COMBINING FREE ENERGY CALCULATIONS WITH MACHINE LEARNING PREDICTIONS OF ADSORPTION TO DESIGN METAL-ORGANIC FRAMEWORKS FOR CHEMICAL SEPARATIONS

by

Ryther M. Anderson

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Chemical Engineering).

Golden, Colorado

Date _____

Signed: _____

Ryther M. Anderson

Signed: _____

Dr. Diego A. Gómez-Gualdrón
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Dr. Anuj Chauhan
Professor and Department Head
Department of Chemical and Biological Engineering

ABSTRACT

The separation of chemical mixtures into their sufficiently pure components is a necessary step in essentially all industrial chemical processes. Presently, the predominant method for industrial chemical separation is distillation, an energy-intensive process. Therefore, the development of less energy-intensive separation technologies can reduce industrial energy usage by a consequential margin. The development of these technologies hinges on the discovery and design of materials (i.e. separation agents) that can selectively interact with mixture components on the molecular level. Metal-organic frameworks (MOFs) are a type of nanoporous crystal that, due to their uniquely tunable pore geometry and chemistry, have particular promise for use as separation agents. MOF tunability results from the modularity of their construction. However, this modularity also results in an intractably large number of candidate MOFs for any given separation. Furthermore, many candidate MOFs are likely to be "hypothetical", meaning they have not been synthesized. A key challenge, then, is to identify, for a given separation, high-performing MOFs that can also be synthesized. Considering the aforementioned structural diversity of MOFs and the variety of conditions encountered in industrial separations, screening MOFs for use as separation agents requires efficient computational prediction of their performance and synthetic likelihood. In this thesis, we explore strategies for the identification of high-performing and thermodynamically accessible MOFs for use in various gas separations. Integral to this work is the development of increasingly universal machine learning models capable of predicting the performance of diverse MOFs in multiple separations. We can use these models both to rapidly identify high-performing MOFs and to derive generalized design rules that can be applied to existing MOF platforms. In addition, we develop methods for the large-scale calculation of MOF free energy, allowing us to derive criteria for MOF synthetic likelihood, at least from a thermodynamic perspective. When used in concert, these methods create a unified approach for the efficient computational discovery and design of MOFs for chemical separations.

TABLE OF CONTENTS

ABSTRACT . . . iii

LIST OF FIGURES . . . x

LIST OF TABLES . . . xxvi

ACKNOWLEDGMENTS . . . xxviii

CHAPTER 1 INTRODUCTION . . . 1

1.1 Chemical Separations . . . 1

1.1.1 Distillation . . . 3

1.1.2 Equilibrium Adsorption Separations . . . 5

1.1.3 Porous Materials for Separations . . . 8

1.2 An Overview of Metal-Organic Frameworks . . . 11

1.3 Computational Screening of MOFs . . . 13

1.3.1 Building MOFs *In Silico* . . . 14

1.3.2 Automated Characterization of MOF Structures . . . 18

1.3.3 Molecular Simulation of MOF Properties . . . 19

1.3.4 Predictive Modeling of MOF Adsorption Performance . . . 21

1.3.5 MOF Synthesizability . . . 22

1.4 MOFs for Gas Separations . . . 26

1.4.1 CO_2 Capture . . . 26

1.4.2 Separation of Alkenes from Alkanes . . . 28

1.4.3 Noble Gas Separation . . . 29

1.4.4 Separations Involving Methane . . . 30

1.5 Improving the Computational Screening of MOFs . . . 32

1.6 Thesis Outline . . . 33

CHAPTER 2 AN OVERVIEW OF COMPUTATIONAL METHODS . . . 35

2.1 Molecular Simulation . . . 35

2.1.1 Background . . . 35
2.1.2 Force Fields . . . 41
2.1.3 Grand Canonical Monte Carlo . . . 43
2.1.4 Equilibrium Molecular Dynamics . . . 44
2.1.5 Thermodynamic Integration . . . 46
2.1.6 Density Functional Theory . . . 47
2.2 Machine Learning . . . 49
2.2.1 Background . . . 49
2.2.2 Tree-based Methods . . . 49
2.2.3 Multilayer Perceptrons . . . 51
2.2.4 Model-based Clustering . . . 54
CHAPTER 3 AN ALGORITHM FOR THE TOPOLOGICALLY GUIDED CONSTRUCTION OF MOFS AND THE IMPORTANCE OF TOPOLOGICAL DIVERSITY . . . 56
3.1 Abstract . . . 56
3.2 Introduction . . . 57
3.3 Computational Synthesis Methodology . . . 60
3.4 General Methods . . . 65
3.4.1 MOF Library Construction . . . 65
3.4.2 MOF Geometry Optimization . . . 67
3.4.3 Grand Canonical Monte Carlo Simulations . . . 68
3.4.4 Mechanical Properties Calculations . . . 68
3.4.5 Textural Properties Calculations . . . 69
3.5 Results and Discussion . . . 69
3.5.1 Topological Dependence for Dispersion Interactions-Driven Gas Storage . . . 69
3.5.2 Topological Dependence for Chemical Separations Driven by "Complexation" Interactions . . . 71
3.5.3 Topological Dependence for MOF Mechanical Properties . . . 73
3.6 Conclusions . . . 76

CHAPTER 4 USING MOLECULAR SIMULATION AND MACHINE LEARNING TO DESIGN MOFS FOR CO_2 CAPTURE . . . 78

4.1 Abstract . . . 78

4.2 Introduction . . . 79

4.3 Computational Methods . . . 81

4.3.1 MOF Construction . . . 81

4.3.2 CO_2 Binding to Functionalized Building Blocks . . . 81

4.3.3 Adsorption in MOFs . . . 83

4.3.4 MOF Properties . . . 84

4.3.5 Training of Machine Learning Models . . . 84

4.4 Results and Discussion . . . 85

4.4.1 Inherent Adsorbate-Adsorbent Interactions . . . 85

4.4.2 General Correlation of Electrostatic Interactions with CO_2 Sorption . . . 85

4.4.3 Variability of Adsorption Mechanisms . . . 87

4.4.4 On the Roles of Pore Chemistry and Topology . . . 89

4.4.5 Predicting the Impact of Functionalization on Altering CO_2 Capture Metrics of Parent MOFs using Decision Trees . . . 91

4.4.6 Correlations Between the Improvement of Different Carbon Capture Metrics . . . 93

4.4.7 Prediction of CO_2 Capture Metrics using Machine Learning . . . 93

4.4.8 Relative Importance of Chemistry- and Topology-Related Factors on Absolute CO_2 Capture Metrics . . . 95

4.4.9 Does Machine Learning Suggest what Chemical and Topological Properties Maximize CO_2 Capture Metrics? . . . 96

4.5 Conclusions . . . 97

CHAPTER 5 PREDICTING SINGLE COMPONENT ADSORPTION ISOTHERMS FOR MULTIPLE MOLECULES IN MOFS USING A SINGLE DEEP LEARNING MODEL . . . 100

5.1 Abstract . . . 100

5.2 Introduction . . . 101

5.3 Data Generation . . . 104

5.3.1 Database Construction . . . 104
5.3.2 Training, Validation, and Test Set Data Generation . . . 105
5.3.3 MOF Fingerprinting . . . 106
5.3.4 Adsorbate Fingerprinting . . . 107
5.3.5 Adsorption Simulations . . . 108
5.4 Neural Network Training . . . 109
5.5 Results and Discussion . . . 110
5.5.1 Model Predictive Ability for Simple Adsorbates . . . 110
5.5.2 Predictive Ability for Full Isotherms . . . 112
5.5.3 Transferability of Model for Predictions on MOFs with Inaccessible Pores . . . 114
5.5.4 Transferability of Model for Predictions on Experimental MOFs . . . 115
5.5.5 Material Ranking Accuracy Based on Performance Metrics . . . 116
5.5.6 Testing the Limits of the Current MLP Model (Extrapolation) . . . 117
5.6 Conclusions . . . 120
CHAPTER 6 METHODS FOR THE LARGE-SCALE COMPUTATION OF MOF FREE ENERGY AND CRITERIA FOR MOF THERMODYNAMIC ACCESSIBILITY . . . 121
6.1 Abstract . . . 121
6.2 Introduction . . . 122
6.3 Computational Methods . . . 124
6.3.1 MOF Database Construction . . . 124
6.3.2 Free Energy Calculations . . . 124
6.4 Results and Discussion . . . 127
6.4.1 Comparison of MOF Free Energy Calculation Methods . . . 127
6.4.2 Comparison Between Free Energies Calculated with Different Force Fields . . . 131
6.4.3 Correlations Between Free Energy and MOF Properties . . . 132
6.4.4 Consequences of MOF Isomorphism on the Application of the Overall Thermodynamic Stability Criterion . . . 136
6.4.5 Can Relative MOF Thermodynamic Stability be Assessed with Alternative Metrics? . . . 138

6.5 Conclusions . . . 141

CHAPTER 7 DEEP LEARNING COMBINED WITH IAST TO SCREEN THERMODYNAMICALLY FEASIBLE MOFS FOR ADSORPTION-BASED SEPARATION OF MULTIPLE BINARY MIXTURES . . . 142

7.1 Abstract . . . 142

7.2 Introduction . . . 143

7.3 Computational Methods . . . 147

7.3.1 Database Construction . . . 147

7.3.2 Adsorption Simulations . . . 148

7.3.3 MLP+IAST Model Construction . . . 148

7.3.4 Model-Based Clustering . . . 149

7.3.5 Free Energy Calculations . . . 150

7.4 Results and Discussion . . . 150

7.4.1 Predicting Single-Component Loadings at Low Fugacities . . . 150

7.4.2 Combining Neural Network Predictions with IAST . . . 153

7.4.3 Predicting Structure-Performance Relationships . . . 155

7.4.4 Identifying for What Separations there may be Privileged MOFs . . . 156

7.4.5 Using MOF Thermodynamic Feasibility Calculations in HTCS . . . 157

7.5 Conclusions . . . 160

CHAPTER 8 GENERAL CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK . 161

8.1 General Conclusions . . . 161

8.2 Recommendations for Future Work . . . 162

8.2.1 Further Development of Machine Learning Models for Predicting Adsorption . . . 162

8.2.2 Towards Universal MOF Synthesizability Criteria . . . 163

REFERENCES . . . 165

APPENDIX A SUPPORTING INFORMATION FOR CHAPTER 3 . . . 199

A.1 The ToBaCCo Net Scaling Algorithm . . . 199

A.2 Calculation of MOF Crystal Mechanical Properties . . . 209

APPENDIX B SUPPORTING INFORMATION FOR CHAPTER 4 211
B.1 Machine Learning Model Training Details 211
B.2 Supplementary Figures 216
APPENDIX C SUPPORTING INFORMATION FOR CHAPTER 5 242
C.1 Adsorbate Force Field Parameters and Fingerprinting 242
C.2 Assessing the Importance of Framework-Adsorbate Electrostatic Interactions 242
C.3 Model Training Details 244
C.4 Additional Prediction Data 245
APPENDIX D SUPPORTING INFORMATION FOR CHAPTER 6 252
D.1 Convergence of the Non-Equilibrium Switching Method 252
D.2 Determining Einstein Crystal Force Constants 253
D.3 Estimating the Error Associated with our FL Path Implementation 255
D.4 Fixed Center of Mass Corrections 255
D.5 Comparison of Calculation Times for QHA Free Energy and FL Path Free Energy 255
D.6 Additional Frenkel-Ladd Path/Quasi-Harmonic Approximation Comparison Data 256
D.7 Temperature Behavior of QHA and FL Free Energies 258
D.8 Mass Weighting Velocities to Calculate Phonon Density of States 258
D.9 Supplementary Tables and Figures 263
APPENDIX E SUPPORTING INFORMATION FOR CHAPTER 7 273
E.1 Model Training Details 273
E.2 Supplementary Figures 277
APPENDIX F A LIST OF THE AUTHOR'S CONTRIBUTIONS 291
F.1 First Authored Publications 291
F.2 Coauthored Publications 292
F.3 Software 293
APPENDIX G COPYRIGHT PERMISSIONS 294

LIST OF FIGURES

Figure 1.1 A decomposition of US industrial energy usage. This figure is redrawn with modifications from ref. . . . 2

Figure 1.2 The relative energy used by major industrial separation technologies. The relative energy use (unit-less) is shown in next to the technology (out of a maximum of 100). This figure is redrawn with modifications from ref. . . . 3

Figure 1.3 A schematic of the Skarstrom cycle, which is most easily read left to right (although since it is a cycle one can read right to left). Adapted from ref. . . . 6

Figure 1.4 A schematic of the IAST pipeline. . . . 8

Figure 1.5 (a) Shows an example of a zeolite (mordenite), and (b) shows and an example of a MOF (MOF-177) . . . 10

Figure 1.6 (a) Shows an example of a MOF with the organic linkers and metal nodes pointed out (b), (c), (d), and (e) are further examples of MOF structures. . . . 12

Figure 1.7 Examples of structure property relationships derived from the large-scale screening of MOFs. On each axis, each point is a MOF. The ΔQ for a 80/20 Kr/Xe separation is plotted on the y-axis versus MOF surface area and largest pore diameter (LPD) at five increasing pressures. Note that the surface area for MOFs with the highest ΔQ values increases with pressure, whereas the analogous peaks for LPD remains nearly stationary. The location and behavior of the peaks on each axis indicate ideal surface areas and LPDs for the separation of 80/20 Kr/Xe at the pressures shown. . . . 15

Figure 1.8 a. An example of a crystallographic net, and b. A MOF build according the the net shown in a. . . . 16

Figure 1.9 (a) What could be considered the "experimental" building blocks of a MOF, namely the linker molecule (left, a hexatopic carboxylate) and the metal (right, in this case Cu). (b) A computationally convenient decomposition of the building blocks in (a). Connection sites (where bonds are formed between building blocks) are shown in yellow, note that the linker molecule in (a) can be reconstructed from combinations of the two benzene-like rings shown in (b), and that the linker carboxylate groups are included in the computational metal node (right). . . . 17

Figure 1.10 Depictions of oft-used MOF textural properties, each property is labelled. Reproduced with permission from ref. . . . 18

Figure 2.1 The Rosenbrock function function sampled using 10,000 (a) random (x,y) points and (b) (x,y) points generated using the Metropolis importance sampling algorithm. The average probability (calculated from the Boltzmann distribution) of the points sampled using each method is shown. Both calculations were done with $T = 1.0$ . . . 40

Figure 2.2 Metropolis sampling of the Rosenbrock function for (from left to right) $T = 0.1$, 1, 10, and 100. . . . 40

Figure 2.3 An example isotherm of Ar in the silicalite framework MOR calculated using a series of GCMC simulations. Each point corresponds to a GCMC simulation at the corresponding pressure and isotherm temperature. Four snapshots of the simulation box showing the equilibrium loading of Ar at the corresponding pressure are shown. . . . 44

Figure 2.4 An example of a decision tree, the colors of each node qualitatively represent to homogeneity of the output values Y_i of each observation sent to that node. The leaf nodes should have the most homogeneous Y_i values. Perfect homogeneity is not often achieved in practice. . . . 51

Figure 2.5 A schematic of a simple MLP, with two hidden layers (white nodes). The magnification shows an individual node, in which the input weights are combined (in this case as a sum), fed into the activation function (a sigmoid is shown), then output with the node weight. . . . 53

Figure 3.1 Percentage of publications corresponding to 15 of the most well-known MOFs out of 42,455 publications as determined from a literature search using ISI Web of Knowledge. Fifteen out of the thousands of MOFs reportedly synthesized to date cover 22.4% (9,489 publications) of all MOF literature. The topology of each MOF is shown in parenthesis. . . . 58

Figure 3.2 The 19 topologies (three letter codes and barycentric unit cells shown) and seven tetratopic linkers (L1-L7), which together with the Cu paddlewheel form the 133 MOFs studied in this work. Atoms colored green indicate connection sites in the linker building blocks. These sites connect to the corresponding connection sites in metal node building blocks. . . . 61

Figure 3.3 A summary of the new ToBaCCo workflow. Algorithmic details on the "scale net" step are provided in A. . . . 64

Figure 3.4 (a) The template kkm along the a vector (left) and the b vector (right), examples of the different vertex types (V1, V2) and edge types (E1, E2) are labelled. (b) The building block assignments used to construct the first kkm MOF (left), and the MOF viewed down the a and b vectors (middle and right, respectively). (c) The building block assignments used to construct the second kkm MOF (left), and the MOF viewed down the a and b vectors (middle and right, respectively). . . . 66

Figure 3.5 (a) Methane deliverable capacity verses linker type each of the 133 MOFs is shown as a point with color/shape corresponding to the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of deliverable capacity and surface area change with linker size. (b) Methane deliverable capacity versus volumetric surface area with the corresponding line of best fit for the MOFs of each linker type, lines and points are colored by linker type. . . . 70

Figure 3.6 Ethene selectivity vs. MOF composition at different ethene:ethane compositions and pressures. Individual MOFs are shown as points, box and whisker plots are shown in grey for each MOF composition (isomorphic MOF set). . . . 71

Figure 3.7 Ethene selectivity vs $1/\rho_{\text{oms}}$ for each ethene:ethane composition and pressure considered. Note that some points are excluded by the plot range, this is to better show the selectivity trend with increasing $1/\rho_{\text{oms}}$. The black lines correspond to a selectivity of one. . . . 73

Figure 3.8 Simulation snapshots showing the location of adsorbed ethene. Left: Ethene selective ($S = 3.1$) sty-L1 MOF. Right: non-selective ($S = 0.9$) lim-L1 MOF. Ethene is shown in yellow (for clarity ethane is not shown). Both MOFs are shown from two perspectives: viewing down the a axis (top) and viewing down the b axis (bottom 74

Figure 3.9 Smallest eigenvalues of the MOF elastic tensor vs. linker type. Each of the 133 MOFs is shown as a point with color/shape corresponding the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of the eigenvalues change with linker size. 75

Figure 3.10 (a) Bulk modulus vs. MOF linker type. Each of the 133 MOFs is shown as a point with color/shape corresponding the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of the bulk moduli change with linker size. (b) Bulk modulus vs. the largest pore diameter (LPD) with points colored by linker type. 76

Figure 4.1 Topological blueprints and connecting molecular building blocks used in the construction of the over 400 MOF designs studied in this work. The $-(H)_4$ case corresponds to the non-functionalized linker with which parent MOFs constructed. 82

Figure 4.2 DFT-calculated optimal binding configuration between CO_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 86

Figure 4.3 Parity plots for GCMC-calculated CO_2 loading at 1 bar vs. DFT-calculated CO_2 binding energies. The plots relate the normalized, aver-age GCMC-calculated CO_2 loading (error bars indicate standard error of the mean) for structures featuring the same functionalized building block against the normalized DFT-calculated CO_2 binding energy with the corresponding building block. (a) Simulations with MBBB charges (SRCC = 0.610, $R^2 = 0.387$) and (b) simulations with no charges (SRCC = 0.022, $R^2 = 0.004$). 88

Figure 4.4 Probability density maps showing average CO_2 loading distribution in three analogous structures with different functionalization. The $-(H)_4$ and $-(CH_3)_2$ cases are outliers in Figure 4.3, while the $-(CN)_4$ is not. The maps indicate a different adsorption mechanism depending on the interaction strength between CO_2 and the functionalities. . 89

Figure 4.5 Effect of MOF functionalization (alteration of pore chemistry) of parent MOFs of different topologies on CO_2 loadings and selectivities. Results are presented for adsorption of a 15:85 CO_2:N_2 mixture at 1 bar (left) relevant to post-combustion CO_2 capture, and for a 20:80 CO_2/H_2 mixture at 20 bar (right) relevant to pre-combustion CO_2 capture. ΔS and ΔCO_2 are the changes in selectivity and CO_2 loading, respectively, due to functionalization. 90

Figure 4.6 Decision tree trained to predict whether parent MOF functionalization would improve the CO_2 selectivity over N_2 of the parent MOF during exposure to a 15:85 CO_2:N_2 mixture at 1 bar. The numbers at the terminal nodes are fractions, but can also be inter-preted as the probability that a MOF falling in that node is improved or not. Variables appearing in this tree are the highest bond dipole moment (HBDM) in functional group, sum of epsilons (SE) in functional group, most negative charge (MNC) in functional group, void fraction (VF) of parent MOF, and topology. 92

Figure 4.7 Correlations among improvements of different CO_2 capture metrics. 1 = CO_2 loading at 1 bar, 2 = CO_2 selectivity over H_2, 3 = CO_2 selectivity over N_2, 4 = CO_2 working capacity during adsorption of CO_2:H_2 mixture, 5 = CO_2 working capacity during adsorption of CO_2:N_2 mixture (a) matrix of SRCCs relating the rankings of topologies for each metric. A higher ranking means functionalization of the parent of such topology improved the metric in question more frequently. (b) Matrix of SRCCs relating the rankings of functional groups (FGs) for each metric. A higher ranking means functionalization of parent MOF with such FG improved the metric in question more frequently. Matrices are symmetric with above-diagonal positions color coded according to the color bar. 94

Figure 4.8 Parity plots comparing the prediction of CO_2 selectivity over N_2 by various trained machine learning algorithms versus the prediction from actual molecular (GCMC) simulations on the test set. The inset numbers indicate the R^2 and the Spearman ranking correlation coefficient (SRCC). MLR = multiple linear regression, DT = conditional inference decision trees, RF = random forests, GBM = gradient boosted machines, SVM = support vector machines, NN = artificial neural networks. 95

Figure 4.9 Relative importance of material descriptors obtained from GBM training. This plot is complementary to ICE plots shown in Figure B.24 to Figure B.28. S = selectivity, WC = working capacity, N = adsorption loading. FG = functional group, VF = void fraction, HDBM = highest dipole moment, MPC = most positive charge, MNC = most negative charge, LPD = largest pore diameter, PLD = limiting pore diame-ter, SE = sum of epsilons, GSA = gravimetric surface area. 96

Figure 4.10 Average GA-calculated descriptor values across all topologies giving maxima for different MOF CO_2 capture metrics. Metric indexes are: 1= CO_2 loading at 1 bar (red), 2 = CO_2 selectivity over H_2 (green), 3 = CO_2 selectivity over N_2 mixture (orange), 4 = CO_2 working capacity for the CO_2:H_2 mixture case (blue), and 5: CO_2 working capacity for the CO_2:N_2 mixture case (pink). Error bars correspond to the standard deviation. Descriptors labels are the same as in Figure 4.9. 98

Figure 5.1 The building blocks used for MOF database construction, dashed lines indicate connections to the rest of the framework; (a) inorganic (metal-containing) nodes, which include Cu (4-connected), Zn (6-connected), Cr (6-connected), and Zr (8- and 12-connected) oxoclusters, (b) organic nodes (the central part of multitopic organic linkers) (c) connecting building blocks (the arms of multitopic organic linkers or the body of ditopic linkers). 105

Figure 5.2 Descriptors constituting the MOF fingerprint. (a) Six textural properties: void fraction, gravimetric surface area, largest pore diameter (LPD: dark blue sphere), pore limiting diameter (PLD: light blue sphere), pore size standard deviation (PSSD), and density. (b) 17 chemical motifs (boxed), for which their respective number density in each MOF was calculated. The percentage of MOFs in the database that contain each motif is listed at the top. 107

Figure 5.3 Histograms of the textural properties used as descriptors in the MOF fingerprint. . . . 108

Figure 5.4 The configuration of our final model. The number of nodes in each hidden layer are shown above the corresponding layer. 110

Figure 5.5 Isotherms for alchemical adsorbates considered in (a) a representative small pore MOF (LPD=6.5 Å), (b) a representative intermediate pore MOF (LPD=13.7 Å), and (c) a representative large pore MOF (LPD=19.5 Å). All the MOFs shown are of the mcn topology. The large pore in each case is illustrated by a yellow sphere. 111

Figure 5.6 Parity plots comparing the predictions of the final MLP model for the six indicated adsorbates versus GCMC-calculated values in the 400 test set MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings). . 113

