G. A., Smith, L. J., Huang, X., Bond, J. Q. & Timko, M. T. Strategies for extending zeolite stability in supercritical water using thermally stable coatings. *ACS Catalysis* **10**, 6623–6634 (2020).

55. Li, G., Xiao, P., Webley, P. A., Zhang, J. & Singh, R. Competition of CO_2/H_2O in adsorption based CO_2 capture. in *Energy Procedia* vol. 1 1123–1130 (2009).

56. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to metal–organic frameworks. *Chemical Reviews* **112**, 673–674 (2012).

57. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **341**, (2013).

58. Farha, O. K., Eryazici, I., Jeong, N. C., Hauser, B. G., Wilmer, C. E., Sarjeant, A. A., Snurr, R. Q., Nguyen, S. T., Yazaydin, A. Ö. & Hupp, J. T. Metal-organic framework materials with ultrahigh surface areas: Is the sky the limit? *Journal of the American Chemical Society* **134**, 15016–15021 (2012).

59. Nandi, S., Collins, S., Chakraborty, D., Banerjee, D., Thallapally, P. K., Woo, T. K. & Vaidhyanathan, R. Ultralow parasitic energy for postcombustion CO_2 capture realized in a nickel isonicotinate metal–organic framework with excellent moisture stability. *Journal of the American Chemical Society* **139**, 1734–1737 (2017).

60. Chanut, N., Bourrelly, S., Kuchta, B., Serre, C., Chang, J. S., Wright, P. A. & Llewellyn, P. L. Screening the effect of water vapour on gas adsorption performance: Application to CO_2 capture from flue gas in metal–organic frameworks. *ChemSusChem* **10**, 1543–1553 (2017).

61. Nugent, P., Giannopoulou, E. G., Burd, S. D., Elemento, O., Giannopoulou, E. G., Forrest, K., Pham, T., Ma, S., Space, B., Wojtas, L., Eddaoudi, M. & Zaworotko, M. J. Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

62. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

63. Wilmer, C. E., Leaf, M., Lee, C. Y., Farha, O. K., Hauser, B. G., Hupp, J. T. & Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry* **4**, 83–89 (2012).

64. Lin, J.-B., Nguyen, T. T. T., Vaidhyanathan, R., Burner, J., Taylor, J. M., Durekova, H., Akhtar, F., Mah, R. K., Ghaffari-Nik, M., Marx, S., Fylstra, N., Iremonger, S. S., Dawson, K. W., Sarkar, P., Hovington, P., Rajendran, A., Woo, T. K. & Shimizu, G. H. K. A scalable metal-organic framework as a durable physisorbant for Carbon Dioxide Capture. *Science (Accepted)* (2022).

65. Rowsell, J. L. C. & Yaghi, O. M. Metal-organic frameworks: A new class of porous materials. *Microporous and Mesoporous Materials* vol. 73 3–14 (2004).

66. Mason, J. A., Sumida, K., Herm, Z. R., Krishna, R. & Long, Jeffrey. R. Evaluating metal–organic frameworks for post-combustion carbon dioxide capture via temperature swing adsorption. *Energy & Environmental Science* **4**, 3030 (2011).

67. Adil, K., Bhatt, P. M., Belmabkhout, Y., Abtab, S. M. T., Jiang, H., Assen, A. H., Mallick, A., Cadiau, A., Aqil, J. & Eddaoudi, M. Valuing Metal–Organic Frameworks for Postcombustion Carbon capture: A benchmark study for evaluating physical adsorbents. *Advanced Materials* **29**, 1–10 (2017).

68. Düren, T., Bae, Y.-S. & Snurr, R. Q. Using molecular simulation to characterise metal-organic frameworks for adsorption applications. *Chemical Society reviews* **38**, 1237–1247 (2009).

69. Getman, R. B., Bae, Y., Wilmer, C. E. & Snurr, R. Q. Review and analysis of molecular simulations of methane, hydrogen, and acetylene storage in metal-organic frameworks. *Chemical Reviews* **112**, 703–23 (2012).

70. Smit, B. & Maesen, T. L. M. Molecular simulations of zeolites: Adsorption, diffusion, and shape selectivity. *Chemical Reviews* **108**, 4125–4184 (2008).

71. Haghpanah, R., Nilam, R., Rajendran, A., Farooq, S. & Karimi, I. A. Cycle synthesis and optimization of a VSA process for postcombustion CO_2 capture. *AIChE Journal* **59**, 4735–4748 (2013).

72. Krishnamurthy, S., Haghpanah, R., Rajendran, A. & Farooq, S. Simulation and optimization of a dual-Adsorbent, two-bed vacuum swing adsorption process for CO_2 capture from wet flue gas. *Industrial and Engineering Chemistry Research* **53**, 14462–14473 (2014).

73. Rajagopalan, A. K., Avila, A. M. & Rajendran, A. Do adsorbent screening metrics predict process performance? A process optimisation based study for post-combustion capture of CO_2. *International Journal of Greenhouse Gas Control* **46**, 76–85 (2016).

74. Maruyama, R. T., Pai, K. N., Subraveti, S. G. & Rajendran, A. Improving the performance of vacuum swing adsorption based CO_2 capture under reduced recovery requirements. *International Journal of Greenhouse Gas Control* **93**, 102902 (2020).

75. Li, J. R., Ma, Y., McCarthy, M. C., Sculley, J., Yu, J., Jeong, H. K., Balbuena, P. B. & Zhou, H. C. Carbon dioxide capture-related gas adsorption and separation in metal-organic frameworks. *Coordination Chemistry Reviews* vol. 255 1791–1823 (2011).

76. Li, J.-R., Kuppler, R. J. & Zhou, H.-C. Selective gas adsorption and separation in metal-organic frameworks. *Chemical Society reviews* **38**, 1477–504 (2009).

77. Khurana, M. & Farooq, S. Integrated adsorbent-process optimization for carbon capture and concentration using vacuum swing adsorption cycles. **63**, 2987–2995 (2017).

78. Krishnamurthy, S., Haghpanah, R., Rajendran, A. & Farooq, S. Simulation and optimization of a dual-adsorbent, two-bed vacuum swing adsorption process for CO_2 capture from wet flue gas. *Industrial and Engineering Chemistry Research* **53**, 14462–14473 (2014).

79. Mission Innovation. Accelerating Breakthrough Innovation in Carbon Capture, Utilization, and Storage. 1–291 (2017).

80. Farmahini, A. H., Krishnamurthy, S., Friedrich, D., Brandani, S. & Sarkisov, L. From Crystal to adsorption column : Challenges in multiscale computational screening of materials for adsorption separation processes. *Industrial & Engineering Chemistry Research* **57**, 15491–15511 (2018).

81. Hasan, M. M. F., First, E. L. & Floudas, C. A. Cost-effective CO_2 capture based on in silico screening of zeolites and process optimization. *Physical Chemistry Chemical Physics* **15**, 17601 (2013).

82. Zhao, Z., Li, Z. & Lin, Y. S. Adsorption and diffusion of carbon dioxide on metal-organic framework (MOF-5). *Industrial and Engineering Chemistry Research* **48**, 10015–10020 (2009).

83. Kramer, M., Schwarz, U. & Kaskel, S. Synthesis and properties of the metal-organic framework Mo 3(BTC)2 (TUDMOF-1). *Journal of Materials Chemistry* **16**, 2245–2248 (2006).

84. Bastin, L., Bárcia, P. S., Hurtado, E. J., Silva, J. A. C., Rodrigues, A. E. & Chen, B. A microporous metal-organic framework for separation of CO_2/N_2 and CO_2/CH_4 by fixed-bed adsorption. *Journal of Physical Chemistry C* **112**, 1575–1581 (2008).

85. Bloch, W. M., Babarao, R., Hill, M. R., Doonan, C. J. & Sumby, C. J. Post-synthetic structural processing in a metal-organic framework material as a mechanism for exceptional CO_2/N_2 selectivity. *Journal of the American Chemical Society* **135**, 10441–10448 (2013).

86. Talu, O. *Needs, status, techniques and problems with binary gas adsorption experiments*. (1998).

87. Krishnamurthy, S., Lind, A., Bouzga, A., Pierchala, J. & Blom, R. Post combustion carbon capture with supported amine sorbents: From adsorbent characterization to process simulation and optimization. *Chemical Engineering Journal* **406**, (2021).

88. Mason, J. A., McDonald, T. M., Bae, T. H., Bachman, J. E., Sumida, K., Dutton, J. J., Kaye, S. S. & Long, J. R. Application of a high-throughput analyzer in evaluating solid adsorbents for post-combustion carbon capture via multicomponent adsorption of CO_2, N_2, and H_2O. *Journal of the American Chemical Society* **137**, 4787–4803 (2015).

89. Polat, H. M., Kavak, S., Kulak, H., Uzun, A. & Keskin, S. CO_2 separation from flue gas mixture using $[BMIM][BF_4]$/MOF composites: Linking high-throughput computational screening with experiments. *Chemical Engineering Journal* **394**, (2020).

90. Jung, M., Park, J., Lee, K., Attia, N. F. & Oh, H. Effective synthesis route of renewable nanoporous carbon adsorbent for high energy gas storage and CO_2/N_2 selectivity. *Renewable Energy* **161**, 30–42 (2020).

91. Vaidhyanathan, R., Iremonger, S. S., Shimizu, G. K. H., Boyd, P. G., Alavi, S. & Woo, T. K. Direct observation and quantification of CO_2 binding within an amine-functionalized nanoporous solid. *Science (New York, N.Y.)* **330**, 650–3 (2010).

92. Zhang, J. & Chen, X. Optimized acetylene / carbon dioxide sorption in a dynamic porous crystal. *Journal of the American Chemical Society Articles* 5516–5521 (2009).

93. Huck, J. M., Lin, L.-C., Berger, A. H., Shahrak, M. N., Martin, R. L., Bhown, A. S., Haranczyk, M., Reuter, K. & Smit, B. Evaluating different classes of porous materials for carbon capture. *Energy Environ. Sci.* **7**, 4132–4146 (2014).

94. Simmons, J. M., Wu, H., Zhou, W. & Yildirim, T. Carbon capture in metal-organic frameworks - A comparative study. *Energy and Environmental Science* **4**, 2177–2185 (2011).

95. Sun, L. B., Li, A. G., Liu, X. D., Liu, X. Q., Feng, D., Lu, W., Yuan, D. & Zhou, H. C. Facile fabrication of cost-effective porous polymer networks for highly selective CO_2 capture. *Journal of Materials Chemistry A* **3**, 3252–3256 (2015).

96. Wang, Z., Goyal, N., Liu, L., Tsang, D. C. W., Shang, J., Liu, W. & Li, G. N-doped porous carbon derived from polypyrrole for CO_2 capture from humid flue gases. *Chemical Engineering Journal* **396**, (2020).

97. Krishna, R. Screening metal–organic frameworks for mixture separations in fixed-bed adsorbers using a combined selectivity/capacity metric. *RSC Advances* **7**, 35724–35737 (2017).

98. Sikora, B. J., Wilmer, C. E., Greenfield, M. L. & Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal--organic frameworks. *Chemical Science* **3**, 2217–2223 (2012).

99. Bloch, W. M., Babarao, R., Hill, M. R., Doonan, C. J. & Sumby, C. J. Post-synthetic structural processing in a metal-organic framework material as a mechanism for exceptional CO_2/N_2 selectivity. *Journal of the American Chemical Society* **135**, 10441–10448 (2013).

100. Chen, J., Loo, L. S. & Wang, K. An ideal absorbed solution theory (IAST) study of adsorption equilibria of binary mixtures of methane and ethane on a templated carbon. *Journal of Chemical and Engineering Data* **56**, 1209–1212 (2011).

101. Carbon-Capture-Technology-Compendium-2020.

102. Freeman, S. A., Dugas, R., van Wagener, D., Nguyen, T. & Rochelle, G. T. Carbon dioxide capture with concentrated, aqueous piperazine. *Energy Procedia* **1**, 1489–1496 (2009).

103. Chung, Y. G., Camp, J., Haranczyk, M., Sikora, B. J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O. K., Sholl, D. S. & Snurr, R. Q. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **26**, 6185–6192 (2014).

2. Chapter 2: Simulation and Modelling Methods

In this chapter, all methods used in this thesis to predict the performance of materials for carbon capture and storage applications are described, along with all techniques employed to optimize conditions and analyze resulting data from the studies discussed in this work.

2.1. Introduction and Overview

The work presented in this thesis involved the study and analysis of metal-organic framework materials (MOFs) with a focus on their applications in post-combustion carbon capture and storage (PoC-CCS).

The materials were studied at the atomistic level through explicit consideration of the atomic structure of these materials and their relationship to gas molecules. The materials were also studied at the process, or engineering, level by simulating a material's performance in industrial pressure swing adsorption (PSA) systems. Finally, extensive datamining of large-scale screenings of MOFs was performed using sophisticated machine learning techniques.

2.2. Atomistic Simulations

The use of simulation techniques to study molecular and periodic systems at the atomistic level has developed and grown as computers become increasingly powerful. The applications span a wide range: from the development of catalysts[1–3] to investigations into battery technologies,[4–6] from protein folding[7–9] to studies of gas adsorption and separation properties.[10–16] Many different simulation techniques exist to model systems at the atomic level, ranging from density functional theory, a class of methods that explicitly describe all electrons in a system with quantum mechanics, to empirical force field methods that use classical mechanics and fitted parameters to model the atomic behaviour. Although density functional theory methods may present a more complete picture of a system, they are time-consuming and computationally expensive. As such, when dealing with simulations requiring the sampling of hundreds of thousands of configurations, for example in Grand Canonical Monte Carlo (GCMC) simulations used to calculate gas adsorption properties,[12,17,18] the use of empirical methods is far more practical.

To perform studies of the interactions between gas molecules and MOFs at the empirical level, a set of parameters to describe dispersion and electrostatic interactions are required to model the

system. Parameters that model the dispersion interactions using the Lennard-Jones potential (see section 2.2.2.4) are readily available and easily transferrable between systems due to their relatively weak interactions with the surrounding atoms. The most common parameters used to model the electrostatic interactions are partial atomic charges placed directly on atoms within the simulation. These charges are significantly more challenging to obtain since the partial atomic charges in a system are heavily dependant on the surrounding atoms.[19] As a result, the electron density which dictates the partial atomic charge assigned to a specific atom cannot simply be generalized to an atom type as is done with the dispersion interactions. Instead, consideration of the entire structure is required to develop appropriate partial atomic charges.

2.2.1. Atomic Charges: The REPEAT Method

The method used in this thesis to calculation partial atomic charges in MOFs is known as the Repeating Electrostatic Potential Extracted Atomic (REPEAT) method.[20] This technique was developed by Woo and coworkers as a means of fitting partial atomic charges to a quantum mechanical electrostatic potential (QM-ESP). As a result, although the use of these charges is considered empirical, a full single-point DFT calculation is required for each material to generate the QM-ESP. In this work, the QM-ESP was generated using VASP (Vienna Ab-Initio Simulation Package),[21] a plane-wave based periodic DFT code, and the REPEAT charges were fit to reproduce the resulting QM-ESP using a least-squares fitting. Although several other techniques exist to calculate partial atomic charges based on the QM-ESP, REPEAT was the first method to do this for periodic systems and in comparative studies, the REPEAT method was found to be the most robust and accurate method for calculating partial atomic charges in periodic systems.[22,23] Once charges are generated for all atoms in a material, gas adsorption properties can be calculated using GCMC. The partial atomic charges on the framework atoms are assumed to be constant despite the presence of the guests in the simulation.

2.2.2. Grand Canonical Monte Carlo Simulations

A Monte Carlo is a simulation technique that uses random numbers to determine the probabilities of different outcomes of a system. One of the simplest examples of a Monte Carlo is an algorithm that simulates the rolling of two 6-sided dice. By “rolling” the two dice thousands of times and recording the resulting sum of those dice at each iteration, the probability of any outcome from rolling two dice can easily be calculated.

This concept is expanded with the Metropolis Monte Carlo, a scheme that targets specific states of a system, and favours points in the simulation which fall closer to that desired state. This technique is used in the Grand Canonical Monte Carlo (GCMC) Simulation method[12,17] discussed in this thesis, in which gas molecules are randomly placed, deleted, and perturbed within the pores of the material. This method uses the principles of statistical mechanics to estimate macroscopic gas adsorption properties such as the gas uptake and isosteric heat of adsorption at a given set of temperature and pressure conditions. The GCMC method samples the Grand Canonical ensemble, meaning that the simulation assumes that the system has a constant chemical potential, volume, and temperature. By randomly perturbing the gas molecules within the pores of a material, and sampling microstates that have favourable potential energies, the simulation can give estimates of the most probable number of gas molecules within those pores.

2.2.2.1. GCMC Simulation Algorithm

The system modelled by GCMC assumes that the gas phase and the adsorbed phase are in a state of equilibrium. This leads to the most important relationship leveraged by the GCMC method shown in equation 2.1. It states that the chemical potential of the gas phase is equal to the chemical potential of the adsorbed phase. As such, the first step in the GCMC algorithm is to calculate the gas-phase chemical potential, μ_{gas}, at the fixed temperature and pressure point being modelled by the simulation. The value of μ_{gas} is calculated using equation 2.2,[24] where μ^0_{gas} is the chemical potential of the gas in a standard state, T is the temperature, p is the pressure of the gas, R is the gas constant, and f is the fugacity of the gas estimated using an equation of state such as the Peng-Robinson.[25]

$$\mu_{gas} = \mu_{adsorbed} \tag{2.1}$$

$$\mu_{gas}(T,p) = \mu^0_{gas} + RTln\left(\frac{f}{f^0}\right) \tag{2.2}$$

Once the chemical potential is calculated, the GCMC simulation inserts a gas molecule in a random position and orientation in the pores of the material and calculates the potential of the system. The system is then perturbed by either randomly repositioning an existing molecule, inserting an additional molecule, or deleting an existing molecule. The change in potential energy of this perturbation is then calculated using equation 2.3 where U_{i} is the potential energy of the current state, and U_{i-1} is the potential energy of the previously accepted state. If $\Delta U \leq 0$ then the new configuration is considered favourable and is automatically accepted.

$$\Delta U = U_i - U_{i-1} \tag{2.3}$$

In the case where $\Delta U > 0$, meaning the new configuration, i, is less energetically favourable than the previous configuration, $i-1$, an acceptance criterion is applied to determine whether configuration i will be accepted. The equation to calculate the acceptance probability varies depending on the nature or the perturbation; different equations are used when a molecule is moved, inserted, or deleted. When the system is perturbed through the movement of an existing gas molecule, the acceptance criteria, A, is calculated using equation 2.4 and takes the form of a Boltzmann factor calculated using ΔU, the Boltzmann constant k, and the temperature of the system, T. In this equation, the ΔU of this new configuration is weighted by the thermal energy of the system kT. The returned value, A, is between 0 and 1, and determines whether the move will be accepted, with smaller values of ΔU having a higher probability of acceptance. A random number between 0 and 1 is then generated by the GCMC algorithm; the move is accepted if this random number is less than or equal to A. Otherwise, the move is rejected, and the system is reverted to the $i-1$ state.

$$A = e^{-\frac{\Delta U}{kT}} \tag{2.4}$$

The acceptance criteria when dealing with insertion and deletion moves are defined by equations 2.5 and 2.6, respectively. Since the ensemble being sampled is the Grand Canonical Ensemble which requires a constant chemical potential, the impact of the addition or removal of a gas molecule to the system needs closer consideration. This is addressed by equations 2.5 and 2.6, which consider the effect of the number of gas molecules in the system, N, the volume of the system, V, and the pressure, p, again weighted by a Boltzmann factor.

$$A = \frac{Vp}{kT(N+1)} e^{-\frac{\Delta U}{kT}} \tag{2.5}$$

$$A = \frac{NkT}{Vp} e^{-\frac{\Delta U}{kT}} \tag{2.6}$$

If a move is accepted, the configuration is saved and is added to the ensemble average, however, if the move is rejected, the system is reverted to configuration $i-1$, which is then added to the ensemble average. Take for example a system where N=5 in state $i-1$ that undergoes the insertion of an additional molecule to create the configuration i where N=6. If this step is rejected and the system is reverted to configuration $i-1$, N=5 would be added to the ensemble average. In this example, step i and $i-1$ both added N=5 to the ensemble average. The entire process of randomly perturbing and evaluating the system is then repeated until a desired number of steps or configurations have been sampled or a convergence criterion has been met. Once the simulation is complete, the ensemble

averages from the simulation are calculated which represent the most probable adsorption properties of the material at the given conditions. The GCMC code used throughout this thesis was written in-house by a previous member of the Woo Lab, Peter Boyd,[26] based on the DL_POLY molecular dynamics code.[27]

2.2.2.2. Calculating the Potential

As discussed in section 2.2.2.1, the most important consideration in simulating the gas adsorption properties of a material is the calculation of the potential energy surface (PES) or the potential energy of the system at any given configuration. The PES calculated in our GCMC simulations take the form shown in equation 2.7, which is the sum of the dispersion interactions modelled by a Lennard-Jones potential, E_{LJ}, and the electrostatic interactions modelled using a Coulomb potential, E_{coul}. The interaction energies E_{LJ} and E_{coul} are calculated for every pair of atoms in the system, however, since the framework atoms are fixed throughout and will therefore not impact the value of ΔU, only the interactions between the guests and the framework and the guests and other guest molecules are considered.

$$U_i = \sum_{i,j}^{\substack{atom\\pairs}} E_{ij}^{LJ}(r_{ij}) \quad + \sum_{i,j}^{\substack{atom\\pairs}} E_{ij}^{coul}(r_{ij}) \tag{2.7}$$

The need to consider all pairs of atoms in the simulation when calculating the potential becomes challenging when dealing with crystalline structures. When simulating crystalline materials such as metal-organic frameworks (MOFs), an important consideration is the periodic boundary conditions. Since MOF structures are modelled as periodic systems, it is assumed that the number of mirror images of the unit cell in each simulation is infinite in all three dimensions. As such, these mirror images need to be considered when calculating the potentials of the system, however, calculating a potential based on an infinite number of atom pairs is not possible. The Lennard-Jones and Coulomb potentials therefore need to employ different techniques to deal with this challenge.

2.2.2.3. Periodic Boundary Conditions

Metal-organic frameworks are infinitely periodic crystalline materials and require a specialized technique known as periodic boundary conditions (PBCs) to be accurately modelled. This technique reduces the structure of the material to the smallest possible repeating unit, known as the unit-cell.

When used in simulations, PBCs allow for the creation of "image-cells", duplicates of the unit-cell, along all three axes surrounding the unit cell. This allows the simulation code to consider interactions at any distance and in any direction, an important consideration when simulating interactions between atoms in periodic systems.

When simulating the short-range interactions between a guest molecule and the framework atoms, a process vital to GCMC Simulations, it is important to avoid self-interaction between the guest molecule in the unit-cell to the same guest in the image-cells. When dealing with exceedingly small unit cells, the guest molecule in an image-cell could compete with the binding of the same guest in the unit-cell due to repulsive steric or electrostatic forces. To remove this competition, the unit cell is extended, adding image cells in all three dimensions to create a "super-cell" such that the length of the cell along each axis exceeds the short-range interaction cut-off distance. The short-range cut-off used in this thesis was 12 Å. Extending the cell parameters to meet or exceed this cut-off value allows the simulation to rely on the *minimum image convention* when calculating the short-range interactions between atoms. The *minimum image convention* reduces the computational cost of the simulation by only calculating the interactions between the atom in the unit cell, and the closest image of the other atom in the surrounding unit- or image-cells. This technique is employed in this thesis when calculating the Lennard-Jones Potential and the short-range contributions to the electrostatic potential.

2.2.2.4. Lennard-Jones Potential

The dispersion and steric interactions modelled in our GCMC simulations take the form of a Lennard-Jones potential defined by equation 2.8[28,29] and whose behaviour is demonstrated in Figure 2.1. In this equation, ϵ_{ij} represents the well depth of the Lennard-Jones potential between atoms i, and j, and is interpreted as the minimum potential energy value, and σ_{ij} is known as the distance scaler between atoms i and j and represents the distance at which the attractive term and the repulsive terms are equal in magnitude making $E_{LJ} = 0$. Both ϵ_{ij} and σ_{ij} are parameters that have been fit to reproduce experimental observations. The final term r_{ij} is the interatomic distance between the atom pairs. This potential is known as a 12-6 potential, in which the ${r_{ij}}^{-12}$ term represents the repulsive contribution and the ${r_{ij}}^{-6}$, the attractive contribution. Since repulsion and attraction scale by $1/{r_{ij}}^{12}$ and $1/{r_{ij}}^{6}$, respectively, the terms decay rapidly as a function of the interatomic distance and converge to 0 at large distances. This rapid decay allows for the use of a cut-off distance in simulations (typically 12 Å) and

therefore justifies the use of the minimum image convention. The Lennard-Jones potential is therefore easily dealt with when using periodic boundary conditions to approximate an infinitely periodic system.

$$E_{LJ} = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.8}$$

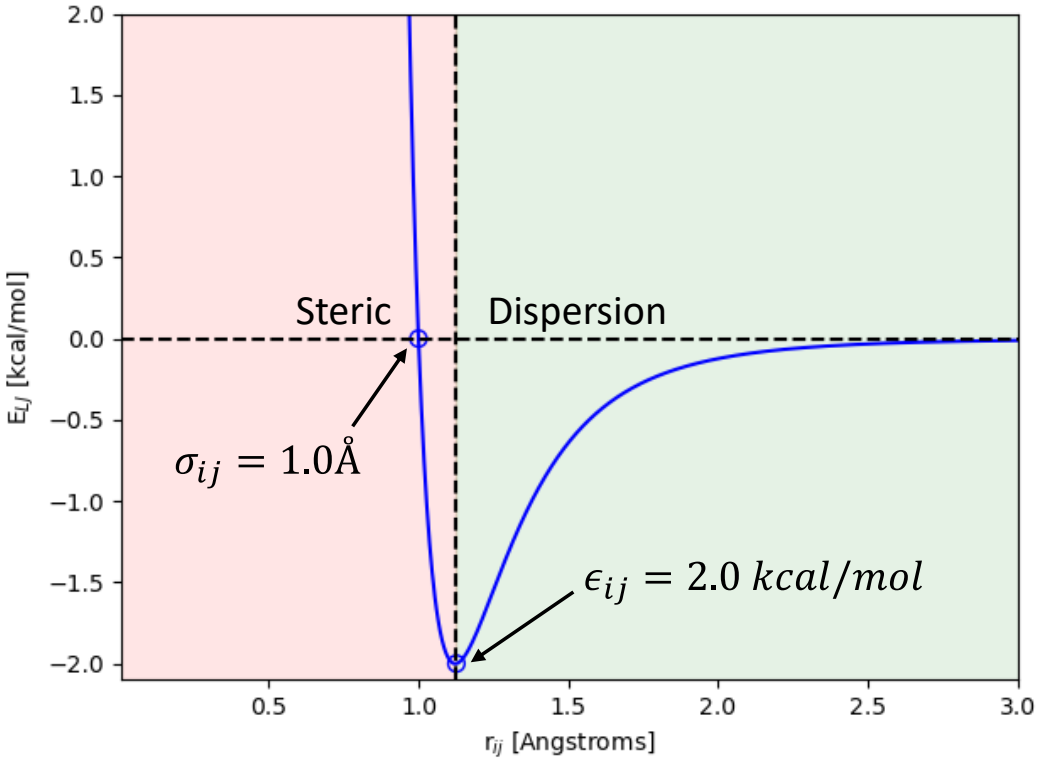


Figure 2.1 Plot showing the Lennard-Jones potential for a function with a $\boldsymbol{\sigma_{ij}}$ of 1.0 Å and an $\boldsymbol{\epsilon_{ij}}$ of 2.0 kcal/mol. The region representing the steric repulsion is shown in red, and the attractive region resulting from dispersion interactions is shown in green.

The parameters used by a Lennard-Jones are fit to reproduce some experimental observable values, such as the gas uptakes of a material,[30] boiling points,[31] or high-level quantum computations.[32] These individual potentials can be assembled to form a suite known as a forcefield. There are many forcefields that exist for public use, but the two main forcefields used throughout this thesis to model the framework atoms are the Universal Force Field (UFF)[33] and the DREIDING forcefield.[34] Although these forcefields contain parameters for every atom on the periodic table, they do not contain the parameters for any hetero-atom pairs. This means that the values of $\sigma_{ij}$and ε_{ij} need to be calculated based on single atom parameters, using equations 2.9 and 2.10,[35] respectively.

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \tag{2.9}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \tag{2.10}$$

2.2.2.5. Electrostatic Potential and the Ewald Summation

The second term in equation 2.7 is the energy associated with the electrostatic interactions between the atoms in the simulation. Normally, a Coulomb potential could be used to calculate the E_{coul}, relying primarily on the partial atomic charges described in section 2.2.1. The most well-known formula to calculate the Coulomb potential is given in equation 2.1, in which q_i refers to the partial atomic charges on atom i, N is the number of atoms in the unit cell, n is the number of mirror images being considered in the calculation, ε_0 is the permittivity of a vacuum, and L is the length of the unit cell.

$$E_{coul}(i,j) = \frac{1}{4\pi\varepsilon_0} \sum_{n} \sum_{i>j}^{N} \sum_{j}^{N} \frac{q_i q_j}{|r_{ij} + nL|} \tag{2.11}$$

Although this equation considers a set of mirror images in the coulomb potential, we are again faced with the challenge of determining the number of unit cells that need to be considered to accurately model the electrostatic interactions without the need to rely on an infinite number of atom pairs. Compared with the Lennard-Jones potential discussed in section 2.2.2.4, the coulomb potential decays at a much slower rate of $1/r_{ij}$, and therefore using a similar cut-off would likely omit important interactions. On the other hand, increasing this cut-off value would lead to an exponentially greater number of atom pairs and result in a much more expensive computation. To mitigate this, a more efficient method at estimating the coulomb potential in periodic systems is used, the Ewald summation,[36] defined in equation 2.12.

$$Ewald(i,j) = \frac{q_i q_j}{4\pi\varepsilon_0} \left(\frac{\mathrm{erf}\,(a r_{ij})}{r_{ij}} + \frac{erfc(a r_{ij})}{r_{ij}} \right) \tag{2.12}$$

$$\mathrm{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r e^{-t^2} dt \tag{2.13}$$

$$erfc(r) = \frac{2}{\sqrt{\pi}} \int_r^\infty e^{-t^2} dt \tag{2.14}$$

To handle the computations of the coulomb potential with periodic boundary conditions, the Ewald summation partitions the potential into two components modelling the long-range interactions, described by the error function erf (equation 2.13) solved in reciprocal space, and the short-range interactions described by a complementary error function $erfc$ (equation 2.14) solved in real space. These two error functions replace the $1/r_{ij}$ term in equation 2.11 to produce a new function used to calculate the energy from the ESP (equation 2.15). Together, these two error functions sum to be

equivalent to the $1/r_{ij}$ term, demonstrated in Figure 2.2, however separating these terms allows for the use of a cut-off similar to that seen with the Lennard-Jones potential to delineate between the short and long-range interactions, optimized using the α parameter. The short-range interactions can therefore be solved directly using the minimum image convention and the complementary error function $erfc$. The long-range interactions can be computed using the long-range error function, erf, to estimate the long-range contributions to the coulomb potential in all mirror images. The energy of the long-range error function rapidly converges in reciprocal space, making this calculation less computationally expensive when compared to a direct summation.

$$E_{ESP} = \frac{1}{4\pi\varepsilon_0}\sum_{n}\sum_{i>j}^{N}\sum_{j}^{N} q_i q_j \left(\frac{\mathrm{erf}(\alpha|r_{ij}+nL|)}{|r_{ij}+nL|} + \frac{erfc(\alpha|r_{ij}+nL|)}{|r_{ij}+nL|}\right) \tag{2.15}$$

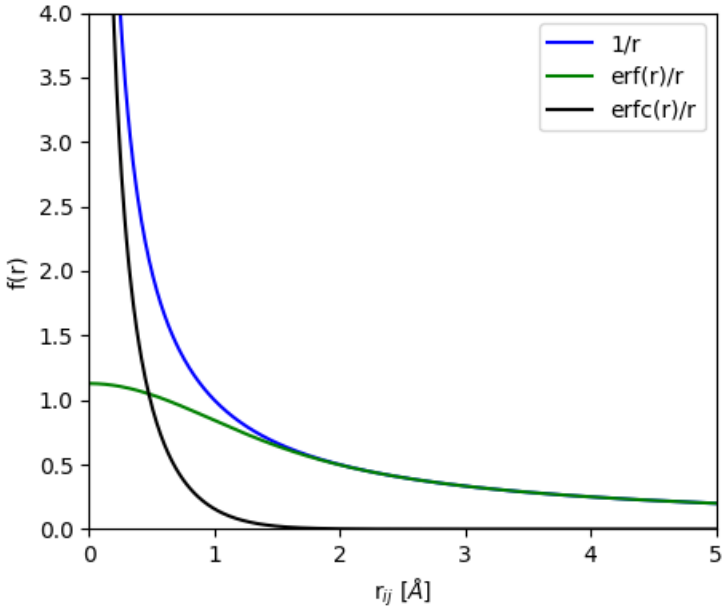


Figure 2.2 Comparison of the decay rates of the $\mathbf{1}/\boldsymbol{r}$, $\mathbf{erf}(\boldsymbol{r})/\boldsymbol{r}$, and $\mathbf{erf}\,\boldsymbol{c}(\boldsymbol{r})/\boldsymbol{r}$ functions used to calculate the coulombic contribution to the potential calculated by GCMC plotted as a function of the interatomic distance $\boldsymbol{r_{ij}}$ between atoms $\boldsymbol{j}$ and $\boldsymbol{j}$.

2.3. Process Simulations

Much of the work discussed in this thesis is centered around a sophisticated pressure swing adsorption (PSA) process simulation algorithm, designed by the research group of Dr. Arvind Rajendran at the University of Alberta, which reproduces the conditions within a gas separation column.[37] This algorithm, specifically developed for pressure/vacuum swing adsorption systems, was able to accurately

reproduce the pressure swing adsorption performance for Zeolite-13X in an 80kg pilot plant[38] and can be easily modified to model a variety of gas separation configurations. This section details the cycle used throughout this thesis to study materials for PoC-CCS systems, and the algorithm used to model that cycle.

2.3.1. 4-Stage Light Product Pressurization Cycle

The work presented in this thesis relied on a 4-stage Light Product Pressurization (LPP) cycle, determined to be the most efficient cycle to separate CO_2 from N_2 at post-combustion carbon capture conditions.[26] This cycle, shown in Figure 2.3, depicts a single column packed with the solid sorbent material being studied throughout the column's operational stages. The operational 4 stages are described as follows: 1) the ***adsorption phase*** in which the flue gas is allowed to flow through the column. The purified N_2 is released to the atmosphere while CO_2 is adsorbed into the pores of the sorbent; 2) the ***blowdown phase***, in which the lower valve is closed, and a vacuum is applied to the top of the column. This stage removes the remaining N_2 trapped in the column; 3) the ***evacuation phase***, where the top valve is closed, and the bottom valve is re-opened. A vacuum is applied to the lower valve to remove the high purity CO_2 from the pores of the sorbent and recover the column; and 4) the ***light-product pressurization*** phase repressurizes the column with purified N_2 reserved from the adsorption phase, pumped into the column from the top valve. This final stage ensures the column is returned to the adsorption phase pressure and pushes the CO_2 front down to the bottom of the column. Each cycle relies on a series of unique variables controlling the operation of the column throughout the separation.

2.3.2. 4-Stage LPP Simulator Input Variables

Although input variables vary depending on the cycle being modelled, the 4-stage LPP cycle requires 7 variables controlling the cycle parameters defined in Table 2.1, and a series of parameters defining the gas adsorption characteristics of the chosen adsorbent material. For the work discussed in this thesis, a dual-site DSL model was used to predict adsorption of CO_2 and N_2, meaning that the PSA simulator required 11 parameters to model the properties of the adsorbent, including 4 isotherm parameters for each gas, heats of adsorption, and crystal density of the adsorbent material. Therefore, each unique PSA simulation required 18 input parameters.

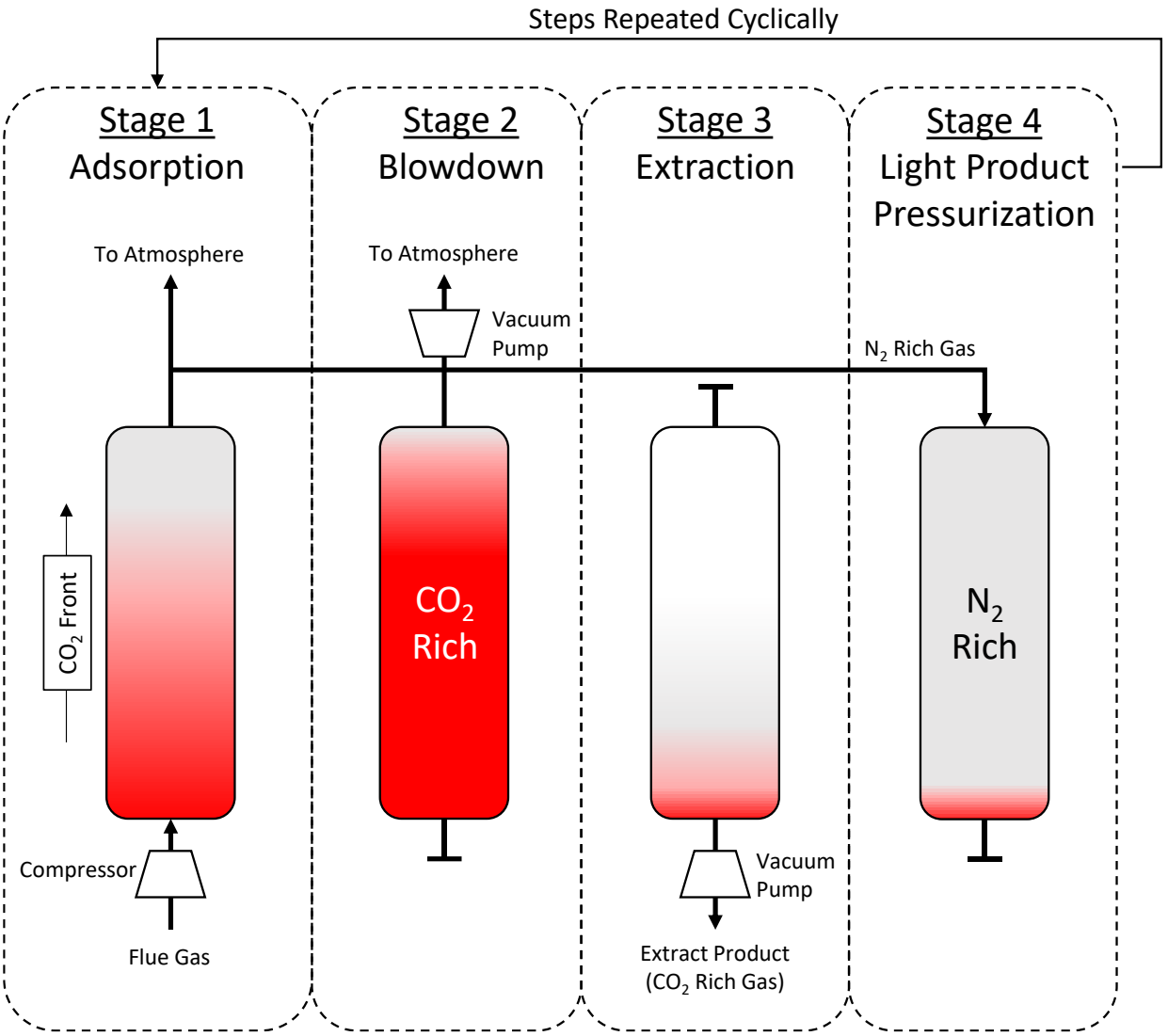


Figure 2.3 Diagram of the 4-stage light product pressurization cycle used in the pressure swing adsorption simulator, depicting a single column over the course of a single cycle. The columns, packed with sorbent material, are pressurized with flue gas, with the CO_2 rich gas depicted in red and the N_2 rich gas depicted in gray.

The 7 input parameters specific to the 4-stage LPP cycle shown in Table 2.1 control the pressure within the column, the time spent on each stage of the cycle, as well as the feed velocity of the flue gas entering through the inlet valve. Importantly, these 7 input variables need to be optimized for each unique material to accurately report a material's ability to separate CO_2 from N_2 in a PoC-CCS PSA system. Prior to performing any PSA simulations, however, a series of key assumptions needed to be made to effectively model the system.

Table 2.1 Descriptions of the 7 process variables used in the PSA simulation specific to the 4-stage LPP cycle.

Process Variable	Description
Adsorption Time (t_{ads}), seconds	Time spent in the adsorption phase
Blowdown Time (t_{blow}), seconds	Time spent in the blowdown phase
Evacuation Time (t_{evac}), seconds	Time spent in the evacuation phase
Blowdown Pressure (P_{blow}), bar	Lowest pressure achieved during blowdown phase
Evacuation Pressure (P_{evac}), bar	Lowest pressure achieved during evacuation phase
Feed Velocity (v_0), m/s	Velocity of flue gas entering column during adsorption phase
Flue Gas Temperature (T_{in}), Kelvin	Temperature of flue gas entering column during adsorption phase

2.3.3. Assumptions

To effectively simulate an industrial pressure/vacuum swing adsorption column, 10 key assumptions needed to be made:

1. The gas-phase obeys the ideal gas law;
2. Particle size used in the column is constant for all sorbent materials;
3. Bulk gas flow can be represented by an axially dispersed plug flow model;
4. Mass transfer in the crystals is assumed to be fast, and the rate-limiting step is the diffusion through the macropores in the structured pelletized spheres;
5. Gas concentrations have no radial gradients meaning the concentrations are constant in the radial direction;
6. Frictional pressure drops in the axial direction along the column can be described by Darcy's law (equation 2.16), where v is the interstitial velocity, μ is the fluid viscosity, ε is the bed voidage, r_p is the particle radius, P is the pressure, and z is the axial coordinate within the column;

$$v = \frac{4}{150\mu}\left(\frac{\varepsilon}{1-\varepsilon}\right)^2 r_p^2\left(-\frac{\partial P}{\partial z}\right) \tag{2.16}$$

7. Thermal equilibrium exists between gas-phase and solid-phase inside the column;
8. The outer surface of the column is held at a constant temperature, and all heat transfer into and out of the column occurs across the column wall;

9. Particle properties and bed voidage are constant across the column;
10. Dynamics of the vacuum pumps, valves, and the piping can be captured using a pressure history relation.

Once these assumptions have been established, the gas can be propagated through the column using the finite volume method.

2.3.4. The Finite Volume Method

The process simulator relies on the finite volume method to propagate the gas through the column. In the finite volume method, demonstrated in Figure 2.4, a column of length L packed with the adsorbent material is partitioned lengthwise into N evenly sized bins. Each bin is assumed to have a uniform gas pressure and composition. The gas mixture is then propagated along the column by solving the component, solid-phase, and overall mass-balance equations defined by equations 2.17, 2.18, and 2.19, respectively, using the ODE23 solver in MATLAB.[39] A full set of definitions for all variables and constants in equations 2.17, 2.18, 2.19, 2.23, and 2.24 can be found in Table 2.2.

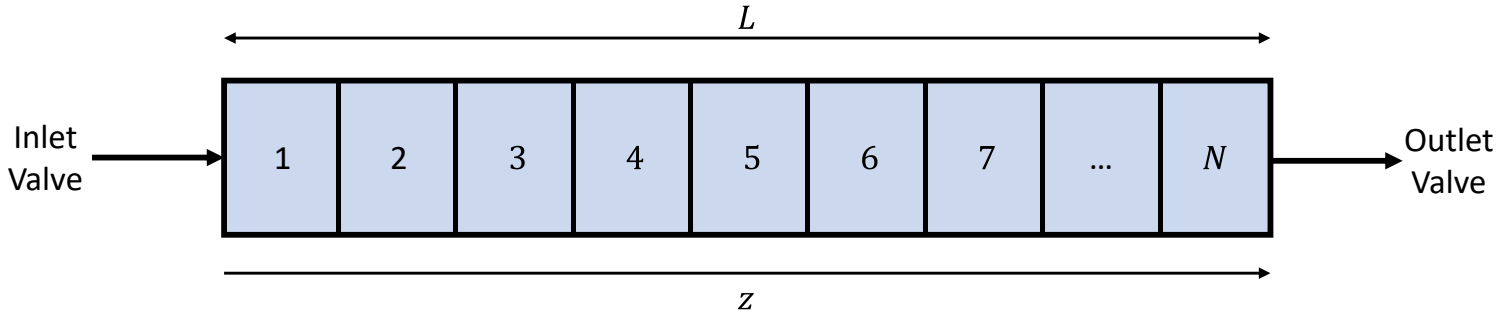


Figure 2.4 Diagram depicting the column of length $\boldsymbol{L}$ packed with a sorbent material divided into $\boldsymbol{N}$ equally sized segments. The segments, which are assumed to have constant gas pressure and composition, and are positioned along the column's axial axis, $\boldsymbol{z}$.

$$\frac{\partial y_i}{\partial t} + \frac{y_i}{P}\frac{\partial P}{\partial t} - \frac{y_i}{T}\frac{\partial T}{\partial t} = \frac{T}{P} D_L \frac{\partial}{\partial z}\left(\frac{P}{T}\frac{\partial y_i}{\partial z}\right) - \frac{T}{P}\frac{\partial}{\partial z}\left(\frac{y_i P}{T} v\right) - \frac{RT}{P}\frac{1-\varepsilon}{\varepsilon}\frac{\partial q_i}{\partial t} \tag{2.17}$$

$$\frac{\partial q_i}{\partial t} = k_i(q_i{}^* - q_i) \tag{2.18}$$

$$\frac{1}{P}\frac{\partial P}{\partial t} - \frac{1}{T}\frac{\partial T}{\partial t} = -\frac{T}{P}\frac{\partial}{\partial z}\left(\frac{P}{T} v\right) - \frac{RT}{P}\frac{1-\varepsilon}{\varepsilon}\sum_{i=1}^{ncomp}\frac{\partial q_i}{\partial t} \tag{2.19}$$

These mass-balance equations all rely on estimates of the amount of each component gas adsorbed into the pores of the material, q_i, using the mole fraction for that component gas, y_i, in each finite volume bin. The amount of each gas adsorbed in each segment, q_i, is calculated using a competitive dual-site Langmuir model, obtained from the solutions to the Ideal Adsorbed Solution Theory (IAST).[40,41] This competitive model is defined in equation 2.20 and demonstrated for a CO_2/N_2 binary gas mixture. In this equation, $q_{sat1_{CO_2}}$ and $q_{sat2_{CO_2}}$ are the fitted saturation uptakes for the two sites and the terms b_{CO_2} and d_{CO_2} are the fitted dual-site Langmuir parameters for CO_2 at a given temperature. Since temperature is not constant throughout the column, this temperature-dependent Langmuir constant is calculated using equation 2.21 which transforms the temperature-independent Langmuir constant b_{0,CO_2} using the internal energy, defined in equation 2.22, and the temperature. In these equations, R refers to the gas constant, T the temperature, P_{CO_2} the partial pressure for CO_2, and ΔH_{ads} the isosteric heat of adsorption for the given guest molecule.

$$q_{CO_2} = \frac{q_{sat1_{CO_2}} b_{CO_2} P_{CO_2}}{1 + b_{CO_2} P_{CO_2} + b_{N_2} P_{N_2}} + \frac{q_{sat2_{CO_2}} d_{CO_2} P_{CO_2}}{1 + d_{CO_2} P_{CO_2} + d_{N_2} P_{N_2}} \tag{2.20}$$

$$b_{CO_2} = b_{0,CO_2} e^{\frac{-\Delta U}{RT}} \tag{2.21}$$

$$\Delta U = \Delta H_{ads} + RT \tag{2.22}$$

As mentioned, the temperature throughout the column is not constant and therefore energy balance equations also need to be solved for accurate predictions to be made. Again, this involves solving ordinary differential equations defining the energy-balance within the column itself to model the temperature at various positions, shown by equation 2.23. The simulation also considers the heat transfer through the walls of the column, which is modelled by solving the ordinary differential equation given by equation 2.24.

$$\left[\frac{1-\varepsilon}{\varepsilon}\left(\rho_s C_{p,s} + C_{p,a}\sum_{i=1}^{ncomp} q_i\right)\right]\frac{\partial T}{\partial t} = \frac{K_z}{\varepsilon}\frac{\partial^2 T}{\partial z^2} - \frac{C_{p,g}}{R}\frac{\partial(vP)}{\partial z} - \frac{C_{p,g}}{R}\frac{\partial P}{\partial t} - \frac{1-\varepsilon}{\varepsilon}C_{p,a}T\sum_{i=1}^{ncomp}\frac{\partial q_i}{\partial t} + \frac{1-\varepsilon}{\varepsilon}\sum_{i=1}^{ncomp}(-\Delta H_i)\frac{\partial q_i}{\partial t} - \frac{2h_{in}(T-T_w)}{\varepsilon r_{in}} \tag{2.23}$$

$$\rho_w C_{p,w}\frac{\partial T_w}{\partial t} = K_w\frac{\partial^2 T_w}{\partial z^2} + \frac{2r_{in}h_{in}}{{r_{out}}^2 - {r_{in}}^2}(T - T_w) - \frac{2r_{out}h_{out}}{{r_{out}}^2 - {r_{in}}^2}(T_w - T_a) \tag{2.24}$$

The boundary conditions needed to solve these ordinary differential equations may vary according to the separation cycle being modelled, however, the work performed in this thesis relied on a 4-stage light product pressurization cycle whose boundary conditions are given in Appendix 2.1.

Table 2.2 Definitions of variables found in equations 2.17, 2.18, 2.19, 2.23, and 2.24.

Variable	Definition
y_i	Composition of component *i*, in the gas phase
P	Pressure, Pa
T	Temperature, K
T_w	Column wall temperature, K
T_a	Ambient temperature, K
ρ_s	Density of the adsorbent, kg m^{-3}
ρ_w	Wall density, kg m^{-3}
$C_{p,s}$	Specific heat capacity of the adsorbent, J kg^{-1} K^{-1}
$C_{p,a}$	Specific heat capacity of the adsorbed phase, J kg^{-1} K^{-1}
$C_{p,g}$	Specific heat capacity of the gas phase, J kg^{-1} K^{-1}
$C_{p,w}$	Specific heat capacity of the column wall, J kg^{-1} K^{-1}
D_L	Axial dispersion, m^2 s^{-1}
z	Axial coordinate, m
R	Gas Constant, Pa m^3 mol^{-1} K^{-1}
K_w	Thermal conductivity of column wall, J m^{-2} K^{-1} s^{-1}
K_z	Effective gas thermal conductivity, J m^{-2} K^{-1} s^{-1}
ε	Bed voidage
q_i	Concentration of component *i* in adsorbed phase, mmol/g
q_i^*	Concentration of component *i* in adsorbed phase in next segment along the *z* axis, mmol/g
r_{in}	Column inner radius, m
r_{out}	Column outer radius, m
h_{in}	Inside heat transfer coefficient, J m^{-2} K^{-1} s^{-1}
h_{out}	Outside heat transfer coefficient, J m^{-2} K^{-1} s^{-1}
k_i	Mass transfer coefficient of component *i*, s^{-1}
t	Time, s

2.3.5. Convergence and Outputs

The pressure swing adsorption simulator is allowed to run until it reaches a cyclic steady state or a maximum number of cycles. In this context, the cyclic steady state is said to be achieved when the mass of gas entering the column is equal to the mass of gas exiting the column per unit time, within a pre-defined error threshold. If the simulator reaches the maximum number of cycles defined by the user without reaching a cyclic steady state, the simulation ends and will not be considered successful.

Once completed, the PSA simulator will output four key process performance metrics: the energy required to run the cycle, the purity of captured gas, the recovery of the captured gas, and the overall productivity of the cycle. The energy of the pressure/vacuum swing adsorption cycle, often referred to as the PSA component of the parasitic energy, is dominated by the energy needed to run the vacuum pumps. The parasitic energy is calculated using equation 2.25 which takes the form of the sum of the energy at each step in the cycle, i, in an S-stage separation cycle. The energy of each individual step in the cycle is the work of isentropic compression associated with running the vacuum pumps and is calculated using equation 2.26. In this equation η is the efficiency of the vacuum pump, γ is the adiabatic coefficient, and Q is the volumetric flow rate. The purity, recovery, and productivity are defined in equations 2.27, 2.28, and 2.29, respectively, where n is the number of moles, t_{cycle} is the total time needed to run the cycle, and $V_{sorbent}$ is the volume of sorbent material in the column. In the context of these calculations, the heavy particle refers to the guest molecule being targeted for capture, and the light particle refers to the other guest molecule present in the simulation.

$$E_{VSA} = \sum_{i}^{S} E_i \tag{2.25}$$

$$E_i = \frac{1}{\eta}\frac{\gamma}{(\gamma-1)}\int_{t=0}^{t=t_{step}} QP_{out}\left[\left(\frac{1}{P_{out}}\right)^{\frac{(\gamma-1)}{\gamma}} - 1\right]dt \tag{2.26}$$

$$Purity = 100\%\left[\frac{n_{heavy,captured}}{n_{heavy,captured} + n_{light,captured}}\right] \tag{2.27}$$

$$Recovery = 100\%\left[\frac{n_{heavy,captured}}{n_{heavy,captured} + n_{heavy,emitted}}\right] \tag{2.28}$$

$$Productivity = \frac{n_{heavy,captured}}{t_{cycle}V_{sorbent}} \tag{2.29}$$

2.4. Optimization Methods

Over the course of this thesis, a variety of methods are implemented for optimizing materials for post-combustion carbon capture applications, and for fitting advanced machine learning models to create predictive tools and to find insights through datamining. In this section, the two most important methods, the Adam optimizer and the Genetic Algorithm are described.

2.4.1. Gradient Descent and the Adam Optimizer

The Adaptive Moment Estimation (Adam) optimizer,[42] a variation on the stochastic gradient descent method, is in widespread use in the fitting of artificial neural networks (ANNs) in which thousands of weights need to be optimized to build a final model. In regular gradient descent,[43,44] a random sampling of the variable space is performed and evaluated using a loss function. A common example of a loss function is the mean absolute deviation (MAD), shown in equation 2.30, which compares the predicted ($y_{predicted}$) values to the measured (or actual) values ($y_{measured}$) for A observations.

$$MAD = \frac{1}{A}\sum_{i=1}^{A}\left|y_{predicted} - y_{measured}\right| \tag{2.30}$$

For this example, the value of $y_{predicted}$ takes the form of a simple linear function $g(x)$, shown in equation 2.31 being fit to a series of observed pairs of x and y, with the gradient descent searching for the best slope parameter, m. Once the values for the loss functions are determined for the random set of points along the possible range of values for m, the gradient at the lowest point is calculated using equation 2.32, which in this example takes the form of the partial derivative of the MAD with respect to m. In the case of more complex optimizations with more than one parameter, this is performed for each of the variables and the solution to equation 2.32 and will be dependant on the function being used to predict $y_{predicted}$.

$$g(x) = mx \tag{2.31}$$

$$\frac{\partial}{\partial m}MAD = \frac{\partial}{\partial m}\left[\frac{1}{A}\sum_{i=1}^{A}\left|g(x) - y_{measured}\right|\right] = \frac{\partial}{\partial m}\left[\frac{1}{A}\sum_{i=1}^{A}\left|mx - y_{measured}\right|\right] \tag{2.32}$$

$$\frac{\partial}{\partial m}MAD(m) = \frac{1}{A}\sum_{i=1}^{A}\frac{x(mx - y_{measured})}{A\left|mx - y_{measured}\right|} \tag{2.33}$$

In the regular gradient descent algorithm, the gradient would be calculated by solving equation 2.33 from every value of x for the best slope m. The value of m is then adjusted according to that gradient using equation 2.34 and weighted by a learning rate α at every step j. The learning rate is a value between 0 and 1 which ensures that the algorithm will take incremental steps and not overshoot the best solution. The algorithm ends after it reaches a predefined number of steps, or some convergence criterion is met.

$$m_{j+1} = m_j + \alpha \frac{\partial}{\partial m} MAD(m_j) \tag{2.34}$$

Stochastic gradient descent is similar to regular gradient descent but better suited for datasets containing large numbers of observations, A. In stochastic gradient descent, the gradients are computed using a random subset of observed $y_{measured}$ values at every step to reduce the number of computations required and increase the speed of the fitting.

The Adam optimizer is a variation on the stochastic gradient descent method that relies on an adaptive learning rate (ALR). Using an ALR over a fixed value removes the need of selecting the best learning rate for each individual optimization. This is of particular importance when fitting ANNs when the user has no prior knowledge of the gradients. An additional benefit to using ALRs is the prevention of a phenomenon known as *exploding gradients*. This phenomenon is defined as an instability that occurs during the fitting of neural networks. This instability is the result of error gradients compounding over the course of complex optimizations resulting in the algorithm performing increasingly large overcorrections ultimately leading to a failed optimization. By dynamically adjusting the learning rates, the use of ALRs prevents these overcorrections and improves the efficiency of the optimization.

The ALR uses an exponentially decaying average of past gradients, n_j, and past squared gradients, v_j. Continuing the above example for optimizing the value of m, use of this ALR is demonstrated in equation 2.35. The values of n_j and v_j are defined in equations 2.36 and 2.38, respectively. However, it was found by the authors of the Adam method that during the initial steps of the optimization, the values of n_j and v_j were heavily biased towards zero and caused the optimizer to stagnate.[42] This can be addressed using equations 2.37 and 2.39, which corrects the bias in the n_j and v_j terms. In the series of equations 2.35 to 2.39, the variables β_1, β_2, and ε are parameterized values defining the decay in the learning rate. The authors recommend values of 0.9, 0.999, and 10^{-8} for β_1, β_2, and ε, respectively.[42]

$$m_{j+1} = m_j - \frac{\alpha}{\sqrt{\hat{v}_j} + \varepsilon} \hat{n}_j \tag{2.35}$$

$$n_j = \beta_1 n_{j-1} + (1-\beta_1)\left[\frac{\partial}{\partial m} MAD(m_j)\right] \tag{2.36}$$

$$\hat{n}_j = \frac{n_j}{1-\beta_1^j} \tag{2.37}$$

$$v_j = \beta_2 v_{j-1} + (1-\beta_2)\left[\frac{\partial}{\partial m} MAD(m_j)\right]^2 \tag{2.38}$$

$$\hat{v}_j = \frac{v_j}{1-\beta_2^j} \tag{2.39}$$

2.4.2. Genetic Algorithm

A genetic algorithm (GA) is a powerful optimization tool that excels at finding reasonable approximations of the best solutions of a system by mimicking Darwinian evolution.[45–49] When dealing with complex systems that require optimization of a large set of variables, a GA, as a stochastic universal sampling method, would be of particular use if a systematic or gradient-based approach is too computationally expensive. A GA functions by first generating a random sampling across the range of allowed values for all variables, creating an initial population. This population is then tested and ranked according to an objective function used to estimate the fitness of the individuals in that generation. Once the fitness of these individuals is known, the more fit individuals are allowed to mate, combining a set of variables (often referred as genes) with the possibility of mutation, and creating a new generation. When performed correctly, each subsequent generation will see an improvement to the average fitness until the optimization converges. Due to the stochastic nature of these optimizations, duplicate runs are often performed to ensure the initial optimization did not converge to a local minimum.

2.4.3. Genetic Algorithm: Gene Selection

The first step in performing a GA optimization is the selection of the variables that will form the genes of your population. These variables are simply the parameters that the GA is attempting to optimize, as well as the range allowed for each variable. For example, we can consider a simple case where a baker needs to find the optimal quantities of ingredients to make a cake. In this case, the variables being optimized are the volumes of flour, sugar, and baking soda needed to produce the best possible cake. When using a GA, the volumes of flour, sugar, and baking soda would form the genes of the optimization and the final cake product will be used to evaluate those genes. Importantly, the baker can use their knowledge and experience to set the ranges of allowable volumes for each individual

ingredient. Although the above example deals with continuous variables, genes can be discrete or continuous, or a mixture of both. For the work presented in this thesis, only continuous variables were optimized. Once the genes are properly defined, the initial generation can be created.

2.4.4. Genetic Algorithm: Latin Hypercube Sampling

Once the genes are selected, the next step in the optimization is the creation of the initial generation. This generation needs to cover an even distribution of the N-dimensional space, where N is the number of genes, to ensure the space is appropriately sampled. The technique used in this thesis to generate the initial sampling of the data is Latin Hypercube Sampling (LHS).[50] In LHS, the N-dimensional space is subdivided into M equally sized bins along each dimensional axis, where M is the size of the population. Random selection of the genes is then performed, ensuring that no bin is sampled more than once. An example comparing purely random selection to LHS selection is demonstrated in Figure 2.5a and b, respectively showing the difference in sampling for 5 random points in two dimensions. This technique is commonly used to generate the initial populations in GAs and ensures that the variables selected for each gene are well distributed across the entire allowable range. Once the initial population has been generated, the optimization can proceed to evaluate the individuals in that population using an objective function.

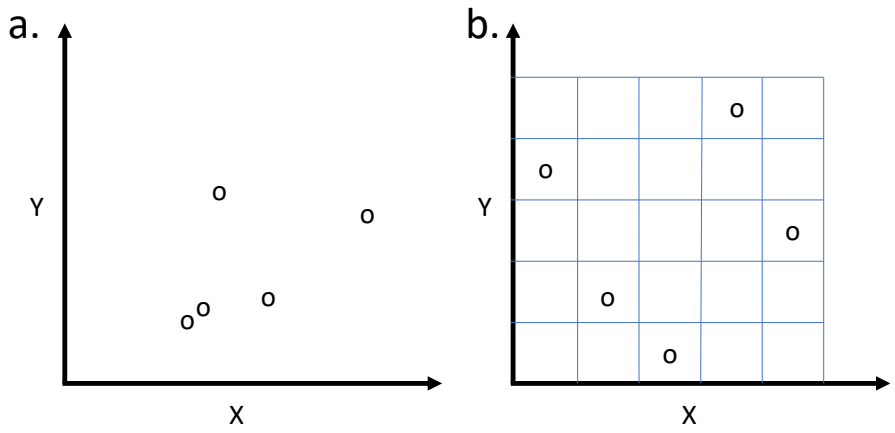


Figure 2.5 a) Example of 5 randomly selected data points and b) 5 datapoints selected using the Latin hypercube sampling technique.

2.4.5. Genetic Algorithm: Objective Functions

Defining the objective function is an essential step in designing a GA and is dependant on the needs of the specific optimization. For single objective optimizations, where the objective is the

property being optimized, the objective function can be as simple as the target value itself. For example, if a GA were designed to optimize the adsorption of CO_2 onto a material, the objective or fitness value could simply be the CO_2 uptake for that material.

When dealing with multi-objective optimizations, such as optimizations of the parasitic energy (PE) and productivity (Pr) of a PSA process, a more complex objective function is required. This can take the form of a relative distance function, shown in equation 2.40, where each term is weighted accordingly using w_1 and w_2 which require tuning to ensure neither target value overpowers the optimization. Additionally, the weights need to compensate for any differences in units between the two variables. Although equation 2.40 was used in this thesis to optimize PE and Pr, a reasonably high target was selected for Pr_{target} based on available data, even though no theoretical maximum value for this property is known. If no prior data was available when designing the fitness function, a different function would have been required to effectively maximize the Pr value. Once the fitness values have been calculated for each individual in a generation, the GA optimization can proceed to the elitism and mating steps.

$$Fitness = \sqrt{\frac{\left(PE_{target} - PE_{individual}\right)^2}{w_1} + \frac{\left(Pr_{target} - Pr_{individual}\right)^2}{w_2}} \tag{2.40}$$

2.4.6. Genetic Algorithm: Elitism and Mating

Once the fitness has been calculated for all individuals in a generation, those individuals are then ranked by that fitness value. In the case of a minimization such as that shown using equation 2.40, the individuals with the smallest fitness value are selected for the elitism step of the optimization. During this step, the top performers are carried forward to the next generation in the optimization with no changes made to the genes. This is controlled by a parameter known as the *elitism rate*, which defines the percentage of the population that will be carried forward via elitism. A common value for this parameter is 10%, meaning the top 10% of a population will be carried forward.

After the individuals are carried forward via elitism, the remainder of the next generation needs to be created through mating. The selection of mating pairs is performed using an algorithm that mimics a roulette wheel[45] where the odds of selection for each individual are proportional to their fitness values. The volume of the slice on the roulette wheel occupied by an individual is defined by equation 2.41 when maximizing the fitness value, or equation 2.42 when minimizing the fitness value. In these equations v_i is the volume individual i will occupy on the roulette wheel and the denominator term

ensures that the entire volume of the wheel will equal 1. This means that the volume term, v_i, also represents the probability that individual i will be selected for mating.

$$v_i = \frac{fitness_i}{\sum_{j=1}^{M} fitness_j} \tag{2.41}$$

$$v_i = \frac{{}^{1}/_{fitness_i}}{\sum_{j=1}^{M} {}^{1}/_{fitness_j}} \tag{2.42}$$

Once the probabilities for each individual in the generation are calculated, the roulette wheel is populated in order of probability demonstrated in Figure 2.6. Random selection is then performed using a random number generator to select a float ranging from 0 to 1. Depending on the scale of the fitness values, a decision needs to be made to determine the number of decimal points allowed during this step, otherwise, some individuals may not have the opportunity to be selected. This is performed twice per mating pair, selecting both parents using the same technique. Individuals are not allowed to mate with themselves as this would be functionally the same as the elitism step and depending on the *size of the population* and the *mutation rate* set by the user, a decision should be made on whether to allow a pair to mate multiple times.

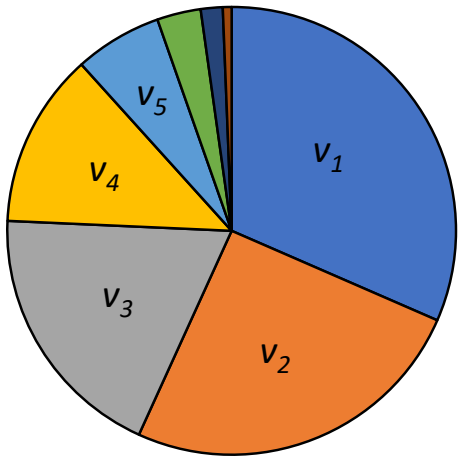


Figure 2.6 Visual demonstration of the roulette wheel used in the GA mating protocol, where v¬i is the volumetric slice occupied by an individual in the generation and represents the probability of selection.

Once a mating pair is selected, the genes of the two individuals are combined using a mating protocol. The protocol used in the work presented in this thesis is designed for optimizing continuous variables and weights the genes according to the relative probabilities of the parents. This is demonstrated for a single gene, z, using equation 2.43 which combines the probabilities calculated

using equation 2.41 or 2.42 and the genes for the parents 1 and 2. Using this method, a mating pair will always produce identical offspring, unless a mutation occurs during the mating step.

$$z_{child} = z_{parent1} + \left(\frac{v_{parent1}}{v_{parent1} + v_{parent2}}\right)\left(z_{parent2} - z_{parent1}\right) \tag{2.43}$$

The GA optimization also simulates a crucial feature found in Darwinian evolution, the random mutation. To ensure the population does not stagnate too early, random mutations are introduced during the mating step of the optimization. The probability and the severity of these mutations are properties controlled by the user, but in the work presented in this thesis a value of 30% was used for both terms. A probability of mutation of 30% means that a child in any mating pair has a 30% chance that a mutation will occur. Once a child has been selected for a mutation, the gene being mutated is selected at random and then perturbed according to equation 2.44. In this equation, the severity of the mutation (*S*) is defined as the fraction of the range of allowable values for that gene, with the direction the mutation chosen at random. It is important to note, that the value of the gene is not permitted to exceed the bounds of the allowable range, and in the case where the mutation falls beyond the limits the value is artificially set to z_{max} or z_{min}, accordingly. Once the new generation is built, the whole process is repeated until the optimization converges.

$$z_{child,mutated} = z_{child,un-mutated} \pm S[z_{max} - z_{min}] \tag{2.44}$$

2.4.7. Genetic Algorithm: Convergence

After a minimum number of generations is performed, the GA algorithm used in this thesis will begin to test for convergence. Once the best individual carried through by elitism remains unchanged for a pre-defined number of steps, the simulation is considered converged, and the optimization ends. The value used throughout this thesis for convergence is 10, meaning that the optimization is considered converged when the best solution remains unchanged after 10 consecutive generations.

The optimizations performed in this thesis which ran for a minimum of 10 generations with 100 individuals per generation created large amounts of data. This data could be mined and analyzed to build predictive models, increase the pace of research, and acquire valuable insights into materials for carbon capture and storage. Due to their ability to rapidly build complex and insightful models based on large datasets, a variety of machine learning techniques were selected to perform this datamining.

2.5. Machine Learning Methods

The massive amounts of data accumulated over the course of the work presented here-in provided an ideal opportunity to make use of a wide range of machine learning techniques. The aim of these machine learning models was two-fold: the first aim was to attempt to develop predictive models which could be used to increase the pace of materials discovery in the field of carbon capture and storage. The second aim was to evaluate a range of parameters and metrics on their ability to predict industrial performance of materials and provide insights into which properties may be targeted to maximize that performance.

2.5.1. Model Types

Although many model types exist in the field of machine learning and artificial intelligence, the two model types used in this work are regressors and classifiers. A regressor is a model designed to reproduce the target variable, whereas a classifier will simply try to bin each datapoint into a pre-defined category. Among these techniques are two crucial paradigms pertinent to this thesis: supervised and unsupervised learning. Although unsupervised learning techniques are only relevant to classification models, understanding both paradigms is important to evaluate the work performed in this thesis.

2.5.2. Supervised Learning

The supervised learning paradigm is one in which the model being fit has prior knowledge of the target or observation it is trying to reproduce; in other words, the model has training data. An example of a supervised learning model would be one that aims to classify the species of butterfly based on some simple characteristics (or features) such as size, colour, and weight. In the case of supervised learning, the model in question would be trained using the characteristics of each butterfly, while trying to reproduce the species labels provided by the user. By providing the labels, the model has prior knowledge of the final values and attempts to reproduce those labels using the three physical characteristics of the butterfly. Once the model is built and can successfully classify the species of butterfly, it can be used in the future to identify the species of new butterflies for which the species are not known. This paradigm is in contrast to unsupervised learning where the labels are not provided to the model.

2.5.3. Unsupervised Learning

The unsupervised learning paradigm is one in which the model being developed has no prior knowledge of the observations or target variables it is trying to reproduce. Using the example discussed above, the model would be provided with size, colour, and weight of each butterfly, but not the species labels. Instead, these models search the feature space to find trends that a researcher can later superimpose over their targets (example: butterfly species) to determine whether any important insights can be extracted. Importantly, running such a model on identical features will yield the same results regardless of any changes in the targets being superimposed.

2.5.4. Feature Scaling

The first challenge in developing machine learning models stems from the large size of the datasets being handled. These datasets need to be preprocessed before any model fittings can take place. This preprocessing often includes the removal of outliers and the scaling of the features, or the descriptors, being used to fit the models.

Scaling is a vital step in preprocessing a dataset for machine learning applications. It involves a transformation to the entire feature space to ensure important features can be effectively identified by the machine learning codes. As an example, two features are being used to fit a predictive model. The first includes values ranging from 1 to 10, while the second ranges from 250 to 700. Although the differences are significant in the absolute values, feature 1 is found to be more predictive of the target property. Without scaling, a machine learning algorithm would likely identify feature 2 due to its large absolute variance over the more important feature 1.