Figure 5.7 Isotherms for methane (top) and ethane (bottom), for the Q1 (0.00), Q2 (0.25), Q3 (0.50), Q4 (0.75), and Q5 (1.00) quantiles of isotherm mean percentage error (IMPE, with corresponding values shown). Points are GCMC simulated values (filled points correspond to fugacities included in training, empty points were not included in training), and orange lines show MLP predictions. 113

Figure 5.8 Parity plots comparing the predictions by the final MLP model for the six indicated adsorbates versus GCMC-calculated values in 1,528 CoRE MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings). . 116

Figure 5.9 Parity plots comparing the predictions of the final MLP model for the six indicated (extrapolated) adsorbates versus GCMC-calculated values in the 400 test set MOFs. These adsorbates possess properties outside the ranges covered by the alchemical adsorbates used during model training. Points color indicates the point density in the plot (the highest density is observed at low loadings). 119

Figure 6.1 The computational building blocks used to construct our 8,500-MOF database. (a) nodular inorganic building blocks (NIBBs). (b) 3-dimensional, nodular organic building blocks (NOIBs). (c) 2-dimensional nodular organic building blocks. (d) connecting building blocks (CBBs). NOIBs approximately correspond to so-called "secondary building units" (SBUs) in MOF synthesis literature. NOIBs correspond to the node of multitopic organic linkers. CBBs approximately correspond to ditopic linkers and/or to the arms of multitopic organic linkers. 125

Figure 6.2 Comparison between free energies calculated with the QHA (F_{QHA}) versus free energies calculated with via the FL path (F_{FL}). (a) Parity plot with point shape and color corresponding to the type of MOF inorganic node. (b) Parity plot illustrating the point density according to the color bar. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE) calculated for adjusted QHA free energies. 128

Figure 6.3 (a) Phonon density of states (PDOS) for two previously synthesized MOFs (names are shown above each plot). (b) Histograms of the contribution of frequencies in five ranges to the total QHA free energy (F_{QHA}) for each MOF (e.g. for most MOFs, frequencies in the 0-667 cm^{-1} range contribute less than 10% to F_{QHA}). (c) Parity plot comparing F_{QHA} and F_{FL} (FL free energies), with point colors indicating low-frequency, high-frequency, and uncategorized MOFs. 130

Figure 6.4 Parity plots comparing FL free energies (F_{FL}) calculated with different force fields. (a) UFF vs. UFF4MOF. (b) Dreiding vs. UFF4MOF. (c) Dreiding vs. UFF. Point shape and color correspond to the type of metal node in each MOF plotted. Different point features indicate the type of metal node in the MOF. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE). . . 131

Figure 6.5 FL free energies (F_{FL}) versus MOF properties. (a) F_{FL} vs. MOF density. b) F_{FL} vs. MOF gravimetric surface area (GSA). (c) F_{FL} vs. metal to linker atoms ratio. (d-f) breakdown of F_{FL} vs. metal to linker atoms ratio plot in (c) by metal node. (d) 6c Cr MOFs. (e) 4c Cu MOFs. (f) 6c Zn MOFs. (g) 8/12c Zr MOFs. Point features indicate the inorganic node in the MOF. Points outlined in black correspond to previously synthesized MOFs found in our 8,500-MOF database. Black lines in plots (d-f) are linear fits (equation shown) for the free energies of previously synthesized MOFs of each inorganic node type, and shaded regions correspond to the 95% confidence interval of each linear fit. 134

Figure 6.6 Relative FL free energy ($\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$) versus metal to linker atoms ratio of each MOF. $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ is obtained by subtracting the corresponding free energy from the linear fits in Figure 6.5d-g from F_{FL}. The horizontal line shows the bound that the 54 previously synthesized MOFs (purple points) in our 8,500-MOF database are all found below. . . . 135

Figure 6.7 A histogram of the free energy differences ΔF_{FL} per atom observed between isomorphs in our 8,500 MOF database. The solid, blue line indicates the mean difference, and the dashed, red line indicates the median difference. 137

Figure 6.8 The experimentally observed NOTT-100 (outlined by the dashed line) along with all their isomorphs found in our database. Isomorphs are labeled by their topology, with their free energy per atom relative to the NOTT-100 (fof topology) presented in parenthesis. 138

Figure 6.9 (a) Percentage of MOFs that can be discarded because they are synthetically inaccessible as a function of tolerance, the dashed line corresponds to the percent of MOFs discarded at the tolerance for free energy (2.5%) that retains all of the experimentally observed isomorphs in our database. (b) Percentage of synthetically accessible isomorphs that are kept as a function of tolerance, the dashed line corresponds to all 34 isomorphs. The tolerance is defined as being within a certain percent of the lowest free, potential or strain energy predicted within an isomorphic series. 139

Figure 7.1 The computational MOF building blocks used to construct our 51,520 MOF database. (a) Connecting (so-called "edge") building blocks. (b) Planar organic nodular building blocks. (c) Non-planar organic nodular building blocks. (d) Nodular inorganic building blocks. Connection sites "x", where bonds are formed with other building blocks, are highlighted in yellow. 147

Figure 7.2 Machine learning (ML) predictions versus GCMC predictions for single-component loadings of the indicated adsorbates in our 1,000 test MOFs. Loading values are compared for each MOF corresponding to 15 different fugacities between 0.02 and 100 bar. The MdAPE calculated across all fugacities is shown for each plot. 151

Figure 7.3 Comparison of our MLP-predicted isotherms for the six indicated adsorbates to GCMC simulated isotherms for the 0.00 (Q_1), 0.25 (Q_2), 0.50 (Q_3), 0.75 (Q_4), and 1.00 (Q_5) quantiles of the IMAPE. The open, blue points are GCMC simulated values and the orange lines are MLP predictions. 153

Figure 7.4 MLP+IAST predictions of the loadings of each component (top) and $\Delta Q_{\mathrm{B/A}}$ (bottom) verses GCMC predictions for the indicated binary mixtures. The loading plots are colored according to the component and the $\Delta Q_{\mathrm{B/A}}$ plots are colored by point density. These plots show values for all 10 pressures (1 to 10 bar) and mixture compositions (20/80, 40/60, 60/40, 80/20) considered. 154

Figure 7.5 The percent of GCMC-predicted top 8000 ΔQMOFs (y-axis) identified (50 top MOFs for each of the 160 conditions considered) when considering the top N MLP+IAST-predicted MOFs for each condition (x-axis). 155

Figure 7.6 Plots comparing the relationship of $\Delta Q_{\mathrm{Kr/Xe}}$ versus MOF void fraction as predicted by GCMC (orange) and our MLP+IAST model (blue). The mean VF of MOFs in the 95th percentile of $\Delta Q_{\mathrm{Kr/Xe}}$ is shown by the solid line for GCMC $\Delta Q_{\mathrm{Kr/Xe}}$, and by the dashed line for MLP+IAST $\Delta Q_{\mathrm{Kr/Xe}}$. 156

Figure 7.7 Clusters of conditions as clustered by their mean high-performing MOF properties. Each cluster is shown as a different shape and color. The clusters are projected onto directions found from a cluster-based dimensionality reduction algorithm that seeks to find linear combinations of the mean MOF properties that results in maximal cluster separation. Here the first three of four directions are shown in (a), (b), and (c). These three directions account for the majority of the structure in the clustered data. Uncertainty boundaries between the clusters are shown in grayscale, the darker the area the higher uncertainty. The specific conditions 80/20 N_2/CH_2 at 5 bar and 80/20 Kr/Xe at 1 bar are directly labeled. 158

Figure 7.8 $\Delta Q_{\mathrm{Kr/Xe}}$ for an 80/20 composition at 1 bar verses $\Delta Q_{\mathrm{N_2/CH_4}}$ for an 80/20 composition at 5 bar, both values were predicted using our MLP+IAST model. Cutoffs for to be in the 90th percentile for both ΔQs are shown and thermodynamically accessible MOFs in both regions are highlighted as yellow stars. . . 159

Figure 7.9 Six high-performing and thermodynamically feasible MOFs identified from our screening study. Topologies are shown in bold with the metal in parenthesis. The corresponding linker molecules (carboxylate form) are shown below each MOF. 160

Figure A.1 Example of faithful vs. unfaithful embedding of a graph. (a) The $K_{3,3}$ graph (known as the utility graph) unfaithfully embedded in $\mathbb{R}^2$. Notice that edges cross as they lie in the same plane, (b) The $K_{3,3}$ graph faithfully embedded in $\mathbb{R}^3$. Notice that edges do not cross. 199

Figure A.2 Building blocks and illustration of the need for anisotropic rescaling to properly map a building block onto a node in a topological blueprint. Connection sites are highlighted in yellow for all shown building blocks. (a) 4-connected Cu-Paddlewheel (a nodular building block), (b) 3-connected phenyl (a nodular building block), (c) a 2-connected diphenyl (a connecting building block), and (d) how building blocks connect with each other during MOF computational synthesis (the anticipated value of "x" is used in rescaling procedures). (e) The vertex geometry of 4-connected nodes in the barycentric representation of the sty topology. The vertex itself is represented in blue and the connected vertices are presented as gray points (the value of "y" is used in rescaling procedure). (f) how anisotropic rescaling of the net allows the node geometry to better fit the Cu-paddlewheel, which in turn allows the paddlewheel to be properly mapped onto the node. (note that the rescaling is such that "y" becomes equal to "x"). 200

Figure A.3 Example of how ToBaCCo anisotropically rescaled the same topological blueprint to build seven MOFs based on linkers of different size and aspect ratio. The original barycentric embedding of the fog topology (chosen as example) is shown to the right. Inorganic nodes and organic linkers used in MOF construction are shown to the left. The anisotropically rescaled net embeddings (all shown to scale) with the corresponding unit cell parameters are shown at the center. Green atoms represent connection sites in the building blocks. 201

Figure A.4 (a) The qtz net unit cell, which has three types of vertices, in its barycentric embedding. (b) The voltage graph with edges labelled e_i. The voltage graph vertices are labelled by the unit cell vertices which are in the quotient groups they represent. . 202

Figure A.5 Components of the stress tensor in each direction. 210

Figure A.6 Plots of ethene selectivity versus the percentage of ethene (out of the total adsorbed ethene) less than 4 Å from a Cu OMS for each combination of pressure and mixture composition considered. 210

Figure B.1 Schematics for the process we employed for training and validating the classification and regression models. Parameter grids were used for the NN, RF, GBM, DT, and SVM models, but not for the MLR models (which do not have model parameters). We used 5-times-repeated 10-fold cross validation. C.V. = cross-validation. 212

Figure B.2 Examples of initial configurations attempted when exploring the interaction of adsorbates with functional groups in MOF building blocks. 216

Figure B.3 DFT-calculated optimal binding configuration between N_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 217

Figure B.4 DFT-calculated optimal binding configuration between H_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 218

Figure B.5 DFT-calculated binding configuration between CO_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 219

Figure B.6 DFT-calculated binding configuration between N_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 220

Figure B.7 DFT-calculated binding configuration between H_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines. 221

Figure B.8 DFT-calculated binding energies for the interaction of the adsorbates (CO_2, N_2, and H_2) with the functional groups (squares), and with the phenyl ring (circles). 221

Figure B.9 Parity plots for GCMC-simulated CO_2 loadings and DFT-calculated binding energies.(a) CO_2 loading in a 20:80 CO_2:H_2 mixture at 20 bar, with (top) and without (bottom) MBBB charges and (b) CO_2 loading in a 15:85 CO_2:N_2 mixture at 1 bar, with (top) and without (bottom) MBBB charges. Error bars correspond to the standard error for the averaged CO_2 loadings. 222

Figure B.10 Parity plots for the GCMC-calculated CO_2 – framework interactions during simulated CO_2 at 1 bar and DFT-calculated CO_2 binding energies for A) dispersion interactions (R2 = 0.411, SRCC = 0.681) and B) electrostatic interactions (R2 = 0.623, SRCC = 0.710). 222

Figure B.11 CO_2 probability density maps showing the average CO_2 loading distribution in ten analogous structures for the remaining functionalizations not included in Figure 4.4. . . 223

Figure B.12 Circles denote fraction of MOFs improved by each functional group (left axis), black squares are the values of each MOF for the variable on the right axis, the white diamonds are the averages of these values for (a) CO_2 loading at 1 bar, (b) CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar, (c) CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar, (d) CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar, (e) CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. 224

Figure B.13 Effect of MOF functionalization (pore chemistry) on the CO_2 loading and CO_2 selectivity of "parent" MOFs of different topologies calculated with "zero" MOF atom charges. Results are presented for adsorption of a 15:85 CO_2:N_2 mixture at 1 bar (left, post-combustion carbon capture) and a 20:80 CO_2/H_2 mixture at 20 bar (right, pre-combustion carbon capture). As a reference, compare this with Figure 4.5 in the main text. 225

Figure B.14 Effect of MOF functionalization (pore chemistry) on the CO_2 capture metrics of "parent" MOFs of different topologies. (a) CO_2 loading at 1 bar (b) CO_2 working capacity for adsorption of CO_2:H_2 mixture. (c) CO_2 working capacity for adsorption of CO_2:N_2 mixture. 226

Figure B.15 Decision tree trained to predict whether parent MOF functionalization would improve the CO_2 loading of the parent MOF during exposure to pure CO_2 at 1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) in the functional group; topology, void fraction (VF), and gravimetric surface area (GSA) of the parent MOF. 227

Figure B.16 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 selectivity over H_2 of the parent MOF during exposure to a 20:80 CO_2:H_2 mixture at 20 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) and most positive charge (MPC) in the functional group; topology, largest pore diameter (LPD), and gravimetric surface area (GSA) of the parent MOF. 227

Figure B.17 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 working capacity the parent MOF for adsorption conditions corresponding to a 20:80 CO_2:H_2 mixture at 20 bar and desorption conditions corresponding to pure CO_2 at 1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) in the functional group and the topology of the parent MOF. 228

Figure B.18 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 working capacity of the parent MOF for adsorption conditions corresponding to a 15:85 CO_2:N_2 mixture at 1 bar and desorption conditions corresponding to a 90:10 CO_2:N_2 mixture at 0.1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) and most positive charge (MPC) in the functional group; the topology and void fraction (VF) of the parent MOF. 228

Figure B.19 Parity plots comparing the predictions on the test set of CO_2 loading at 1 bar by each machine learning algorithm to those obtained by GCMC simulations. GBM emerged as the most accurate algorithm. 229

Figure B.20 Parity plots comparing the predictions on the test set of CO_2 selectivity for adsorption of 20:80 CO_2:H_2 mixture at 20 bar by each machine learning algorithm to those obtained by GCMC simulation. GBM emerged as the most accurate algorithm. The three largest GBM prediction outliers are all fcu MOFs with functionalization -$(CN)_2$, -$(CN)_4$, and, -$(OH)_2$. 229

Figure B.21 Parity plots comparing the predictions on the test set for CO_2 working capacity in the CO_2:H_2 mixture case by each machine learning algorithm, showing the GBM model to be the most accurate. 230

Figure B.22 Parity plots for regression model predictions on the test set of CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Again, the GBM model is the most accurate on the test set. The top 5 largest GBM prediction outliers (for largest to smallest) are acs1 MOF with –$(CN)_2$ functionalization, bcs1 MOF with –$(H)_4$, bcs1 MOF with –$(NO_2)_2$, lvtb1 MOF with –$(CN)_4$, and lvtb3 MOF with –$(SH)_2$. Working capacity was calculated by setting the gas-phase compositions during desorption conditions to a 90:10 CO_2:N_2 mixture, which is an approximation based on the assumption of high CO_2/N_2 selectivity during adsorption stage. A negative working capacity is an indication that such assumption is incorrect for the specific material. 230

Figure B.23 Cross-validation metrics: mean absolute error (MAE), root mean squared error (RMSE), and R2 ("Rsquared"), from regression model training for (a) CO_2 loading at 1 bar, (b) CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar, (c) CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar, (d) CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar, (e) CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Error bars correspond to the 95% confidence interval. 231

Figure B.24 ICE plots for each variable from the GBM model of CO_2 loading at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations. 232

Figure B.25 ICE plots for each variable from the GBM model of CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations. 233

Figure B.26 ICE plots for each variable from the GBM model of CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations. 234

Figure B.27 ICE plots for each variable from the GBM model of CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations. 235

Figure B.28 ICE plots for each variable from the GBM model of CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations. 236

Figure B.29 Genetic algorithm maximization values (bars) along with GCMC data (black squares). The maximum values from GCMC simulations are white diamonds. (a) CO_2 loading at 1 bar, (b) CO_2 selectivity over H_2, (c) CO_2 selectivity over N_2, (d) CO_2 working capacity for the CO_2:H_2 mixture case, (e) CO_2 working capacity for the CO_2:N_2 mixture case. Error bars are the standard error of the mean across five maximizations. . 237

Figure B.30 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 loading at 1 bar (black lines). Green regions indicate the standard deviation of the optimum values. . 238

Figure B.31 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 selectivity over H_2 (black lines). Green regions indicate the standard deviation of the optimum values. 238

Figure B.32 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 selectivity over N_2 mixture (black lines). Green regions indicate the standard deviation of the optimum values. 239

Figure B.33 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 working capacity for the 20:80 CO_2:H_2 mixture case (black lines). Green regions indicate the standard deviation of the optimum values. 239

Figure B.34 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 working capacity for the CO_2:N_2 mixture case (black lines). Green regions indicate the standard deviation of the optimum values. 240

Figure B.35 Histograms of properties encountered in the MOFs constructed for this study, and which were used to train the machine learning algorithms. 241

Figure C.1 (a) Shows the force-field model used for single atom adsorbates, where only Lennard-Jones parameters are assigned to these adsorbates. (b) Shows the force-field model used for two-atom adsorbates, where each atom is assigned LJ parameters and a charge, the atoms are separated by a distance rbond, and a charge-balancing pseudo atom is placed at the bond center (note that q can be zero). 242

Figure C.2 The adsorbate fingerprints used as input for the MLP. a. Shows the fingerprint for single atom adsorbates and b. shows the fingerprint for two atom adsorbates, where $\sigma_{\text{effective}}$ is taken to be the average of the adsorbate end-to-end and top-to-bottom distances. . . . 242

Figure C.3 A comparison of N2 loading in the 400 test set MOFs calculated with (y-axis) and without (x-axis) framework charges. . . . 244

Figure C.4 (a) Shows the learning curve of our final model, demonstrating that our training set is representative of the validation set (and vice-versa) and that the model was not overfit by fitting over too many epochs at the learning rate used. (b) shows the error variability in 500 replications of our final model trained with different random seeds. . . . 246

Figure C.5 Plots analogous to Figure 5.7 in the main text for predictions on the remaining simple adsorbates, with adsorbates: (a) argon, (b) krypton, (c) xenon, and (d) nitrogen. Fugacities corresponding to filled, blue points were included in training, empty points correspond to pressures not included in training, orange lines correspond to model predictions. . . . 247

Figure C.6 Model predictions for methane loading vs GCMC simulated loading with blocking spheres for the 400 MOF test set.(a) Shows predictions using MOF textural properties corrected for the presence of blocking spheres,(b) shows predictions using uncorrected MOF textural properties. Note that this plot only shows loadings in the range that changed due to the presence of blocking spheres, i.e. over a lower loading range as the blocking spheres only affected the loading in certain small-pore MOFs. . . . 247

Figure C.7 Showing the distribution of percent error for loadings predictions in the test set MOFs (indicated, orange) and the CoRE MOFs (indicated, blue). These distributions include predictions made on all adsorbates at all fugacities considered. . . . 248

Figure C.8 Plots analogous to Figure 5.7 in the main text for CoRE MOF predictions, with adsorbates: (a) argon, (b) methane, (c) krypton, (d) xenon, (e) nitrogen, and (f) ethane. Filled, blue points correspond to GCMC simulated values, orange lines correspond to model predictions. Only the first four and five points (respectively) of the methane Q1 and Q2 isotherms are shown as higher-pressure points surpass the y-axis scale. . . . 249

Figure C.9 The percent of MOFs correctly placed by the MLP (y-axis) out the top N MOFs where N is the value on the x-axis for all six adsorbates considered for (a) the 400 test set MOFs, and for (b) the CoRE MOF subset. . . . 250

Figure C.10 Parity plots comparing the predictions of the final MLP model for (a) propane, (b) n-butane, and (c) isobutane versus GCMC-calculated values. Points are colored by how close they are to saturation loading (1.00 being at saturation). . . . 251

Figure D.1 FL path free energy convergence with (a) switching time and (b) minimum number of atoms in the simulation supercell. . . . 253

Figure D.2 FL free energies calculated without any restriction on the force constants versus those calculated with a force constant lower bound of 10 kcal/mol Å^2. R and S are the Pearson and Spearman ranking correlation coefficients, respectively, and MAE is the mean absolute error. . . . 254

Figure D.3 Comparison of total calculation times in core minutes (defined as the product of calculation time and the number of cores used for the calculation) for QHA (blue) and FL (orange) free energies in our MOF database. 257

Figure D.4 Histogram of the absolute deviation between QHA and FL free energies (ΔF) after removing the apparent QHA systematic error. The dashed and solid vertical lines indicate the median and absolute deviations, respectively. 258

Figure D.5 Phonon density of states (PDOS) for select previously synthesized MOFs. Each PDOS was individually normalized to vary between 0 and 1 before plotting. 259

Figure D.6 Absolute Error for QHA free energy relative to FL free energy in all MOFs in our database. High-frequency and low-frequency MOFs are highlighted with different color. Note that larger errors but more "systematic" errors occur for high-frequency MOFs, while smaller but less "systematic" errors occur for low-frequency MOFs. 259

Figure D.7 (a) QHA free energies versus FL free energies for each MOF in our database. The 100 points farthest from the parity line are colored blue, while the 100 closest to the parity line are colored red. (b) Superimposed PDOS for each of the points shown in (a) with corresponding color. (c) QHA free energies versus FL free energies when using modes with frequency up to 2669 cm^{-1} in the QHA calculations. 260

Figure D.8 QHA and FL free energies over a range of temperatures between 20 and 600 K for the indicated MOFs. 260

Figure D.9 Analogous to Figure 6.3 in the main text, but for QHA and FL free energies calculated at 100 K. 261

Figure D.10 Analogous to Figure 6.3 in in the main text but for PDOS and QHA free energies calculated using non-mass-weighted velocities. 262

Figure D.11 Analogous to Figure D.6 but for QHA free energies calculated with non-mass-weighted velocities. (a) Shows three groups of MOFs each colored according to their distance from the parity line, blue corresponds to the 100 MOFs farthest from the parity line, green to the 100 MOFs nearest the average distance from the parity line, and red the 100 MOFs closest to the parity line. (b) shows the superimposed PDOS for all 100 MOFs in each category, colored accordingly, and (c) is analogous to (a) but with the QHA free energy calculated only using vibrations with frequencies less than 2669 cm^{-1}. 262

Figure D.12 (a-c) Parity plots comparing FL free energies (F_{FL}) calculated with different force fields. (a) UFF vs. UFF4MOF. (b) Dreiding vs. UFF4MOF. (c) Dreiding vs. UFF. Point shape and color correspond to the type of metal node in each MOF plotted. (d-f) Parity plots comparing FL free energies (F_{FL}) calculated with different force fields, illustrating the point density. (d) UFF vs. UFF4MOF. (e) Dreiding vs. UFF4MOF. (f) Dreiding vs. UFF. This figure is made using the same data as Figure 6.4 in the main text. 263

Figure D.13 FL free energies (F_{FL}) versus MOF properties. (a) F_{FL} vs. MOF void fraction. (b) F_{FL} vs. MOF largest pore diameter. (c) F_{FL} vs. inverse density. (d-f) breakdown of F_{FL} vs. density plot (Figure 6.5a in the main text) by metal node. d) 6c Cr MOFs. e) 4c Cu MOFs. (f) 6c Zn MOFs. (g) 8/12c Zr MOFs. Point features indicate the inorganic node in the MOF. Points outlined in black correspond to previously synthesized MOFs found in our 8,500-MOF database. Black lines in plots (d-f) are linear fits for the free energies of previously synthesized MOFs of each inorganic node type, and shaded regions correspond to the 95% confidence interval of each linear fit. . 264

Figure D.14 Analogous to Figure 6.5 but for free energies calculated using the UFF. 264

Figure D.15 Analogous to Figure 6.5 but for free energies calculated using the Dreiding forcefield. . 265

Figure D.16 Analogous to Figure 6.6 but for for UFF (a) free energies and Dreiding (b) free energies. 265

Figure D.17 Example MOFs selected from outside the "synthesizability" region of the plot of relative free energy ($\Delta_{LM}F_{FL}$) versus MOF metal to linker atoms ratio. $\Delta_{LM}F_{FL}$ is obtained by subtracting the free energy from the linear models shown in Figure 6.5d-g from the corresponding absolute from F_{FL}. The orange line delimits the region where the $\Delta_{LM}F_{FL}$ of the 54 previously synthesized MOFs (purple points) in our 8,500-MOF database are found below. 266

Figure D.18 Parity plots of FL path free energy calculated at 100, 300, 350, and 400 K, point densities for each temperature are shown on the leading diagonal. Given the close correlation of free energies at 300, 350, and 400 K Figure 6.5 and Figure 6.6 in the main text will not appreciably change across these temperatures. 267

Figure D.19 The spread of $\Delta_{LM}F_{FL}$ values within the 32 studied isomorphic MOF series. The yellow points correspond to the experimentally observed isomorph for each series. The horizontal, red line corresponds to the upper boundary in $\Delta_{LM}F_{FL}$ above which MOFs have low synthetic likelihood. Accordingly, note that many MOFs that would be considered stable may not be experimentally observed, simply because a more stable isomorph is more stable. 269

Figure D.20 The linkers observed in the experimentally observed isomorphs for the 32 isomorphic MOF series discussed in Chapter 6. 270

Figure D.21 Parity plots comparing potential energy at 300 K (U_{300K}) and 0 K (U_{0K} , strain energy) with free energy F_{FL} (at 300 K) (a) U_{300K} vs F_{FL}. (b) U_{0K} vs F_{FL}. Point features indicate the type of metal node. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE). 271

Figure D.22 The region of synthetic likelihood outlined in Chapter 6 but using (a) strain energy and (b) potential energy at 300 K 272

Figure E.1 (a) Learning curve for our MLP trained to predict single component isotherms. (b) Parity plot of our model predictions versus GCMC values for our validation set (all points shown). (c) is a zoomed-in version of (b), only showing predicted loadings up to 7 mol/kg. 274

Figure E.2 Comparison of our MLP predictions of the validation and test sets on a log-log scale. . 276

Figure E.3 Error metrics for all the adsorbate/fugacity combinations for our single component isotherm testing data. Each error metric was calculated for all 1,000 test set MOFs. (a) Median Absolute Errors (MdAEs), (b) Median Absolute Percentage Errors (MdAPEs), (c) Pearson correlation coefficients, and (d) Spearman rank correlation coefficients. 277

Figure E.4 Comparison of our MLP-predicted isotherms for the six indicated adsorbates to GCMC simulated isotherms on a log-log scale for the 0.00 (Q_1), 0.25 (Q_2), 0.50 (Q_3), 0.75 (Q_4), and 1.00 (Q_5) quantiles of the IMAPE. The open, blue points are GCMC simulated values and the orange lines are MLP predictions. 278

Figure E.5 Error metrics for Ar/Kr binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{\mathrm{B/A}}$. 278

Figure E.6 Error metrics for Kr/Xe binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{\mathrm{B/A}}$. 279

Figure E.7 Error metrics for CH_4/C_2H_6 binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{\mathrm{B/A}}$. 280

Figure E.8 Error metrics for N_2/CH_4 binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{\mathrm{B/A}}$. 281

Figure E.9 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for Ar/Kr. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line. 282

Figure E.10 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for CH2/C_2H_6. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line. 283

Figure E.11 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for N_2/CH_4. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line. 284

Figure E.12 The means (points) and standard deviations (bars) of the VFs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations. . 285

Figure E.13 The means (points) and standard deviations (bars) of the LPDs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations. . 286

Figure E.14 The means (points) and standard deviations (bars) of the PLDs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations. . 287

Figure E.15 The means (points) and standard deviations (bars) of the GSAs MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations. . 288

Figure E.16 The means (points) and standard deviations (bars) of the densities of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations. 289

Figure E.17 Pairwise parentage overlaps of high-performing MOFs (MOFs in the 95th percentile of ΔQ) for the indicated conditions. Conditions are shown as space-separated strings of the mixture, fraction of the weakly adsorbed component, and pressure in kPa. 290

LIST OF TABLES

Table 5.1 Model performance metrics of our final model for loading predictions made on the test set MOFs. . . . 112

Table 5.2 Comparison of MOF rankings between MLP predictions and GCMC simulations as indicated by the Spearman rank correlation coefficients (S) and the number of MOFs in the top-100 from GCMC simulation that were also found in the top-100 from MLP predictions. . . . 118

Table 6.1 Correlation metrics between different energies calculated for all MOFs in our database. R and S are the Pearson and Spearman Ranking correlation coefficients, respectively, and (MAE) is the mean absolute error. . . . 140

Table C.1 Parameters for single-site alchemical adsorbates used for *training*; all possible combinations were considered. . . . 243

Table C.2 Parameters for single-site alchemical adsorbates used for *validation*; all possible combinations were considered. . . . 243

Table C.3 Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) used for *training*; all possible combinations were considered. . . . 243

Table C.4 Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) used for *validation*; all possible combinations were considered. . . . 243

Table C.5 Lennard-Jones parameters for single site real adsorbates. . . . 243

Table C.6 Version information for Python packages used for MLP training. . . . 244

Table C.7 The optimal hyperparameter values for our final MLP found using a grid search. Topology is denoted by $N_1 - \cdots - N_M$, where N_i is the number of nodes in each hidden layer. . . . 245

Table C.8 Model performance metrics of our final model for loading predictions made on the CoRE MOFs. . . . 248

Table C.9 Model performance metrics of our final model for loading predictions made on adsorbates with properties outside of range spanned by the training set (extrapolative predictions). . . . 250

Table D.1 Strain energy for each metal node present in our database. These energies were calculated by optimizing the isolated metal node in LAMMPS using UFF4MOF. The carboxylate groups on each node where capped with hydrogen. . . . 263

Table D.2 The number of MOFs with $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ higher than the highest $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ among the previously synthesized MOFs in our database for both an absolute cutoff and a cutoff with a 20% tolerance for UFF4MOF, UFF, and Dreiding free energies. . . . 265

Table D.3 List of the experimentally observed MOF isomorphs for the 32 isomorphic MOF series discussed in Chapter 6. The experimentally observed MOF is described by the "Linker" (as per Figure D.20), "Metal" and "Observed topology" columns. The remaining columns indicate the topology that it is predicted on the basis of having lowest free energy (F_{FL}), potential energy ($U_{300\mathrm{K}}$), and strain energy ($U_{0\mathrm{K}}$). Cases where the predicted topology differ from the experimentally observed one include a percent value in parenthesis, which indicates the deviation between the F_{FL} (or $U_{300\mathrm{K}}$ or $U_{0\mathrm{K}}$) of the predicted and experimentally observed topologies 268

Table D.4 Experimental isomorph identification metrics for free and potential energy at different temperatures (strain energy metrics do not change with temperature). 269

Table E.1 The fugacities used for GCMC training, validation, and test data for our MLP trained to predict single component isotherms. 274

Table E.2 The MOF fingerprint used to train our MLP for predicting single-component isotherms. The first 20 descriptors shown are MOF chemical motif number densities. Each motif is described using the atom type nomenclature used by UFF. To obtain the number density of the motif, the count of each atom type was normalized by the MOF volume in Å^3. The last six MOF descriptors listed are MOF textural properties. 275

Table E.3 The hyperparameters used during the training of our MLP for predicting single component isotherms. Topology is denoted by $N_{\ldots-N_M}$, where Ni is the number of nodes in each hidden layer. 275

Table E.4 Versions for all the Python packages used for our MLP training and IAST calculations. . 276

ACKNOWLEDGMENTS

I hereby thank my advisor, Prof. Diego Gómez-Gualdrón, for his guidance, encouragement, knowledge, (all the other words), etc. Without Diego, this thesis would either not exist or be a faint echo of its current self in some alternate universe. I also thank Ben Schweitzer for his capacious knowledge of most things and for his support during the first years of my studies. Special thanks to the Colorado School of Mines High-Performance Computing department for allowing me to wreak havoc on their systems for years. Finally, I express my gratitude to my parents, Beckie and Sam Anderson, for their unreserved support, for putting up with my nonsense, and for paying my cell phone bill.

Additional thanks are granted, in some particular but unspecified order, to Caroline Lima Salles, for her friendship; Jimmy Crawford, for his ridiculous notions/behavior and a series of valuable discussions; Arthur Korous for his consistent reminders that doctoral degrees are optional and rock climbing is fun; Prof. Annette Bunge for her encyclopedic knowledge of The Black Hole; and Kevin Prince for being just a real good guy. To everybody else I could thank (and there are a lot of them), my implicit gratitude will have to suffice.

CHAPTER 1

INTRODUCTION

1.1 Chemical Separations

The overwhelming majority of world energy originates from the burning of fossil fuels (natural gas, oil, and coal) [1]. Fossil fuels are finite and produce harmful emissions when burned, making them doubly unsuitable as a long-term energy source, especially when considering that, at the time of writing, the U.S. and world energy consumption has been steadily increasing for decades [1, 2]. There are fundamentally two ways to meet the challenge of generating sufficient energy to power the world while reducing harmful emissions. The first (*i*) is to replace the burning of fossil fuels with renewable methods of energy generation (i.e. solar, wind, and geothermal technologies). The second (*ii*) is to reduce the amount of energy consumed in the first place, which reduces the emissions from burning fossil fuels and simultaneously makes the switch from non-renewable to renewable energy sources easier. Here, I focus on (*ii*). Specifically, replacing high-energy intensity separation technologies with technologies that require (ideally markedly) less energy will lead to substantial reductions in U.S. (and world) energy consumption. Tangentially, separation technologies can be used to separate and sequester harmful emissions from industrial processes. For example, CO_2 can be removed from combustion products to limit the greenhouse gas emissions from the burning of fossil fuels.

As evidenced in Figure 1.1, approximately half (45-55%) of the energy consumed by U.S. industry is used to separate mixtures, half again of this energy is used by distillation processes [3]. Currently, 90-95% of the separations in the U.S. petroleum and chemical industries are accomplished using distillation [4]. Other common, energy-intensive separation technologies (e.g. drying and evaporation) also require a relatively large input of heat. However, there are promising low energy-intensity alternatives to these energy-intensive methods. Given this, and considering the ubiquity of distillation in key industries, it is clear that replacing distillation with less energy-intensive alternatives will reduce U.S. industrial energy consumption by a substantial amount. Of course, the exact amount depends on the relative energy use of the alternative processes.

The eight main industrial separation technologies are presently: distillation, evaporation, drying, extraction, adsorption, membrane, crystallization, and physical property-based processes (PPBP, e.g. flotation) [4]. Distillation, adsorption, and membrane separations are discussed in more detail in the subsequent sections. As stated previously, distillation, evaporation, and drying are considered high-energy intensity methods, whereas the remaining methods (extraction, adsorption, membrane, crystallization, and

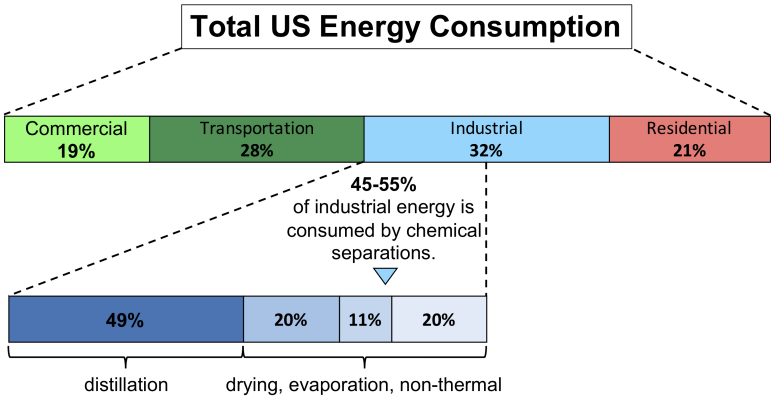

Figure 1.1 A decomposition of US industrial energy usage. This figure is redrawn with modifications from ref. [3].

PPBP) are considered mid- to low-energy intensity alternatives. Figure 1.2 shows the relative energy usage of each of these technologies. Notably, extraction, adsorption, and membrane separation require 31, 25, and 15%, respectively, of the energy used by distillation. Therefore, according to the data in Figure 1.1 and Figure 1.2, if all distillations are replaced by any of these low-energy intensity alternatives, the total U.S. energy consumption would be reduced by approximately 5-6%. Clearly, replacing *every* distillation with a low-energy alternative is unrealistic, however, replacing even a fraction would still have a marked effect on U.S. energy usage. Furthermore, hybrid systems that augment existing high-energy methods with low-energy methods can be implemented in cases where total replacement is difficult or impossible.

The main challenge associated with implementing the extraction, adsorption, and membrane separation methods is finding suitable materials for use as separation agents. Separation agents interact with mixture components on the micro or molecular level to retain one component (ideally) abundantly and selectively, these include solvents, adsorbents (e.g. porous crystals), and membranes (e.g. molecular sieves). In principle, each separation requires a material that is specifically designed for the separation application, although there are cases where materials can be used for multiple separations. Broadly speaking, there are two "sets" of considerations that should be taken into account when designing a material for use as a separation agent. First (i), the material must achieve adequate selectivity, or purity of the isolated component(s), and throughput, or the rate of pure component production. Designing a material with these capabilities hinges on the ability to precisely identify and/or control the material/mixture interactions, i.e. to precisely control the chemistry of the material. Second (ii), candidate materials must be reasonably durable and resist fouling at the relevant conditions, and their production should be economically viable. The work herein is on the forefront of consideration (i), including only limited (and implicit) consideration of (ii). This is because, while operating conditions and economies of scale can be changed, the physical

properties of a mixture, and thus the properties of a material that can selectively separate large enough quantities of that mixture, are unalterable.

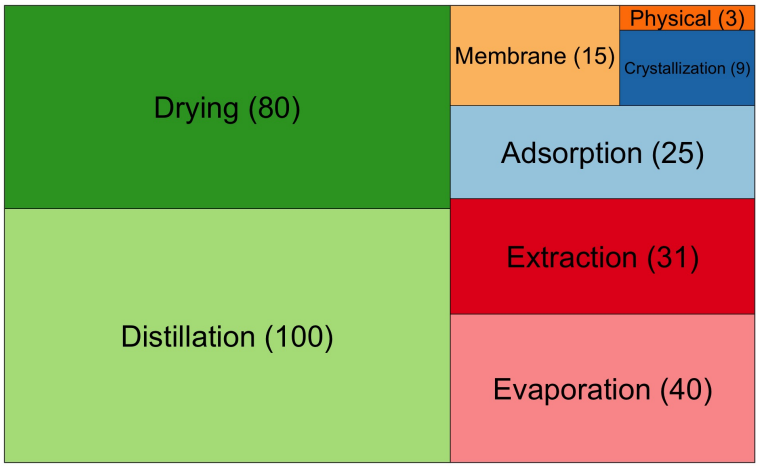

Figure 1.2 The relative energy used by major industrial separation technologies. The relative energy use (unit-less) is shown in next to the technology (out of a maximum of 100). This figure is redrawn with modifications from ref. [4].

In this thesis, I focus on the design and discovery of materials (specifically metal-organic frameworks, or MOFs) that interact with mixture components on the molecular scale. Thus, when considering specific mixtures, my focus is on mixtures of small gasses that are relevant to industrial processes. This is due to a number of reasons. First, MOFs are particularly well-suited for adsorbing large quantities of gas with high specificity (the reasons for this are discussed in subsequent sections). Second, many gas mixtures are difficult to separate as the components tend to have similar physical properties and cryogenic boiling points (most industrial gasses are currently separated using cryogenic distillation) [5]. Third, the separation of small gasses is ubiquitous in industry [5], to use the pure component(s) in some process and to limit harmful emissions.

1.1.1 Distillation

In distillation, liquid mixture components are separated based on (differences in) their boiling points. Essentially, the more volatile component(s) of a mixture are vaporized using heat, these volatile components then occupy a larger "part" of the gas phase, and, therefore, the more volatile component(s) can be concentrated by condensing the vapor phase. Here, I breifly describe the principles of distillation, assuming that Raoult's law applies (i.e the solution being separated is at least close to ideal). Raoult's law states that

$$p_i = p_i^* x \tag{1.1}$$

where p_i is the vapor phase partial pressure of the ith component of a mixture, p_i^* is the vapor pressure of the pure component i, and x is the mole fraction of component i in the liquid phase. When the boiling points of a mixture are vastly different, the vapor phase can simply be condensed, as the vapor phase will be mostly the more volatile component. However, for mixtures in which the components have similar boiling points (i.e. similar vapor pressures), both components will have non-negligible partial pressures in the vapor phase, in these cases fractional distillation is used. In fractional distillation uses iterations of evaporation and condensation to progressively separate a mixture in a column (just a cylinder of some inert solid). There are numerous fractional distillation setups, depending on the properties of the mixture being separated, but fundamentally: the liquid mixture heated is at the bottom of the column and the rising vapor is then condensed on a series of surfaces (usually a plate with a high surface area). At each condensation point, the mole fraction of the volatile component(s) increases in the liquid phase, given that the vapor being condensed has a higher partial pressure of the volatile component(s). At some point (height of the column), the volatile component(s) will reach acceptable purity and can be collected. Mixtures are input to the column at some point along the height of the column, so the liquid phase falls, gaining concentration of the less volatile component(s), and the vapor phase rises, gaining concentration of the more volatile component(s).

The boiling points of small gasses are far below ambient temperatures, meaning that these mixtures must be drastically cooled for distillation (to produce a liquid phase). This process of cooling the gas to allow for a liquid phase during distillation is called cryogenic distillation, which is used for several key separations in industry, for example, N_2/O_2, N_2/CH_4, $CH_4/C_2/H_6$, Kr/Xe, and alkene/alkane separations [4–6]. In a cryogenic distillation process, the gas mixture is cooled and compressed to form a large volume of the liquid phase, which is then separated using the same principles as any distillation procedure. These cooling/compression steps require large inputs of energy. To illustrate this, it is useful to examine the case of ethene production. Ethene is a chemical used in a variety of industries; it is mostly produced by steam cracking (i.e. the conversion of ethane to ethene), an extremely high-energy intensity process [7]. Part of ethene production is its separation from other cracking products (e.g. methane, ethane, and heavier hydrocarbons) using cryogenic distillation. Of the total energy used for ethene production, 22% is consumed by the separation process, and essentially all of this 22% is consumed for the cooling/compression of the gas mixture [4]. There are further losses of energy (not included in this 22%) required to convert energy to mechanical work for use in the refrigeration process [4]. Replacing or augmenting the cryogenic distillation steps of ethene production with membrane- or adsorption-based methods could potentially save 56 trillion Btu per year, which equates to 0.5-0.6% of the total U.S. energy consumption in 2019 [2, 4].

1.1.2 Equilibrium Adsorption Separations

Adsorption separations harness the interaction of an adsorbent material with a gas (or liquid) mixture to isolate the mixture component(s). Essentially, one or more mixture components are concentrated on the surface of the adsorbent material, often a micro- or nano-porous solid, the concentrated component(s) are then recovered from the surface of the adsorbent. For this to work, the adsorbent must interact more strongly with one or more mixture components than with the others, ensuring that the adsorbed phase has a higher concentration of certain components than the gas (or liquid) phase. This is analogous to Raoult's law (Equation 1.1), but the concentration of component i in the adsorbed phase is proportional to the strength of its interaction with the adsorbent, rather than its vapor pressure. One measure of the strength of an adsorbate/adsorbent interaction is the enthalpy of adsorption (ΔH_{ads}), or the heat released up the adsorbate (adsorbed species) binding. Note that the adsorption process is always exothermic. This is because for (any reasonable amount of) adsorption to occur, it must be spontaneous, i.e. the Gibbs free energy

$$\Delta G_{\text{ads}} = \Delta H_{\text{ads}} - T\Delta S_{\text{ads}} \tag{1.2}$$

must be negative. The entropy difference ΔS_{ads} between the gas phase and the adsorbed phase is negative since the translation and rotational degrees of freedom of the adsorbate are limited in the adsorbed phase. Consequently, $\Delta G_{\text{ads}} < 0$ implies $\Delta H_{\text{ads}} < 0$, or an exothermic process. Adsorbates may be weakly bound to the adsorbent via intermolecular interactions (dispersion and charge-charge interactions), which is called physisorption, or the adsorbate may be covalently bound to some moiety in the adsorbent, which is called chemisorption. Physisorption is characterized by a relatively small ΔH_{ads}, around 10-15 kcal/mol; chemisorption is characterized by a relatively large, ΔH_{ads}, or > 15 kcal/mol [8, 9]. Notably, very large ΔH_{ads} values are generally undesirable as ΔH_{ads} is related to the energy required to regenerate the adsorbent [10]. In other words, if ΔH_{ads} is too large the isolated mixture component(s) cannot be easily retrieved. Therefore, physisorption is the primary method considered when designing or selecting materials for adsorption separation.

The amount of an adsorbate present in an adsorbent at any operating condition is dictated by the equilibrium of the adsorbed phase with the gas phase, the equilibrium condition being that the temperature and chemical potential of the gas inside the adsorbent are equal to those of the gas outside the adsorbent. Assuming the adsorbent does not undergo any phase changes, the amount of adsorbed molecules will increase with pressure and decrease with temperature. So, once equilibrium is reached between the gas and adsorbed phase, the amount of adsorbed molecules will no longer change. This constraint of equilibrium implies that adsorption-based separations are inherently cyclical, with each cycle consisting of equilibration

of the mixture/adsorbent system followed by removal of the adsorbed phase from the adsorbent surface (in the absence of the gas mixture), a process called regeneration. Regeneration is usually accomplished by decreasing the pressure (pressure swing adsorption, or PSA) or increasing the temperature (temperature swing adsorption, or TSA) [9]. A basic, and still commonly used, PSA procedure is the Skarstrom cycle [9, 11]. The Skarstrom cycle uses a two-bed setup and two main steps: (*i*) adsorption at high pressure and (*ii*) purging (desorption) at low pressure. In more detail: the first bed is pressurized with the gas feed while the second bed is depressurized (this is called "blow down"), once the first bed is pressurized, the high pressure adsorption has begun. While the first bed is filling, a small amount of the purified tail gas from the first bed is fed to the second at atmospheric pressure to purge it of the adsorbed component(s). Finally, the second bed is pressurized and the first bed goes through blow down, repeating the cycle. A schematic of the Skarstrom cycle is shown in Figure 1.3. Note that the Skarstrom is designed to purify the weakly adsorbed component of a mixture, to also purify the strongly adsorbed component, it must be extracted during the desorption phase, often under vacuum [11]. Consequently, purifying the strongly adsorbed component usually takes significantly more energy than the weakly adsorbed component.