To address this issue and ensure the machine learning algorithm can easily identify key features, all features are rescaled using the Standard Scaler, also known as Z-score normalization.[51] This standard scaler fits equation 2.45 to the feature distribution and returns the transformed value for each individual point in the feature space. In equation 2.45, z is the transformed value, x is the feature value being scaled, u is the mean of the feature distribution, and s is the standard deviation of the feature distribution. The new transformed features will have a mean of 0 where values of 1 and -1 are equal to one standard deviation away from the mean in the original feature distribution. By applying this standard scaler to all features, the relative variance in the feature space is standardized, which in turn makes the identification of key features by the machine learning algorithm faster and more effective.

$$z = \frac{(x - u)}{s} \tag{2.45}$$

2.5.5. Cross-Validation

There are several techniques that can be used to perform validation of fitted machine learnings models. The most basic example is the partitioning of the dataset into two subsets: a training and validation set. Once separated, the training set is used to train the feature scaler and the machine learning models, and the validation set is used to test the accuracy of those models. When dealing with smaller datasets (datasets which typically have fewer than 1000 points) the technique often employed is cross-validation.[52,53] This is commonly referred to N-Fold cross-validation, where the N represents the number of equally sized "chunks" into which the training data is partitioned. One chunk is then removed from the training set, and the model is trained on the remaining N-1 chunks. Once trained, the model is then validated on the removed "chunk". This process is repeated N times, with each "chunk" acting as the validation set once. The N-fold cross-validation performance is then presented as the average of the N model validations. The number of "folds" or "chunks" used ranges from 5 to 10 and is often selected based on the size of the dataset being studied. As the key objective of this validation technique is to ensure the accuracy of the model while preventing overfitting, a larger validation set is often favoured. As a result, smaller training sets should favour fewer folds, while larger sets can increase the number of folds to improve the confidence of the validation results. This trade-off is an important consideration when using N-fold cross-validation.

2.5.6. Linear Discriminant Analysis

One of the simplest forms of machine learning is a technique called Linear Discriminant Analysis (LDA).[54] This is a supervised classification technique that relies on the fitting of a line or planar surface to separate data points according to their known classifications in multi-dimensional space. In this thesis, one and two-dimensional LDAs are used to discriminate classes in datasets.

During linear discriminant analysis, the single or multi-dimensional data is reduced to a single dimension through projection onto a line. The equation of the line is then optimized to best discriminate between the classes present in the dataset. This process is demonstrated in Figure 2.7, where the initial dataset is split into two classifications, red and blue, and can be plotted in two dimensions according to its features A & B (Figure 2.7a). Using LDA, a line is generated and the data points are projected onto

that line, shown in Figure 2.7b, making the dataset one dimensional (Figure 2.7c). Once the dimensions of the data are reduced in this way, each classification can be assigned two values. The first is the scatter defined as the range of values along the 1-dimensional line denoted by the symbol s in Figure 2.7c, and the second is the mean value along the 1-dimensional line denoted by μ.

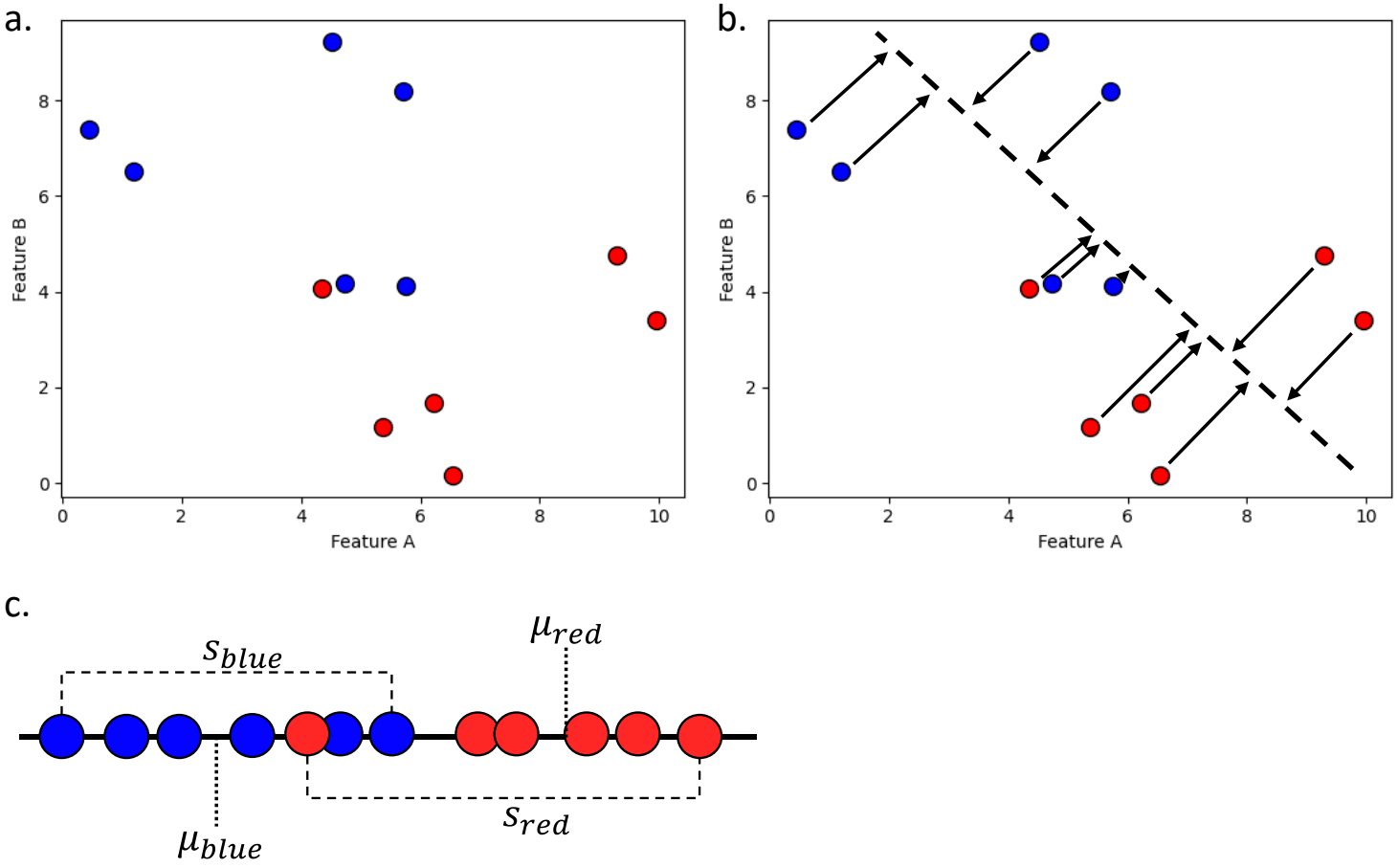


Figure 2.7 a) Generic dataset with two classifications defined by the colour of the ball defined by two features A and B. b) The same generic dataset as in a) with a dashed line representing the line fit through linear discriminant analysis and the arrows demonstrating the projection of the 2-dimensional data onto the line. c) The representation of the dataset shown in a) and b) projected onto the dashed line in figure b), where the variable s is the scatter of the class and μ is the mean of the class along the line's frame of reference

To optimize the LDA model, the distance between the mean of the two classifications is maximized while the scatter is simultaneously minimized for each class. This is done by maximizing the objective function shown defined by equation 2.46. Once a good solution for the line is found, a value along that line is selected to distinguish between the two classifications. Once the model is constructed, it can be used as a predictive model to discriminate new data with no labels.

$$LDA\ Objective = \frac{(\mu_{red} - \mu_{blue})^2}{s_{blue}^2 + s_{red}^2} \quad (2.46)$$

2.5.7. Principal Component Analysis

Principal component analysis (PCA)[43] is an unsupervised machine learning technique that relies on similar concepts to the linear discriminant analysis, where multi-dimensional data is reduced to a lower dimensionality through single variable decomposition. Single variable decomposition is a method that reduces an NxN square matrix to a single variable or factor, where N is the number of features being used to fit that model. This matrix is then decomposed to a single value forming a transformed property called a principal component (PC). Similar to the method used for the LDA, this decomposition is performed using the projections of the data onto a fit line which acts as the axis for the PC. Unlike the LDA which projects the data onto a single line, the PCA will generate projections onto N orthogonal lines where N is the number of features in the dataset. Furthermore, unlike LDAs the PCA has no prior knowledge of the classifications within the data, and thus cannot perform the same procedure to fit those lines.

The first step in the principal component analysis is to centre the data, performing a translation in N dimensions setting the mean of the data along each axis to 0. This is important as each of the lines which form the principal components (PCs) must pass through the origin of the plot. Once centred, a line is then fit to the data, and the data points are projected onto that line.I This fitting aims to maximize the eigenvalue which takes the form of the sum of the squared distance from each projected point on the line to the origin. Once the first PC line is fit, the model moves on to the next second PC line which must pass through the origin and be orthogonal to the first PC line. This fitting follows the same procedure, maximizing the eigenvalue. This step is then repeated until N orthogonal PC lines are successfully fit. By virtue of the fitting procedure, these PCs are ranked according to their importance with PC1 > PC2 > … > PCN.

As an unsupervised learning technique, the PCA method excels at classification problems. Using the example of the butterflies discussed in sections 2.5.3, by fitting the PCA to a variety of characteristics such as size, weight, and colour, the known species of those butterflies can then be overlayed on the resulting PCs to determine whether any clustering occurs. When a PCA generates a model with distinct clustering according to the known species label, that model can then be applied to butterflies whose

species are not known. The value of the new butterflies' PCs can determine where they fall on the PCA plot to predict their species. Although this technique is more qualitative in nature, a more quantitative insight can be obtained using PCAs. Importantly, when distinct clustering occurs among the known labels, the features that most contribute to the fit PCs can be extracted to provide the user with a ranking of feature importance. Again, using the butterflies as an example, when distinguishing between two species, after extracting the feature contributions to the PCs it was found that the colour of the butterflies most contributed to the value of the PCs and is therefore the best indicator of species amongst the provided characteristics. Although the example of butterflies used in this section is quite simplistic, the PCA excels when dealing with large datasets with many features where simply extracting the individual features by hand would be cumbersome or unfeasible. The concept of the PCA can also be expanded to use non-linear kernels when the separation between the known labels does not occur using the linear fittings described in this section.

2.5.8. Kernel-Based Principal Component Analysis

Kernel-based principal component analysis (kPCA)[44] expands on the concept of the PCA described in section 2.5.7. kPCA is a powerful tool when dealing with data that cannot be linearly separable using their original features but can be made separable when using a non-linear kernel. Although many kernels are available, the most commonly used is the radial-basis function (rbf), a modified Gaussian kernel defined by equation 2.47 where x_i and x_j are the values of some feature x for observation i and j, d is the Euclidean distance between x_i and x_j, and l is the length scale defining the width of the kernel. This means that to perform the projection of data onto the rbf kernel, a covariant matrix needs to be generated for each feature. This covariant matrix is an $n \times n$ matrix where n is the number of observations (or individual data points) in the training set.

$$k(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right) \tag{2.47}$$

Once the data is projected into the non-linear kernel space, a regular PCA is performed. Due to the generation of the covariant matrix needed to project the data into higher dimensional space, the number of principal components generated is proportional to the number of observations in the dataset as opposed to the number of features, as was the case with the regular PCA. This means that for n

observations, the kPCA method will generate n^2 principal components due to the need to generate an $n \times n$ kernel matrix.

2.5.9. Decision Trees and Random Forest

2.5.9.1. Decision Trees

Decision trees[55] are a conceptually simple yet powerful supervised machine learning tool that can rapidly and effectively find trends in large datasets. This method can be implemented as either a classifier or a regressor, however, due to the discrete nature of the decision tree's structure, the output of a regressor would not be continuous. As a result, decision trees are more commonly used as a classification tool when not part of a random forest algorithm.

2.5.9.2. Decision Trees: General Structure

The structure of a decision tree follows that of a branching flowchart beginning with a "root" node (Figure 2.8). At the root node, a feature is selected, and the data is partitioned into two "child" nodes based on a chosen value of that feature that best splits the data. This process is then repeated until all nodes contain only one classification or the maximum depth set by the user is reached. The structure of a decision tree with a maximum depth of 2 is demonstrated in Figure 2.8, where an arbitrary dataset is separated into two unique classifications, *Class 1* and *Class 2*, based on the feature space x. At each decision node, two new nodes are created according to the criteria at that decision node. As an example, in Figure 2.8, the *Root Node* splits the data according to whether the feature x_0 is greater than or equal to some value a.

2.5.9.3. Decision Trees: Splitting the Data

Although the concept of a decision tree is easily understood, the process of fitting data to such a model is not as straightforward. For a binary classification model like the one demonstrated in Figure 2.8, the splitting is controlled by the Gini impurity value[56] , calculated using equation 2.48. In equation 2.48, N_1, N_2, and N_{all} represent the number of points in the first created node, the second created node, and the total number of points in the decision node, respectively, and $Gini_1$ and $Gini_2$ are the Gini metric values of the first and second created nodes, respectively.

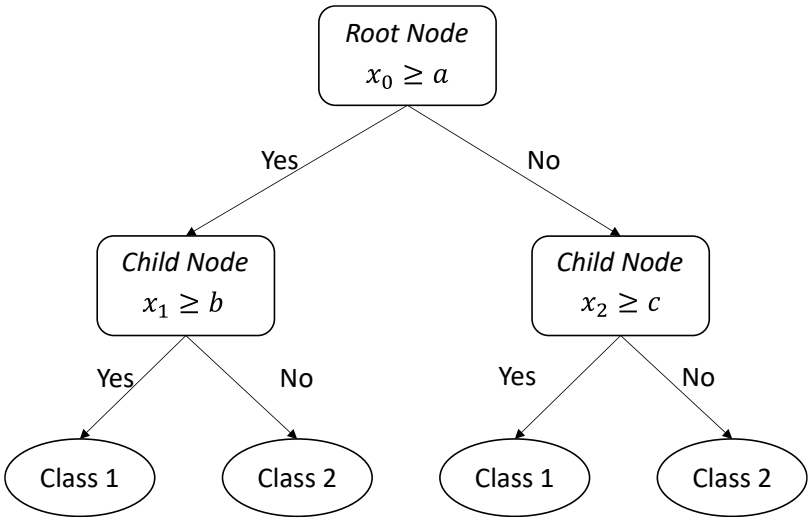


Figure 2.8 General structure of a decision tree classifier with a maximum depth of 2, dividing an arbitrary dataset into two classifications: Class 1 and Class 2. This decision tree is using the feature space, x, where x0¸ x1, and x2 are the features selected at each node, and the variables a, b, and c represent the values used to split the data.

$$Gini\ Impurity = \frac{N_1}{N_{all}} Gini_1 + \frac{N_2}{N_{all}} Gini_2 \tag{2.48}$$

The Gini impurity is based on a Gini metric which represents the purity of the data in a single node, described by equation 2.49, where p_i is the probability of randomly selecting a value with a specific classification i from the entire population of that node (C). For a binary classifier, this Gini value ranges from 0 to 0.5, where 0 means that the node contains only a single classification, and 0.5 means that there is a perfect 50/50 split between the two classes.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2.49}$$

The features in the dataset are tested systematically at each node to identify which feature and at what value best minimizes the Gini impurity by discriminating between the two classifications. This is repeated for every node in the decision tree until the fitting is complete, starting from the *root* node.

2.5.10. Random Forest

A random forest model[57] is an ensemble method that relies on a series of individual decision trees to generate a prediction by taking the mode of the classifications when building a classifier, or the

average of the predictions when building a regressor. Unlike simple decision trees, the use of multiple trees and the implementation of an average allows for a more continuous output when performing regression, and as such, random forest models have proven useful when dealing with regression problems.[58,59]

The key difference in fitting decision trees for individual models vs decision trees that will form part of the random forest model is the selection of features in each individual tree. In the classical decision tree (described in section 2.5.9), the tree searches the entire feature space for the feature which best splits the data by minimizing its Gini impurity. A random forest model introduces an element of stochastic selection to this process by only allowing each individual tree to search a random subset of the features. If, for example, there are M features in the dataset being used to fit a random forest model, each individual tree will only be allowed to use a random subset consisting of $\sqrt{M}$ features. Without implementing this random selection, the individual decision trees in the random forest would be identical, and therefore running the fitting multiple times would not improve the results. By randomly removing features from the decision space, the random forest model allows individual trees to identify other important features which would otherwise be ignored. This strategy has been shown to drastically improve the predictive power of decision trees.[60,61]

2.5.11. Gradient Boosted Decision Trees

Another machine learning technique similar to decision trees and the random forest is the gradient boosted decision tree (GBDT).[62] The GBDT is an ensemble method that relies on a series of decision trees run in sequence, each attempting to improve upon the solution of the previous tree. This is in contrast to the random forest model, which simply generates multiple predictions and estimates the value based on some central tendency of those estimates. The GBDT consists of many individual decision trees, commonly referred to as *estimators*, which attempt to predict the difference between the current predicted value and the actual value for each point in the data.

2.5.11.1. Loss Function

The loss function is a function used by the fitting algorithm to determine the accuracy of the model's parameters and improve the predictions. The gradient boosted decision tree regression model relies on two separate loss functions during the fitting. The first is the loss function guiding the creation

of the nodes in individual estimators, while the second evaluates the predictions generated by the GBDT model and provides targets for subsequent estimators.

The first of these loss functions used in this thesis is the Mean Squared Error function (MSE) shown in equation 2.50, where N is the number of observations in the training set, y_i is the observed (or "actual") value for observation i, and γ_i is the predicted value for observation i on the given node or leaf. This function is used to fit the individual decision trees by selecting a feature and cut-off value that minimizes the MSE for each decision node.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \gamma_i)^2 \tag{2.50}$$

The second, and most important, loss function used in GBDT models in this thesis is a pseudo-least squares (LS) function, shown in equation 2.51. This LS function differs from a typical least-squares function due to its coefficient of 1/2, which simplifies the derivative $dLS/d\gamma$ shown in equation 2.52 by removing the factor of 2. This loss function is used to fit the final prediction values, and its derivative is the gradient referred to in the gradient boosted decision tree's name. The fitting of a GBDT model's overall goal is to generate a series of decision trees aimed at minimizing this LS function by improving on the previous guess.

$$LS(y,\gamma) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \gamma_i)^2 \tag{2.51}$$

$$\frac{dLS(y,\gamma)}{d\gamma} = -1\sum_{i=1}^{N}(y_i - \gamma_i) \tag{2.52}$$

2.5.11.2. Initial Guess

The first estimator in the GBDT is the only estimator which is not a decision tree. Instead, it creates an initial guess, γ_0, which is identical for every point in the training data. The initial guess applied to the set is the value that minimizes the loss function, which can be found by simply setting equation 2.52 equal to 0 and solving for γ_0 (equation 2.53). Once provided with an initial guess, subsequent estimators taking the form of decision trees are fit to improve upon that guess for each individual point in the training set.

$$\sum_{i=1}^{N}(\gamma_0 - y_i) = 0 \tag{2.53}$$

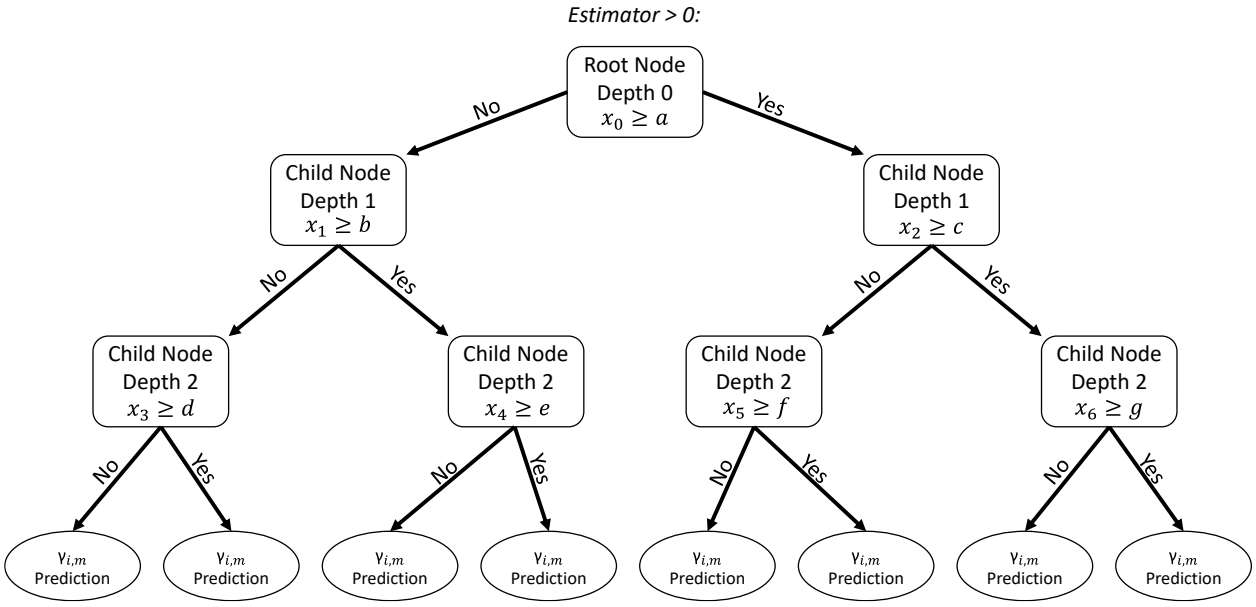


Figure 2.9 General structure of a decision tree estimator from a gradient boosted decision tree with a maximum depth of 3, with each branch terminating in a prediction "γ" _(i,m) where i is the individual in the set, and m is the estimator. This decision tree is using the feature space, x, where xj is the feature selected in node j, and the variables a, b, c, d, e, f, and g represent the values used to split the data.

2.5.11.3. Structure of Estimators (m>0)

Once an initial guess is generated, the gradient for each observation is calculated from the derivative of the loss function (equation 2.53). In the case of the LS function, this derivative can be simplified to become the difference between the *initial guess* and the observed value for each point in the set. This difference is then used as the target in the fitting of the next *estimator* or decision tree (m=1) which generates a new prediction, $\gamma_{i,m}$, for each observation i. The structure of the estimators where m>0 is shown in Figure 2.9. A new prediction is then generated for every observation using equation 2.54, where each value of $\gamma_{i,m}$ is weighted by a learning rate, α, to ensure the model makes incremental improvements over the previous prediction without overshooting the best solution. In this equation, $F_{i,m}(x)$ is the prediction for observation i after running m estimators, and when m=1, $F_{i,m-1}(x)$ is equal to γ_0 for all values of i. Once M estimators have been successfully fit, the final predictions are made using equation 2.55. Although the GBDT method excels at modelling non-linear trends in complex data, when more challenging problems are presented with massive datasets available, a more advanced technique called the artificial neural network (ANN) can be used.

$$F_{i,m}(x) = F_{i,m-1}(x) + \alpha\gamma_{i,m} \tag{2.54}$$

$$F_{i,M}(x) = \gamma_0 + \alpha \sum_{m=1}^{M} \gamma_{i,m} \tag{2.55}$$

2.5.12. Artificial Neural Networks

The artificial neural network (ANN)[63] is an advanced machine learning technique that has exploded in popularity in recent years due to its ability to model complex systems ranging from regression models to more advanced natural language processing.[64] ANNs are composed of a series of hidden layers, with each layer containing a set of nodes. The overall structure is meant to emulate a portion of the human brain, with the nodes in each hidden layer representing a neuron. By providing data to the input layer of the ANN, the network transforms and processes the data through the activation of individual neurons, which carries a "signal" to the output layer – providing the user with a prediction.

2.5.12.1. Multilayer Perceptron

The type of ANN used in this thesis is also known as a Multilayered Perceptron (MLP) or a feed-forward neural network.[65] An example of a generic MLP is presented in Figure 2.10, which shows an MLP with 4 hidden layers and 50 nodes (or neurons) per hidden layer. The data is fed into the input layer on the left and passed along each of the straight lines in the network until they reach a node, represented by a circle. The data passed from layer to layer is transformed using a linear function taking the form of the matrix multiplication shown in Equation 2.56. In this equation, the features are donated by the variable $x_{n,m}$, the feature number is denoted by the variable m, and the index of the individual point in the training set is denoted by the variable n. The weights, denoted by the variable w, are unique to each individual node, and each unique point (n) in the set passing though the network receives a single value for each node. This transformation will be performed for each node in the hidden layers, and once for the output layer. This means that a single MLP like the one shown in Figure 2.10 will perform the linear transformation 50(m +1)+3(50^2) times for each point in the data fed into the input layer. Therefore, a model based on 5 input features with the architecture shown in Figure 2.10 will perform 7,800 linear transformations for each point in the dataset.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ x_{3,1} & x_{2,2} & \cdots & x_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} x_{1,i} w_i \\ \sum_{i=1}^{m} x_{2,i} w_i \\ \sum_{i=1}^{m} x_{3,i} w_i \\ \vdots \\ \sum_{i=1}^{m} x_{n,i} w_i \end{bmatrix} \quad (2.56)$$

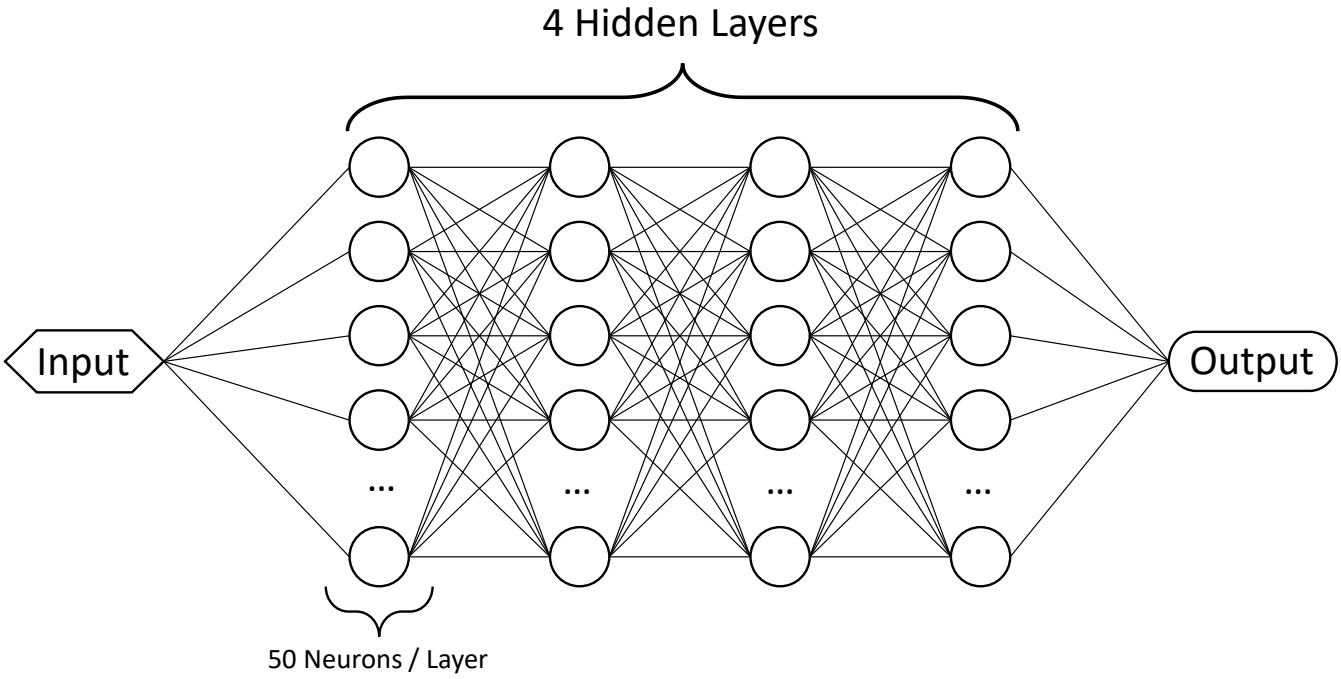


Figure 2.10 Diagram of an example multi-layered perceptron neural network including 4 hidden layers with 50 neurons / hidden layer. All lines connecting the nodes in the network represent the linear transformation of the data entering a neuron (or node) represented by a circle.

Once the linear transformation is complete to form the new matrix, the data is and processed at each neuron using an activation function. This activation function determines whether that node is "activated" in the same way that a neuron in a brain gets activated by an incoming signal. When a node is activated, the signal or value passed to the node is allowed to be propagated through the network. Conversely, when a node fails to activate a value of 0 is propagated from that node through the network and the output of that node does not contribute to the final output value of the MLP. The most

common activation function used in MLP models is the linear rectifier (ReLU) function shown in equation 2.57, whose behaviour is demonstrated in Figure 2.11. In this equation, the value x is the linearly transformed feature passed into the node from a previous layer. The resulting value from the neuron is then passed to the next layer in the MLP and the process is repeated.

$$ReLU(x) = \max(0, x) \tag{2.57}$$

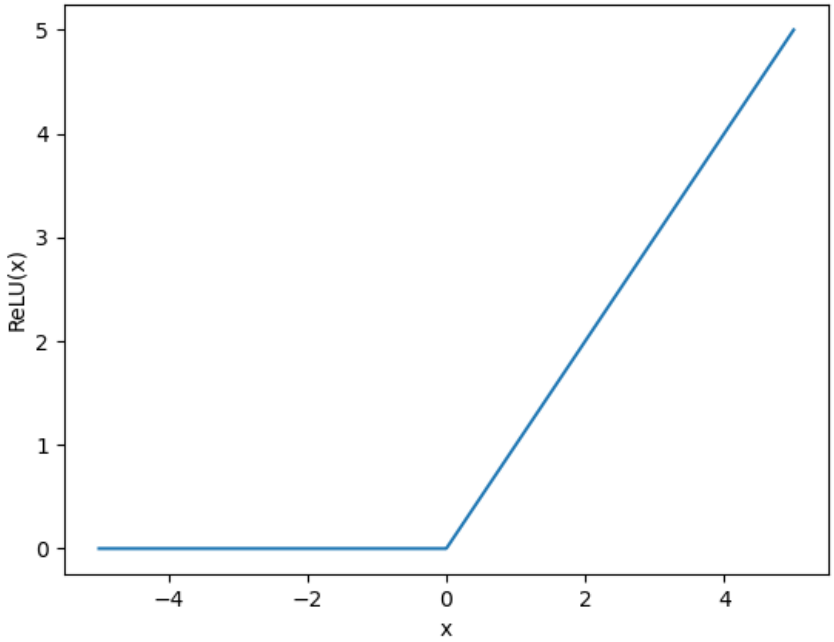


Figure 2.11 Plot of the Linear Rectifier (ReLU) function for values of x ranging from -5 to 5.

2.5.12.2. MLP Fittings and Backpropagation

The fitting of an MLP neural network is a time-consuming and computationally expensive process due to the large number of calculations required in the predictions. However, the conversion of the weights and features into tensors allows the required matrix operations to be performed rapidly using linear algebra libraries. These fittings require a step called backpropagation to adjust and test the weights implemented during the linear transformation steps. At the beginning of the optimization, an initial guess of the weights in the network is generated at random and the initial predictions are made. The algorithm then moves backwards through the network from the output layer in the direction of the input layer. An adjustment to the weight is then calculated using the Adams optimizer (see section 2.4.1)

and applied to the existing weight using equation 2.58 where w_{new} is the adjusted weight, w_{old} is the previous weight, δw is the calculated adjustment to the weight, and α is the learning rate. The learning rate is a parameter that controls the magnitude of the adjustments to the weights, ensuring gradual improvements to the model are made to avoid overshooting favourable solutions.

$$w_{new} = w_{old} + \alpha \delta w \tag{2.58}$$

Although the calculations required to fit and run neural networks appear to be complex and expensive, the reliance of ANNs on matrix operations allows the calculations to be easily handled with established linear algebra libraries. Therefore, these calculations are highly parallelizable and can be performed rapidly and easily on graphical processing units (GPUs). The ability of the ANN to effectively capture complex and non-linear relationships, their efficiency when fit on GPUs, and their rapid and easy deployment once trained has led to the ANN becoming an overwhelmingly popular tool. They have become ubiquitous in everyday life, powering technologies ranging from voice recognition on cell phones to self-driving cars.

2.6. References

1. Mesa, C. A., Francàs, L., Yang, K. R., Garrido-Barros, P., Pastor, E., Ma, Y., Kafizas, A., Rosser, T. E., Mayer, M. T., Reisner, E., Grätzel, M., Batista, V. S. & Durrant, J. R. Multihole water oxidation catalysis on haematite photoanodes revealed by operando spectroelectrochemistry and DFT. *Nature Chemistry* **12**, 82–89 (2020).

2. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature Materials* **5**, 909–913 (2006).

3. Bhandari, S., Rangarajan, S. & Mavrikakis, M. Combining computational modeling with reaction kinetics experiments for elucidating the in situ nature of the active site in catalysis. *Accounts of Chemical Research* **53**, 1893–1904 (2020).

4. Wang, C. Y. & Srinivasan, V. *Computational battery dynamics (CBD)-electrochemical/thermal coupled modeling and multi-scale modeling*.

5. Jankowski, P., Lastra, J. M. G. & Vegge, T. Structure of magnesium chloride complexes in ethereal systems: Computational comparison of THF and glymes as solvents for magnesium battery electrolytes. *Batteries and Supercaps* **3**, 1350–1359 (2020).

6. Nishijima, M., Ootani, T., Kamimura, Y., Sueki, T., Esaki, S., Murai, S., Fujita, K., Tanaka, K., Ohira, K., Koyama, Y. & Tanaka, I. Accelerated discovery of cathode materials with prolonged cycle life for lithium-ion battery. *Nature Communications* **5**, (2014).

7. Adhikari, U., Mostofian, B., Copperman, J., Subramanian, S. R., Petersen, A. A. & Zuckerman, D. M. Computational estimation of microsecond to second atomistic folding times. *Journal of the American Chemical Society* **141**, 6519–6526 (2019).

8. Levitt, M. & Warshel, A. *Computer simulation of protein folding*. *IO Philpotts, A. R., £con. Geol* vol. 54 (1960).

9. White, P., Haysom, S. F., Iadanza, M. G., Higgins, A. J., Machin, J. M., Whitehouse, J. M., Horne, J. E., Schiffrin, B., Carpenter-Platt, C., Calabrese, A. N., Storek, K. M., Rutherford, S. T., Brockwell, D. J., Ranson, N. A. & Radford, S. E. The role of membrane destabilisation and protein dynamics in BAM catalysed OMP folding. *Nature Communications* **12**, (2021).

10. Burns, T. D., Pai, K. N., Subraveti, S. G., Collins, S. P., Krykunov, M., Rajendran, A. & Woo, T. K. Prediction of MOF Performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

11. Zhang, H. & Snurr, R. Q. Computational study of water adsorption in the hydrophobic metal-organic framework ZIF-8: Adsorption mechanism and acceleration of the simulations. *Journal of Physical Chemistry C* **121**, 24000–24010 (2017).

12. Snurr, R. Q., Bell, A. T. & Theodorou, D. N. Prediction of adsorption of aromatic hydrocarbons in silicalite from grand canonical Monte Carlo simulations with biased insertions. *J. Phys. Chem* vol. 97 (1993).

13. Gómez-Gualdrón, D. A., Wilmer, C. E., Farha, O. K., Hupp, J. T. & Snurr, R. Q. Exploring the limits of methane storage and delivery in nanoporous materials. *Journal of Physical Chemistry C* **118**, 6941–6951 (2014).

14. Wilmer, C. E., Leaf, M., Lee, C. Y., Farha, O. K., Hauser, B. G., Hupp, J. T. & Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry* **4**, 83–89 (2012).

15. Simon, C. M., Kim, J., Gomez-Gualdron, D. A., Camp, J. S., Chung, Y. G., Martin, R. L., Mercado, R., Deem, M. W., Gunter, D., Haranczyk, M., Sholl, D. S., Snurr, R. Q. & Smit, B. The materials genome in action: Identifying the performance limits for methane storage. *Energy and Environmental Science* **8**, 1190–1199 (2015).

16. Li, J., Wang, Y., Chen, Z. & Rahman, S. S. Simulation of adsorption–desorption behavior in coal seam gas reservoirs at the molecular level: A comprehensive review. *Energy & Fuels* **34**, 2619–2642 (2020).

17. Getman, R. B., Bae, Y., Wilmer, C. E. & Snurr, R. Q. Review and analysis of molecular simulations of methane, hydrogen, and acetylene storage in metal-organic frameworks. *Chemical Reviews* **112**, 703–23 (2012).

18. Vaidhyanathan, R., Iremonger, S. S., Shimizu, G. K. H., Boyd, P. G., Alavi, S. & Woo, T. K. Direct observation and quantification of CO_2 binding within an amine-functionalized nanoporous solid. *Science (New York, N.Y.)* **330**, 650–3 (2010).

19. Torrisi, A., Bell, R. G. & Mellot-Draznieks, C. Functionalized MOFs for enhanced CO_2 capture. *Crystal Growth and Design* **10**, 2839–2841 (2010).

20. Campañá, C., Mussard, B. & Woo, T. K. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *Journal of Chemical Theory and Computation* **5**, 2866–2878 (2009).

21. Kresse, G. & Furthmüller, J. Vienna ab initio simulation package (VASP). (2001).

22. McDaniel, J. G., Li, S., Tylianakis, E., Snurr, R. Q. & Schmidt, J. R. Evaluation of force field performance for high-throughput screening of gas uptake in metal-organic frameworks. *Journal of Physical Chemistry C* **119**, 3143–3152 (2015).

23. Manz, T. A. & Sholl, D. S. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *Journal of Chemical Theory and Computation* **6**, 2455–2468 (2010).

24. Atkins, P. & de Paula, J. Physical Chemistry. eight ed. (2006).

25. Peng, D.-Y. & Robinson, D. B. A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals* **15**, 59–64 (1976).

26. Boyd, P. G. Computational high throughput screening of metal organic frameworks for carbon dioxide capture and storage applications. (2015).

27. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

28. Lennard-Jones, J. E. On the determination of molecular fields. II. From the equation of state of gas. *Proc. Roy. Soc. A* **106**, 463–477 (1924).

29. Huggins, M. L. & Mayer, J. E. Interatomic distances in crystals of the alkali halides. *The Journal of Chemical Physics* **1**, 643–646 (1933).

30. García-Sánchez, A., Ania, C. O., Parra, J. B., Dubbeldam, D., Vlugt, T. J. H., Krishna, R. & Calero, S. Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

31. Martin, M. G. & Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of n-Alkanes. *Journal of Physical Chemistry B* **102**, 2569–2577 (1998).

32. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: A reactive force field for hydrocarbons. *Journal of Physical Chemistry A* **105**, 9396–9409 (2001).

33. Coupry, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal-organic frameworks. *Journal of Chemical Theory and Computation* (2016).

34. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (1990).

35. Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Annalen der physik* **248**, 127–136 (1881).

36. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der physik* **369**, 253–287 (1921).

37. Haghpanah, R., Nilam, R., Rajendran, A., Farooq, S. & Karimi, I. A. Cycle synthesis and optimization of a VSA process for postcombustion CO_2 capture. *AIChE Journal* **59**, 4735–4748 (2013).

38. Krishnamurthy, S., Rao, V. R., Guntuka, S., Sharratt, P., Haghpanah, R., Rajendran, A., Amanullah, M., Karimi, I. A. & Farooq, S. CO_2 capture from dry flue gas by vacuum swing adsorption: A pilot plant study. *AlChE* **60**, 1830–1842 (2014).

39. MATLAB. *MATLAB R2016a*. (The MathWorks Inc., 2016).

40. Chen, J., Loo, L. S. & Wang, K. An ideal absorbed solution theory (IAST) study of adsorption equilibria of binary mixtures of methane and ethane on a templated carbon. *Journal of Chemical and Engineering Data* **56**, 1209–1212 (2011).

41. Simon, C. M., Smit, B. & Haranczyk, M. PyIAST: Ideal adsorbed solution theory (IAST) Python package. *Computer Physics Communications* **200**, 364–380 (2016).

42. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. (2014).

43. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901).

44. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319 (1998).

45. Baker, J. E. & others. Reducing bias and inefficiency in the selection algorithm. in *Proceedings of the second international conference on genetic algorithms* vol. 206 14–21 (1987).

46. Whitley, D. *A genetic algorithm tutorial*. vol. 4 (1994).

47. Chipperfield, A. J. & Fleming, P. J. The MATLAB genetic algorithm toolbox. (1995).

48. Hooper, J., Ismail, A., Giorgi, J. B. & Woo, T. K. Computational insights into the nature of increased ionic conductivity in concentrated samarium-doped ceria: A genetic algorithm study. *Physical Chemistry Chemical Physics* **12**, 12969–12972 (2010).

49. Le, T. C. & Winkler, D. A. Discovery and optimization of materials using evolutionary approaches. *Chemical Reviews* vol. 116 6107–6132 (2016).

50. McKay, M. D., Beckman, R. J. & Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **42**, 55–61 (2000).

51. Kochendörffer, R. Kreyszig, E.: Advanced Engineering Mathematics. J. Wiley & Sons, Inc., New York, London 1962. IX + 856 S. 402 Abb. Preis s. 79.—. *Biometrische Zeitschrift* **7**, (1965).

52. Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* **36**, 111–133 (1974).

53. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. http://robotics.stanford.edu/~ronnyk (1995).

54. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer New York, 2002).

55. Quinlan, J. R. *Induction of Decision Trees*. *Machine Learning* vol. 1 (1986).

56. Ceriani, L. & Verme, P. The origins of the Gini index: Extracts from Variability to Mutability (1912) by Corrado Gini. *Journal of Economic Inequality* **10**, 421–443 (2012).

57. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P. & Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947–1958 (2003).

58. Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S. B., Janes, C. R., Lanphear, B. P., McCandless, L. C., Takaro, T. K., Venners, S. A., Webster, G. M. & Allen, R. W. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental Pollution* **245**, 746–753 (2019).

59. Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C. W., van den Bossche, P., van Mierlo, J. & Omar, N. Random forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy* **232**, 197–210 (2018).

60. Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y. & Pham, B. T. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment* **662**, 332–346 (2019).

61. Chen, W., Zhang, S., Li, R. & Shahabi, H. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Science of the Total Environment* **644**, 1006–1018 (2018).

62. Hastie, T., Tibshirani, R. & Friedman, J. *Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*. (Springer, 2009).

63. Mcculloch, W. S. & Pitts, W. *A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*. *BULLETIN OF MATHEMATICAL BIOPHYSICS* vol. 5 (1943).

64. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* vol. 521 436–444 (2015).

65. Haykin, S. S. & Haykin, S. S. *Neural networks and learning machines*. (Prentice Hall/Pearson, 2009).

2.7. Appendix 2.1: 4-Stage Light-Particle Pressurization Boundary Conditions

Table A2.1 Boundary conditions for the 4-stage LPP PSA/VSA simulations

Step	$z = 0$	$z = L$	
Adsorption	$D_L \frac{\partial y_i}{\partial z}\Big\|_{z=0} = -v\|_{z=0}(y_{i,feed} - y_i\|_{z=0})$	$\frac{\partial y_i}{\partial z}\Big\|_{z=L} = 0$	(A.2.1)
	$K_Z \frac{\partial T}{\partial z}\Big\|_{z=0} = -\varepsilon v\|_{z=0}\rho_g C_{p,g}(T_{feed} - T\|_{z=0})$	$\frac{\partial y_i}{\partial z}\Big\|_{z=L} = 0$	(A.2.2)
	$K_Z \frac{\partial T}{\partial z}\Big\|_{z=0} = -\varepsilon v\|_{z=0}\rho_g C_{p,g}(T_{feed} - T\|_{z=0})$	$\frac{\partial T}{\partial z}\Big\|_{z=L} = 0$	(A.2.3)
	$T_w\|_{z=0} = T_a$	$T_w\|_{z=L} = T_a$	(A.2.4)
	$v\|_{z=0} = v_{feed}$	$P\|_{z=L} = P_H$	(A.2.5)
Blowdown	$\frac{\partial y_i}{\partial z}\Big\|_{z=0} = 0$	$\frac{\partial y_i}{\partial z}\Big\|_{z=L} = 0$	(A.2.6)
	$\frac{\partial T}{\partial z}\Big\|_{z=0} = 0$	$\frac{\partial T}{\partial z}\Big\|_{z=L} = 0$	(A.2.7)
	$T_w\|_{z=0} = T_a$	$T_w\|_{z=L} = T_a$	(A.2.8)
	$v\|_{z=0} = 0$	$P\|_{z=L} = P_{INT} + (P_H - P_{INT})e^{-\alpha t}$	(A.2.9)
Evacuation	$\frac{\partial y_i}{\partial z}\Big\|_{z=0} = 0$	$\frac{\partial y_i}{\partial z}\Big\|_{z=L} = 0$	(A.2.10)
	$\frac{\partial T}{\partial z}\Big\|_{z=0} = 0$	$\frac{\partial T}{\partial z}\Big\|_{z=L} = 0$	(A.2.11)
	$T_w\|_{z=0} = T_a$	$T_w\|_{z=L} = T_a$	(A.2.12)
	$P\|_{z=0} = P_L + (P_{INT} - P_L)e^{-\alpha t}$	$v\|_{z=L} = 0$	(A.2.13)
Feed Pressurization	$D_L \frac{\partial y_i}{\partial z}\Big\|_{z=0} = -v\|_{z=0}(y_{i,feed} - y_i\|_{z=0})$	$\frac{\partial y_i}{\partial z}\Big\|_{z=L} = 0$	(A.2.14)
	$K_Z \frac{\partial T}{\partial z}\Big\|_{z=0} = -\varepsilon v\|_{z=0}\rho_g C_{p,g}(T_{feed} - T\|_{z=0})$	$\frac{\partial T}{\partial z}\Big\|_{z=L} = 0$	(A.2.15)
	$T_w\|_{z=0} = T_a$	$T_w\|_{z=L} = T_a$	(A.2.16)
	$P\|_{z=0} = P_H - (P_H - P_{PRESS})e^{-\alpha t}$	$v\|_{z=L} = 0$	(A.2.17)
Light Product Pressurization	$\frac{\partial y_i}{\partial z}\Big\|_{z=0} = 0$	$D_L \frac{\partial y_i}{\partial z}\Big\|_{z=L} = -v\|_{z=L}(y_{i,ads} - y_i\|_{z=L})$	(A.2.18)
	$\frac{\partial T}{\partial z}\Big\|_{z=0} = 0$	$K_Z \frac{\partial T}{\partial z}\Big\|_{z=L} = -\varepsilon v\|_{z=L}\rho_g C_{p,g}(T_{ads} - T\|_{z=L})$	(A.2.19)
	$T_w\|_{z=0} = T_a$	$T_w\|_{z=L} = T_a$	(A.2.20)
	$v\|_{z=0} = 0$	$v\|_{z=L} = \frac{v_{ads}P_{ads}\|_{z=L}}{P\|_{z=L}}$	(A.2.21)

3. Chapter 3: Guest Atom Localization Algorithm

The work discussed in this chapter was a collaborative effort between me and former members of the research lab of Dr. Tom Woo and is the subject of a manuscript currently in progress. For full details on my contributions to this work, see section 3.5 – Author Contributions.

3.1. Abstract

An automated method for locating binding site locations in nanoporous materials based on grand canonical Monte Carlo (GCMC) was developed and optimized for use in high-throughput screening applications. This method, called the Guest Atom Localization Algorithm (GALA) smooths the noisy probability distributions generated by GCMC and follows a series of steps to select maxima and fit guest molecules to those maxima. To use GALA in high-throughput screening applications, the control parameters needed to be optimized for the individual guest molecules being studied. In this chapter, the control parameters were optimized for use with CO_2, chosen due to its importance in post-combustion carbon capture applications. Using these optimized parameters, the binding sites generated by GALA were compared to experimentally determined CO_2 sites to confirm the accuracy of the algorithm.

3.2. Introduction

Nanoporous metal organic framework (MOF) materials[1,2] have received significant attention as materials for gas purification systems to be used in clean energy technologies, in particular, post-combustion carbon capture and storage (CCS).[3–12] MOFs have been discovered that possess 'world-record' internal surface areas of >5000 m^2/g that selectively and reversibly capture CO_2.[13,14] Such materials could be used as the solid sorbents in pressure and/or temperature swing adsorption (PSA/TSA) systems[15,16] for large scale and energy efficient post-combustion CO_2 scrubbing. Compared to other porous materials such as zeolites, MOFs promise greater opportunities for tuning the material's properties due to the nearly endless combinations of metal ions, organic linker groups and linker substituents that can make up the MOF structure. The tunability would allow for the optimization of MOFs for applications such as post-combustion CO_2 capture, where a high selectivity for CO_2 over N_2 is desired, or precombustion CO_2 capture, where a high selectivity for CO_2 over H_2 is desired. Despite their potential, the rational design of MOFs for gas adsorption applications is hampered by the microscopic

nature of the interactions being generally inaccessible by experiment. Without this detailed information, the potential to develop design principals is limited.

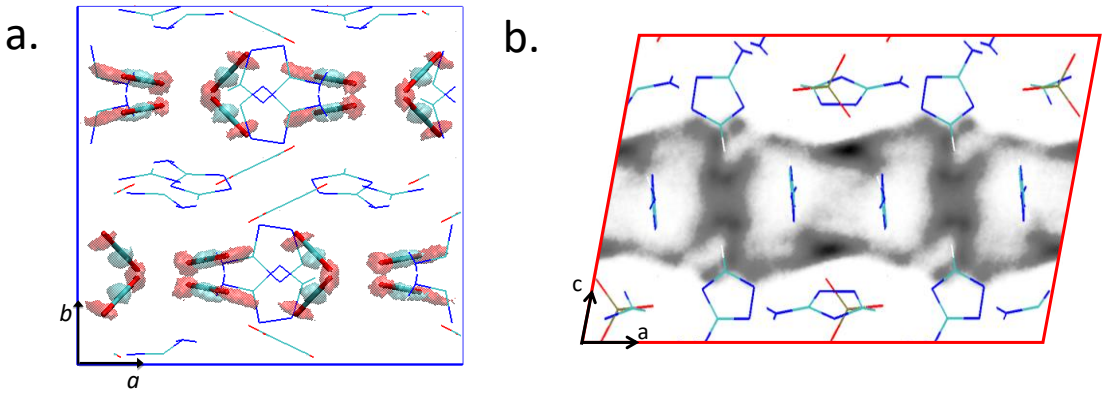


Figure 3.1 a) 3D-isosurfaces of the CO_2 probability distributions (cyan – carbon; red – oxygen) in CALF-15 (Zn2(3-amino-1,2,4-triazole)2(oxalate)), determined from a GCMC simulation.17 Also shown in tube representation are the experimental CO_2 binding sites determined from X-ray analysis. b) Centre of mass probability density plots of CO_2 molecules in CALF-16 (Zn3(3-amino-1,2,4-triazole)3(PO4)). In both a) and b), the framework of the MOF is shown in line representation with H and Zn atoms removed for clarity.

Experimentally, the gas adsorption properties of a MOF are usually measured by evaluating the gas uptake as a function of the partial pressure of the guest at constant temperature, which gives a gas adsorption isotherm. Molecular simulation has developed into a powerful tool to examine gas adsorption in MOFs[18–20] and simulated adsorption isotherms determined from the GCMC simulations are now frequently reported alongside their experimental counterparts. In a GCMC simulation, the uptake capacity is computed in a 'brute force' manner, where the guest-host interaction energies are sampled via a Monte Carlo scheme. A typical GCMC simulation will sample millions of guest configurations within the host to generate a single point on the adsorption isotherm. In addition to the net gas uptake, the sampling of the guest configurations in a GCMC simulation also yields the probability distribution of the gas within the MOF. These distributions can be discretized on a three-dimensional (3D) grid and are often visualized with 3D-isosurfaces or 2D contour/density plots as shown in Figure 3.1a and 1b, respectively. In many cases, the probability density is localized in specific regions, which we term 'binding sites', responsible for the majority of the gas capacity. The guest molecule binding sites can often be established by visual inspection if the simulated probability distributions are localized and well converged. Figure 3.1a shows a probability distribution that is highly localized, whereas that shown in Figure 3.1b is more diffuse.

Since the gas adsorption isotherms are typically well reproduced by the GCMC simulations (albeit often with some adjustment of the simulation parameters), one might also expect the binding sites to be well reproduced by the simulations. Unfortunately, the binding sites are rarely determined experimentally, particularly when the binding is physisorptive. In one of the first cases where the CO_2 binding sites were experimentally located in a MOF, they were indeed found to be in excellent agreement with those extracted from the simulated probability density distributions. Figure 3.1a, shows the agreement between the simulated and experimental CO_2 binding sites in the MOF CALF-15. In this study, and a follow-up study on the related MOF CALF-16,[21] the CO_2 binding sites were determined from the calculated probability distributions by identifying maxima manually, a labour-intensive exercise demonstrated in Figure 3.2.

Recently, high throughput virtual screening of MOFs has become possible.[22–24] For example, in a pioneering work, Wilmer and Snurr screened ~130,000 MOFs for their methane and CO_2 uptake capacity with GCMC simulations.[22,24] In these studies, structure-property relationships have been established using geometric features of the MOFs such as the void volume and the maximum pore sizes. Further studies on MOF databases have been published attempting to relate equilibrium adsorption to process level pressure swing adsorption systems, linking atomistic and process scale simulations. Analysis of the nature of the binding sites in the highest performing MOFs resulting from such large-scale screening could establish key structural and chemical features that chemists could use for the rational design of MOFs. Manual localization of the binding sites (Figure 3.2) based on GCMC probability distributions on such large datasets is not practical and would require an algorithm that can consistently locate the binding sites from the generated probability distributions in an automated fashion.

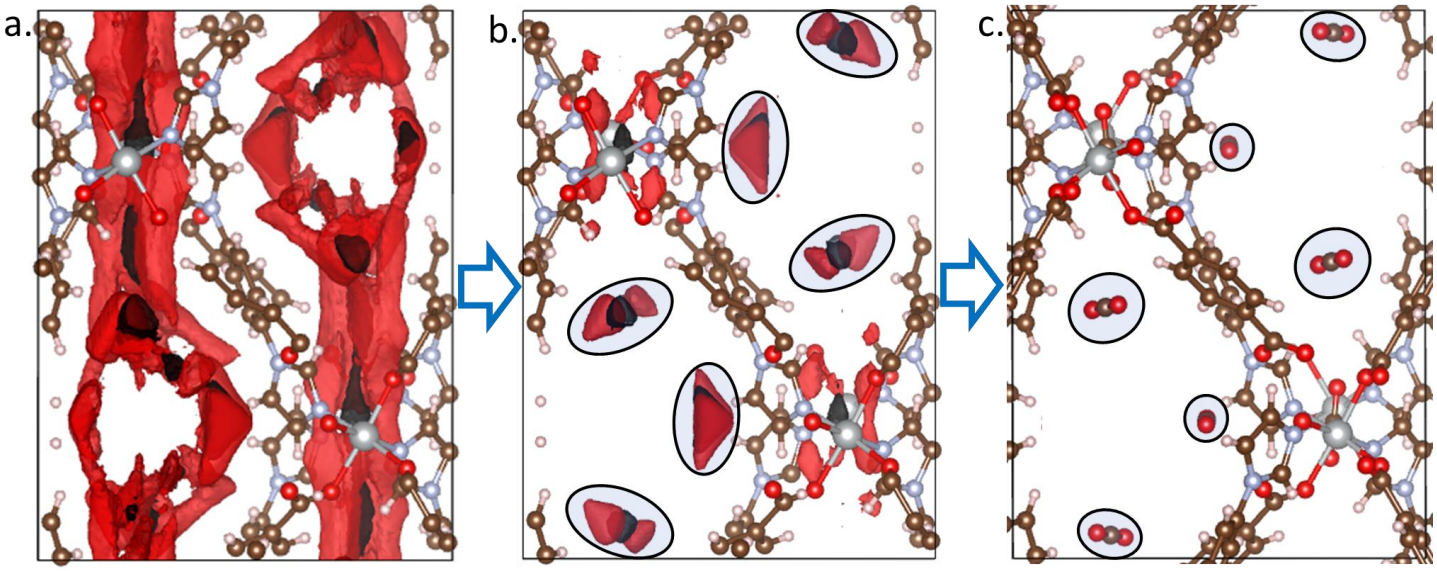


Figure 3.2 a) Unrefined CO_2 isosurfaces with the oxygen (red) surfaces and carbon (black) surfaces, generated from a simulation of hypothetical MOF str_m19_o180_ubt with 10 million GCMC cycles. b) The same isosurfaces as in (a) but with a reduced isosurface value revealing areas of high probability corresponding to CO_2 binding sites (circled). c) The same MOF with CO_2 molecules placed by GALA (circled).

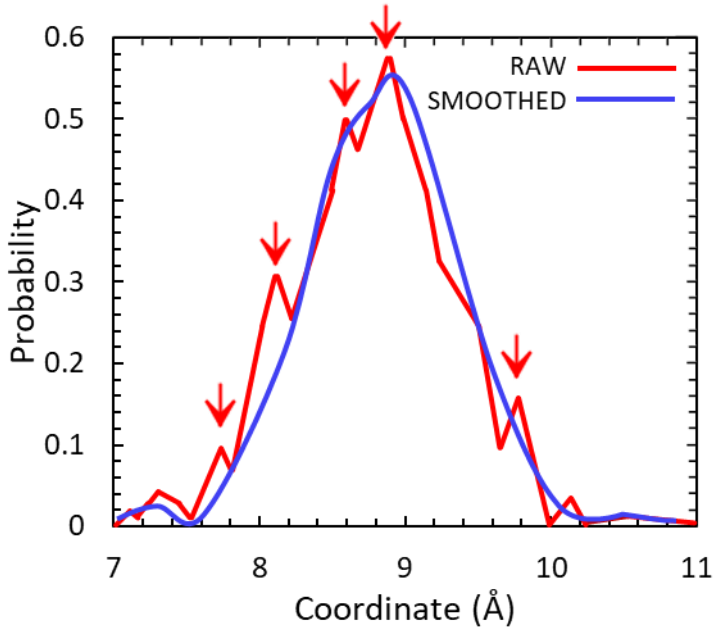


Figure 3.3 1D probability distribution of the carbon atom derived from GCMC simulation of CO_2 gas adsorption in MOF CALF-15 (red). The probability is plotted along a line which passes through one of the binding sites. The red arrows point to local maxima in the raw probability distribution. The blue line is the result of a 'smoothing' of raw probability distribution with a frequency filter (this work).

There are several challenges that an automated binding site localization algorithm must overcome. The most obvious challenge is that the data is noisy since the data points are generated with a stochastic Monte Carlo scheme. In principle a smoothed probability distribution can be obtained with enough sampling, however, when screening hundreds of thousands of MOFs, long GCMC runs are undesirable. For example, the red line shown in Figure 3.2 is the probability distribution in the vicinity of the binding site from a simulation on CALF-15 where the error in the CO_2 uptake is determined to be only 2.1% (4.69±0.10 mmol CO_2/g MOF). Although the uptake is well converged, there are numerous local maxima (red arrows in Figure 3.3) in the raw probability distribution (red solid line in Figure 3.3) in the vicinity of the CO_2 binding site. It is not unusual for the raw GCMC-derived probability distributions to possess thousands of local maxima for a MOF with fewer than 20 'true' binding sites. This suggests that an algorithm will be presented with noisy data and that a 'smoothing' or noise reduction algorithm needs to be employed.

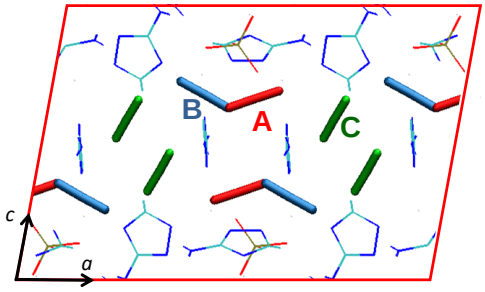


Figure 3.4 Select CO_2 binding sites determined from maxima in the calculated probability distributions in MOF CALF-16. CO_2 molecules are shown in tube representation while the MOF framework is shown with line representation with the H and Zn atoms removed for clarity.

Another anticipated problem with locating the binding sites in an automated fashion is the possibility of the presence of spatially close or even overlapping binding sites. For example, in a study of CALF-16[25] it was found that a number of binding sites, some of which are depicted in Figure 3.4 and labeled A through C. These were obtained by manually locating maxima in the probability distributions (Figure 3.1b). The binding sites labeled A and B in Figure 3.4 are overlapping and mutually exclusive, with site A having a significantly higher probability than site B. Molecular dynamics simulations were consistent with this picture showing that a CO_2 molecule in binding site A would occasionally 'slide' into site B and back. Thus, the binding site algorithm should not only be able to handle noisy input probability distributions, but also handle overlapping binding sites.

To handle the probability distributions in large MOFs, a binding site location algorithm must also be able to efficiently handle a large number of data points. For example, to have a grid resolution of 0.1 Å, the probability distribution of well-studied MOF-5, which has a cubic unit cell dimension of 25.7 Å, would contain over 10^7 grid points. MOF-5 is a prototypical MOF and does not possess a large unit cell.

Considering the challenges, the goal in this work is to develop a binding site localization algorithm that can be used to locate the binding sites of physisorbed guests from GCMC derived probability distributions in an automated manner. Once the algorithm is developed, the input parameters controlling these calculations need to be tuned to allow for implementation in high-throughput screening. Finally, the algorithm needs to be tested on systems with experimentally determined or manually determined binding sites to prove its accuracy.

Ideally, such an algorithm should have some of the following features: a) the ability to handle noisy data and avoid identification of "false" maxima; b) a minimal number of parameters that are only required for fine-tuning and are intuitive with physical meaning; c) the ability to deal with periodic nature of MOFs and identify maxima that may occur at the simulation cell boundaries; d) the ability to handle large datasets (>10^7 points) efficiently; e) to provide some meaningful ranking of binding sites so that insignificant sites can easily be ignored.

In this work, the methodology for automatic localization of guest-molecule binding sites from GCMC-derived probabilities which satisfies the aforementioned criteria that we call GALA (Guest Atom Location Algorithm) is formulated and implemented. The developed procedure enables automation of the binding site localization and is suitable for interfacing with high throughput methods for MOF screening. The parameters associated with this algorithm have been optimized for high throughput screening of CO_2 in nanoporous materials and validated on over 300 MOFs with a range of geometric and adsorption properties where the binding sites are known experimentally or have been found manually.

3.3. Methodology & Results

3.3.1. How GALA Works

3.3.1.1. Overview

In section 3.3.1, a brief outline of the GALA methodology is given followed by full details of each step including descriptions of the user adjustable parameters that affect the outcome of the binding site calculations:

1. Within the GCMC code, 3-dimensional probability distributions are generated for each atom type in the guest molecule using an equitable binning procedure to reduce noise.
2. A Gaussian noise filter is applied to the probability distributions to further reduce stochastic noise.
3. Local maxima in the probability distributions are identified. Low occupancy maxima are discarded along with maxima that are in close proximity to maxima with higher occupancies.
4. The guest molecule is best fit to the location of the maxima.
5. Guest molecule positions are optionally geometry optimized and the binding energies calculated.

Following the identification of binding sites, one can then report these in a user accessible format ranked by the probability values at each maximum. The GALA code was written in Python 2.7.

3.3.1.2. Generation of the Probability Distributions

GALA uses a probability distribution in which the volume of the simulation cell is uniformly discretized along each cell vector, thereby forming a 3D grid in which the center of each discretization volume forms a grid point. To generate the probability distribution during a MC simulation, the positions of each of the atoms of the guest molecules are binned into the closest grid points, generating a running total over the course of the simulation, updated after each MC step. We define the occupancy, ρ_i, at the grid point 'i' as given by equation 3.1. The volume of the bin in equation 3.1 is determined assuming a triclinic unit-cell, calculated using in equation 3.2 where a, b, and c are the lengths of the three unit-cell vectors, α, β, and γ are the angles between the unit-cell vectors, and N is the number of bins in a unit cell.

$$\rho_i = \frac{number\ of\ steps\ guest\ atom\ binned\ at\ point\ i}{(total\ number\ of\ MC\ steps)\cdot(volume\ of\ bin)} \quad (3.1)$$

$$Volume\ of\ Bin = \frac{abc\sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma}}{N} \quad (3.2)$$

3.3.1.3. Initial Smoothing Technique: Equitable Binning

We found that the manner in which the 3D probability distributions are generated from a GCMC run can strongly influence the effectiveness of the binding site location. Most importantly, we determined that an equitable binning procedure[22] greatly reduces the noise in probability distributions that are generated. Figure 3.5 compares a normal binning procedure with an equitable binning procedure. In normal binning, the position of an atom is assigned wholly to the grid volume it falls within and a value of 1 is added to that grid volume's bin. When using equitable binning, the distance to the center of each of the closest 8 grid volumes are identified and the value of 1.0 is distributed proportionally based on the proximity of the atom to each grid point. Those fractions are then accumulated into the eight individual bins, as shown in the 2D case in Figure 3.5b. Equitable binning effectively decreases the noise in the probability distributions thereby reducing the length of the simulations required to locate the binding sites.

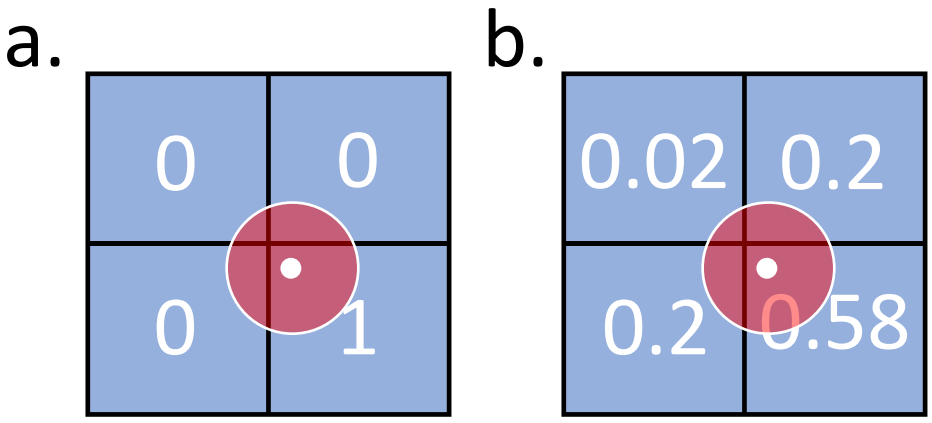


Figure 3.5 A 2D representation of (a) normal binning procedure and (b) equitable binning. The brown circle with the central white dot depicts the position of the atom. The numbers in each grid area show the amount that the atom contributes to the probability distribution in each grid area for the two binning procedures.

3.3.1.4. Grid Resolution and Grid Size

The resolution of the probability distribution is controlled by the grid spacing. We introduce a *grid size* parameter, which is simply the length of the discretization of the simulation cell volume along each cell vector. Although a higher resolution grid is intuitively desirable, a more coarse grid does have a smoothing effect. This is because with a larger grid spacing, the guest positions are binned into larger grid volumes. However, if the grid spacings are too large, resolution is lost, and the locations of the maxima can be made inaccurate. Through some optimization to be discussed later in the *Results and Discussion* section, a good working value of the *grid size* was found to be 0.15 Å, which provides a balance between smoothing of the distributions and the accuracy of the binding site locations.

3.3.1.5. Probability Plot Convergence: Tanimoto Coefficient

One practical aspect in generating the probability distribution for locating binding sites involves how long one should run the GCMC simulation before the probability distribution is considered converged. This not only depends on the combination of the guest and material, but will also depend on various simulation parameters such as the maximum size of the MC perturbations, the nature of the MC moves, etc. As a result, we attempted to develop a convergence metric that would be independent of such parameters and variables. One can of course simply run a very long simulation to ensure one has a well converged set of probability distributions. However, in high throughput screening scenarios, this may not be practical when one wants to identify the binding sites in tens of thousands of materials.

The convergence criteria we settled on involves comparing the probability distribution from two or more GCMC chains that are run in parallel or by comparing the probability distribution of two equivalent volumes in a supercell simulation. If the distributions are similar enough, the probability distributions are considered converged. To evaluate the similarity between two probability distributions A and B, the *Tanimoto coefficient*, T, is used as defined in equation 3.3 where the summations are over all equivalent grid points. The Tanimoto coefficient ranges between 0 and 1 where a value of 0 means that the two distributions are completely dissimilar, while a value of 1 means that the two distributions are identical. A recommended value of the Tanimoto coefficient which balances binding site accuracy with simulation length, the determination of which is shown in the *Results and Discussion* section, was found to be 0.725.

$$T = \frac{\sum A_i B_i}{\sum A_i A_i + \sum B_i B_i - \sum A_i B_i} \tag{3.3}$$

If GALA is to be implemented, the GCMC code will have to be modified to generate probability distributions as previously outlined using a uniform discretization of the unit cell volume, an equitable binning procedure, and a Tanimoto test of convergence. We note that one probability distribution is required for every atom type in the guest molecule, as well as the center of mass of the guest. For example, if CO_2 is the guest molecule, a separate probability distribution is generated for the carbon and oxygen atoms. In that case the center of mass and carbon atom positions are the same.

The remainder of the GALA method is independent of the simulation code that generates the probability distribution. We have implemented this portion of GALA in our own code written in Python 2.7 using the NumPy and SciPy libraries. It reads a probability distribution in Gaussian CUBE format, performs smoothing operations, and places guest molecules on the selected maxima. Code is free upon request.

3.3.1.6. Smoothing of the Probability Distribution

The next step of GALA involves further reduction of the stochastic noise in the probability distribution. Without smoothing, the noise in the probability distribution will result in too many maxima being located giving 'false' binding sites. On the other hand, excess smoothing can result in the loss of meaningful maxima that correspond to relevant binding sites. GALA uses a standard Gaussian noise filter to smooth the probability distributions. Gaussian filters are low pass filters that remove high frequency noise that work efficiently in real space. The Gaussian kernel given in equation 3.4 is used to smooth the data in each cell axis.

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} \tag{3.4}$$

Sigma, σ, controls the level of smearing and is an adjustable parameter in GALA. As *sigma* increases, the width of each Gaussian kernel increases and the amount of smoothing increases. To allow for high-throughput screening, an ideal sigma value of 2.0 Å was selected following an optimization described in the *Results and Discussion* section of this thesis.

3.3.1.7. Placement of the binding sites

Once the smoothing of the distributions is complete, the binding sites can be located. This involves locating all maxima in the smoothed probability distribution using a simple grid search

algorithm where a grid point is identified as a maximum when the occupancy in all 26 adjacent grid points is lower. Maxima that fall below a user adjustable *occupancy cut-off* are also discarded. The *occupancy cut-off* is an important parameter as we find that there are often many low occupancy maxima that appear in the void spaces of medium and large pore materials, particularly at adsorption pressures of less than 1 bar. Optimization of this *cut-off* parameter is also performed concurrently with the optimization of the *sigma* parameter.

3.3.1.7.1. Vetting the Maxima

Maxima are further vetted by removing those that are within a user adjustable *exclusion radius* of a maximum with a higher occupancy. Specifically, if more than one maximum is within the *exclusion radius* of other maxima, only the maximum with the highest occupancy is retained. The ideal *exclusion radius* will depend on the nature of the guest molecules and the atom types. For high-throughput screening of CO_2 an *exclusion radius* of 0.675 Å was found to be a good working value and eliminates the risk of placing equivalent binding sites.

3.3.1.7.2. Rules of Guest Placement

Following identification and elimination of the local maxima in the probability distributions, the guest molecules must be placed in the location of the maxima. For single atom guests, such as Ar, the atom is simply placed on the maxima. For diatomic molecules, one atom is first placed on the maximum with the highest occupancy. The next highest occupancy maxima that is within 20% of the equilibrium bond distance of the guest molecule is chosen as the second local maxima. The second atom is then positioned with the bond aligned with the direction vector between the two maxima. The guest molecule is then slid along the bond vector to give an occupancy weighted best fit to the position of the two maxima. Polyatomic linear molecules are placed using a similar methodology as diatomic molecules. However, because the maxima may not lie in a perfect straight line, the guest molecule is rotated and translated from the initial placement to give an occupancy weighted best fit to all atoms of the molecule. For non-linear molecules, the center of mass (**A1**) is placed on the maximum. An arbitrary second atom (**A2**) of the guest is positioned in the same way the second atom of a diatomic is positioned. A third maxima is then identified corresponding to an arbitrary third atom **A3**. The molecule is then rotated about the **A1**-**A2** bond to position the third atom to align the **A1**-**A2**-**A3** molecular plane with the plane of the three maxima. The molecule is then translated and rotated about the center of mass to give an occupancy weighted best fit to all atoms.