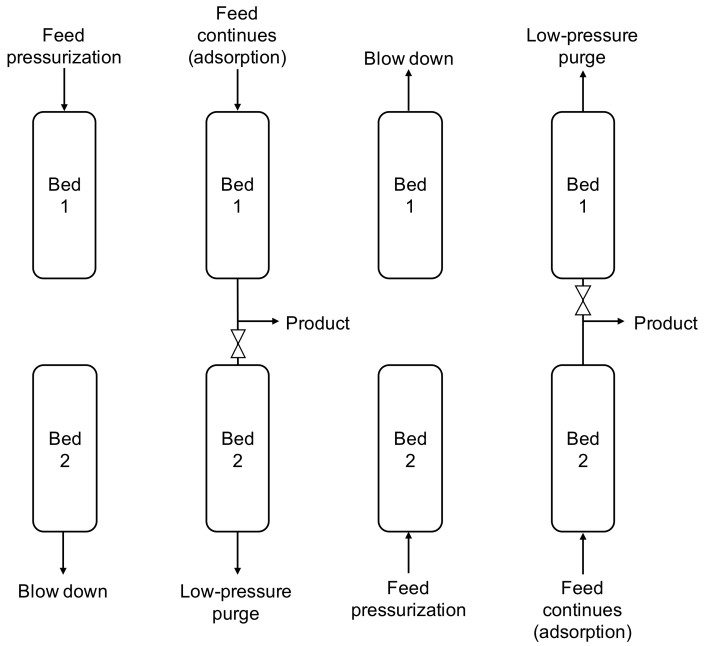

Figure 1.3 A schematic of the Skarstrom cycle, which is most easily read left to right (although since it is a cycle one can read right to left). Adapted from ref.[9]

There are a number of metrics used to assess a material for use in a PSA process. Here I discuss three widely used metrics: adsorption selectivity, adsorption capacity, and productivity. Selectivity is a measure of how pure the adsorbed phase is, for example, a highly selective material will almost exclusively adsorb

only one mixture component. For a binary (A/B) mixture, the selectivity is defined as

$$S_{\mathrm{A/B}} = \frac{q_{\mathrm{A}}/q_{\mathrm{B}}}{x_{\mathrm{A}}/x_{\mathrm{B}}} \tag{1.3}$$

where q_{A} and q_{B} are the loadings of the mixture components in the adsorbent and x_{A} and x_{B} are the mole fractions of the components in the bulk gas phase. Capacity is the amount of the strongly adsorbed species that can fit in the adsorbent at equilibrium conditions, usually it is reported as the amount of adsorbate (moles, mass, or volume at standard temperature and pressure) per amount of adsorbent (mass or volume). Productivity is the amount of product produced per amount of adsorbent per cycle [9]. Selectivity and capacity are measurable via molecular simulation or through experiment (if some assumptions are made). Productivity can be measured experimentally or simulated (with some difficulty); it depends on both capacity and selectivity. Productivity is the primary metric used to assess adsorbents for industrial usage. To understand why selectivity and capacity are difficult to measure experimentally, it is first necessary to discuss adsorption isotherms.

An adsorption isotherm is a measure of the amount of an adsorbate present in an adsorbent as a function of pressure. For multicomponent mixtures, each component has an associated isotherm, which depends on the composition of the mixture. Single component isotherms are routine measurements, but binary or multicomponent isotherms are difficult and time consuming to measure directly. Therefore, the selectivity and capacity of an adsorbent for a gas mixture, which rely on the loadings reported from multicomponent isotherms, are also difficult and time consuming to measure directly [12]. For this reason, analytical models of mixture adsorption, such as Ideal Adsorbed Solution Theory (IAST) [13, 14], are often used to estimate multicomponent loadings. IAST takes as input single component isotherms and the mixture composition and outputs the multicomponent loadings at a specified pressure or fugacity [13–15]; a schematic of this pipeline is shown in Figure 1.4. The multicomponent loadings derived from IAST can then, in turn, be used to calculate adsorbent selectivity and capacity. An alternative to this approach is to use molecular simulation to calculate multicomponent loadings directly, which has been used successfully for a variety of adsorbent materials [16–22].

Selectivity and capacity are not unrelated metrics. A high selectivity does not imply a high capacity, and vice versa. Many materials that exhibit the highest selectivity for certain small gasses have relatively low capacities [19, 23]. Consequently, ranking the performance of adsorbents based on these metrics individually often leads to ambiguous results. Because of this, a variety of "combined" selectivity/capacity metrics have been proposed [11, 23–27]. These metrics seek to identify adsorbents that have an ideal tradeoff between selectivity and capacity. Arguably, the most useful of these metrics is the separation

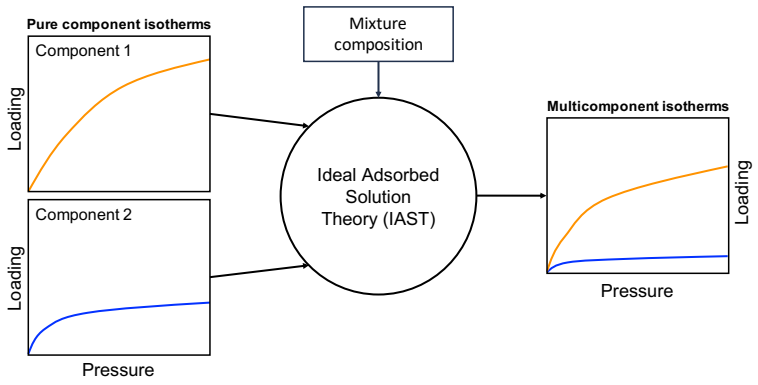

Figure 1.4 A schematic of the IAST pipeline.

potential ΔQ recently proposed by Krishna [11, 23], which is defined as

$$\Delta Q_{\mathrm{B/A}} = q_{\mathrm{A}} \frac{x_{\mathrm{B}}}{x_{\mathrm{A}}} - q_{\mathrm{B}} \tag{1.4}$$

for a binary mixture. ΔQ is particularly useful for two reasons: (i) it requires only the individual component loadings from molecular simulation or IAST calculations, and (ii) it is linearly related to the productivity of Component B (the weakly adsorbed component) calculated via breakthrough simulations [11, 23]. Another form of ΔQ that is linearly related to the productivity of Component A (the strongly adsorbed component) can alternatively be used, specifically

$$\Delta Q_{\mathrm{A/B}} = q_{\mathrm{A}} - q_{\mathrm{B}} \frac{x_{\mathrm{A}}}{x_{\mathrm{B}}} \tag{1.5}$$

Not only do both these ΔQ values vary linearly with their corresponding productivities, they are also an upper bound for those productivities. This means that ΔQ is useful for ranking candidate adsorbent materials and it can be used in the initial process design. To clarify why ΔQ is considered a combined capacity selectivity metric one can rewrite ΔQ in terms of the selectivity towards Component B, thus:

$$\Delta Q_{\mathrm{B/A}} = q_{\mathrm{A}} \frac{x_{\mathrm{B}}}{x_{\mathrm{A}}} \left(1 - \frac{1}{S_{\mathrm{B/A}}} \right) \tag{1.6}$$

If $S_{\mathrm{B/A}}) \gg 1$ then $\Delta Q_{\mathrm{B/A}}$ does not depend much on $S_{\mathrm{B/A}}$. However, for separations characterized by low to moderate selectivities, i.e $1 < S_{\mathrm{B/A}} < 10$, $\Delta Q_{\mathrm{B/A}}$ can change dramatically with $S_{\mathrm{B/A}}$.

1.1.3 Porous Materials for Separations

In the previous section, I discussed the metrics that are used to assess the performance of an adsorbent in an adsorption separation. Here, I explain what types of materials tend to be high-performing adsorbents. Adsorption occurs on the surface of a material, so materials with a large internal surface area, which allows for high capacity, are frequently desirable. Similarly, large pores and void fractions (or empty

space within the material) often lead to large capacities. For gas storage, having a large surface area and pore volume tends to guarantee (at least reasonably) high performance since capacity is of primary importance. Selectivity can be achieved mainly via three mechanisms: (*i*) the steric exclusion of mixture components from the adsorbent, (*ii*) slower diffusion of the species through the material (kinetic separation), and (*iii*) one or more component mixtures binding the adsorbate more strongly (equilibrium separation or the topic of the previous section). (*i*)-(*iii*) An adsorbent can separate mixtures based on any combination of these mechanisms. One should bear in mind that these mechanisms are not "separate". For example, a more strongly bound component will also often diffuse more slowly through an adsorbent. Furthermore, steric separation could be considered an extreme case of equilibrium separation, where the energy penalty of the excluded component binding is so high that it never happens. Steric or kinetic exclusion is a useful mechanism when the target component is more polarizeable (has stronger electrostatic or dispersion interactions). These mechanisms allow the material to be selective towards the less polarizeable component, thus producing the purified target component in the effluent and leading to greater energy efficiency. For example, N_2 can be separated from O_2 (during air separation) using a zeolite that binds the more polarizeable N_2 more strongly. However, a carbon molecular sieve that is selective towards O_2 can be used for the same separation at a quarter of the energy cost [9], despite the much higher adsorption capacity of the zeolite. This reduction in energy intensity is especially pronounced when N_2 is the desired product (see previous section).

Two important types of materials for adsorption separation were mentioned in the previous paragraph: carbon molecular sieves (CMS) and zeolites. CMS are carbon based materials synthesized to have micro- or ultramicro-pores (pore sizes of 3-9 Å), they are typically produced via oxidation and thermal treatment of high-rank coal or anthracite [8]. Their surface areas and pore volumes tend to be small compared to the other materials discussed in this section, but have small pores and are chemically and mechanically stable, making them particularly suitable for steric and kinetic separations. Zeolites are a class of microporous crystal primarily formed from SiO_4, if the zeolite is pure SiO_4 it is called a silicalite. The substitution of SiO_4 (Si^{4+}) tetrahedra with AlO_4 (Al^{3+}) tetrahedra creates an anion in the tetrahedral building unit, which can be balanced with an ion (e.g. Na^+) or hydrogenated to form an acidic site. Zeolites formed from combinations of SiO_4 and AlO_4 are called aluminosilicates. The polar structures formed in aluminosilicates lead to strong adsorption of gasses with large dipole (e.g. H_2O) or quadrupole (e.g. CO_2) moments. Silicalites, on the other hand, are relatively nonpolar and are therefore hydrophobic. One of the main advantages of zeolites is the number of structures that can be formed. 248 silicalite frameworks are known to exist [28, 29], and thousands more structures can be made by incorporating other elements (such as Al) into the silicalite framework. This leads to a wide variety of possible pore shapes, sizes, and chemistries,

the best combination of which can be selected for a particular adsorption application. Zeolites have intermediate surface areas in the 100-1,000 m^2/g range, and pore sizes of approximately 3-15 Å. Finally, since zeolites are crystals, they have predictable properties and well-controlled pore geometry/chemistry, further enhancing their tunability. An example zeolite (mordenite) is shown in Figure 1.5.

Metal-organic frameworks are another class of microporous crystal that are promising for many adsorption separations. They are comprised of metals or metal clusters in regular arrangements connected via coordination bonds with organic linker molecules, the result being a continuous framework 1-, 2-, or 3-dimensional framework. Being crystals, MOFs have many of the same advantages of zeolites, but given the number of possible metal/organic linker combinations, they are far more diverse. This and the modular nature of their construction make MOFs "designer" materials, meaning they can be (relatively) easily rationally designed to suit a particular application. For example, the linker can be functionalized with a polar moiety to enhance the adsorption of polar molecules, or the linker length can be decreased to shrink the dominant pore size. MOFs have high surface areas (ranging from 100 to over 7,000 m^2/g) and pore diameters in the 3-40 Å range. MOFs, being the focal materials of this thesis, are discussed in greater detail in the next section. An example MOF (MOF-177 [30]) is shown in Figure 1.5.

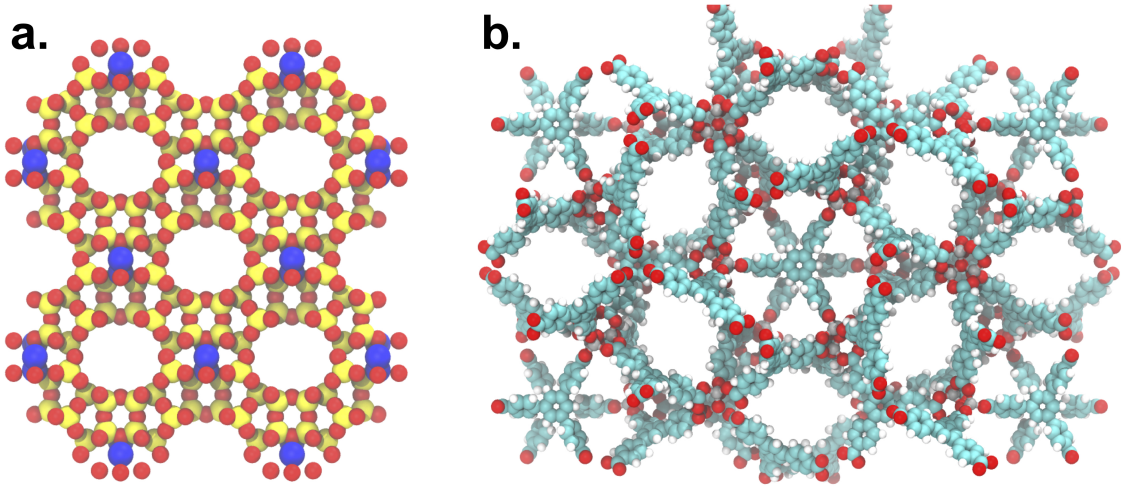

Figure 1.5 (a) Shows an example of a zeolite (mordenite), and (b) shows and an example of a MOF (MOF-177 [30])

There are numerous other types of materials that have been used or show promise for adsorption separation, including covalent organic frameworks (COFs)[31], porous organic cages (POCs) [32], and porous polymer networks (PPNs) [33]. COFs and POCs are both (mostly) crystalline materials, with many of the same advantages of MOFs and zeolites, however neither are as diverse as MOFs. COFs are

conceptually similar to MOFs, but are connected with only covalent bonds. POCs are a subset of porous organic molecules, which are a class of organics that form pores when packed, rather than filling all available void space. POCs are formed from discrete organic cage molecules (i.e. each molecule defines an intrinsic pore) that self-assemble into a crystalline material. Unlike MOFs or COFs, POCs are not connected by a continuous network of bonds, rather individual cages are held together by intermolecular interactions. PPNs are considered "quasi-ordered", exhibiting short-range order but not crystallinity [33], which makes their adsorption properties less predictable. PPNs are essentially extensions of the design principles of COFs (i.e. modular construction) but without the constraint of crystallinity. However, they do exhibit high surface area and pore volume and are relatively cheap to synthesize (compared to MOFs).

1.2 An Overview of Metal-Organic Frameworks

Metal-organic frameworks (MOFs) were introduced briefly in the previous section. Here, they are described in more detail. Fundamentally, the beauty, diversity, and usefulness observed in MOFs over the past 20 years highlight the power of combining organic and inorganic chemistry to produce structures with well-defined design rules. Significant effort has been allocated towards understanding the underlying structure of MOFs on an abstract level [34–36], yielding a robust set of mathematical design rules that have been used directly in the synthesis of some of the most important MOFs in the field. This and the modular nature of MOF synthesis have allowed for the computational construction of MOF databases (containing both existing and theoretical structures) containing hundreds of thousands of structures [37–39]. Screening these databases using molecular simulation and other modeling techniques has allowed researchers to identify and synthesize previously theoretical or unknown MOFs with high performance in a number of applications while also identifying performance limits and generalized design principles [38, 40–43]. Various iterations of this screening methodology, referred to as large-scale or high-throughput computational screening (LSCS or HTCS), have become ubiquitous in the MOF field over the past decade. Figure 1.6 shows several examples of MOFs, highlighting the metal cluster/organic linker construction.

MOFs are a type of coordination polymer, meaning they are formed from a continuous network of organic ligands bonded to metal ions. The ligands can be bound to a single metal ion or a cluster of metal ions, where each ligand can be bound to multiple metals. Throughout this thesis, I refer to MOF metal centers (whether individual metals or clusters of metals) as metal clusters or metal nodes. Commonly observed metals include, Zn, Cu, Cd, Co, and Mn. A wide variety of linker molecules can be used in the synthesis of MOFs, but the linkers need to have the capability of forming coordination bonds. Accordingly, many MOFs are constructed from linkers containing carboxylates and azoles or other heterocycles. Linkers can connect to different numbers of metals or metal clusters, depending on the number of sites per linker

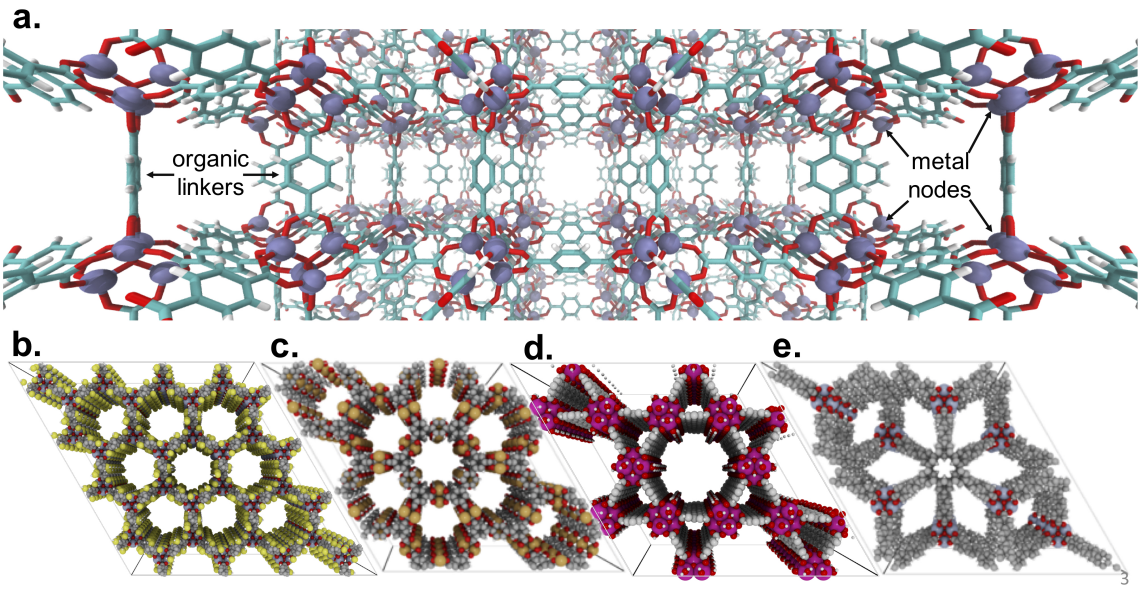

Figure 1.6 (a) Shows an example of a MOF with the organic linkers and metal nodes pointed out (b), (c), (d), and (e) are further examples of MOF structures.

capable of forming coordination bonds. A linker that connects two clusters is called ditopic, a linker connecting three clusters is called tritopic, etc. A MOF can also be synthesized from more than one metal [44–46] or linker [30] further increasing the number of possible MOF structures. MOFs are synthesized solvothermally, the assumption being that the framework formation is thermodynamically favorable, but potentially with a large kinetic barrier. Essentially, the MOF linker and metal (as salts) are dissolved in a solvent, the solution is heated allowing the framework to form, and the excess solvent is removed (filtration or drying are common methods). At this point one is left with the MOF crystal with pores full of solvent and/or other small molecules produced during synthesis, which need to be removed to obtain a porous structure, this process is called activation. Common activation methods are similar to those used to regenerate an adsorbent during adsorption separation; these are heating, lowering the pressure (applying a vacuum), or purging with another molecule (supercritical CO_2 is frequently used). Microwave and ultrasonic irradiation have also been used in lieu of direct heating for MOF synthesis [47, 48]. Note that an increasing number of MOFs have been shown to be synthesizeable at room temperature [49–51].

Among the most striking features of MOFs are their ultrahigh surface areas and void fractions. MOFs with internal surface areas of >7,000 m^2/g have been synthesized and much larger areas could be achieved in theory [42, 52]; MOFs with void fractions up to 90% have also been synthesized [30]. As one would expect, these enormous surface areas and pore volumes lead to enormous adsorption capacities, making MOFs the leading materials for the high-pressure storage of small gasses such as hydrogen and methane

[41, 42, 53]. High-density storage of hydrogen and methane is necessary for the operation of fuel tanks for methane or hydrogen powered vehicles. However, the pressures required to compress these gasses to the required densities are not safe for use in a vehicle. Therefore, the densification (at safe pressures) of hydrogen and methane via adsorption in porous materials has been a focal point of energy-related research over the past decade. This is perhaps what MOFs are currently most recognized for, however, there are still large (and probably fundamental) gaps between the Advanced Research Project Agency-Energy (ARPA-E) target for methane storage and the Department of Energy (DOE) target for hydrogen storage and the capabilities of the leading materials (MOFs) [53–55]. This is in spite of numerous computational screening studies considering hundreds of thousands (potentially millions) of MOF structures. Perhaps due, in part, to these seemingly fundamental limits, researchers have been increasingly focused on discovering other potential uses for MOFs.

MOFs have numerous other applications including catalysis [56–59], molecular sensing [60–62], drug delivery [63–66], and chemical separations [19, 43, 67–69, 69]. HTCS has great potential to identify high-performing MOFs in these applications, and significant inroads have already been made in the screening of MOFs for chemical separation. However, while gas storage depends on single-component adsorption capacities at established temperatures and pressures, adsorption separation depends on mixture adsorption capacities at a variety of mixture compositions, temperatures, and pressures (depending on the source of the mixture). This makes the screening process inherently more complex and time consuming, and previous studies have focused on, at most, a few operating conditions (related to key mixtures at industrial operating conditions) at a time. To further complicate matters, only a few small-scale studies have focused on what makes a MOF composition physically realizable, that is, why one composition is experimentally observed over other possibilities. This means that there is uncertainty as to which (if any) of the high-performing MOFs identified from a database containing theoretical structures are synthesizeable. As a result, synthesis attempts of hypothetical MOFs have been limited to structures that are similar to existing MOFs. An alternative approach is to constrain screening studies to databases of existing MOFs. This approach, while useful, does not expand the scope of existing MOFs. These issues and the current progress towards their solutions, are discussed in the subsequent sections.

1.3 Computational Screening of MOFs

As mentioned in the previous section, computational screening (often on a large scale) of MOFs has become commonplace over the past decade. This is because finding the "best" MOF for a certain application is a significant challenge, given the number of possible MOF structures. Experimental screening of more than a few structures is prohibitively expensive time consuming. This is because each MOF must

be synthesized, then characterized to verify that the correct structure was made, and finally its performance for the application being screened must be measured. Each of these steps can take days to weeks depending on the MOF and the application. However, given that MOFs are crystals and can be fully described by their unit cell, many of their properties can be modeled *in silico* using relatively small representations (a few hundred to several thousand atoms), which allows for rapid calculations. Depending on the application being modeled, each simulation can take hours to days, significantly faster than experimental screening and with less user interaction (since computational tasks can be heavily automated). Furthermore, the properties of theoretical or hypothetical MOFs, i.e. MOFs built *in silico* that do not exist experimentally, can be calculated, which allows new, high-performing MOFs to be discovered. Numerous screening studies considering databases of existing and hypothetical MOFs have been conducted since the seminal study of Wilmer et al. [37] in 2011, but the procedure remains fundamentally the same. First, a database of MOFs in built, either by mining MOFs from repositories of experimental structures [70, 71] or by building hypothetical MOFs with a construction algorithm [37–39]. Next, the performance of each MOF is calculated using molecular simulation or another type of model, this involves at least one calculation per MOF in the database. Finally, the high-performing MOFs and design rules can be identified. Design rules identified from computational screening studies often hinge on MOF structure versus performance relationships. These relationships show how well MOFs with certain properties are expected to perform for the application being screened. Researchers can then use these relationships to identify alterations to existing MOFs that are likely to increase performance in the application being screened or to guide future MOF exploration. An example structure versus performance relationship is shown in Figure 1.7, illustrating how simple design rules can be extracted from the screening of a large number of MOFs. Each step of a typical computational screening study is described in detail in the following sections.

1.3.1 Building MOFs *In Silico*

There are several existing algorithms for the computational construction of hypothetical MOFs that can be used to build large databases for screening studies. While the implementations of these algorithms vary, they must be able to build a sufficiently structurally and chemically diverse database. This is because molecular simulations on diverse structures can uncover new and useful material properties which lead to high performance in a variety of important applications, whereupon attempts can be made to replicate these properties experimentally. Fundamentally, more robust structure-function relationships can be elucidated from diverse databases, especially when considering database members with extreme properties and/or unique property combinations. Put simply, one is more likely to miss interesting or important phenomena when screening a homogeneous database.