Once the guest molecules are placed, their positions can be optionally optimized using the same force field used in the MC simulations, where a single guest molecule is placed in the empty material. Such an optimization can act to fine tune the binding sites and to symmetrize equivalent binding sites. However, it is important to note that this single molecule potential energy surface may not closely match the free energy surface given by the probability distribution. For example, the potential energy surface does not account for entropic, cooperative, or competitive binding effects. For this reason, when geometry optimization is performed, we flag any binding sites whose center of mass changes by more than 0.2 Å following optimization.

3.3.1.7.3. Binding Energy Calculation

Once the placement of the binding sites is complete, they are ranked according to their occupancies and the binding energies are calculated. This is performed using a single point molecular dynamics simulation with the DL POLY software package,[27,28] where the guest molecule is placed in the empty framework, and the binding energy is using the same force fields used in the GCMC simulation. Again, caution should be used when interpreting the binding energy since it does not include other guest molecules and therefore does not account for any cooperate or competitive effects the guest molecules may experience within the pores of the MOF.

3.3.2. Optimization of Parameters

3.3.2.1. Grand-Canonical Monte Carlo Simulations

The first step in optimizing GALA's control parameters was the generation of the probability distributions using GCMC. All GCMC simulations were performed using an in-house[13] code modelled based on the DL_POLY Classic code.[14] For these simulations, the Lennard-Jones parameters for the framework atoms were taken from the Universal Force Field (UFF)[30] and partial atomic charges were fit to the framework atoms using the REPEAT method.[31] Single component simulations were run to model the adsorption of CO_2 at 298 K and 0.15 bar using the Garcia-Sanchez parameters to model CO_2 adsorption.[32] Fugacities were calculated using the Peng-Robinson Equation of State.[20] Simulations were performed using 30,000 equilibration cycles and a variable number of production cycles as this was one of the parameters to be optimized. This methodology has been shown to reproduce experimental isotherms.[17,21,34]

3.3.2.2. Preparation of Datasets

To ready the GALA code for use in high-throughput screening applications, a generalized set of control parameters needed to be found which would allow the algorithm to run with little to no user intervention. A full list of these parameters along with the final recommended values optimized for CO_2 can be found in Table 3.1. As many of the parameters are interdependent, meaning the best value for one parameter may depend on the value of another, the optimizations could not be performed independently. As such, the optimization of the control parameters was performed in three stages. The first stage involved optimizing the grid-spacing in the probability plot, labeled as *grid size*, in Table 3.1. The second involved the curation of a ***small set*** composed of experimentally realized and hypothetical MOFs aimed at optimizing the *Tanimoto coefficient* and the *exclusionary radius*. The third and final stage focused on the optimization of the *sigma* and *occupancy cut-off* parameters using a ***large set*** of hypothetical MOFs. This diversity of the MOFs in the ***small*** and ***large sets*** is demonstrated in Figure 3.6a which shows the range of volumetric uptakes vs the gravimetric uptakes (calculated at 1 bar and 298 K) for the ***small set*** (yellow squares) and the ***large set*** (blue circles). The diversity in the physical properties is demonstrated in Figure 3.6b, which shows a plot of the internal surface area against the largest pore diameter for all MOFs in both sets.

Table 3.1 GALA control parameters and their optimized values for high throughput screening of CO^2 adsorption in nanoporous materials.

Control parameter	***description***	***recommended value***
grid size, g	Controls the discretization size of the probability distribution along each cell vector.	0.15 Å
Tanimoto, T	Used to determine when the probability distribution is converged enough to stop the GCMC simulation.	0.75
sigma, σ	Initial Guess for smoothing parameter of Gaussian noise filter. A larger value increases smoothing.	2.0 Å
exclusion radius, R_{xc}	Radius to exclude overlapping binding sites.	0.675 Å
occupancy cut-off, O_{min}	Initial guess for minimum occupancy value for which local maxima are considered. Eliminates low probability sites.	10.0%

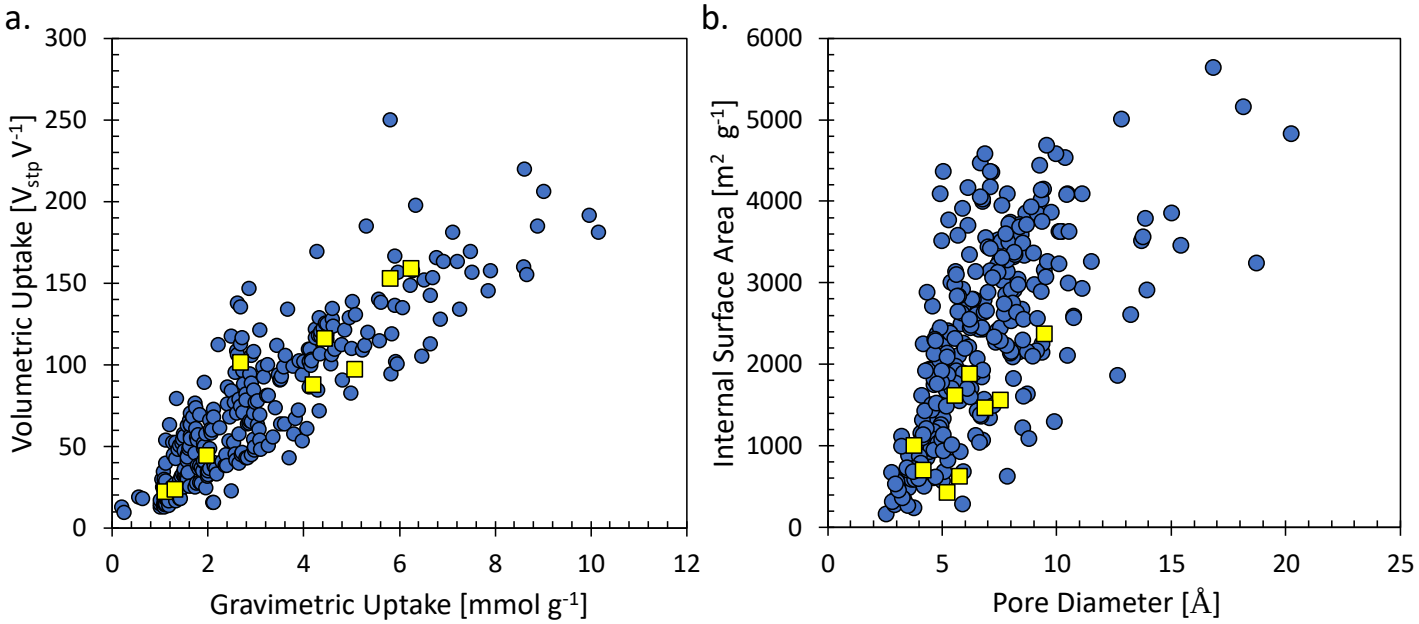


Figure 3.6 Adsorption and geometric properties of the Small Set (yellow squares) and the Large Set (blue circles). Scatter plots of a) the calculated volumetric versus the gravimetric CO_2 uptake (at 1 bar and 298 K) of the MOFs and b) the surface area versus the maximum pore diameter.

3.3.2.3. By Hand Binding Site Identification

With such a small pool of MOFs with experimentally determined CO_2 binding sites available, tuning of GALA control parameters relied heavily hypothetical MOF structures generated in lab, while the majority of experimentally determined binding sites were kept for final validation. This required the determination of CO_2 binding sites 'by-hand' based on high quality probability distributions. Using a probability plot spacing of 0.15 Å, GCMC simulations were run using an excessive number of production steps (20 million) to ensure the plots were sufficiently converged, as the *Tanimoto coefficient* had yet to be optimized. This was performed on 7 MOFs used in the ***small set*** and 298 MOFs used in the ***large set***. Once the probability distributions were generated, the by hand binding site locations could be found through visual inspection of the probability plots. This process begins by inspection of the noisy probability plot shown in Figure 3.2a. The minimum occupancy value shown in the plot, or the isosurface value, can then be increased to resolve areas of high probability within the MOF. These areas of high probability, shown in Figure 3.2b, correspond to binding site locations. These areas of high probability are then used to identify binding site locations, shown in Figure 3.2c.

3.3.2.4. Optimization Stage1: Probability Plot Spacing

The first step in the optimization of the GALA control parameters was the selection of an appropriate *grid size*. Although smaller *grid sizes* result in higher resolution plots, larger *grid sizes* provide an initial smoothing effect to the probability distribution. This means that the selection of an appropriate *grid size* parameter is associated with several important practical considerations, the first of which is the resource cost to running the GCMC simulation. For example, the MOF NU-110 which boasts a large internal surface area,[35] has unit cell dimensions of a=b=c=48.6 Å. If we assume a *grid size* of 0.1 Å, the probability plot for a single atom type would contain 110 million grid points, exceeding one gigabyte when stored in double precision format. Thus, any value below 0.1 Å may be impractical, particularly when considering heteroatomic guest molecules which require different probability plots for each atom type and centre of mass. As GALA runs GCMC simulations in parallel on two separate cores to check for convergence, the memory allocation would effectively be doubled as each subprocess in the simulation would store these plots in memory.

The second practical consideration is associated with the risk of over-smoothing the probability distributions, removing important peaks associated with unique binding motifs. As such, a *grid size* value which minimizes the memory requirements of the simulation, while simultaneously providing the algorithm with a smoothed probability distribution representative of experimental binding motifs was required.

The key to this optimization was maximizing the *grid size* value to sufficiently smooth the raw probability distribution with no loss or distortion in key binding sites. This target meant that the plot spacing needed to be large enough to reduce the computational cost of the screening and provide an initial smoothing effect, while being small enough to capture accurate binding environments. To tune this parameter and ensure the probability plots were not over smoothed, 4 MOFs were selected as a test set. The MOFs included two MOFs with experimentally determined CO_2 binding sites, CALF-15 and MAF-2, and two hypothetical MOFs, hmof-1 and hmof-2, with binding sites located by hand using an excessive number of GCMC production steps (500,000 cycles) with a grid-spacing of 0.1 Å. These MOFs were selected as they covered a wide range of surface areas and CO_2 uptakes, shown in Table 3.2 along with the number of maxima identified for each MOF as a function of the *grid size*. Importantly, no filtering was performed on the probability distributions to locate the maxima listed in Table 3.2 and were generated using 500,000 GCMC cycles.

The results in Table 3.2 show that as grid size increases, the number of maxima identified decreases, demonstrating the smoothing effect of the *grid size* parameter. The last three rows in Table 3.2 denote the number of expected (or ***reference***) binding sites, defined by the number of binding sites found experimentally or through by hand analysis; the volumetric surface area of the MOF; and the gravimetric CO_2 uptake. When binding sites were calculated using a grid size of 0.175 Å, important sites in CALF-15 were absent due to displacement of the maxima as a result of the larger grid spacing. A value of 0.15 Å was therefore chosen as a reasonable compromise between the memory requirements and the smoothing effect as the probability distributions will be further smoothed using a Gaussian Kernel.

Table 3.2 Number of unsmoothed maxima identified in the probability distribution of CO_2 carbon at flue gas conditions as a function of the grid size.

	MOF			
grid size (Å)	***CALF-15***	***MAF-2***	***hmof-1***	***hmof-2***
0.075	33,731	23,540	68,746	20,441
0.100	10,647	6,588	31,514	5,539
0.125	3,627	2,141	15,950	2,024
0.150	1,491	853	8,850	874
0.175	705	375	5,070	435
0.200	419	209	3,041	243
0.225	285	135	1,984	148
0.250	247	88	1,207	102
0.275	210	80	780	85
0.300	177	66	523	73
0.320	175	58	367	59
Expected	16	20	20	16
Surface Area (m^2/cm^3)	133.65	690.61	1396.15	784.81
Uptake (mmol/g)	2.68	0.63	1.09	3.67

3.3.2.5. Built-in Parameter Optimizer

To facilitate optimization of the remaining control parameters, an optimizer was designed to select appropriate *sigma* and *occupancy cut-off* values based a binding site target provided by the user. Two key relationships between these control parameters and the number of binding sites generated can be inferred from the description of those parameters in Table 3.1. Firstly, the knowledge that an increase in *sigma* will result in more smoothing by the Gaussian kernel means the number of peaks in the probability plot will be reduced. Similarly, increasing the *occupancy cut-off* eliminates lower occupancy peaks from consideration in binding site placement. In both cases, reducing the number of peaks reduces the number of potential binding sites within the MOFs. The optimizer makes use of this relationship to nudge the parameters towards the solution that yields a number of binding sites closest to the target value, without falling below that value. Emphasis was placed on GALA exceeding the number *reference* sites as it was found that the important *reference* binding sites could be present as a subset of the GALA binding sites. Conversely, if GALA returned a lower number of binding sites when compared to the *reference* sites, it would imply that important binding site motifs are not being captured. As such, any combination of *sigma* and *occupancy cut-off* which yielded fewer binding sites than the *reference* sites was rejected.

This optimization, illustrated in Figure 3.7, starts by providing an initial guess of the *sigma* and *occupancy cut-off* values. Once the binding sites are calculated using GALA for the initial point, the number of sites is compared to the target value. If the number of GALA binding sites is equal to the target value, the optimization ends. However, if the target value is greater than the number of binding sites, *the occupancy cut-off* is decreased by the defined interval, denoted by I_c. Conversely if the target value is lower than the number of GALA sites, the *occupancy cut-off* is increased by I_c. Once no improvement can be made by adjusting the *occupancy cut-off*, the *sigma* parameter is then adjusted following the same rules as the *occupancy cut-off* and the process is repeated. The optimization ends once no improvements can be made by adjusting the *sigma* parameter, or an exact match between the GALA and target number of binding sites is achieved. The ranges and intervals for the control parameters are given in Table 3.3.

Table 3.3 The maximum, minimum, and interval values for the parameters being optimized by the built-in GALA optimization scheme.

Parameter	Minimum Value	Maximum Value	Interval (I)	Initial Guess
Sigma (σ)	0.4	4.0	0.05	2.0
Occupancy cut-off (c)	0.025	1.0	0.025	0.1

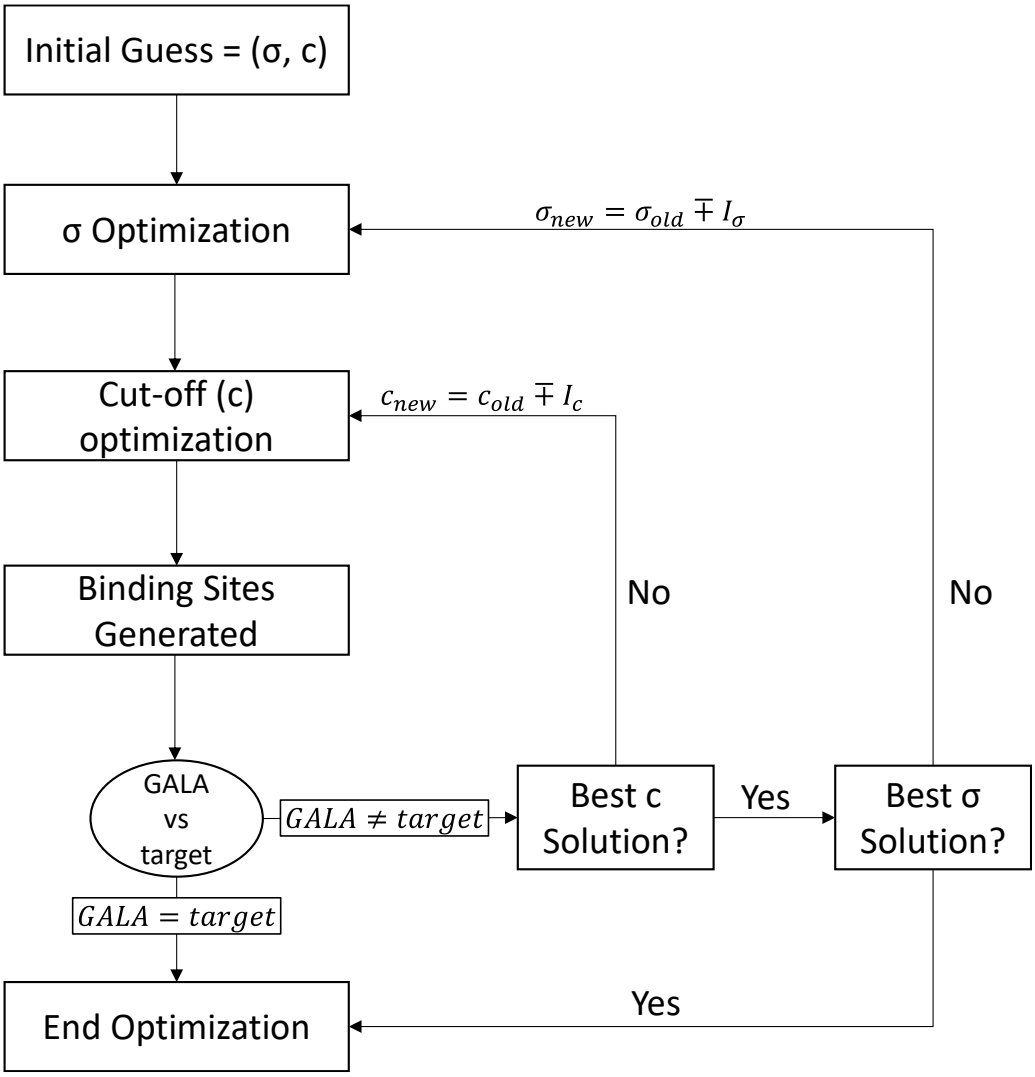


Figure 3.7 Schematic diagram of the built-in sigma and occupancy cut-off parameter optimization scheme which relies on a user defined target value to select the best sigma and occupancy cut-off value for individual MOFs.

3.3.2.6. Optimization Stage 2: Tanimoto Coefficient and Exclusionary Radius

The *Tanimoto coefficient* and *exclusionary radius* were chosen to be optimized together due to their observed impact on the positions of the binding sites generated by GALA. For example, a low *Tanimoto coefficient* could indicate that insufficient sampling of the pore space was performed during the GCMC simulation. Insufficient sampling of the pore space would result in an unconverged probability distribution, meaning the binding sites generated could be inaccurate and unreproducible.

The *exclusionary radius*, which eliminates lower probability maxima in the distribution near a selected peak, controls GALA's ability to identify overlapping but unique binding sites. An example of such a site is shown in Figure 3.8, which demonstrates a binding motif involving two overlapping CO_2 molecules in the MOF CALF-16. While these two binding sites are mutually exclusive, they share similar occupancy values meaning both binding motifs can be found within the MOF at the adsorption conditions studied. Selection of a large *exclusionary radius* would result in loss of one of these important binding sites, however selection of a small *exclusionary radius* could result in the identification of superfluous sites that represent the same binding motif. As such, an *exclusionary radius* needs to be selected such that superfluous sites are eliminated while capturing important unique binding motifs.

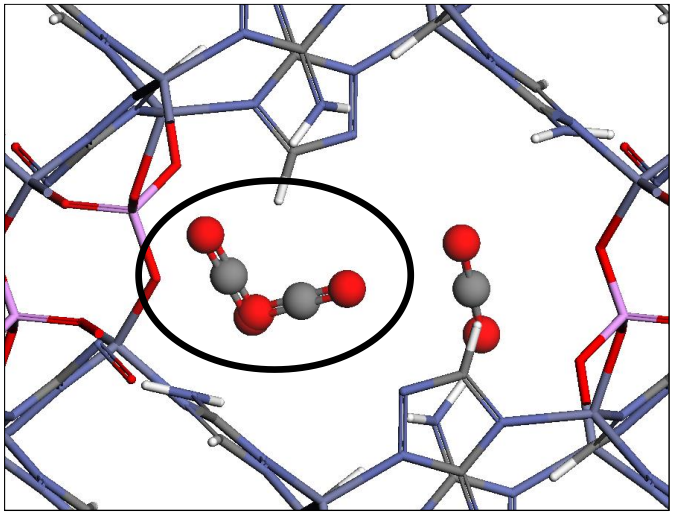


Figure 3.8 CO_2 binding sites calculated in GALA for CALF-16 with an emphasis on the material's competitive binding sites (circled). Binding sites were calculated using the GALA parameters found in Table 3.1.

3.3.2.7. Optimization Stage 2: Parameter Optimization using the Small Training Set

Using a ***small set*** of MOFs composed of 2 experimental and 7 hypothetical MOF structures, the *Tanimoto coefficient* and *exclusionary radius* were optimized. CALF-15 and MAF-2, the two experimental MOFs selected for this set, were chosen due to CALF-15's small pores and very well-defined binding sites, and MAF-2's larger pores with a greater number of binding sites. The two MOFs represent the extremes of the experimentally determined binding sites in the experimental structures. Although the six remaining experimental structures were reserved for final validation of the GALA parameters, the decision was made to include these two experimental MOFs to ensure experimental data was used in the parameter fittings. The performance data and physical properties of the seven hypothetical MOFs is shown in Table 3.4 and demonstrate that a diverse set of MOFs was chosen for this optimization.

Table 3.4 CO_2 performance data and physical properties for the seven hypothetical MOF structures used in the small set during parameter optimization.

	Gravimetric Uptake (mol/g)	Volumetric Uptake (V_{stp}/V)	Isosteric Heat of Adsorption (kJ/mol)	Maximum Pore Size (Å)	Gravimetric Surface Area (m^2 / g)
hmof-1	6.24	159.01	42.93	4.15	696.26
hmof-2	1.09	22.19	24.35	7.52	1558.77
hmof-3	1.96	44.37	30.92	5.74	620.38
hmof-4	4.42	115.86	41.55	5.19	427.65
hmof-5	1.3	23.26	24.02	9.44	2368.33
hmof-6	5.05	97.04	31.76	6.19	1877.95
hmof-7	4.18	87.58	30.25	5.54	1616.16

3.3.2.8. Optimization Stage 2: Fitness Function

Before any parameter optimization could be performed, a metric representing the *fitness* of the located binding sites needed to be established. It was determined that the overall *fitness* of a combination of parameters would be determined using equation 3.5. This equation is the sum of individual fit values for N MOFs in the ***small set***. The *fit* variable in equation 3.5 is determined by the number of binding sites GALA locates relative to the number of *target* sites needed. The number of *target* sites is simply the number of *reference sites* determined experimentally or by-hand sites. When

the number of GALA sites exceeds the number of target sites, the fit is determined by equation 3.6, however if GALA predicts fewer binding sites when compared to the *target* value, a penalty function is applied using equation 3.7. This ensures that any combination of control parameters will heavily favour overprediction over underprediction in the number of binding sites. This is important as it was found that when GALA overpredicts, all *reference* sites are present as a high probability subset of the GALA binding sites, however when GALA generates fewer sites, valuable binding motifs are lost.

$$Fitness = \sqrt{\sum_{i}^{N} Fit_i} \tag{3.5}$$

$$Fit = \left(\frac{GALA - target}{target}\right)^2 \quad \text{when GALA} \geq \text{target} \tag{3.6}$$

$$Fit = 10{,}000 * (target - GALA) \quad \text{when GALA} < \text{target} \tag{3.7}$$

3.3.2.9. Optimization Stage 2: Grid-search of Control Parameters

The optimization of the *Tanimoto coefficient* and *exclusionary radius* were performed using a 2-dimensional grid-search. The probability plots with *Tanimoto coefficients* ranging from 0.325 to 0.9 at intervals of 0.025 were generated for all 9 MOFs in the ***small set***. For each generated plot, GALA was run using a range of *exclusionary radii* from 0.3 to 1.0 Å at intervals of 0.025 Å, creating a 25 x 29 grid. For each point in the grid, the built-in optimizer was used to determine the ideal *sigma* and *occupancy cut-off* for each point, with the saturation uptake used as the target. This optimization allowed selection of the best *Tanimoto coefficient* and *exclusionary radius* independent of the *sigma* and *cut-off* parameters with the *reference* sites used as targets. The *fit* for each MOF was calculated, and an overall *Fitness* was assigned using equations 3.4, 3.5, and 3.6. The results of this grid search are shown in Figure 3.9 and demonstrate that as *Tanimoto coefficient* and *exclusionary radius* are increased, the overall fitness value decreases. Interestingly, the fitness plateaus at a *Tanimoto coefficient* of roughly 0.8 and an exclusionary radius of 0.675 Å. Although an exclusionary radius of 0.675 Å is reasonable, consideration in the selection of a final recommended *Tanimoto coefficient* is more complex. Since the *Tanimoto coefficient* controls the amount of sampling performed during the GCMC simulation and consequently the duration of the GCMC simulation, a concession was made with a recommended *Tanimoto coefficient* of 0.75. This value was chosen to balance the accuracy of the simulation with the computational cost, as a *Tanimoto* of 0.8 was in practice time consuming to achieve, whereas a value of

0.75 could be obtained more rapidly with a reasonable degree of accuracy. This relaxation in *Tanimoto coefficient* requirement reduced the simulation length by and average of 27% across all MOFs in the ***small set***. A final visual inspection was performed which confirmed that the *reference* binding sites identified for the MOFs were present as the highest occupancy subset of the binding sites in the GALA output.

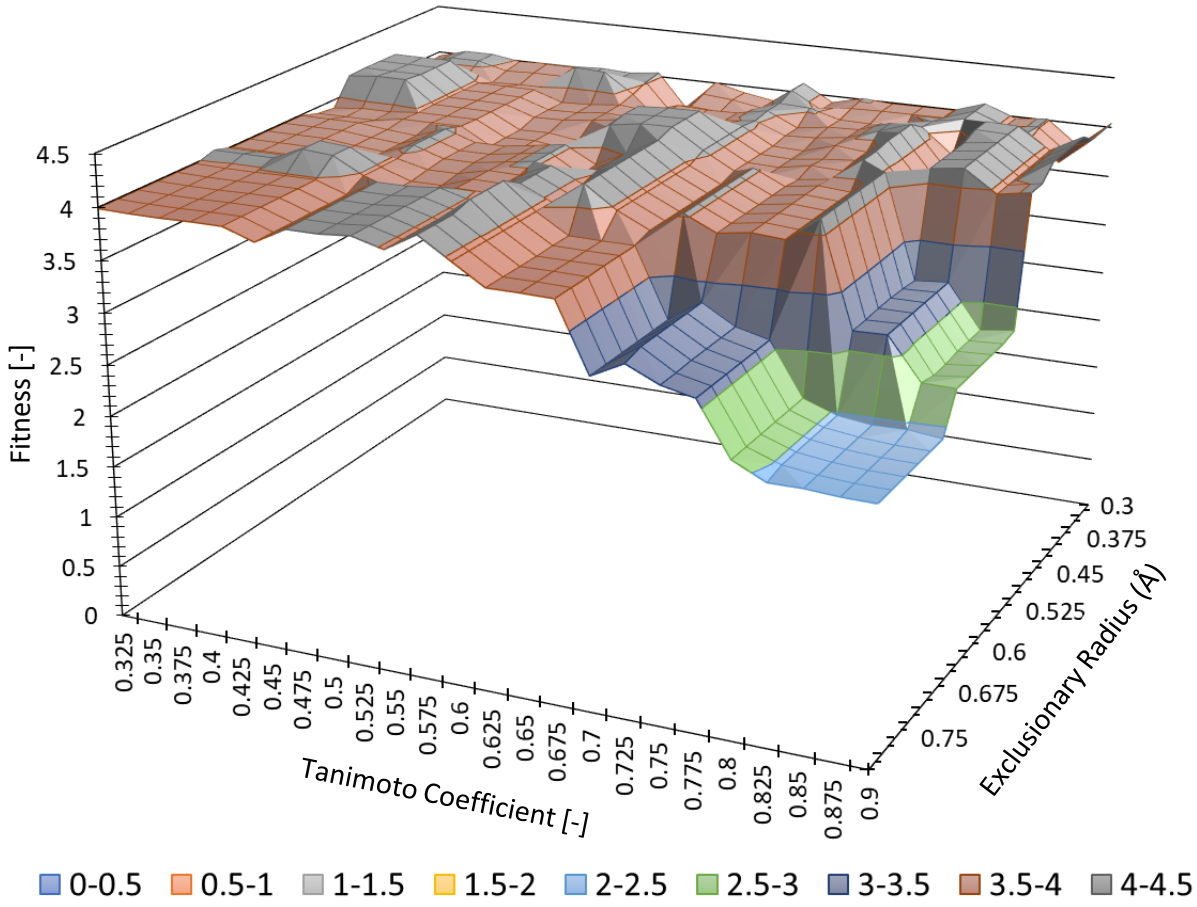


Figure 3.9 3-Dimensional surface plot of the fitness values for every point in the 2-dimensional grid-search of the *Tanimoto* and the *exclusion radius.*

3.3.2.10. Optimization Stage 3: Parameter Optimization using the Large Training Set

The final stage in the optimization of GALA's control parameters for use in high throughput screening was to optimize the final parameters: *sigma* and the *occupancy cut-off*. These parameters were chosen as the last to be optimized since they had little impact on the position of the binding sites, and simply controlled the final number of sites generated by GALA. As such, these parameters are used to eliminate superfluous low probability binding sites.

3.3.2.11. Optimization Stage 3: Large Set

Before any tuning could be performed, a ***large set*** consisting of 298 hypothetical MOFs was assembled and CO_2 probability plots were then generated for all MOFs with a *Tanimoto coefficient* of 0.75, consistent with the results from the ***small set***. Using the number of *reference* sites in the MOFs found through by-hand analysis, the best *sigma* and *occupancy cut-off* values were located using the built-in optimizer. This optimization was automated and performed on all MOFs in the ***large set*** and the "best" *sigma* and *occupancy cut-off* values were extracted and tabulated. Histograms of the best parameters can be found in Figure 3.10a and b for the *sigma* and *occupancy cut-off*, respectively. Based on the results of the MOFs in the large set, 61.4% optimized to a *sigma* of 2.0 Å and 36.6 % *occupancy cut-off* of 0.1, with the next most common frequencies being 2.3 % and 9.4 % for sigma and occupancy, respectively. There is notably more variance in the *occupancy cut-off* distribution when compared to the *sigma* distribution, however most of this variance exists in the region above the 0.1 initial guess, with only 16.8% of MOFs optimizing to a value below 0.1. This result has two implications: 1) that the occupancy cut-off can be used to fine tune the calculation when needed, and 2) the desired binding sites were present as a high occupancy subset of the sites within the MOFs exceeding this value. Since 61.4 % of the MOFs in the ***large set*** optimized to the same *sigma* value, while only 16.8 % of MOFs would require human intervention by means of modification of the *occupancy cut-off* to locate the *reference* sites, the final recommended values for *sigma* and the *occupancy cut-off* were therefore determined to be 2.0 Å and 0.1 (i.e. 10% of the maximum occupancy), respectively, with the *occupancy cut-off* to be used as a single parameter to "fine-tune" the GALA output when needed.

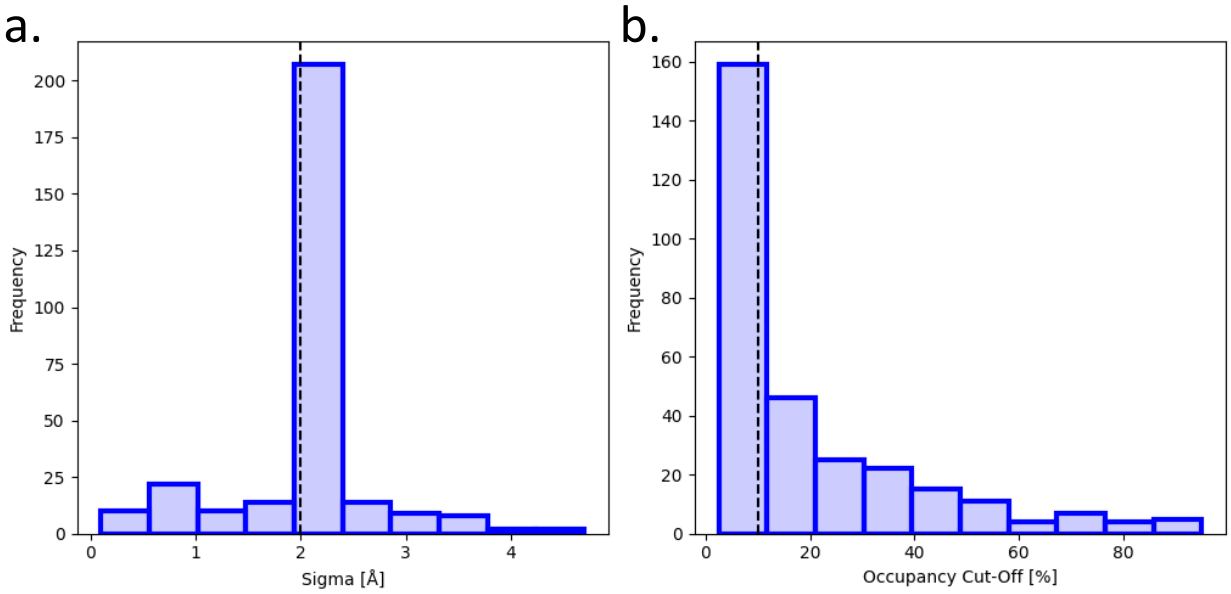


Figure 3.10 Histograms of the optimized parameter for the Large Set of 298 hypothetical MOFs for a) sigma, and b) the occupancy cut-off. For both plots, the most common value is denoted by a dashed line with a sigma value of 2.0 Å and occupancy cut-off of 10%.

3.3.3. Binding Site Accuracy

3.3.3.1. Comparison to Experimentally Determined Binding Sites

Once the parameters of GALA were optimized for high throughput screening, validation against the experimentally realized CO_2 binding sites was performed using the optimized parameters in Table 3.1. For this validation step, the MOFs with experimentally determined CO_2 binding sites were considered. All 9 MOFs are tabulated in Table 3.5 along with the conditions used to determine the experimental binding sites, the number of sites determined experimentally, and the number of binding sites calculated using GALA. This table demonstrates that the number of binding sites located using GALA is in excellent agreement with the number sites determined experimentally. Further evidence for the algorithms ability to locate binding sites can be found in Figure 3.11, which shows the number of binding sites identified by GALA vs the number of sites identified through visual inspection, or *reference sites*, for all MOFs in the ***Large Set.*** The GALA binding sites described in Figure 3.11 were found using the optimized GALA parameters, with fine tuning performed using the occupancy cut-off, and are in excellent agreement with the number of *reference sites* identified by hand.

Table 3.5 Table of MOFs comparing the number of CO_2 binding sites determined experimentally compared to those determined by GALA.

MOF	ref.	temp. (K)	pressure (bar)	no. of binding sites	
				experiment	GALA
CALF-15	[17]	173	0.85	16	16
MAF-2	[36]	195	10.13	18	18
MAF-2	[36]	195	20.27	18	18
$[Rh(II)_2(bza)_4(pyz)]_n$	[37]	298	35	3	3
$[Rh(II)_2(O_2CPh)_4(pyz)]_n$	[38]	213	101	4	6
$[Cu_2(bza)_4(pyz)]_n$	[39]	93	0.32	3	3
FeHCOO	[40]	293	1.01	7	8
$[Rh(II)_2(bza)_4(2\text{-epyz})]_n$	[41]	298	64	1	1
PbSDB	[42]	77	N/A*	4	4

* PbSDB was loaded with a known amount of $^{13}CO_2$ and submerged in liquid nitrogen.

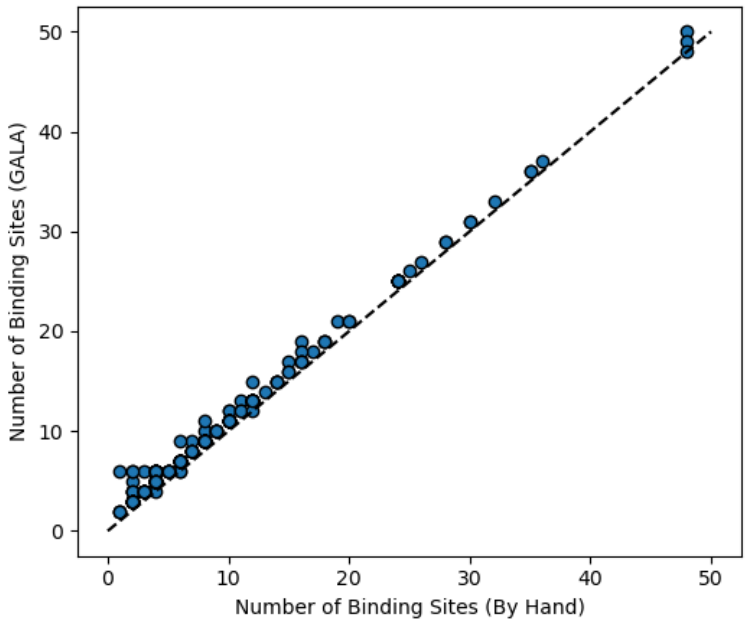


Figure 3.11 Plot of the number of GALA binding sites vs the number of binding sites determine by hand for all MOFs in the Large Set, including a 1:1 line represented by the black dashed line.

Although the number of sites generated by GALA is in good agreement with the number of experimental, or *reference*, sites for each MOF, Table 3.5 provides no confirmation that the binding site positions are in good agreement with experiment. As such, a secondary check was performed through visual comparison of the GALA sites against the *reference* sites. A comparison for all 8 MOFs is shown in Figure 3.12, with GALA sites shown in orange and reference sites in green. Importantly, all tuning of GALA over the course of this work was only performed using 2 out of the 8 MOFs with experimentally characterized CO_2 binding sites, CALF-15 and MAF-2. In other words, the favourable results determined

with the remaining 6 MOFs was not due to the fact that they we used to fine-tune the GALA parameters. For all 8 MOFs, the positions of the binding sites found using GALA are in excellent agreement with the experimental binding site locations, confirming the efficacy of this method at reproducing experimentally determined binding motifs (also suggesting that the force fields used in this work are relatively accurate).

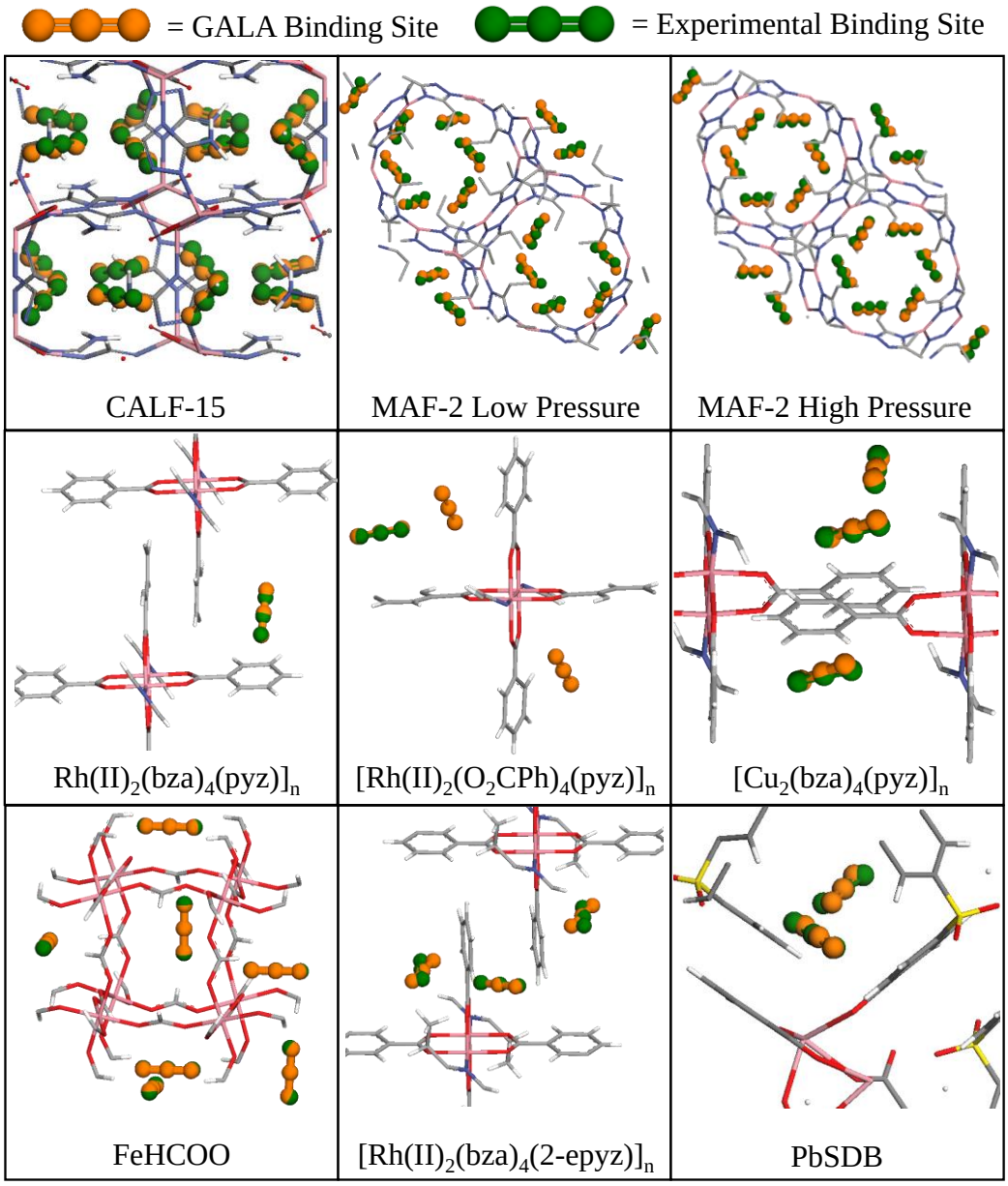


Figure 3.12 Binding site location comparison between experimental binding sites (green) and GALA binding sites (orange) in six of the 8 experimental MOFs. Hydrogen atoms have been omitted from the MAF-2 structures to increase visual clarity. Atoms in this figure: Carbon (gray), Oxygen (red), Nitrogen (blue), Sulfur (yellow), Hydrogen (white), and Metals (pink).

3.3.3.2. Other Guest Molecules

A final test was performed on GALA to determine the efficacy of these parameters on alternate guest molecules. The goal was to determine whether reasonable sites could be located without the need to re-optimize the GALA control parameters for different guests. Using the GALA parameters optimized for CO_2, binding sites for Argon, CH_4, and N_2 were generated using the self-consistent Argon,[43] CH_4-TraPPE,[44] and N_2-NIMF[45] forcefield parameters respectively. The binding sites were then compared to the calculated uptakes for the corresponding guests for each MOF in the ***large set***. The results of this analysis are presented in Figure 3.13a, b, and c for Argon, CH_4, and N_2, respectively, and show that for nearly all MOFs in the ***large set*** the parameters optimized for CO_2 identify more GALA sites than the predicted uptake, with the number of Ar and CH_4 binding sites falling reasonably close to the 1:1 line. Since few points fall below this line for all three guests, the minimum number of required binding sites within the MOF, we demonstrate these parameters can be used for all three guests without re-optimization. When considering the results for the N_2 binding sites shown in Figure 3.13c, it is evident that GALA massively overpredicts the number of binding sites. This is likely due to the very low N_2 uptakes for all MOFs in the set, having no MOF with an N_2 uptake above 2 molecules / unit cell, resulting in weak and poorly defined binding sites.

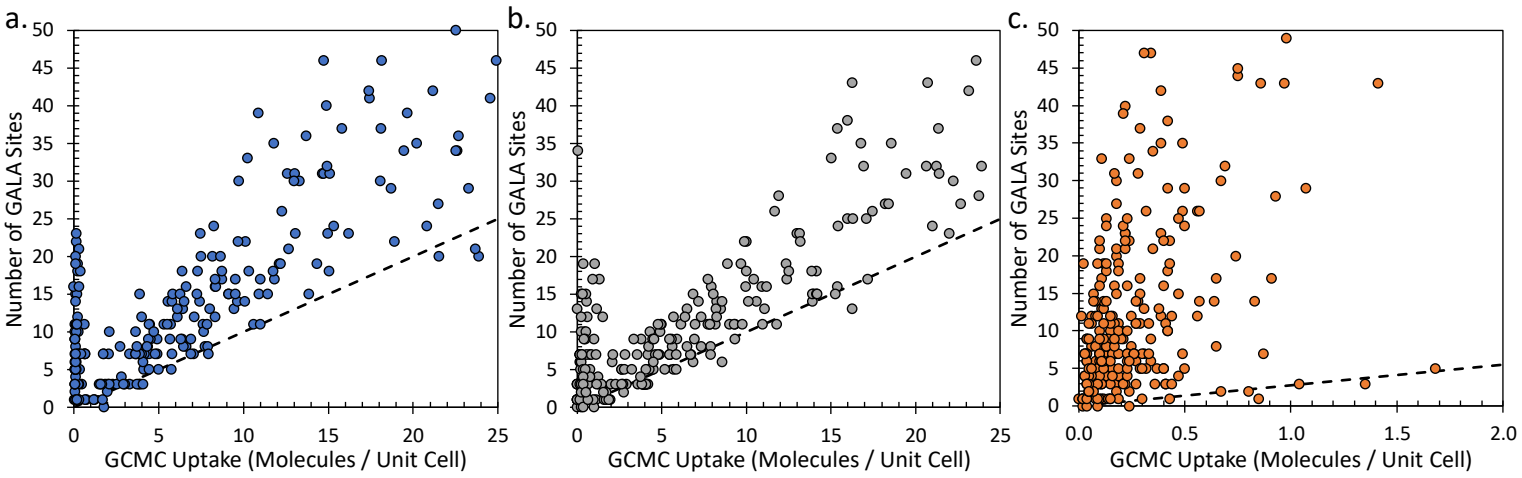


Figure 3.13 Number of GALA binding sites generated plotted against the GCMC uptake for a) Argon, b) Methane, and c) molecular Nitrogen. The dashed line represents the 1:1 line between the x and y axes.

3.4. Conclusions

An accurate and robust method has been developed for locating the guest binding sites from the probability distributions generated from GCMC simulations of the gas adsorption process in MOFs. It was shown to accurately reproduce experimental results as well as reproduce the binding sites determined manually from a range of hypothetical MOFs. The input parameters have been optimized for carbon dioxide, due to its importance in the field of carbon capture and allows for the use of high-throughput screening of a large database of materials with little to no user intervention. In rare cases when the user would need to fine tune the binding site localization, a single parameter, the *occupancy cut-off*, can be adjusted to increase or decrease the number of generated binding sites. GALA has been shown to provide accurate CO_2 binding sites, both in quantity and position that are in very good agreement with the binding sites seen in experimental MOFs determined through x-ray diffraction or NMR experiments. It has also been shown to be able to reproduce or exceed the number of binding sites of argon, methane, and N_2.

3.5. Author Contributions

When I joined this project, much of the initial framework and methodology was already in place to run the GALA code, including the equitable binning, gaussian smoothing, and the logic behind the placement of the binding sites within the framework. My initial objective was to optimize the code for high-throughput screening. The addition of the Tanimoto coefficient for convergence, the optimization codes written into the main GALA code, and all parameter optimizations were performed entirely by me. I performed curation of the MOFs used to tune the algorithm, and the identification of binding sites performed by hand. I also made further improvements to the code itself, including optimizations to improve overall performance – specifically the code needed to be re-organized with large amounts being re-written to allow the code to run smoothly on a High-Performance Computer Cluster. The Grand Canonical Monte Carlo simulation code also needed to be modified to allow for the implementation of the Tanimoto convergence metric, with a supporting framework added to the GALA code itself.

3.6. References

1. Eddaoudi, M., Moler, D. B., Li, H., Chen, B., Reineke, T. M., O'Keeffe, M. & Yaghi, O. M. Modular chemistry: Secondary building units as a basis for the design of highly porous and robust metal-organic carboxylate frameworks. *Accounts of Chemical Research* **34**, 319–330 (2001).

2. Rowsell, J. L. C. & Yaghi, O. M. Metal–organic frameworks: a new class of porous materials. *Microporous and Mesoporous Materials* **73**, 3–14 (2004).

3. Wang, B., Côté, A. P., Furukawa, H., O'Keeffe, M. & Yaghi, O. M. Colossal cages in zeolitic imidazolate frameworks as selective carbon dioxide reservoirs. *Nature* **453**, 207–11 (2008).

4. Thallapally, P. K., Tian, J., Kishan, M. R., Fernandez, C. A., Dalgarno, S. J., McGrail, P. B., Warren, J. E. & Atwood, J. L. Flexible (breathing) interpenetrated metal-organic frameworks for CO_2 separation applications. *Journal of the American Chemical Society* **130**, 16842–16843 (2008).

5. Millward, A. R. & Yaghi, O. M. Metal Organic Frameworks with Exceptionally High Capacity for Storage of Carbon Dioxide at Room Temperature. *Journal of the American Chemical Society* **127**, 17998–17999 (2005).

6. Caskey, S. R., Wong-Foy, A. G. & Matzger, A. J. Dramatic tuning of carbon dioxide uptake via metal substitution in a coordination polymer with cylindrical pores. *Journal of the American Chemical Society* **130**, 10870–10871 (2008).

7. Phan, A., Doonan, C. J., Uribe-Romo, F. J., Knobler, C. B., O'Keeffe, M. & Yaghi, O. M. Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks. *Accounts of chemical research* **43**, 58–67 (2010).

8. Walton, K. S., Millward, A. R., Dubbeldam, D., Frost, H., Low, J. J., Yaghi, O. M. & Snurr, R. Q. Understanding inflections and steps in carbon dioxide adsorption isotherms in metal-organic frameworks. *Journal of the American Chemical Society* **130**, 406–407 (2008).

9. Demessence, A., D'Alessandro, D. M., Foo, M. L. & Long, J. R. Strong CO_2 binding in a water-stable, triazolate-bridged metal-organic framework functionalized with ethylenediamine. *Journal of the American Chemical Society* **131**, 8784–8786 (2009).

10. An, J., Geib, S. J. & Rosi, N. L. High and selective CO_2 uptake in a cobalt adeninate metal-organic framework exhibiting pyrimidine- and amino-decorated pores. *Journal of the American Chemical Society* **132**, 38–39 (2010).

11. Li, J. R., Ma, Y., McCarthy, M. C., Sculley, J., Yu, J., Jeong, H. K., Balbuena, P. B. & Zhou, H. C. Carbon dioxide capture-related gas adsorption and separation in metal-organic frameworks. *Coordination Chemistry Reviews* vol. 255 1791–1823 (2011).

12. Sumida, K., Rogow, D. L., Mason, J. A., McDonald, T. M., Bloch, E. D., Herm, Z. R., Bae, T. H. & Long, J. R. Carbon dioxide capture in metal-organic frameworks. *Chemical Reviews* **112**, 724–781 (2012).

13. Chae, H. K., Siberio-Pérez, D. Y., Kim, J., Go, Y., Eddaoudi, M., Matzger, A. J., O'Keeffe, M. & Yaghi, O. M. A route to high surface area, porosity and inclusion of large molecules in crystals. *Nature* **427**, 523–527 (2004).

14. Furukawa, H., Ko, N., Go, Y. B., Aratani, N., Choi, S. B., Choi, E., Yazaydin, a O., Snurr, R. Q., O'Keeffe, M., Kim, J. & Yaghi, O. M. Ultrahigh porosity in metal-organic frameworks. *Science (New York, N.Y.)* **329**, 424–8 (2010).

15. Ho, M. T., Allinson, G. W. & Wiley, D. E. Reducing the cost of CO_2 capture from flue gases using pressure swing adsorption. *Industrial & Engineering Chemistry Research* **47**, 4883–4890 (2008).

16. Mulgundmath, V. & Tezel, F. H. Optimisation of carbon dioxide recovery from flue gas in a TPSA system. *Adsorption* **16**, 587–598 (2010).

17. Vaidhyanathan, R., Iremonger, S. S., Shimizu, G. K. H., Boyd, P. G., Alavi, S. & Woo, T. K. Direct observation and quantification of CO_2 binding within an amine-functionalized nanoporous solid. *Science (New York, N.Y.)* **330**, 650–3 (2010).

18. Düren, T., Bae, Y.-S. & Snurr, R. Q. Using molecular simulation to characterise metal-organic frameworks for adsorption applications. *Chemical Society reviews* **38**, 1237–1247 (2009).

19. Getman, R. B., Bae, Y., Wilmer, C. E. & Snurr, R. Q. Review and analysis of molecular simulations of methane, hydrogen, and acetylene storage in metal-organic frameworks. *Chemical Reviews* **112**, 703–23 (2012).

20. Smit, B. & Maesen, T. L. M. Molecular simulations of zeolites: Adsorption, diffusion, and shape selectivity. *Chemical Reviews* **108**, 4125–4184 (2008).

21. Vaidhyanathan, R., Iremonger, S. S., Shimizu, G. K. H., Boyd, P. G., Alavi, S. & Woo, T. K. Competition and cooperativity in carbon dioxide sorption by amine-functionalized metal-organic frameworks. *Angewandte Chemie - International Edition* **51**, 1826–1829 (2012).

22. Wilmer, C. E., Leaf, M., Lee, C. Y., Farha, O. K., Hauser, B. G., Hupp, J. T. & Snurr, R. Q. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry* **4**, 83–89 (2012).

23. Lin, L.-C., Berger, A. H., Martin, R. L., Kim, J., Swisher, J. a., Jariwala, K., Rycroft, C. H., Bhown, A. S., Deem, M. W., Haranczyk, M. & Smit, B. In silico screening of carbon-capture materials. *Nature Materials* **11**, 633–641 (2012).

24. Wilmer, C. E., Farha, O. K., Bae, Y.-S., Hupp, J. T. & Snurr, R. Q. Structure–property relationships of porous materials for carbon dioxide separation and capture. *Energy & Environmental Science* **5**, 9849 (2012).

25. Iremonger, S. S., Liang, J., Vaidhyanathan, R., Martens, I., Shimizu, G. K. H., Daff, T. D., Aghaji, M. Z., Yeganegi, S. & Woo, T. K. Phosphonate monoesters as carboxylate-like linkers for metal organic frameworks. *Journal of the American Chemical Society* **133**, 20048–51 (2011).

26. Brehm, M. & Kirchner, B. TRAVIS - A free analyzer and visualizer for Monte Carlo and molecular dynamics trajectories. *Journal of Chemical Information and Modeling* **51**, 2007–2023 (2011).

27. Smith, W., Todorov, I. T. & Leslie, M. *The DL_POLY molecular dynamics package*. *Zeitschrift für Kristallographie* vol. 220 (2005).

28. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

29. Boyd, P. G. Computational High Throughput Screening of Metal Organic Frameworks for Carbon Dioxide Capture and Storage Applications. (2015).

30. Coupry, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal-organic frameworks. *Journal of Chemical Theory and Computation* (2016) doi:10.1021/acs.jctc.6b00664.

31. Krykunov, M., Demone, C., Lo, J. W. H. & Woo, T. K. A New split charge equilibration model and REPEAT electrostatic potential fitted charges for periodic frameworks with a net charge. *Journal of Chemical Theory and Computation* **13**, (2017).

32. García-Sánchez, A., Ania, C. O., Parra, J. B., Dubbeldam, D., Vlugt, T. J. H., Krishna, R. & Calero, S. Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

33. Peng, D.-Y. & Robinson, D. B. A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals* **15**, 59–64 (1976).

34. Nandi, S., de Luna, P., Daff, T. D., Rother, J., Liu, M., Buchanan, W., Hawari, A. I., Woo, T. K. & Vaidhyanathan, R. A single-ligand ultra-microporous MOF for precombustion CO_2 capture and hydrogen purification. *Science Advances* **1**, e1500421–e1500421 (2015).

35. Farha, O. K., Eryazici, I., Jeong, N. C., Hauser, B. G., Wilmer, C. E., Sarjeant, A. A., Snurr, R. Q., Nguyen, S. T., Yazaydin, A. Ö. & Hupp, J. T. Metal-organic framework materials with ultrahigh surface areas: Is the sky the limit? *Journal of the American Chemical Society* **134**, 15016–15021 (2012).

36. Zhang, J. & Chen, X. Optimized acetylene / carbon dioxide sorption in a dynamic porous crystal. *Journal of the American Chemical Society Articles* 5516–5521 (2009).

37. Takamizawa, S., Nakata, E., Saito, T. & Kojima, K. Structural determination of physisorbed sites for CO_2 and Ar gases inside an organometallic framework. *CrystEngComm* **5**, 411 (2003).

38. Takamizawa, S., Nakata, E., Yokoyama, H., Mochizuki, K. & Mori, W. Carbon dioxide inclusion phases of a transformable 1D coordination polymer host [Rh2(O2CPh)4(pyz)]n. *Angewandte Chemie* **115**, 4467–4470 (2003).

39. Takamizawa, S., Nakata, E. & Saito, T. Structural determination of copper(II) benzoate–pyrazine containing carbon dioxide molecules. *Inorganic Chemistry Communications* **7**, 1–3 (2004).

40. Tian, Y. Q., Zhao, Y. M., Xu, H. J. & Chi, C. Y. CO_2 template synthesis of metal formates with a ReO_3 net. *Inorganic Chemistry* **46**, 1612–1616 (2007).

41. Takamizawa, S., Kojima, K. & Akatsuka, T. Channel-switching crystal with guest stress drive. *Inorganic chemistry* **45**, 4580–2 (2006).

42. Chen, S., Lucier, B. E. G., Boyle, P. D. & Huang, Y. Understanding the fascinating origins of CO_2 adsorption and dynamics in MOFs. *Chemistry of Materials* **28**, 5829–5846 (2016).

43. Boato, G. & Casanova, G. A self-consistent set of molecular parameters for neon, argon, krypton and xenon. *Physica* (1961).

44. Martin, M. G. & Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *Journal of Physical Chemistry B* **102**, 2569–2577 (1998).

45. Provost, B. An Improved N_2 Model for Predicting Gas Adsorption in MOFs and Using Molecular Simulation to Aid in the Interpretation of SSNMR Spectra of MOFs. (2014).

4. Chapter 4: Pores to Process

The work discussed in this chapter was a collaborative effort between the research groups of Dr. Tom Woo at the University of Ottawa and Dr. Arvind Rajendran at the University of Alberta. This work formed the basis of a paper published in 2020: Burns, T. D. et al. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for postcombustion CO_2 Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. Environmental Science and Technology 54, 4536–4544 (2020).[1] For full details of my contributions to this project, see section 4.7 Author Contributions.

4.1. Abstract

Although GCMC simulations provide crucial gas separation information, full process level pressure swing adsorption (PSA) simulations are required to give a more complete view of a material's industrial performance. Thousands of materials were screened for use in post-combustion carbon capture and storage (PoC-CCS) according to their compositions, pore geometrics, and adsorption properties using conditions typical in a coal-fired powerplant with a PSA system. Using sophisticated atomistic and process level simulations, the first large scale screening of metal-organic frameworks (MOFs) from the atomistic to the process engineering scale was performed. Using a custom genetic algorithm, 1,022 MOFs were fully optimized to minimize the energetic cost of running a capture plant and maximize the amount of capture CO_2 per unit volume of sorbent (productivity) while adhering to strict purity and recovery targets defined by the US Department of Energy (Us-DoE). 482 MOFs were identified which could meet the purity-recovery targets (DoE-PRT). Of those materials, 223 fell below the energetic cost target set by the US-DoE for a PSA system. The MOFs were ranked, and the top 10 materials presented, with the best overall material IISERP-MOF-2, exhibiting an exceptionally low energetic cost and favourable productivities. Traditional atomistic level performance metrics were compared to the best energetic costs and productivities, and it was concluded that no single metric was predictive of either key process performance value.

4.2. Introduction

The need for viable solutions to slow the effects of climate change is a key driving force behind gas separation and purification research. The application of these technologies for the filtration of greenhouse gases from various emission gas mixtures has been at the forefront of research interests for

the past few decades,[2–5] with a wide variety of proposed solutions ranging from calcium looping[6] to pressure swing adsorption systems. As roughly 35% of the anthropogenic CO_2 emissions arise from fossil fuel burning powerplants,[7,8] substantial focus has been placed on developing systems to scrub CO_2 out of post-combustion flue gases from these large stationary sources. Furthermore, as the transition to carbon neutral sources of energy production has been slow, coupled with the widespread use of coal-fired power plants in underdeveloped and developing nations, an inexpensive system which allows for the capture and storage of CO_2 will be vital in meeting crucial emission targets and minimizing the impacts of climate change.[7]

Although several pilot and industrial scale post-combustion carbon capture and storage (PoC-CCS) projects exist and currently capture millions of tonnes of CO_2 per year,[9] they utilize solvent-based capture systems. These systems require diverting considerable amounts of energy from the powerplant to recover the capture medium through the heating of large volumes of aqueous solutions. This heating is performed to break the covalent bonds between captured CO_2 and the solvent molecules. This incurs an energetic cost to the powerplant and reduces the electrical output. This energetic cost is known as the parasitic energy which also includes the cost of compressing the captured CO_2 to 150 bar for transportation to a storage site.[10] The parasitic energy is therefore a fundamental performance metric of a PoC-CCS system used to determine the operational cost of CO_2 capture. To date, these solvent-based systems are prohibitively expensive, at an estimated cost of 45 USD/tonne of CO_2 (57 CAD/tonne of CO_2 as of January 2022) captured for a state-of-the-art retrofit capture plant.[11] This cost is in contrast to the current carbon taxes imposed on emitters in Canada at 40 CAD/tonne of CO_2 captured,[11] giving little to no financial incentive for emitters to adopt this technology. The high costs of capture due to the technology's reliance on the formation and breaking of covalent bonds, the heating of an aqueous solution, and the rapid degradation of the capture medium through amine foaming,[12] means that widespread adoption of this technology is unlikely.

Several alternative technologies are being explored to replace the expensive solvent-based capture systems. Solid sorbent-based technologies are of particular interest as they often rely on the physical adsorption of the target gas molecules into the pores of a crystalline material. Instead of breaking covalent bonds to recover the capture medium, these technologies rely on either an increase of temperature, commonly known as temperature swing adsorption (TSA); the decrease in pressure, known as pressure swing adsorption (PSA); or a combination of the two (T/PSA) to remove captured CO_2 from the pores of the material. As a result, T/PSA systems would theoretically have lower parasitic

energies when compared to their solvent-based competitors. Due to this reduction in cost, T/PSA systems are at the forefront of PoC-CCS research.[2–5,13,14]

A major hurdle in the development of these technologies is the selection of an appropriate sorbent material, which would allow the capture unit to selectively remove CO_2 from a flue gas mixture while allowing the N_2 to freely pass through. Since the parasitic energy includes a consideration for the compression of the captured CO_2 gas, to minimize the energetic cost of capture, a high purity of CO_2 in the captured gas would be required. The United States Department of Energy (US-DoE) has set specific targets that PoC-CCS technologies would need to meet to be considered viable for industrial use. These targets include considerations for both the purity of the captured CO_2 and the fraction of CO_2 removed from the flue gas. The US-DoE has determined that PoC-CCS technologies need to capture (or recover) 90% of the CO_2 from with flue gas stream while maintaining a purity of at least 95% CO_2.[15] These targets have become known as the US-DoE's purity-recovery targets (DoE-PRT). Additional targets have been set by the US-DoE with respect to the parasitic energy of T/PSA systems, requiring a 30% reduction of parasitic energy compared to existing solvent-based capture systems in order to justify the cost of developing and scaling the technology.[15] A coal-fired powerplant using a retrofitted solvent-based PoC-CCS system has a parasitic energy of approximately 369 kWh/tonne CO_2 captured,[9] meaning that to meet the US-DoE's parasitic energy target, a sorbent-based capture system would need to have a parasitic energy below 258 kWh/tonne CO_2.

Although parasitic energy is a vital metric in determining the cost of capture in a PoC-CCS system, additional information would be required to estimate the costs of building a PoC-CCS system in order to estimate the capital costs. For solid sorbent-based T/PSA systems, this consideration includes a prediction of the quantity of sorbent needed to continually capture the CO_2 being emitted by the powerplant. This metric is known as the productivity of the material and describes the amount of CO_2 captured per cubic meter of sorbent in a single day. This therefore gives researchers three key performance metrics to consider when vetting materials for this application: 1) it's ability to meet the DoE-PRT, 2) the parasitic energy, and 3) the productivity. Additional considerations are required when designing a carbon capture system, such as the replacement costs and the durability of the sorbent material, however those considerations are beyond the scope of this work. Since financial estimates of the cost of capture would vary drastically depending on the geography, economics, and political environment at the location of each individual plant, estimating a dollar value for the cost of capture is

not practical. As such, the work in this chapter relies on the three key performance metrics discussed here-in to provide a generalized comparison of materials for a PSA process.

Metal-organic frameworks (MOFs) are a class of material often studied for the PoC-CCS application due to their large internal surface areas and high tunability.[16–23] However, a major challenge in the development of materials for the PoC-CCS applications is the disconnect between the research performed at the bench scale to the research performed at the engineering, or process level. At the bench scale, chemists and materials scientists synthesize and test MOFs using a series of equilibrium methods to determine simple metrics. These metrics may include the uptake of CO_2 within the pores of a material at a given set of conditions, the selectivity of adsorption for CO_2 over N_2, and the heats of adsorption of both gases with little consideration on the impact these metrics have on performance at the process level. As a result, a wide range of equilibrium performance metrics are used by researchers with no real consensus on how they relate to the three key performance criteria needed to vet the material.[13,19–22,24,25]

The work needed to calculate the purity, recovery, parasitic energy, and productivity metrics has historically involved bench and pilot scale experiments performed by process engineers. These experiments are typically performed using bench scale columns packed with the structured sorbent material with the purity, recovery, parasitic energy, and productivity measured directly. At the time this work was performed, however, no large collaborative efforts between the MOF scientists and process engineers existed. Experiments at the atomic and process scales were often performed independent of one another, with the process engineers relying on materials and isotherms curated from literature. This gap in the field was acknowledged in the 2018 *Mission Innovation* report: "Accelerating Breakthrough Innovation in Carbon Capture, Utilization, and Storage", which stated the major challenge in the development of sorbent materials was to ""Understand the relationship between material and process integration to produce optimal capture designs for flexible operation – bridging the gap between process engineering and materials science."[26,27]

The work presented in this chapter aims to bridge the gap between these atomistic-scale and process level experiments, performing a large-scale screening of MOFs for PoC-CCS in a PSA system from pores to process. Using the structures of experimentally realized MOFs from the Computationally Ready Experimental (CoRE) MOF database,[28] grand-canonical Monte Carlo (GCMC) simulations were combined with sophisticated process level simulations to screen over 4,000 MOFs for their potential for use in a PoC-CCS PSA system using a 4-stage light-product pressurization cycle (LPP),[29] shown in Figure 2.3. For

every MOF, 7 key operating parameters needed to be optimized which control different aspects of the operation of the PSA column. In this chapter, the methodology and results for a multi-stage screening to study MOFs from pores to process is described, as well as an initial attempt at relating conventional equilibrium metrics to process level performance.

4.3. Methodology

The optimizations required for this study were expensive, often taking up to a full week running points in parallel on a high-performance computing (HPC) cluster. Due to the complexity and computational cost of running the required simulations, the screening outlined in this chapter was performed in phases to prioritize MOFs for optimization. The first phase involved applying a series of filters to the MOF database, removing materials according to their adsorption behaviour, pore geometries, and composition. In the second phase, the remaining 1,632 MOFs underwent a rough grid-search using the sophisticated PSA simulation code described in section 2.3 of this thesis, testing process conditions known to harbour high performing points. Once the grid-searches were complete, the MOFs were ranked according to their best parasitic energy (PE) points and underwent full genetic algorithm optimization of the process variables in order of their rankings.

4.3.1. CoRE Database & Named Materials

For this work, the CoRE MOF database[28] was screened in a high-throughput manner. This database consisted of 4,375 experimentally realized neutral MOF structures, as of 2017. Prior to performing this screening project, the database needed to be vetted. Although the CoRE database boasts its *computational readiness*, in practice many structures within the database were found to be problematic, often missing atoms (mostly hydrogen atoms) or containing disorder in the structures that required cleaning. On top of the 4,375 MOFs in the CoRE database, an additional set of 46 nanoporous materials commonly found in literature that includes MOFs, zeolites, and carbonaceous materials, were added to the set. (The full list of named materials can be found in Appendix 4.1.) For simplicity, all materials tested over the course of this work are referred to as MOFs.

4.3.2. DFT Calculations

Prior to performing any simulations on the CoRE database, single-point DFT calculations were performed on the entire cleaned CoRE database to generate a quantum mechanical electrostatic potential (QM-ESP). This QM-ESP was required to calculate high quality partial atomic charges to be

used in the Grand Canonical Monte Carlo (GCMC) simulations. These DFT calculations were performed using the Vienna Ab-Initio Simulation Package (VASP),[30] using the PBE functional[31,32] and planewave cut-off of 400eV. The decision to use VASP to generate the QM-ESPs was made due to the packages robustness and ease in terms of performing DFT calculations on periodic systems such as MOFs. VASP provides high quality pseudo-potentials that are required to do the calculations for most elements on the periodic table.

4.3.3. Grand-Canonical Monte Carlo Simulations

The next step in screening the CoRE database was running the Grand-Canonical Monte Carlo (GCMC) simulations to generate isotherms for every MOF in the set. All GCMC simulations were performed using an in-house[33] code modelled based on the DL_POLY Classic code.[34] For these simulations, the Lennard-Jones parameters for the framework atoms were taken from the DREIDING forcefield,[35] and in cases where atom types were not available, the Universal Force Field (UFF)[36] was used. Partial atomic charges were fit to the framework atoms using the REPEAT method[37] from the QM-ESP generated by VASP simulations. Single component simulations were run to model the adsorption of CO_2 and N_2 sampling 18 pressure points at 298 Kelvin between 0.01 and 1.20 bar. The Garcia-Sanchez parameters were used to model CO_2 adsorption,[38] and the N2-NIMF (Nitrogen in metal-organic frameworks) parameters, fit in-house, were used to model N_2 adsorption.[39] The N_2 parameters can be found in Appendix 4.2. For all pressure points, fugacities were calculated using the Peng-Robinson Equation of State.[40] For every pressure point, simulations were performed using 30,000 equilibration cycles and 30,000 production cycles. This methodology has been shown to reproduce experimental isotherms.[41–43]

4.3.4. Isotherm Fittings

The screening of the CoRE database required two sets of isotherm fittings: single-component single-site Langmuir isotherms (equation 4.1) and to be used for the initial evaluation of MOFs, and dual-site Langmuir isotherms (equation 4.2) to be used in the Pressure Swing Adsorption simulator. In both equations, the adsorption (Θ) at a given pressure (P) is determine by the fitted parameters defining the saturation uptakes (Q, Q_1, and Q_2), and the strength of the adsorption (given by *b* and *d*).

$$\theta = \frac{QbP}{1 + bP} \tag{4.1}$$

$$\theta = \frac{Q_1 bP}{1 + bP} + \frac{Q_2 dP}{1 + dP} \quad (4.2)$$

The fitting algorithm aimed to minimize the Root Mean Squared Error (RMSE), given by equation 4.3, between the 18 individual pressure points calculated in GCMC and the corresponding values at each pressure from the isotherm function. In equation 4.3, Θ_{fit} corresponds to the adsorption predicted by the function, Θ_{GCMC} is the uptake at the specific pressure point, and N corresponds to the number of pressure points being compared (N=18). This fitting was performed using code written in-house in Python 2.7.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_{fit} - \theta_{GCMC}\right)^2} \quad (4.3)$$

4.3.5. Geometric Calculations

The geometric properties for each MOF needed to be calculated to ensure the pores of the MOFs could accommodate CO_2 and N_2. For example, if a MOF's channels were too narrow to accommodate CO_2, adsorption into the pores and channels of the MOF would be impossible. The calculations of the pore sizes, channel geometries, and crystal densities for each MOF were performed using the Zeo++ package.[44] These calculations were performed using 1.71 Å and 1.80 Å probe sizes corresponding to the kinetic radii of CO_2, and N_2, respectively.