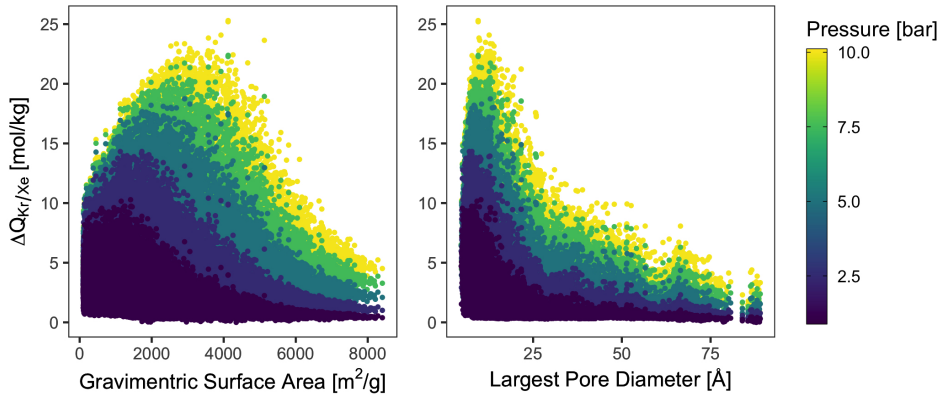

Figure 1.7 Examples of structure property relationships derived from the large-scale screening of MOFs. On each axis, each point is a MOF. The ΔQ for a 80/20 Kr/Xe separation is plotted on the y-axis versus MOF surface area and largest pore diameter (LPD) at five increasing pressures. Note that the surface area for MOFs with the highest ΔQ values increases with pressure, whereas the analogous peaks for LPD remains nearly stationary. The location and behavior of the peaks on each axis indicate ideal surface areas and LPDs for the separation of 80/20 Kr/Xe at the pressures shown.

Two methods have mainly been used to computationally construct MOFs. In the "bottom-up" approach, MOF building blocks (i.e. metal clusters and linkers) are extracted from existing MOFs and combinatorically rearranged into new MOFs [37, 72]. In the "top-down" approach, the construction process starts with a "blueprint" of the MOF describing its underlying connectivity, then arranges the building blocks (these are provided by the user to the algorithm) according to the blueprint [72–77]. Note that both the bottom-up and top-down approaches can be used to build other classes of 1-, 2-, and 3-dimensional porous crystals besides MOFs, including COFs and POCs. The blueprints used in the top-down approach are well-defined objects called crystallographic nets or topologies, which are embeddings of graphs in Euclidean space. To better describe what a crystallographic net is, it is necessary to introduce some formal definitions.

Definition 1 A *graph* is set of vertices (also called nodes) $\{v_0, v_1, ...\}$ together with a set of edges $\{e_0, e_1, ...\}$ whose components are tuples of the elements of the vertex set. Two vertices v_i and v_j are said to be connected if there is an element $e_i = (v_i, v_j)$ in the edge set.

In crystallography,*net* is often used interchangeably with graph. Per Definition 1 cannot have multiple edges connecting the same two vertices or edges connecting a vertex to itself. *Multi-graphs* are defined analogously to graphs but can have multiple edges between the same two vertices and edges connecting a vertex to itself.

Definition 2 An *embedding* of a graph is a representation of the graph in a topological space. In the context of crystallography, this topological space is always Euclidean space (to the extent of the author's knowledge). Therefore, an embedding is an assignation of coordinates (e.g. (x, y, z) in the 3-dimensional case) to the vertex set of a graph. Edge coordinates are then the centroids of the vertices they connect.

Definition 3 An *n-dimensional* graph is a graph that can be embedded in n-dimensional Euclidean space without any edge intersections, but which cannot be embedded in $n-1$ dimensional space without edge intersection [78].

Definition 4 An *n-periodic* graph has an embedding in Euclidean space which has infinite translational symmetry along each vector in a basis of the Euclidean space in which it is embedded [78].

Definition 5 A *crystallographic net* is a 3-periodic, 3-dimensional net. [78]. A crystallographic net can be defined by its unit cell, where each vertex is assigned fractional coordinates within the unit cell and edges are placed at the centroids of the vertices they connect.

Thus, a crystallographic net, or topology (the terms are used interchangeably from here on) describes the metal cluster/organic linker connectivity of the MOF. For each unique MOF, the underlying net has a possibly unique embedding in Euclidean space (typically 3-dimensional, although 2-dimensional MOFs exist). An example of a crystallographic net and a MOF built according to that net are shown in Figure 1.8

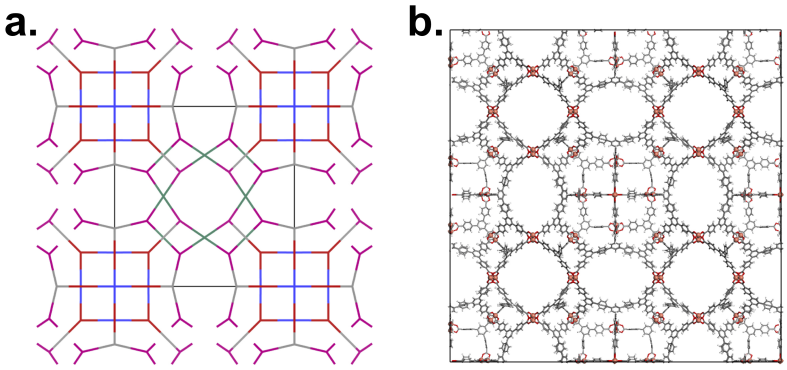

Figure 1.8 a. An example of a crystallographic net, and b. A MOF build according the the net shown in a.

One of the key subroutines in a top-down MOF construction algorithm is, therefore, to find this embedding (or close to it) according to the provided building block geometries. Once this embedding is found, the building blocks can be translated and rotated to the correct positions and bonded together to form the final MOF. The use of alchemical MOF building blocks is often convenient for this procedure to work efficiently, meaning that the provided building blocks do not match the experimental linker molecule/metal

combinations. An example of this is shown in Figure 1.9, where a large linker can be built from computationally convenient, alchemical building blocks. Topologies used to build MOFs can be extracted from the Reticular Chemistry Structure Resource (RSCR) [79], a database of crystallographic nets, where each topology referred to by a unique three- or four-letter code (**pcu**, for example).

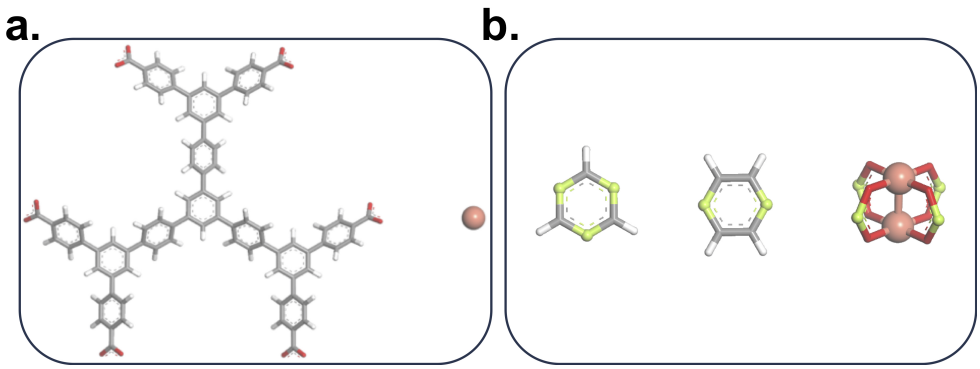

Figure 1.9 (a) What could be considered the "experimental" building blocks of a MOF, namely the linker molecule (left, a hexatopic carboxylate) and the metal (right, in this case Cu). (b) A computationally convenient decomposition of the building blocks in (a). Connection sites (where bonds are formed between building blocks) are shown in yellow, note that the linker molecule in (a) can be reconstructed from combinations of the two benzene-like rings shown in (b), and that the linker carboxylate groups are included in the computational metal node (right).

With this information in mind, a generalized procedure for building a MOF database using the top-down approach can be outlined. First, a set of MOF building blocks is created and converted into a format readable by the construction algorithm (the ubiquitous approach is to extract these building blocks from diverse existing MOFs). Second, a set of crystallographic nets is compiled. Third, the set of building blocks and the set of topologies are input to the construction algorithm, which identifies compatible combinations of building blocks and topologies, finds the correct net embedding, and then constructs a MOF for each combination. As long as enough topologies and building blocks are considered, an enormous variety of structures, that are both chemically and geometrically diverse, can be built this way. Given that most top-down construction algorithms build MOFs based on topological and geometrical rules, without explicitly considering the energetics, optimizing the geometry of each created structure according to an atomistic forcefield is necessary. This process ensures that the structures have (close to) the correct geometry by fixing bond lengths, angles, and dihedrals to chemically realistic values. Forcefields commonly employed for this purpose are the Universal Force Field (UFF) [80], UFF4MOF (an adaption of UFF specifically for MOFs) [81, 82], and the Dreiding forcefield [83]. The process of optimizing MOF geometry using forcefields is described in greater detail in Chapter 2. Our top-down computational construction algorithm (the Topologically Based Crystal Constructor or ToBaCCo), its working principles, and its uses

are the focus of Chapter 3.

1.3.2 Automated Characterization of MOF Structures

Once a database of MOFs is built or extracted from a repository of experimental structures, it is useful to characterize the chemical and structural properties of each MOF (whether to a greater or lesser extent). Given that MOF structures are stored as (x, y, z) coordinates of each atom labelled by the corresponding element symbol, many chemical properties are straightforward to derive. Certain chemical motifs, such as functional groups, can be identified via pattern matching. If the framework atom charges are available, the electrostatic potentials for the guest molecules of interest can be extracted using Coulomb's law. Linker Simplified Molecular-Input Line-Entry System (SMILES) strings can be calculated and compared with databases of organic molecules to extract the chemical properties of the linkers. Besides the numerous chemical properties that can be extracted from computational models of MOFs there are established methods for calculating their textural properties, e.g. surface area. Visual depictions of various MOF textural properties are shown in Figure 1.10.

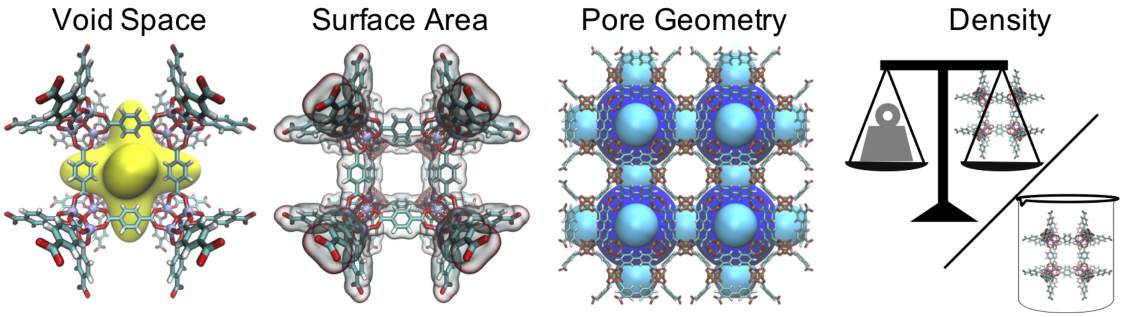

Figure 1.10 Depictions of oft-used MOF textural properties, each property is labelled. Reproduced with permission from ref. [84].

Often calculated and useful textural properties include the surface area (SA), void fraction (VF), pore size distribution (PSD), largest pore diameter (LPD), and pore limiting diameter (PLD). Surface areas are often calculated as the area "accessible" to guest molecules of a certain size, typically N_2 (since experimental surface areas are most often extracted from N_2 isotherms). Similarly, the accessible pore volumes can be calculated for various probe molecule sizes. Voronoi decomposition can be used to calculate MOF, SA, VF, and pore sizes [75, 85, 86]. This method finds the volume around each MOF atom that is closer to that atom than any other, the so-called Voronoi cell. The volume of these cells, in combination with atomic radii, can then be used to identify the occupied/unoccupied volume and SA within the MOF unit cell. The availability of this volume and surface area to a probe, a sphere of x radius, can then be used

to calculate the accessible surface area and volume as a function of x. The software package zeo++ [85] uses Voronoi decomposition to calculate the SA, VF, PSD, LPD, and PLD of a given crystal.

Void fraction can also be calculated using Widom insertions of a probe molecule, most often helium, again to correspond to experimental VF measurements [87]. Accessible SA can alternatively be calculated by "rolling" a probe sphere along the surface of the material. This procedure is implemented using a Monte Carlo technique that inserts a probe around each framework atom and calculates the fraction of these insertions in which the probe does not overlap with any framework atom [88, 89]. This fraction and the size of the probe can then be used to calculate the accessible SA [88, 89]. Finally, PSDs can alternatively be calculated by randomly selecting points within the MOF and finding the radius of the largest sphere that contains the randomly selected point and does not overlap with any MOF atoms [90]. The methods described in this paragraph are implemented in the Monte Carlo software package RASPA [91].

1.3.3 Molecular Simulation of MOF Properties

Probably the most time consuming part of screening any large MOF database is the prediction of MOF performance using molecular simulation. Essentially, researchers target a process that MOFs are expected to be useful for and attempt to replicate the process (directly or indirectly) using molecular simulation(s). Performance metrics can then be extracted from these simulations. For example, an oft-used metric for gas storage is deliverable capacity, which is the difference between adsorption loading at high pressure (storage pressure) and the loading at low pressure (release pressure). Therefore, the loadings of the stored molecules are calculated at each pressure (two total simulations) for each MOF in the database. As another example, ΔQ is a metric used to assess MOF adsorption separation performance (see above). The loadings of each component of a gas mixture at a specific composition of the gas phase are required to calculate ΔQ. These loadings can be predicted using a single grand canonical Monte Carlo (GCMC) simulation and then used in Equations 1.5 or 1.4. An approach used to reduce the computational cost of large-scale molecular simulations is to run "low-accuracy" (i.e. low-cost) simulations on the full database to eliminate unpromising structures, then run "high-accuracy" (i.e. high-cost) simulations on only the more promising MOFs. This elimination step can be repeated several times, each time running progressively more accurate simulations, a procedure called hierarchical screening. Alternatively, the initial low-accuracy simulations can be replaced with model (often derived from a machine learning algorithm) predictions, which are discussed briefly in the next section and in more detail in Chapters 2, 4, 5, and 7.

Many MOF properties including adsorption loading, diffusion coefficients of guest molecules, elastic constants, energy, and vibrational spectra, can be predicted using molecular simulation. Molecular simulations fall into predominantly two categories: quantum mechanical (QM) and classical simulations. In

QM simulations, electrons are considered explicitly, requiring the system to be modeled according to the Schrödinger equation. In classical simulations, atoms are the smallest object considered, allowing systems to be modeled according to classical statistical mechanics. The energies used in classical simulations are obtained via forcefields, or sets of semiempirical functions that describe the energetics of inter and intramolecular interactions. Multiple types of QM simulations exist, of greater or lesser accuracy, and all with the goal of solving the time-independent Schrödinger equation. There is no way to obtain an exact solution to the Schrödinger equation for systems of practical importance, so QM methods seek a solution using approximations. What approximations are made is essentially the separating element between different QM methods. The preeminent challenge in any QM calculation is the treatment of electron-electron interactions. Some methods (Hartree-Fock, Møller-Plesset, and configuration interaction) calculate atomic or molecular wavefunctions to solve these interactions, the more accurate of which are extremely computationally intensive for systems containing more than a few atoms . Density functional theory (DFT), however, only requires the computation of the electron density and does not require the computation of the wavefunction (although in practice parts of the wavefunction are often calculated) [92], making it fast and accurate enough for calculations on larger systems. DFT calculations on systems containing hundreds of atoms are routine on modern computers, making them suitable for small MOFs and MOF building blocks. Minimum energy structures, binding energies, orbital occupancies, elastic constants, transition states, and vibrational spectra are examples of material properties commonly calculated using DFT. Despite its computational efficiency compared to other QM methods, DFT it is still prohibitively expensive when considering thousands of structures or systems with thousands of atoms. One of the most promising uses of QM calculations in the MOF field is to parametrize forcefields using DFT calculated energies [93]. This is because MOF metal interactions with adsorbates and the unique coordination environments observed in MOFs are often not modeled accurately using the previous generation of forcefields.

Various thermodynamic properties can be calculated using classical simulations. Monte Carlo (MC) and molecular dynamics (MD) simulations are the most frequently used algorithms for classical simulations. In MC, the system configurations are sampled over a series of cycles where, in each cycle, "moves" of each atom or molecule (e.g. translation or rotation) are proposed and accepted if the move lowers the total energy of the system. If the move does not lower the total energy of the system, it is accepted with some (variable) probability. As long as a sufficient number of cycles are run, the system evolves through many of its most probable states in this manner. In molecular dynamics, a system is allowed to evolve over time according to Newton's equations of motion. In MD, the total time is approximately analogous to the number of cycles in an MC simulation, so a system will visit many of its

most probable states during an MD simulation given sufficient time. Various thermodynamic quantities can be output from both MC and MD simulations as averages over the number of cycles (for MC) or the simulation time (for MD). GCMC simulations can be used to calculate the enthalpies of adsorption and adsorption loadings (at specified pressures and temperatures) for single adsorbates and for mixtures. MD can be used to calculate the diffusion coefficients of guest molecules in MOFs, free energy, elastic constants, and vibrational spectra. As stated above, classical simulations of MOFs require forcefields that yield the energetics of intraframework interactions and interactions between framework atoms and guest molecules. One common assumption for GCMC simulations is that the MOF is rigid, meaning the MOF atoms are not allowed to move, thus removing the need for a forcefield describing intraframework interactions. This assumption has been proven accurate in many instances, although it is not sound for some flexible MOFs. The methods mentioned in this section are gone over in greater detail in Chapter 2, including key equations and theoretical considerations.

1.3.4 Predictive Modeling of MOF Adsorption Performance

As mentioned in the previous section, molecular simulations can sometimes be replaced by a model capable of predicting MOF performance based on some representation of the MOF and the process being considered. Such a model can be thought of as a "black-box" function which takes as input variables describing the MOF and application (if relevant) and outputs a performance metric. Generally, model predictions are (many) orders of magnitude faster than molecular simulation predictions. For example, a trained machine learning (ML) model can make predictions of adsorption loadings in fractions of a second per MOF, whereas a GCMC simulation is likely to take several minutes to several hours per MOF (depending on the MOF and application). A popular approach is to train ML models on data obtained from GCMC simulations. This allows large and diverse training sets to be used, increasing the accuracy and transferability of the model.

One of the main difficulties in training any ML algorithm to predict MOF performance data is how to represent MOFs to the algorithm. There is often a trade-off between the complexity of the variables (also called descriptors) considered (i.e the time taken to calculate the variables) and the accuracy of the model. Ideally, the model takes as input relatively easy-to-calculate descriptors while maintaining sufficiently low error. The MOF textural properties described above were some of the first variables used to describe MOFs to ML algorithms [94] and are still in use today, as they tend to yield accurate models when used in concert with other variables. Woo and coworkers have developed a set of MOF descriptors called atomic-property-weighted radial distribution functions (AP-RDFs) [95], which simultaneously describe MOF geometry and chemistry and have been used effectively (often in combination with MOF textural

properties) to predict both pure CO_2 [95] adsorption in MOFs and CO_2 adsorption from mixtures [96, 97]. Energy-based MOF descriptors have also shown promise for predicting hydrogen storage metrics on MOFs [53, 98], these descriptors are derived from the adsorption energies of a guest molecule calculated using short simulations. Thus, models trained using this data from rapid simulations are able to predict the results of more time-consuming simulations. Textural properties combined with simple descriptors of MOF chemistry (e.g. the number densities of certain functional groups, metals, etc.) have also been proven accurate to predict a variety of adsorption data [84, 99, 100]. Increasing the "universality" of adsorption-predicting ML models is the key to their future usefulness. Most of the models trained so far can make predictions for a single gas or gas mixture at only one operating condition. Such models can be used to screen new MOF databases for the same application, but a new model must be trained to screen existing databases for different applications, severely limiting the long-term usefulness of single-purpose models. Models that can make predictions at multiple operating conditions and/or for multiple adsorbates are starting to be developed [84, 101–103]; their development hinges on the use of variables describing the adsorbate(s) and operating conditions during training. These increasingly universal models are likely to be the primary future direction of ML prediction of adsorption in MOFs.

1.3.5 MOF Synthesizability

This section is reproduced with permission from: R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. Gómez-Gualdrón. "Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: from molecular simulation to machine learning" *Chem. Mater.* 30, 18, 6325–6337. Copyright 2018 American Chemical Society.

MOF researchers often consider the abstract concept of MOF space, a volume occupied by all possible MOF structures, where directions are related to MOF composition. For example, one direction could be the metal element, another direction the organic linker, and a third direction the synthesis conditions (temperature, solvent, etc.). These example directions span MOF space, many other directions can be conceptualized, they are only meant to illustrate the concept of MOF space. Each point in MOF space, represented in a manner directly analogous to points in Euclidean space (as a list of positions along spanning vectors), defines a MOF. When screening MOFs for an application, we extract a subspace from MOF space and map each point onto a set of performance metrics. Thus, creating a new "design space" of the materials considered during screening and the performance metrics relevant to the application. If we allow MOF space to include points that correspond to "impossible" MOFs, or MOFs that follow the rules of chemistry but cannot physically exist, then the space of MOFs that are "possible" (i.e. that are "synthesizable") must be a subspace of this greater volume. This subspace of synthesizable MOFs is what

we are primarily concerned during a computational screening study. Impossible MOFs may be useful for deriving structure versus performance relationships or for training robust machine learning algorithms but are clearly of little importance in the direct discovery of new materials.

But, how can we tell if a hypothetical MOF is possible? Rigorously, for a MOF to be synthesizable, there must be some mapping of the point in MOF space to our physical universe. That is, there must be some set of physical interactions that produce the MOF in question. In theory, we should be able to find this set of interactions for any synthesizable MOF. Unfortunately, we are far from this ability in practice. However, we can use existing theories to progress towards differentiating between possible and impossible MOFs. We know from thermodynamics that the high energy configurations of a system are less probable (sometimes to the point of impossibility) than low energy configurations. So we can surmise that a hypothetical MOF with high free energy (relative to some idealized state) is unlikely to form from its constituents, the assumption being that the free energy of the unorganized MOF building blocks is lower than that of the "built" MOF. Furthermore, for MOFs with the same linker and node but different topologies, referred to as isomorphs, the MOF (i.e. isomorph) with lower free energy is more likely to be observed experimentally. For systems as complex as MOF synthesis solutions, there are plenty of other reasons a MOF may not be synthesizable. For example, a MOF could be kinetically inaccessible. That is, it may be thermodynamically favorable, but it is separated from the synthesis solution by an insurmountable free energy barrier. Thus, it will never be formed on a reasonable timescale. MOFs can also be "practically impossible"; for example, a certain MOF may require a linker that is chemically possible but not feasible to synthesize.

The goal of screening databases of experimental MOFs is to identify previously synthesized structures that are likely to be high-performing in new applications. Given that there is no uncertainty as to whether the identified high performing MOFs can be synthesized, such studies are more likely to have a quick impact, as new experimental methods do not need to be developed and the experimental characterization of the identified material has already been completed (at least to some extent). However, screening databases of hypothetical MOFs expands the boundaries of MOF space and considers new designs that may work for a purpose via a unique mechanism or outperform any existing material. However, one of the primary issues with screening databases of hypothetical MOFs is precisely the question posited in the previous paragraph: how can we tell if a hypothetical MOF is possible? Currently, synthesis attempts of hypothetical MOFs are guided by their similarity to already existing MOFs and chemical intuition. Candidates are selected out of the high-performing MOFs predicted by a screening study if they, for example, have the same topology and node as an existing MOF, but a slightly different linker. This is a reasonable procedure: it is likely to result in a new, high-performing MOF, and does not require expensive,

exploratory experiments. Nevertheless, it limits the experimental exploration of MOF space to regions close to existing MOFs and often misses out on potentially higher performing MOFs. The tendency to under-explore MOF space is highlighted in Chapter 3. There, data is presented indicating that out of over 40,000 articles on MOFs (published at the time), approximately 50% of these focus on 30 MOFs. Hence, despite the primary appeal of MOFs being their unparalleled structural and chemical diversity, research has mainly focused on fewer MOFs than there are zeolites.