4.3.6. Pressure Swing Adsorption Simulator

A sophisticated PSA simulation code written by collaborators at the University of Alberta was used in this screening.[29] This code simulated a 4-stage Light Particle Pressurization (LPP) cycle, described in section 2.3.1 of this thesis. This cycle was determined to be the most efficient to separate CO_2 from N_2 at post-combustion carbon capture conditions.[29]

This simulation code, which captures the most important physics and dynamics of the separation process, has been shown to reproduce experimental results for both transient and steady-states for a pilot-scale column packed with 80 kg of Zeolite 13-X pellets.[45] In the process simulations, MOF crystals were assumed to be bound using an inert structuring agent to form spherical pellets 1 mm in diameter which reduces the overall density of the MOF by 25% when compared to its pure crystalline form. The column being modelled had a diameter of 0.3 m and a length of 1 m to be consistent with the scale of

the pilot-plant for which the simulations were validated on. The PEs were calculated using well established efficiencies for vacuum pumps, blowers, and compressors. To simulate a material, the PSA simulator required an input of single component dual-site Langmuir parameters for both guests (CO_2 & N_2), the isosteric heats of adsorption for each guest, and the structured density of the MOF. The simulator could then be used to predict the *Purity* of the captured CO_2, the percent of CO_2 removed from the flue gas by the PSA unit (the *Recovery*), the PE of the separation, and the productivity of the material (*Prod*). The PSA simulator was written in MATLAB Version R2016a and compiled using MATLAB Runtime R2016a for use on Compute Canada high-performance computing clusters.[46]

4.3.6.1. Modifications to the Pressure Swing Adsorption Simulator

To prepare the PSA simulation code for use in high-throughput screening, major modifications needed to be made. The original code relied on hardcoded values for all operating conditions, column and bed configurations, and material properties. This means that for every individual material and set of process conditions, without modification the code would need to be edited and recompiled to accommodate these new parameters. Over the course of this study, millions of process points were run, which would not have been possible under these conditions. The code was therefore modified to read a custom input file controlling all variables within the simulator. Additional modifications were made to incorporate the compression term, described in section 4.3.7, to the parasitic energy into the actual PSA code, and the output files were modified to provide a breakdown of the parasitic energy by its components.

4.3.7. Compression Energy Calculations

Once the CO_2 has been removed from the flue gas stream, it needs to be compressed to 150 bar for transport to the storage site.[10] The amount of energy required for this compression step is non-trivial and needs to be considered when calculating parasitic energy. Although the PSA simulator provides a prediction of parasitic energy, prior to modifications describe in section 4.3.7, this value only considered the energy required by the separation unit and did not consider the energy required to compress the captured gas to transport conditions.

Since the transport pressure of CO_2 exceeds the conditions at which CO_2 condenses to a supercritical fluid (80 bar at 313.15 Kelvin), the compression term of the parasitic energy needed to be split into components: a compressor term for the energy required to increase the pressure of gaseous

CO_2 to 80 bar, and a pump term to complete the pressurization after the CO_2 condenses into a supercritical fluid. The overall compression term for the parasitic energy is given by equation 4.4.

$$E_{compression} = E_{compressor} + E_{pump} \tag{4.4}$$

The first term in equation 4.4, the contribution of the compressor to the parasitic energy, is calculated using equation 4.5. The value of S in equation 4.5 is defined as the number of compression cycles required to reach the critical pressure of 80 bar from an initial pressure of 1 bar. The number or cycles can be calculated using equation 4.6, where r_T and r_i are the total compression ratio and the maximum compression ratio achievable in a single stage, respectively. A single stage compressor has a maximum compression ratio 3,[47] however a maximum compression ratio of 3 would cause extreme temperature increases of the gas, making further compression more challenging. A value of 2.5 was chosen to avoid these extreme temperature increases[48] and decrease the overall energy required for the compression. The Y term in equation 4.5 is calculated using equation 4.7 and a description of the remaining variables along with the values used in the calculation can be found in table 4.1. The compressor energy is the dominant term in equation 4.4 accounting for over 90% of the total energy required to compress the captured gas to storage conditions when the molar purity exceeds 60% CO_2.

$$E_{Compressor} = S\left[\frac{n_{total}RT}{\eta_{Compressor}}\left(\frac{\gamma}{\gamma-1}\right)\left[Y^{\frac{\gamma-1}{\gamma}}-1\right]\right] \tag{4.5}$$

$$r_i{}^S = r_T = \frac{P_{high}}{P_{low}} \tag{4.6}$$

$$Y = r_T{}^{1/S} \tag{4.7}$$

The second term in equation 4.4 defines the energy required to run a pump to compress the sequestered CO_2/N_2 mixture to the desired final storage pressure of 150 bar once the gas mixture has condensed into a supercritical fluid at around 80 bar. The pump's contribution to the parasitic energy was calculated using equation 4.8.[47]

$$E_{Pump} = \frac{\Delta P\, m_{tot}}{\eta_{pump}\rho} \tag{4.8}$$

Table 4.1 Definitions and values of variables used in equation 4.4, to calculate energy of compression of the captured gas mixture from 1 to 80 bar.

Variable	Definition	Value Used
n_{total}	total number of moles in gas mixture	From Simulation (moles)
R	Gas Constant	8.314 m^3 Pa mol^{-1} K^{-1}
T	temperature of Compression	313.15 K
γ	adiabatic coefficient (C_p/C_V) calculated using NIST[49]	1.4
$\eta_{Compressor}$	efficiency of gas compressor[47]	0.85

The variables m_{tot} and η_{pump} are defined as the total mass of the fluid being compressed (calculated during the simulation) and the efficiency of the pump, respectively. For the pump efficiency a value of 0.75 was used.[50] The ΔP term in equation 4.8 represents the change in pressure resulting from the use of the pump, a value of $7.0x10^6$ Pa was used to compress the supercritical fluid from 80 to 150 bar. The pressures were converted to Pascals so that the resulting energy was in Joules. The density of the supercritical fluid, ρ, was estimated using the average density of the fluid in kg/m^3 over the change in pressure, calculated using a function of based on the CO_2 purity.

4.3.7.1. Density of Supercritical Fluid Mixture

The density of the supercritical fluid (ρ) in equation 4.8, can be calculated as a function of CO_2 purity. To determine the average fluid density over the pressure ranges being tested, a quadratic equation, shown in equation 4.9, was fit to density values obtained from the NIST Standard Reference Database.[49] The values in the equation being used were fit specifically for the defined pressure range at 313.15 Kelvin. The fit is representative of a range of CO_2 molar ratios from 0.6 to 1 was chosen to allow for an accurate prediction of the fluid density at high CO_2 purities. This equation cannot be used to accurately predict the fluid density below 60% CO_2 purity. This limitation was considered acceptable since any point that falls below a CO_2 purity of 60% is greatly below the DoE-PRT and therefore was not considered in this analysis.

$$\rho = \alpha x^2 + \beta x + \delta \tag{4.9}$$

In equation 4.9 the values for α, β, and δ are fitted values and are shown in Table 4.2. The plot of the fitted data with a Pearson R^2 of 0.9998 is shown in Figure 4.1. The values acquired from the NIST database used to fit the plot can be found in Appendix 4.3.

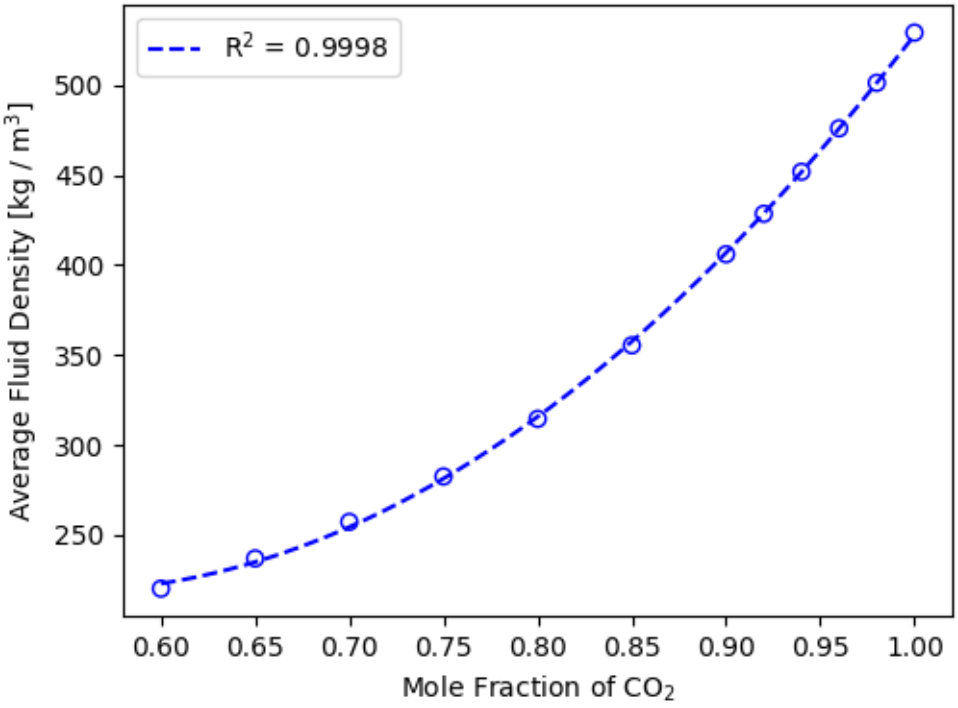


Figure 4.1 Plot of the average density of a supercritical fluid mixture of CO_2 and N_2 as a function of mole fraction of CO_2. The calculated densities are presented as blue circles, and the fitted line based on equation 4.9 is shown as a blue dashed line with a Pearson R^2 of 0.9998.

Table 4.2 Fitted from equation 4.9, fit to reproduce the average fluid density of a CO_2/N_2 supercritical fluid over a range of CO_2 mole fractions.

Parameter	Fitted Value (kg/m³)
α	1491.2
β	-1624.1
δ	660.63

4.3.8. Tiered Screening Approach

The large-scale screening of the CoRE database was predicted to be computationally expensive due to the large computational cost associated with the optimization of a single material. The time predicted to optimize a single material was between 5 and 7 days per MOF using a high-performance computing cluster. Since the CoRE database consisted of over 4,000 MOFs, a tiered approach was adopted to pre-filter and prioritize materials for full optimization.

4.3.8.1. Initial Filters

Before any optimizations were performed on the CoRE database, isotherms were calculated using GCMC and fit to single-site Langmuir isotherms. Those single-site Langmuir parameters were then used to generate *Non-Linearity* plots. Secondary filters were then applied to further reduce the size of the MOF database which removed structures based on their pore geometries and atomic composition. These primary filters reduced the number of MOFs to be screened from the initial 4,375 to 1,632.

4.3.8.2. Pore Size Exclusion

A simple structural filter was then applied to the remaining MOFs. This screening aimed to exclude any MOFs which did not contain pores and channels large enough to accommodate the gasses in the mixture. Any MOF which had a channel diameter below 3.30 Å, the kinetic diameter of CO_2, was removed from the set.

4.3.8.3. Removal of Toxic and Expensive Elements

A second structural filter was applied to the set of MOFs, which aimed to exclude materials that would not be considered for industrial application for practical reasons. Any MOFs which contained expensive, toxic, or radioactive elements were removed from the set, since they would likely not be viable for use at an industrial scale. The remaining atom types in the screening include H, Li, B, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, V, Mn, Fe, Co, Ni, Cu, Zn, Br, Zr, C, Sn, and I.

4.3.8.4. Non-Linearity Plots

Prof. Arvind Rajendran's research group at the University of Alberta, who developed the sophisticated PSA simulator, had found that single-site Langmuir isotherm parameters could provide some insight into a material's best obtainable purity and recovery values. More specifically, it was found

that the non-linearity value, denoted by *b* in equation 4.1, played an important role in determining these purity and recovery points. These values could be compared to a reference material, in this case Zeolite-13X to determine regions of maximum purity-recovery points. This is demonstrated in Figure 4.2, which compares the relative *b* values for CO_2 and N_2. This plot was split into regions, with the green region able to meet the DoE-PRT, the orange region able to meet a more relaxed 90% purity and 90% recovery, and the red region failing to meet either those targets. For this work, all MOFs in the CoRE database were plotted in this *Non-Linearity Plot*, and any materials which did not fall within the 95/90 or 90/90 purity/recovery regions were removed.

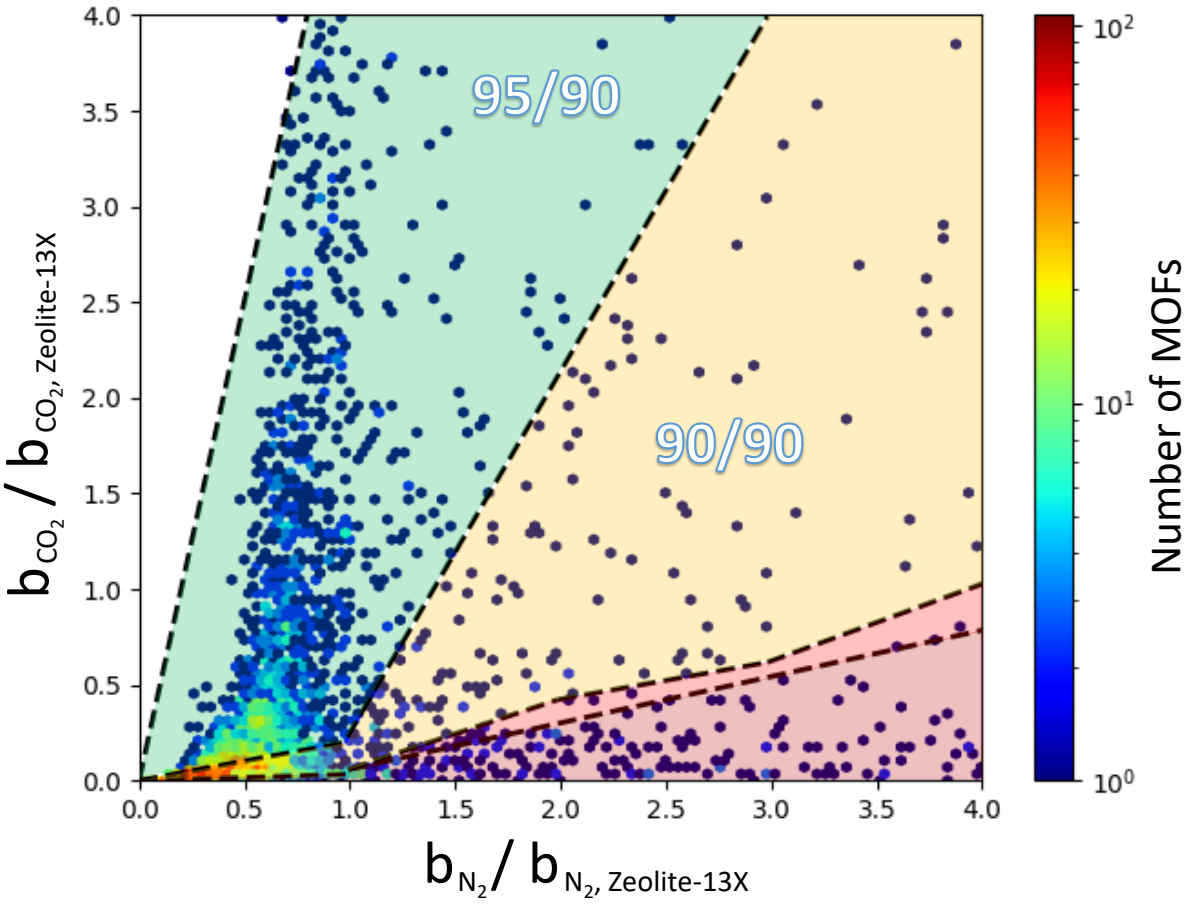


Figure 4.2 Non-linearity plot denoting the non-linearity values, b, for all MOFs relative to Zeolite-13X. All MOFs in the set are plotted as a heatmap and divided into regions denoting the best purity/recovery obtainable for those materials. The green region corresponds to MOFs which are able to achieve a purity of 95% with a recovery of 90% with same point. The orange region corresponds to MOFs which are able to meet a more relaxed 90% purity and 90% recovery, whereas the regions in red indicate materials which are unable to meet those targets.

4.3.8.5. Initial Grid-Search Screening

After the initial set of filters had been applied, a secondary step in the pre-screening was performed. The aim of this step was to rapidly approximate the process performance of the remaining 1,632 MOFs by performing a rough grid-search of the process variables. For this analysis, 3 out of the 7 process variables were modified to sample regions known to harbour high performing points. The ranges and intervals of the grid are shown in Table 4.3 along with descriptions of all 7 process variables, along with the fixed values selected for the remaining process variables. Since the points identified in the grid-search would likely not be the *best* performance points, a relaxed set of criteria was applied to the data. Instead of adhering to the rigid DoE-PRT of 95% purity and 90% recovery, the resulting data was filtered according to a more relaxed criteria of 90% purity and 85% recovery. All process points which fell below this threshold were discarded, and the remaining points were sorted according to their parasitic energies. This allowed for an approximate ranking of the MOFs by parasitic energy and served as a prioritized list of candidates for full optimization.

Table 4.3 List of all process variables controlling the pressure swing adsorption simulation, including ranges and intervals tested, and the constant values used in the grid-search.

Process Variable	Description	Minimum Value	Maximum Value	Interval
Adsorption Time (t_{ads}), seconds	Time spent in the adsorption phase	30	120	10
Blowdown Time (t_{blow}), seconds	Time spent in the blowdown phase	30		Fixed
Evacuation Time (t_{evac}), seconds	Time spent in the evacuation phase	t_{ads} + 10		Fixed
Blowdown Pressure (P_{blow}), bar	Lowest pressure achieved during blowdown phase	0.06	0.15	0.01
Evacuation Pressure (P_{evac}), bar	Lowest pressure achieved during evacuation phase	0.03		Fixed
Feed Velocity (v_0), m/s	Velocity of flue gas entering column during adsorption phase	0.1	1.0	0.1
Flue Gas Temperature (T_{in}), Kelvin	Temperature of flue gas entering column during adsorption phase	298.15		Fixed

4.3.8.6. Genetic Algorithm Optimization

Using the prioritized ranking generated from the grid-search, the MOFs which met the relaxed DoE-PRT targets of 90% purity and 85% recovery were optimized in order of their parasitic energies, from lowest to highest. This allowed for the prioritization of MOFs with the potential to be high-performing materials to be tested, and therefore improve the efficiency of the screening. The MOFs were optimized using a custom genetic algorithm (GA) written in Python. The goal of the GA was to find the 7 process variables that minimized parasitic energy (or maximized productivity), while also meeting the DoE-PRT for each MOF. A population size of 100 individuals was chosen, and a two-phase optimization scheme (described below) was devised to minimize parasitic energy and maximize productivity while ensuring the DoE-PRT. The DOE-PRT constraint of 95% purity and 90% recovery was maintained through the application of a penalty function.

4.3.8.6.1. Defining the Genes

In a genetic algorithm, the *genes* are the variables being optimized. The genes used are the 7 parameters controlling the PSA simulation. All 7 genes used in these optimizations and their allowed ranges are listed in Table 4.4. The number of genes and the ranges for the allowable values were expanded from those presented Table 4.3 to search the variable space more thoroughly for the best possible performance points.

Table 4.4 List of all process variables controlling the pressure swing adsorption simulation used as genes in the genetic algorithm, with their minimum and maximum allowable values.

Process Variable	Minimum Value	Maximum Value
Adsorption Time (t_{ads}), seconds	20	200
Blowdown Time (t_{blow}), seconds	20	50
Evacuation Time (t_{evac}), seconds	20	200
Blowdown Pressure (P_{blow}), bar	0.03	0.15
Evacuation Pressure (P_{evac}), bar	0.01	0.06
Feed Velocity (v_0), m/s	0.1	2.0
Flue Gas Temperature (T_{in}), Kelvin	293.15	328.15

4.3.8.6.2. Two Phase Optimization

After testing the genetic algorithm on several materials, it became apparent that an optimization designed to minimize parasitic energy while maximizing productivity operating under rigid constraints to the purity and recovery would not yield high-performance points. Most MOFs struggled to find any points that could meet the DoE-PRT. To remedy this a two-phase approach was implemented, with the first phase aimed at optimizing the purity and recovery of the MOF, and the second phase aimed at optimizing the parasitic energy and productivity of the MOF.

4.3.8.6.3. Phase 1: Purity & Recovery

During the first phase of this optimization, the GA attempted to optimize the purity and recovery of the MOF. This was done through the application of the objective function shown in equation 4.10. This objective function took the form of a relative distance function, where P_i was the calculated CO_2 purity at the tested conditions, R_i was the calculated CO_2 recovery at the tested conditions, P_{target} and R_{target} were both set to 100%, and P_{DoE} and R_{DoE} were the targets set by the DoE-PRT of 95% and 90% for purity and recovery, respectively. The value calculated by this objective function is known as the *Fitness*, a value which the GA will attempt to minimize. Once the optimizer found that 20% of a generation could meet the DoE-PRT, the optimization would proceed to phase 2.

$$Fitness = \sqrt{\left(\frac{P_{target} - P_i}{P_{target} - P_{DoE}}\right)^2 + \left(\frac{R_{target} - R_i}{R_{target} - R_{DoE}}\right)^2} \tag{4.10}$$

4.3.8.6.4. Phase 2: Parasitic Energy & Productivity

During the second phase of the optimization, the GA switches from optimizing purity and recovery to optimizing productivity and parasitic energy. Before a new generation could be created by the GA, the fitness of every point sampled in phase 1 was recalculated using equation 4.11. An important difference between this objective function and that defined by equation 4.10 is the presence of the penalty functions P_{err} and R_{err}, defined by equations 4.12 and 4.13, respectively. These functions were added to impose a steep penalty to any points which did not meet the DoE-PRT, and therefore ensured that the GA would heavily favour points that fell within the Purity-Recovery constraints. The variables in equation 4.11 E_i and Pr_i are defined as the parasitic energy (including the compression term) and the

productivity of the given point being tested. The target and minimum values for the energy and productivity are given in Table 4.5. The ***Target*** values provided in Table 4.4 were selected as arbitrary targets beyond the range of known MOF performance values prior to this work, while the ***Minimum*** values for parasitic energy and productivity were chosen to balance the two terms in equation 4.11, ensuring both properties would equally influence the fitness.

$$Fitness = \sqrt{\left(\frac{E_{target} - E_i}{E_{target} - E_{min}}\right)^2 + \left(\frac{Pr_{target} - Pr_i}{Pr_{target} - Pr_{min}}\right)^2 + P_{err} + R_{err}} \quad (4.11)$$

$$P_{Err} = \begin{cases} if\ P_i < P_{DoE}, & 10{,}000 \times (P_{DoE} - P_i) \\ if\ P_i \geq P_{DoE}, & 0 \end{cases} \quad (4.12)$$

$$R_{Err} = \begin{cases} if\ R_i < R_{DoE}, & 10{,}000 \times (R_{DoE} - R_i) \\ if\ R_i \geq R_{DoE}, & 0 \end{cases} \quad (4.13)$$

Table 4.5 Values used in equation 4.11 to balance the fitness function.

Parameter	Minimum (min)	Target	Units
Parasitic Energy	250	130	kWh/tonne CO_2
Productivity	0.5	3.5	mol / m^3 / second

4.3.8.6.5. Initial generation:

At the beginning of any GA optimization, an initial generation of 'individuals' needs to be created. In this case each individual is a set of the 7 process parameters, which need to span the entire range of possible values for the process parameters. Additionally, since some knowledge of regions of high performance for each MOF had already been gathered from the grid-search, the initial generation was seeded using the 10 best points from the grid-search. The remainder of the 90 individuals were chosen using the Latin Hypercube Sampling technique, using code written in Python. This ensured that the ranges for all 7 parameters, listed in Table 4.3, were appropriately sampled and a diverse initial generation was created.

4.3.8.6.6. Elitism rates

Elitism is a technique employed by GA methods to retain the best performing samples from each generation to the next. The custom GA used in this study employed a 10% elitism rate, meaning that the 10 most fit individuals in a generation were carried over to the next, allowing for additional opportunities for them to mate.

4.3.8.6.7. Mating protocols

The mating protocols used by this GA can be divided into two steps: the first is the selection of mating pairs, and the second is the combination of the genes in each mating pair. During the first step of the mating protocol, the fitness value of all the individuals in a generation are combined to create a "roulette wheel", with each individual assigned a score between 0 and 1. This score, calculated using equation 4.14, represents the probability the individual will be chosen to be mated. In this equation, the inverse value of fitness value of individual i is divided by the sum of all inverse fitness values for the N individuals in the generation. This means that the sum of all scores in a generation will always equal to 1.

$$score_i = \frac{\left({}^{1}/_{fitness_i}\right)}{\sum_{j=1}^{N}\left({}^{1}/_{fitness_j}\right)} \tag{4.14}$$

Once 90 mating pairs have been selected, the second step can be performed: combining the genes. The value of each gene is compared and the difference between the two "*parents*" is calculated. A new value for this gene is assigned using equation 4.15, where *X* is the gene, *child* is the new individual, and i and j are the "*parents*", or the two individuals in the mating pair. Importantly, the influence a parent has on the new genes is weighted by the fitness of both parents. This means that in an example where parent i is more fit than parent j, the genes of the child will more closely resemble those of parent i over parent j.

$$X_{child} = X_i + \left(\frac{fitness_i}{fitness_i + fitness_j}\right)(X_i - X_j) \tag{4.15}$$

4.3.8.6.8. Mutations

A key mechanism used in this GA to prevent the optimizer from getting stuck in a local minimum is mutation. This mutation causes random perturbations to the genes during the mating step, altering them by a pre-defined amount. The odds of mutation occurring in an individual was set to 30%, meaning that roughly 1 out of 3 "*children*" in a new generation will undergo a mutation on one of its genes, selected at random. This mutation will randomly add or subtract up to 30% from the full allowable range of values for that gene, described in Table 4.4, however these values will not be allowed to exceed those ranges. The odds and severity rate for mutations was set to 30% based on expert judgment from previous experience with continuous number genetic algorithms.

4.3.8.6.9. Convergence

The optimizations were considered to be converged when the best solution remained unchanged for 10 consecutive generations. If the best solution is not improved upon over those generations, it is assumed that the GA has found a minimum and will not improve beyond that point. This convergence criteria was chosen as it implies that the GA optimization has stagnated and is not likely to improve upon the current best solution.

4.3.8.7. Duplicate Runs

Due to the stochastic nature of this optimization technique, and the fact that the optimizer can get stuck in a local minimum, GA optimizations are typically run multiple times. In this case, a full GA optimization was performed 3 times, each with a different starting generation. Although the first optimization relied on seeding from the grid-search results, in subsequent optimizations all 100 individuals in the starting generation were created using Latin Hypercube Sampling.

4.4. Results and Discussion

4.4.1. Results From Screening

4.4.1.1. MOFs Meeting DoE-PRT

Over the course of this screening, emphasis was placed on locating materials which could meet the DoE-PRT of 95% purity of captured CO_2 and 90% recovery of CO_2 from the flue gas.[15] This initial

target is considered the minimum requirement a material needs to meet to be considered viable for use in an industrial PSA system. Out of the 1,022 MOFs optimized over the course of this work, 482 MOFs were identified as being able to meet the DoE-PRT. Once this target is met, the goal of these optimizations was to locate materials with parasitic energies below the DoE target of 258.0 kWh/tonne CO_2 while maximizing the productivity.[15] Out of the 482 MOFs that met the DoE-PRT, 223 were found that are able to achieve parasitic energies below this target threshold. The key challenge of this multi-objective optimization was the antagonistic relationship between the parasitic energy and the productivity. The result of this relationship was such that as the parasitic energy of the separation decreases, so does the productivity. To get a full picture of a material's performance, all three important points: the best parasitic energy, best productivity, and best overall fitness points, need to be considered. Importantly, only the parasitic energy and productivity values from process points which meet the DoE-PRT were considered in this analysis.

4.4.1.2. Best Parasitic Energy MOFs

Although optimizations relied on an objective function that considered both parasitic energy and productivity, at the time this study was performed no targets were given defining the productivity of a material. This is important since the parasitic energy is closely related to the operating cost of a capture plant, and therefore gives insight into how expensive the capture process will be, whereas the productivity of the material is closely related to the initial capital cost of the capture unit and is used to determine the size and complexity of the capture plant. Since the actual cost of capture calculation is a complex topic and may vary significantly according to plant type, country of operation, and the physical location of the plant, the parasitic energy is considered as an initial indicator of the operational cost of capture. This parasitic energy, which can be defined as the amount of electrical energy diverted by the power plant to run the capture unit, needs to be minimized to reduce the operational cost of capture. The best parasitic energies for all 482 materials which met the DoE-PRT are given in Figure 4.3, plotted against the process point's productivity value. To contextualize these results, Figure 4.3 includes the lowest theoretical parasitic energy achievable through a PSA system of 171 kWh/tonne CO_2, shown by the green line. The assumptions used in calculating the theoretical limit can be found in Appendix 4.4. Also included in Figure 4.3 is the performance of a retrofitted liquid amine capture system of 369 kWh/tonne CO_2 (red line),[51] and the DoE target of 258 kWh/tonne CO_2 (orange line).[15] The process points shown in Figure 4.3 demonstrate that nearly all of the optimized materials outperform the liquid amine capture system with 223 materials that are able to meet the DoE target.

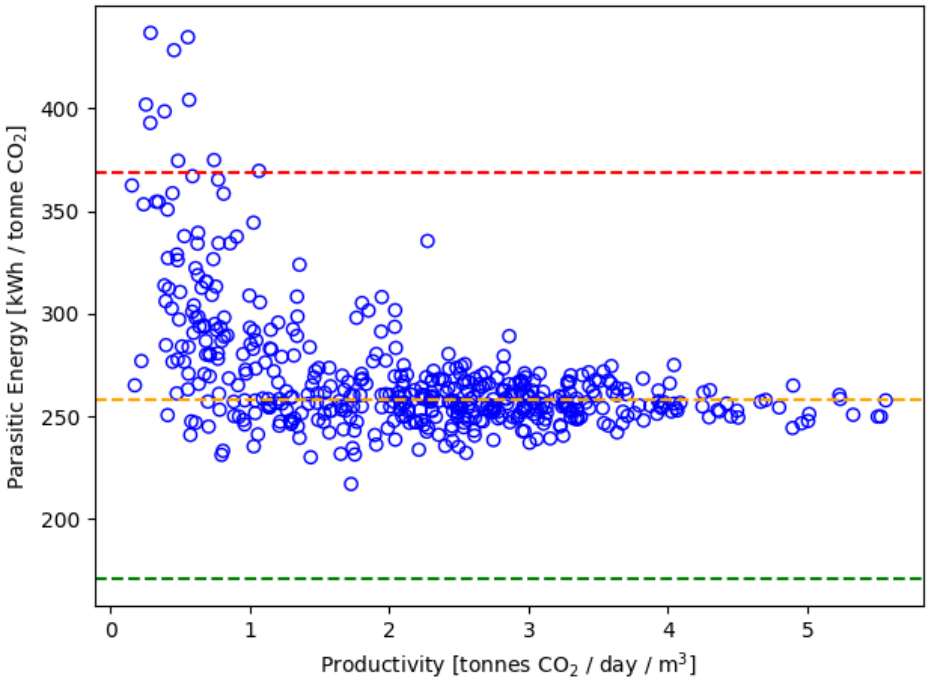


Figure 4.3 The parasitic energy plotted against the productivity for the lowest parasitic energy process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line).

The 482 materials were then ranked according to their parasitic energies, with the 10 best materials presented in Table 4.6, along with their corresponding productivities at that process point. The material with the absolute lowest parasitic energy out of all materials tested was determined to be IISERP-MOF-2, a Ni-based MOF, which had a parasitic energy value of 217.06 kWh / tonne CO_2 captured. Additionally, all 10 materials shown in Table 4.6 improve upon the DoE target by over 8% the target energy's value, with the best material IISERP-MOF-2 falling nearly 16 % below the target value. This result demonstrates that MOFs are able to meet all requirements set out by the US-DoE for sorbent-based post-combustion carbon capture systems. Although these results are promising, another key factor requires consideration: the productivity of the materials.

Table 4.6 Top 10 materials ranked according to the best parasitic energy points found during the optimization. The productivity values shown are the productivities of the materials at the process point that minimizes the parasitic energy. With the exception of IISERP-MOF-2, zif-36-frl, and UTSA-16, the names provided for each MOF are the designations given in the CoRE database.

MOF	*Energy [kWh / tonne CO_2]*	*Productivity [tonnes CO_2 / m^3 / day]*
IISERP-MOF-2	217.06	1.71
LABGAY	230.07	1.44
PURRIE	231.22	0.80
zif-36-frl	231.40	1.75
BENXOJ	231.74	1.64
HAFWUI	232.17	2.55
FEVDIV	233.22	0.80
UTSA-16	233.76	2.21
AMOYOR	234.55	1.75
RAXDAX	235.16	2.51

4.4.1.3. Best Productivity MOFs

Although no specific targets have been set for the productivities of these materials, this property remains vital in determining the capital cost of building a sorbent-based post-combustion capture system. The process points shown in Table 4.6 perform exceedingly well when only considering parasitic energy, however they generally perform poorly when considering their corresponding productivity values, all falling well below the productivity of Zeolite-13x (4.25 tonnes CO_2 / m^3 / day), a well studied sorbent material currently in use in many PSA systems.

The second set of points to consider when vetting these materials is the best productivity points located during the genetic algorithm optimizations. As the relationship between parasitic energy and productivity is antagonistic, we expect an increase in parasitic energy as the productivity increases. This is demonstrated in Figure 4.4, which again shows the parasitic energy plotted against the productivity for all 482 MOFs which meet the DoE-PRT, however this time considering the best productivity process point for each material. When comparing these results to those seen in Figure 4.4, it becomes apparent that maximizing the productivity has resulted in a much smaller set of materials exceeding the DoE parasitic energy target. Instead of the original value seen found using the best parasitic energy points of 223 materials which have parasitic energies below 258 kWh / tonne CO_2, by selecting the best productivity point, this value has fallen to 32 materials.

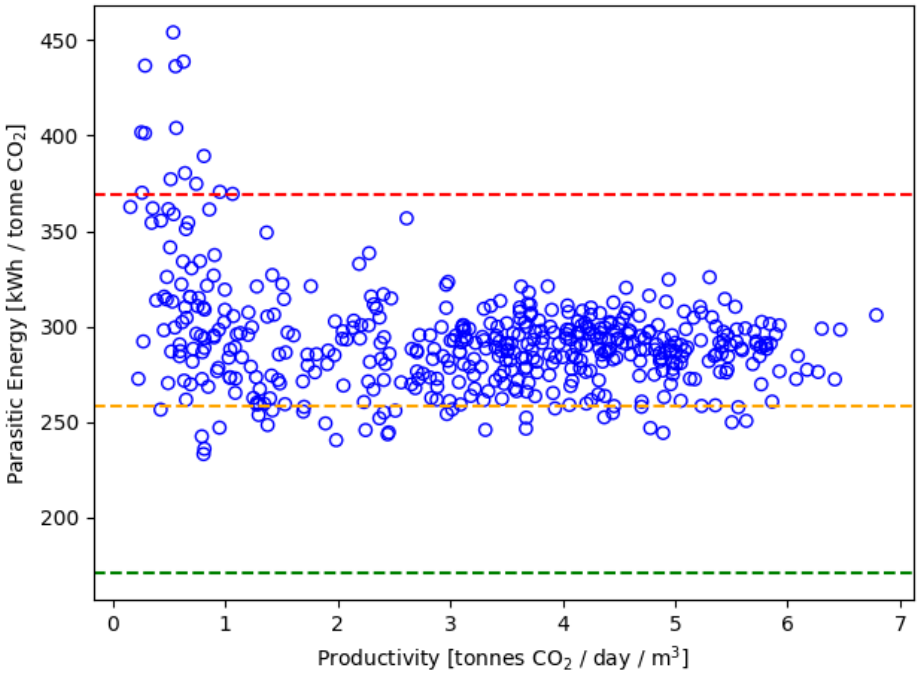


Figure 4.4 The parasitic energy plotted against the productivity for the highest productivity process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line).

Again the 482 materials which met the DoE-PRT were ranked according to their best productivity points with the top 10 shown in Table 4.7. Analysis of the results in Table 4.7 shows that the top 10 points by productivity all have parasitic energies above the DoE target value of 258 kWh / tonne CO_2, and therefore would not be considered viable for use in industrial capture systems. These results further reinforce the need to consider both the parasitic energy and productivity in vetting materials.

4.4.1.4. Best Overall Fitness MOFs

To effectively consider both the parasitic energy and productivity in vetting materials for sorbent-based PoC-CCS systems, the overall fitness of the points can be considered. This fitness is defined by equation 4.11 and balances the parasitic energy with the productivity. The best fitness points for each of the 482 materials is shown in Figure 4.5, again along with the theoretical minimum parasitic energy (green line), the DoE target parasitic energy (orange line), and the parasitic energy of a retrofit solvent-based capture system (red line). The results shown in Figure 4.5 demonstrate that a large number of favourable parasitic energy points can still be found when constraining the productivity, with 194 materials able to meet the DoE target for parasitic energy.

Table 4.7 Top 10 materials ranked according to the best productivity points found during the optimization. With the exception of NaA and UTSA-16, the names provided for each MOF are the designations given in the CoRE database.

MOF	*Energy* *[kWh / tonne CO_2]*	*Productivity* *[tonnes CO_2 / m^3 / day]*
QEJYIP	305.99	6.81
FIJDIM02	298.38	6.46
NaA	272.34	6.43
QISVEU	298.91	6.31
TOKDON	276.24	6.27
FAGREM	277.38	6.16
KANCIN	284.62	6.08
GALCAZ	272.51	6.08
UTSA-16	276.57	5.93
TEDSIG	300.64	5.93

The top 10 materials according to the most fit process point are shown in Table 4.8, including parasitic energy and productivity for those points. Although the parasitic energies of the top performing materials are higher than those seen in Table 4.6, and the productivities are lower than those in Table 4.7, all 10 materials are able to meet the DoE target for parasitic energy while achieving a productivity of over 3.7 tonnes of CO_2 captured / day / m^3 of sorbent, a large improvement over the points in Table 4.5. Interestingly, IISERP-MOF-2 again emerges as the material with the best overall fitness score, with a parasitic energy of 221.36 kWh / tonne CO_2 compared to the "best" parasitic energy point shown in Table 4.5 of 217.06 kWh / tonne CO_2. This process point has only a marginal increase in parasitic energy, while increasing the productivity by a factor of 2.18.

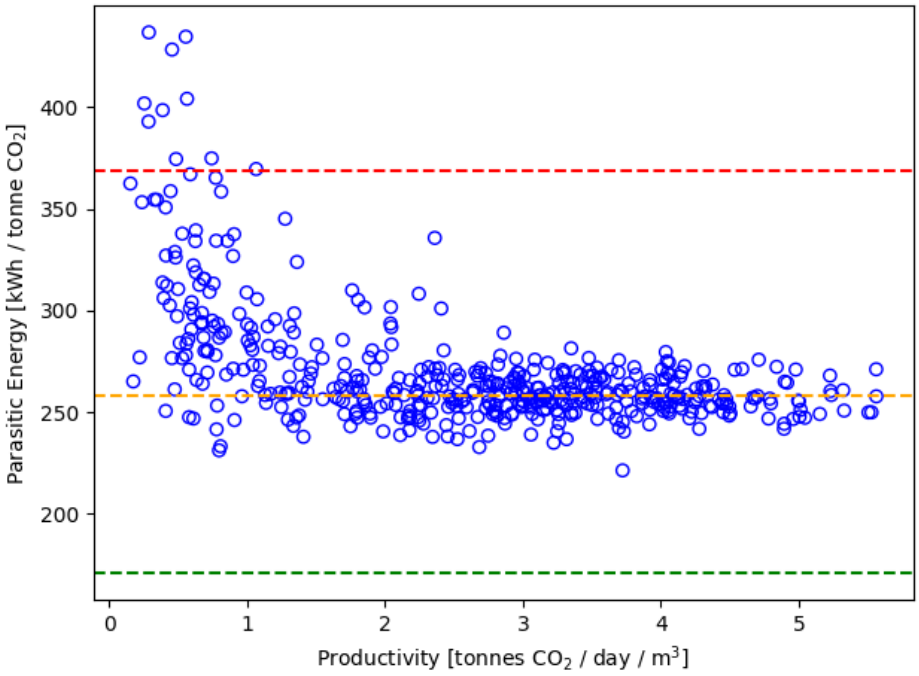


Figure 4.5 The parasitic energy plotted against the productivity for the best fitness process point for the 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line).

4.4.1.5. Pareto Fronts

The overall goal of this screening was to identify materials which were able to meet the DoE-PRT and rank them according to their parasitic energies and productivities. The key challenge encountered over the course of these optimizations was the antagonistic relationship between the parasitic energy and the productivity. This meant that for each material optimization, any improvement to the parasitic energy was met with a penalty to the productivity, and vice-versa. This antagonistic behaviour is demonstrated in Figure 4.6, which shows the pareto front between the parasitic energy and the productivity of six important materials: IISERP-MOF-2 (blue circles), the top performing material in terms of parasitic energy and overall fitness; QEJYIP (green triangles), the top performing material in terms of productivity; Zeolite NaA (red squares) and UTSA-16 (black circles), materials that can be found in all three top 10 lists; Zeolite-13X, a material currently used in industrial PSA systems;[52] and Mg-MOF-74, a material commonly considered to be "high-performing" within the MOF community.[53,54] These pareto fronts represent the performance limits of the materials under the defined Purity-Recovery constraints,

and demonstrate the importance considering both target values in the analysis. The pareto fronts shown in Figure 4.7 indicate that the top performing material in terms of overall fitness and parasitic energy, IISERP-MOF-2, outperforms the other 5 across the entire front, and allows for the selection of a point which appropriately balances the parasitic energy with the productivity. Additionally, it can be noted that the famous MOF Mg-MOF-74 and the best productivity MOF QEJYIP, completely fail to meet the DoE parasitic energy target across the entire pareto fronts. The Zeolites shown in Figure 4.6: Zeolite NaA and Zeolite-13X are both able to meet the DoE target, however Zeolite NaA performs well overall compared to other materials. This finding is significant as Zeolites are relatively cheap with known pathways for large scale production,[55,56] however a key challenge in using Zeolites is the reduction in CO_2 adsorption in wet conditions[56] and therefore additional considerations would need to be made to accommodate the additional costs of drying the flue gas prior to capture.

Table 4.8 Top 10 materials ranked according to the best overall fitness points found during the optimization. With the exception of IISERP-MOF-2, NaA, and UTSA-16, the names provided for each MOF are the designations given in the CoRE database.

MOF	***Energy*** ***[kWh / tonne CO_2]***	***Productivity*** ***[tonnes CO_2 / m^3 / day]***
IISERP-MOF-2	221.36	3.73
GALCAZ	241.83	4.91
IGAHED02	244.31	4.92
GAYFOD	249.90	5.51
WUNSII	249.87	5.51
AXOHIE	247.24	5.06
IMISIH	246.63	4.94
NaA	241.79	4.30
YEZFIU	247.73	5.02
UTSA-16	249.10	5.13

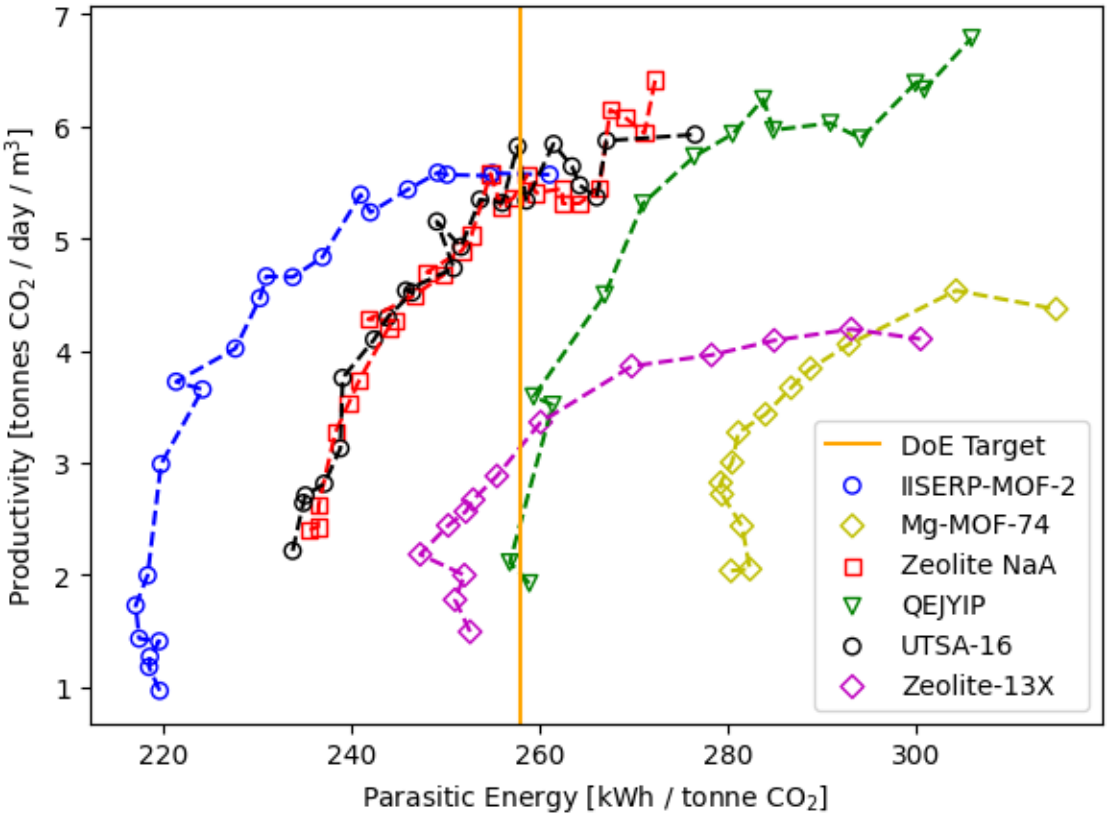


Figure 4.6 Pareto fronts showing the performance boundary between the best productivity and parasitic energy points sampled for IISERP-MOF-2 (blue circles), Mg-MOF-74 (yellow diamonds), Zeolite NaA (red squares), QEJYIP (green triangles), UTSA-16 (black circles), and Zeolite-13x (purple diamonds). The DoE target for parasitic energy of 258 kWh/tonne CO_2 is shown as a solid orange line.

4.4.1.6. Contextualizing Results

Additional context can be given to the performance of these materials, as simply considering parasitic energy and productivity can be abstract and the relationship with process cost may not be immediately apparent. By assuming a 100 MW coal-fired powerplant, one can estimate the number and size of columns used, as well as the energy penalty applied to the plant's electrical output. For this analysis, it is assumed that burning coal produced 120 kJ / mol CO_2 emitted,[57] which is then captured by cylindrical columns 4 meters in height and 1 meter in diameter, the dimensions of conventional industrial PSA columns. According to these assumptions, the calculated CO_2 emission rate from a 100 MW coal-fired powerplant is 3168.72 tonnes CO_2 / day. Based on this value, the parasitic energy can be converted to MW electricity / day, and the number of columns can be calculated using the material's

productivity. The results of this analysis are shown in Figure 4.8a, b, and c for the best parasitic energy, productivity, and overall fitness points, respectively. This figure demonstrates that running a sorbent-based capture system would require diverting anywhere from 28.7% to 60.0% of a powerplant's electricity output to run the capture plant. For context, a state-of-the art retrofit liquid amine system like the one described in section 4.2 would require an approximate energy penalty of 49 % (369 kWh / tonne CO_2)[51] and at an estimated cost of 45 USD/tonne of CO_2 (57 CAD/tonne of CO_2 as of January 2022). [11] Out of the 1,022 materials fully optimized over the course of this work, only 9 are unable to optimize to an energy penalty below that of a retrofitted liquid amine plant, and therefore a PSA system packed with a MOF material has the potential to substantially reduce the cost of capture of CO_2. However, it should be noted that the viability of implementing a plant with these parasitic energies would need to be evaluated on a case-by-case basis, as the cost of the electricity as well as any taxes on emissions would vary significantly according to the location of the powerplant.

Furthermore, the energy penalty discussed above only considers the operational cost of running the capture plant. This technology will require between 192 to over 8000 columns operating simultaneously to capture at least 90% of the CO_2 in the flue gas stream. Although no costing analysis is performed using these estimates, it is apparent that even the smallest capture setup consisting of 192 columns would require an enormous amount of sorbent material, approximately 422 m^3 of sorbent assuming a 25% packing loss in the columns.[58–60] This implies a significant initial capital cost in construction and maintenance of the plants, however this cost cannot easily be estimated as it will vary significantly depending on the sorbent being used. Although many of the process points in Figure 4.7 have energy penalties below the DoE target value, the initial capital cost and the likely ongoing maintenance of the capture plant are the most probable barriers to wide-spread use of this technology in coal-fired powerplants. Further studies need to be performed to determine the viability of using this technology for different applications, as it may be well suited for smaller emitters such as cement and steel production plants.

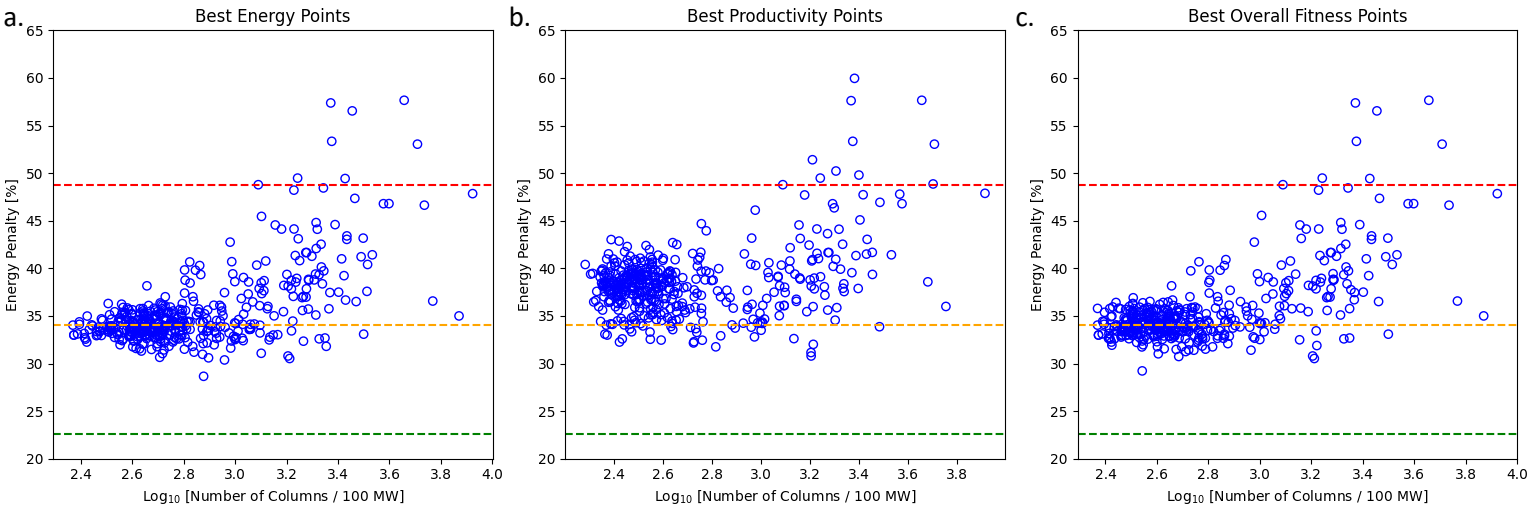


Figure 4.7 Plots of the energy penalty vs the number of columns required to run a continuous capture in a 100 MW coal-fired powerplant for (a) the best parasitic energy process points, (b) the best productivity process points, and (c) the best overall fitness process points for all 482 materials that meet the DoE-PRT. Included in the figure is the theoretical thermodynamic limit for the energy penalty (green line), the DoE target for the energy penalty (orange line), and the energy penalty from a retrofitted liquid amine capture system (red line).

4.4.2. Relationship to Common Metrics

A common challenge in materials discovery at the atomic scale is the lack of consensus towards what constitutes a high performing material. Two of the most prominent metrics used to describe a material for post-combustion CO_2 capture are the selectivity for CO_2 over N_2 and the CO_2 working capacity.[19–23] The results gathered over the course of this screening provided a unique opportunity to evaluate if common metrics used to evaluate MOFs are good predictors of high performance in an industrial PSA system. The CO_2/N_2 selectivity and CO_2 working capacity were compared to the best parasitic energy and best productivity points for the 482 materials that met the DoE-PRT, and the results are shown in Figure 4.8. This figure demonstrates that high performing materials exist along a wide range of selectivities, spanning several orders of magnitude, with no correlation to the parasitic energy or productivity. Furthermore, working capacities, which intuitively one might assume to be related to the material's productivity, shows no apparent correlation between the metric and the materials' PSA performance.

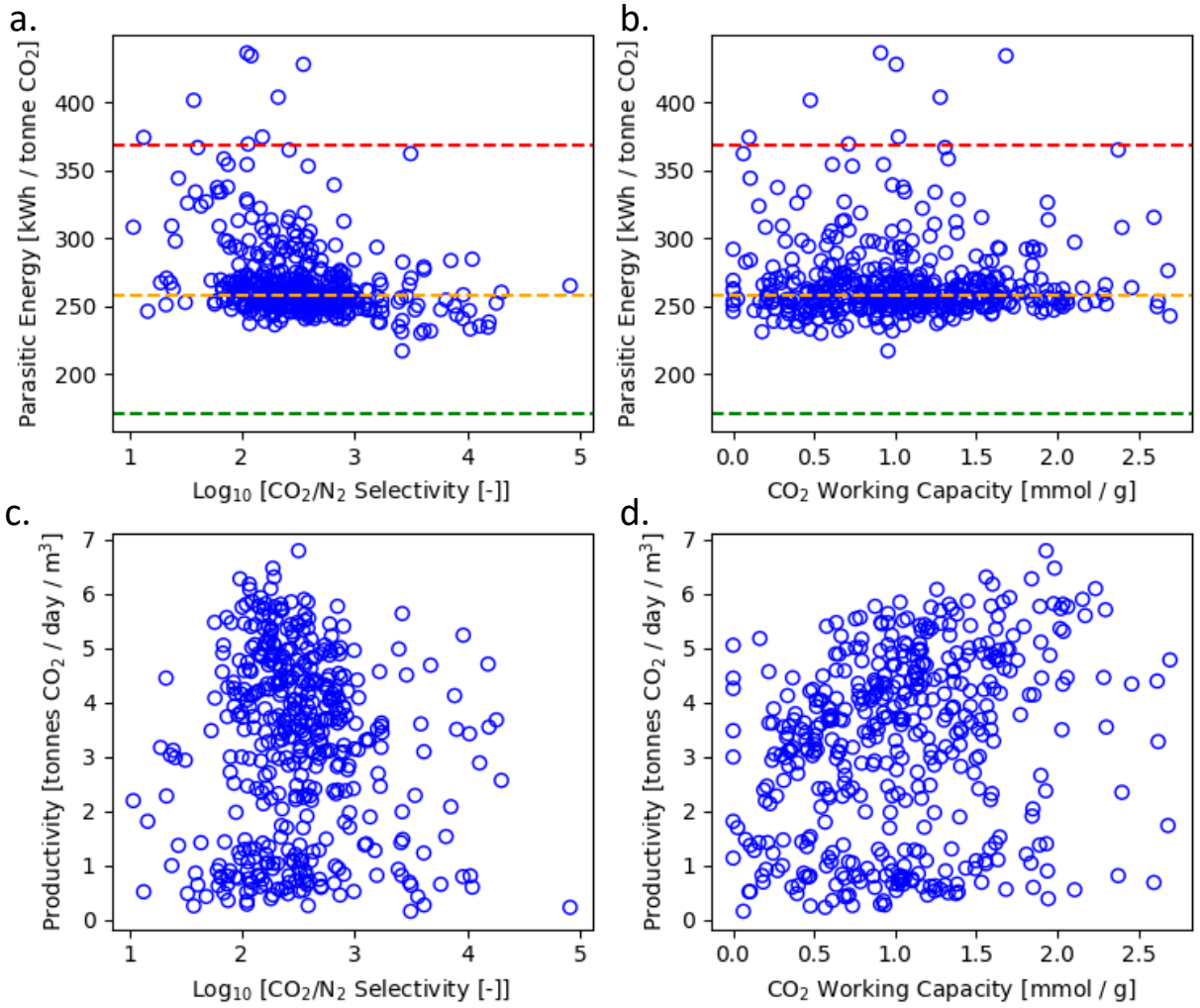


Figure 4.8 (a) The plot of the best parasitic energy for all 482 materials that meet the DoE-PRT plotted against the CO_2/N_2 selectivity, (b) the plot of the best parasitic energy for all 482 materials that meet the DoE-PRT plotted against the CO_2 working capacity, (c) the best productivity for all 482 materials that meet the DoE-PRT plotted against the CO_2/N_2 selectivity, and (d) the best productivity for all 482 materials that meet the DoE-PRT plotted against the CO_2 working capacity. Included in the figures (a) and (b) is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line).

This search was expanded to include a wide range of atomic level performance metrics, from simple direct observations such as the CO_2 loading capacity, to more complex metrics such as the composite metric proposed by Chung and coworkers, the Adsorbent Performance Score.[13] To compare the 8 fitted DSL parameters, 22 adsorption metrics, and 7 geometric properties to PSA performance, Pearson Correlation R^2 and Spearman Correlation R^2 values were calculated for each metric. R^2 values were calculated to compare the best parasitic energy and productivity values for each material. The results from this analysis are shown in Table 4.9 and demonstrate that no correlation exists between any

of the metrics and PSA performance. This is evidenced by the highest Spearman R^2 achieved for any of the metrics having a value of 0.15, between the Yang Figure of Merit (FOM)[61] and the Parasitic Energy. Interestingly, the Huck Parasitic Energy,[48] a metric designed by Huck and coworkers to estimate the process parasitic energy based on atomistic simulations, completely fails to predict the parasitic energy from the PSA simulator. The lack of insight provided by these atomistic metrics and geometric properties demonstrates the need for a holistic approach to screening materials, requiring both the atomistic and process level simulations to predict a material's performance effectively. A more in-depth description of the composite metrics and their relationship to PSA performance is explored in Chapter 5.

Table 4.9 The Pearson and Spearman correlation R^2 values comparing the best parasitic energies and productivities to 22 common adsorption metrics and 7 geometric properties used to evaluate materials for post-combustion carbon capture.

Metrics	PE Pearson R^2	PE Spearman R^2	Productivity Pearson R^2	Productivity Spearman R^2
Single component CO_2 uptake [mmol/g]	0.00	0.00	0.03	0.02
Competitive CO_2 uptake [mmol/g]	0.01	0.00	0.03	0.02
Single component N_2 uptake [mmol/g]	0.01	0.04	0.00	0.00
Competitive N_2 uptake [mmol/g]	0.08	0.10	0.00	0.01
CO_2 working capacity [mmol/g]	0.00	0.00	0.06	0.08
Competitive CO_2 working capacity [mmol/g]	0.00	0.00	0.06	0.08
N_2 Working Capacity [mmol/g]	0.01	0.04	0.00	0.00
Competitive N_2 Working Capacity [mmol/g]	0.08	0.10	0.00	0.01
Single component CO_2/N_2 selectivity [-]	0.02	0.05	0.02	0.06
CO_2/N_2 Selectivity [-]	0.01	0.10	0.01	0.00
CO_2 Heat of adsorption [kcal/mol]	0.02	0.00	0.02	0.04
N_2 Heat of adsorption [kcal/mol]	0.00	0.01	0.01	0.03
Huck's PE [kWh/tonne CO_2]	0.00	0.00	0.00	0.07
Henry Selectivity [-]	0.00	0.02	0.00	0.01
Adsorbent performance score [mmol/g]	0.01	0.09	0.01	0.01
Percent Regenerability [%]	0.01	0.00	0.02	0.03
Separation Potential [mmol/g]	0.01	0.00	0.03	0.02
Sorbent Selection Parameter [mmol/g]	0.00	0.12	0.01	0.00
Notaro's FOM [mmol/g]	0.00	0.11	0.01	0.00
Ackley's FOM [mmol/g]	0.01	0.10	0.01	0.01
Yang's FOM [-]	0.01	0.15	0.00	0.01
Wiersum's FOM [mol^3/J/kg]	0.01	0.02	0.03	0.06
Crystal density [kg/m^3]	0.00	0.00	0.01	0.01
Maximum accessible pore diameter [Å]	0.01	0.00	0.00	0.01
Maximum channel diameter [Å]	0.01	0.00	0.00	0.02
Maximum pore diameter [Å]	0.01	0.00	0.00	0.01
Volumetric surface area [m^2/cm^3]	0.02	0.03	0.01	0.02
Gravimetric surface area [m^2/g]	0.01	0.02	0.02	0.02
Gravimetric void volume [cm^3/g]	0.01	0.01	0.00	0.01

4.5. Conclusions

The work presented in Chapter 4 yielded several important insights into using MOFs in a PoC-CCS PSA system. Over the course of this study, 1,022 MOFs were fully optimized using a sophisticated PSA simulator, based on isotherms calculated using grand canonical Monte Carlo simulations. This was the first large scale screening of MOFs for PoC-CCS which effectively bridged the gap between atomistic level studies to process level design. By combining computational chemistry techniques with process engineering simulations, 482 MOFs were identified with were able to meet Purity-Recovery targets set by the US DoE. Of those 482 MOFs, 223 could meet the US-DoE's Parasitic Energy targets. This demonstrated that MOFs used in post-combustion PSA systems are excellent candidates for industrial use. The size of the capture plants needed was also explored and it was concluded that many materials would require large and complex capture plants, which would likely result in prohibitively expensive initial capital investment and operational cost. Although PSA may nots be practical for use with powerplants, it may in fact be feasible for use on smaller operations such as cement and steel production plants.

Importantly, all flue gases tested over the course of this work did not include any water. The decision to exclude water, an important by-product of combustion, was made due to the large computational cost of calculating adsorption isotherms for water with GCMC, and a lack of available and reliable the force-field parameters for water adsorption at the time this study was performed. Furthermore, at the time this work was performed, the PSA simulation code was unable to account for water in the gas stream. Finally, the analysis in this chapter is a first pass assessment of these materials, aimed at identifying promising materials for the process at an early stage of development. This means that prior to use in an industrial PSA unit, the materials identified here-in would need to be tested experimentally to ensure the adsorption behaviour of CO_2 and N_2 remains unchanged in the presence of water, as was reported for IISERP-MOF-2.[62]

Finally, several traditional atomistic performance metrics were considered and compared to the process performance values. It was found that none of these metrics were predictive of the parasitic energy or productivity of materials, and therefore could not be directly used to estimate the operational or initial capital costs of capture. This is significant as these performance metrics are often used to identify high performing materials. Some of these metrics, like the Huck's PE,[48] were designed specifically to estimate the industrial PSA performance and were found to be poor indicators of that

performance. It was therefore concluded that to effectively vet a material for a PoC-CCS application, a full holistic approach needs to be taken, studying the material from pores to process.

4.6. References

1. Burns, T. D. *et al.* Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

2. Lin, L.-C. *et al.* In silico screening of carbon-capture materials. *Nature Materials* **11**, 633–641 (2012).

3. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry* **4**, 83–89 (2012).

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

5. Fernandez, M., Trefiak, N. R. & Woo, T. K. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. *Journal of Physical Chemistry C* **117**, 14095–14105 (2013).

6. Erlach, B., Schmidt, M. & Tsatsaronis, G. Comparison of carbon capture IGCC with pre-combustion decarbonisation and with chemical-looping combustion. *Energy* **36**, 3804–3815 (2011).

7. IPCC. Climate Change 2014 Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change 1–112 (2014).

8. Rogelj, J. et al. Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development. Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, (2018).

9. D'Alessandro, D. M., Smit, B. & Long, J. R. Carbon Dioxide Capture: Prospects for New Materials. *Angewandte Chemie International Edition* **49**, 6058–6082 (2010).

10. Aspelund, A. & Jordal, K. Gas conditioning-The interface between CO_2 capture and transport. *International Journal of Greenhouse Gas Control* **1**, 343–354 (2007).

11. Giannaris, S., Bruce, C., Jacobs, B., Srisang, W. & Janowczyk, D. Implementing a second generation CCS facility on a coal fired power station – results of a feasibility study to retrofit SaskPower's Shand power station with CCS. *Greenhouse Gases: Science and Technology* **10**, 506–518 (2020).

12. Giannaris, S. et al. SaskPower's Boundary Dam Unit 3 Carbon Capture Facility-The Journey to Achieving Reliability. (2021).

13. Chung, Y. G. *et al.* In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances* **2**, e1600909 (2016).

14. Hasan, M. M. F., First, E. L. & Floudas, C. A. Cost-effective CO_2 capture based on in silico screening of zeolites and process optimization. *Physical Chemistry Chemical Physics* **15**, 17601 (2013).

15. 2020 Carbon capture program R&D compendium of carbon capture technology, National Energy Technology Laboratory, United States Department of Energy (2020).

16. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to metal–organic frameworks. *Chemical Reviews* **112**, 673–674 (2012).

17. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **341**, (2013).

18. Moghadam, P. Z. *et al.* Development of a Cambridge structural database subset: A collection of metal-organic frameworks for past, present, and future. *Chemistry of Materials* vol. 29 2618–2625 (2017).

19. Nugent, P. *et al.* Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

20. Liang, L. *et al.* Carbon dioxide capture and conversion by an acid-base resistant metal-organic framework. *Nature Communications* **8**, (2017).

21. Jiang, J. *et al.* Higher symmetry multinuclear clusters of metal–organic frameworks for highly selective CO_2 capture. *Journal of the American Chemical Society* **2** (2018).

22. McDonald, T. M. *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

23. Yu, J. *et al.* CO_2 capture and separations using MOFs: Computational and experimental studies. *Chemical Reviews* **117**, 9674–9754 (2017).

24. Dzubak, A. L. *et al.* Ab initio carbon capture in open-site metal-organic frameworks. *Nature Chemistry* **4**, 810–816 (2012).

25. Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature* **576**, 253–256 (2019).

26. Mission Innovation. Accelerating Breakthrough Innovation in Carbon Capture, Utilization, and Storage. 1–291 (2017).

27. Walton, K. S. & Sholl, D. S. Research challenges in avoiding "showstoppers" in developing materials for large-scale energy applications. *Joule* **1**, 208–211 (2017).

28. Chung, Y. G. *et al.* Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **26**, 6185–6192 (2014).

29. Haghpanah, R., Nilam, R., Rajendran, A., Farooq, S. & Karimi, I. A. Cycle synthesis and optimization of a VSA process for postcombustion CO_2 capture. *AIChE Journal* **59**, 4735–4748 (2013).

30. Kresse, G. & Furthmüller, J. Vienna ab initio simulation package (VASP). (2001).

31. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868 (1996).

32. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **105**, 9982–9985 (1996).

33. Boyd, P. G. Computational high throughput screening of metal organic frameworks for carbon dioxide capture and storage applications. (2015).

34. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

35. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (1990).

36. Coupry, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal-organic frameworks. *Journal of Chemical Theory and Computation* (2016).

37. Krykunov, M., Demone, C., Lo, J. W. H. & Woo, T. K. A new split charge equilibration model and REPEAT electrostatic potential fitted charges for periodic frameworks with a net charge. *Journal of Chemical Theory and Computation* **13**, (2017).

38. García-Sánchez, A. *et al.* Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

39. Provost, B. An Improved N_2 Model for predicting gas adsorption in MOFs and using molecular simulation to aid in the interpretation of SSNMR spectra of MOFs. (2014).

40. Peng, D.-Y. & Robinson, D. B. A new two-constant equation of state. *Industrial & Engineering Chemistry Fundamentals* **15**, 59–64 (1976).

41. Nandi, S. *et al.* A single-ligand ultra-microporous MOF for precombustion CO_2 capture and hydrogen purification. *Science Advances* **1**, e1500421–e1500421 (2015).

42. Vaidhyanathan, R. *et al.* Competition and cooperativity in carbon dioxide sorption by amine-functionalized metal-organic frameworks. *Angewandte Chemie - International Edition* **51**, 1826–1829 (2012).

43. Vaidhyanathan, R. *et al.* Direct observation and quantification of CO_2 binding within an amine-functionalized nanoporous solid. *Science (New York, N.Y.)* **330**, 650–3 (2010).

44. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **149**, 134–141 (2012).

45. Krishnamurthy, S. *et al.* CO_2 capture from dry flue gas by vacuum swing adsorption: A pilot plant study. *AlChE* **60**, 1830–1842 (2014).

46. MATLAB. *MATLAB R2016a*. (The MathWorks Inc., 2016).

47. Ulrich, D. G. & Vasudevan, P. T. Chemical engineering process design and economics: A practical guide. (2004).

48. Huck, J. M. *et al.* Evaluating different classes of porous materials for carbon capture. *Energy Environ. Sci.* **7**, 4132–4146 (2014).

49. Shen, V. K., Siderrius, D. W., Krekelberg, W. P. & Hatch, H. W. *NIST Standard Reference Simulation Website, NIST Standard Reference Database Number 173*. (National Institute of Standards and Technology).

50. Level, D. European best practice guidelines for assessment of CO_2 capture technologies. (2011).

51. Fout, T. *et al.* Cost and performance baseline for fossil energy plants volume 1a: Bituminous coal (PC) and natural gas to electricity revision 3. *National Energy Technology Laboratory (NETL)* **1a**, 240 (2015).

52. Kumar, R. Vacuum swing adsorption process for oxygen production - A historical perspective. *Separation Science and Technology* vol. 31 877–893 (1996).

53. Valenzano, L. *et al.* Computational and experimental studies on the adsorption of CO, N_2, and CO_2 on Mg-MOF-74. *The Journal of Physical Chemistry C* **114**, 11185–11191 (2010).

54. Qasem, N. A. A. & Ben-Mansour, R. Energy and productivity efficient vacuum pressure swing adsorption process to separate CO_2 from CO_2/N_2 mixture using Mg-MOF-74: A CFD simulation. *Applied Energy* **209**, 190–202 (2018).

55. Weitkamp, J. et al. Industrial applications of zeolite catalysis. Studies in Surface Science and Catalysis vol. 84 (1994).

56. Heard, C. J. *et al.* Zeolite (In)Stability under Aqueous or Steaming Conditions. *Advanced Materials* vol. 32 (2020).

57. Tomeczek, J. & Palugniok, H. Specific heat capacity and enthalpy of coal pyrolysis at elevated temperatures. *Fuel* **75**, 1089-1093 (1996).

58. Simon, C. M. *et al.* The materials genome in action: Identifying the performance limits for methane storage. *Energy and Environmental Science* **8**, 1190–1199 (2015).

59. Konstas, K. *et al.* Methane storage in metal organic frameworks. *Journal of Materials Chemistry* vol. 22 16698–16708 (2012).

60. Li, B., Wen, H. M., Zhou, W., Xu, J. Q. & Chen, B. Porous Metal-Organic Frameworks: Promising Materials for Methane Storage. *Chem* vol. 1 557–580 (2016).

61. Rege, S. & Yang, R. A simple parameter for selecting an adsorbent for gas separation by pressure swing adsorption. *Separation Science and Technology* **36**, 3355–3365 (2001).

62. Nandi, S. *et al.* Ultralow parasitic energy for Postcombustion CO_2 capture realized in a nickel isonicotinate metal–organic framework with excellent moisture stability. *Journal of the American Chemical Society* **139**, 1734–1737 (2017).

4.7. Author Contributions

The work discussed in this chapter was a collaborative effort between several members of the research lab of Dr. Tom Woo at the University of Ottawa and Dr. Arvind Rajendran at the University of Alberta. The initial Pressure Swing Adsorption code was provided by Dr. Arvind Rajendran.

My contributions to this work were extensive and touched on every aspect of the project. Dr. Sean Collins, a former PhD student of Dr. Tom Woo, and I performed the GCMC simulations and isotherm fittings. The genetic algorithm used in the optimizations was originally written by Sean Collins, however it required extensive tuning for this application. I developed and implemented the 2-phase optimization to the Genetic algorithm, as well as made several modifications to the code to make it compatible with the PSA simulation code.

I worked in collaboration with Nagesh Kasturi, a PhD student of Dr. Arvind Rajendran, to understand the workings of the PSA simulator. I was then responsible for generalizing the PSA simulation code, allowing for it's use in high-throughput screening applications, compiling the updated code and porting it to a high-performance computing cluster. I performed extensive edits to the PSA simulation code, adding the compression calculations, generalizing the input parameters, and adding compatibility with other isotherm types. I worked in collaboration with Gokul Subraveti, a student of Dr. Arvind Rajendran, to develop the compression term to the parasitic energy.

I was responsible for executing the geometric and chemical filters as well as developing the grid-search algorithm, running the code, and the final extraction and processing of the resulting data. This step included the vetting, selection, and ranking of MOFs for full Genetic Algorithm optimization. The values and ranges used in the grid-search were provided by Dr. Rajendran. I performed the full genetic algorithm optimizations on all CoRE MOFs discussed in this chapter, extracted the data, and performed all post-processing shown herein. Again, the ranges of the operating parameters testing in the genetic algorithm were provided by Dr. Rajendran. I calculated all geometric and uptake properties discussed in this chapter for the MOFs in the CoRE database and performed the comparison of those metrics to the final process conditions.

4.8. Appendix

4.8.1. Appendix 4.1: List of Named Materials

Table A4.1. List of the 46 named materials optimized over the course of this study.

ZIF-68	Cu-bttri	ppn-6-CH_2TAEA
ZIF-69	ppn-6-CH_2DETA	ppn-6-CH_2TETA
sifsix-3-Cu	ZIF-115-mer	ZIF-40-gis
ZIF-36-frl	sifsix-3-Zn	ZIF-116-cag
IISERP-MOF-2	NaX	ZIF-8
MOF-177	ZIF-79	ZIF-116-sod
Co-MOF-74	ZIF-78	CaX
ZIF-82	ZIF-81	ppn-4
ppn-6-CH_2Cl	magic-mof	ppn-6
Mg-MOF-74	MgA	Cu-btc
Ni-MOF-74	UMCM-1	ppn-6-SO_3Li
ps-mfi	ZIF-70	mme-cubttri
ZIF-36-cag	ppn-6-SO_3H	HMOF-5
MgX	ZIF-39-dia	Zn-MOF-74
ZIF-39-zni	UTSA-16	CaA
Cu-bttri	NaA	

4.8.2. Appendix 4.2: N2-NIMF Lennard-Jones Parameters

Table A4.2. Nitrogen in Metal-organic frameworks (N2-NIMF) Lennard-Jones parameters used in GCMC simulations.

Atom	bl (Å)*	q (e)	(ε/k_B) (K)	σ (Å)
N	0.5500	-0.4820	39.996	2.4549
COM	0.0000	0.9640	0.000	0.0000

*bl is the distance from the atom to the centre of mass (COM).

4.8.3. Appendix 4.3: Density of Supercritical Fluid Mixture

Table A4.3. Average density of a supercritical mixture of CO_2 and N_2 as a function of mole fraction of CO_2 from the NIST-REFPROP database.

Temperature [K]	CO_2 mole composition	Density @ 80 bar [kg/m^3]	Density @150bar [kg/m^3]	Average density [kg/m^3]
303.15	0.6	139.83	300.54	220.185
303.15	0.65	147.11	326.8	236.955
303.15	0.7	155.35	359.22	257.285
303.15	0.75	164.82	400.19	282.505
303.15	0.8	175.99	453	314.495
303.15	0.85	189.62	521.12	355.37
303.15	0.9	207.19	604.74	405.965
303.15	0.92	215.96	640.87	428.415
303.15	0.94	226.26	677.25	451.755
303.15	0.96	238.82	712.99	475.905
303.15	0.98	254.96	747.42	501.19
303.15	1	277.9	780.23	529.065

4.8.4. Appendix 4.4: Theoretical Thermodynamic Limit

To calculate the theoretical thermodynamic limit of separation several assumptions were made:

1. The only term in the $\Delta G_{separation}$ was the $\Delta S_{separation}$, meaning this limit only considers the entropic cost of separating the gases.
2. The gas mixture is ideal and $\Delta S_{separation} = -\Delta S_{mixing}$
3. The gas in the column at the time of evacuation had a CO_2 purity of 100%.
4. All compressors used in the separation and final compression to transport conditions were 100% efficient.

The final calculation was made using equations A4.1 and A4.2, where the $E_{final,compression}$ term is calculated using equation 4.4.

$$E_{thermodynamic\ limit} = -T\Delta S_{\text{separation}} + E_{pumps,evacuation} + E_{final\ compression} \quad \text{(A4.1)}$$

$$E_{thermodynamic\ limit} = nRT\left[\left(x_{CO_2} lnx_{CO_2} + x_{N_2} lnx_{N_2}\right) - ln\left(\frac{P_{desorption}}{P_{adsorption}}\right)\right] + E_{final\ compression} \quad \text{(A4.2)}$$

5. Chapter 5: Data Mining of Detailed Process Simulations

In this chapter, I discuss the work done to study the common equilibrium adsorption metrics used in the field of MOF research and their relation to PSA process performance. The work presented in this chapter is partially based on work published in 2020: *Burns, T. D. et al. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion CO_2 Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. Environmental Science and Technology **54**, 4536–4544 (2020)*.[1] however the work presented herein is more complete and was performed on a larger dataset. All work presented in this chapter is my own.

5.1. Abstract

In this chapter, the results from the screening discussed in Chapter 4 were analyzed with the goal of vetting conventional equilibrium performance metrics commonly used by researchers in the field of materials discovery. The aim was to determine whether any of the conventional metrics were predictive of industrial pressure swing adsorption (PSA) performance. Over the course of this chapter, 37 metrics including fitting isotherm parameters, simple adsorption metrics, complex composite metrics, and geometric properties were used as descriptors in a range of machine learning classification models. These models aimed to classify materials according to their ability to meet the US Department of Energy's purity-recovery target (DoE-PRT), their parasitic energies, or their productivities. It was concluded that none of the 37 metrics were predictive of a material's parasitic energy or productivity, but that several metrics could be used to determine a material's ability to meet the DoE-PRT relating atomistic to industrial level performance for the first time. It was found that the most important metrics in determining a MOF's ability to meet the DoE-PRT were the N_2 isotherm parameters, a significant result given the intuitive belief among MOF researchers that CO_2 adsorption behaviour of a material plays the most crucial role in PSA performance.

5.2. Introduction

Due to the disconnect that exists between material chemists and chemical engineers when studying materials for post-combustion carbon capture, the relationship between the performance of a MOF at the bench scale based on equilibrium experiments and the performance of that same material in a pressure swing adsorption (PSA) column is not well understood. As such, a lack of consensus exists among MOF chemists as to which equilibrium performance metric best represents a MOF's PSA

performance. This lack of consensus has led to many researchers reporting different performance metrics for their discovered materials, ranging from simple metrics like the CO_2 uptake capacity or CO_2/N_2 selectivity to more complex calculated metrics such as the Adsorbent Performance Score (APS) and the Sorbent Selection Parameter (SSP).[2–9] The range of performance metrics available to vet a material has allowed researchers to pick whichever metric is most complimentary to their objectives. This lack of insight has resulted in a substantial knowledge gap in the field of materials discovery for the post-combustion PSA process. The data obtained in Chapter 4 provided a unique opportunity to study these equilibrium metrics and compare them to their industrial PSA performance values. The goal of this work was to determine which atomic level properties, if any, can be used to identify high performing materials at an early stage of development and help streamline materials discovery.

It was shown in Chapter 4 that no direct relationship exists between the process performance of a material and the commonly used equilibrium performance metrics. This became evident when observing the correlation coefficients between equilibrium metrics and two important process performance metrics: the parasitic energy (PE) and the productivity (Prod). Although no direct relationship was found, the size of the dataset assembled over the course of the multi-scale screening provided the opportunity to data mine the results and explore whether more complex relationships exist.

In this chapter these complex relationships were studied using an array of tools ranging from simple univariate analysis to more sophisticated Machine Learning (ML) techniques. These techniques allow researchers to identify trends imbedded in large datasets which may not be apparent to humans, and range in complexity from the relatively simple linear discriminant analysis, which attempts to separate classifications using a single line, to sophisticated statistical ensemble models like the Random Forest method. Using these techniques, the MOFs studied in Chapter 4 were classified into several sets based on their PSA performance. The first set attempted to separate MOFs by their ability to meet the US Department of Energy's Purity and Recovery Targets (DoE-PRT)[10] of 95% purity of captured CO_2 and 90% recovery of CO_2 from the flue gas stream, the minimum requirement for the technology to be considered viable. The second set attempted to separate MOFs by their parasitic energy with classifications of "*high*" or "*low*" based on the median value, and the third attempted to separate the MOFs according to the materials' productivity using the same classification method described for the PE.

5.3. Methodology

5.3.1. Data Set Classification

To appropriately analyse the data from the screening performed in Chapter 4, several subsets of the data were created. The ***fail set*** included all materials that failed to meet the DoE-PRT. A ***pass set*** was generated and included all MOFs which were able to meet the DoE-PRT. When combined, these two subsets are known as the ***DoE-PRT*** binary classification set. Several additional sets were also generated over the course of this analysis. The ***top-150*** set was composed of the top 150 MOFs ranked by their best parasitic energy points. Importantly, since parasitic energy is ill defined when the MOF is unable to meet the DoE-PRT, this subset only included MOFs from the ***pass set***, ranked by the best parasitic energy points found given that constraint.

To study whether any trends exist in parasitic energy, the ***pass set*** was subdivided into two subsets along the *median* parasitic energy. All MOFs whose best parasitic energy points fell below the *median* were added to the ***low PE*** set, and all MOFs whose best parasitic energy points exceeded the *median* were placed in the ***high PE***. Finally, two subsets were created using a similar method to the ***low PE*** and ***high PE*** sets, instead dividing the MOFs along *median* productivity, making the ***low Prod*** and ***high Prod*** sets. A summary of all sets used is provided in Table 5.1.

Table 5.1 Description of the four main classification sets, and subsets used for data analysis throughout this chapter.

SET	SUBSETS	NUMBER OF MOFS	DESCRIPTION
TOP-150	-	150	The top 150 MOFs ranked according to the best parasitic energy (PE) achieved during the optimizations.
DOE-PRT	Fail Set	532	MOFs which failed to meet the DoE-PRT of 95% purity and 90% recovery.
	Pass Set	443	MOFs which met the DoE-PRT of 95% purity and 90% recovery.
PE	Low PE	222	MOFs which met the DoE-PRT and had a parasitic energy (PE) below the median value of 259.3 kWh / tonne CO_2 captured.
	High PE	221	MOFs which met the DoE-PRT and had a parasitic energy (PE) above the median value of 259.3 kWh / tonne CO_2 captured.
PROD	Low Prod	222	MOFs which met the DoE-PRT and had a parasitic energy (PE) below the median value of 3.5 Tonnes CO_2 / day / m^3 of adsorbent.
	High Prod	221	MOFs which met the DoE-PRT and had a productivity (Prod) above the median value of 3.5 Tonnes CO_2 / day / m^3 of adsorbent.

5.3.2. Generating Isotherms

The first step in the analysis was the comparison of the isotherms between the ***fail set***, the ***pass set***, and the ***top-150*** sets. The isotherm parameters from the MOFs in the set were used to generate full isotherms from 0 to 2 bar at 298.15 Kelvin. The range of isotherm values as well as the average adsorption values were then plotted. This analysis was performed using code written in Python 3.8 and was repeated for the ***PE*** and ***Prod sets***.

5.3.3. Conventional Performance Metrics

Throughout this chapter, 37 performance metrics given in Table 5.2 were evaluated. The metrics in Table 5.2 have been subdivided into four groups: group 1 contains the fitted dual-site Langmuir isotherm parameters generated from GCMC data, group 2 contains the simple adsorption metrics, group 3 contains the composite adsorption metrics, and group 4 consists of geometric properties of the MOFs. This section describes how the metrics were computed for this study.

The metrics in group 1 were fit using data generated single component GCMC simulations at 18 different pressure points ranging from 0.01 to 1.20 bar at 298 Kelvin for each guest. The GCMC simulations used to model the adsorption at individual pressure points was performed using a code developed in-house[11] based on the DL_POLY Classic code.[12] GCMC simulations model the interactions between a gas molecule and the framework atoms by calculating the sum of the Lennard-Jones and Coulomb potentials (see section 2.2.2). The Lennard-Jones parameters for the frameworks atoms were taken from the DREIDING[13] forcefield when available, and supplemented by the Universal Force Field[14] when unavailable. The atomic charges on the framework atoms were calculated using the REPEAT method,[15] which fits partial atomic charges to the quantum mechanical (QM) electrostatic potential (ESP) of the MOF. The QM ESP was calculated using the Vienna ab-inito Simulation Package (VASP)[16,17] using the PBE functions[18,19] with an energy cutoff of 400 eV. The CO_2 molecules were modelled using the Garcia-Sanchez[20] parameters, while the N_2 molecules were modelled using the N2-NIMF[21] parameters. Simulations were performed using 30,000 equilibration cycles and 30,000 production cycles. Once all 18 pressure points were generated using GCMC, the isotherm parameters were fit to minimize the Root Mean Squared Error (RMSE) using code written in Python3.8.

Metrics in group 2 in Table 5.2 were calculated assuming a dry post-combustion flue gas from a coal-fired powerplant at 298 Kelvin, 0.85 bar N_2, and 0.15 bar CO_2. The desorption conditions used assumed a temperature of 298 Kelvin, and pressures of 0.015 bar CO_2 and 0.085 bar N_2. The gas adsorption values at

each set of conditions were calculated using the same GCMC parameters described for group 1. The composite adsorption metrics in group 3 were calculated using the values calculated from group 2 and the equations in Appendix 5.1. Finally, the geometric properties, found in group 4, were calculated using the Zeo++ package.[22]

Table 5.2 List of all features tested, along with the corresponding ID used in Figure 5.4. These features are separated into four groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties.

ID	Group	Feature
1	1	CO_2 Q_1 - Site 1 saturation uptake [mmol/g]
2		CO_2 Q_2 - Site 2 saturation uptake [mmol/g]
3		CO_2 b_{298} – Site 1 Langmuir parameter [1/bar] @ 298 K
4		CO_2 d_{298} – Site 2 Langmuir parameter [1/bar] @ 298 K
5		N_2 Q_1 - Site 1 saturation uptake [mmol/g]
6		N_2 Q_2 - Site 2 saturation uptake [mmol/g]
7		N_2 b_{298} – Site 1 Langmuir parameter [1/bar] @ 298 K
8		N_2 d_{298} – Site 2 Langmuir parameter [1/bar] @ 298 K
9	2	CO_2 Heat of adsorption [kcal/mol]
10		N_2 Heat of adsorption [kcal/mol]
11		Single component CO_2 uptake [mmol/g]
12		Competitive CO_2 uptake [mmol/g]
13		CO_2 working capacity [mmol/g]
14		Competitive CO_2 working capacity [mmol/g]
15		Single component N_2 uptake [mmol/g]
16		Competitive N_2 uptake [mmol/g]
17		N_2 working Capacity [mmol/g]
18		Competitive N_2 working Capacity [mmol/g]
19		Single component CO_2/N_2 selectivity [-]
20		CO_2/N_2 Selectivity [-]
21	3	Henry Selectivity [-][23]
22		Huck's PE [kWh/tonne CO_2][24]
23		Adsorbent performance score [mmol/g][8]
24		Percent Regenerability [%][25]
25		Separation Potential [mmol/g][26]
26		Sorbent Selection Parameter [mmol/g][9]
27		Notaro's FOM [mmol/g][27]
28		Ackley's FOM [mmol/g][28]
29		Yang's FOM [-][9]
30		Wiersum's FOM [mol^3/J/kg][29]
31	4	Crystal density [kg/m^3]
32		Maximum accessible pore diameter [Å]
33		Maximum channel diameter [Å]
34		Maximum pore diameter [Å]
35		Volumetric surface area [m^2/cm^3]
36		Gravimetric surface area [m^2/g]
37		Gravimetric void volume [cm^3/g]

5.3.4. Preparation of Univariate Plots

A univariate analysis was performed on 37 conventional metrics shown in Table 5.2, commonly used to assess the viability of metal-organic frameworks for use in post-combustion carbon capture. This analysis involved the comparison of the normalized histograms of the fail, the pass, and the top-150 sets. The histograms were generated, normalized, and smoothed, using code written in Python 3.8, with the matplotlib and SciPy modules.[30] All histograms were smoothed using a Gaussian Kernel and normalized so that the area under the curve would equal to 1. Once the histograms were generated, they were compared visually to determine whether any individual metrics could be used to differentiate between the sets. This methodology was then repeated on the subsets in the PE set and the Prod set.

5.3.5. Machine Learning Techniques

To search for complex relationships between the equilibrium metrics and the PSA performance results, a series of machine learning techniques were applied. These techniques ranged in complexity from 1-dimensional linear discriminant analysis to complex ensemble methods like the Random Forest, encompassing both supervised and unsupervised learning techniques.

5.3.5.1. Supervised and Unsupervised Learning

Two machine learning paradigms were employed over the course of this work: supervised and unsupervised learning. The term supervised learning implies the method being used makes use of provided labels for each item in the dataset to locate trends. This paradigm is typical of classification and regression models in which the model is explicitly trying to reproduce the provided value or classification. Conversely, unsupervised methods are given no indication of the labels the user is trying to predict, and instead searches for trends inherent in the feature space independent of those labels. Once the unsupervised model is fit, the user then applies the labels to verify whether the model is predictive.

5.3.5.2. Classification Models

The supervised machine learning models employed in this work were all classification models. These classification models were fit to differentiate between the subsets in the ***DoE-PRT set***, the ***PE set***, and the ***Prod set***. The decision to rely on classification over regression models for continuous values such as the PE and Prod was made under the assumption that any dataset that would perform well using a regression model would yield a high accuracy classification model. As such, classification was used as a first pass

analysis with the expectation that regression could be applied if high accuracy classification models were generated.

5.3.5.3. Feature Scaling

Prior to attempting any fittings, the features, or the properties being used to fit the models, needed to be appropriately scaled. This scaling ensured that the influence of a single feature would not dominate, and instead the models would consider the relative variance of those features. For this work, the Standard Scaler from the Scikit-Learn Preprocessing package[31] was used, and the scaler was fit using only the training data to ensure no prior knowledge of the validation data could be inferred by the models. After fitting was complete, the validation set data was transformed using the same scaler prior to passing those data to the fitted model. All model validations were performed using 5-fold cross-validation balanced accuracies.

5.3.5.4. Balanced Accuracy

The choice to use balanced accuracies as the performance metric for these models was made to compensate for any imbalances in the data set while providing intuitive results. Since the ***DoE-PRT*** set consists of all 975 MOFs, 443 of which meet the DoE-PRT, a slight imbalance exists between the two classifications, with roughly 45% of the MOFs meeting the DoE-PRT. This means that a model that assumes all MOFs fail the DoE-PRT will have an overall accuracy of 55%. Balanced accuracy is used to correct this imbalance, as the same hypothetical model will always return a balanced accuracy of 50%, regardless of the number of materials in each classification. This means that any model with a balanced accuracy at or below 50% performs the same or worse than a completely random model. Balanced accuracy is calculated using equation 5.1 and can be described as the average prediction accuracy of *N* unique classifications.

$$Balanced\ Accuracy(BA) = \frac{100\%}{N}\sum_{i=1}^{N}\frac{n_i^{True}}{n_i^{All}} \tag{5.1}$$

5.3.5.5. 5-Fold Cross-Validation

Five-fold cross-validation was used to avoid overfitting, optimize the machine learning hyperparameters and validate the machine learning models fit throughout this chapter. This technique was employed due to the relatively small size of the available dataset and ensured an accurate representation of the model performance could be provided. All model performance values presented in this chapter were

calculated using 5-fold cross-validation (see section 2.5.5), using a custom code written in house in Python 3.8.

5.3.5.6. Linear Discriminant Analysis (LDA)

The first machine learning technique employed was linear discriminant analysis (LDA), described in Section 2.5.6. This technique was used to assess whether a single value or a simple line could be used to separate the sets into their respective subsets using one or two features. As such, 1-dimensional and 2-dimensional LDAs were applied to all metrics in Table 5.2 and pairs of metrics, respectively.

5.3.5.6.1. One Dimensional LDA

Linear discriminant analysis was performed on all 37 metrics individually to determine whether any single property could be used to differentiate between MOFs in the subsets with the sets described in Table 5.1. These calculations were performed using the SciKit-Learn Linear Discriminant Analysis package[31] in Python 3.8.

5.3.5.6.2. Two Dimensional LDA

Two-dimensional linear discriminant analysis was performed on all pairs of the 37 features to determine whether any pair of properties could be used to differentiate between subsets of MOFs in the ***DoE-PRT***, ***PE***, *or* ***Prod sets***. These calculations were performed using the SciKit-Learn Linear Discriminant Analysis package[31] in Python 3.8.

5.3.5.7. Kernel Principal Component Analysis

A linear and non-linear Principal Component Analysis was used to determine whether any of the 29 parameters or 8 isotherm parameters could be used to differentiate between the sets of MOFs. Since the PCA technique is unsupervised, this fitting only needed to be performed twice on the entire set of MOFs. The first fitting was performed on all 975 MOFs used in the analysis to analyze the ***DoE-PRT*** set, however when considering the ***PE*** and ***Prod sets*** which only include MOFs which meet the DoE-PRT, a second fitting was necessary and was performed on the 443 MOFs which met the DoE-PRT. Once the fitting was complete, the two main principal components were plotted against on the X and Y-axis. The labels according to the sets being tested were then applied to the data and visually compared to determine whether any groupings between classifications became apparent. The plots are then checked for clusters of

like labels (for example, those that meet the DOE-PRT and those that don't) to determine whether any of the features can be used to differentiate between the classifications. This work was performed using the Scikit Learn Kernel Principal Component package[31] in Python 3.8, using the linear, radian basis function (rbf), and cosine kernels while varying the gamma parameter from 10^{-6} to 1.0 for the rbf kernel.

5.3.5.8. Decision Trees & Random Forest

Decision trees and random forest model classifiers were trained using the Scikit-Learn Random Forest package.[31] The decision trees and random forest models were optimized to distinguish between the subsets in the ***DoE-PRT***, ***PE***, and ***Prod sets***. Once fit, the individual decision trees in the random forest models were extracted and analyzed to determine the importance of individual features. All model accuracies reported herein were calculated using on 5-fold cross-validation.

5.4. Results and Discussion

5.4.1. Data Set Preparation

The data sets outlined in Table 5.1 include the ***DoE-PRT set***, which contains 443 MOFs which meet the DoE-PRT and 532 MOFs which did not meet the DoE-PRT. The 443 MOFs which meet the DoE-PRT were subdivided into subsets to form the ***PE*** and ***Prod sets***. Distributions of the best PE and Prod values for the 442 MOFs which meet the DoE-PRT are shown in Figure 5.1a and b. A full list of the metrics being studied is given in Table 5.2.

5.4.2. Isotherm Analysis

The first step in this analysis was to compare the isotherms being used in the PSA simulator to determine whether any trends are evident among the sets. For this analysis, the ***DoE-PRT*** and the ***Top-150 sets*** were used, and the range of the isotherms as well as the average isotherm for each subset is plotted in Figure 5.2. In this analysis, the average isotherm was calculated by taking the average adsorption value at each pressure point. When the ***Top-150*** set is compared to the subset of MOFs from the ***DoE-PRT set*** which meet the DoE-PRT (***pass set***), the CO_2 isotherms shown in Figure 5.2a and b show that the ***top-150*** set range falls within the bounds of the ***pass set***. Additionally, the average isotherm in these plots, showed by the solid lines, indicates that the average isotherm for the two sets is almost identical. This can be contrasted with the average CO_2 isotherm shown in Figure 2c for the subset of MOFs from the ***DoE-PRT set*** which fail to meet the DoE-PRT (***fail set***) which has an average isotherm with a lower saturation capacity

and is more linear when compared to the average CO_2 isotherms from the other two sets. Although this appears a promising result, the range of isotherms (shown by the shaded regions of Figure 5.2) for the ***fail set*** appears to completely overlap the ***pass*** and ***Top-150 sets***. As such, a more in-depth analysis on the isotherms would be required before a conclusion can be drawn.

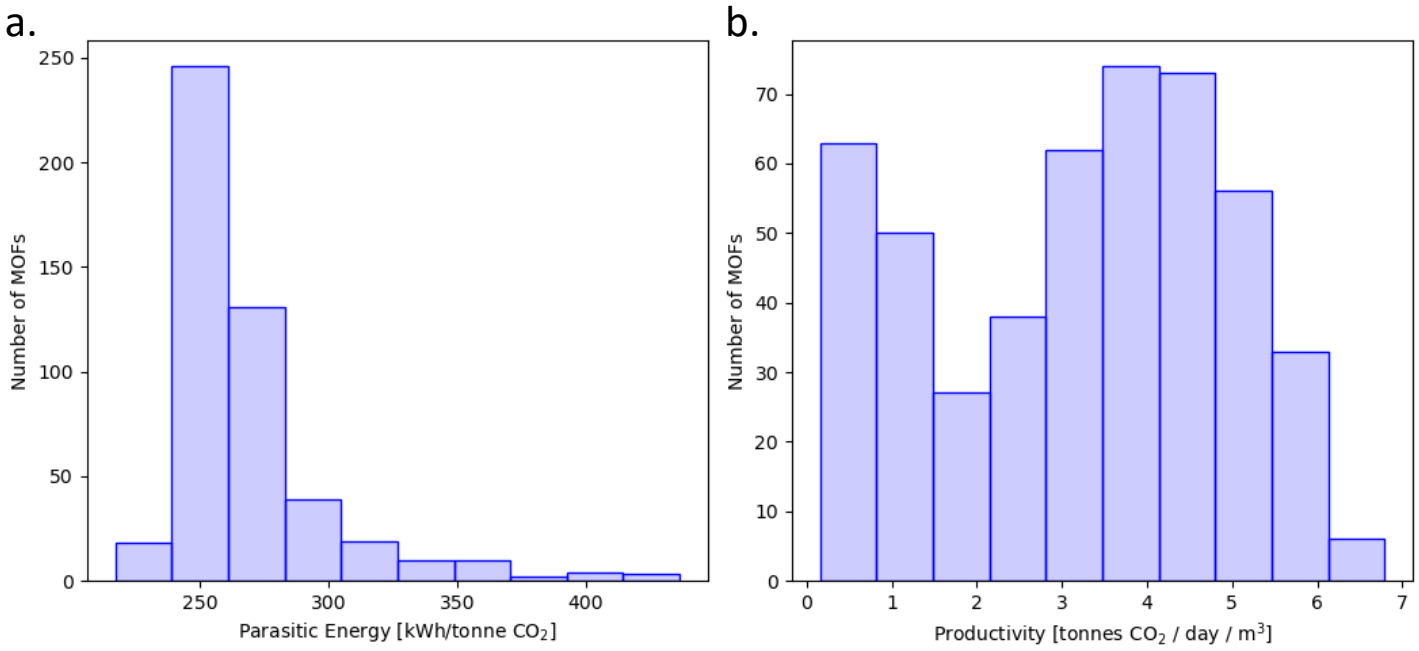


Figure 5.1 Histograms of the (a) best parasitic energy and (b) best productivity for all 1,022 fully optimized MOFs which meet the DoE-PRT.

When considering the N_2 isotherm behaviours in Figure 5.2d, e, and f for the ***top-150***, ***pass***, and ***fail sets***, respectively, little difference can be observed between the average isotherms, however the N_2 isotherm range for the ***fail set*** extends higher and lower than the other two sets. This may imply that the N_2 isotherm behaviour could also play a role in determining a MOF's ability to meet the DoE-PRT. As with the CO_2 isotherm behaviour, additional analysis is required to confirm this relationship.

A similar analysis was performed to compare subsets of the ***PE*** and the ***Prod sets***. The results of this analysis were presented in Appendix 5.2, however there is no clear distinction between the subsets.

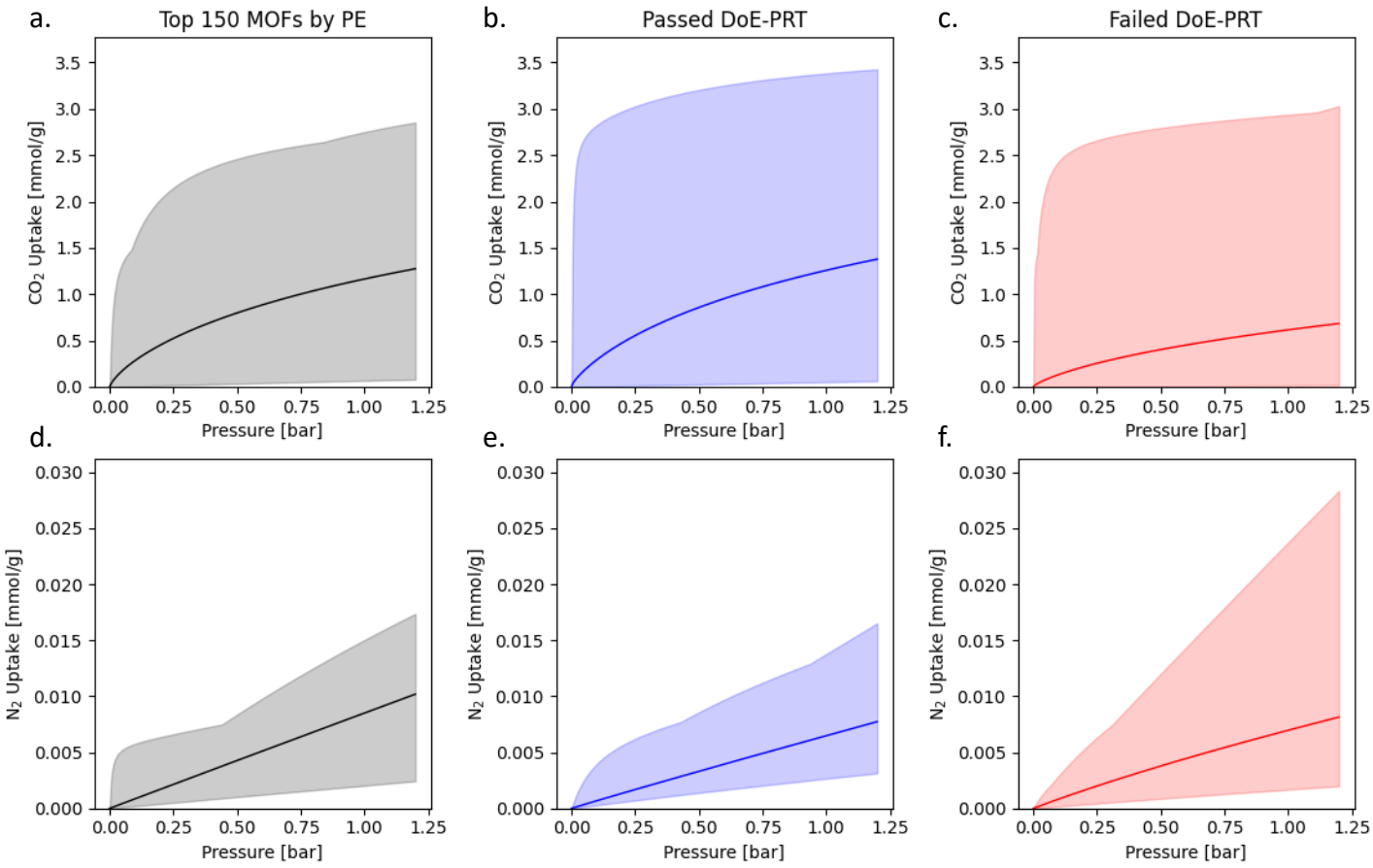


Figure 5.2 Plots showing the range of the isotherms, denoted using a shaded area, and the average isotherm shown in a solid line for (a) CO_2 isotherms of the top 150 MOFs ranked by parasitic energy shown in black, (b) CO_2 isotherms of the MOFs which meet the DoE-PRT shown in blue, (c) the CO_2 isotherms of the MOFs which fail the DoE-PRT shown in red, (d) N_2 isotherms of the top 150 MOFs ranked by parasitic energy shown in black, (e) N_2 isotherms of the MOFs which meet the DoE-PRT shown in blue, and (f) the N_2 isotherms of the MOFs which fail the DoE-PRT shown in red.

5.4.3. Univariate Analysis

To perform a more in-depth analysis into the isotherms and the various adsorption metrics listed in Table 5.2, a univariate analysis was performed. This aimed to compare single metrics in isolation and determine whether they provide any insight into the materials' process performance. By plotting normalized histograms of each property and separating the MOFs according to their subset (example ***pass*** and ***fail DoE-PRT subsets***), a visual comparison of the sets allows for an initial assessment of the features' ability to separate the MOFs into their respective classifications.

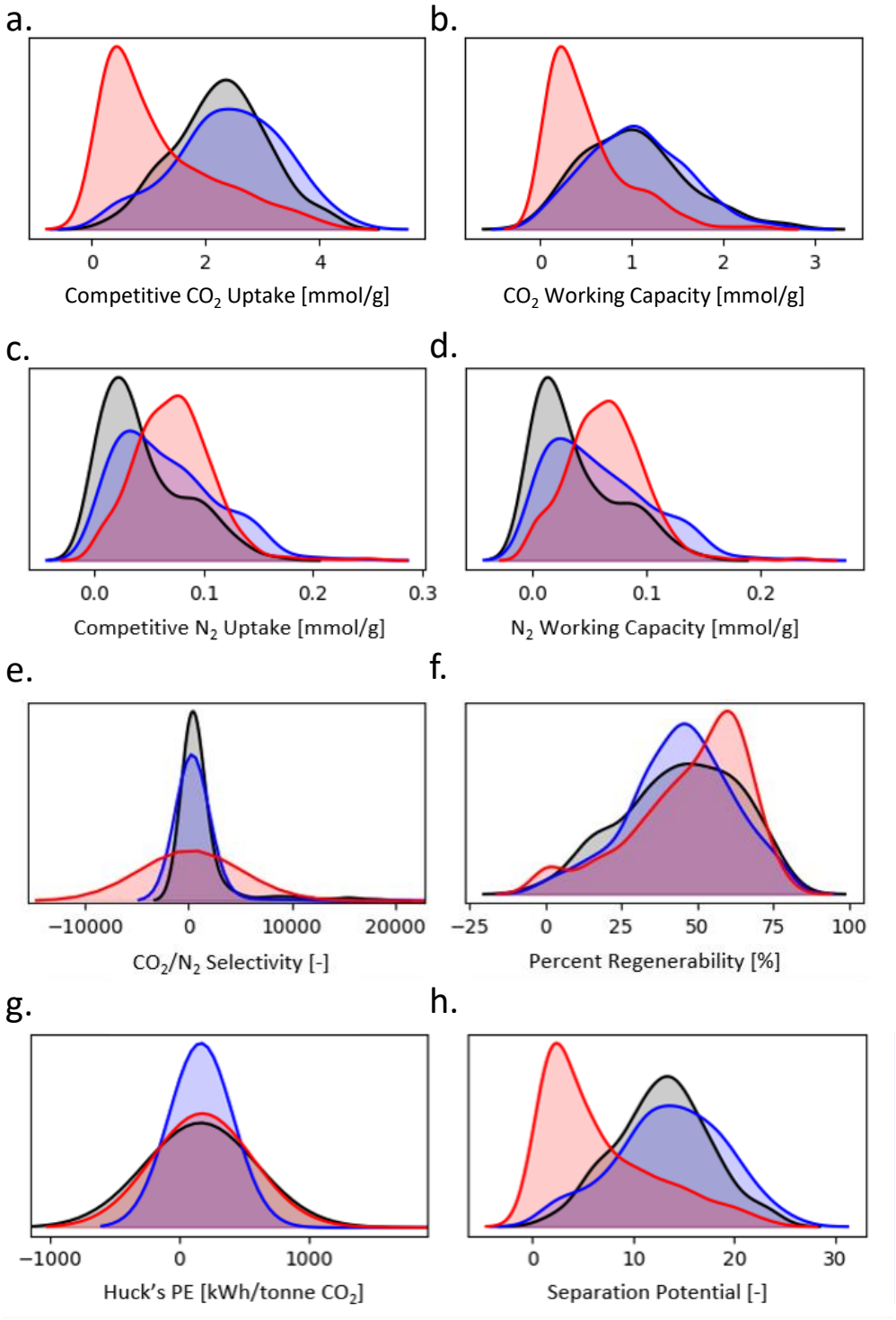


Figure 5.3 Univariate plots showing smoothed distributions of (a) competitive CO_2 uptake, (b) CO_2 working capacity, (c) competitive N_2 uptake, (d) N_2 working capacity, (e) CO_2/N_2 selectivity, (f) percent Regenerability, (g) Huck's parasitic energy, and (h) the separation potential. In this figure, the top 150 MOFs ranked by parasitic energy are shown in black, the MOFs which meet the DoE-PRT are shown in blue, and the MOFs which fail the DoE-PRT are shown in red. All histograms are normalized so that the area under the curve is equal to 1.

For this analysis, MOFs in the ***pass***, ***fail***, and ***top-150*** sets were compared. The histograms for all 37 metrics can be found in Appendix 5.2, with several important metrics included in Figure 5.3. The metrics most discussed in literature to denote a material's performance are the CO_2 uptakes, CO_2 working capacity, and CO_2/N_2 selectivity, shown in Figure 5.3a, b, and e, respectively. It should be noted that although several plots shown in Figure 5.3 contain negative values such as working capacity, selectivity, and Huck's PE, these negative values are a biproduct of the gaussian kernel used to smooth the distributions and no negative values exist in the original unsmoothed distributions. Furthermore, although little to no separation can be observed between the ***top-150*** (black) and ***pass set*** (blue) peaks for the CO_2 uptake and CO_2 working capacities, a significant separation between the ***pass*** and ***fail sets*** is apparent for both metrics. This indicates that the CO_2 uptake capacity and working capacity may be used to inform on a MOF's ability to meet the DoE-PRT at the extremes, however when a MOF displays moderate CO_2 uptake capacity or working capacity it's ability to meet the DoE-PRT cannot be inferred. Top overlap of the ***top-150*** and ***pass set*** distributions for these two metrics indicates that they do not provide any useful insight into the materials' best parasitic energies. When comparing the same peaks for the CO_2/N_2 selectivity in Figure 5.3e, the histograms completely overlap indicating this metric cannot be used to predict a material's process performance on its own.

A similar comparison can be made for the N_2 uptake capacity and working capacity in Figure 5.3c and d, respectively. Although the N_2 peaks show more overlap than the CO_2 adsorption metrics, some separation is apparent, particularly when comparing the ***top-150*** and ***fail set***. This indicates the potential for these two metrics to be used in the prediction of a material's ability to meet the DoE-PRT, and the separation between the ***top-150 set*** from the other two sets indicates the potential for these features to be used to predict the material's parasitic energy.

Included in Figure 5.3 is three additional metrics: the percent Regenerability,[25] Huck's parasitic energy,[24] and the separation potential.[26] These metrics are known as composite adsorption metrics and are calculated based on simple adsorption metrics such as uptake capacities and working capacities. The histograms for the Percent Regenerability and the Huck parasitic energy, shown in Figures 5.3f and g, respectively, show significant overlap between all three sets. This overlap implies the metrics' inability to differentiate between the sets. The histograms for the separation potential, however, shows significant separation between the ***pass*** and ***fail sets***, indicating the potential for this metric to be used to predict a materials ability to meet the DoE-PRT.

Histograms were also plotted for the remaining metrics in Appendix 5.2, but none of the metrics are able to differentiate the subsets in the ***Prod*** and ***PE sets***.

5.4.4. Linear Discriminant Analysis

The univariate analysis demonstrated that some metrics may be used to distinguish between MOFs which can and cannot meet the DoE-PRT. To quantify those results, linear discriminant analysis (LDA) was performed on the individual metrics (1-dimensional LDA), and on pairs of metrics (2-dimensional LDA). The aim of the 1-demensional LDA was to explore whether a single value could be used to separate the MOFs into their subsets, while the aim of the 2-dimensional LDAs was to explore whether a straight line in 2-dimsional data could separate the MOFs into their subsets, with a primary focus on predicting whether a MOF could meet the DoE-PRT. The results of this analysis are shown in Figure 5.4, a heatmap showing the 5-fold cross-validation balanced accuracies for all 1 and 2-dimensional LDAs. The numbers along the axis of this figure denote the IDs of the metrics corresponding to the IDs listed in Table 5.2, while the diagonal of this plot represents the 1-dimensional LDAs. The metrics in Figure 5.4 have been subdivided into four groups: group 1 contains the fitted dual-site Langmuir isotherm parameters, group 2 contains the simple adsorption metrics, group 3 contains the composite adsorption metrics, and group 4 consists of geometric properties of the MOFs.

The balanced accuracies shown in Figure 5.4 along the diagonal, representing the 1-D LDAs, range from 49% to 75%. This indicates that using a single metric in the prediction, a material's ability to meet the DoE-PRT can be accurately predicted 75% of the time. The 1-D LDA results for metrics in group 1 and group 4 show that using any of the dual-site Langmuir parameters or geometric properties will never yield a balanced accuracy greater than 56% percent, and therefore barely outperform a completely random model. The group with the highest success in the 1-dimensional LDA models was group 2, which boasted balanced accuracies ranging from 50 to 75%. In agreement with the univariate analysis, the 1-dimensional LDAs for competitive CO_2 uptake capacity (ID 12) was able to predict a MOF's ability to meet the DoE-PRT 75% of the time, while the competitive CO_2 working capacity (ID 14) was also able to make the same prediction with a balanced accuracy of 72%, demonstrating the importance of these metrics in predicting this basic process performance target. Interestingly, when comparing single component (ID 11) to competitive CO_2 uptake (ID 12), the single component CO_2 uptakes underperforms with a balance accuracy of 69%. This is significant since, although competitive uptake values can easily be determined through simulation, experimental determination is significantly more challenging without the use of a competitive model.

Another metric highlighted in the univariate analysis was the competitive N_2 uptake (ID 16). The 1-dimensional LDA for this metric only returned a balanced accuracy of 62%, indicating that using a single

uptake value to separate the classifications was only 12% more effective than a completely random model. Further, the competitive N_2 working capacity (ID 18) had a similar balanced accuracy of 64%, only marginally better than the competitive CO_2 uptake. Interestingly, the single component N_2 uptake and working capacity (ID 15 and 17, respectively) yielded high balanced accuracies when compared to their competitive counterparts, indicating that the N_2 adsorption behaviour in the presence of CO_2 is less predictive than the metrics from experiments relying on pure N_2 gas.

Other metrics that need to be highlighted from this plot are the CO_2/N_2 selectivity (ID 20) which completely fails to predict whether a MOF can meet the DoE-PRT, and the separation potential (ID 25) which was able to predict the MOF's ability to meet the DoE-PRT with 75% balanced accuracy. This accuracy is on par with the competitive CO_2 uptake, however due to the increased complexity in calculating the metrics in Group 3, this result leads to an early conclusion that using composite metrics on their own does not provide any additional insights into a material's process performance.

Next the 2-dimensional LDA results can be examined, which consider whether the classes in the set can be effectively separated by a single line. Overall, when metrics from group 2 are considered in the 2-D LDAs, higher balanced accuracies are obtained, with the highest balanced accuracies reaching 80%. This improvement when adding a single feature to the prediction models demonstrates that simple modelling techniques may be insufficient to capture the process performance of a material, and more complex relationships between the metrics and a material's process performance may exist. The best results from the 2-D LDAs were found when combining single component and competitive CO_2 loading, single component and competitive CO_2 working capacity, and the single component N_2 uptakes and working capacities. The result from the CO_2 metrics is interesting as it implies that not only the competitive adsorption behaviour needs to be considered, but the change in adsorption behaviour when the MOF is exposed to either a pure gas or a mixture of gases. Additionally, the results from the N_2 adsorption metrics indicate that the absolute N_2 loadings when the MOF is exposed to pure N_2 gas play a more vital role in process performance than previously thought.

Similar analysis was performed on the ***PE*** and ***Prod sets*** however no metric or combination of metrics were able to achieve balanced accuracies greater than 63%. The work performed on the ***PE*** and ***Prod sets*** demonstrates that predicting high performance of a material based on *PE* and *productivity* is more challenging compared to predicting its ability to meet the DoE-PRT. The heatmaps for the ***PE*** and ***Prod sets*** can be found in Appendix 5.4.

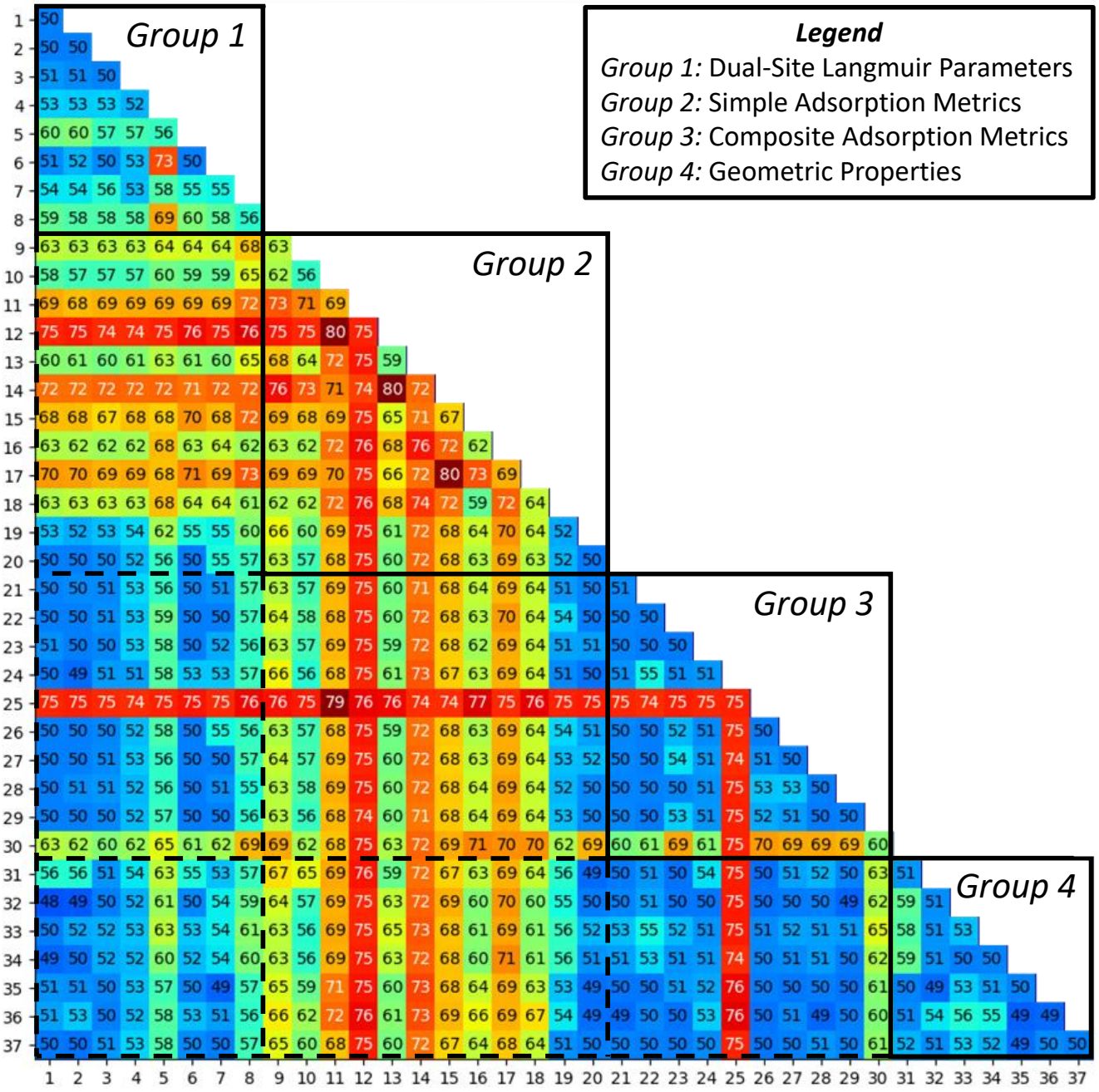


Figure 5.4 A heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional linear discriminant analysis models fit to classify MOFs in the DoE-PRT set according to their subsets, with the diagonal representing the balanced accuracy for the 1-dimensional linear discriminant analysis of each feature. The feature IDs on the axes correspond to the features in Table 5.1 and can be separated into four distinct groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties.

5.4.5. Principal Component Analysis

The initial results from the LDA were promising, however LDAs rely entirely on linear relationships between metrics. Since a key conclusion from the LDA was the potential for the existence of more complex relationships between the metrics and process performance, principal component analysis (PCA) was employed to explore these relationships. For this work, both regular principal component analysis (PCA), which again explores linear relationships between metrics based on the variance in the feature distributions, and kernel principal component analysis (kPCA) which explores non-linear relationships between the variances in the feature distributions, were applied to all sets of MOFs. The LDA analysis showed that these metrics can be used to distinguish between MOFs in the ***DoE-PRT*** set, the aim of the PCA analysis was to explore whether any features could be used to distinguish between subsets in the ***PE*** and ***Prod sets***.

A PCA model, when correctly fit, returns an alternate array of features known as principal components (PC). These PCs represent lines or kernels along the path of greatest variance through the feature space, with the first and second PCs being the principal component vectors (or kernels) with the largest and second largest variance. This means that the PCs determined by the PCA and kPCA models are ranked according to their importance. As this form of unsupervised machine learning does not rely on any random permutations or selection of features, and as a result a PCA with identical features will always return the same result. Once the fittings are complete and the PCs are generated, the *target* property can then be applied to the data to determine whether the model can be used for prediction. For a classification scheme, this can be visualized by plotting the first and second PCs in a scatter plot and assigning colour to the points corresponding to their classifications.

The PCA analysis was performed attempting to distinguish between three different sets: ***DoE-PRT, PE,*** and ***Prod sets*****,** the results of which are presented in Figure 5.5a, b, and c, respectively. Although the models have no prior knowledge of the MOF classifications, the PCA results for the ***DoE-PRT*** set is noticeably different from the ***PE*** and ***Prod sets*** since the former consists of the full 975 MOFs, whereas the ladder only contains the 443 MOFs which met the DoE-PRT. Consistent with the LDA results, the linear kernel in the PCA was able to distinguish between the MOFs which passed the DoE-PRT (blue) from those which failed (red), however with significant overlap in the sets. To quantify these results, an LDA was performed on the top two principal components found using the linear kernel, where the model achieved a balanced accuracy of 74.0 %. Although some separation occurred, when the contributions of the metrics to the two principal components were extracted, it was found that no single metric dominated the PCs. When

considering the ***PE*** and ***Prod sets,*** the PCA was unable to distinguish the subsets – with the ***Low*** and ***High*** subsets completely overlapping.

The kPCA was similarly performed attempting to distinguish between the same three sets: ***DoE-PRT***, ***PE***, and ***Prod sets,*** the results of which are presented in Figure 5.d, e, and f, respectively. For this analysis, the radian basis function (rbf) kernel is presented with a gamma value of 0.1 as it returned the best separation between the subsets. Again, an LDA was performed on the top two principal components found using the rbf kernel and separated the MOFs in the ***DoE-PRT set*** with a balanced accuracy of 75.1%, only improving upon the linear kernel by 1.1%. Figures containing the results from the cosine kernel can be found in Appendix 5.5, along with the alternate gamma values tested for the rbf kernel. The results shown in Figure 5.5d, e, and f are all consistent with those seen with the PCA: some separation between MOFs which pass and fail the DoE-PRT, but no separation between the parasitic energy and productivity subsets.

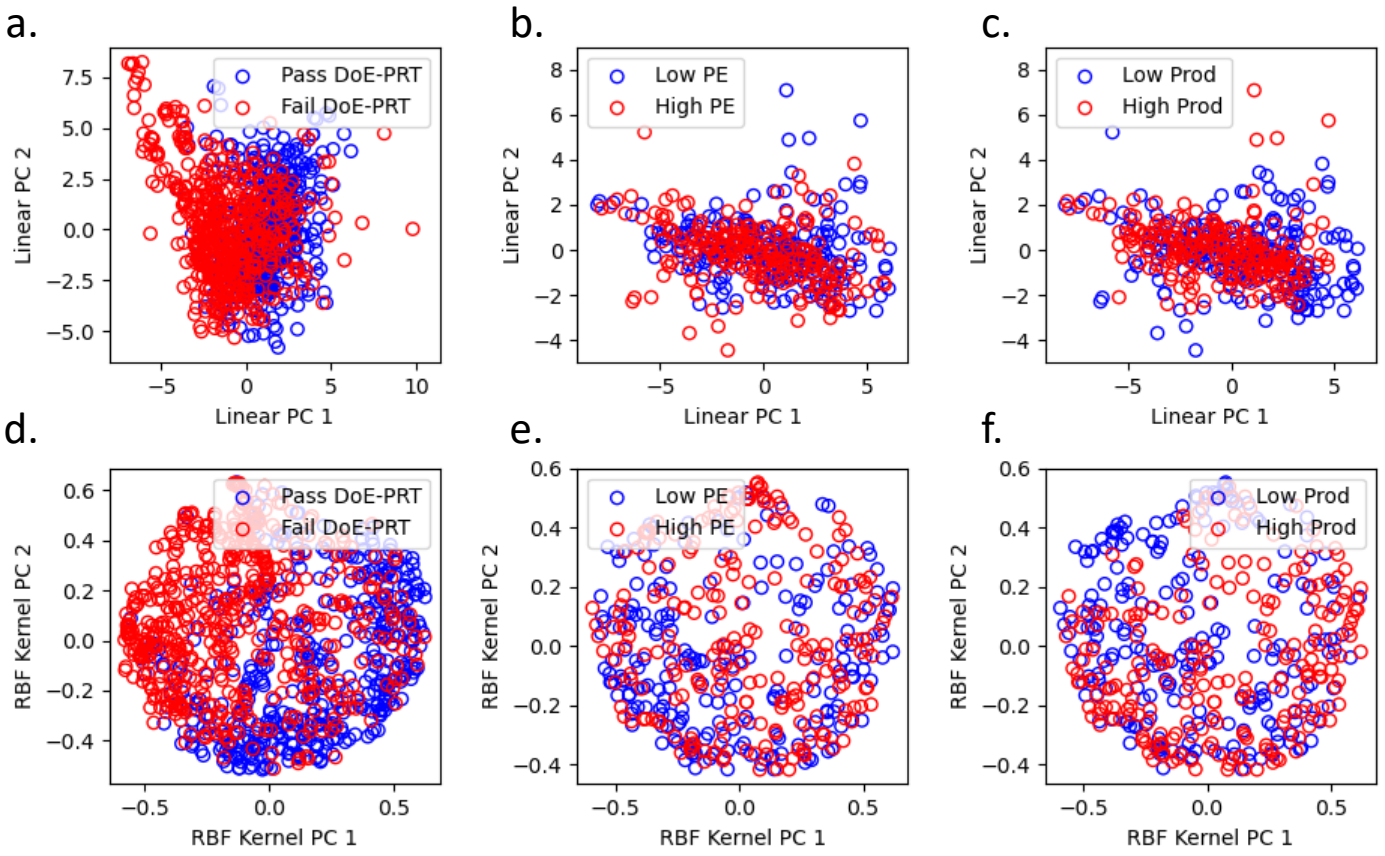


Figure 5.5 Scatter plots of first two principal components from principal component analysis (a-c) and kernel principal component analysis (d-f) with a gamma of 0.1 using the rbf kernel. The PCAs were run on several sets: (a & d) MOFs which meet the DoE-PRT (blue) and do not meet the DoE-PRT (red), (b & e) MOFs with low PE (blue) and MOFs with high PE (red) divided by the median PE, and (c & f) MOFs with low productivity (red) and high productivity (blue) divided by the median productivity value.

The results of the PCA analysis led to the conclusion that equilibrium metrics are not easily relatable to the best parasitic energy or best productivity values for a material. All results presented do indicate that a strong correlation between these atomistic metrics and the materials ability to meet the DoE-PRT exists and that a predictive model is possible.

5.4.6. Decision Trees & Random Forest Modelling

The results of all prior analysis from the Univariate to the PCA, indicate the potential for a predictive model which could estimate a MOF's ability to meet the DoE-PRT. The main goal of this work was to find equilibrium properties which could be used to select high performing materials at an early stage of discovery, and to ideally have a model which provides interpretable results. Although deep neural networks can extract complex relationships, they are often not interpretable. As a result, the decision was made to explore the use of the Random Forest (RF) decision tree method. In a RF, a series of decision trees are fit to predict the provided target property, and the result is the average of all trees in the forest. The primary advantage of this model paradigm is the ability to extract interpretable rule sets used by the model for analysis.

Using all features listed in Table 5.2, a random forest model containing 300 decision trees with a maximum depth of 3 was fit. The 5-fold cross-validation accuracy of this model in predicting a material's ability to meet the DoE-PRT was 83.0%. This value, although only 3% higher than the best LDA models, showed that prediction of the target property was possible with a 33% improvement over a completely random model. As mentioned previously, RFs allow for the analysis of the features used by the model to make these predictions, and as such a ranking of feature importance can be generated.

A breakdown of the node compositions can be found in Figure 5.6, which shows the frequency of use of each feature at all three depths of the decision trees. In a decision tree, the nodes in the first depth (or decision layer) of the tree are considered most important as the first split contains the full set of MOFs used in the fitting. At all three depths, however, the two most important features are the N_2 b_{298} – the strong site fitted dual-site Langmuir parameters for N_2, and N_2 q_1 – the strong site saturation uptake for N_2. For both depth 1 and depth 2, these two features make up over 20% of the nodes at their depths, and 16% at depth 3. The consistency and high frequency of use of these two features in the prediction of the material's ability to meet the DoE-PRT indicates that these two properties are the most important in the feature set, with other features such as the CO_2 uptake and working capacity being ranked significantly lower. This result is opposite to the LDA results and implies that when an ensemble approach is taken to classification, the CO_2 adsorption metrics are no longer the most important.

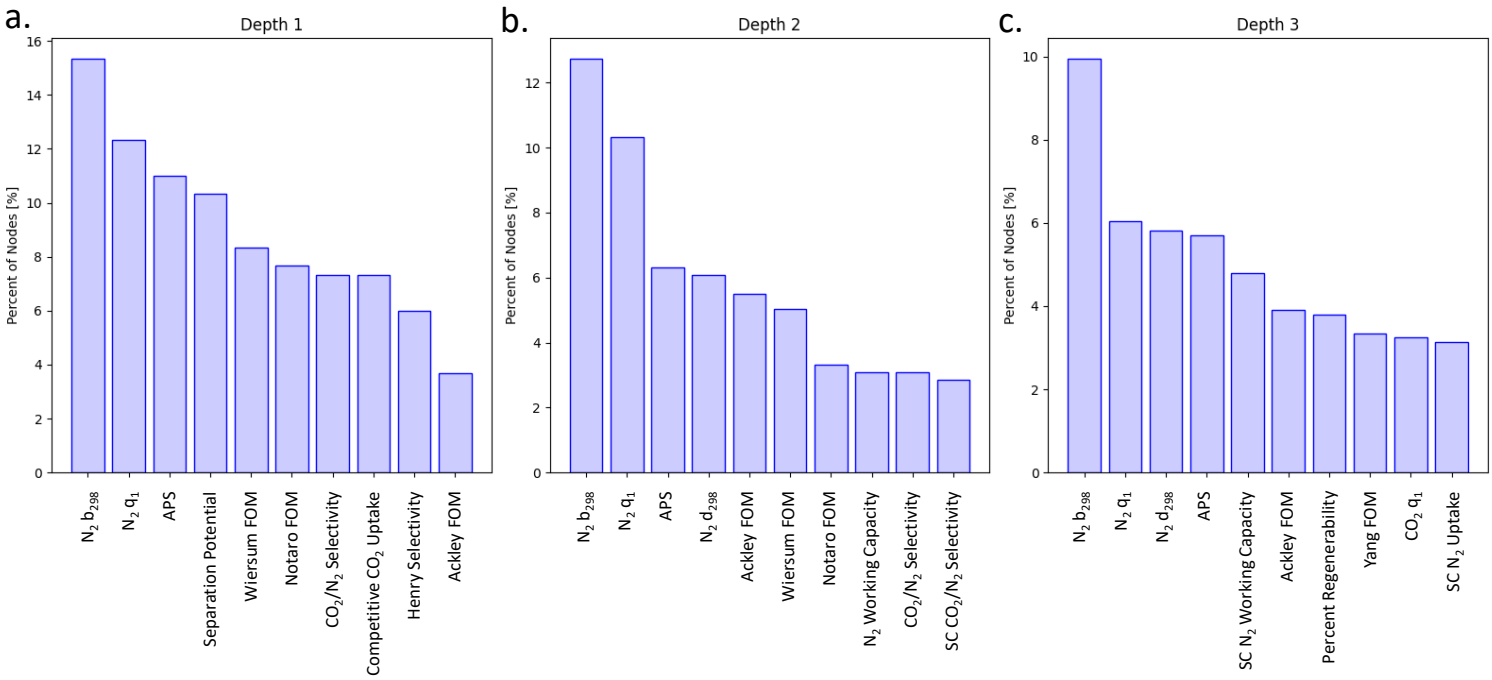


Figure 5.6 Bar plots demonstrating the percent of nodes showing the top 10 features found in the random forest classifier model, with the percentage of nodes relying on the features shown in the y-axis for (a) the first nodes in the trees, (b) the nodes in the second layer of the decision trees, and (c) the nodes in the third layer of the decision trees.

This hypothesis needs to be tested, however, before a conclusion can be drawn. To test this, RF models were fit using only specific features chosen to test this hypothesis. Models were fit using only the dual-site Langmuir parameters of either CO_2 or N_2, the results of which are shown in Table 5.3. When comparing models fit using the DSL parameters, the N_2 model performed 10% higher than the CO_2 model. Since the LDA results did not show good correlation with the DSL parameters, performing this comparison was incomplete. As such, additional RF models were fit using the single component and competitive uptakes and working capacities for either CO_2 or N_2. Using these features, the CO_2 model outperformed the N_2 model, however the CO_2 uptake model still performs roughly 5% worse than the N_2 DSL model. These results provide strong evidence towards confirming the conclusion that N_2 adsorption behaviour plays the most important role in determining the material's ability to meet the DoE-PRT, however the CO_2 adsorption behaviour may still be of some importance. This insight is significant since historically researchers only considered the CO_2 adsorption behaviour, only mentioning the N_2 behaviour when discussing the CO_2/N_2 selectivity, a metric shown in this work to completely fail at predicting the materials' ability to meet the DoE-PRT. For a full list of rankings of features from the RF model, see Appendix 5.6.

Table 5.3 Balanced accuracies of random forest classification models fit using alternate sets of features comparing CO_2 and N_2 specific features to the original random forest model for the DoE-PRT set, labeled All Features. Also included are the balanced accuracies from the best PE-set and Prod-set classification models.

Data Set	***Feature Set***	***5-fold Cross-validation Balanced Accuracy [%]***	
		CO_2	***N_2***
DoE-PRT	DSL Parameters	73.6	83.6
DoE-PRT	Uptakes and Working Capacities	79.0	77.1
DoE-PRT	All Features	83.0	
PE	All Features	64.7	
Prod	All Features	59.9	

Additional RF models were fit in an attempt to classify the ***PE*** and ***Prod sets*** into their respective subsets; however, the best model fitting returned a balanced accuracy of 64.7 %, and therefore were not considered successful. The balanced accuracy results from the best PE and Prod classification models can be found in Table 5.3. Finally, since the RF model was only able to improve upon the LDA results by 3%, a multi-layered perceptron neural-network model was fit to classify the MOFs in the ***DoE-PRT set*** and returned a balanced accuracy of 89.8 % based on 5-fold cross-validation, demonstrating the use of a more complex model as a screening tool in the early stages of MOF development. The confusion matrix for this model can be found in Appendix 5.7.

5.5. Conclusions

Over the course of the work presented in this chapter, 37 unique metrics were analyzed and vetted for their ability to predict the three key pressure swing adsorption performance metrics: a material's ability to meet the US Department of Energy's purity-recovery targets (DoE-PRT), the parasitic energy of running the capture unit and storing the capture gas to transport conditions, and the productivity of the material used to determine the size and complexity of the capture plant. The results of this analysis showed that none of the 37 metrics could be used to predict whether a material would have a favourable parasitic energy or productivity, however it was found that several metrics, including competitive CO_2 uptake and the N_2 isotherm parameters, were predictive of the material's ability to meet the DoE-PRT. Additionally, several metrics that are commonly used as benchmarks for high performance, such as the CO_2/N_2 selectivity, were not able to distinguish between MOFs which could or could not meet the DoE-PRT.

Although metrics such as the competitive CO_2 uptake may provide reasonable estimates of a MOF's ability to meet the 95/90-DoE-PRT, none of the metrics tested in this work appear to be effective at predicting whether a material will be high performing in terms of its parasitic energy or productivity. Thus, the results from this study suggest that full detailed process models are required to screen materials and identify high performers. The metrics commonly used in the field of materials research for pressure swing adsorption systems therefore do not provide a complete picture of a material's performance, and no single performance metric should be considered alone.

Interestingly, it was found that several metrics and combinations of metrics could be used to predict a material's ability to meet the DoE-PRT, the minimum requirement for the technology to be viable. This result was significant at the time of publication as it was the first time that process level performance had been quantitatively related to atomic level performance of a material. Further, the analysis of the random forest models revealed that the N_2 isotherm values play an important role in a material's ability to meet the DoE-PRT, a non-intuitive result as the focus of MOF research for PoC-CCS has historically prioritized metrics relating to CO_2. This insight was significant as the key focus of the materials discovery process was maximizing CO_2 adsorption, while only implicitly considering N_2 adsorption via the CO_2/N_2 selectivity, a metric which completely failed at predicting all three industrial PSA performance metrics. Although the random forest models performed reasonably well when relying solely on CO_2 metrics, when only the N_2 isotherm parameters were provided the models outperformed those fit using the CO_2 metrics. This indicated that although CO_2 adsorption is an important consideration in assessing a material for post-combustion PSA capture, the N_2 adsorption behaviour is of greater importance. At the time of submission of the manuscript, this result was contrary to the belief within the MOF community and showed the need to shift the focus of materials discovery from locating materials that adsorb large amounts of CO_2 to materials that adsorb little to no N_2. After initial submission of these results, but prior to publication, Snurr and co-workers published a similar study which performed a techno-economic analysis multi-scale screening of 369 MOFs.[32] The conclusions of that study led to a similar conclusion that N_2 adsorption behaviour, not CO_2 behaviour, is the most important metric in determining a material's ability to meet the DoE-PRT, in agreement with the conclusions discussed in this chapter.

5.6. References

1. Burns, T. D., Pai, K. N., Subraveti, S. G., Collins, S. P., Krykunov, M., Rajendran, A. & Woo, T. K. Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

2. Dzubak, A. L., Lin, L. C., Kim, J., Swisher, J. A., Poloni, R., Maximoff, S. N., Smit, B. & Gagliardi, L. Ab initio carbon capture in open-site metal-organic frameworks. *Nature Chemistry* **4**, 810–816 (2012).

3. Boyd, P. G., Chidambaram, A., García-Díez, E., Ireland, C. P., Daff, T. D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S. M., Maroto-Valer, M. M., Reimer, J. A., Navarro, J. A. R., Woo, T. K., Garcia, S., Stylianou, K. C. & Smit, B. Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature* **576**, 253–256 (2019).

4. Nugent, P., Giannopoulou, E. G., Burd, S. D., Elemento, O., Giannopoulou, E. G., Forrest, K., Pham, T., Ma, S., Space, B., Wojtas, L., Eddaoudi, M. & Zaworotko, M. J. Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

5. Liang, L., Liu, C., Jiang, F., Chen, Q., Zhang, L., Xue, H., Jiang, H. L., Qian, J., Yuan, D. & Hong, M. Carbon dioxide capture and conversion by an acid-base resistant metal-organic framework. *Nature Communications* **8**, (2017).

6. Jiang, J., Lu, Z., Zhang, M., Duan, J., Zhang, W., Pan, Y. & Bai, J. Higher symmetry multinuclear clusters of metal-organic frameworks for highly selective CO_2 capture. *Journal of the American Chemical Society* **140**, 17825–17829 (2018).

7. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

8. Chung, Y. G., Gómez-Gualdrón, D. A., Li, P., Leperi, K. T., Deria, P., Zhang, H., Vermeulen, N. A., Stoddart, J. F., You, F., Hupp, J. T., Farha, O. K. & Snurr, R. Q. In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances* **2**, e1600909 (2016).

9. Rege, S. & Yang, R. A simple parameter for selecting an adsorbent for gas separation by pressure swing adsorptions. *Separation Science and Technology* **36**, 3355–3365 (2001).

10. Carbon-Capture-Technology-Compendium-2020.

11. Boyd, P. G. Computational high throughput screening of metal organic frameworks for carbon dioxide capture and storage applications. (2015).

12. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

13. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (1990).

14. Rappé, A. K. K., Casewit, C. J. J., Colwell, K. S. S., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).

15. Campañá, C., Mussard, B., Woo, T. K., Campañá, C., Mussard, B. & Woo, T. K. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *Journal of Chemical Theory and Computation* **5**, 2866–2878 (2009).

16. Kresse, G. & Furthmüller, J. Vienna ab initio simulation package (VASP). (2001).

17. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **54**, 11169–11186 (1996).

18. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868 (1996).

19. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **105**, 9982–9985 (1996).

20. García-Sánchez, A., Ania, C. O., Parra, J. B., Dubbeldam, D., Vlugt, T. J. H., Krishna, R. & Calero, S. Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

21. Provost, B. An improved N_2 model for predicting gas adsorption in MOFs and using molecular simulation to aid in the interpretation of SSNMR spectra of MOFs. (2014).

22. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **149**, 134–141 (2012).

23. Fischer, M., Hoffmann, F. & Fröba, M. Metal–organic frameworks and related materials for hydrogen purification: Interplay of pore size and pore wall polarity. *RSC Advances* **2**, 4382 (2012).

24. Huck, J. M., Lin, L.-C., Berger, A. H., Shahrak, M. N., Martin, R. L., Bhown, A. S., Haranczyk, M., Reuter, K. & Smit, B. Evaluating different classes of porous materials for carbon capture. *Energy Environ. Sci.* **7**, 4132–4146 (2014).

25. Basdogan, Y., Sezginel, K. B. & Keskin, S. Identifying highly selective metal organic frameworks for CH_4 /H_2 separations using computational tools. *Industrial & Engineering Chemistry Research* **54**, 8479–8491 (2015).

26. Krishna, R. Screening metal–organic frameworks for mixture separations in fixed-bed adsorbers using a combined selectivity/capacity metric. *RSC Advances* **7**, 35724–35737 (2017).

27. Notaro, F., Mullhaupt, J. T., Wells, F. W. & Ackley, M. W. Adsorption process and system using multilayer adsorbent beds. **3**, (1997).

28. Ackley, M. W., Stewart, A. B., Henzler, G. W., Leavitt, F. W., Notaro, F. & Kane, M. S. PSA Apparatus and process using adsorbent mixtures. (2000).

29. Wiersum, A. D., Chang, J. S., Serre, C. & Llewellyn, P. L. An adsorbent performance indicator as a first step evaluation of novel sorbents for gas separations: Application to metal-organic frameworks. *Langmuir* **29**, 3301–3309 (2013).

30. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M. & Duchesnay, É. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

32. Leperi, K. T., Chung, Y. G., You, F. & Snurr, R. Q. Development of a general evaluation metric for rapid screening of adsorbent materials for postcombustion CO_2 capture. *ACS Sustainable Chemistry and Engineering* **7**, 11529–11539 (2019).