Considering the discussion in the previous paragraphs, one way to modify typical HTCS protocols to accelerate MOF discovery (and encourage wider experimental exploration of MOF-space) is to predict the "synthesizability" of MOF prototypes in a database. The development of effective criteria for synthesizability is tied to understanding how MOF synthesis occurs. Although recent experiments have provided some insights into the fundamental processes controlling MOF nucleation and growth [104] there remains a lack of consensus on exactly what factors control whether a MOF can be synthesized or not and whether these factors are thermodynamic or kinetic in nature. Molecular simulations can potentially clarify how and why certain MOFs are observed experimentally, and others are not. Yet, compared to the number of studies that focus on MOF property prediction, only a few simulation studies have investigated MOF properties directly related to their synthesis.

In 2004, Mellot-Draznieks et al. used Monte Carlo simulations to anneal building blocks (starting from a disordered arrangement) into MOFs [105, 106]. Although the method was successful in predicting potential MOF structures for a given set of building blocks, few conclusions about MOF stability or formation mechanisms could be made due to the absence of dynamics and an unrealistic description of building block interactions. Since 2015, researchers have started to use molecular dynamics (MD) simulations to follow kinetic aspects of MOF self-assembly. Starting with disordered building blocks in an implicit solvent, Yoneya et al. were able to form a crystalline framework over a long MD simulation by tuning the metal-linker interactions [107]. However, the effects that this artificial tuning of the building block interactions might have on the formation mechanism remain unclear. Biswal and Kusalik used equilibrium MD to observe the early formation stages of MOF-2/MOF-3 in an explicit solvent. However, only a primarily amorphous phase formed during the simulation [108, 109]. Kollias et al. [110] used metadynamics simulations to probe the free-energy landscape during the formation of crystal-like secondary building units (SBUs). Notably, Colón et al. [111] used MD simulations and the finite string method to probe the free-energy landscape during the assembly of 27 SBUs into an approximately 2.5 nm MOF-5 crystallite. Evidently, the use of biased MD to model MOF self-assembly is encouraging, but questions remain on how the choice of interaction parameters, reaction coordinates, and model simplifications could impact conclusions.

Another set of simulation studies have focused on the energetics of fully assembled MOFs, often with the intending to understand the synthetic preference for certain isomorphs. MOF isomorphs, while all having the same composition, can have considerably different properties [40, 77, 112] because of the differences in how the building blocks are arranged in space, that is, because of the differences in crystal topology [34–36]. Accordingly, when a high-performing MOF prototype is identified during HTCS, there is a chance that a worse-performing isomorph is actually favored during synthesis. Using molecular mechanics, Amirjalayer et al. [113] showed that the experimentally observed HKUST-1/Cu-BTC [114] (**tbo** topology) has lower strain energies than its unobserved pto isomorph. Analogous findings were reported for the experimentally observed MOF-14 [115] (of the catenated, **pto-c** topology) and its unobserved **tbo-c** isomorph [113]. Impeng et al. showed that three experimentally observed Cu-paddlewheel MOFs of the nbo-b topology had lower strain energy than four hypothesized isomorphs, respectively [116].

Using density functional theory (DFT) energies, Cai et al. [117] explained the preference between the **tbo** and **fmj** topologies for four MOFs (one of them being Cu-BTC) depending on how the BTC linker was functionalized. Arhangelskis et al. used DFT energies to determine the most stable isomorph among eight zeolitic imidazole framework (ZIF) prototypes based on a CF_3IM linker, which was subsequently synthesized [118]. On the basis of the similarity of DFT formation energies, Liu and Truhlar explained the often reported coexistence of the MOF isomorphs NU-901 and NU-1000 (**scu** and **csq** topology, respectively) [119]. Curiously, an inspection of the literature [120–122] seems to indicate that synthesis more readily yields the reportedly less stable isomorph, NU-1000, which could be due to a number of synthetic factors not being considered in the calculations.

Although strain/lattice energies have shown some success in explaining the synthetic preference for a given MOF isomorph, [113, 116–118] these quantities do not account for thermal effects. On the other hand, there are instances where temperature (while keeping other synthesis conditions constant) has been reported to change the topology of the synthesized MOF [123–125]. Thus, in principle, a temperature dependent quantity such as free energy should be considered in MOF synthetic accessibility prediction. Gee and Sholl [126] calculated the harmonic free energy and free energy of solvation for 25 ZIF isomorphs to assess their relative stability. A similar evolution in prediction methods can be seen in the pharmaceuticals field, where free energy calculations [127–131] are taking the place of traditional lattice energy calculations [132–134], as the former increasingly appear necessary to predict experimentally observed structures among energetically close organic crystal isomorphs [127].

1.4 MOFs for Gas Separations

MOFs are promising materials for a variety of separations due, in part, to their high adsorption capacities and well-defined pore structures, but mainly due to their unparalleled tunability. This tunability arises directly from the number of possible MOF structures, which arises, in turn, from the modular nature of MOF construction. MOF pore geometry and chemistry can be rationally altered by using linkers of a certain size and/or chemistry, allowing researchers to tune MOFs to be selective towards certain molecules based on the molecule size, shape, and chemical interactions. Zeolites, whose pore geometry is limited by the size/shape of their tetrahedral building units and their relatively few 248 topologies [29], have potential in a limited number of separation applications. That is, if one is targeting zeolites for use in a separation, a zeolite with the correct pore size(s) must already exist, otherwise it is unlikely any zeolite will be sufficiently selective. Such a constraint does not exist for MOFs, since their pore geometry is limited only by the building blocks used (of which there are innumerable choices) and the (potentially) thousands of topologies available to MOFs. A similar limitation exists for zeolite pore chemistry. The hydrophobicity and reactivity of zeolite pores can be tuned to some extent by changing the Al^{3+} content, which introduces dipoles or acidic sites. However, the spatial distribution of these active sites is difficult to control, and the accessibility of these sites to guest molecules depends heavily on the zeolite topology. Again, in contrast to zeolites, MOF pore chemistry can be altered in numerous ways. For example, by functionalizing the organic linker molecules, using different metals, by changing the spatial arrangement of active sites, or via combinations of these methods. Research on using MOFs for separations has been broad, but mostly in the experimental realm, although there have been computational modeling/screening studies on MOFs for several important separations. A brief overview of the research on MOFs for selected, important gas separations, both experimental and computational, is given herein.

1.4.1 CO_2 Capture

Research on MOFs for CO_2 capture, due to concerns over harmful emissions contributing to anthropogenic climate change, has been the focus of intense study over the last decade. The typical composition of postcombustion flue gas at a power plant is 15% CO_2 and 85% N_2 [135]. This flue gas is emitted from a combustion process at ambient conditions, at which the majority of MOFs are selective towards CO_2 (a material could only be selective towards N_2 via a size exclusion mechanism), making postcombustion adsorption of CO_2 using MOFs a promising method to limits CO_2 emission. For this separation, the CO_2/N_2 adsorption selectivity of the adsorbent is less important than the CO_2 deliverable capacity, which is how much CO_2 the absorbent can release at the regeneration temperature (for a TSA cycle) or pressure (for a PSA cycle) [135]. MOFs have also been considered for precombustion carbon

capture, which can be used in the integrated gasification combined cycle (IGCC) process. In an IGCC process, a fuel source is converted (over a series of steps) to a gas mixture of approximately 40% CO_2 and 60% H_2, the CO_2 is captured from this mixture and the remaining H_2 is combusted [136].

One mechanism that allows MOFs to be particularly selective towards CO_2, for pre or postcombustion capture, is the strong binding of CO_2 with MOF open metal sites (OMS). These sites are exposed under-coordinated metals that often interact strongly with adsorbates. Due to this mechanism, experimental research on MOFs for CO_2 capture has put significant effort towards developing MOF platforms with a high density of OMS. In 2008, Matzger and coworkers demonstrated that MOF-74 (a MOF that can be synthesized with various metals) has greater CO_2 uptake at 298 K and atmospheric pressure when synthesized with Mg over Co, Ni, or Zn [137]. More recently, Zhai et al. demonstrated that CO_2 uptake and CO_2/N_2 selectivity could be tuned upwards by altering the metal composition of another high-performing MOF platform, CPM-200 [138]. Some MOFs have topologies and pore geometries that allow for the cooperation of multiple OMS in the binding of a single CO_2 molecule. The effect of these "single molecule traps" can be modulated using linker size, as shown by by Zhou and coworkers with their MOF PCN-88 [139]. The metal species and spatial arrangement of OMS are not the only features that can be tuned to enhance CO_2 capture performance. Additionally, MOF linkers can be functionalized with groups expected to interact strongly with CO_2. For example, Wang et al. increased the CO_2 selectivity and capacity in a postcombustion environment of the MOF UIO-67 using modified linkers with larger dipole moments [140].

Many computational studies of MOFs for CO_2 capture have also been conducted over the past decade. Frequently, these studies focus on specific adsorption mechanisms in a small set of MOFs that enhance CO_2 capacity or selectivity. However, here, I will discuss a few studies focused on large-scale screening, as these are more pertinent to the themes of this thesis. In 2014, Woo and coworkers trained a machine learning algorithm capable of classifying a MOF as high-performing or low-performing as assessed by CO_2 capacity; they proposed that this algorithm could be used to identify the top 10% of MOFs from a large database for further screening using molecular simulations [99]. More recently, in 2019, Woo and coworkers expanded on their previous work to develop a machine learning model capable of the direct prediction of CO_2 working capacity and selectivity for precombustion carbon capture, this model was able to predict both metrics with sufficient accuracy to be used as a first step in a hierarchical screening study [96]. Finally, in their most recent work (2020), Woo and coworkers trained a set of neural networks for predicting CO_2 working capacity (for a 1 bar to 0.15 bar PSA cycle) and selectivity in postcombustion conditions, again showing the model to have sufficiently low error for reducing the number of MOFs screened with more computationally intensive methods [97]. In a recent work by Boyd et al. MOFs with a

unique cooperative binding mechanism that allowed them to maintain high CO_2/N_2 selectivity in the presence of water (wet flue gas environment) were identified from a large MOF database [39]. Two of these MOFs were then synthesized and demonstrated (experimentally) to maintain high CO_2 capture performance even with high relative humidity [39].

1.4.2 Separation of Alkenes from Alkanes

Alkenes such as ethene and propene are used in the manufacturing of plastics and are, therefore, some of the most important industrial chemicals on the planet. Alkenes are mostly produced through hydrocarbon cracking (this process is briefly outlined above for ethene production), which produces mixtures of alkenes and alkanes, from which the valuable alkenes must be separated. This separation is currently accomplished with cyrogenic distillation, which accounts for a significant percentage of the total energy usage of alkene production [4]. Alkene/alkane separation is particularly difficult as the chemical properties of alkenes are similar to those of alkanes. Unlike CO_2 which has a large quadropole moment allowing it to be separated from weakly interacting gases like N_2 and H_2, gaseous alkenes and alkanes both primarily interact through dispersion interactions. However, MOFs have been found to bind ethene and propene strongly via a π-back donation interaction, in which MOF open metal sites (OMS) donate an electron to the antibonding orbital of a π-bonds. This mechanism is often referred to as π-complexation.

One of the first demonstrations of using the π-complexation interaction to selectively bind alkenes was a study by Yoon et al. who demonstrated that the enthalpy of adsorption for propene in MIL-100 could be dramatically increased when MIL-100 was activated in a manner to produce a large quantity of OMS (while the enthalpy of adsorption for propane remained the same) [141]. Bao et al. found that Mg-MOF-74 had much higher affinity for ethene over ethane (estimated selectivity of 15.3) and propene over propane (estimated selectivity of 18.7) due to the high density of OMS in Mg-MOF-74 yielding many π-complexation interaction sites [142]. Bao et al. followed up on this study by testing different versions of MOF-74 synthesized with metals besides Mg, finding much greater selectivity towards propene in Co-MOF-74 (estimated selectivity of 46) or Mn-MOF-74 (estimated selectivity of 24) than in Mg-MOF-74 [143]. Yang et al. found that NOTT-300 exhibited very high ethene/ethane selectivity (estimated at 48.7) due to interactions between the ethene π-electrons and -OH framework groups [144]. Furthermore, they found the lowest energy configuration of ethene within the NOTT-300 framework lead to π-π stacking between multiple adsorbed ethene molecules, allowing for a greater ethene uptake capacity.

There have been relatively few computational screening studies considering MOFs for alkene/alkane separations, in large part due to the difficulty of describing the π-complexation interaction, a quantum mechanical mechanism, with classical forcefields. Therefore, significant effort has been allocated towards

developing forcefields that accurately describe this interaction using DFT-calculated interaction energies [145–147]. Jorge et al. and Kulkarni and Sholl were able to obtain a reasonable matching between experimentally measured ethene/propene and GCMC calculated adsorption isotherms in MOFs with Cu OMS using this procedure [145–147]. Kulkarni and Sholl also performed a small screening study of existing MOFs with Cu OMS using their DFT-derived forcefield, identifying several promising MOFs for use as adsorbents for ethene/ethane and propene/propane separation [145]. Altintas and Keskin screened 278 existing MOFs for ethene/ethane selectivity and capacity, finding ethene adsorption selectivity to decrease with increasing pressure and with increasing pore size [148]. They also screened for kinetic separation performance by calculating the diffusion coefficients of ethene and ethane in each of the 278 MOFs, finding kinetic selectivity to be similar or lower than adsorption selectivity in the MOFs they studied [148].

1.4.3 Noble Gas Separation

Noble gases such as krypton and xenon have many industrial applications and, given their inertness, are difficult to separate. Kr is used in fluorescent lamps and xenon has several applications in lighting, laser manufacturing, and medicine [135, 149]. Kr and Xe, are also rare, being extremely dilute (both comprise less than a single ppm in the atmosphere), making them high value chemicals. Kr and Xe are mainly produced via the cryogenic distillation of air. In this process, an 80/20 Kr/Xe mixture is isolated from the other components of air, after being heated and compressed, this mixture must then undergo another cryogenic distillation to produce pure Kr and Xe [5]. Kr/Xe mixtures are also present in the off-gas stream of nuclear powerplants [135, 150]. The radioactive isotope ^{85}Kr, produced by fission of uranium or plutonium, has a half-life of over 10 years, whereas the Xe isotopes produced have much shorter half-lives. Hence, the Xe in spent nuclear fuel can be reused quickly, but only if it separated from ^{85}Kr. Given that Xe is larger and more polarizeable than Kr, the primary mechanism used to separate Kr from Xe is via selective adsorption, however, size exclusion mechanisms can also be used.

Mueller et al. were the first (2006) to use a MOF (HKUST-1) to separate a Kr/Xe mixture, using a PSA cycle and showed that highly pure Kr could be obtained in the tail stream [151]. This combined with the high capacity of HKUST-1 for Xe highlighted the potential of MOFs for separating noble gases [151]. As in CO_2 capture and alkene/alkane separation studies, researchers have also tested the MOF-74 platform with various metals for Kr/Xe separation. In 2014, Perry et al. showed experimentally that Co-MOF-74 has higher Xe selectivity than MOF-74 synthesized with Ni, Mg, or Zn [152]. They also investigated the effect of pore size by measuring Xe selectivity in a set of isoreticular MOFs (MOFs that have the same topology and node but different linkers), finding that the smallest-pore MOF, NOTT-100, had the highest Xe selectivity [152]. More recently, in 2017, Carreon and coworkers used the MOF ZIF-8 in a membrane

separation process with an overall selectivity towards Kr, despite ZIF-8 having having equilibrium adsorption selectivity towards Xe [153]. This is because the ZIF-8 membrane was kinetically selective towards Kr [153]. In other words, ZIF-8 has a pore window smaller than the kinetic diameter of Xe, making the energy barrier of diffusion through ZIF-8 higher for Xe than for Kr.

Kr/Xe separation has been the focus of a few important computational screening studies, starting with the work of Snurr and coworkers in 2012 [19]. They screened a database of 137,000 hypothetical MOFs for Xe selectivity and capacity for an 80/20 Kr/Xe mixture at 1,5, and 10 bar using GCMC simulations, finding that MOFs with uniform pores slightly larger than the size of a Xe atom (or 4-5 Å in diameter) had the highest Xe selectivity, but relatively low Xe capacity, highlighting that the most selective materials are not necessarily the most capacious [19]. Additionally, they found that slightly larger pores, up to around 8 Å, were best for an ideal selectivity/capacity tradeoff [19]. In 2016, Thallapy and coworkers screened a database of 125,000 MOFs, containing around 5,000 existing MOFs for Xe selectivity at the dilute conditions relevant to spent nuclear fuel reprocessing [43]. They found that one of the existing MOFs, SBMOF-1, had an exceedingly high predicted Xe selectivity (>70) as well as reasonably high Xe capacity [43]. After identifying SBMOF-1 computationally, the authors tested it experimentally in a breakthrough column for separating Xe from a gas mixture representative of nuclear off-gas, finding that the breakthrough time for Xe was over an hour in their lab-scale setup.

1.4.4 Separations Involving Methane

Natural gas is an important transitional fuel source between petroleum and coal and renewable energy sources. The U.S. has abundant natural gas reserves and methane combustion produces less harmful emissions than petroleum [154]. As a result, a large portion of computational MOF research has been focused on screening materials for methane storage in vehicle fuel tanks. However, many natural gas deposits contain 10-20% (or more) N_2 [155], which must be reduced below 5% before use. Currently, this is accomplished using cryogenic distillation. Given that CH_4 is still the majority component in these N_2 rich natural gas deposits, N_2 selective adsorbents, which rely on a sieving mechanism to exclude CH_4, are more energy efficient, as the product stream will contain the purified CH_4 in this case. Given that the adsorption selectivity of an adsorbent will (almost) always be in favor of CH_4 (so long as it can fit in the adsorbent), the exploration of MOFs, whose pores are generally larger than the kinetic diameter of methane, for N_2/CH_4 separation has been limited. However, CH_4 selective adsorbents are still desirable to purify dilute sources of methane, such as N_2/CH_4 mixtures that are >50% N_2. For example, N_2 rich streams from some cryogenic distillation processes can contain large percentages of methane (approximately 25%) [156], which cannot be emitted to the atmosphere, as methane is a greenhouse gas. A methane selective adsorbent can

be used to purify this N_2 rich tail gas before it is released to the atmosphere, the CH4 rich gas mixture retained in the adsorbent can then be recycled back through the cryogenic distillation unit [156]. Landfill gas (LFG) is another source of (relatively) dilute methane emissions, concentrating the methane in LFG in an adsorbent material is a potential way to limit methane emissions while obtaining another source of fuel [157, 158]. While N_2/CH_4 separations have garnered significant attention with the increasing demand for natural gas, the separation of methane from other light hydrocarbons is also an important industrial process currently accomplished with cryogenic distillation [159]. For example, methane/ethane separation is of interest in natural gas processing, as natural gas often contains significant percentages of ethane, which is a precursor in the synthesis of ethene (an important industrial chemical) [7]. Besides the cases specifically discussed in this paragraph, there are numerous other industrially-relevant separations involving methane and other hydrocarbons or small gases.

An interesting study by Yoon et al. in 2017 showed that the Cr MOF MIL-100(Cr) was selective towards N_2 via an equilibrium absorption mechanism, the first (and still only, to the authors knowledge) such material [160]. The strong binding of N_2 was attributed to the complexation of N_2 with the abundant open metal sites (OMS) in MIL-100(Cr), similar to the complexation of ethene/propene with OMS discussed in the previous section [160]. However, Krishna found that the N_2 selectivity (between 8 and 20 depending on the conditions) of MIL-100(Cr) is not sufficient to be useful in industry, since a large percentage of the CH_4 in the mixture will still adsorb with the N_2, making the yield of 96% pure CH_4 unacceptably low [11]. MOFs, however, are promising as methane-selective adsorbents to purify nitrogen rich streams. Another study by Li et al. in 2018 found that a newly synthesized squarate-based MOF had much higher methane selectivity (approximately 12.5) than 14 other MOFs for which comparable measurements were available, making their MOF one of the most promising existing structures for upgrading natural gas from unconventional sources or removing methane from emissions [158]. In 2014, Jia et al. investigated a series of isoreticular MOFs (JUC-100, JUC-103, and JUC-106) for the separation of C_3H_8/CH_4 and $C_2H_6CH_4$, finding all three MOFs to be promising structures with high selectivities and capacities for the heavier hydrocarbons. In a 2020 simulation study, Ponraj and Borah demonstrated the efficacy of using a popular MOF, HKUST-1, for separating methane/ethane and methane/propane mixtures, as both an adsorbent material and as a membrane [161]. Ponraj and Borah also screened a large database of experimental MOFs in 2021 (over 12,000 structures) for separation of methane from ethane and propane, considering selectivity, working capacity, and adsorbent regenerability [162]. Regenerability is a measure of how much pore space can be continuously regenerated in a cyclic adsorption process, the expectation being that some on the strongly adsorbed component will remain in the adsorbent pores ad infinitum. They found a tradeoff relationship between ethane/propane selectivity and capacity, similar to

that discovered by Snurr and coworkers for Kr/Xe separation [19], and identified several promising structures for separating various methane/ethane/propane mixtures [162].

1.5 Improving the Computational Screening of MOFs

Throughout this introduction, I have introduced several problems with the pipelines currently used for the high-throughput computation screening (HTCS) of MOFs. Here, those problems are compiled and summarized.

Problem 1 *There is a lack of topological diversity in the hypothetical MOFs primarily considered in HTCS studies.* This problem was more relevant at the time when I began my Ph.D. studies, this was mainly due to limitations in the methods being used for computational MOF construction. Since then, recent screening studies have considered MOFs built according to hundreds or thousands of topologies. Nonetheless, the importance of topological diversity in HTCS is still relevant and should be considered in a successful screening study.

Problem 2 *There is a lack of diversity in the applications considered in HTCS studies.* MOFs have large capacities for small gasses at high pressures, but their performance is still well below ARPA-E targets for methane and hydrogen storage, and likely cannot be pushed noticeably higher with new screening studies [38, 53, 54, 163]. After many studies on methane adsorption considering, in total, an enormous number and variety of MOFs, no materials close to the ARPA-E target have been found, there is likely a theoretical limit for MOF methane capacity that is not much higher than the current world record. The case of H_2 storage is similar [53]. However, MOFs are still promising materials for adsorption-based separations, and given the large number of gas separations currently accomplished using cryogenic distillation, there are opportunities to discover new MOFs and repurpose existing ones for diverse gas separation applications. Despite this, only a few separations have been considered when screening MOFs on a large scale.

Problem 3 *HTCS for adsorption applications using molecular simulation is often prohibitively computationally intensive when considering multiple operating conditions and/or adsorbates.* This is especially true when screening for separation performance, which frequently requires several GCMC simulations per MOF. An increasingly popular solution is to replace GCMC simulations with machine learning (ML) predictions.

Problem 4 *Most existing ML models are "single-purpose".* That is, they can only make predictions for a single adsorbate or mixture at a single operating condition (temperature and pressure). These models can be used to screen new MOFs for a single purpose. However, as noted previously, the main limitation in MOF screening is not the number of structures, but the number of applications considered.

Problem 5 *There is uncertainty as to what hypothetical MOFs can be synthesized.* A central goal of any HTCS study of hypothetical MOFs is to identify new materials that are high-performing in an important application. However, since there is uncertainty as to what hypothetical MOFs can be experimentally synthesized, synthesis attempts have been constrained to new MOFs that are similar to the existing ones. This limits the exploration of MOF space and misses out on potentially higher-performing materials.