5.7. Appendix 5.1: Equations of Composite Adsorption Metrics

$$\text{Henry Selectivity } [-] = \frac{K_{H,CO_2}}{K_{H,N_2}} = \frac{Q_{1,CO_2} b_{298,CO_2} + Q_{2,CO_2} d_{298,CO_2}}{Q_{1,N_2} b_{298,N_2} + Q_{2,N_2} d_{298,N_2}} \quad \text{(S 5.1)}$$

$$\text{Huck's PE } \left[\frac{\text{kWh}}{\text{tonne } CO_2}\right] = Q_{thermal} + W_{compression} = \left[\frac{C_p m_{adsorbent} \Delta T}{m_{CO_2}} + \frac{\Delta h_{CO_2} \Delta\sigma_{CO_2} + \Delta h_{N_2} \Delta\sigma_{N_2}}{m_{CO_2}}\right] + \left[E_{compressor} + E_{pump}\right] \quad \text{(S 5.2)}$$

$$\text{Adsorbent performance score (APS) } \left[\frac{\text{mmol}}{\text{g}}\right] = \Delta\sigma_{ads,CO_2} Sel_{CO_2/N_2,ads} \quad \text{(S 5.3)}$$

$$\text{Percent Regenerability } [\%] = 100\% \frac{\Delta\sigma_{CO_2}}{n_{CO_2}^{mixture}} \quad \text{(S 5.4)}$$

$$\text{Separation Potential } \left[\frac{\text{mmol}}{\text{g}}\right] = Q_A \frac{y_A}{1 - y_A} - Q_B \quad \text{(S 5.5)}$$

$$\text{Sorbent Selection Parameter } \left[\frac{\text{mmol}}{\text{g}}\right] = \frac{\Delta n_{CO_2}}{\Delta n_{N_2}} \left(\frac{n_{sat,CO_2} b_{CO_2}}{n_{sat,N_2} b_{N_2}}\right) \quad \text{(S 5.6)}$$

$$\text{Notaro's FOM } \left[\frac{\text{mmol}}{\text{g}}\right] = \Delta\sigma_{CO_2} \frac{Sel^{\,2}_{CO_2/N_2,ads}}{Sel_{CO_2/N_2,des}} \quad \text{(S 5.7)}$$

$$\text{Ackley's FOM } \left[\frac{\text{mmol}}{\text{g}}\right] = \Delta\sigma_{ads,CO_2} \frac{Sel_{CO_2/N_2,ads}}{Sel_{CO_2/N_2,des}} \quad \text{(S 5.8)}$$

$$\text{Yang's FOM } [-] = Sel_{CO_2/N_2,adsorption} \frac{\Delta\sigma_{CO_2}}{\Delta\sigma_{N_2}} \quad \text{(S 5.9)}$$

$$\text{Wiersum's FOM } \left[\frac{\text{mol}^3}{\text{J kg}}\right] = 1000 \frac{\sqrt{Sel_{CO_2/N_2,ads} - 1} \times \Delta\sigma_{CO_2}}{\Delta H_{CO_2}^{ads}} \quad \text{(S 5.10)}$$

5.8. Appendix 5.2: Isotherm Plots for PE and Prod sets

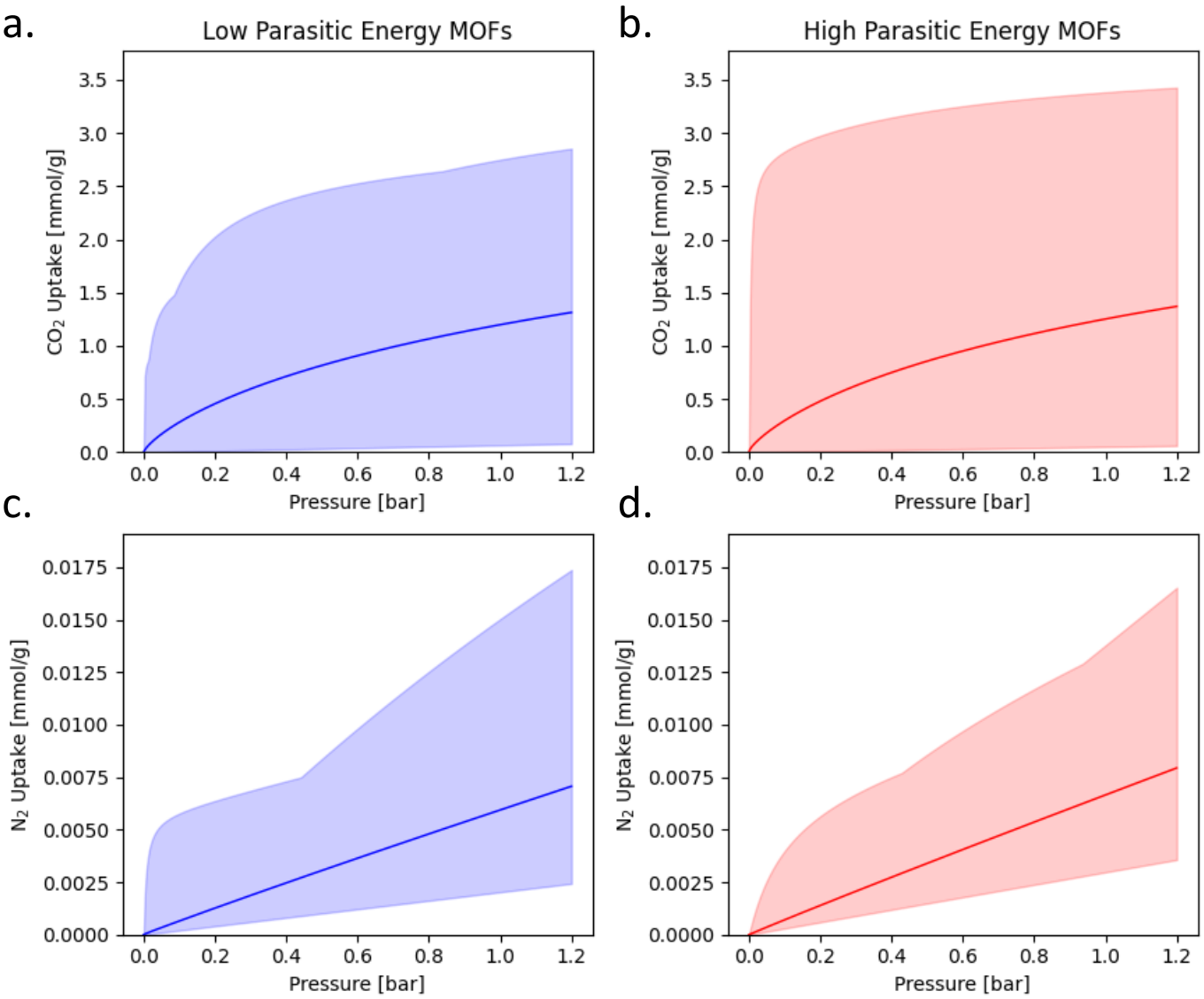


Figure A5.1. Plots showing the range of the isotherms, denoted using a shaded area, and the average isotherm shown in a solid line for (a) CO_2 isotherms of the MOFs in the low PE subset shown in blue, (b) the CO_2 isotherms of the MOFs in the high PE subset shown in red, (c) N_2 isotherms of the MOFs in the low PE subset shown in blue, and (d) the N_2 isotherms of the MOFs win the high PE subset shown in red.

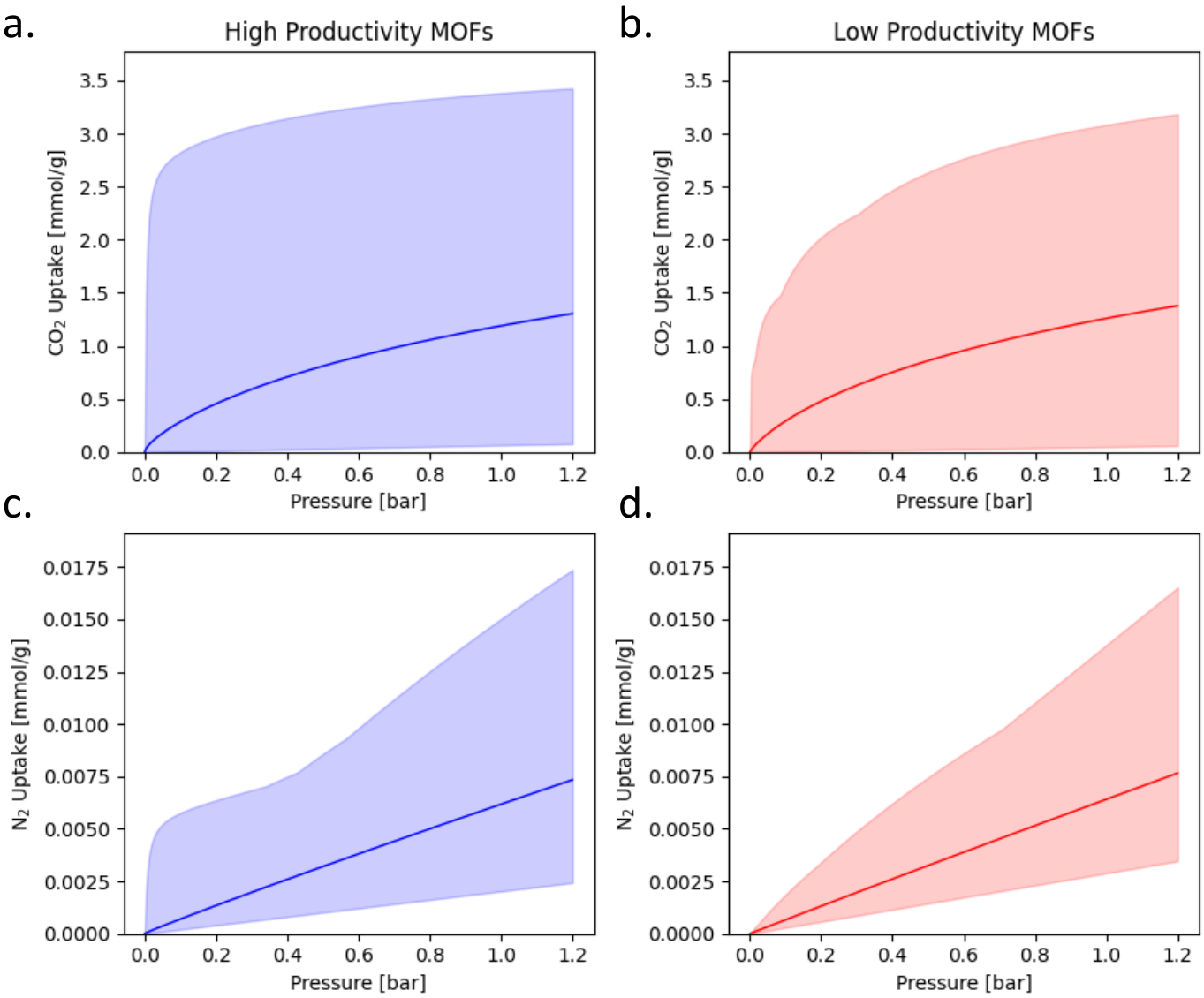


Figure A5.2. Plots showing the range of the isotherms, denoted using a shaded area, and the average isotherm shown in a solid line for (a) CO_2 isotherms of the MOFs in the high Prod subset shown in blue, (b) the CO_2 isotherms of the MOFs in the low Prod subset shown in red, (c) N_2 isotherms of the MOFs in the high Prod subset shown in blue, and (d) the N_2 isotherms of the MOFs win the low Prod subset shown in red.

5.9. Appendix 5.3: All Univariate Plots

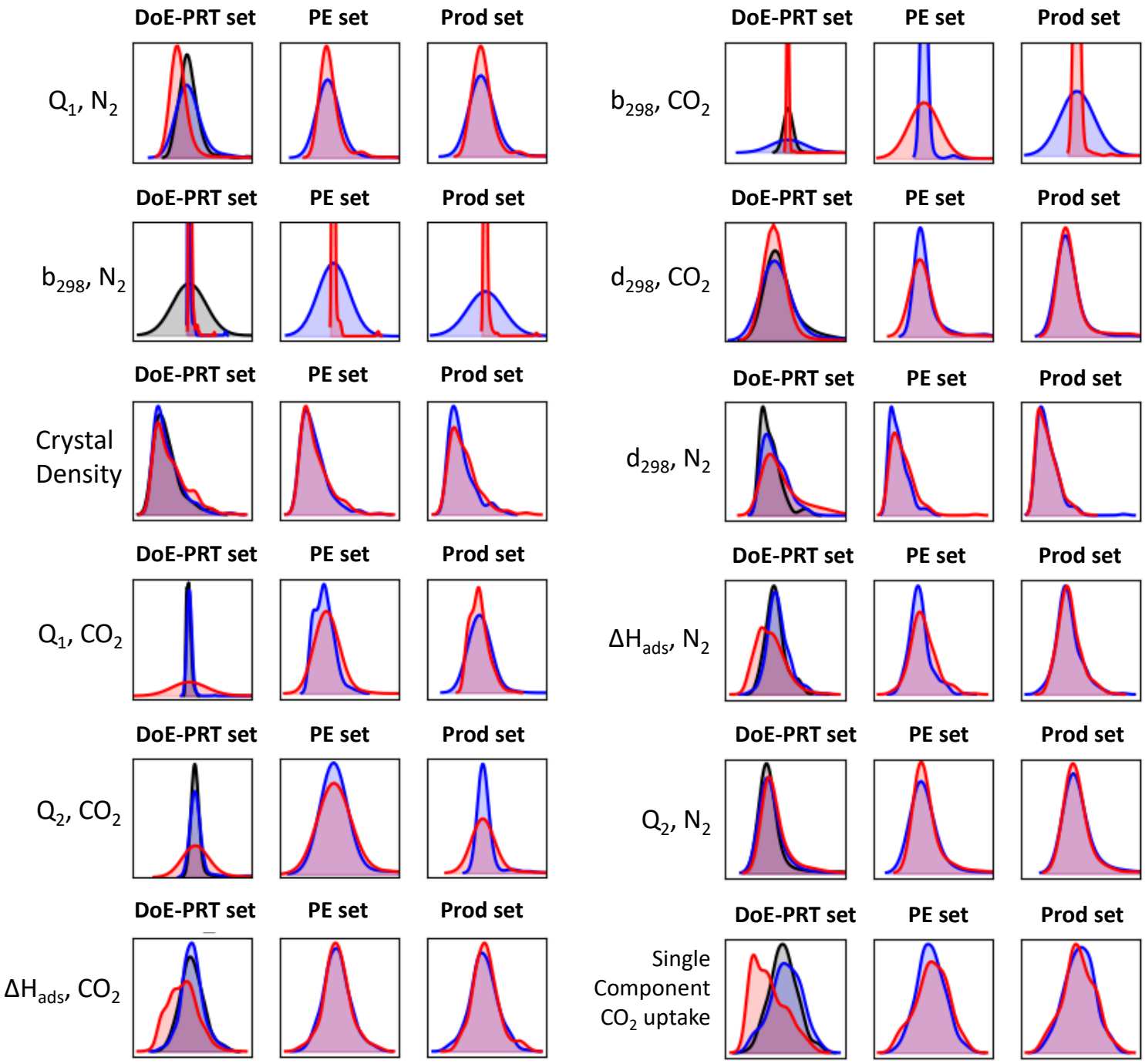


Figure A5.3. Univariate analysis performed on the *DoE-PRT*, *PE*, and *Prod sets* for the first 12 conventional metrics.

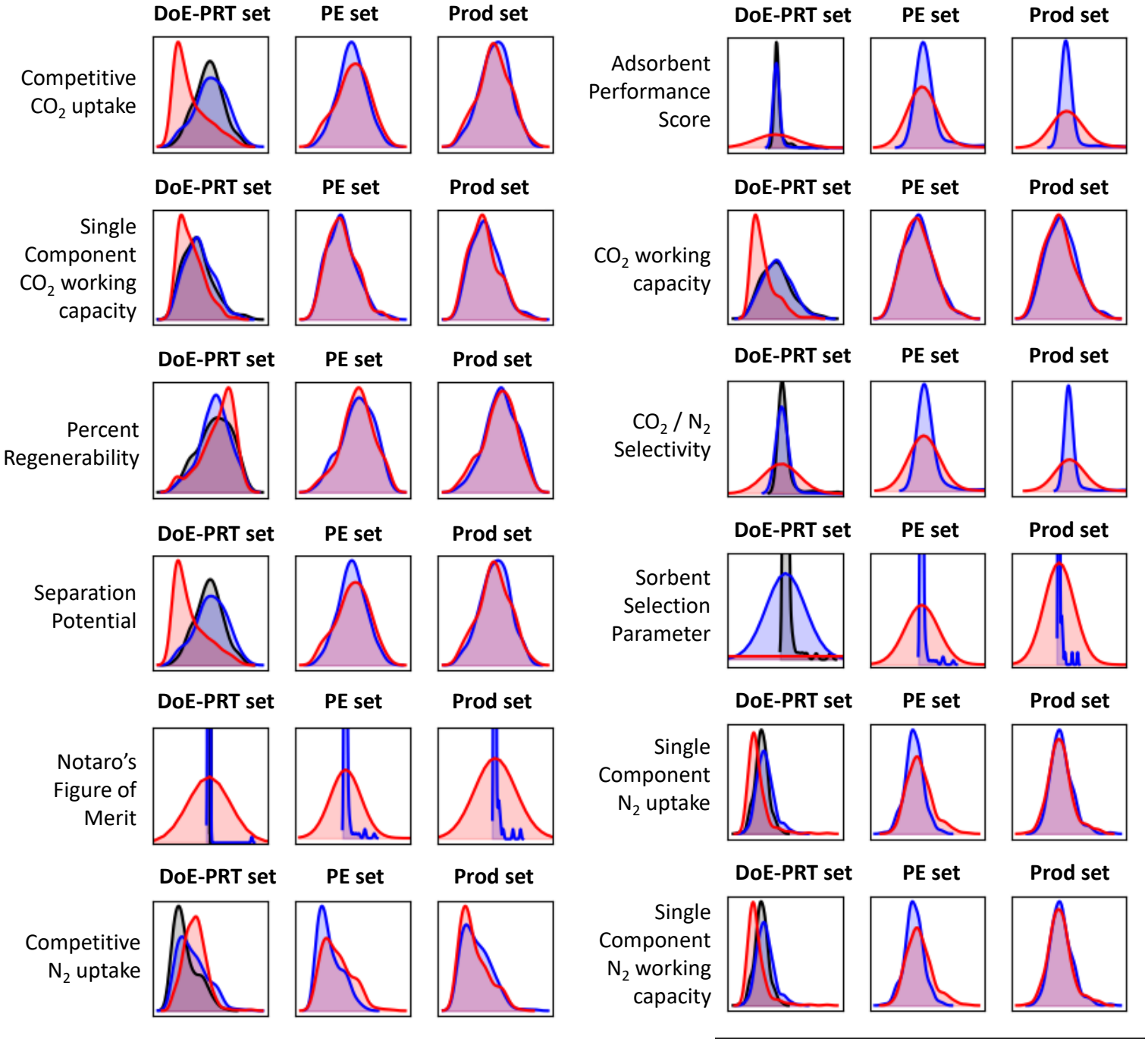


Figure A5.4. Univariate analysis performed on the *DoE-PRT*, *PE*, and *Prod sets* for 12 conventional metrics.

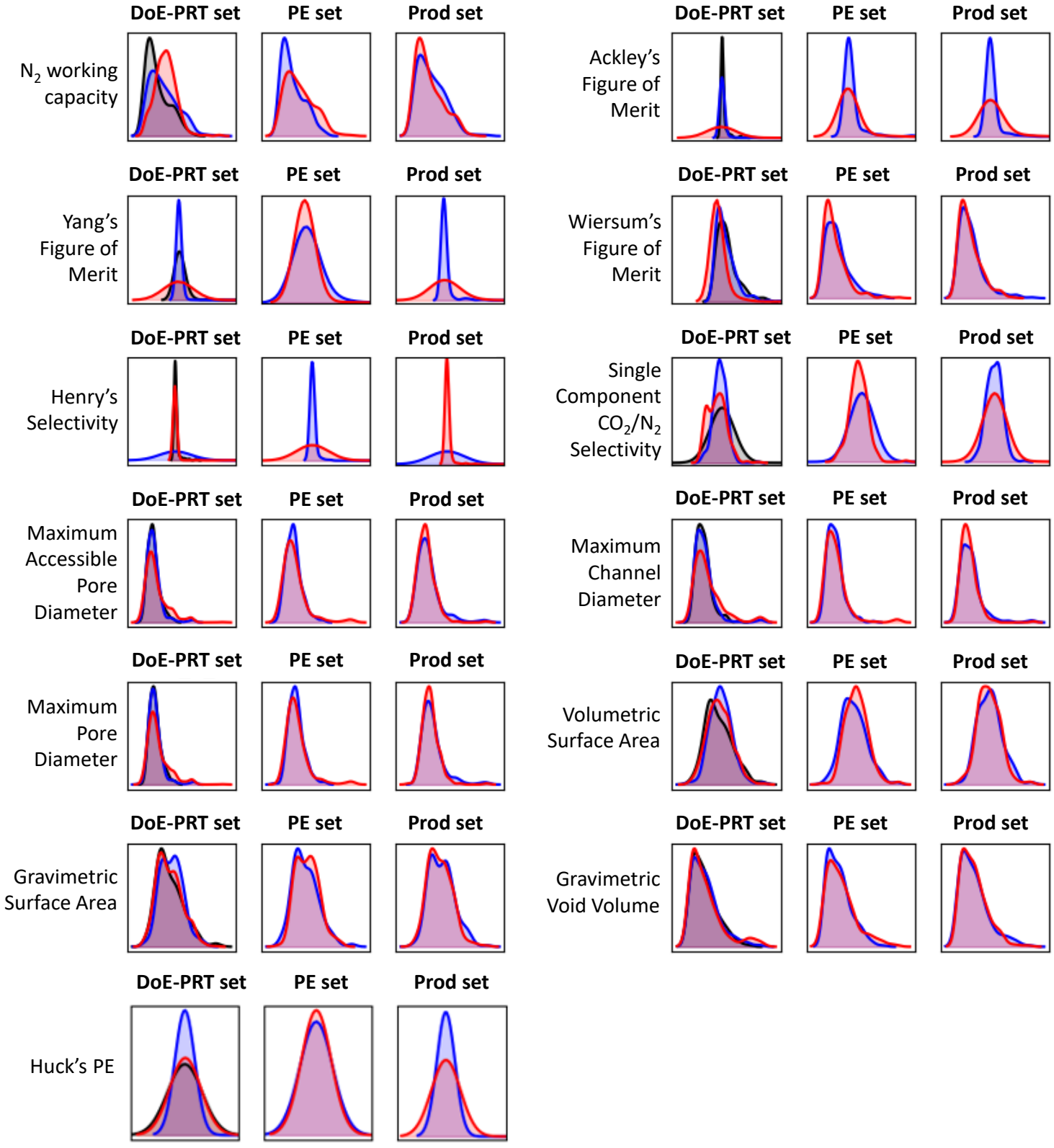


Figure A5.5. Univariate analysis performed on the *DoE-PRT*, *PE*, and *Prod sets* for the 13 conventional metrics.

5.10. Appendix 5.4: LDA Heatmaps for PE and Prod sets

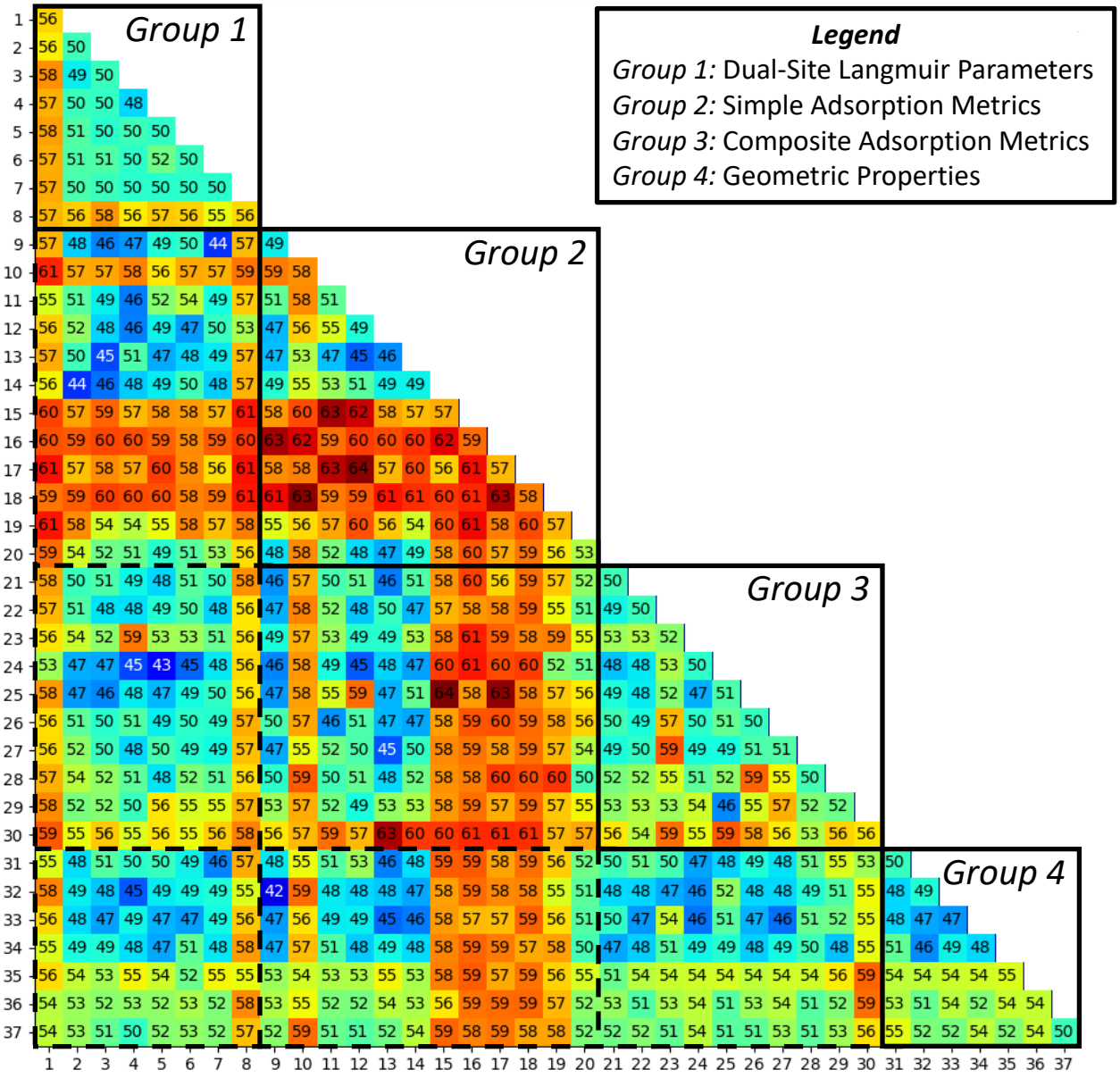


Figure A5.6. A heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional linear discriminant analysis models fit classify MOFs in the *PE set* according to their subsets, with the diagonal representing the balanced accuracy for the 1-dimensional linear discriminant analysis of each feature. The feature IDs on the axes correspond to the features in Table 5.1 and can be separated into four distinct groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties.

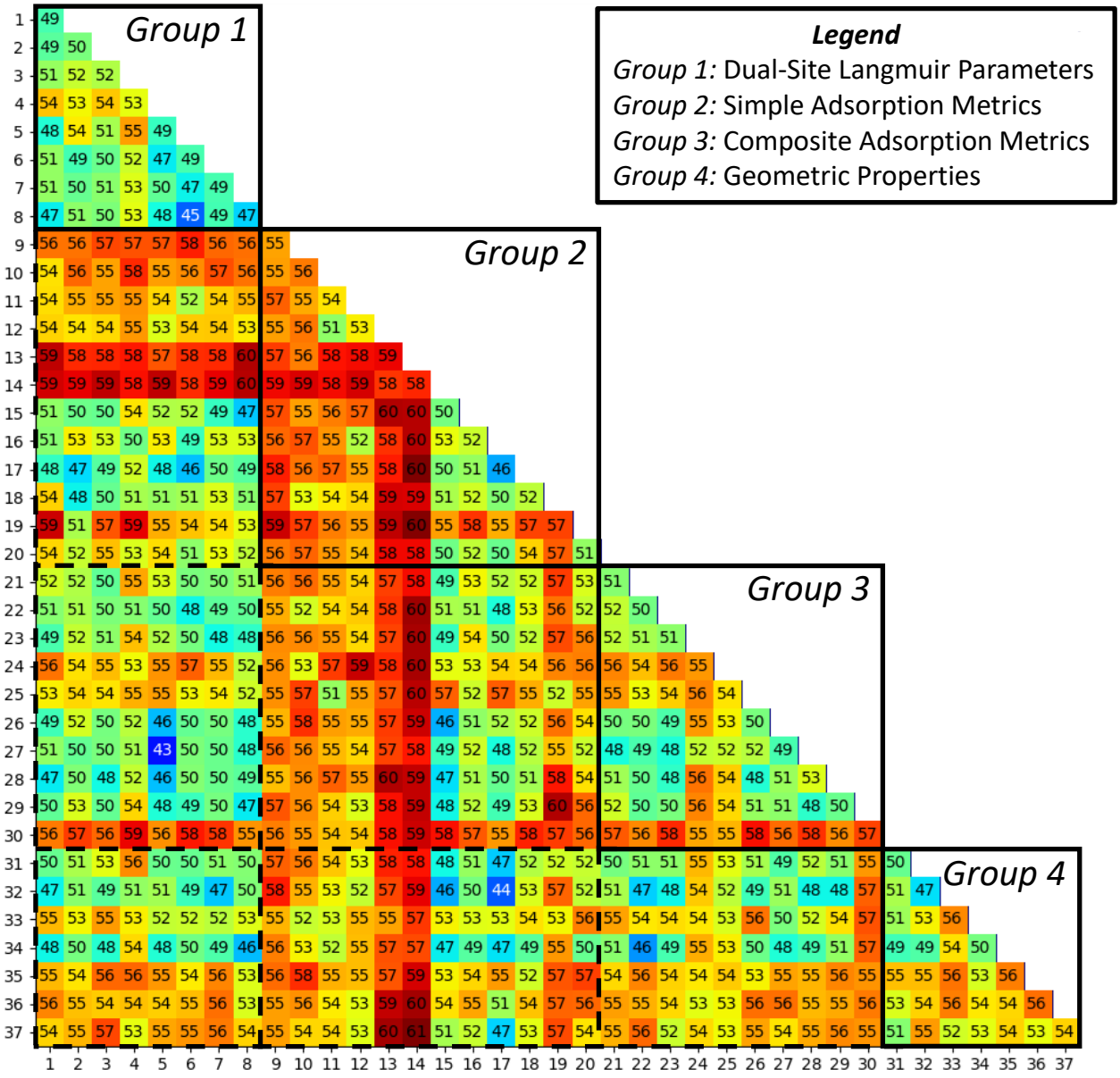


Figure A5.7. A heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional linear discriminant analysis models fit to classify MOFs in the *Prod set* according to their subsets, with the diagonal representing the balanced accuracy for the 1-dimensional linear discriminant analysis of each feature. The feature IDs on the axes correspond to the features in Table 5.1 and can be separated into four distinct groups: group 1 consists of the fitted dual-site Langmuir isotherm parameters, group 2 consists of simple adsorption metrics, group 3 consists of advanced composite adsorption metrics, and group 4 consists of geometric properties.

5.11. Appendix 5.5: Additional PCA Plots

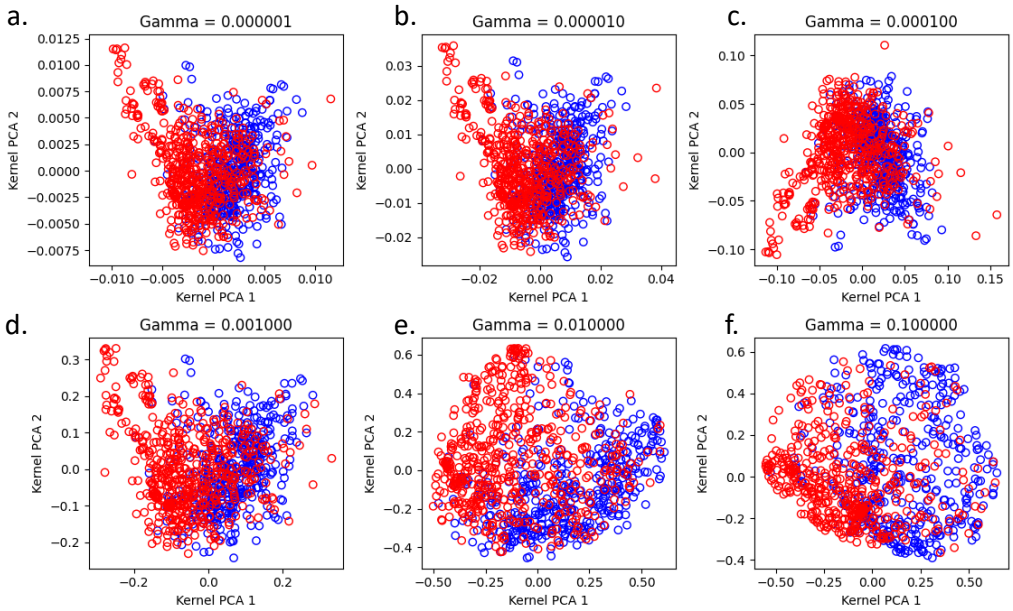


Figure A5.8. Results from modifying the RBF gamma parameter for the Kernel Principal Component Analysis separating the subsets in the *DoE-PRT set*, showing results for gamma values of (a) 10^{-6}, (b) 10^{-5}, (c) 10^{-4}, (d) 10^{-3}, (e) 10^{-2}, and (f) 10^{-1}, with MOFs in the *Pass set* shown in blue, and MOFs in the *Fail set* shown in red.

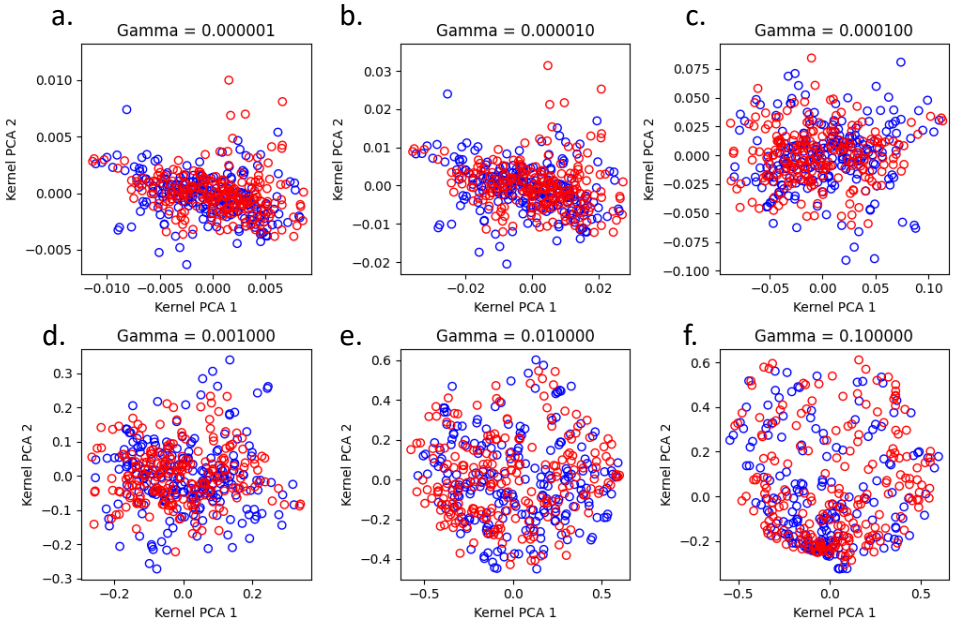


Figure A5.9. Results from modifying the RBF gamma parameter for the Kernel Principal Component Analysis separating the subsets in the *PE set*, showing results for gamma values of (a) 10^{-6}, (b) 10^{-5}, (c) 10^{-4}, (d) 10^{-3}, (e) 10^{-2}, and (f) 10^{-1}, with MOFs in the *Low PE set* shown in blue, and MOFs in the *High PE set* shown in red.

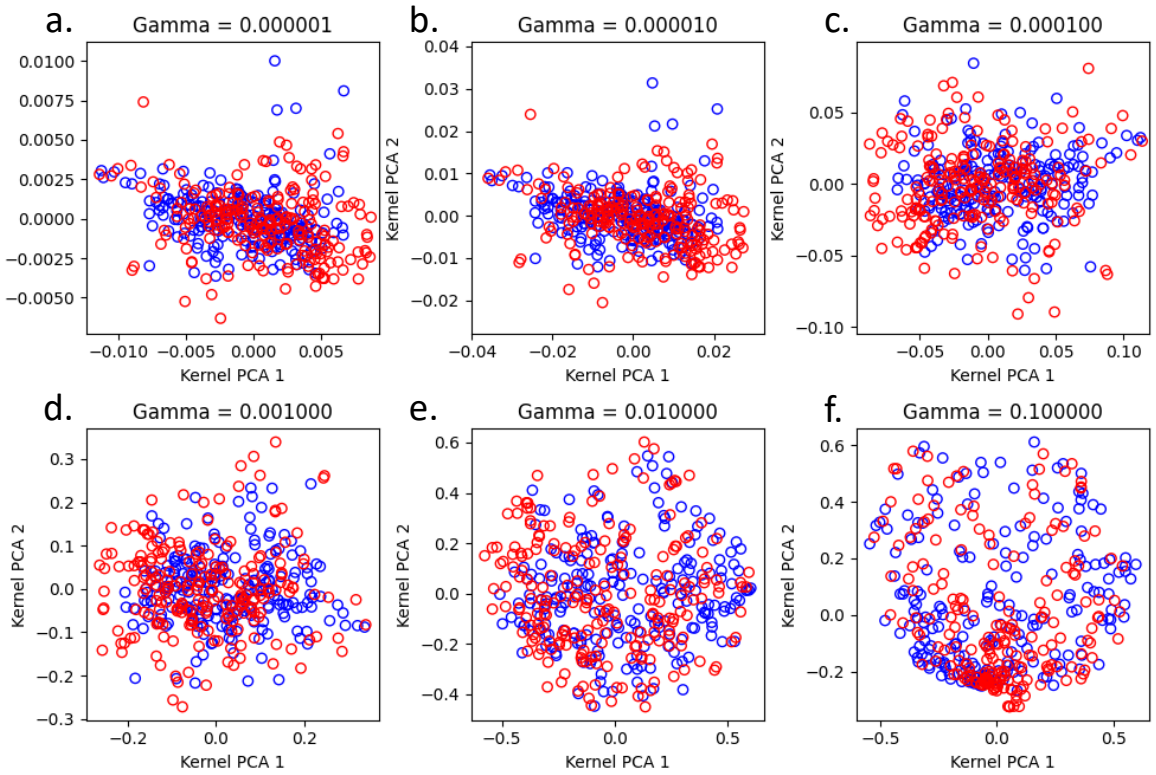


Figure A5.10. Results from modifying the RBF gamma parameter for the Kernel Principal Component Analysis separating the subsets in the *Prod set*, showing results for gamma values of (a) 10^{-6}, (b) 10^{-5}, (c) 10^{-4}, (d) 10^{-3}, (e) 10^{-2}, and (f) 10^{-1}, with MOFs in the *High Prod set* shown in blue, and MOFs in the *Low Prod set* shown in red.

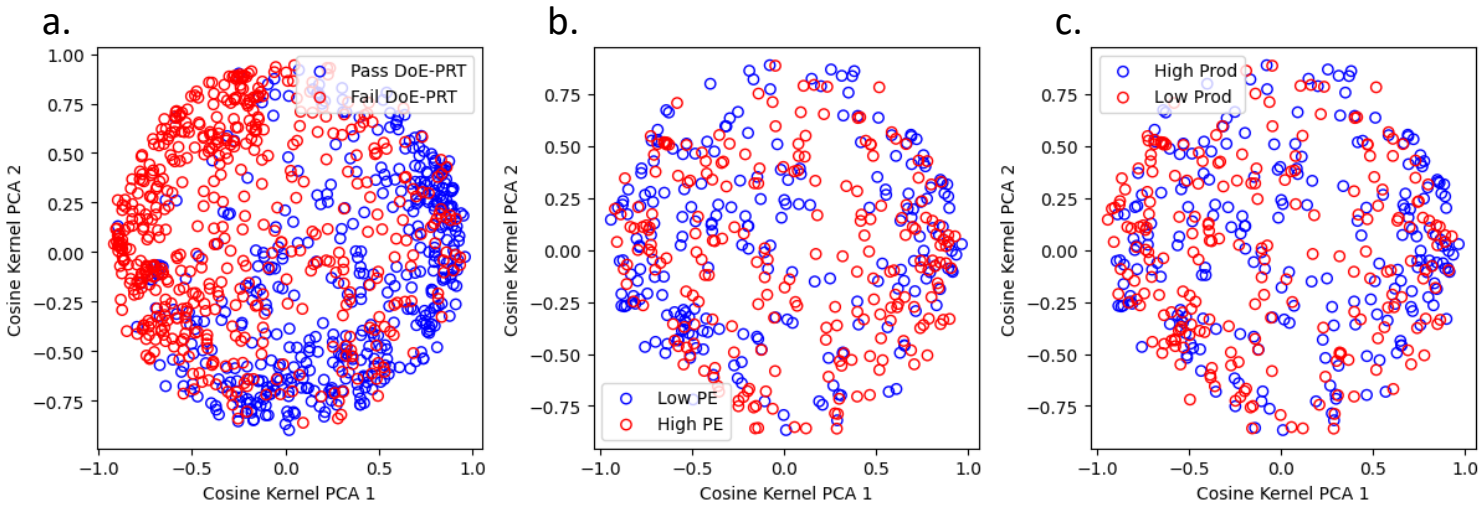


Figure A5.11. Results from using the cosine kernel in the principal component analysis, for the (a) *DoE-PRT*, (b) *PE*, and (c) *Prod sets*.

5.12. Appendix 5.6: Random Forest Feature Rankings

Table A5.1. Random forest feature ranking, ordered by the first decision layer ranking and determined by the prevalence of the feature in the decision nodes. Included in the table is the overall ranking, calculated using the sum of instances across all three decision layers.

	First Layer Ranking	*Second Layer Ranking*	*Third Layer Ranking*	*Overall Ranking*
b_{298}, N_2	1	1	1	1
Q_2, N_2	2	2	2	2
Adsorbent Performance Score	3	3	4	3
Separation Potential	4	23	36	13
Wiersum's FOM	5	6	19	6
Notaro's FOM	6	7	13	7
CO_2/N_2 Selectivity	7	9	21	9
Competitive CO_2 uptake	8	24	33	20
Henry Selectivity	9	16	20	10
Ackley's FOM	10	5	6	5
Single component CO_2 uptake	11	26	26	25
Single component N_2 Working Capacity	12	13	5	8
Yang's FOM	13	12	8	11
b_{298}, CO_2	14	15	15	16
Single component N_2 uptake	15	11	10	12
Single component CO_2/N_2 selectivity	16	10	11	15
N_2 working capacity	17	8	14	14
CO_2 working capacity	18	25	22	24
Q_2, N_2	19	14	16	18
Δh_{ads} CO_2	20	28	32	32
d_{298}, N_2	Not Used	4	3	4
Δh_{ads} N_2	Not Used	17	24	23
Q_1, CO_2	Not Used	18	9	19
Sorbent Selection Parameter	Not Used	19	37	27
Competitive N_2 uptake	Not Used	20	17	22
Percent Regenerability	Not Used	21	7	17
d_{298}, CO_2	Not Used	22	12	21
Q_2, CO_2	Not Used	27	34	31
Maximum pore diameter	Not Used	29	27	28
Crystal density	Not Used	30	31	33
Maximum channel diameter	Not Used	31	25	29
Gravimetric surface area	Not Used	32	28	34
Maximum accessible pore diameter	Not Used	33	35	36
Single component CO_2 working capacity	Not Used	34	29	35
Volumetric surface area	Not Used	35	23	30
Huck's PE	Not Used	36	18	26
Gravimetric void volume	Not Used	37	30	37

5.13. Appendix 5.7: Neural Network Confusion Matrix

Figure A.12. Confusion matrix showing the 5-fold cross-validation results for an MLP model fit to distinguish between two sets of MOFs which pass or fail the DoE-PRT in the *DoE-PRT set*. This neural network was composed of 4 hidden layers with 50 nodes per layer and used a learning rate of 0.001.

		Actual DoE-PRT	
		Pass	*Fail*
Predicted DoE-PRT	*Pass*	275	2
	Fail	1229	78

6. Chapter 6: N_2 Binding Sites

The work discussed in this chapter expands on the analysis performed in Chapter 5 and will be used as the basis for a manuscript that is in preparation for publication. The work presented in this chapter is entirely my own.

6.1. Abstract

The results obtained from the pores-to-process large scale screening of MOFs showed that N_2 adsorption behaviour[1] played a vital role in determining a material's ability to meet the US Department of Energy's Purity-Recovery targets (DoE-PRT) of 95% purity of capture CO_2 and 90% recovery of CO_2 from the flue gas.[2] The relationship between N_2 adsorption behaviour and process performance was explored in more depth through the study of CO_2 and N_2 binding sites within the pores of the material. Binding sites were calculated using the Guest Atom Localization Algorithm (GALA), and a comparison was performed between the single component (pure gas) binding sites to the binary component (gas mixture) binding sites. This comparison was performed using *Tanimoto coefficients* and uptake ratios to study the competition between the two guests at important binding sites within the materials. It was found that MOFs in which N_2 binding sites were displaced by CO_2 frequently saw a decrease in N_2 uptake capacity. This displacement and subsequent decrease in uptake was determined to be an important feature driving a material's ability to meet the DoE-PRT, equating atomistic behaviour to industrial process performance for the first time.

6.2. Introduction

Relating atomistic properties of a metal-organic framework (MOF) to the performance in an industrial pressure swing adsorption system (PSA) remains a challenge. Although attempts have been made to relate the behaviour of gas molecules within the pores of the MOFs to PSA performance,[3–5] the disconnect which exists between atomistic studies of materials for gas separations and process level engineering was a key barrier in understanding the underlying mechanisms driving high performance.

The results presented in Chapters 4 and 5 provided a unique opportunity to study the mechanisms of gas adsorption within the pores of MOFs. The conclusions discussed in Chapter 5 described the N_2 isotherm behaviour's role in a MOF's ability to meet the US Department of Energy's Purity-Recovery (DoE-PRT) targets for a post-combustion carbon capture and storage (PoC-CCS) system – that is a 95% purity of the captured CO_2 and recovery of 90% of the CO_2 from the flue gas.[2] The results

indicated that the N_2 adsorption behaviour, rather than the CO_2 adsorption behaviour, is the most important factor in a material's ability to meet the DoE-PRT.[1,6] This is a non-intuitive result which goes against the current belief held by most MOF chemists.[5,7–13] This conclusion therefore suggests that studying the behaviour of N_2 molecules within the pores of the MOFs may provide valuable insights into the mechanisms driving PSA performance. Using the Guest Atom Localization Algorithm discussed in Chapter 3, it was possible to study the nitrogen binding sites within the pores of MOFs and search for relevant trends in the binding environments. This idea led to the *first hypothesis* discussed in this chapter: *The N_2 binding sites play an important role in determining a MOF's ability to meet the DoE-PRT.*

Another insight from the linear discriminant analysis (LDA) discussed in Chapter 5 was the importance of comparing the adsorption behaviour of a pure gas (single component) to the adsorption behaviour of a binary gas mixture (binary component). The presence of another gas in the pores of a MOF can often lead to competition for binding sites and plays an important role in gas uptakes for both guests. As such the goal of the work discussed in this chapter was to study the effects of competition on CO_2 and N_2 binding sites, and their relationship to a MOF's ability to meet the DoE-PRT.

6.3. Methods

6.3.1. Data Sets

The data sets used in this analysis were drawn from the results of our previous large-scale screening of the Computation-Ready Experimental (CoRE) MOF database,[14] discussed in Chapter 4. This multi-scale screening relied on single component isotherms generated by Grand Canonical Monte Carlo (GCMC) simulations. These isotherms were then used as inputs into a sophisticated Pressure Swing Adsorption (PSA) simulator coupled to a custom Genetic Algorithm (GA) to optimize the process conditions of the separation.

For this analysis, a set of 704 MOFs were selected randomly from the MOFs which underwent full PSA optimizations. Of the MOFs selected, 423 were able to meet the DoE-PRT and 281 were not. The set of MOFs which do meet the DoE-PRT was larger than the set of MOFs which did not since MOFs with the potential for high performance were prioritized for optimization in the initial study. This decision to prioritize MOFs was made due to the high computational cost of the optimizations, since a single material could take up to 5 days of high-performance computing time to fully optimize.

6.3.2. Grand Canonical Monte Carlo Simulations

To analyze the binding environments of the guest molecules, grand canonical Monte Carlo (GCMC) simulations were performed to generate probability distributions, using a GCMC code written in-house[15] based on the DL_POLY classic code.[16] To effectively study the binding environments of N_2 in the presence of CO_2, simulations were performed which contained both molecules and are known as *binary component* (BC) simulations. *Single component* (SC) simulations, or simulations containing only CO_2 or N_2, were also performed to be used as a point of comparison in studying the competitive binding between the guests. For this study, both binary component CO_2/N_2 simulations with a 15:85 mole ratio, as well as single component CO_2 and N_2 simulations were performed for each MOF. The binary simulations were performed at 298 Kelvin and 1 bar of total pressure, and the single component simulations were run at 298 Kelvin and 0.15 bar and 0.85 bar for CO_2 and N_2, respectively. These conditions recreate the dry flue gas compositions from a coal-fired powerplant equipped with a post-combustion carbon capture and storage (PoC-CCS) system.[17]

The guest molecules in this simulation were modelled using the Garcia-Sanchez[18] and the N_2-NIMF[19] (Nitrogen in Metal-Organic Frameworks) parameters for CO_2 and N_2, respectfully. The dispersion interactions were modelled using a Lennard-Jones potential, with the framework atoms parameterized using a mixture of the Dreiding Force-Field[20] for available atom types supplemented with the Universal Force-Field (UFF).[21] The Coulomb interactions were modelled using atomic charges calculated with the REPEAT method[22] fit to quantum mechanical electrostatic potentials (ESP). QM ESPs were calculated using the Vienna Ab Initio Simulation Package (VASP)[23,24] using the PBE functional[25,26] and an energy cutoff of 400 eV. To ensure sufficient sampling was performed to generate high quality probability plots, 20,000 GCMC equilibration cycles were used with an excessive number of 10,000,000 GCMC production steps. This number of production steps ensured that the probability distributions were reproducible and well converged.

The probability distributions generated by the GCMC simulations were represented on a 3-dimensional grid. Every time a guest atom is sampled within a grid-point, that information is saved and added to the corresponding occupancy value. The occupancy is calculated using equation 6.1 and is defined as the fraction of steps in the simulation that the grid-point was occupied by the guest atom divided by the volume of the grid-point. The occupancy value can also be thought of as the probability of finding the guest atom within the bin per unit volume. For these simulations, we used a grid-spacing of

0.15 Å resulting in a volume of 3.3 x 10^{-3} Å^3 per grid-point. GCMC generates a unique probability plot for each atom type in a guest molecule, therefore a simulation containing both CO_2 and N_2 will generate three probability plots: a Carbon, an Oxygen, and a Nitrogen plot. This analysis involved running 3 unique simulations for each MOF: a binary component simulation containing CO_2 and N_2, a single component simulation containing only CO_2, and a single component simulation containing only N_2. As a result, six probability plots were generated for each MOF.

$$Occupancy_i = \frac{N_i}{N_{Total} V_i} \quad (6.1)$$

where:

N_i = the number of steps guest occupied grid point i

V_i = The volume of grid-point i

N_{Total} = Total number of production steps

6.3.3. Guest Atom Localization Algorithm

To generate binding sites based on the GCMC probability distributions, the Guest Atom Localization Algorithm (GALA) discussed in Chapter 3 was used. This algorithm, which locates binding sites by smoothing the probability distribution and fitting the guests to the remaining peaks, has been shown to accurately reproduce the binding sites seen in experimental MOFs. Binding sites were calculated for both guest molecules present in the simulations, CO_2 and N_2, using parameters optimized to reproduce experimental CO_2 binding sites, discussed in Chapter 3. For this analysis, binding sites were generated from single component CO_2 and N_2 GCMC simulations as well as from binary CO_2/N_2 simulations. The conditions used in these simulations are outlined above.

6.3.4. Tanimoto Coefficient

Once probability plots were generated from single component CO_2 and N_2 and binary component CO_2/N_2 simulations, the plots were compared using the *Tanimoto* coefficient[27,28] given in equation 6.2. The *Tanimoto* is a common tool in chemo-informatics for assessing the similarity of two substances based on a molecular fingerprint and returns a value between 0 and 1. A *Tanimoto* coefficient of 0 indicates the two substances are completely dissimilar, and 1 indicates that they are identical. GALA relies on a *Tanimoto* similarity metric to compare two concurrently run probability plots

as a check for convergence in the simulation, as such, the *Tanimoto* has proven to be a useful metric in comparing probability distributions.

The choice to use a *Tanimoto coefficient* to compare the probability plots was made based on the assumption that plots with low similarity indicate a change in binding environments within the pores of the MOF. Conversely, plots with a high *Tanimoto coefficient* would indicate little to no change in binding environments occurred between the two states. To study the change in binding environments of CO_2 in the presence of N_2, a *Tanimoto* was calculated to compare the single and binary component probability distributions for both guests. Since CO_2 is a hetero-atomic molecule, two distributions were generated for each CO_2, whereas only a single probability distribution was generated for N_2. As a result, three *Tanimoto* values were generated, one for each atom type present in the simulations, for each MOF in the set.

$$Tanimoto = \frac{a \cdot b}{(a \cdot a) + (b \cdot b) - (a \cdot b)} \tag{6.2}$$

where:

a = probability plot 1 vector

b = probability plot 2 vector

6.3.5. Uptake Ratios

A secondary comparison was performed on the distributions once the results from the *Tanimoto* coefficients were analyzed. For each MOF in the dataset, an uptake ratio was calculated comparing the single (SC) and binary component (BC) uptakes for both guests. The goal was to study whether any *displacement* of binding sites in the MOFs corresponded to changes in guest uptake. These ratios were calculated using equation 6.3 based on the uptakes from the initial GCMC simulations.

$$Uptake\ Ratio = \frac{Uptake_{BC}}{Uptake_{SC}} \tag{6.3}$$

6.3.6. Bootstrapping

Bootstrapping is a powerful statistical method commonly used to estimate population statistics, such as a population mean, and confidence intervals based on a limited sample dataset. This method operates under the assumption that the sample distribution is a random sampling of the global distribution done without bias. By randomly sampling a dataset allowing for replacements (IE duplicates), the bootstrapping method mimics the initial sampling used to generate the data. By repeating this re-sampling 10,000 times a distribution of possible population statistics is generated which can be used to estimate the desired population statistic. For this work, bootstrapping was used to estimate the 99% confidence intervals (CI) of the mean for the distributions of *Uptake Ratios* and *Tanimoto Coefficients*. These CIs were used to determine whether a statistically significant difference between MOFs which do and do not meet the DoE-PRT exists. The 99% confidence intervals were used to compare the single and binary component distributions for MOFs which pass and fail the DoE-PRT. If the 99% confidence intervals of the mean for the two distributions do not overlap, the difference between the distributions is considered statistically significant. The bootstrapping was performed using a code written in-house in Python, and the confidence intervals were calculated using equation 6.4, where μ is the mean and σ is the standard error of the bootstrapped distribution.

$$99\%\ CI = \mu \pm 2.576\sigma \tag{6.4}$$

6.3.7. Linear Discriminant Analysis (LDA)

Further analysis was performed on the resulting distributions using Linear Discriminant Analysis (LDA). LDA is a machine learning method used to split N-dimensional data into two categories using a single linear kernel. LDA was performed on the resulting data to quantify the separation between MOFs which did and did not meet the DoE-PRT. This analysis was performed using the Sci-kit Learn module in Python.[29] The results presented herein are models fit using 5-fold cross-validation.

6.4. Results and Discussion

6.4.1. Effect of Tanimoto on Binding Sites

The work presented in this chapter assumes that a low *Tanimoto* coefficient corresponds to a change in binding sites within the MOF. To visualize the effect of the *Tanimoto* on the displacement of

N_2 binding sites, 20 MOFs were selected at random along a wide range of *Tanimoto* values (0.03 to 0.99). The full list of MOFs can be found in Appendix 6.1. Figure 6.1a and 6.1b show the N_2 binding sites for the single component and binary component simulations, respectively, in the MOF with the Cambridge Crystallographic Datacentre (CCDC) designation FASQUN. FASQUN has a low N_2 *Tanimoto* coefficient value of 0.20, indicating low similarity between the SC and BC binding environments. The binding sites (circled in red) in Figure 6.1a are clearly replaced by CO_2 binding sites in Figure 6.1b (also circled). The N_2 binding sites located in the BC simulation are displaced from their SC positions and instead occupy less favourable binding sites surrounding the original site. This is evidenced by the reduction in occupancy values, where the SC sites have occupancy values of $2.5x10^{-3}$ $Å^{-3}$ compared to the sites in the BC simulation which have occupancy values of about $6.4x10^{-4}$ $Å^{-3}$. The presence of CO_2 therefore reduces the highest occupancy value to about 25% of the SC value, indicating a significantly less favourable binding environment.

Alternately, the SC and BC binding sites for the MOF with the CCDC designation CUGLTM02 are shown in Figures 6.1c and 6.1d, respectively. CUGLTM02, which has an N_2 *Tanimoto* coefficient of 0.99, shows no change in N_2 binding sites in the presence of CO_2. In contrast to the MOF FASQUN, CUGLTM02's highest SC binding site has an occupancy of $2.6x10^{-3}$ $Å^{-3}$ compared to the highest BC occupancy of $2.5x10^{-3}$ $Å^{-3}$, which indicates the probability of the site being occupied by an N_2 is almost unchanged in the presence of CO_2, reaffirming the assumption that *Tanimoto* coefficients are an indication of binding environment similarity.

6.4.2. Probability Plot Comparisons

Probability plots generated from Grand Canonical Monte Carlo (GCMC) were used to study the effects of competition on the binding environments of CO_2 and N_2 at post-combustion carbon capture conditions. The *single component* (SC) and *binary component* (BC) probability plots were compared for CO_2 and N_2 using a *Tanimoto* similarity metric for each atom type. The data was then separated according to the MOF's ability to meet the DoE-PRT. The histograms of the *Tanimotos* for this comparison are shown in Figure 6.2a, 6.2b, and 6.2c for the carbon, oxygen, and nitrogen probability plots, respectively. These plots have been normalized so that the area under the curve is equal to 1, allowing for a more direct comparison of the sets with a different number of MOFs. Upon initial analysis of Figures 6.2a and 6.2b, the histograms for MOFs which pass the DoE-PRT are nearly identical to those that fail to meet the DoE-PRT. Additionally, the means of the distributions all appear to exceed a

Tanimoto coefficient of 0.95, indicating that there is little to no change in the binding environments of the CO_2 guest molecules in the presence of N_2.

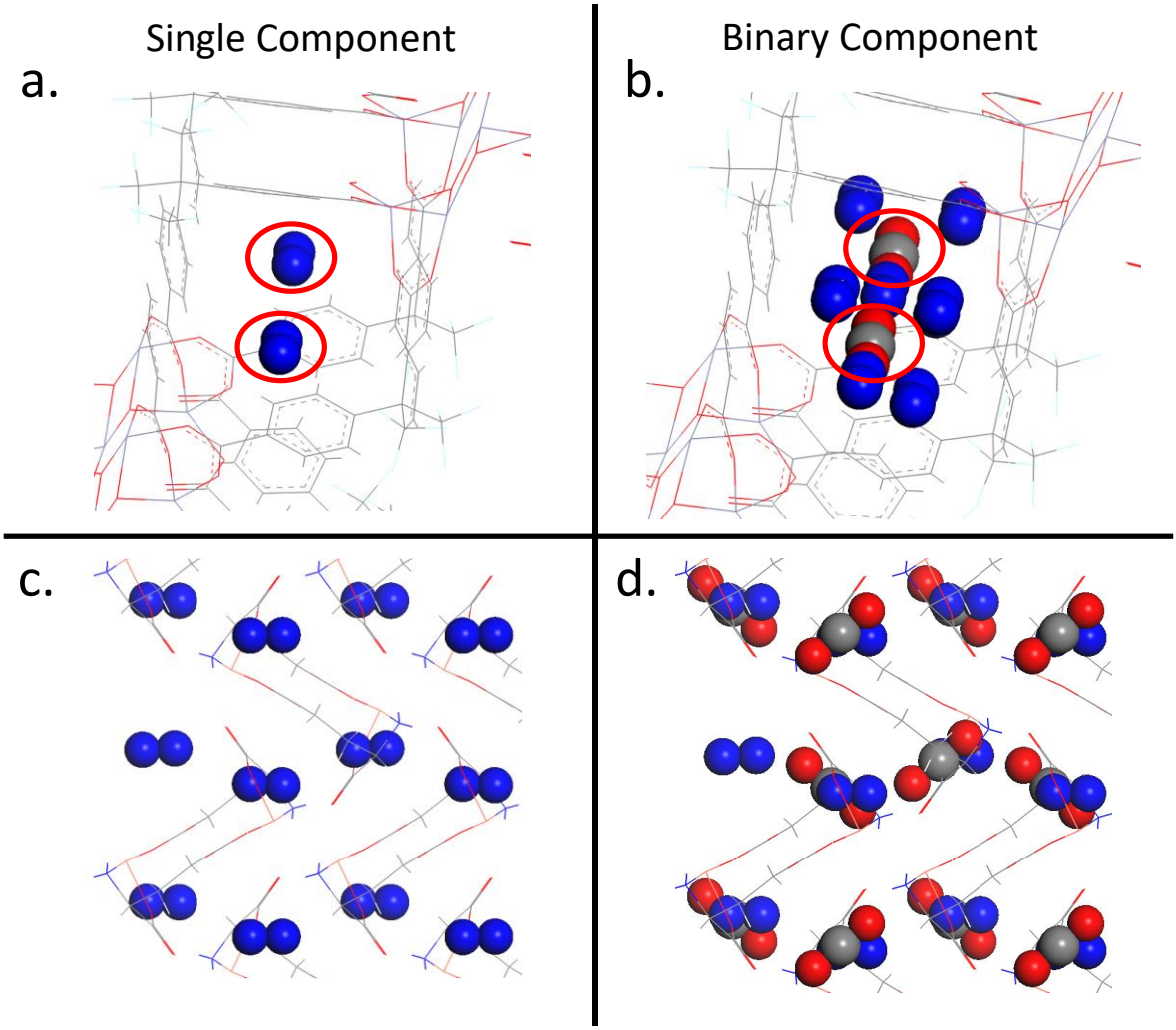


Figure 6.1 Visualization of the binding sites within the pores of MOFs (a) FASQUN for the single component N_2 simulation, (b) FASQUN for the binary component CO_2/N_2 simulation, (c) CUGLTM02 single component N_2 simulation, and (d) CUGLTM02 binary component CO_2/N_2 simulation. In this representation, the framework atoms are shown in a stick representation, the N_2 binding sites are displayed in CPK format in blue, and the CO_2 binding sites are displayed in CPK format with the oxygen atoms in red and carbon in grey.

Upon visual inspection of Figure 6.2c, however, there is a notable separation between the distributions of MOFs which do and do not meet the DoE-PRT. The distribution for the MOFs which do meet the DoE-PRT appears to favour lower *Tanimoto* coefficient values, indicating that the N_2 binding environment undergoes a drastic change in the presence of CO_2. Further, MOFs which fail to meet the DoE-PRT are skewed towards higher *Tanimoto* coefficient values, indicating that the N_2 environment is more likely to remain unchanged in the presence of CO_2 for those MOFs.

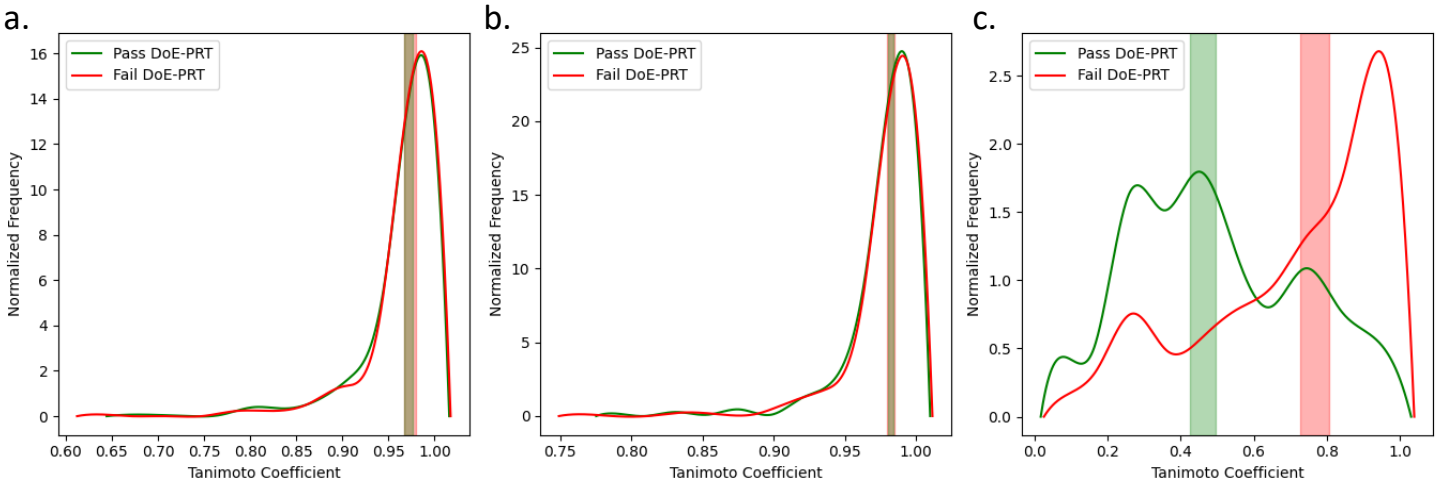


Figure 6.2 Histograms of the Tanimoto coefficients for (a) carbon, (b) oxygen, and (c) nitrogen probability plots, comparing the single component probability distributions to the binary component probability distributions. In these plots, the distributions for MOFs which meet the DoE-PRT are represented by the green lines, whereas MOFs which do not meet the DoE-PRT are represented by the red lines. The 99% confidence intervals of the mean Tanimoto coefficients are shown as green and red bars for the distributions of MOFs which passed and failed the DoE-PRT, respectively. All histograms have been normalized so that the area under the curve is equal to 1.

Although the distributions in Figure 6.2c appear separate, a significant amount of overlap still exists between the two sets of MOFs. As such, a more quantitative approach was taken to examine whether a statistically significant difference between these two distributions existed. To test whether the difference between these distributions was statistically significant, the 99% confidence interval of the mean was calculated for all 6 distributions in Figure 6.2 using the bootstrapping method. These confidence intervals are shown in the plots as green and red vertical bars. Using this method, the two confidence intervals are compared, checking for overlap. If the two confidence intervals do not overlap, the difference between these distributions is considered statistically significant.

The 99% confidence intervals of the mean completely overlap shown in Figures 6.2a and 6.2b, for the carbon and oxygen probability plot *Tanimoto* coefficients, respectively. The overlapping confidence intervals imply that no statistically significant difference between the distributions exists. On the other hand, those shown in Figure 6.2c for the nitrogen probability distributions, are separate with a large gap between the two intervals. This provides quantitative evidence that the difference between the two distributions is statistically significant. This difference in the plots is evidence that the displacement of N_2 by CO_2 within the pores of the material is an important mechanism in determining process level performance, linking atomistic behaviour to process performance for the first time. This

result led to a *secondary hypothesis: the displacement of N_2 by CO_2 results in reduced N_2 uptakes which correlates to the materials' ability to meet the DoE-PRT.*

6.4.3. Uptake Ratios

To explore this secondary hypothesis: that the displacement of N_2 by CO_2 drives the separation performance by reducing the N_2 loading, uptake ratios were calculated comparing SC and BC uptakes for both CO_2 and N_2, shown in equation 6.3. The resulting histograms from this analysis are shown in Figures 6.3a and 6.3b for CO_2 and N_2, respectively. Comparing CO_2 uptake ratio distributions for MOFs which do and do not meet the DoE-PRT in Figure 6.3a is analogous to the *Tanimoto* distributions seen in Figures 6.2a and 6.2b, with the two distributions completely overlapping, along with their corresponding 99% confidence intervals of the mean. Comparing the N_2 uptake ratio distributions in Figure 6.3b reveals an apparent separation of peaks with some significant overlap between the distributions of MOFs which do and do not meet the DoE-PRT. Again, the 99% confidence intervals of the mean do not overlap, indicating that the difference between these distributions is statistically significant.

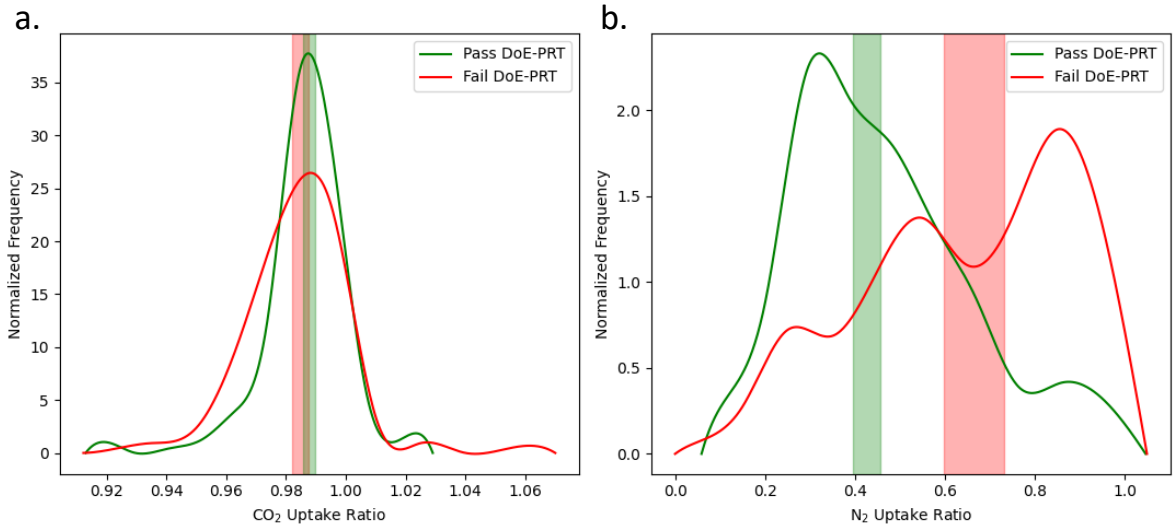


Figure 6.3 Histograms of the uptake ratios for (a) carbon dioxide, (b) molecular nitrogen, comparing the binary component uptakes to single component uptakes. In these plots, the distributions for MOFs which meet the DoE-PRT are represented by the green lines, whereas MOFs which do not meet the DoE-PRT are represented by the red lines. The 99% confidence intervals of the mean uptake ratios are shown as green and red bars for the distributions of MOFs which passed and failed the DoE-PRT, respectively. All histograms have been normalized so that the area under the curve is equal to 1.

The differences observed in the *Tanimoto* coefficient, and N_2 uptake ratio distributions implies that the competitive binding of CO_2 onto N_2 binding sites displaces the N_2 molecules and inhibits adsorption onto the surface of the material. The inhibition of N_2 binding within the pores of the material appears to be an important mechanism in determining a MOF's ability to meet the DoE-PRT. Although significant, the considerable overlap between the distributions in Figures 6.3a and 6.3b also implies that this competition is not the only mechanism controlling the process performance.

6.4.4. Correlation between Tanimoto and Uptake Ratio

Although both N_2 *Tanimoto* coefficients and N_2 uptake ratios display statistically significant differences between the classes of MOFs which do and do not meet the DoE-PRT, the second hypothesis has not been thoroughly vetted, and the similarity between these distributions may be coincidental. The relationship is explored in Figure 6.4, with the N_2 uptake ratios plotted against the N_2 *Tanimoto* coefficients. A linear tread line was fit to the full dataset, with a Pearson R^2 of 0.84, indicating that a correlation exists between these two properties. The separation between the two classes observed in the distributions as well as the correlation between these properties provides compelling evidence to confirm the *second hypothesis*, as well as the potential to exploit these relationships for use in a predictive tool.