1.6 Thesis Outline

The purpose of this thesis is to improve the methods used for high-throughput computational screening (HTCS) of MOFs, enabling more efficient materials discovery for diverse gas separation applications. I present the methods we developed for building diverse sets of MOFs *in silico*, data and algorithms we used to better understand how MOF platforms can be improved in the context of CO_2 capture, our development of increasingly "universal" models for the rapid prediction of adsorption loadings in MOFs, and large-scale free energy calculations which can be used to derive criteria for MOF synthetic likelihood. While the focus throughout is on MOFs, many of the methods and ideas described here are applicable, in many cases directly, to other classes of porous crystals.

In Chapter 2, the primary methods used in Chapters 3 through 7 are described. I discuss the principles of the molecular simulation methods used, specifically grand canonical Monte Carlo, molecular dynamics, thermodynamic integration, and density functional theory. Following this, I briefly outline the machine learning algorithms employed in Chapters 4, 5, and 7.

In Chapter 3, our algorithm, the Topologically Based Crystal Constructor (ToBaCCo), for constructing MOFs *in silico* is described. The working principles of the algorithm and the relevant graph theory background are included. Following this, data is shown that demonstrates the importance of considering diverse topologies when screening MOFs. Specifically, the data shows how several important MOF properties can vary widely across MOFs with the same composition, but different topologies (i.e. isomorphic MOFs).

Chapter 4 is a study on the influence of pore chemistry and topology on the pre and postcombustion CO_2 capture performance of MOFs. The influence of various MOF properties, including topology and linker functionalization, on five CO_2 capture metrics is determined using a combination of GCMC and DFT simulations, and machine learning. Generally applicable criteria for improving the CO_2 capture capabilities of existing MOFs are derived.

Chapter 5 focuses on the development of a unique deep learning model capable of predicting the isotherms of multiple small gas adsorbates in diverse MOFs. This was the first example of a single model capable of predicting adsorption loadings at multiple pressures for multiple adsorbates, without any

modification. Its development hinged on the use of alchemical adsorbates (adsorbate models not representative of any existing molecule). By using alchemical adsorbates, we were able to produce training data using GCMC simulations with a broad and regular array of adsorbate properties, which allowed the model to precisely learn the behavior of adsorption loading with adsorbate properties.

In Chapter 6, different methods for the large-scale calculation of MOF free energies are outlined. Specifically, we discuss and compare two methods for calculating the free energies of crystals: the quasiharmonic approximation and the Frenkel-Ladd path. We conclude that the Frenkel-Ladd path is the better method for the calculation and subsequent comparison of the free energies of diverse MOFs. Following this, we use the free energies calculated on a large MOF database (that includes both existing and hypothetical structures), to derive two criteria that can be used to determine if a MOF is thermodynamically accessible, that is, if it is likely to be synthesizable from a thermodynamic perspective.

Finally, in Chapter 7, a new model capable of predicting mixture adsorption at multiple pressures and for multiple mixtures is developed. This model uses single-component isotherms predicted by an updated version of the neural network described in Chapter 5 as input for an Ideal Adsorbed Solution Theory (IAST) calculation. IAST calculates the mixture loadings given the single component isotherms of each mixture component. First, we prove that this combined neural network/IAST model makes sufficiently accurate predictions for use in a hierarchical screening study. Then, we use this combined model in an example screening of 50,000 MOFs, showing how it can be used to identify MOFs that are simultaneously high performing for multiple separations. Finally, we use the criteria outlined in Chapter 6 to identify several synthetically likely MOFs that are predicted to be high-performing for two important gas separations.

CHAPTER 2

AN OVERVIEW OF COMPUTATIONAL METHODS

2.1 Molecular Simulation

As stated in the introduction, molecular simulation can be (mostly) divided into two main categories: quantum mechanical (QM) and classical simulations. QM simulations seek solutions to the time-independent Schrödinger equation under certain assumptions that (ideally) yield sufficiently accurate results. The set of assumptions used defines a QM simulation "type". For example, density functional theory (DFT) assumes that a system can be modeled using electron density (a function of three variables), rather than a $3N$ dimensional wavefunction (N is the number of electrons in the system). In classical simulations, atoms are the smallest objects considered. The energies used in a classical simulation are obtained from a forcefield, which is a set of functions that essentially takes as input the coordinates and velocities of the atoms in the system and outputs the system energy. This type of computation requires less computational effort per atom than a QM simulation, which allows for the consideration of larger and/or more systems. With modern computers, systems of hundreds of atoms are routinely used in QM simulations, however, systems of thousands, tens of thousands, or even millions of atoms can be considered in classical simulations. Given the focus on large-scale screening, which requires the consideration of thousands of (relatively) large systems, the primary methods used in this thesis fall into the category of classical simulation, so four of the six sections herein (2.1.2 to 2.1.5) deal with classical simulation methods. In the background section, some of the theory that is relevant to understanding molecular simulation is introduced, this theory applies to both quantum and classical systems, but is only practically applied to classical systems. Theory and background that is specifically relevant only to QM (in the context of DFT) simulations is presented in 2.1.6.

2.1.1 Background

The purpose of a molecular simulation is to measure the properties of a many-particle system, regardless of whether those particles are electrons, atoms, or groups of atoms. The calculation of these properties is accomplished using the framework of statistical mechanics. In statistical mechanics, many thermodynamic properties are derived from the partition function, a statistical description of a system in equilibrium. A heuristic derivation of the partition function for a system of N particles and constant volume in thermal equilibrium with a large heat bath (i.e. a system at a constant temperature) will help with understanding the subsequent sections. For this derivation, following Frenkel and Smit [164], it is helpful to consider a

system that is a collection of discrete states with energy E_i, that is, a quantum system. The partition function of a quantum system can then be related to that of a classical system (i.e. a system where energy is a continuous function of the particle positions and momenta) via a straightforward analogy.

Imagine a system A in thermal equilibrium with a large heat bath. The total energy of A and the heat bath is $U = E_B + E_A$, with E_A being the energy of A and E_B being the energy of the bath. Suppose that A can be prepared in an energy eigenstate with energy E_i, and let the degeneracy of E_i be $\Omega_A(E_i)$ (the degeneracy of an energy level is the number of states with that energy level). Therefore, the probability of the system being in stat i is

$$P_i = \frac{\Omega_A(E_i)}{\sum_j \Omega_A(E_j)} \tag{2.1}$$

which is simply the number of states with energy E_i divided by the total number of states. We would like to know the probabilities P_i of each state available to the system since these probabilities can be used to calculate the expectation values (or averages) of various thermodynamic variables. For example, the average energy of the system can be computed as $\langle E \rangle = \sum_i E_i P_i$, which is just the definition of the expectation value of a random variable (in the case E) with some discrete probability distribution yielding the P_i. How can we compute P_i? According to Equation 2.1 we require $\Omega_B(E_i)$, which cannot, in general, be calculated directly. However, if we note that $\Omega_A(E_i) = \Omega_B(U - E_i)$ where $\Omega_B(U - E_i)$ is the degeneracy of the heat bath when A is in state i, and note that calculating $\log \Omega_B(U - E_i)$ is equivalent to calculating $\Omega_B(U - E_i)$, we can derive a convenient expression for P_i using a Taylor expansion of $\log \Omega_B(U - E_i)$ around $E_i = 0$ [164], thus:

$$\log \Omega_B(U - E_i) = \log \Omega_B(U) - E_i \frac{\partial \log \Omega_B(E)}{\partial E} + \mathcal{O}(E^{-1}) \tag{2.2}$$

The definitions of entropy and temperature together imply that $\partial \log \Omega_B(E)/\partial E = \beta = 1/(k_B T)$, where k_B is Boltzmann's constant and T is temperature. Specifically, for a system with a constant number of particles N, constant volume V, and constant total energy E, entropy $S = k_B \log \Omega$. Moreover, temperature can be defined as $1/T = (\partial S/\partial E)_{V,N}$, therefore:

$$k_B \left(\frac{\partial \log \Omega}{\partial E} \right)_{V,N} = \left(\frac{\partial S}{\partial E} \right)_{V,N} = \frac{1}{T} \tag{2.3}$$

Finally, combining 2.2 and 2.3 we obtain:

$$\Omega_B(E_i) = e^{-\beta E_i} + \mathcal{O}(E^{-1}) \tag{2.4}$$

implying that

$$P_i = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} \tag{2.5}$$

Equation 2.5 is the Boltzmann distribution for a system at temperature T. The partition function Q is defined from Equation 2.5 as $\sum_j e^{-\beta E_j}$. As stated previously, various thermodynamic quantities can be calculated from Q. Recall that Equation 2.5 tells us the probability of finding a constant T system in an energy eigenstate with energy E_i. Note that $Q = \sum_j e^{-\beta E_j}$ only applies to systems with a constant number of particles N, constant volume V, and at a constant temperature T. Other partition functions can be derived in cases where different system properties are held constant; the properties held constant define an "ensemble". For example, (*i*) the constant N, V, and T ensemble we have been discussing is called the canonical or NVT ensemble, (*ii*) the constant N, pressure P, and T ensemble is called the isothermal/isobaric or NPT ensemble, and (*iii*) other ensembles can be used depending on the purpose of the simulation (such as the grand canonical ensemble discussed in the next section). Regardless of the ensemble, we can use the probabilities calculated using a partition function to calculate the thermal average of an observable A, for example:

$$\langle A \rangle = \frac{\sum_i e^{-\beta E_i} A_i}{\sum_j e^{-\beta E_j}} \tag{2.6}$$

where A_i is the expected value of A in quantum state i, or more precisely $A_i = \langle i|A|i\rangle$. We can calculate $\langle A \rangle$ using Equation 2.6 only if we have a solution to the Schrödinger equation for our system, allowing us to calculate the A_i. Unfortunately, for any system of practical size this is not possible. Even if we could solve the Schrödinger equation for an N particle system, the number of states that must be averaged in Equation 2.6 would be too large to hold in any computer's memory. Consider, for example, one mole of Ar at room temperature and atmospheric pressure. The entropy of this amount of Ar at these conditions is 154.846 J/K [165], which means the *logarithm* of the degeneracy of this Ar system is on the order of 10^{25}. Conveniently, however, we can derive an expression analogous to for a classical system, which is of practical use. Frenkel and Smit [164] show that this expression is

$$\langle A \rangle = \frac{\int d\mathbf{p}^N d\mathbf{r}^N \exp\left[-\beta\left(\sum_i p_i^2/(2m_i) + V(\mathbf{r}^N)\right)\right] A(\mathbf{p}^N\mathbf{r}^N)}{\int d\mathbf{p}^N d\mathbf{r}^N \exp\left[-\beta\left(\sum_i p_i^2/(2m_i) + V(\mathbf{r}^N)\right)\right]} \tag{2.7}$$

and

$$Q_{\text{classical}} = C \int d\mathbf{p}^N d\mathbf{r}^N \exp\left[-\beta\left(\sum_i \frac{p_i^2}{2m_i} + V(\mathbf{r}^N)\right)\right] \tag{2.8}$$

where C is a constant ensuring that the quantum partition function approaches the classical partition function in the classical limit (i.e. when $\hbar \to 0$, $\hbar$ being the reduced Planck's constant), $\mathbf{p}^N$ and $\mathbf{r}^N$ are the momenta and positions of all N particles, respectively, and m_i is the mass of the ith particle. The $1/N!$ term arises because identical particles cannot be distinguished [164]. Note that the operator A has been replaced by a function $A(\mathbf{p}^N\mathbf{r}^N)$ of the particle positions and momenta. The exponential term in

Equations 2.7 and 2.8 can more conveniently be expressed as

$$\exp\left[-\beta\left(\sum_i \frac{p_i^2}{2m_i} + V(\mathbf{r}^N)\right)\right] = \exp\big(-\beta H(\mathbf{p}^N, \mathbf{r}^N)\big)$$

where H is the system Hamiltonian, a function of the particle positions and momenta that outputs the total energy $U = K + V$ with $K = \sum_i p_i^2/(2m_i)$ being the kinetic energy and $V = V(\mathbf{r}^N)$ being the potential energy. The space of the N particle coordinates and momenta is called the phase space. The phase space for a 3-dimensional system has $6N$ dimensions (each particle has 3 coordinates and 3 momemta). Equation 2.7 can be evaluated for systems of a practically useful size, but only using certain techniques. We could attempt to evaluate the integrals in Equation 2.7 using numerical quadrature, however, this is not feasible even for small systems. If x points are taken along each dimension of phase space, the total number of points that must be evaluated in x^{6N}, which is intractably large even for small x and N. One can imagine that a similar problem would if one tries to sample the integrand using random points in phase space. However, we know that certain points in phase space are more probable than others, and that the probability of each point can be calculated using Equation 2.5. Therefore, we can restrict our attention to the regions of phase space that are probable, this is called importance sampling.
Unfortunately, we cannot use Equation 2.5 directly as it requires knowledge of the partition function Q, which can only be calculated by enumerating all states of the system of interest (which is not possible, see above). Yet, if we know the relative probabilities of each point in phase space, we can consider only the volumes of phase space that are *more* probable than others. Conveniently, we can calculate the relative probabilities of two states i and j as:

$$P_i/P_j = e^{-\beta E_i}/e^{-\beta E_j} \tag{2.9}$$

Equation 2.9 is the basis for the Metropolis Monte Carlo (MC) algorithm (as it is used in molecular simulation) [166]. In this algorithm, a system is prepared in state i, a trial "move" is then attempted which shifts the system to state j (this move could be particle translation, molecular rotation, etc.), if j is determined to be more probable than i using Equation 2.9 (i.e. if j has lower energy than i) then the move is accepted, otherwise (if $P_j < P_i$) the move is accepted with probability P_i/P_j. This has the effect of moving the system towards more likely states. If we generate M points in phase space in this manner then $\langle A \rangle$ can be calculated as

$$\langle A \rangle \approx \frac{1}{M}\sum_i^M A(\mathbf{r}_i^N) \tag{2.10}$$

The particle momenta are not included in Equation 2.10 since the Metropolis MC algorithm is used to model time-independent phenomena. Molecular dynamics (MD) simulations are usually used to model

time-dependent phenomena (e.g. diffusion). Equilibrium MD simulations can also be considered as importance sampling algorithms. This is because, in MD, the system of interest is allowed to evolve through time by calculating the forces exerted on each particle, and these forces cannot be high enough to force the system into a high-energy (improbable) state in an equilibrium MD simulation. Clearly, the accuracy of $\langle A \rangle$ calculated via Equation 2.10 depends on the number of points M generated. In practice, increasing values of M can be tested to see at what value $\langle A \rangle$ stops changing.

To conclude this section, two figures are shown to briefly illustrate some of the concepts outlined earlier. Figure 2.1 shows an example of the Metropolis Monte Carlo method compared with random sampling for the Rosenbrock function:

$$z = (1 - x)^2 + 100(y - x^2)^2 \tag{2.11}$$

which has a minimum at the point (1,1) and a shallow "basin" region surrounding this minimum. If we imagine the energy of some 2-dimensional system to be given by the Rosenbrock function (i.e. take z in Equation 2.11 to be energy) then the probability of each (x,y) state is given by $e^{-\beta z}$. Note that for the random sampling, the average probability of the points sampled was 0.0086 versus 0.5951 for the Metropolis sampling. Additionally, note that the system is limited to the shallow energy "basin" around (1,1) when using Metropolis sampling. Given the temperature dependence of the Boltzmann distribution, we expect high energy states to become more accessible at higher temperatures. An example of this is shown by Figure 2.2, where the results of Metropolis sampling the Rosenbrock function at four increasing temperatures are summarized. Clearly, at higher temperatures, more points of phase space are available to our example Rosenbrock function system.

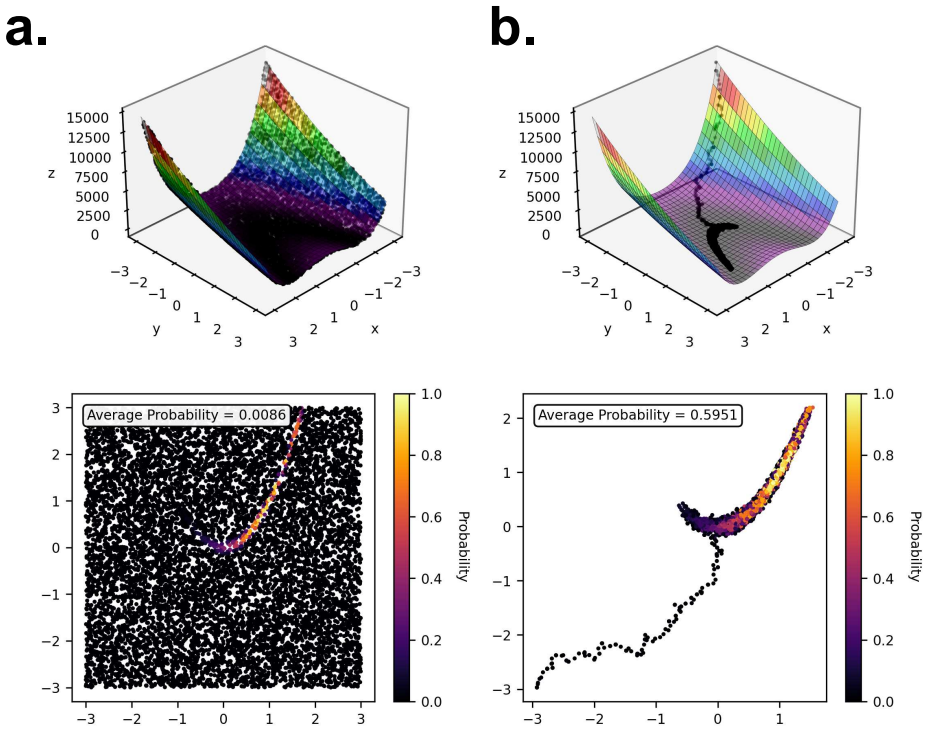

Figure 2.1 The Rosenbrock function function sampled using 10,000 (a) random (x,y) points and (b) (x,y) points generated using the Metropolis importance sampling algorithm. The average probability (calculated from the Boltzmann distribution) of the points sampled using each method is shown. Both calculations were done with $T = 1.0$

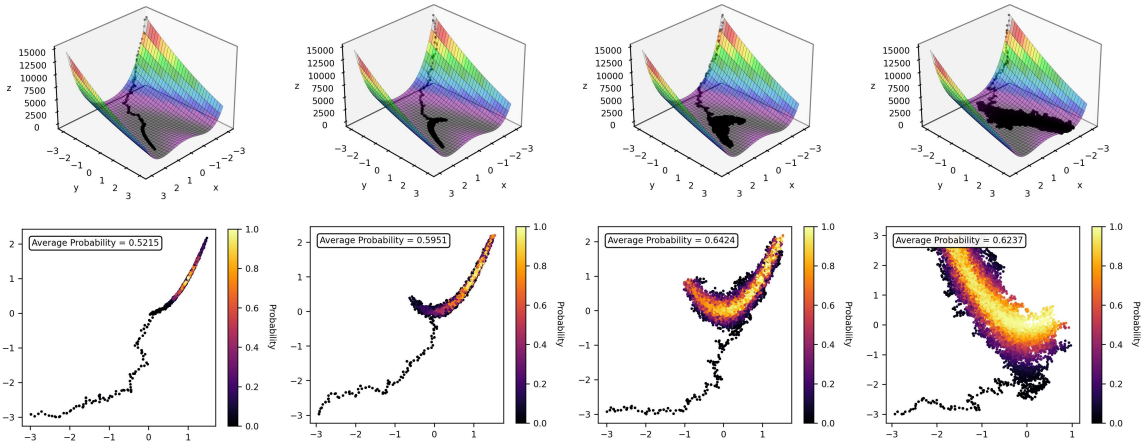

Figure 2.2 Metropolis sampling of the Rosenbrock function for (from left to right) $T = 0.1$, 1, 10, and 100.

2.1.2 Force Fields

Of paramount importance to classical simulations is how system energies are calculated. In a Monte Carlo (MC) simulation, the total energy of each trial system configuration determines how likely the system is to evolve to that configuration, and in a molecular dynamics (MD) simulation, the forces experienced by individual particles are calculated from energies. Kinetic energy is straightforward to calculate. Per-particle kinetic energy depends only on the mass and momentum of the particle, and the total kinetic energy is just the sum of the kinetic energies of each particle in a system. Potential energy, on the other hand, is some function $V(\mathbf{r}^N)$ of the particle coordinates. $V(\mathbf{r}^N)$ can be complex depending on the system being modeled. Forcefields are sets of functions that, when combined, yield $V(\mathbf{r}^N)$. Systems of individual particles (i.e. particles that are not part of a molecule) can be modeled considering only non-bonded interactions. Common functional forms used to describe non-bonded interactions are the Lennard-Jones (LJ) and Coulomb potentials. The LJ potential is used to calculate the energy of van der Waals (dispersion) interactions between particle i and particle j, thus:

$$E_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2.12}$$

where r_{ij} is the particle-particle distance, ϵ is a variable parameter describing the strength of the interaction and σ is a variable parameter dictating at what distance the interaction becomes repulsive. The Coulomb potential is used to calculate the charge-charge interaction energy between pairs of particles, thus:

$$E_C = C \frac{q_i q_j}{r_{ij}} \tag{2.13}$$

where C is a constant, q_i is the charge of particle i, q_j is the charge of particle j, and r_ij is the particle-particle distance. The total LJ and electrostatic energies of the system of interest can then be computed by summing the interaction energies of each ij pair. Typically, LJ interactions, which vanish as $r_{ij} \to \infty$ are not considered past a certain cutoff. For molecular simulations, this cutoff is in the 8-14 Å range (12.5 Å is a popular choice). Charge interactions can also be neglected past a certain cutoff, but, because these interactions decrease more slowly with r_{ij} than LJ interactions, a more precise solution is to use long-range electrostatic solvers, such as Ewald summation, which can efficiently consider electrostatic interactions at distances past a certain cutoff [167].

Intermolecular interactions require additional functions used to calculate the energies of bond stretching, bond angle bending, and dihedral rotation. In addition to dihedral angles, so-called "improper" terms are used to maintain certain molecular motifs in the correct equilibrium configuration. Impropers are often used to keep aromatic rings and trigonal planar motifs flat and maintain tetrahedral geometry. The functional forms of these intramolecular potentials depend on the forcefield being used. Bond stretching is

usually modeled with a harmonic potential:

$$E_B = K_B(r_B - r_0)^2 \tag{2.14}$$

where E_B is the bond energy, K_B is a force constant dictating the steepness of the harmonic well, and r_B is the bond length, and r_0 is the equilibrium bond length. A large force constant K_B corresponds to a stiff bond, and vice versa. The equilibrium bond length corresponds to the lowest (zero) energy configuration. For example, a C-C single bond would have an equilibrium bond length of 1.54 Å. The angles between two adjacent bonds are usually modeled similarly to bond stretching:

$$E_A = K_A(\theta - \theta_0)^2 \tag{2.15}$$

The terms in this equation are directly analogous to those in Equation 2.14, but with θ being the angle between two adjacent bonds, rather than a bond length. The energies associated with dihedral rotation are usually modeled with a periodic function, since there are often multiple energy minima for a dihedral angle. A representative function for dihedral angle energy is

$$E_D = K_D[1 - d\cos(n\phi)] \tag{2.16}$$

K_D is analogous to the bond/angle force constants, d and n determine the location and number of the energy minima, respectively (d is restricted to 1 or -1 in this case), and ϕ is the dihedral angle. A commonly used improper term is the energy of the out-of-plane bending of a group of four trigonal planar atoms i,j,k, and l. Atom i is usually taken to be central to the triangle, that is, i is bound to j, k, and l, but j, k, and l are not bound to eachother. Such a motif is encountered in a benzene ring, for example, with i being a carbon atom, j and k also being carbon, and l being H. The atoms in such a motif form a plane, any deviation from planarity of this group can be measured by the angle ω between the bond connecting an atom l to i and the plane formed by atoms i, j, and k. Then, a positive energy can be assigned to any values of ω not equal to zero. For example

$$E_I = K_I[1 - \cos(\phi)] \tag{2.17}$$

has a minimum at 0 radians (degrees), 2π radians (360 degrees), etc. Note that Equations 2.12 to 2.17 depend only on the system particle positions, as any bond angle, dihedral angle, or improper angle can all be calculated from atom positions. The total potential energy of a system containing both inter and intramolecular interactions can then be calculated as

$$V(\mathbf{r}^N) = E_{LJ} + E_C + E_B + E_A + E_D + E_I \tag{2.18}$$

with each term being implicitly summed over all corresponding interactions in the system. That is, E_{LJ} is the sum over all pairwise LJ interactions, E_B is the sum of the energies of all bonds, etc. The functional

form of each potential and the parameters associated with of these functions are the forcefield used to describe the system. For example, the parameters for the functions shown here are the LJ ϵ and σ for each pair interaction, the particle charges, the bond/angle force constants (K_B and K_A) and equilibrium values (r_0 and θ_0), the dihedral force constants K_D, periodicity n, and energy minima locations d, and the improper force constants K_I.