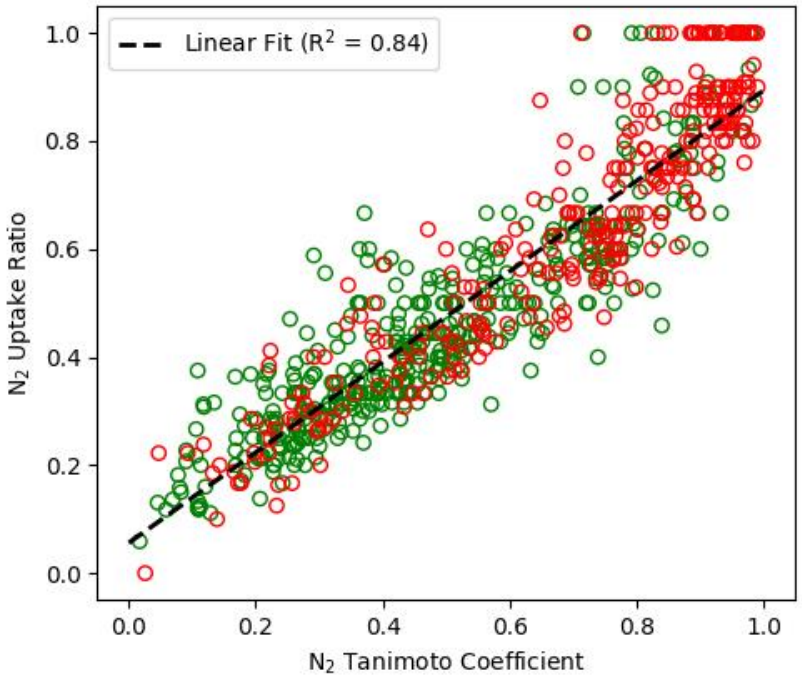


Figure 6.4 Plot of N_2 uptake ratio against N_2 Tanimoto coefficients for MOFs which meet the DoE-PRT (green circles) and MOFs which do not meet the DoE-PRT (red circles) with a linear tread line fit to the full data set (black dashed line).

6.4.5. LDA Prediction of DoE-PRT

To explore using N_2 *Tanimoto* coefficients and N_2 uptake ratios as a predictive tool in determining whether a material can meet the DoE-PRT, 2-dimensional LDA was performed on the data set using the two metrics. The best model fit was able to predict a MOF's ability to meet the DoE-PRT with a balanced accuracy of 68.6 % based on 5-fold cross-validation results. This means that this model is 18.6 % more predictive than a completely random assignment. This analysis was expanded to 38 additional metrics introduced in Chapters 4 and 5, pairing each metric with the N_2 *Tanimoto* coefficients and N_2 uptake ratios. The results of the 5-fold cross-validation of these models found that combining either the N_2 *Tanimotos* or N_2 uptake ratios with the percent regenerability metric returned a balanced accuracy of 75.9 %, an accuracy that is 25.9 % higher than a completely random model. Although coupling these metrics with the new N_2 parameters introduced in this chapter yielded improved accuracies in the LDAs, these models do not outperform the best LDA model discussed in Chapter 5, which had a balanced accuracy of 80.0 %. The N_2 *Tanimoto* coefficients and N_2 uptake ratios were then added to random forest models discussed in Chapter 5, and similarly yielded no improvement over the existing results of 83 % balanced accuracy. Although the performance of these models is not insignificant, these new metrics yielded no improvements over existing models. The high computational cost associated with calculating the *Tanimoto* coefficients for a predictive model means that the N_2 *Tanimoto* value's usefulness when predicting a MOF's ability to meet the DoE-PRT is limited.

6.4.6. Design Principles

The results from this study highlight the potential for the extraction of design principals, or chemical and structural features within the pores of MOFs which result in high performance. Attempts were made to study the binding sites generate by GALA to extract these design principles, utilizing techniques ranging from an analysis of the elemental composition of the binding pockets to more in-depth maximum clique analysis. Due to the relatively small size of the available dataset, with an under-representation of MOFs which do not meet the DoE-PRT, no useful design principles were extracted. Despite this lack of success, if given a sufficiently large set of binding sites for both high- and low-performing MOFs, it would be reasonable to assume that such an analysis would yield insightful design principles for researchers to target, guiding the development of MOFs for the PoC-CCS process.

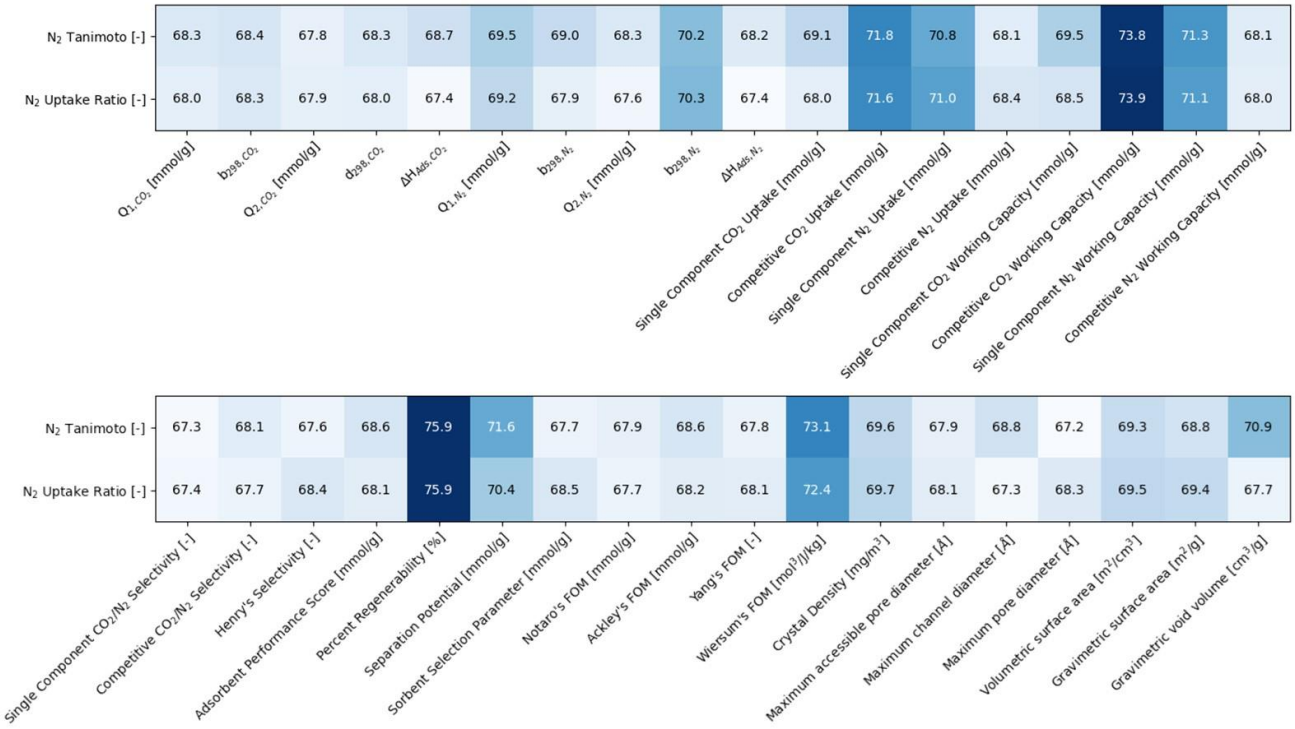


Figure 6.5 Heatmap of the 5-fold cross-validation balanced accuracies of the 2-dimensional LDAs including the N_2 Tanimoto coefficient and N_2 uptake ratio parameters combined with 38 metrics tested in Chapters 4 and 5 of this thesis.

6.5. Conclusions

The goal of this work was to explore the relationship between the N_2 binding sites within the pores of MOFs and study their relationship to the materials' ability to meet the US Department of Energy's Purity-Recovery Targets (DoE-PRT). Using the process performance data gathered in the large-scale screening of MOFs discussed in Chapter 4, and sophisticated techniques to extract and analyze the resulting binding sites, the change in binding environments for CO_2 and N_2 were analyzed. This analysis was performed with an emphasis on the change in binding in the presence of a competing gas. Using the Tanimoto coefficient to compare the probability distributions generated using GCMC, it was concluded that the N_2 binding environments for MOFs which met the DoE-PRT were significantly altered in the presence of CO_2. In contrast, the MOFs which did not meet the DoE-PRT were more likely to see little to no change in N_2 binding environments in the presence of CO_2. This discovery led to the first conclusion: MOFs which undergo a change in N2 binding sites when exposed to CO_2 are more likely to meet the

DoE-PRT. This change in binding environment is characterized by the displacement of N_2 by CO_2 from key binding pockets.

Further analysis of the displaced N_2 revealed that the shifts in binding environments were correlated to a reduction in N_2 uptake in the presence of CO_2. The N_2 uptake ratio between the single component and binary component simulations demonstrated that the displacement of N_2 resulted in a decrease in N_2 uptake capacity. This uptake ratio was also found to be an indicator of a MOF's ability to meet the DoE-PRT, and in combination with the N_2 *Tanimoto* coefficient could predict the material's ability to meet the DoE-PRT with a 68.6% balanced accuracy. The correlation between the N_2 uptake ratios and *Tanimoto* coefficients along with the statistically significant difference in the uptake ratio distributions led to a *second conclusion:* MOFs which meet the DoE-PRT are more likely to experience a reduction in N_2 uptake capacity due to competition between the N_2 and CO_2 binding sites.

These two conclusions are significant as they are the first instance of atomistic behaviour of gases within the pores of the MOFs being directly related to industrial separation performance. Although the conclusions of Chapter 5 indicated that N_2 adsorption behaviour was linked to a material's ability to meet the DoE-PRT, with the work discussed in this chapter we have shown that a direct link exists between the material's ability to meet the DoE-PRT and the physical binding sites of the two gases in the pores of a MOF.

Importantly, although N_2 binding environments have proven significant, it should be noted that there exists significant overlap between the distributions of N_2 *Tanimoto* coefficients and N_2 uptake ratios for the two classifications. Additionally, although the results from the linear discriminant analysis showed that using both properties improved prediction by 18.6% when compared to a randomized model, this performance is not high enough for use as a predictive tool. The overlap in both distributions and the performance of the LDA models can therefore lead to a *third conclusion*: although competition for binding sites is an important factor in determining a MOF's ability to meet the DoE-PRT, other unknown factors also play a role in that determination.

6.6. Future Work

This work can form the basis of future studies to explore the relationship between the N_2 binding sites and a material's process performance. Although the conclusions of this chapter discussed the relationship between competition for binding sites within the MOF, the composition and structural makeup of those binding sites was not explored. To perform an analysis of the binding site

compositions, additional simulations involving thousands of materials would need to be performed, calculating high quality CO_2 and N_2 binding sites from both single component and binary component GCMC simulations. Once completed and the binding sites generated, a quantitative method of identifying which sites within the MOFs experienced competition between CO_2 and N_2, with CO_2 as the dominant guest, would need to be developed and refined for high-throughput screening. Once the relevant binding sites are identified, a maximal clique analysis, a graph theoretical method for identifying common structural features, would need to be performed on the whole set of identified binding sites. This analysis would ideally extract common structural features of the sites and allow for the identification of common structural and chemical motifs that result in the desired competition. If such motifs can be successfully identified, they would provide valuable insights into the underlying chemistry driving the separations and provide design targets for MOF chemists working in the field of materials discovery for CO_2/N_2 separation applications.

6.7. References

1. Burns, T. D., Pai, K. N., Subraveti, S. G., Collins, S. P., Krykunov, M., Rajendran, A. & Woo, T. K. Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

2. Carbon-Capture-Technology-Compendium-2020.

3. Basdogan, Y., Sezginel, K. B. & Keskin, S. Identifying highly selective metal organic frameworks for CH_4/H_2 separations using computational tools. *Industrial & Engineering Chemistry Research* **54**, 8479–8491 (2015).

4. Krishna, R. Screening metal–organic frameworks for mixture separations in fixed-bed adsorbers using a combined selectivity/capacity metric. *RSC Advances* **7**, 35724–35737 (2017).

5. Rege, S. & Yang, R. A simple parameter for selecting an adsorbent for gas separation by pressure swing adsorption. *Separation Science and Technology* **36**, 3355–3365 (2001).

6. Leperi, K. T., Chung, Y. G., You, F. & Snurr, R. Q. Development of a general evaluation metric for rapid screening of adsorbent materials for postcombustion CO_2 capture. *ACS Sustainable Chemistry and Engineering* **7**, 11529–11539 (2019).

7. Boyd, P. G., Chidambaram, A., García-Díez, E., Ireland, C. P., Daff, T. D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S. M., Maroto-Valer, M. M., Reimer, J. A., Navarro, J. A. R., Woo, T. K., Garcia, S., Stylianou, K. C. & Smit, B. Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature* **576**, 253–256 (2019).

8. Nugent, P., Giannopoulou, E. G., Burd, S. D., Elemento, O., Giannopoulou, E. G., Forrest, K., Pham, T., Ma, S., Space, B., Wojtas, L., Eddaoudi, M. & Zaworotko, M. J. Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

9. Liang, L., Liu, C., Jiang, F., Chen, Q., Zhang, L., Xue, H., Jiang, H. L., Qian, J., Yuan, D. & Hong, M. Carbon dioxide capture and conversion by an acid-base resistant metal-organic framework. *Nature Communications* **8**, (2017).

10. Jiang, J., Lu, Z., Zhang, M., Duan, J., Zhang, W., Pan, Y. & Bai, J. Higher symmetry multinuclear clusters of metal–organic frameworks for highly selective CO_2 capture. *Journal of the American Chemical Society* **2**, jacs.8b07589 (2018).

11. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

12. Chung, Y. G., Gómez-Gualdrón, D. A., Li, P., Leperi, K. T., Deria, P., Zhang, H., Vermeulen, N. A., Stoddart, J. F., You, F., Hupp, J. T., Farha, O. K. & Snurr, R. Q. In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances* **2**, e1600909 (2016).

13. Dzubak, A. L., Lin, L. C., Kim, J., Swisher, J. A., Poloni, R., Maximoff, S. N., Smit, B. & Gagliardi, L. Ab initio carbon capture in open-site metal-organic frameworks. *Nature Chemistry* **4**, 810–816 (2012).

14. Chung, Y. G., Camp, J., Haranczyk, M., Sikora, B. J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O. K., Sholl, D. S. & Snurr, R. Q. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **26**, 6185–6192 (2014).

15. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **43**, 5735–5749 (2014).

16. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

17. Bae, Y. S. & Snurr, R. Q. Development and evaluation of porous materials for carbon dioxide separation and capture. *Angewandte Chemie - International Edition* vol. 50 11586–11596 (2011).

18. García-Sánchez, A., Ania, C. O., Parra, J. B., Dubbeldam, D., Vlugt, T. J. H., Krishna, R. & Calero, S. Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

19. Provost, B. An improved N_2 model for predicting gas adsorption in MOFs and using molecular simulation to aid in the interpretation of SSNMR spectra of MOFs. (2014).

20. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897–8909 (1990).

21. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **114**, 10024–10035 (1992).

22. Campañá, C., Mussard, B., Woo, T. K., Campañá, C., Mussard, B. & Woo, T. K. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *Journal of Chemical Theory and Computation* **5**, 2866–2878 (2009).

23. Kresse, G. & Furthmüller, J. Vienna ab initio simulation package (VASP). (2001).

24. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **54**, 11169–11186 (1996).

25. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868 (1996).

26. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **105**, 9982–9985 (1996).

27. Willett, P. *Similarity-based approaches to virtual screening*. *AstraZeneca R&D Charnwood* (2003).

28. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, (2015).

29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

6.8. Appendix 6.1: Comparison of N_2 Single and Binary Component Binding Sites

Table A6.1. Comparison of the single component N_2 and binary component binding sites for five CoRE MOFs.

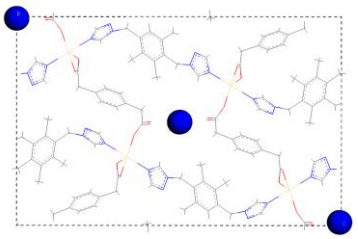
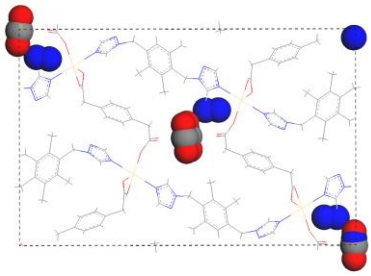
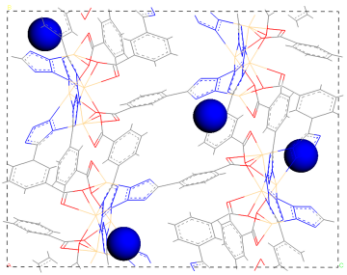
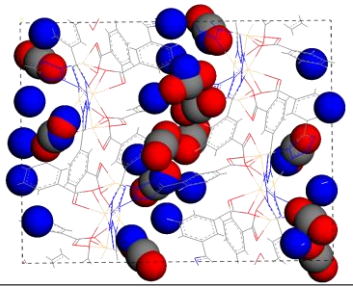
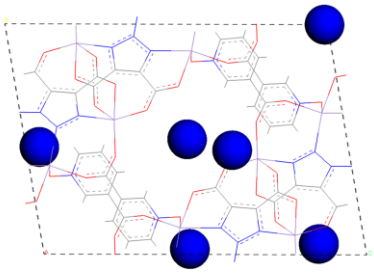
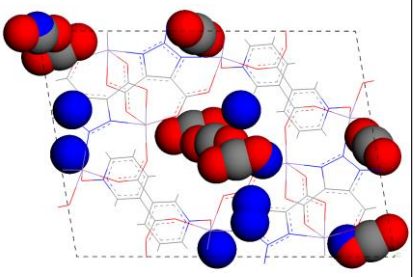
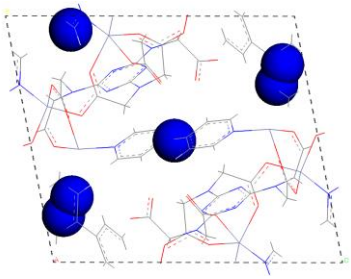
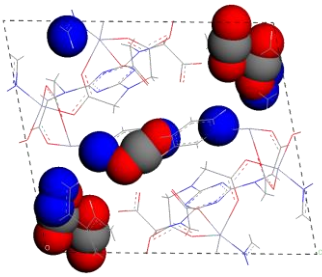
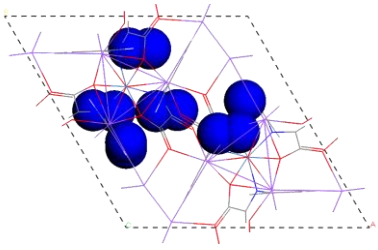
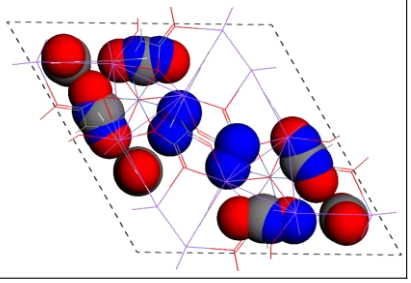
MOF	*N_2 Tanimoto Coefficient*	*Single Component N_2 Binding Sites*	*Binary Component CO_2/N_2 Binding Sites*
FIHXUR	0.026		
PELXIP	0.080		
VIPZAX	0.13		
OSOMIT	0.23		
VEHNED	0.28		

Table A6.2. Comparison of the single component N_2 and binary component binding sites for five CoRE MOFs.

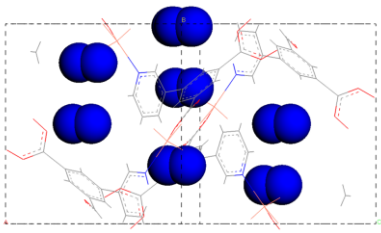
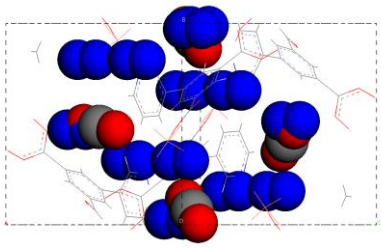
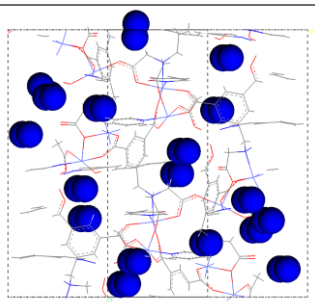
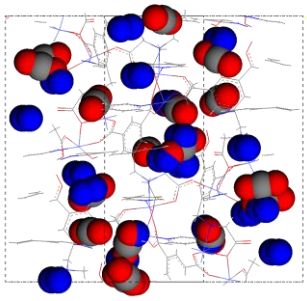
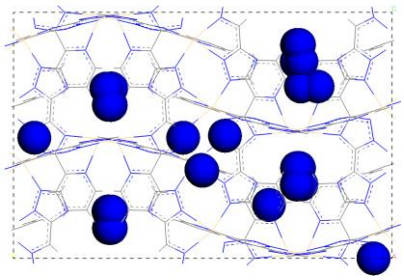
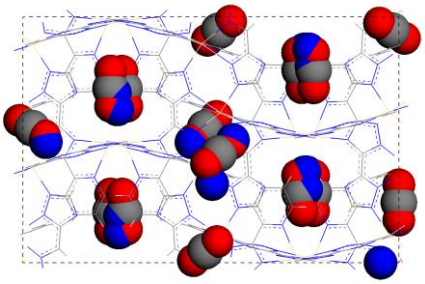
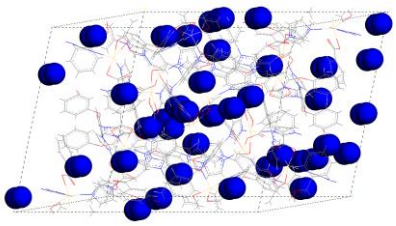
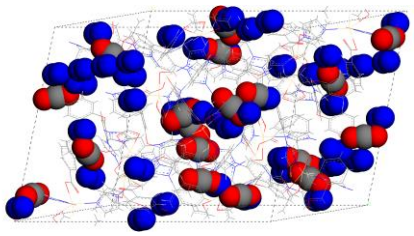
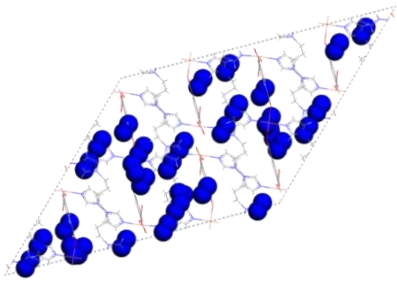
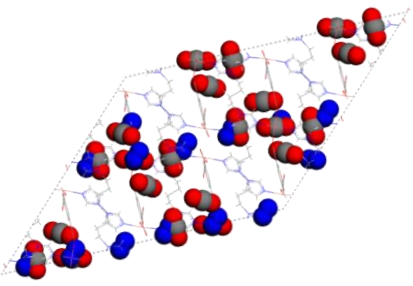
MOF	*N_2 Tanimoto Coefficient*	*Single Component N_2 Binding Sites*	*Binary Component CO_2/N_2 Binding Sites*
IDUDES	0.32		
FIFMIS	0.35		
LEDCAA	0.36		
OKIPUU	0.39		
LEDPIU	0.42		

Table A6.3. Comparison of the single component N_2 and binary component binding sites for five CoRE MOFs.

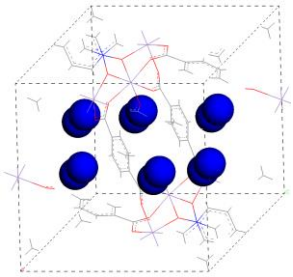
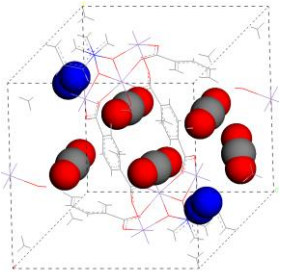
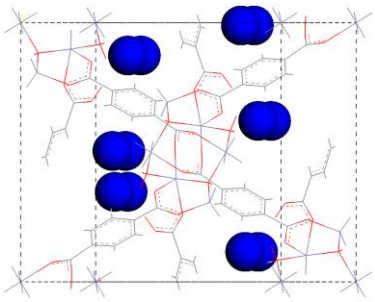
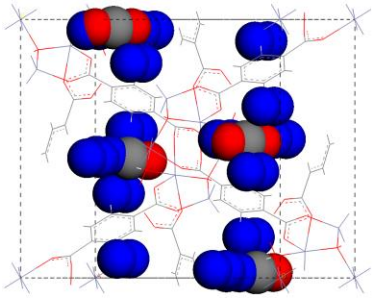
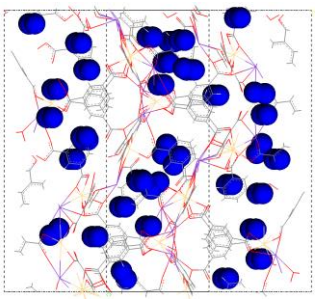
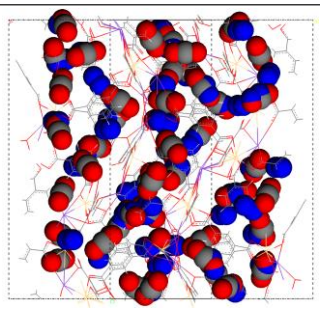
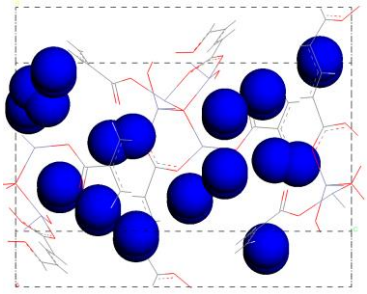
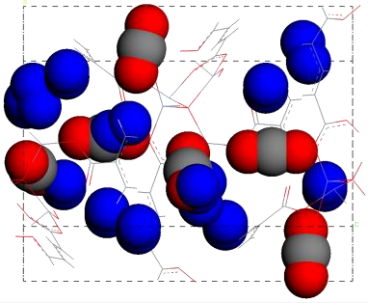
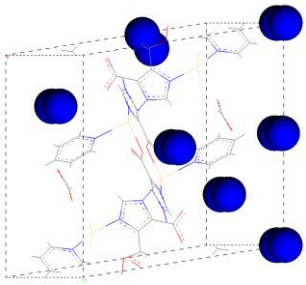
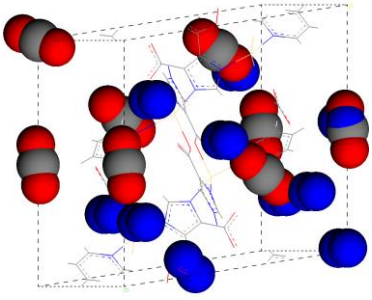
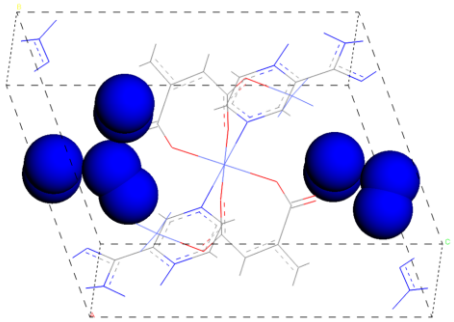
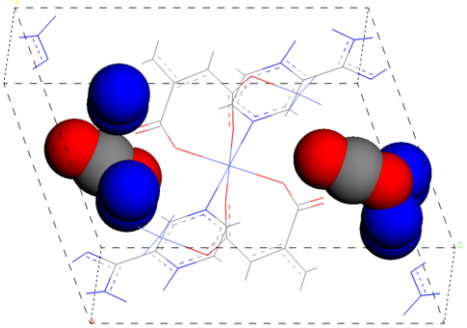
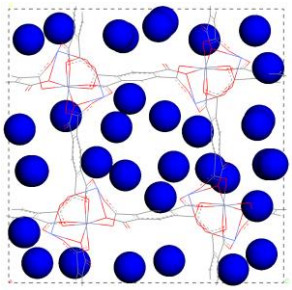
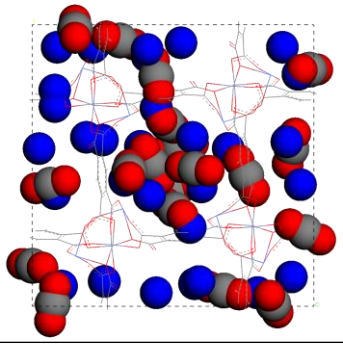
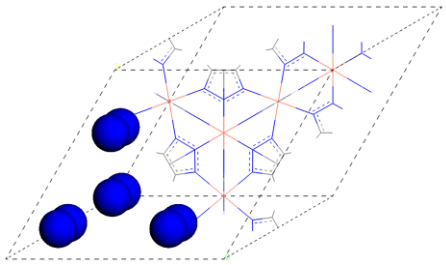
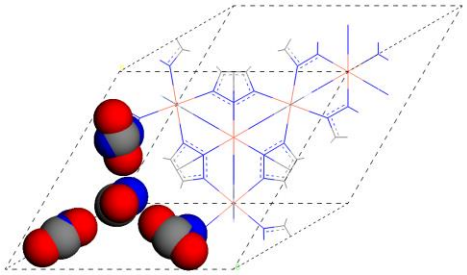
MOF	*N_2 Tanimoto Coefficient*	*Single Component N_2 Binding Sites*	*Binary Component CO_2/N_2 Binding Sites*
LUSHUD	0.44		
REDROI	0.46		
VEZMIY	0.50		
XOHLEM	0.59		
QEJYIP	0.65		

Table A6.4. Comparison of the single component N_2 and binary component binding sites for three CoRE MOFs.

MOF	***N_2 Tanimoto Coefficient***	***Single Component N_2 Binding Sites***	***Binary Component CO_2/N_2 Binding Sites***
EJONOH	0.70		
EGATOW	0.80		
CAYBAH01	0.90		

7. Chapter 7: Interpolation Model in the PSA Simulator

The work presented in this chapter is an expansion on the work presented in Chapter 4 of this thesis and forms the basis of a manuscript currently in progress. The work described in this chapter is entirely my own.

7.1. Abstract

The sophisticated pressure swing adsorption simulation code used in Chapter 4 relies on a simplistic competitive dual-site Langmuir model to estimate competitive loadings during the simulation. This competitive model uses single-component adsorption data calculated from grand canonical Monte Carlo (GCMC) simulations, despite GCMC's ability to directly calculate competitive loadings. In this chapter, an alternative method to estimate competitive loading using multi-component GCMC simulations with a simple linear interpolator was tested against the competitive dual-site Langmuir for its ability to predict loadings at conditions relevant to pressure swing adsorption (PSA) and temperature swing adsorption (TSA) systems. Using GCMC simulations, a grid was generated for 85 MOFs at PSA conditions, and 101 MOFs at TSA conditions, whereby the allowed ranges of temperature, pressure, and mole fractions for a CO_2/N_2 binary gas mixture were evenly sampled. Linear interpolators were fit to the resulting grids, known as adsorption cubes, and used to predict a set of 1000 loadings located between the grid-points. These loadings were compared to those calculated using the competitive dual-site Langmuir and GCMC simulations. For both PSA and TSA systems, it was found that the linear interpolation better reproduces the GCMC results when compared with the competitive dual-site Langmuir models. This improvement in competitive loading predictions demonstrates that a linear interpolator can be used to replace the competitive dual-site Langmuir model using 27 GCMC simulations, while simultaneously improving the predictive potential of the PSA and TSA simulation codes.

7.2. Introduction

The pores-to-process screening presented in Chapter 4 relied on a sophisticated pressure swing adsorption (PSA) simulation code. This PSA code used single-component isotherm parameters fit to data generated from grand canonical Monte Carlo (GCMC) simulations to predict industrial PSA performance metrics, accurately reproducing the results of an 80 kg pilot plant packed with Zeolite-13X.[1]

An important criticism of this simulation technique is the reliance on single-component adsorption data, a decision that was made during the simulator's development due to the lack of

experimental multi-component isotherms. Since no multi-component data was available, single-component isotherms are used to estimate the competition between the molecules in the gas mixture to be adsorbed onto the sorbent materials.[2] More specifically, once provided with single-component isotherms, the PSA simulator applies a competitive model derived from ideal adsorbed solution theory (IAST)[3] to predict competitive loadings at a variety of conditions. However, as discussed in Chapter 2 of this thesis, GCMC can directly simulate the adsorption of gas mixtures and produce accurate competitive isotherms for each component gas in the system. This led to the question: if competitive loadings can be directly calculated using GCMC, can these loadings be used in the PSA simulator and remove the reliance on the competitive isotherm model?

Although this question appears to be straight-forward, the implementation of such a fundamental change to the simulation code is not trivial. In section 2.3 of this thesis, a detailed description of the PSA simulation code is provided. Section 2.3 describes the way that the competitive isotherm model is used to predict the loadings of each gas at any given set of conditions with different temperatures, pressures, and mole ratios. The implementation of this competitive dual-site Langmuir model, which relies on temperature independent isotherm parameters (equations 2.20 to 2.22), means that loadings can be estimated for any combination of conditions. If competitive loadings are to be calculated directly through GCMC, the biggest challenge would be determining what set of conditions to use as inputs to accurately model all possible conditions found within the column that might occur during the process optimization.

Since the loadings at any given point within the columns cannot be predicted in advance, providing them directly using GCMC would require running the GCMC simulations on the fly for every point in the simulation. Over the course of a PSA simulation, the code will need to estimate loadings in the column thousands of times to accurately predict the propagation of the gas components through the column. Since a single GCMC simulation used to estimate loadings can range in simulation time from a matter of minutes to over an hour, running these simulations on an ad hoc basis would be too time consuming. This is compounded through the need to run the simulator hundreds of times for each material to effectively optimize the process performance. Therefore, running a GCMC simulation to directly predict the loadings at every step in the simulation is not practical.

In this chapter, I explore a possible solution to this challenge: using a linear interpolation model fit to a grid of GCMC adsorption data along a range of temperatures, pressures, and mole ratios. By generating a grid of adsorption data, herein referred to as an *adsorption cube*, representing the range of

possible conditions within a PSA column, the loadings for each gas can be interpolated for any set of conditions. The effect of this interpolation method is explored comparing the accuracy of the predicted loadings to binary GCMC simulation results and the competitive dual-site Langmuir model based on single-component isotherms. This analysis was performed considering the ranges present in a PSA column as well as the ranges present within a temperature swing adsorption (TSA) column, due to its relevance to future works being performed in the Woo Lab.

7.3. Methods

7.3.1. Metal-Organic Framework Sets

For this analysis, two sets of MOFs were curated through random selection from the sets optimized during the large-scale screening detailed in Chapter 4. The set used to test the PSA system consisted of 85 MOFs, and the set used to test the TSA system consisted of 101 MOFs. This selection was performed at random to ensure no bias in the MOF performance and build a generalized description of the performance of the competitive dual-site Langmuir model.

7.3.2. Grand Canonical Monte Carlo Simulations

To generate the data necessary to test the efficacy of the interpolation model and compare it to other methods of calculating competitive uptake, a series of Grand Canonical Monte Carlo (GCMC) simulations were performed using a code written in-house[4] based on the DL_POLY molecular dynamics[5] code. To perform a test representative of the range of conditions present within the gas separation column, careful consideration was taken to ensure ranges of appropriate conditions were sampled.

This analysis was performed for two separation systems: Pressure Swing Adsorption (PSA) and Temperature Swing Adsorption (TSA). The key difference between these two systems is in the mechanism for removing the adsorbed CO_2 during the evacuation phase of the cycle. In a PSA system, the pressure in the column is reduced which reduces the bulk-phase chemical potential of the gas, shifting the system's equilibrium of the CO_2 from the adsorbed phase to the gas-phase. In a TSA system, instead of reducing the pressure, the temperature of the column is increased which causes a similar shift from adsorbed phased to gas phase seen in the PSA system. A general set of parameters could therefore not be probed to cover both systems, and instead two separate comparisons were performed.

The simulations were carried out assuming a binary mixture of CO_2 and N_2 were present within the column, ranging from a mole fraction of 1.0 to 0.0 CO_2. An 11x11x11 grid probing the possible range of temperatures, pressures, and mole fractions present within the column at the given time was generated for each MOF using GCMC. The grid-points were selected to be evenly spaced, with the exception of the pressure range which was partitioned along the Log_{10} scale. This was done to allow for a better sampling of the low-pressure regions, which are vitally important during these separations. The full range of conditions probed are shown in Table 7.1 for both systems.

Table 7.1 Ranges of conditions testing using GCMC for the PSA system and the TSA system.

		Pressure Swing Adsorption (PSA)	Temperature Swing Adsorption (TSA)
Temperature (K)	Max	313.15	395.15
	Min	298.15	295.15
Pressure (bar)	Max	2.0	3.0
	Min	1.0×10^{-4}	1.0×10^{-4}
Mole Fraction CO_2	Max	1.0	1.0
	Min	0.0	0.0

The guest molecules in this simulation were modelled using the Garcia-Sanchez parameters for CO_2,[6] and the N_2-NIMF (Nitrogen in Metal-Organic Frameworks) parameters for N_2.[7] The framework atoms were modelled using the Universal Force-Field (UFF)[8,9] to estimate the dispersion interactions, and the partial atomic charges used to model the electrostatic interactions were calculated from a DFT calculation using the REPEAT method.[10] To ensure sufficient sampling was performed on each grid-point, 20,000 equilibration cycles were performed followed by 1,500,000 production steps in the GCMC simulations. The decision to use production steps over the more conventional cycles was made to ensure sufficient sampling of the space within the MOFs was performed when the mole fraction of one of the component gases was very small. Alternate grid configurations were also generated using the same methodology to create 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids for each MOF.

7.3.3. Latin Hypercube Sampling

Once adsorption cubes were generated for each MOF, the efficacy of the interpolation model needed to be tested by sampling the space between the grid-points. To ensure a well distributed sampling of the space between grid-points, Latin Hypercube Sampling (LHS) was performed to generate 1000 points, sampling the 3-dimensional space described by Table 7.1. This LHS sampling was performed

using code written in Python and the lhsmdu (Latin Hypercube Sample with Multi-Dimensional Uniformity)[11,12] module. A unique set of points was generated based on the PSA conditions and the same points were used for all 85 MOFs within the PSA set. Likewise, another unique set of 1000 points was generated based on the TSA conditions and were used for all 101 MOFs within the TSA set.

GCMC simulations were then carried out using the same methodology used to generate the grids on each individual point within the LHS sets. These GCMC uptakes were calculated to determine the "true" gas uptake values at every point.

7.3.4. Competitive Dual-Site Langmuir

Once the "true" values for the uptakes had been determined using GCMC, the uptakes at the points in the LHS sets needed to be determined using the competitive dual-site Langmuir (DSL) model. This model relies on the fitted isotherm parameters based on single component CO_2 and N_2 uptake data, shown in equation 7.1, and assumes the gas occupies two unique binding sites. In this equation, θ is the uptake (or loading) of gas, i, P is the partial pressure of gas i, Q_1 and Q_2 are the saturation uptakes, and b and d are the DSL constants of gas, i. Q_1, Q_2, b, and d are all parameters that are fit using single site uptake data generated by GCMC via a modified IAST.[3,13] The fittings performed by this IAST program aimed to minimize the root mean squared error (RMSE).

$$\theta_i = \frac{Q_{1,i} b_i P_i}{1 + b_i P_i} + \frac{Q_{2,i} d_i P_i}{1 + d_i P_i} \tag{7.1}$$

The competitive DSL model combines the fitted parameters from both component gases. The key assumption made by this model is that the two guest molecules compete for the same binding sites within a MOF, where the stronger site (described by Q_1 and b) for CO_2 competes with the stronger site for N_2, and the weaker site (described by Q_2 and d) for CO_2 compete with the weaker site for N_2. This results in the competitive model shown in equation 7.2, which calculates the uptake of gas *i* in the presence of gas *j*. Using this model and the same DSL parameters used in Chapter 4, the uptakes at every point in the LHS sets were calculated.

$$\theta_i = \frac{Q_{1,i} b_i P_i}{1 + b_i P_i + b_j P_j} + \frac{Q_{2,i} d_i P_i}{1 + d_i P_i + d_j P_j} \tag{7.2}$$

7.3.5. Linear Interpolation

A linear interpolator was fit for each MOF using the 11x11x11 adsorption grid. The code was written in Python using the LinearNDInterpolator module in the SciPy package.[14] Once the interpolation model was fit using the adsorption cube, loadings were calculated for every point in the LHS Sets. This process was then repeated for the 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids for each MOF.

7.3.6. Percent Mean Absolute Deviations

Once the uptake predictions had been generated using the competitive DSL model and the linear interpolation model, the uptakes needed to be compared to the "true" uptakes calculated from the multi-component GCMC simulations. The comparison was performed using a Percent Mean Absolute Deviation (MAD), shown in Equation 7.3, where Θ is the uptake and N is the number of points in the LHS set.

$$Percent\ MAD = \frac{100\%}{N}\sum_{i=1}^{N}\frac{\left|\theta_{predicted} - \theta_{true}\right|}{\theta_{true}} \tag{7.3}$$

7.4. Results and Discussion

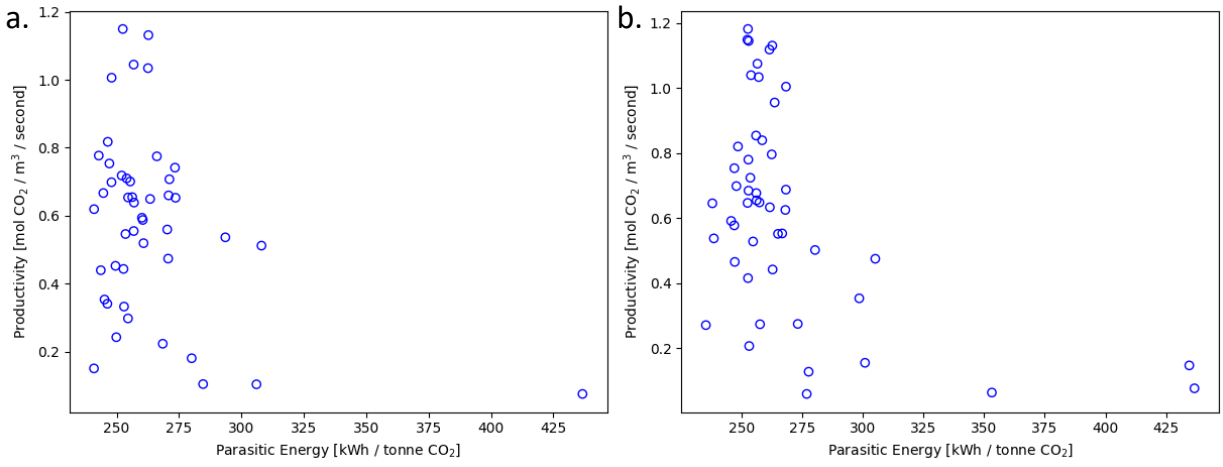


Figure 7.1 Plots of the productivity vs parasitic energy for MOFs which meet the DoE-PRT in (a) the Pressure Swing Adsorption (PSA) set, and (b) the Temperature Swing Adsorption (TSA) sets. All points displayed are the best parasitic energy point from the optimizations in Chapter 4.

7.4.1. Performance of Test Sets

The performance of the MOFs found in the PSA and TSA test sets are found in Figure 7.1a and 7.1b, respectively. The PSA set consisted of 80 MOFs, 68.75% of which failed to meet the DoE-PRT, and the TSA set consisted of 101 MOFs, 74.26% of which failed the DoE-PRT. The over-representation of MOFs which do not meet the DoE-PRT within both sets is due to the high proportion of MOFs which failed to meet the DoE-PRT from the CoRE database, based on the work discussed in Chapter 4. The distributions of performance for the remaining MOFs in the set are shown in Figure 7.1a, and 7.1b for the PSA and TSA set, respectively. Of the MOFs in both sets which met the DoE-PRT, the parasitic energy ranges from 235 to 437 kWh/tonne CO_2 captured, while the productivity ranges from 0.06 to 1.18 moles CO_2 captured / m^3 of adsorbent / second. These ranges indicate the materials tested covered a wide range of PSA performances and we can therefore assume that any results discussed in this chapter are similarly representative.

7.4.2. Pressure Swing Adsorption

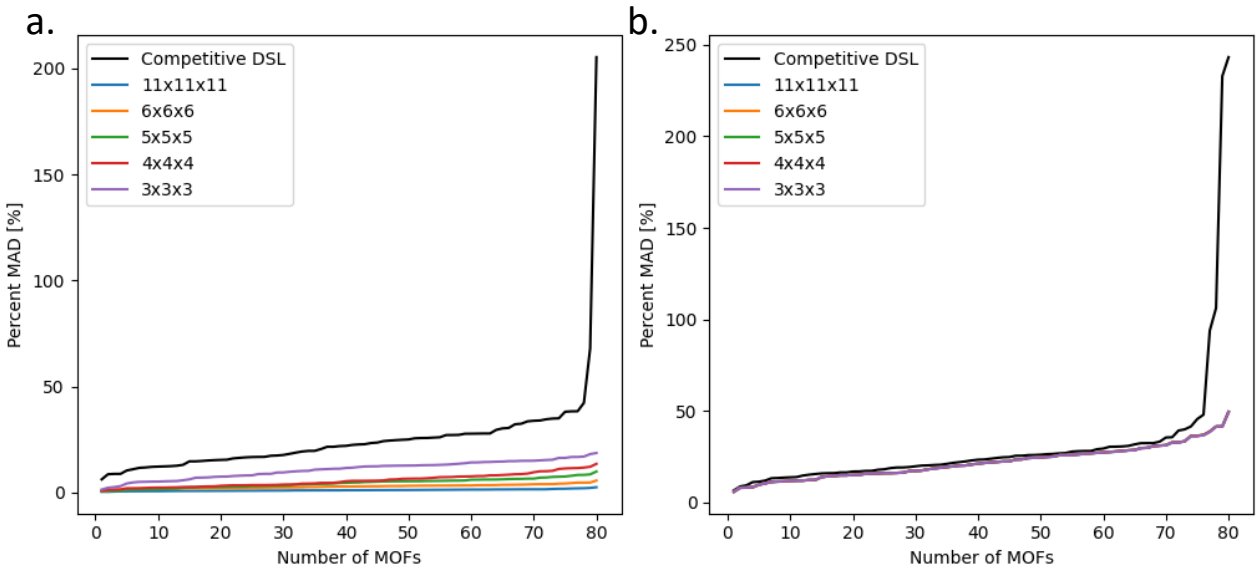


Figure 7.2 Plots of the Percent Mean Absolute Deviation (MAD) for (a) CO_2 uptake and (b) N_2 uptake for all 80 MOFs in the PSA set. The black line represents the MAD for the loadings calculated using the competitive DSL model, and the purple, red, green, orange, and blue lines are the percent MADs from the loadings calculated using the linear interpolator using the 11x11x11, 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids, respectively. The percent MAD values are ordered from lowest to highest and are plotted as a function of the number of MOFs. This means that at any given point, the number of MOFs that fall below a certain Percent MAD indicated by the y-axis can be determined by the corresponding point on the x-axis.

The results from the PSA analysis are shown in Figure 7.2a and 7.2b for the CO_2 uptakes and N_2 uptakes, respectively. The plots show the Percent MAD from lowest to highest, with the number of MOFs given in the x-axis. To interpret these plots, one can look at the value in the y-axis, take for example a Percent MAD of 20%, and find the corresponding point in the x-axis, say 40 MOFs, this means that 40 MOFs from the set had a Percent MAD equal to or lower than 20%. When the predictions of the competitive DSL model are compared to the interpolator's predictions, it becomes clear that even the coarse 3x3x3 grid linear interpolation model outperforms the competitive DSL for the uptake of CO_2. This coarse-grid interpolator model had on average a Percent MAD 14.1% lower than that of the competitive DSL model. A finer grid, like the 11x11x11 grid, on average had a Percent MAD 23.8% lower than the competitive DSL. These improvements indicate that even using a coarse grid in PSA simulator would be beneficial in increasing the accuracy of the process simulations and one could selecting an appropriate grid size to control the level of accuracy needed for a specific application.

Interestingly, the results in Figure 7.2b, which compares the grid interpolation results to the competitive DSL models for N_2 uptakes yields very different results. Although the grid interpolator does outperform the competitive DSL for every MOF, it only displays an improvement of 8.6 % on average for the 11x11x11 grid and the different grid-sizes appear to be almost entirely overlapping. This overlap indicates that no significant improvements can be made from increasing the number of grid-points.

This behaviour is explained by observing the nitrogen isotherm shapes for the MOFs in the set. Figures 7.3a and 7.3b show all isotherms tested at 298.15 Kelvin for CO_2 and N_2, respectively from the PSA set. The CO_2 isotherms in Figure 7.3a are highly diverse in both uptakes, ranging from under 0.1 mmol/g to over 3 of CO_2 adsorbed at 2 bar with similarly diverse shapes ranging from linear to highly non-linear. When compared to the N. isotherms in Figure 7.3b, the maximum loadings only range from about 0.002 to 0.03 mmol/g N_2 adsorbed at the highest pressure. The shapes of the N_2 isotherms are also highly linear. The ranges of the isotherms as well as their average shapes are shown in Figure 7.3c and 7.3d for CO_2 and N_2, respectively. The average CO_2 isotherm has a clear curve, indicating that on average there are stronger interactions between the CO_2 and the MOF frameworks. Conversely the average nitrogen isotherm is very linear, indicating weaker interactions between the N_2 molecules and the MOF frameworks. Nitrogen's weak loadings coupled with their linear isotherms means that prediction via linear interpolation will yield accurate and consistent results. The linear shapes of these isotherms also explain the low sensitivity to the grid sizes being tested.

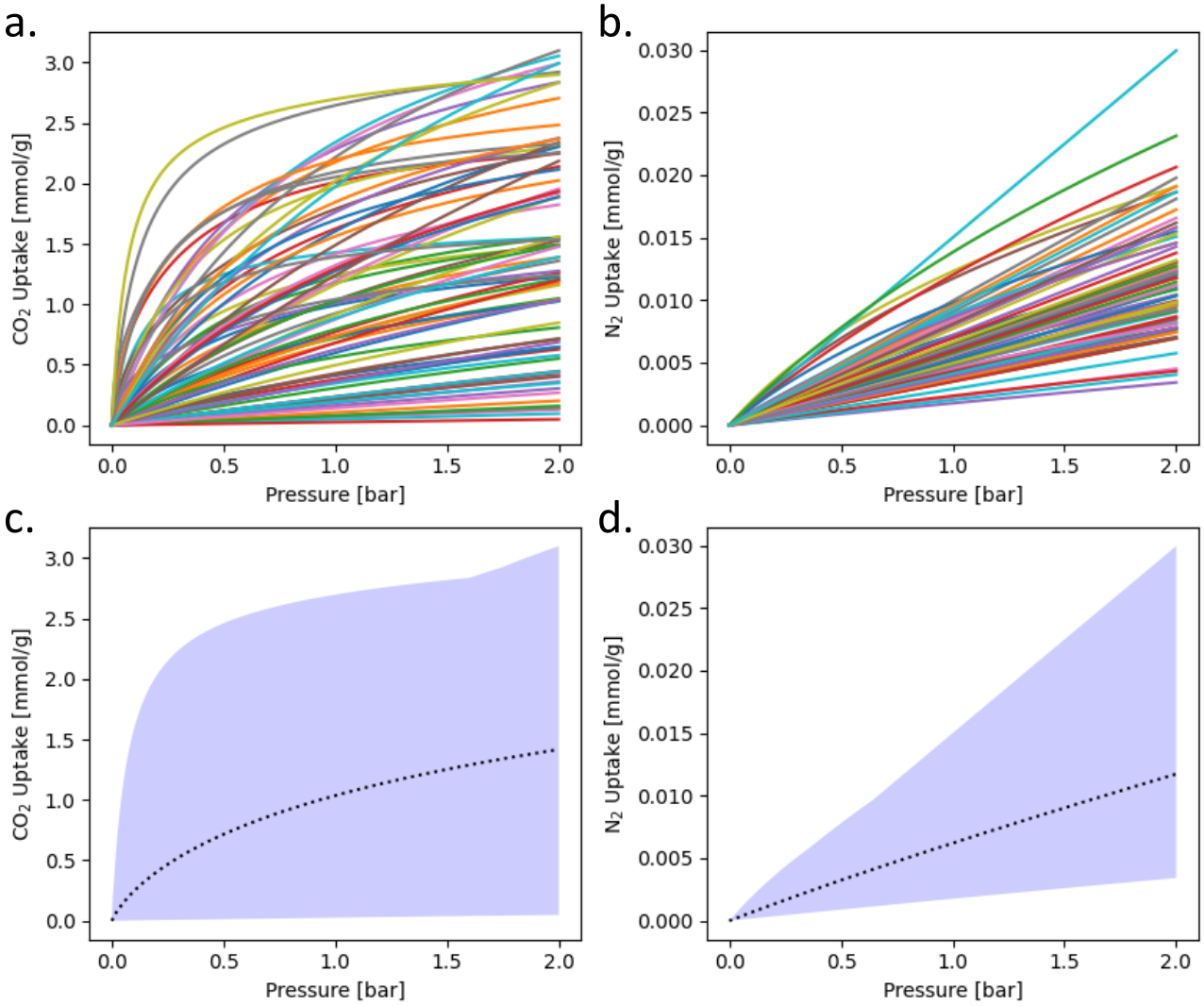


Figure 7.3 a) Plot of all CO_2 isotherms used in the PSA set, showing the adsorption of CO_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. b) Plot of all N_2 isotherms used in the PSA set, showing the adsorption of N_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. c) The range of CO_2 isotherms tested (blue area) showing the upper and lower bounds of CO_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the PSA set, and the average uptake as a function of pressure (black dotted line). d) The range of N_2 isotherms tested (blue area) showing the upper and lower bounds of N_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the PSA set, and the average N_2 uptake as a function of pressure (black dotted line).

7.4.3. Temperature Swing Adsorption

The results from the TSA analysis are shown in Figures 7.4a and 7.4b for the CO_2 uptakes and N_2 uptakes, respectively. The results appear analogous to the PSA set, however the differences between the competitive DSL model uptakes and the interpolator models are not as extreme as those seen in the PSA set. For instance, the sharp grid (11x11x11) linear interpolation model had an average Percent MAD

of 15.86 % when compared to the competitive DSL model. This contrasts with the 28.3 % reduction in percent MAD observed with the PSA set. Despite this, the CO_2 predictions based on the interpolation models are still an overall improvement when compared to the competitive DSL model.

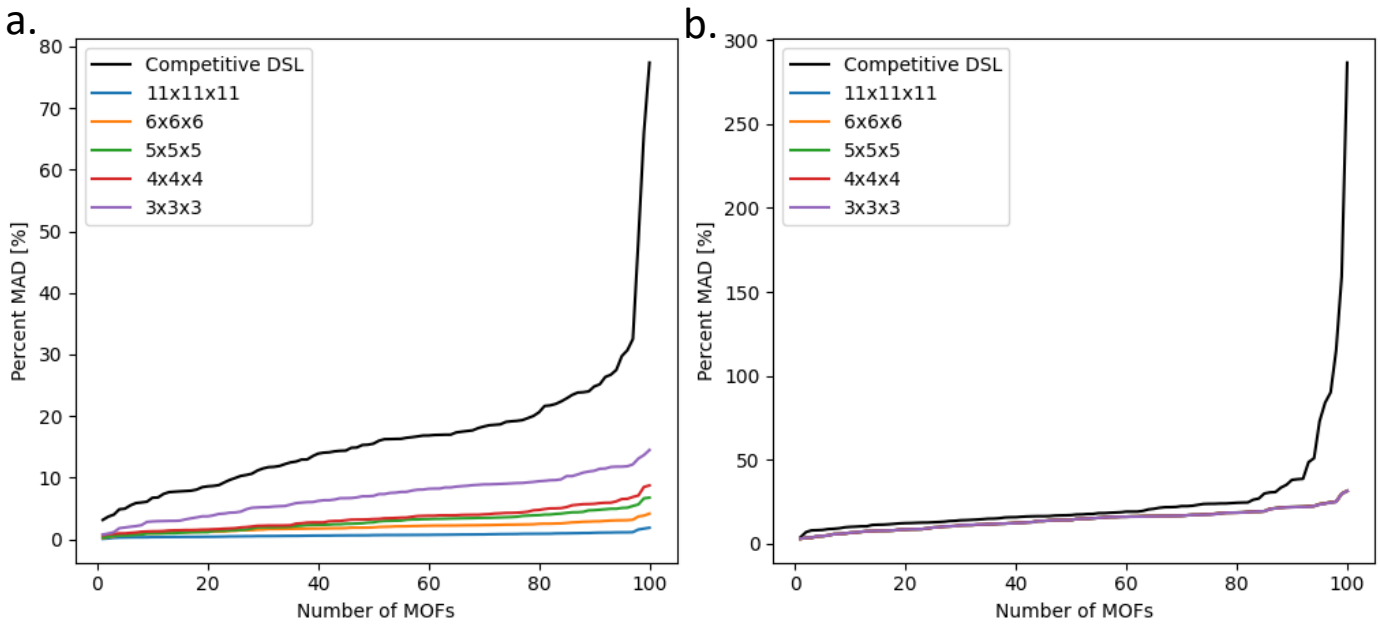


Figure 7.4 Plots of the Percent Mean Absolute Deviation (MAD) for (a) CO_2 uptake and (b) N_2 uptake for all 101 MOFs in the TSA set. The black line represents the MAD for the loadings calculated using the competitive DSL model, and the purple, red, green, orange, and blue lines are the percent MADs from the loadings calculated using the linear interpolator fit using the 11x11x11, 6x6x6, 5x5x5, 4x4x4, and 3x3x3 grids, respectively. The percent MAD values are ordered from lowest to highest and are plotted as a function of the number of MOFs. This means that at any given point, the number of MOFs that fall below a certain Percent MAD indicated by the y-axis can be determined by the corresponding point on the x-axis.

In Figure 7.4b, we again see little to no change in the Percent MAD for the N_2 uptakes when we alter the grid-spacing, with the interpolation models only improving upon the competitive DSL by 11.3 % on average. This behaviour is again analogous to that seen in the PSA set. The behaviours of the isotherms are shown in Figure 7.5, where Figure 7.5a and 7.5b at 298.15 K for CO_2, and N_2 respectively, for all 101 MOFs in the TSA set. It should be noted that the improvement of the linear interpolator over competitive DSL model when modelling N_2 loadings is slightly higher in the TSA set (11.3 % in the TSA set compared to 8.6 % in the PSA set) which is likely due in part to the higher curvature in the N_2 isotherms, illustrated by Figure 7.5b and 7.5d, where the highest adsorption in the low-pressure ranges is less linear than those in the PSA set. However, due to the random nature of the selection of MOFs within the sets, this difference was not deemed to be significant.

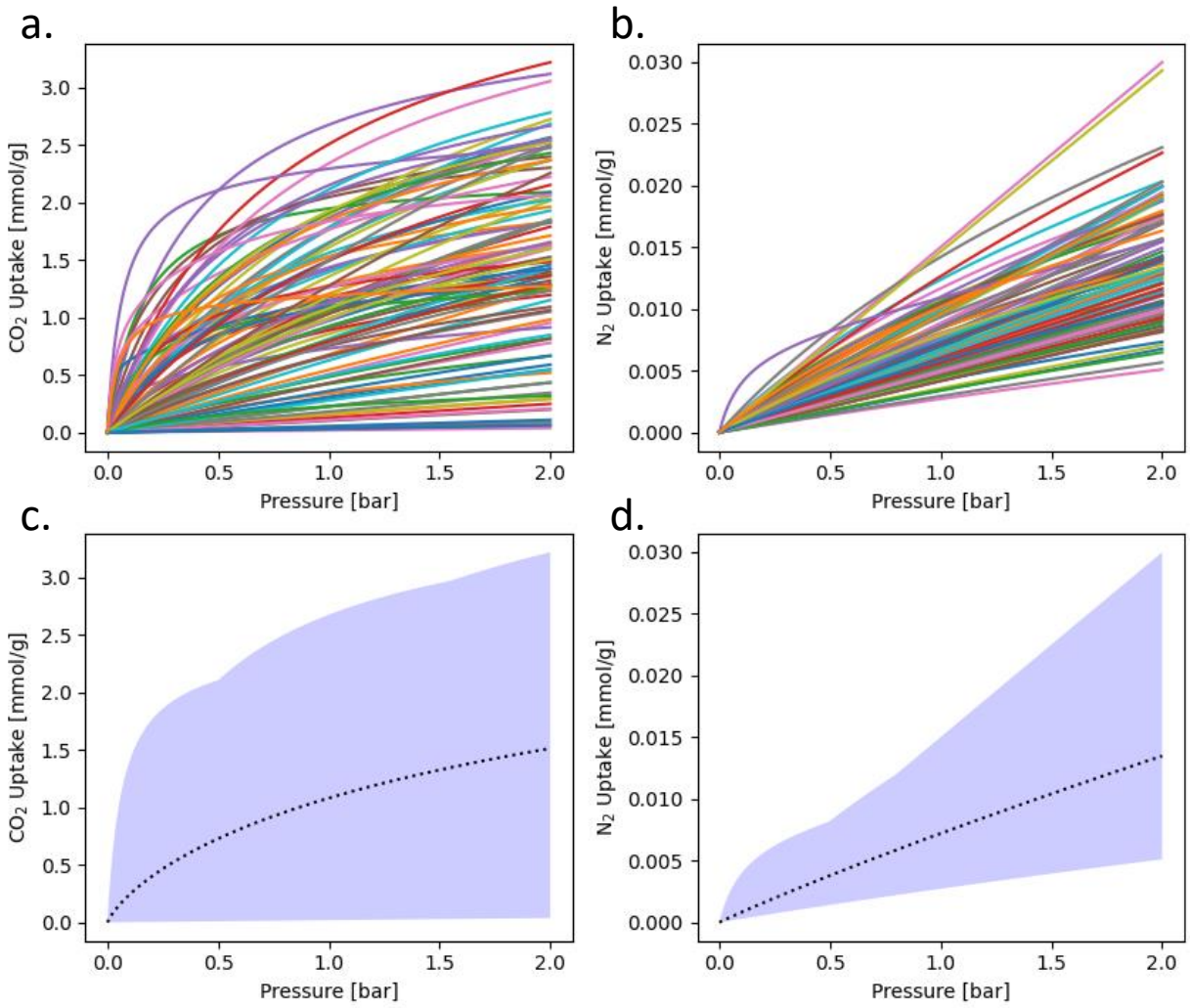


Figure 7.5 a) Plot of all CO_2 isotherms used in the TSA set, showing the adsorption of CO_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. b) Plot of all N_2 isotherms used in the TSA set, showing the adsorption of N_2 as a function of pressure up to 2.0 bar at 298.15 Kelvin. c) The range of CO_2 isotherms tested (blue area) showing the upper and lower bounds of CO_2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the TSA set, and the average uptake as a function of pressure (black dotted line). d) The range of N2 isotherms tested (blue area) showing the upper and lower bounds of N2 adsorption uptake as a function of pressure at 298.15 Kelvin for all MOFs in the TSA set, and the average N2 uptake as a function of pressure (black dotted line).

7.5. Conclusions

The sophisticated PSA simulator used in the large-scale screening of MOFs discussed in Chapter 4 relies on a competitive DSL model derived from single component simulation data. This DSL model is used to estimate the concentrations of component gases within a PSA column at each time step of the simulation. An exploration of an alternative paradigm to predict gas uptakes during this simulation was explored using a linear interpolator, fit to multi-component simulation data. The linear interpolator was

tested against the competitive DSL model in an attempt to predict uptakes from GCMC simulations at a range of conditions relevant to a PSA and TSA process.

Two sets were assembled to test the interpolation paradigm for a PSA system and TSA system. In both systems, it was determined that the linear interpolation models reliably outperform the competitive DSL model at predicting CO_2 and N_2 loadings. These improvements were present even when using a coarse 3x3x3 grid, requiring only 27 GCMC simulations, demonstrating that an interpolation model would be a valuable addition to the PSA simulation code. Although significant modifications to the PSA simulation code will need to be performed to implement this interpolation method, the improvements in gas adsorption predictions within PSA and TSA systems could improve the overall accuracy of the simulations.

To fully test the PSA simulator with an incorporated linear interpolation model, the code would need to be heavily modified and tested. The ultimate test of this new methodology would be to optimize materials using the interpolation model and compare the results to the competitive DSL model. This would determine whether the new methodology significantly impacts the reported performance of the materials. Upon receiving the results discussed in this chapter, the Rajendran group attempted modifications of their PSA simulation code, however they reported that the use of the interpolator increased the time required to optimize the system by an order of magnitude. Therefore, without overcoming this technical challenge, the implementation of a linear interpolation model into the PSA code is unlikely despite the improvements to the predicted gas uptakes at various points within the column.

7.6. References

1. Krishnamurthy, S., Rao, V. R., Guntuka, S., Sharratt, P., Haghpanah, R., Rajendran, A., Amanullah, M., Karimi, I. A. & Farooq, S. CO_2 capture from dry flue gas by vacuum swing adsorption: A pilot plant study. AlChE **60**, 1830–1842 (2014).

2. Haghpanah, R., Nilam, R., Rajendran, A., Farooq, S. & Karimi, I. A. Cycle synthesis and optimization of a VSA process for postcombustion CO_2 capture. *AIChE Journal* **59**, 4735–4748 (2013).

3. Chen, J., Loo, L. S. & Wang, K. An ideal absorbed solution theory (IAST) study of adsorption equilibria of binary mixtures of methane and ethane on a templated carbon. *Journal of Chemical and Engineering Data* **56**, 1209–1212 (2011).

4. Boyd, P. G. Computational high throughput screening of metal organic frameworks for carbon dioxide capture and storage applications. (2015).

5. Smith, W. & Forester, T. R. DL_POLY_2.0: A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* **14**, 136–141 (1996).

6. García-Sánchez, A., Ania, C. O., Parra, J. B., Dubbeldam, D., Vlugt, T. J. H., Krishna, R. & Calero, S. Transferable force field for carbon dioxide adsorption in zeolites. *Journal of Physical Chemistry C* **113**, 8814–8820 (2009).

7. Provost, B. An improved N_2 model for predicting gas adsorption in MOFs and using molecular simulation to aid in the interpretation of SSNMR spectra of MOFs. (2014).

8. Coupry, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal-organic frameworks. *Journal of Chemical Theory and Computation* (2016) doi:10.1021/acs.jctc.6b00664.

9. Rappé, A. K. K., Casewit, C. J. J., Colwell, K. S. S., Goddard III, W. A. & Skiff, W. M. UFF, a Full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).

10. Krykunov, M., Demone, C., Lo, J. W. H. & Woo, T. K. A new split charge equilibration model and REPEAT electrostatic potential fitted charges for periodic frameworks with a net charge. *Journal of Chemical Theory and Computation* **13**, (2017).

11. Deutsch, J. L. & Deutsch, C. v. Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning and Inference* **142**, (2012).

12. Sahil Moza. sahilm89/lhsmdu: Latin hypercube sampling with multi-dimensional uniformity (LHSMDU): Speed boost minor compatibility fixes. (2020).

13. Simon, C. M., Smit, B. & Haranczyk, M. PyIAST: Ideal adsorbed solution theory (IAST) Python package. *Computer Physics Communications* **200**, 364–380 (2016).

14. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

8. Chapter 8: FoCAS Surrogate PSA Model

The work performed in this chapter relied heavily on the data accumulated in Chapter 4 and forms the basis of a manuscript currently in progress, with the goal of publication in the near future. The work described in this chapter is entirely my own.

8.1. Abstract

Using advanced machine learning techniques, a suite of models was fit to act as a replacement, or surrogate to the sophisticated pressure swing adsorption simulator used in Chapter 4. These models were fit using the results of over 5 million individual process calculations and are able to predict the purity of captured CO2, recovery of CO2 from the flue gas stream, parasitic energy of capture, and productivity of the material with a high degree of accuracy. Using these models, an application called the Fossil fuel combustion carbon Capture And Storage (FoCAS) A.I. was developed, which uses a customizable genetic algorithm to fully optimize a material for post-combustion carbon capture applications using the conditions present in a coal-fired power plant. The use of FoCAS was demonstrated by optimizing 4,433 materials which were sorted by their ability to meet the US Department of Energy's purity-recovery targets (DoE-PRT) and ranked according to their best parasitic energy and productivity values. It was found that approximately 82% of the MOFs can meet the DoE-PRT, 459 of which can obtain parasitic energies below the DoE's target of 258 kWh/tonne CO2.

8.2. Introduction

The work performed throughout this thesis, and the work being performed in the field of materials discovery to identify candidate materials for post-combustion carbon capture,[1–5] has required a tremendous amount of time and computational resources to complete. The nearly infinite number of potential materials[5–10] that could be used for post-combustion gas separation means that the task of vetting every material at a high level is impossible. The large-scale screening of materials discussed in Chapter 4 of this thesis involved the use of a sophisticated pressure swing adsorption (PSA) simulator[1,11] coupled to a custom genetic algorithm to optimize the operating conditions of a PSA column at flue gas conditions typical of a coal-fired power plant. The goal of these optimizations was to find the best parasitic energy and productivity achievable for each individual material. Although the use of this simulation code yielded important insights into the behaviour of metal-organic frameworks (MOFs) in industrial PSA systems, the optimization of the operating conditions required up to 5 days to complete

for each individual material. As a result, the screening discussed in Chapter 4 only involved the full optimization of 1,022 MOFs and took over a year to complete on one of Compute Canada's large high performance computing clusters.

Although roughly 25% of the Computation Ready Experimental (CoRE) MOF database[12] was fully optimized in Chapter 4, over 5.6 million individual process points were tested using the PSA simulator code. This large dataset provided an opportunity to build a surrogate model using sophisticated machine learning techniques. In this context, a surrogate model is a statistical machine learning model which aims to replace the PSA simulator during the optimizations. Since the simulations for each individual process point, or set of operational parameters for each MOF, being run through the PSA simulator is the bottleneck of the screening process, replacing the PSA code with a surrogate model would result in a significant improvement to the speed of optimization and ultimately increase the pace of materials discovery.

The PSA simulator used in Chapter 4 calculates four important performance metrics for each MOF at every set of process conditions: the *Purity* of captured CO_2, the *Recovery* of CO_2 from the flue gas stream, the *Parasitic Energy* of separation which comprises both the energetic cost of running the separation and the compression of captured gas to transport conditions (150 bar), and the *productivity* of the material. To simplify the work of building a surrogate model to predict these four outputs, a suite of four machine learning models was developed, each predicting one of these important output values. Although the parasitic energy is the sum of the energy required to run the capture unit and the energy to compress the captured gas to transport conditions, only the energy of the capture unit is included in the surrogate model. This is because the energy of compression is entirely dictated by the purity of the captured CO_2 (see Chapter 4), and as such predicting the energy of compression and the purity of captured CO_2 would be redundant.

In this chapter, four artificial neural network models were developed to predict the purity, recovery, parasitic energy (excluding compression), and productivity of a material given any set of process conditions. These models were then combined into a suite and connected to a genetic algorithm (GA) to optimize the process conditions for each MOF. This suite of artificial intelligence models coupled with the GA, called the **Fo**ssil fuel combustion carbon **C**apture **A**nd **S**torage (FoCAS) A.I., will be made available for use by researchers on the website of the research group of Professor Woo.

Using FoCAS, the entire 2014 CoRE database[12] (4,433 MOFs) was fully optimized to maximize productivity and minimize the total parasitic energy (including compression) for each MOF. This was done while constraining the purity and recovery to meet the US Department of Energy's Purity-Recovery Targets (DoE-PRT) of 95% purity of captured CO_2 and a recovery of at least 90% of CO_2 from the flue gas,[13] the minimum requirement for a material to be considered viable for the post-combustion carbon capture process.

8.3. Methodology

8.3.1. Dataset

Over the course of the work performed in Chapter 4, a large dataset consisting of 5,653,350 unique process points were run using the Pressure Swing Adsorption (PSA) simulator. To train the neural networks, the dataset was randomized and 80% of the process points were used for training. The remaining 20% was separated into two equal subsets to be used as a development and test set. Separate models were developed to predict the purity of the captured CO_2, recovery of CO_2 from the flue gas, the productivity of the separation, and the parasitic energy of separation (excluding compression). The features used to train the code are identical to those used as inputs into the PSA simulator and can be separated into two distinct categories: material specific features and the operating conditions of the PSA column. Among the material specific features are the 8 fitted isotherm parameters for CO_2 and N_2, the structured density of the materials, and the internal energies for CO_2 and N_2 derived from the isosteric heats of adsorption. All material specific features were directly pulled from the work performed in Chapter 4. The operating conditions include the 7 process parameters which control the 4-stage light particle pressurization process: inlet temperature, intermediate pressure, evacuation pressure, initial flow rate, and times spent on each step of the cycle. In total 20 features were used to describe each process point: the full list of features and their descriptions can be found in Appendix 8.1.

8.3.2. Removing Outliers

To improve the models, outliers in the parasitic energy fittings were removed from the dataset. These outliers were removed for being several orders of magnitude larger than the mean, exceeding a value of 1.2×10^{11} kWh / tonne CO_2 (excluding compression energy). Since parasitic energies of this size are too large to ever be considered viable for use in a PSA system, accurately predicting values in this range does not provide the user with any useful insights. As such, a filter was applied to the parasitic

energy points, removing the upper 10% of the distribution. This value was selected to reduce the maximum parasitic energy in the training data to 329.49 kWh / tonne CO_2 (excluding compression energy), as any material with a parasitic energy that falls above this value will not be considered viable. This reduced the new training set for parasitic energy fittings from 5,653,350 to 5,088,015 points. Removal of outlier points was only performed in the parasitic energy fittings, as no outliers were found for *purity, recovery,* or *productivity*.

8.3.3. Fitting Subsets

8.3.3.1. Initial Fitting Attempts

As mentioned in section 8.3.1, the full dataset used in this work included over 5.6 million individual process points. Initial attempts were made to generate neural network models using the full dataset for training, development, and validation, for all four target performance metrics. Although the models generated using this set resulted in high Pearson R^2 values ranging from 0.940 to 0.997, it was discovered that the models regularly overpredicted the purity and recovery values resulting in a greater number of points meeting the DoE-PRT. This overprediction regularly occurred despite the high R^2 values calculated on the test sets. This behaviour is easily explained through consideration of the imbalance found within the purity and recovery values used in the training of these models. Since the optimizations performed in Chapter 4 heavily penalized process points which fell below the DoE-PRT, the dataset of sampled points is heavily skewed towards high purity and recovery. This is demonstrated in Figure 8.1a and b for purity and recovery, respectively. Since so many of the process points used in the fittings favoured high purity and recovery, the Pearson R^2 calculated for the predictions on the test set was artificially inflated. It became apparent that subsets would need to be created to balance the distributions of the purity and recovery values to improve the models' performance.

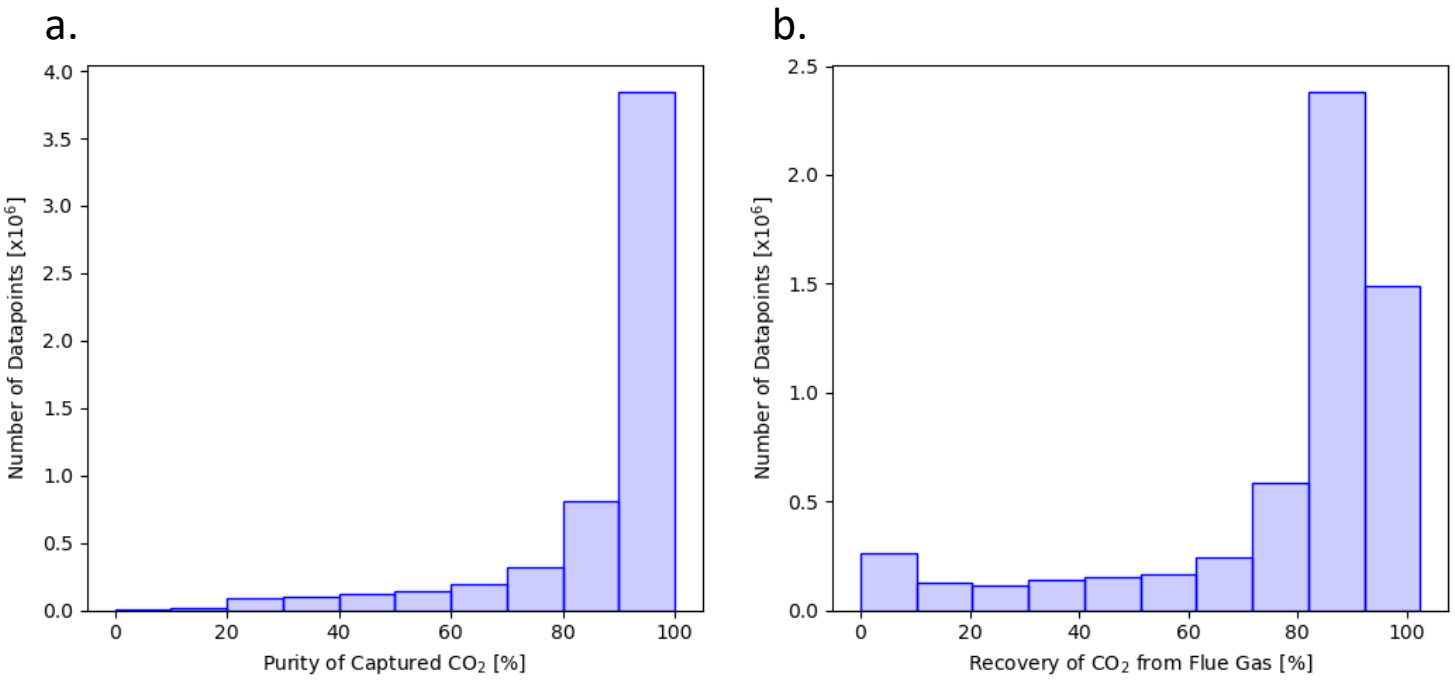


Figure 8.1 Histograms of the (a) Purity of captured CO_2 and the (b) Recovery of CO_2 from flue gas in the 5.6 million process points used in the initial fittings of the neural network models.

8.3.3.2. Purity Subset

The level of imbalance in the initial dataset when dealing with the purity of the captured CO_2, shown in Figure 8.1a, resulted in a significant challenge in flattening the distribution. In the original dataset, 68% of the process points had purity values greater than 90% while only 6% of the process points had purity values below 50%. To remedy this imbalance, the dataset was separated into 10 equal bins, and a maximum of 12,000 values were selected at random from each bin. The distribution of purity values in the new subset generated from this sampling is shown in Figure 8.2, which has a more evenly distributed set of purity values when compared to Figure 8.1a. Due to the heavy imbalance in the initial dataset, this sampling resulted in a large reduction of datapoints available for training, with the entire subset containing 107,046 individual process points, well below the total number of 5.6 million. This subset was further partitioned into training, development, and test sets using the same ratios discussed in section 8.3.1.

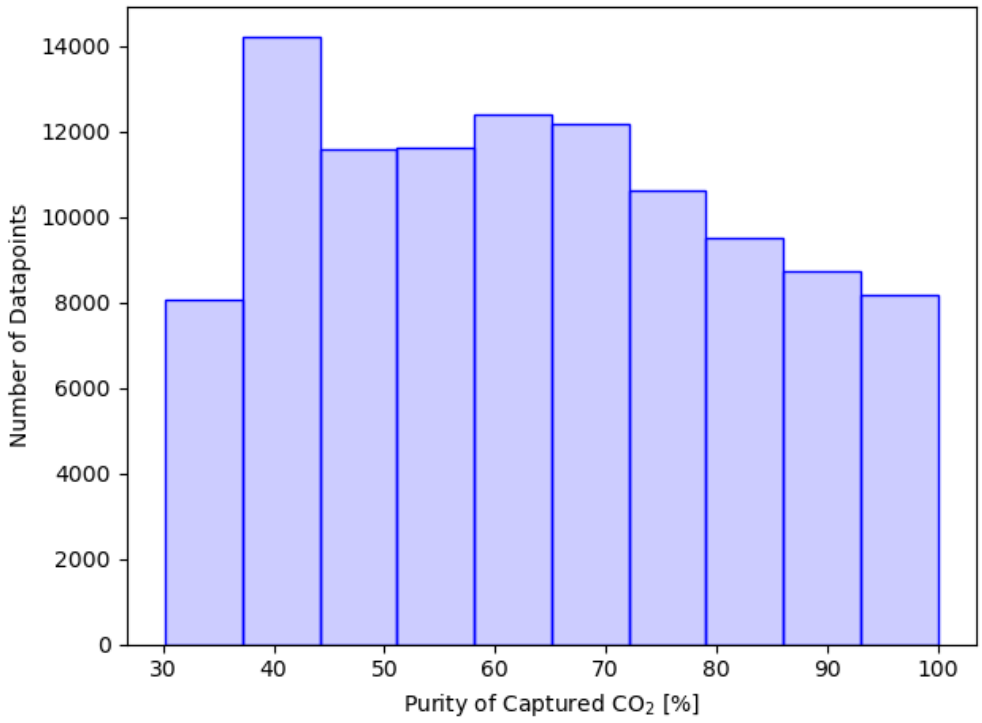


Figure 8.2 Histogram of the purity values present in the new purity subset used to train the neural network model for purity.

8.3.3.3. Recovery Subset

A similar imbalance was found in the values for CO_2 recovery in the initial dataset, although to a lesser degree when compared to the purity of captured CO_2. The distribution of recovery values shown in Figure 8.1b shows higher frequencies of recovery values in the lower range when compared to the purity in Figure 8.1a. This higher frequency allowed for a larger subset to be created to model the recovery values. The full dataset was partitioned into 10 evenly sized bins based on the recovery values and the subset was generated by random selection of a maximum of 50,000 datapoints per bin. The final subset consisted of 500,000 datapoints and the distribution of recovery values is shown in Figure 8.3. This subset was further partitioned into training, development, and test sets using the same ratios discussed in section 8.3.1.

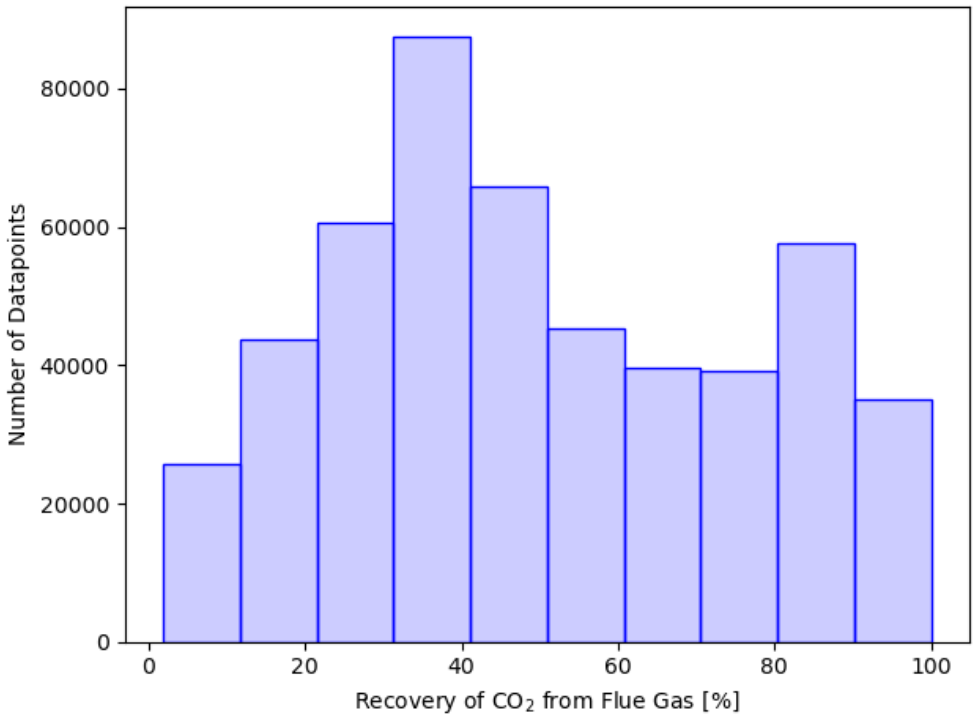


Figure 8.3 Histogram of the recovery values present in the new recovery subset used to train the neural network model for recovery.

8.3.3.4. Parasitic Energy Subset

Initial attempts at fitting neural network models to predict the parasitic energy of the PSA process (excluding compression) yielded lower Pearson R^2 values with a maximum value 0.94 when compared to the other three performance metrics. A decision was made to reduce the training set size and focus on process points which met the DoE-PRT. This decision was made based on the knowledge that any process points that do not meet the DoE-PRT are not considered viable and as a result calculating an accurate parasitic energy for those points is essentially meaningless. The model was therefore fit using only process points which met the DoE-PRT and included 1,050,450 process points. The distribution of parasitic energies is shown in Figure 8.4. Although this distribution is not flat, the range of values in this distribution are spread over a much wider range than seen for the initial purity and recovery sets, resulting in a more balanced dataset. Further, the decision was made to not flatten this distribution as the value added from the inclusion of additional datapoints was deemed to be of greater importance when dealing with this more challenging regression problem. Such a distribution will yield a model that

favours predictions in the range of 150 to 225 kWh / tonne CO_2 (excluding compression), a much larger range of values when compared to purity and recovery, and which will result in the predictions being more conservative. This subset was further partitioned into training, development, and test sets using the same ratios discussed in section 8.3.1.

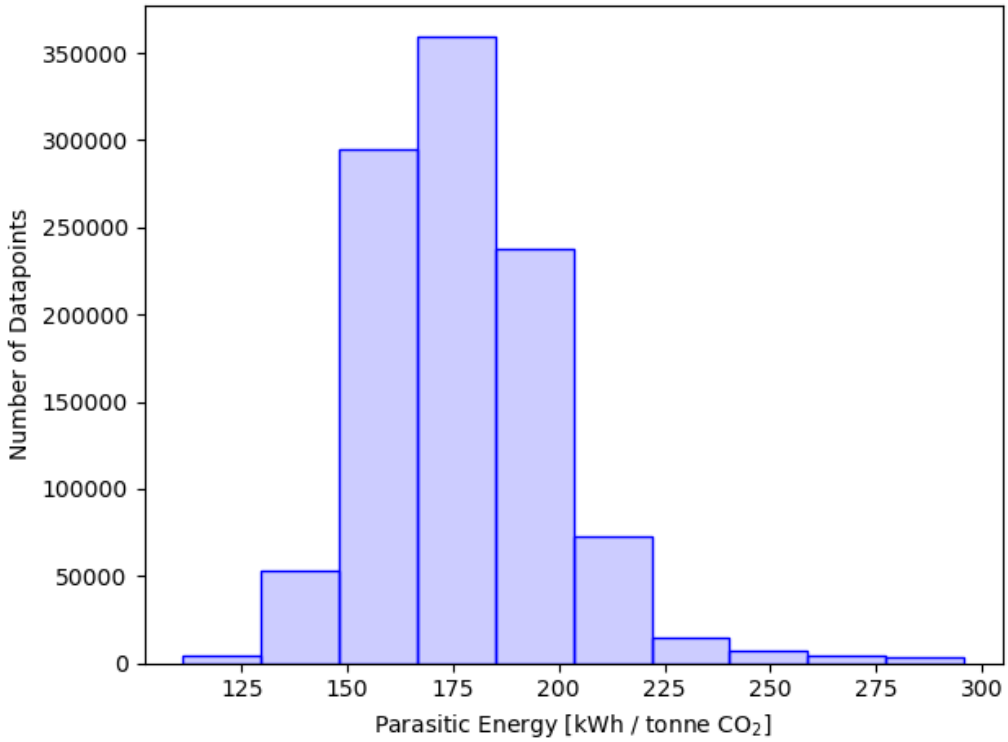


Figure 8.4 Histogram of the parasitic energy values (excluding compression) present in the new recovery subset used to train the neural network model for parasitic energy.

8.3.3.5. Productivity Subset

Although some initial success was obtained from using the entire dataset for the models used predict of the productivity of individual process points, a subset was created including only the points which met the DoE-PRT to be consistent with the methodology used for the parasitic energy points. This subset included 1,050,450 individual process points with a distribution shown in Figure 8.5. This subset was further partitioned into training, development, and test sets using the same ratios discussed in section 8.3.1.

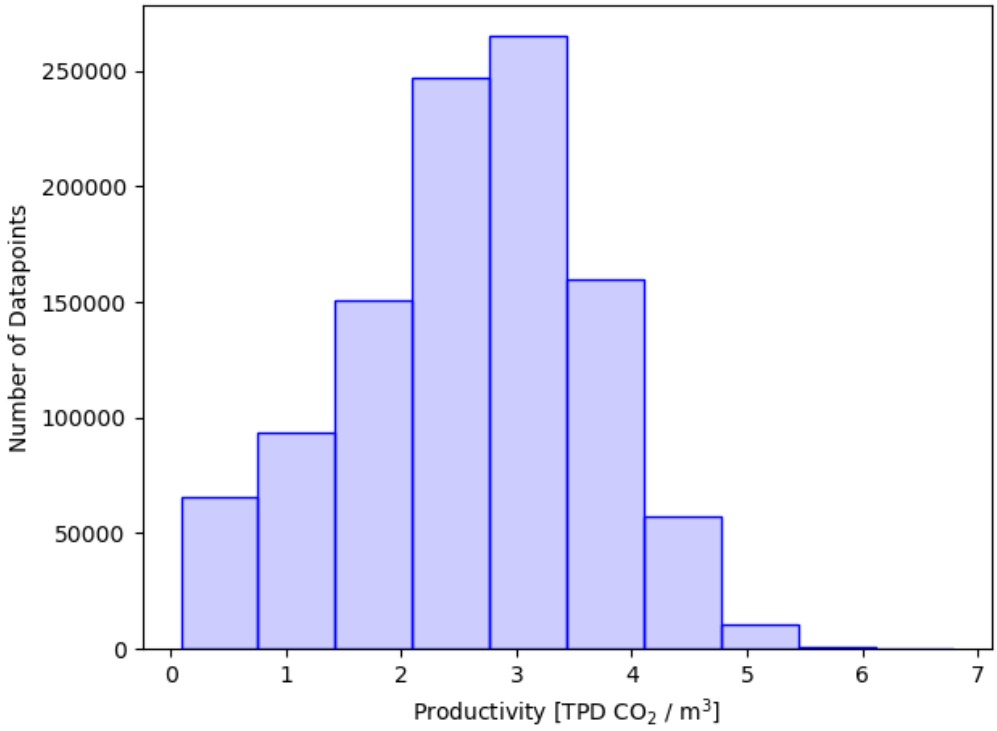


Figure 8.5 Histogram of the productivity values present in the new recovery subset used to train the neural network model for productivity.

8.3.4. Scaling Features

To ensure features with larger absolute values but lower relative variance did not dominate the fittings, the features were scaled to ensure important features would be properly identified. Feature scaling was performed using the standard scaler in the Scikit-Learn Preprocessing package,[14] in Python. To improve fittings and feature distributions, several features were $\log_{10}$ transformed prior to scaling. The full list of features, and their transformations, can be found in Appendix 8.1.

8.3.5. Feature Correlation

Model fittings which include an excessive number of features may suffer from an abundance of noise in the feature space, which may lead to poor model performance. As such, the feature set was analyzed for redundant descriptors, with the aim of removing excess noise from the fittings. This was done by checking for feature correlation, operating under the assumption that a feature which is found to be highly correlated to another feature in the set may be removed to reduce the noise in the fitting

without sacrificing performance. To determine whether strong correlations existed between features, a Pearson R^2 was calculated for every pair of features in the dataset. This would allow for the removal of redundant features and improve the fittings. This analysis was performed using code written in Python and the Pearson R^2 values were calculated using the SciPy Stats package.[15]

8.3.6. Gradient Boosted Decision Trees

To gauge the feasibility of building surrogate models for this process, Gradient Boosted Decision Tree (GBDT) regressors were fit using a Learning Rate of 0.1 and 100 estimators. These models were written in Python and relied on the Scikit-Learn Ensemble packages.[14] These models provided a lower bound of possible model performance, operating on the assumption that an artificial neural network will be able to reproduce or improve upon a less complex machine learning technique such as a GBDT regressor.

8.3.7. Neural Networks – Multi-Layer Perceptron

The neural networks fit over the course of this work were written in Python using the PyTorch package.[16] The networks structures ranged from 1 to 4 hidden layers, each containing 5 to 20 hidden nodes. The non-linear transformations were performed using the Linear Rectifier (ReLU) function. The fittings were run for a maximum of 20,000 epochs using the Adam optimizer to minimize the L1 Loss Function (mean absolute error). To reduce the risk of overfitting, a drop-out probability was implemented into the networks, randomly assigning zeroes to the weights for individual nodes.

8.3.8. Iterative Grid-Search

To improve initial model fittings, the hyperparameters needed to be optimized. The hyperparameters included the number of layers, the number of nodes per hidden layer, the dropout probability, and the learning rate. To perform this optimization, an iterative grid search was performed using a custom code written in Python. The ranges for all hyperparameters were partitioned into 3 grid-points, generating a grid containing 3^6 unique points. Once all points in the 6-dimensional grid were run, the ranges were halved, and a new grid was generated centred on the best point from the previous step. The optimization was considered complete after an iteration did not improve upon the previous best solution. The full range of allowable values for the 6 hyperparameters can be found in Appendix 8.2.

8.4. Results and Discussion

8.4.1. Model Development

8.4.1.1. Feature Correlation

To determine whether any correlation exists between the features or the target values prior to model fittings, Pearson R^2 correlation coefficients were calculated. The resulting correlations are shown in Figure 8.6, with the highest R^2 value of 0.47 for a single pair indicating poor correlation between the input features. This result indicates that none of the input features are redundant and will all be needed in the development of the surrogate models. Additionally, the lack of correlation between the input features and the target values indicates that no single feature can be used to predict the target values.

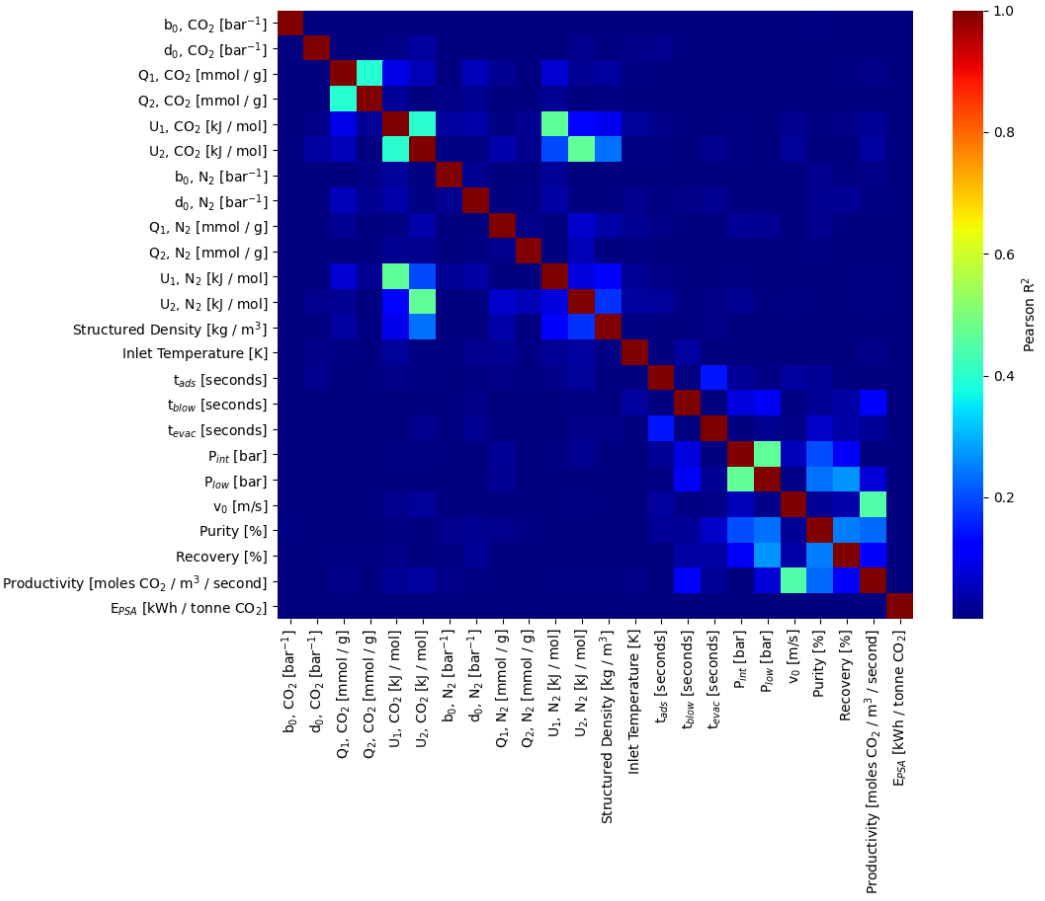


Figure 8.6 Heatmap of the feature pairs, where each grid entry is the Pearson R^2 correlation coefficient for each feature pair.

8.4.1.2. Gradient Boosted Decision Trees

The low correlations between the features and the target variables indicate that no simple relationship exists that can be exploited by a surrogate model. Although no simple relationship exists, since the Pressure Swing Adsorption (PSA) simulator relies solely on these features to predict all four target values, a more complex relationship must exist. To date, the best machine learning paradigm for modelling complex relationships is the Artificial Neural Network (ANN), however such methods are computationally expensive, requiring the use of powerful Graphical Processing Units (GPUs). The expected expense in the development of such models needs to be justified by proving such relationships exist prior to fitting any ANNs.

To prove the existence of these complex relationships, Gradient Boosted Decision Trees (GBDT) were fit to the full data set with a generalized set of hyperparameters. The results of the fittings are shown in Table 8.1 for the training and test sets. Even without optimization, all four target values exceed test set R^2 values of 0.78 which indicates the complex relationship can be modelled using machine learning. Although the performance of the GBDT models indicate a strong correlation can be achieved, the performance of these models is too low to be used as an effective surrogate model to the PSA simulator. This justifies the use of more complex models, such as Artificial Neural Networks, to achieve the desired performance.

Table 8.1 Table of the R^2 coefficients and mean absolute errors (MAE) showing the prediction accuracy of the Gradient Boosted Decision Trees for the training and test sets of the four target variables.

Target Value	Training Set R^2	Test Set MEA	Test Set R^2	Test Set MAE
Purity [%]	0.84	4.242	0.84	4.240
Recovery [%]	0.82	6.784	0.82	6.763
Productivity [mol CO_2 / m^3 / second]	0.90	0.073	0.90	0.073
Parasitic Energy [kWh / tonne CO_2]	0.79	12.394	0.78	12.413

8.4.1.3. Artificial Neural Networks

To model these complex relationships, four ANN models were optimized using an iterative grid-search approach to predict purity, recovery, parasitic energy, and productivity. The plots of the predicted vs actual values for all four targets from the optimized ANN model test sets are shown in

Figure 8.2, and the Pearson R^2 values for the test sets are shown in Table 8.7. Using these advanced models, all four targets were predicted with Pearson R^2 value exceeding 0.97, indicating high accuracy of the predicted results. These models can be combined with a custom genetic algorithm, allowing a researcher to screen a material for post-combustion carbon capture in a matter of minutes instead of days.

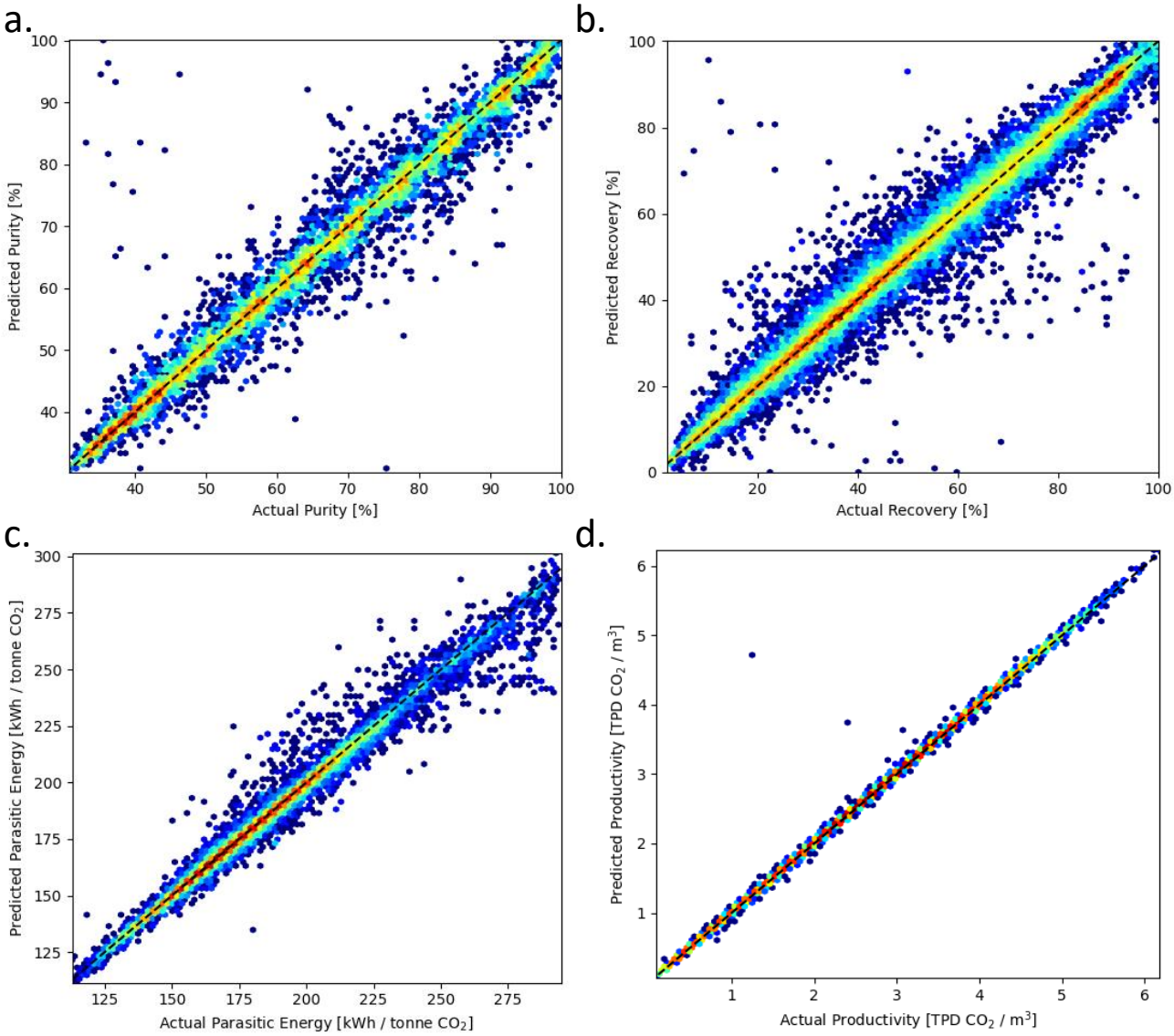


Figure 8.7 Heatmaps of the test set predicted vs actual values from the ANN models for (a) purity, (b) recovery, (c) parasitic energy, and (d) productivity. The 1:1 line is shown as a black dashed line while the colourmap is shown in the Log_{10} scale.

For all four neural networks, the results from the iterative grid-search optimized to the same parameters: 4 layers with 50 hidden nodes per layer, a dropout probability of 0, and a learning rate of

0.001. This result is interesting as these values represent the extremes of the allowable ranges, meaning that a better model could potentially be obtained by allowing the optimizer to explore the space beyond those boundaries. The practical consideration with such a test, however, is that beyond this range the memory requirements for the model fittings becomes too cumbersome and the code crashes on the computing resources available. Given the high performance of the four neural network models presented in Table 8.2, the performance at these extremes was deemed to be sufficient.

Table 8.2 Table of the R^2 coefficients and mean absolute errors (MAE) showing the prediction accuracy of the optimized ANNs for the training test sets of the four target variables.

Target Value	Test Set R^2	Test Set MAE
Purity [%]	0.9730	1.672
Recovery [%]	0.9840	1.694
Productivity [TPD CO_2 / $m^3_{adsorbent}$]	0.9997	0.012
Parasitic Energy [kWh / tonne CO_2]	0.9918	0.985

8.4.1.4. FoCAS A.I. Application

To package and distribute these models, a simple application was designed to allow a user to rapidly optimize the pressure swing adsorption conditions. This application was called the **Fo**ssil fuel combustion carbon **C**apture **A**nd **S**torage (FoCAS) A.I. and relies on a customizable genetic algorithm to test different conditions to locate the best Parasitic Energy, Productivity, Purity, Recovery, or any combination of those four values for a material. A screenshot of the application's graphical user interface is shown in Figure 8.8 and takes as input the fitted single component dual-site Langmuir isotherm parameters for the material at 298.15 Kelvin, the isosteric heat of adsorption, and the material's crystal density. This software will be made available on the Woo lab's website for any MOF researcher to vet new materials rapidly and efficiently, reducing the time to optimize a single material from 5 days to under 3 minutes.

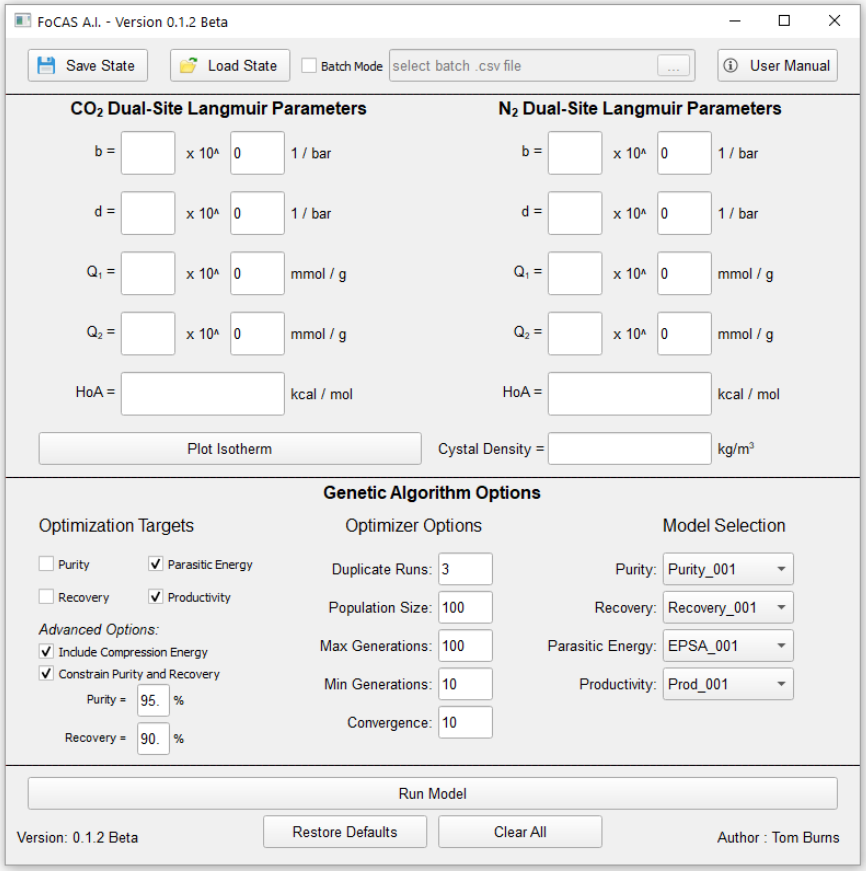


Figure 8.8 Screenshot of the FoCAS A.I. application window.

8.4.2. Comparisons to Detailed PSA Simulations

Prior to running a full-scale screening of the CoRE database using the FoCAS A.I. algorithm, the outputs from FoCAS A.I. and the detailed PSA simulator were compared for a set of process points. For this test, the 20 best parasitic energy, productivity, and overall fitness points were extracted from sample calculations. Single point calculations were performed using the PSA simulator to calculate the purity, recovery, parasitic energy (excluding compression), and productivity for all 60 process points. The results from this spot check are shown in Figure 8.9a, b, c, and d for the purity, recovery, parasitic energy, and productivity, respectively. These results show that although work was done in selecting appropriate training subsets for the parasitic energy and recovery, the models overpredict the process points' ability to meet the DoE-PRT, demarked by the green dashed line in Figure 8.9a and b. Although

better agreement is seen in Figure 8.9c, FoCAS still regularly underpredicts the parasitic energy. Finally, the productivity, shown in Figure 8.9d, shows excellent agreement between the FoCAS points and the PSA simulator.

The overprediction of a MOF's ability to meet the DoE-PRT provides an explanation for the exceptionally high productivity located by FoCAS. Since the constraints of purity and recovery are relaxed, lower parasitic energies and higher productivity can be achieved. Although there is some misclassification with respect to a MOF's ability to meet the DoE-PRT and some under-estimations of the parasitic energies, the main goal of the FoCAS algorithm was to provide a rapid tool to researchers as an initial screening for large databases of materials. As such, it should be noted that the rankings from the FoCAS screening located many of the high performing materials identified from Chapter 4 including IISERP-MOF-2 and PURRIE. This is further evidenced by comparing the rankings of the top 100 materials according to parasitic energy, productivity, and overall fitness from the screening in Chapter 4 to the one performed using FoCAS. To perform this comparison, the rankings from Chapter 4 are compared to the corresponding rankings calculated using FoCAS. These results are shown in Table 8.3 and show that when ranked by parasitic energy 49 of the best MOFs from Chapter 4 are in the top 100 of the FoCAS ranking, 70 in the top 200, and 81 in the top 300. The comparison for the rankings of overall fitness is comparable to the parasitic energy, however the productivity sees a significant drop when searching for high performers. This drop is again likely due to the models' overestimation of a process point's ability to meet the DoE-PRT, allowing for much higher productivity values. Regardless, many of the high performing materials are successfully identified by FoCAS. This means that researchers would be able to use FoCAS to screen a large database of materials and generate a prioritized list of MOFs for high throughput screening with more comprehensive mechanistic models, such as the PSA simulator.

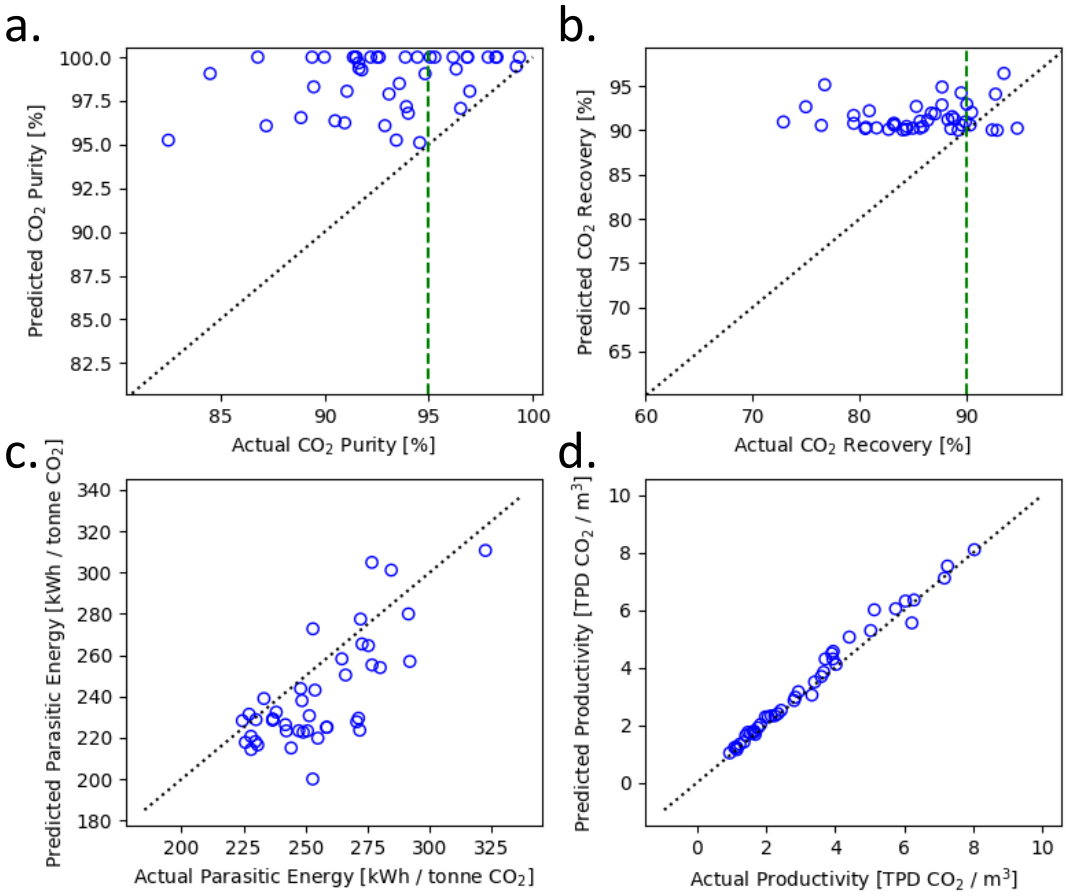


Figure 8.9 Comparison of the predicted vs the actual (a) purity, (b) recovery, (c) parasitic energy (including compression), and (d) productivity. In this figure, the predicted values are calculated using the FoCAS surrogate model and the actual values are calculated using the PSA simulator.

Table 8.3 The number of the top 100 materials based on the rankings from the screening of 1,022 MOFs in Chapter 4 according to parasitic energy, productivity, and overall fitness in the top 100, 200, and 200 rankings of those same MOFs using FoCAS.

	Number of MOFs from Chapter 4 in FoCAS rankings		
Chapter 4 Top 100 ranked by	***Top 100***	***Top 200***	***Top 300***
Parasitic Energy [kWh / tonne CO_2]	49	70	81
Productivity [TPD / m^3]	3	22	46
Overall Fitness [-]	35	59	75

8.4.3. Screening of CoRE Database

A large-scale screening of the CoRE database was performed using FoCAS. In this screening, 4,433 materials were fully optimized using the ANN models described here-in by means of a genetic algorithm with the same fitness functions described in Chapter 4, to maximize the productivity of the material and minimize the parasitic energy. This fitness function included penalty functions to ensure the genetic algorithm favoured points which met the DoE-PRT.

The results from the optimization of the CoRE database found that 3,633 MOFs are able to meet the DoE-PRT, nearly 82% of the MOFs in the CoRE database. The points corresponding to the best parasitic energy, best productivity, and best overall fitness for each of the 3,633 MOFs is shown in the heatmaps in Figures 8.10a, b, and c, respectively. Although over 3600 MOFs were found to meet the DoE-PRT, only 459 MOFs were found that would also meet the DoE's parasitic energy target of 258 kWh / tonne CO_2.[13]

Recall that during the PSA screening performed in Chapter 4, approximately 50% of the 1,022 MOFs which underwent full optimization were able to meet the DoE-PRT. When comparing the results from the FoCAS algorithm for the same set of 1,022 MOFs studied in Chapter 4, 46% are able to meet the DoE-PRT. This result is consistent with the results of the previous screening; however, it should be noted that many of these materials may have been used in fitting the models albeit with randomly selected process points. Although the individual process points are generated using the genetic algorithm, 11 out of the 20 features for these materials can be found in the training set. This is the result of the training being performed on set randomly selected process points from the original screening, and not on a per-MOF basis. That means that to fully test the efficacy of these models, a separate series of calculations to test the MOFs not included in the initial set would need to be optimized using the PSA simulator and the results compared to those found by FoCAS. Since the materials optimized in Chapter 4 were prioritized according to their potential for high performance, the value of 82% may have several implications. The first is that the assumptions made in Chapter 4 to rank materials may not be valid. The second and more probable implication is that the artificial neural network models used in this screening tend to overpredict a material's ability to meet the DoE-PRT. Exploration of the assumptions used to vet materials and prioritize them for screening, as well as the in-depth comparison of FoCAS results for new materials to their PSA process performance could be the subject of future work, however it would be a significant project and is beyond the scope of this chapter.

The 10 best materials found over the course of this screening ranked by their best parasitic energy process points, productivity process points, and overall fitness process points are presented in Table 8.4a, b, and c, respectively. Of particular interest are the values shown in Table 8.4a, which displays 3 MOFs with parasitic energy points below that of IISERP-MOF-2. As IISERP-MOF-2 was determined to be the MOF with the lowest parasitic energy in the screening from Chapter 4, this result is significant. Additionally, both materials with CCDC designations OKIVIO (200.007 kWh / tonne CO_2) and PURRIE (215.09 kWh / tonne CO_2) were present in the initial 1,022 set with parasitic energies of 249.56 and 231.22 kWh / tonne CO_2, respectively. The energies located by FoCAS are therefore significantly lower than those found during the initial PSA optimizations. Additionally, the highest productivity MOF located using FoCAS, PULDOQ, obtained a productivity value of 10.30 tonnes CO_2 captured per day (TPD) / m^3 of adsorbent. This is in contrast to the best productivity located during the screening in Chapter 4 of 6.81 TPD / m^3, although it should be noted that PULDOQ was not present in the original set of 1,022 optimized MOFs.

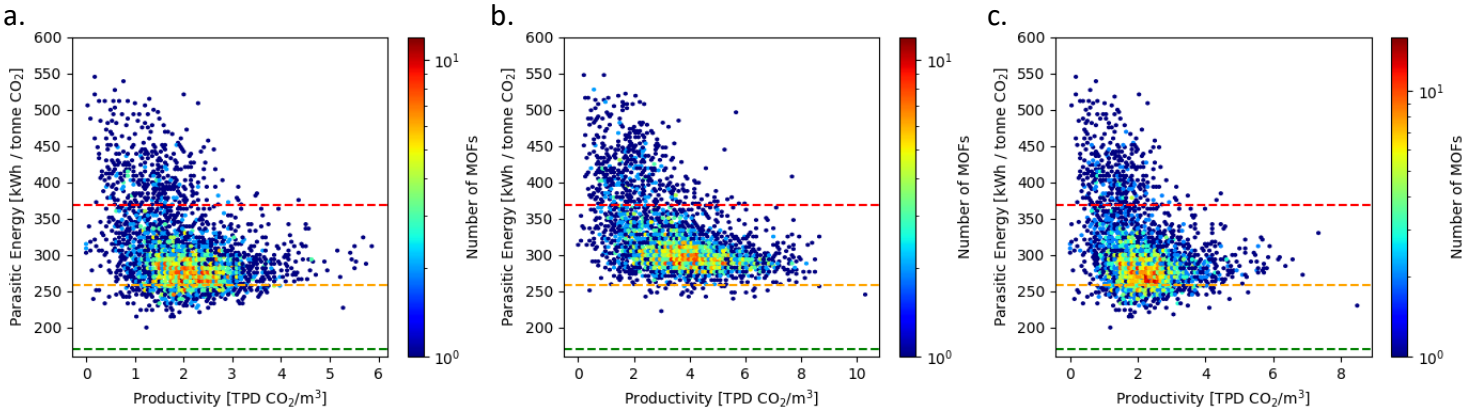


Figure 8.10 Heatmaps of the parasitic energy vs productivity for MOFs found to meet the DoE-PRT showing (a) the best parasitic energy process points, (b) the best productivity process points, and (c) the best overall fitness process points. Included in the figure is the theoretical thermodynamic limit for the parasitic energy (green line), the DoE target for the parasitic energy (orange line), and the parasitic energy from a retrofitted liquid amine capture system (red line).

Table 8.4 The parasitic energy and productivity of a single process point for the top 10 MOFs ranked by (a) parasitic energy, (b) productivity, and (c) overall fitness.

a.

MOF	*Parasitic Energy [kWh / tonne CO_2]*	*Productivity [TPD CO_2 / m^3]*
OKIVIO	200.07	1.22
MIZKOW	214.35	1.71
PURRIE	215.09	1.79
IISERP-MOF-2	216.55	1.03
TAMJEY	217.81	1.75
MOYNEU	219.82	1.90
GUSJOU	220.73	2.32
BERFIP	223.37	1.33
HICVUM	223.40	1.22
ZIF-3	223.44	1.64

b.

MOF	*Parasitic Energy [kWh / tonne CO_2]*	*Productivity [TPD CO_2 / m^3]*
PULDOQ	244.40	10.30
JUCXEK	256.91	8.67
IJASES	324.34	8.67
CAXZOS	275.46	8.48
HAKCAZ	293.22	8.48
YOCMAF	317.36	8.44
VULKIX	277.84	8.29
IDUDIW	287.94	8.21
WECSAZ01	314.97	8.21
NOCLUN	288.20	8.18

c.

MOF	*Parasitic Energy [kWh / tonne CO_2]*	*Productivity [TPD CO_2 / m^3]*
OKIVIO	200.07	1.22
PULDOQ	228.64	8.48
MIZKOW	214.35	1.71
PURRIE	215.09	1.79
IISERP-MOF-2	218.17	3.04
TAMJEY	217.81	1.75
GUSJOU	220.73	2.32
MOYNEU	219.82	1.90
HICVUM	226.35	4.11
FEVDIV	223.67	1.75

8.5. Model Limitations and Domain Considerations

The first and most significant limitation of these models was the overrepresentation of high performing MOFs in the training data. Since materials were identified according to their potential for high performance prior to running any PSA simulations (see Chapter 4), the materials included in the training sets are those that were deemed more likely to be high performing. As a result, an imbalance may exist in the original dataset which may cause the models to over-predict the performance of a material. To mitigate this effect, full optimizations using the PSA simulator code would need to be performed on low performing materials to supplement the dataset and improve the global performance of the models.

Further, as machine learning models perform exceptionally well with interpolation, they often fail to be predictive when there is a need for extrapolation. As such, when using FoCAS A.I., it is important to consider whether the material being tested is within the domain of the training data. The domain range for the material parameters used for all four models can be found in Appendix 8.3. By considering whether a material being tested is within the domain of these models, any issues caused by the over-representation of high performing materials can be reduced.

8.6. Conclusions

A series of artificial neural networks were successfully designed to predict the purity of captured CO_2, recovery of CO_2 from the flue gas stream, the parasitic energy (excluding compression), and the productivity of the material. These neural networks act as a surrogate model to the sophisticated Pressure Swing Adsorption (PSA) Simulator used in Chapter 4 when combined into a suite. Using a

custom genetic algorithm which optimizes process conditions, these codes form an application called the **Fo**ssil fuel combustion carbon **C**apture **A**nd **S**torage (FoCAS) A.I., which will fully optimize the process conditions for a material in roughly 3 minutes, 1,440 times faster than using the PSA simulator. Using FoCAS, the 4,433 materials were fully optimized, and 3,633 MOFs were found to be able to meet the DoE-PRT. Of those 3,633 MOFs, 459 materials were identified that could meet the DoE's parasitic energy target of 258 kWh/tonne CO_2.[13] Although it was shown that the FoCAS application may overpredict the material's ability to meet the DoE-PRT, it remains a powerful tool to be used as a first pass screening of large databases of materials, allowing researchers to rapidly screen and prioritize materials for more in-depth study, and significantly improve the current pace of materials discovery.

8.7. References

1. Burns, T. D., Pai, K. N., Subraveti, S. G., Collins, S. P., Krykunov, M., Rajendran, A. & Woo, T. K. Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models. *Environmental Science and Technology* **54**, 4536–4544 (2020).

2. Bae, Y. S. & Snurr, R. Q. Development and evaluation of porous materials for carbon dioxide separation and capture. *Angewandte Chemie - International Edition* vol. 50 11586–11596 (2011).

3. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **43**, 5735–5749 (2014).

4. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

5. Nugent, P., Giannopoulou, E. G., Burd, S. D., Elemento, O., Giannopoulou, E. G., Forrest, K., Pham, T., Ma, S., Space, B., Wojtas, L., Eddaoudi, M. & Zaworotko, M. J. Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

6. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to metal–organic frameworks. *Chemical Reviews* **112**, 673–674 (2012).

7. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **341**, (2013).

8. Moghadam, P. Z., Li, A., Wiggin, S. B., Tao, A., Maloney, A. G. P., Wood, P. A., Ward, S. C. & Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A collection of metal-organic frameworks for past, present, and future. *Chemistry of Materials* vol. 29 2618–2625 (2017).

9. Jiang, J., Lu, Z., Zhang, M., Duan, J., Zhang, W., Pan, Y. & Bai, J. Higher symmetry multinuclear clusters of metal–organic frameworks for highly selective CO_2 capture. *Journal of the American Chemical Society* **2**, jacs.8b07589 (2018).

10. Yu, J., Xie, L. H., Li, J. R., Ma, Y., Seminario, J. M. & Balbuena, P. B. CO_2 capture and separations using MOFs: Computational and experimental studies. *Chemical Reviews* **117**, 9674–9754 (2017).

11. Krishnamurthy, S., Rao, V. R., Guntuka, S., Sharratt, P., Haghpanah, R., Rajendran, A., Amanullah, M., Karimi, I. A. & Farooq, S. CO_2 capture from dry flue gas by vacuum swing adsorption: A pilot plant study. *AlChE* **60**, 1830–1842 (2014).

12. Chung, Y. G., Camp, J., Haranczyk, M., Sikora, B. J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O. K., Sholl, D. S. & Snurr, R. Q. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **26**, 6185–6192 (2014).

13. Carbon-Capture-Technology-Compendium-2020.

14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

15. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

16. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z. D., Research, A. I., Lin, Z., Desmaison, A., Antiga, L., Srl, O. & Lerer, A. *Automatic differentiation in PyTorch*.

8.8. Appendix

8.8.1. Appendix 8.1: Features used in Neural Networks

Feature Name	Description	Log Transformed?
b_0, CO_2	Generalized fitted Langmuir parameter for the strong site	Yes
Q_1, CO_2	Saturation uptake for the strong site	Yes
d_0, CO_2	Generalized fitted Langmuir parameter for the weak site	Yes
Q_2, CO_2	Saturation uptake for the weak site	Yes
b_0, N_2	Generalized fitted Langmuir parameter for the strong site	Yes
Q_1, N_2	Saturation uptake for the strong site	Yes
d_0, N_2	Generalized fitted Langmuir parameter for the weak site	Yes
Q_2, N_2	Saturation uptake for the weak site	Yes
ΔU, CO_2	CO_2 internal energy	No
ΔU, N_2	N_2 internal energy	No
Structured Density	Structured density of the material, 0.75*crystal density	No
Adsorption Time (sec)	Time spent in the adsorption phase of the cycle	No
Blowdown Time (sec)	Time spent in the blowdown phase of the cycle	Yes
Evacuation Time (sec)	Time spent in the evacuation phase of the cycle	No
Blowdown Pressure (bar)	Lowest pressure achieved during the blowdown phase	Yes
Evacuation Pressure (bar)	Lowest pressure achieved during the evacuation phase	Yes
Feed Velocity (m/s)	Velocity of the flue gas feed entering the separation column	No
Flue Gas Temperature (K)	Temperature of the flue gas stream	No

8.8.2. Appendix 8.2: Neural Network Hyperparameters

Hyperparameter	Description	Maximum	Minimum
Layers	The number of hidden layers in the neural network	4	1
Nodes in Layer 1	The number of nodes in the first hidden layer	20	1
Nodes in Layer 2	The number of nodes in the second hidden layer	20	1
Nodes in Layer 3	The number of nodes in the third hidden layer	20	1
Nodes in Layer 4	The number of nodes in the fifth hidden layer	20	1
Dropout Probability	The probability of the output from a neuron in network will be ignored	1.0	0.0
Learning Rate	Tuning parameter that controls the size the adjustment along the calculated gradient during fittings	10^{-5}	10^{-2}

8.8.3. Appendix 8.3: Domain of the four Neural Network Models

Feature Name	Purity [%]		Recovery [%]		Parasitic Energy [kWh / tonne CO_2]		Productivity [TPD / m^3]	
	Max	Min	Max	Min	Max	Min	Max	Min
b_0, CO_2 (m^3/mol)	0.0078	2.4×10^{-10}	0.0078	2.38×10^{-10}	0.0078	2.38×10^{-10}	0.0078	2.38×10^{-10}
Q_1, CO_2 (mmol / g)	373.75	0.0	373.75	0.0	150.0	0.0	150.0	0.0
d_0, CO_2 (m^3/mol)	6.9×10^{-6}	0.0	6.9×10^{-6}	0.0	6.52×10^{-6}	0.0	6.52×10^{-6}	0.0
Q_2, CO_2 (mmol / g)	386.38	0.0	386.39	0.0	350.03	0.0	350.03	0.0
b_0, N_2 (m^3/mol)	0.083	3.7×10^{-8}	0.083	3.69×10^{-8}	0.083	9.48×10^{-8}	0.083	9.48×10^{-8}
Q_1, N_2 (mmol / g)	55.00	0.0	55.0	0.0	55.0	3.17×10^{-9}	55.0	3.17×10^{-9}
d_0, N_2 (m^3/mol)	0.0009	0.0	0.00093	0.0	0.00033	0.0	0.00033	0.0
Q_2, N_2 (mmol / g)	43.18	0.0	53.75	0.0	53.75	0.0	53.75	0.0
ΔU, CO_2 (J / mol)	56165.66	8425.66	56165.66	8425.66	56165.66	8425.66	56165.66	8425.66
ΔU, N_2 (J / mol)	24703.77	574.34	24703.77	574.34	24225.58	574.34	24225.58	574.34
Structured Density (kg / m^3)	1911.99	243.75	1911.99	243.75	1888.87	321.75	1888.87	321.75

9. Chapter 9: Conclusions and Future Work

9.1. Conclusions

The main goal of the work presented in this thesis was to study metal-organic frameworks (MOFs) for post-combustion carbon capture and storage in coal-fired powerplants by bridging the gap between materials science and process engineering. This was accomplished using a series of simulation techniques, which allowed the study of gas adsorption in MOFs from their crystal structures to their industrial pressure swing adsorption performance. These techniques aimed to find materials that minimize the cost of capture, by reducing the parasitic energy while maximizing the materials' productivity. The secondary goal of this work was to study the process performance of MOFs to determine what chemical features or materials science performance metrics result in energetically inexpensive and efficient CO_2 capture in industrial systems.

9.1.1. Chapter 3: Guest Atom Localization Algorithm (GALA)

In Chapter 3, a technique for studying the binding motifs of gas molecules called the Guest Atom Localization Algorithm (GALA) was developed and optimized for CO_2, due to its importance in carbon capture and storage. The methodology used to locate binding sites based on Grand Canonical Monte Carlo simulations was described in detail, along with an explanation of the various control parameters needed to accurately locate binding sites in MOFs. These control parameters were optimized for CO_2 to allow the use of the GALA code in high-throughput screening applications. The full optimization of these parameters was described in detail, with a demonstration of GALA's accuracy using a series of MOFs with experimentally determined binding sites.

9.1.2. Chapter 4: Pores to Process

In Chapter 4, a large-scale screening of MOFs, studying the gas adsorption properties from the crystal scale to industrial pressure swing adsorption systems was performed on the Computation Ready Experimental (CoRE)[1] MOF database. In this study a tiered approach was used to screen over 4,000 MOFs for use in a pressure swing adsorption system to filter CO_2 emissions at coal-fired powerplant conditions. Using only the crystal structures of the materials, Grand Canonical Monte Carlo (GCMC) simulations were used to predict gas adsorption isotherms and isosteric heats of adsorption for CO_2 and N_2. These properties were then used as inputs into a sophisticated pressure swing adsorption simulator, whose process parameters were optimized using a custom genetic algorithm to determine the best parasitic energy, productivity, purity, and recovery points for each MOF. Of the 1,022 materials that underwent a full optimization using the genetic algorithm, 482 were identified that were able to meet

the department of energy's purity and recovery targets of 95% purity of captured CO_2 and 90% recovery of CO_2 from the flue gas stream. Of those 482 MOFs, 223 were identified that were also able to meet the US DoE's target for parasitic energy.[2] A comparison of the resulting process performance for the 1,022 MOFs was then compared to a series of materials science performance metrics. These metrics, commonly used in the field to identify high performing materials,[3–9] were not correlated with the parasitic energy or productivity at the industrial scale.

9.1.3. Chapter 5: Datamining PSA Results

In Chapter 5, the results from the screening in Chapter 4 were studied in more detail to search for complex relationships between the materials science performance metrics and the industrial PSA performance for each material. Using a series of machine learning techniques such as Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and the Random Forest (RF) method, 37 metrics were used to build models to predict parasitic energy, productivity, and the MOF's ability to meet the US DoE's purity-recovery targets (DoE-PRT).[2] It was found that none of the metrics were predictive of the parasitic energy or productivity, however many metrics could be used to estimate the materials ability to meet the DoE-PRT. It was found that some individual metrics were able to distinguish between MOFs which could and could not meet the DoE-PRT with a balanced accuracy of 75 %. Among them, the most important metric identified was the strength of the N_2 binding site defined by the fitted Langmuir parameter. When used in random forest fittings, the N_2 isotherm parameters were able to accurately predict a MOF's ability to meet the DoE-PRT with a balanced accuracy of 83 %. This indicates that an important feature that a material needs to meet the DoE-PRT is weak N_2 binding, contrary to the belief in the materials sciences community that CO_2 binding plays the most important role. The results from Chapter 5 demonstrate that materials science performance metrics, easily calculated using GCMC, can be predictive of process level performance.

9.1.4. Chapter 6: N_2 Binding Sites

In Chapter 6, the insights from Chapter 5 relating the strength of N_2 binding to a material's ability to meet the DoE-PRT, were further explored using GALA. In this chapter, CO_2 and N_2 binding sites were calculated for 704 MOFs, randomly selected from the 1,022 MOFs optimized in Chapter 4. Among these MOFs, 423 were able to meet the DoE-PRT, while 281 could not. The probability distributions, used by GALA to locate binding sites, were compared for both guests. The single-component (single gas) and binary-component (gas mixture) binding environments were compared using a Tanimoto similarity metric, where it was found that MOFs which were able to meet the DoE-PRT showed a significant

change in the N_2 probability distributions when compared to MOFs which did not meet the DoE-PRT. This led to the first conclusion, that the displacement of N_2 binding sites in the presence of CO_2 increases the material's chance of meeting the DoE-PRT. This conclusion led to a hypothesis that the displacement of N_2 binding sites resulted in a decrease in N_2 adsorption, driving the material's ability to separate CO_2 from N_2. By considering the N_2 adsorption ratios, or the ratio of the N_2 binary-component loading to single-component loading, it was noted that N_2 adsorption ratio distributions for MOFs which met the DoE-PRT showed a statistically significant decrease in N_2 adsorption ratios when compared to MOFs which did not. These two metrics were found to be correlated, with a Pearson R^2 of 0.84. Finally, the assumption that that Tanimoto coefficient is related to binding site displacement was validated by visualization of binding sites along a range of Tanimoto values, where it was confirmed that MOFs with low Tanimoto values in the N_2 probability distributions showed significant changes in the N_2 binding sites. This work related the behaviour at the molecular scale to industrial process performance for the first time.

9.1.5. Chapter 7: Linear Interpolation in PSA Simulator

In Chapter 7, an improvement to the Pressure Swing Adsorption simulator code is proposed and explored. This change to the simulation seeks to replace the competitive dual-site Langmuir model based on single-component isotherms with a linear interpolation model fit using binary-component GCMC loadings. The aim was to explore the interpolation method's ability to predict loadings at any given parameters present within a pressure swing adsorption (PSA) and temperature swing adsorption (TSA) column. To fit the linear interpolator models, a grid of adsorption data was generated using GCMC for every MOF in the test sets, which sampled temperature, pressure, and mole fraction ranges present in the separation column. Using Latin Hypercube Sampling (LHS), 1000 points between the grid-points were selected, and the loadings were determined using GCMC, the linear interpolator model, and the competitive dual-site Langmuir model for every MOF in the set. The interpolated and the competitive dual-site Langmuir model loadings were compared to the GCMC using a mean absolute deviation. It was found for both PSA and TSA systems, that the interpolation model more accurately predicted both the CO_2 and N_2 loadings regardless of the coarseness of the grids. This result demonstrates that binary-component GCMC data can be directly used in the PSA simulator, however further work is required to determine the impact such a model would have on the final PSA simulation results.

9.1.6. Chapter 8: FoCAS

In Chapter 8, I explored the use of a surrogate model to replace the PSA simulator, built using a suite of artificial neural networks. In this chapter, I used a dataset consisting of over 5 million unique process simulations to build machine learning models to predict the purity of captured CO_2, recovery of CO_2 from the flue gas, parasitic energy, and productivity of the sorbent material. Prior to the machine learning fittings, the feature space being used to model the separation column was explored, searching for correlated descriptors that could be considered redundant. This search found that none of the features were highly correlated, indicating all descriptors needed to remain for the model fittings. To determine whether the data could be predictive of the target performance metrics, and ensure the cost associated with fitting and optimizing multilayer perceptron neural networks was justified, an initial set of fittings were performed using gradient boosted decision trees. These unoptimized trees were able to reproduce the target process performance metrics on the validation sets with Pearson R^2 values ranging from 0.78 to 0.90. Based on these initial results, a multilayer perceptron neural network was optimized using an iterative grid-search approach for each of the target variables. All four models exceeded Pearson R^2 values of 0.97 when tested on the validation sets, demonstrating that these models could indeed be used to replace the PSA simulator. A piece of software called the **Fo**ssil fuel combustion carbon **C**apture **A**nd **S**torage (FoCAS) A.I. was constructed to run these models. A full screening of the CoRE database was performed to demonstrate the FoCAS application which mirrored the screening in Chapter 4. Although it was found that the model tended to over-estimate the purity and recovery of the materials, it proved a powerful tool to be used as a first pass screening of large databases of materials. By reducing the time needed to fully optimize a single material, FoCAS succeeds in accelerating the pace of materials discovery for Post-Combustion Carbon Capture and Storage.

9.2. Future Work

Although much was accomplished over the course of this thesis, there will always be more that can be achieved. In this section, possible future directions for each project are discussed.

9.2.1. Chapter 3: Guest Atom Localization Algorithm (GALA)

Given additional time and resources on this project, the GALA code could be optimized using the same procedure for a variety of important guest molecules, including O_2, N_2, CH_4, and H_2O. By optimizing the algorithm for a wider variety of guest molecules, the GALA code could be used more dynamically to study a diverse set of systems, including but not limited to, water adsorption studies, methane storage applications, and oxy-fuel separation.

9.2.2. Chapters 4 & 5: Pores to Process

There are several future directions that can be performed based on the work presented in Chapter 4. For example, work presented in this Chapter caught the attention of TotalEnergies. This resulted in a $4 million collaborative research project for the research lab of Dr. Tom Woo which is currently underway performing a similar study for temperature swing adsorption systems. Additionally, although this study was performed considering the full CoRE database at the time the work was performed, that database of MOFs has since been significantly expanded from roughly 4000 MOFs at the time of this study to over 14,000. This work could be continued on, considering the thousands of additional materials now present in the experimental database. Finally, a similar study could be performed on pressure swing adsorption systems, considering alternate cycle configurations to determine whether the 4-stage light particle pressurization is the best choice for industrial PSA systems. Finally, future work performed on this topic can involve the inclusion of an important third guest molecule, water. Water was omitted during the initial study due to the difficulty associated with simulating H_2O isotherms,[10–12] however recent advancements in water simulations may indicate the possibility of including water in future studies.

9.2.3. Chapter 6: N_2 Binding Sites

Additional work could be performed based on the results in Chapter 6 by expanding the test set to improve the confidence in the results. Additionally, the significant overlap between the materials which pass and fail the DoE-PRT implies that the displacement of N_2 binding sites is not the only contributing factor in determining a material's ability to separate CO_2 from N_2. Additional work would need to be performed to explore other possible contributing factors controlling the mechanics of the gas separation.

9.2.4. Chapter 7: Linear Interpolation in PSA Simulator

The work presented in Chapter 7 on the interpolation of multi-component GCMC simulation data has many possible paths forward for future research. Since it was shown that the linear interpolator improved upon the competitive dual-site Langmuir model, the next steps would involve incorporating the interpolation models into the PSA and TSA simulators to determine the overall effects on the simulator outputs. The results from such a study would ultimately determine the next steps, if for example the use of the interpolator does not significantly impact the simulators' outputs, the widespread implementation of this method may not provide any valuable improvements. However, if

the results are significant, these improved simulators could be used to perform future screenings with greater confidence in the simulation results.

9.2.5. Chapter 8: FoCAS

The next steps and possible future directions for this project would be to continue work on model development to improve the purity and recovery predictions. Since the dataset used to build the models currently included in the FoCAS A.I. software suite came from the work performed in Chapter 4, which put significant emphasis on finding materials that met the DoE-PRT, there was a significant imbalance in the dataset in favour of points with high purity and high recovery. Although steps to mitigate this imbalance in the initial fittings were taken, performing simulations using the PSA simulator to locate points in the low and middling purity and recovery ranges would help improve predictions in those ranges and reduce the over-estimation of those targets. Furthermore, the FoCAS algorithm could be used to perform initial screenings of additional MOF databases, including the expanded CoRE database,[1] and the hypothetical Tabasco database.[13]

9.3. References

1. Chung, Y. G., Camp, J., Haranczyk, M., Sikora, B. J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O. K., Sholl, D. S. & Snurr, R. Q. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials* **26**, 6185–6192 (2014).

2. Carbon-Capture-Technology-Compendium-2020.

3. Dzubak, A. L., Lin, L. C., Kim, J., Swisher, J. A., Poloni, R., Maximoff, S. N., Smit, B. & Gagliardi, L. Ab initio carbon capture in open-site metal-organic frameworks. *Nature Chemistry* **4**, 810–816 (2012).

4. Boyd, P. G., Chidambaram, A., García-Díez, E., Ireland, C. P., Daff, T. D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S. M., Maroto-Valer, M. M., Reimer, J. A., Navarro, J. A. R., Woo, T. K., Garcia, S., Stylianou, K. C. & Smit, B. Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature* **576**, 253–256 (2019).

5. Nugent, P., Giannopoulou, E. G., Burd, S. D., Elemento, O., Giannopoulou, E. G., Forrest, K., Pham, T., Ma, S., Space, B., Wojtas, L., Eddaoudi, M. & Zaworotko, M. J. Porous materials with optimal adsorption thermodynamics and kinetics for CO_2 separation. *Nature* **495**, 80–84 (2013).

6. Liang, L., Liu, C., Jiang, F., Chen, Q., Zhang, L., Xue, H., Jiang, H. L., Qian, J., Yuan, D. & Hong, M. Carbon dioxide capture and conversion by an acid-base resistant metal-organic framework. *Nature Communications* **8**, (2017).

7. Jiang, J., Lu, Z., Zhang, M., Duan, J., Zhang, W., Pan, Y. & Bai, J. Higher symmetry multinuclear clusters of metal–organic frameworks for highly selective CO_2 capture. *Journal of the American Chemical Society* **2**, jacs.8b07589 (2018).

8. McDonald, T. M., Mason, J. A., Kong, X., Bloch, E. D., Gygi, D., Dani, A., Crocellà, V., Giordanino, F., Odoh, S. O., Drisdell, W. S., Vlaisavljevich, B., Dzubak, A. L., Poloni, R., Schnell, S. K., Planas, N., Lee, K., Pascal, T., Wan, L. F., Prendergast, D., *et al.* Cooperative insertion of CO_2 in diamine-appended metal-organic frameworks. *Nature* **519**, 303–308 (2015).

9. Chung, Y. G., Gómez-Gualdrón, D. A., Li, P., Leperi, K. T., Deria, P., Zhang, H., Vermeulen, N. A., Stoddart, J. F., You, F., Hupp, J. T., Farha, O. K. & Snurr, R. Q. In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Science Advances* **2**, e1600909 (2016).

10. Sarkisov, L., Centineo, A. & Brandani, S. Molecular simulation and experiments of water adsorption in a high surface area activated carbon: Hysteresis, scanning curves and spatial organization of water clusters. *Carbon* **118**, 127–138 (2017).

11. Ghosh, P., Kim, K. C. & Snurr, R. Q. Modeling water and ammonia adsorption in hydrophobic metal-organic frameworks: Single components and mixtures. *Journal of Physical Chemistry C* **118**, 1102–1110 (2014).

12. Paranthaman, S., Coudert, F. X. & Fuchs, A. H. Water adsorption in hydrophobic MOF channels. *Physical Chemistry Chemical Physics* **12**, 8123–8129 (2010).

13. Boyd, P. G. & Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **18**, 3777–3792 (2016).