2.1.3 Grand Canonical Monte Carlo

Grand canonical Monte Carlo (GCMC) simulations are used when the number of atoms or molecules in a system of interest is required to change throughout the course of the simulation. More specifically, GCMC simulations correspond to the "grand canonical" (or μVT) ensemble, where the chemical potential μ, volume V, and temperature T are held constant. The simulation box (containing the system of interest) is kept in equilibrium with an infinite gas reservoir at constant μ, and particles are inserted into (from this reservoir) or deleted from the simulation box according to the Metropolis algorithm. That is, a particle is inserted/deleted into/from the simulation box if that insertion/deletion lowers the total system (the simulation box and gas reservoir) energy, otherwise it is inserted with some probability. These insertions and deletions are called Monte Carlo "moves". Other moves considered in GCMC simulations are translation and rotation of particles or molecules already within the simulation box. A common practice is to specify the pressure and temperature of the gas reservoir are specified during a GCMC simulation, the pressure is then converted to fugacity using an equation of state (the Peng-Robinson equation of state is a popular choice), and then the chemical potential is calculated from the fugacity and temperature.

From the description above, one can imagine how adsorption isotherms in porous crystals are calculated using GCMC simulations. The simulated system is the crystal unit cell (or a supercell of the unit cell) in equilibrium with a gas reservoir of the adsorbate(s) of interest, which defines a fixed volume. Each isotherm point in an isotherm corresponds to the average number of adsorbates calculated from a GCMC simulation at the corresponding pressure (chemical potential), in which adsorbates are inserted into and deleted from the crystal pores (see Figure 2.3). In practice, each GCMC simulation is conducted over a set of n cycles, where in each cycle a certain number of moves m is attempted; m is usually the number of adsorbates present in the system with some minimum value (i.e. each cycle may consist of one attempted move per adsorbate). The total number of attempted moves is, then, $m \times n$. The number of cycles and the ratio of move types are tunable parameters of the simulation. Typically, the system is allowed to equilibrate over a set number of cycles, during which time no averages are calculated. After equilibration, production cycles are run in which averages are calculated. This is to avoid including states outside of equilibrium in averaging. For example, most GCMC simulations of adsorption are initialized with no

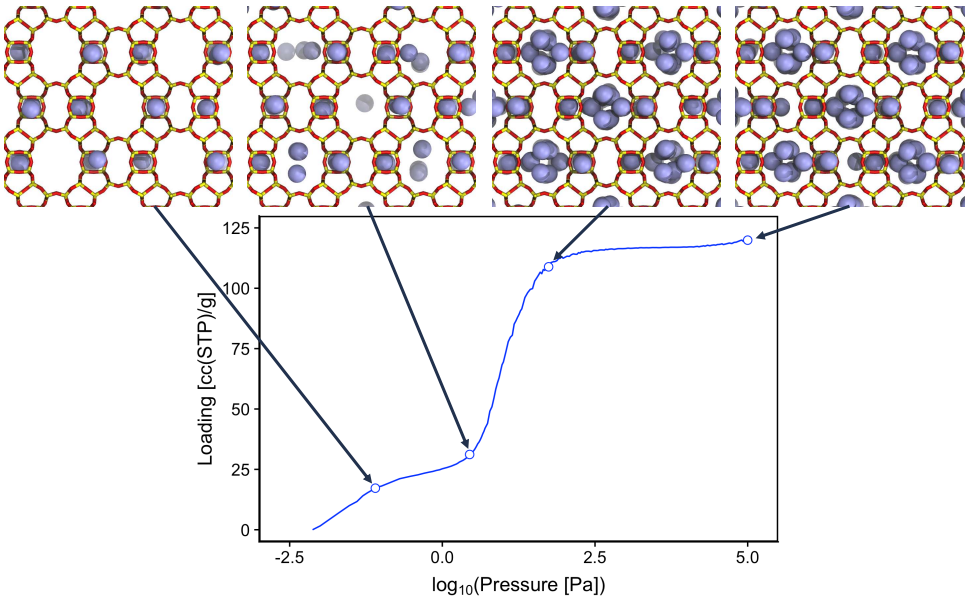

Figure 2.3 An example isotherm of Ar in the silicalite framework MOR [28] calculated using a series of GCMC simulations. Each point corresponds to a GCMC simulation at the corresponding pressure and isotherm temperature. Four snapshots of the simulation box showing the equilibrium loading of Ar at the corresponding pressure are shown.

adsorbates in the adsorbent, meaning that the loading will increase from zero to near the equilibrium value over some number of cycles. The loadings that are still increasing should not be included in averaging. Multicompontent isotherms can easily be computed using GCMC simulations, by considering a multicomponent gas reservoir of fixed composition, where each component has a separate (but still constant) chemical potential. When multiple adsorbates are considered, an additional MC move is included, which swaps the identity of two adsorbates of different types.

2.1.4 Equilibrium Molecular Dynamics

Molecular dynamics (MD) simulations are similar to real experiments. We consider a model system whose properties we want to measure, prepare that system, and allow it to equilibrate, and then observe its behavior (measure its properties) over some time. The output of any MD simulation is a trajectory of the system through time, that is, the particle positions and momenta over a series of timesteps. Any observable quantity that can be expressed as a function of these positions and momenta can then be calculated from this trajectory. This calculation is usually a time average of the instantaneous values of the observable. When calculating averages in MD we assume that enough states are visited over time that the time-average converges to the ensemble average (an average over all possible states of a system, which can be estimated from a MC simulation). This assumption is called the ergodic hypothesis. The ergodic

hypothesis does not apply in all cases. For example, systems with metastable states separated by energy barriers that cannot be overcome by thermal fluctuations are unlikely to be ergodic. In an MC simulation, the system can "hop" between energy minima (skipping over energy barriers), however, in MD the system is likely to remain in a single minima (unless some bias is introduced). Nevertheless, the ergodic hypothesis is plausible for many systems [164], and the prevailing assumption in the literature is that most systems are (sufficiently) ergodic for time-averaging to produce accurate results unless a phenomenon that relies on crossing relatively large energy barriers is being simulated.

MD, in essence, is an algorithm that integrates Newton's equations of motion over a series of discrete (but small) timesteps. The Newtonian equation of motion is Newton's second law

$$F_i = \frac{dp_i}{dt} \tag{2.19}$$

where F_i the net force acting on particle i and p_i is the momentum of a particle i. For an N particle system the net force acting on particle i is, then

$$F_i = \frac{dp_i}{dt} = \sum_{i \neq j}^{N} F_{ij} + F_E \tag{2.20}$$

where F_{ij} is the force applied to i by particle j and F_E is an external force (which is often zero). Thus, the position and momentum of particle i as functions of time can be obtained by integrating Equation 2.20. Other equations of motion that allow for more efficient integration, which are, in principle, equivalent to the Newtonian equations of motion, are often used in practice. The basic steps of an MD algorithm are as follows:

1. Initialize the system of N particles
2. Over a series of timesteps (Δt) starting at $t = 0$:
    2.1. Compute the forces on all particles
    2.2. Integrate Equation 2.20 for each particle
    2.3. Let time progress ($t = t + \Delta t$)
    2.4. Record data (i.e. update averages, etc.)
3. End simulation (compile any recorded data, etc.)

In the majority of cases, the most computationally intensive step is calculating the forces. This requires evaluation of the position derivatives of the forcefield equations outlined above for each particle in the simulation. Therefore, efficient distance calculation and function evaluation are integral to a practically

useful MD program. While efficient force calculations are key for the speed of an MD program, the algorithm used to integrate the equations of motion is key to obtaining reliable results [164]. The algorithm should be reasonably fast, accurate for reasonably long timesteps (allowing for fewer force evaluations), and it should (perhaps most importantly) conserve energy over short and long timescales. Based on these requirements, the Verlet algorithm is a good choice in most situations. The starting point of a derivation of the Verlet algorithm are two Taylor expansions of a particle coordinate around time t:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{F(t)\Delta t^2}{2m} + \frac{d^3r}{dt^3}\frac{\Delta t^3}{3!} + \mathcal{O}(\Delta t^4)$$

and

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{F(t)\Delta t^2}{2m} - \frac{d^3r}{dt^3}\frac{\Delta t^3}{3!} + \mathcal{O}(\Delta t^4)$$

Summing these two expansions and neglecting $\mathcal{O}(\Delta t^4)$ terms we obtain an expression for the particle position at time $t + \Delta t$:

$$r(t+\Delta t) \approx 2r(t) - r(t-\Delta t) + \frac{F(t)\Delta t^2}{m} \tag{2.21}$$

Once Equation 2.21 is applied to all particles in the system being simulated, the positions for $t - \Delta t$ can be discarded and the "current" particle positions can be updated to those calculated at $t + \Delta t$. The error in the Verlet algorithm is $\mathcal{O}(\Delta t^4)$, so a smaller timestep (Δt) results in less error. Usually, one finds a balance between efficiency (a larger Δt that requires fewer force computations for a simulation of total time t_{total}), and error (making sure that energy is conserved over t_{total}).

2.1.5 Thermodynamic Integration

Free energy cannot be measured directly from a simulation, this is because thermal quantities such as free energy or entropy, depend on the volume available to a system in phase space. In (perhaps) more clear terms, there is no expression for free energy that is in the form of an average over a collection of points in phase space (i.e. the type of average that can be calculated over time or over a finite series of Monte Carlo moves). However, free energy differences can be computed using thermodynamic integration. For example, the equation

$$\left(\frac{\partial F}{\partial V}\right)_{N,T} = -P$$

where F is free energy, can be used to calculate the change in free energy due to expanding or compressing a system. The volume of the system is changed over a series of equilibrium MD simulations from V_1 to V_2, allowing the derivative $(\partial F/\partial V)_{N,T}$ to be calculated and then integrated to obtain the free energy difference between the system at V_1 and the system at V_2. Experimentally, thermodynamic integration can

only be performed along physical pathways, like pressure or temperature changes. However, in simulations we can integrate along unphysical pathways: any parameter in a potential energy function can be used as a variable to construct an integration path [164]. This is accomplished by switching "on" or "off" forcefield interaction parameters according to a coupling parameter λ. The parameter λ is constructed such that at $\lambda = 0$ the potential energy is U_1 and at $\lambda = 1$ the potential energy is U_2 and the potential energy any intermediate values of λ is

$$U(\lambda) = (1 - \lambda)U_1 + \lambda U_2 \tag{2.22}$$

Then by progressing λ from 0 (system 1) to 1 (system 2) the free energy difference between system 1 and system 2 (the work required to switch from system 1 to system 2) can be calculated as

$$F_2 - F_1 = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{2.23}$$

where $\langle \partial U(\lambda) \partial \lambda \rangle_\lambda$ is the ensemble average (or time average for an MD simulation, these are assumed to be equivalent for an ergodic simulation) of the derivative of $U(\lambda)$ with respect to λ [164]. In practice, a discrete number of λ values between 0 and 1 are chosen and an equilibrium simulation is run for each of these values, during which the integrand of Equation 2.23 is calculated. The free energy difference can then be obtained via numerical integration of the discrete values of $\langle \partial U(\lambda) \partial \lambda \rangle_\lambda$. This procedure guarantees that the switching path is reversible, since each intermediate point is at equilibrium. Alternatively, system 1 can be continuously switched to system 2 during a single non-equilibrium simulation, making the switching irreversible. However, if the switching is slow enough, the heat released will be negligible and the error in the calculated free energy difference will also be negligible [168]. In addition, by switching forward and backward the error due to heat release can be cancelled (see Chapter 6 for more details) [169].

2.1.6 Density Functional Theory

Density functional theory is a class of electronic structure calculation method (i.e. a quantum mechanical method), that can be used to accurately predict a variety of material and molecular properties. Similar to classical simulations, the main goal of a DFT simulation can be considered the calculation of energies. The accurate calculation of energies then allows other properties to be derived (e.g. geometries, interaction and reaction energies, and vibrational spectra). The total electronic energy of an atom, ion, or molecule is

$$E_{\text{total}} = E_K + E_V + E_J + E_X + E_C \tag{2.24}$$

where E_K is kinetic energy, E_V is the potential energy due to electron-nucleus Coulomb attraction, E_J is the potential energy of electron-electron Coulomb repulsion, E_X is the the total electron exchange energy,

and E_C is is the total electron correlation energy. The E_X term corrects for the correlated motion of electrons of the same spin and removes the unphysical self-repulsion terms used to calculate E_J and the E_C term corrects for the weakly correlated motions of electrons with opposite spins [92]. The computation of the exchange and correlation terms is the main difficulty in calculating the total electronic energy.

In contrast to other QM methods in which system properties are derived from computed wavefunctions (approximate solutions to the time-independent Schrödinger equation), DFT uses electron density to calculate the total electron energy and other properties. Hohenberg and Kohn demonstrated that the total electronic energy of a molecule (or atom or ion) is a functional (a functional is a function of a function) of the electron density and that the ground state electron density when input to this functional yields the ground state energy of the molecule. Stated mathematically:

$$E_0 = F[n_0(r)] \tag{2.25}$$

where E_0 is the ground state energy, $n_0(r)$ is the ground state electron density (a function of position r), and F is the functional relating the two [170]. Starting from the work of Hohenberg and Kohn [170], Kohn and Sham [171] developed an equation allowing for the calculation of total electronic energy from the electron density $n(r)$ (including the exchange and correlation energies). This equation is the Schrödinger equation for an unphysical system of noninteracting electrons, but its solution produces the same electron density as that of the corresponding system with interacting electrons [92]. In a DFT simulation, the Kohn-Sham equation is solved. The solution yields the ground state electronic energy and ground state electron density of the system of interest. Essentially, an initial electron density is guessed, allowing the total electronic energy to be calculated, the total electronic energy is then minimized as a function of electron density.

Electron density can be represented with a variety of functions. Planewaves are used for calculations on infinitely periodic models of crystalline structures. For calculations on individual atoms or molecules, Gaussian functions centered on atom coordinates are typically used to represent orbitals $\psi(r)$, thus:

$$\psi(r) = \sum_{i=1}^{N} C_i e^{-\alpha_i r^2} \tag{2.26}$$

where C_i and α_i are parameters that can be varied (in practice only the C_i are change) to obtain a solution to the Kohn-Sham equation. The electron density corresponding to the orbital ψ can then be calculated as $n(r) = \psi(r)^2$ [92]. The electron density of a molecule can, therefore, be calculated by finding Gaussian representations for all the orbitals in that system (by varying the C_i). The number of Gaussian functions to use per orbital and the α_i parameters form a basis set. Different basis sets (of greater or lesser complexity) can be used depending on the properties/system one seeks to model.

2.2 Machine Learning

2.2.1 Background

Machine learning (ML), for the purposes of this contribution, is a method that builds predictive statistical models based on the patterns and relationships present in a set of training data. The trained model can then make predictions on new data using what it "learned" from the training data. The ML algorithms discussed herein can be categorized as supervised or unsupervised. During supervised training, the model is given a set of observations i. For each i there is a set of input variables X_i (often the X_i are vectors of n variables) and a set of output (or response) variables Y_i. The training algorithm then seeks to reproduce the Y_i based only on the X_i. The Y_i can be categorical, meaning a model would be trained to categorize observations based on the X_i, producing a "classification" model. If the Y_i are numerical, then the result of training is a "regression" model. An enormous variety of supervised ML algorithms exist, most of which can be used either for classification or regression. Most algorithms work by minimizing the errors between the actual training set Y_i and the predicted Y_i as a function of some parameters. For example, linear regression can be regarded as a simplistic supervised machine learning algorithm. For some input x_i, y_i a model $f(x) = ax + b$ is fit by altering the parameters a and b so that the error (in most cases the sum of the squared residuals) $\sum_i (f(x_i) - y)^2$ is minimized. The resulting linear model can then be used to make predictions for new x values. Clearly, this model will only work for input/response relationships that are approximately linear. More advanced models, such as neural networks, can be trained to make accurate predictions with high-dimensional input data and complex response variables, but the working principles remain the same. Each neural network has a set of parameters that influence its predictions, and these parameters are tuned to minimize the error between the predicted and actual response variables. Unsupervised machine learning algorithms seek to learn patterns in "unlabeled" data or data without any output variables Y_i. For example, clustering analysis (a class of unsupervised machine learning tasks) seeks to group observations represented by X_i based on some similarity metric(s). The properties of the resulting clusters can then be analyzed for useful information and new observations can be categorized into the clusters obtained from a training set. In this section, the machine learning algorithms used in the subsequent chapters are described briefly and on a high level. The practical aspects of model training are provided in the individual chapter methods sections or the corresponding appendix.

2.2.2 Tree-based Methods

This section is reproduced with permission from: R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. Gómez-Gualdrón. "Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: from molecular simulation to machine learning" *Chem. Mater.* 30, 18, 6325–6337. Copyright 2018 American

Chemical Society.

Decision trees (DTs) are perhaps the most-well known supervised learning methods for classification, but they can also be used for quantitative prediction. DTs work by sequentially "splitting" observations into two or more groups at a time using the value of one of the input variables X_j as a splitting criterion. This procedure creates a "tree" with many "nodes". Typically, at each node, the observations are divided into two branches (binary trees). The first splitting occurs at the top (root) node, and splitting continues at each of the subsequent nodes except for the terminal (leaf) nodes. The DT algorithm attempts to find the optimal splitting criteria that create a tree where all observations grouped at the same terminal node have the same output value Y. Note that the nature of DT algorithm results in quantitative predictions of Y being quantized.

A natural improvement over the DT algorithm, the RF algorithm, essentially consists of an ensemble of individual DTs. For quantitative predictions, the RF predicts the output value Y based on the input values X_j as the ensemble average of the value Y predicted by each individual DT. For classification predictions, the RF predicts the value Y as the most common value predicted by each individual DT. A significant advantage of the RF algorithm over the individual DT algorithm is the reduced output sensitivity to small changes in input values [172]. Another advantage of the RF algorithm, particular to quantitative predictions, is that, unlike the individual DT, it does not produce "quantized" predictions of Y.

Decision trees are also frequently used in gradient boosting machines (GBMs). Gradient boosting is a technique applied to "boost" the prediction capabilities of "weak learner" algorithms (such as individual decision trees). Starting from a single weak learner, the model is iteratively adjusted by adding new learners, which are constructed in such a way so the total model descends the gradient of the model error on the training set. The gradient boosting algorithm is described in more detail in Appendix B. A shematic representation of a decision tree is shown in Figure 2.4.

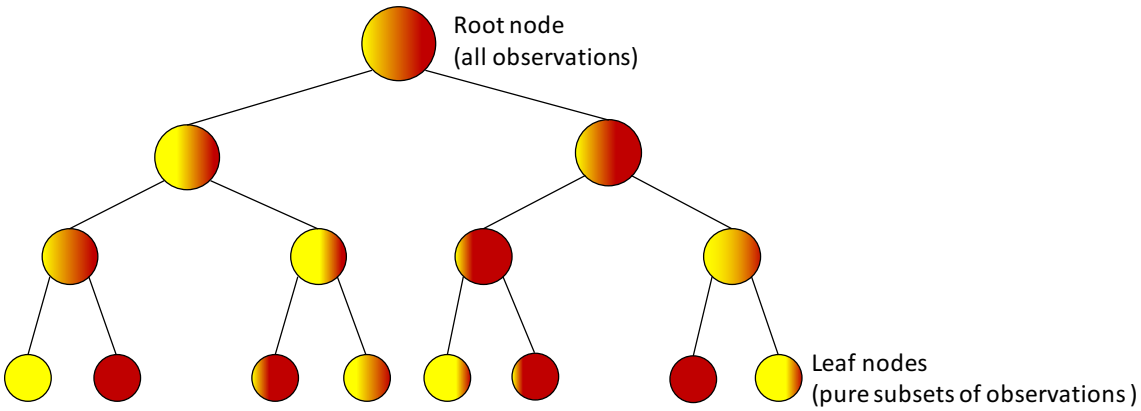

Figure 2.4 An example of a decision tree, the colors of each node qualitatively represent to homogeneity of the output values Y_i of each observation sent to that node. The leaf nodes should have the most homogeneous Y_i values. Perfect homogeneity is not often achieved in practice.

2.2.3 Multilayer Perceptrons

This section is reproduced with permission from: R. Anderson, A. Biong, D. Gómez-Gualdrón. "Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model" *J. Chem. Theory Comput* 16, 2, 1271–1283. Copyright 2020 American Chemical Society.

A multilayer perceptron (MLP) is a type of feedforward artificial neural network, which consists of at least one hidden layer (for a total of a minimum of three layers including the input and output layer). The goal of the MLP, as with any other machine learning algorithm, is to predict a response Y given a set X of variables. For instance, when the set X_i for observation i is fed into the input layer (one node for each variable in the set X_i), the network (once trained) will provide the corresponding response Yi as output. The output is typically a real number (when doing regression) or an integer (when doing classification).

Each hidden layer consists of N nodes, which are assigned an activation function (a non-linear function) and a weight. Each node in a hidden layer receives a weighted summation of the outputs of the preceding layer, and then applies the activation function to it. The weighted output of each node is included in a summation that is received by the nodes in the next layer. The process continues until the output layer is reached, which consists of one node with a linear (for regression) or step (for classification) activation function.

The training of an MLP consists of finding the weights assigned to the output of each node, which result in matching as closely as possible an array of known responses i given the corresponding variable sets X_i (these known responses and variables sets are referred to as the training set). Thus, during training, an MLP learns (unveils) the relationship between variable sets and their corresponding response from a

(hopefully) broad and diverse dataset. The trained MLP can then be applied to new variable sets X_i for which the response is unknown.

The optimal weights are found using a supervised learning algorithm known as backpropagation [173]. The weights of each node are initialized, usually randomly or according to some probability distribution, and the prediction is tested for a single input or a set of inputs. The error of this prediction is then calculated (i.e. the loss function) and used to calculate the partial derivate of the model error with respect to each weight in the network [173]. The calculation of these derivatives is backpropagation. The calculated derivatives (i.e. the gradient of error) can then be used in an optimization function which minimizes the loss function.

There are many so-called "hyperparameters" which influence the final weights, and therefore the output of an MLP model. Here, the hyperparameters tuned during the training of our models discussed in Chapters 5 and 7 are briefly described. These are (*i*) the architecture or topology of the net, which is the number of hidden layers and the number of nodes in each layer, (*ii*) the activation function applied to the incoming weights for each node (e.g. sigmoid, ReLU, or linear), (*iii*) how the initial weights are assigned, i.e. the initialization procedure, (*iv*) the optimizer used to minimize the loss function, (*v*) the number of times (epochs) all of the training data is passed through the backpropagation algorithm, and (*vi*) batch size which is the number of observations passed through the net before updating the weights (between one and the total number of observations). An schematic of an MLP is shown in Figure 2.5

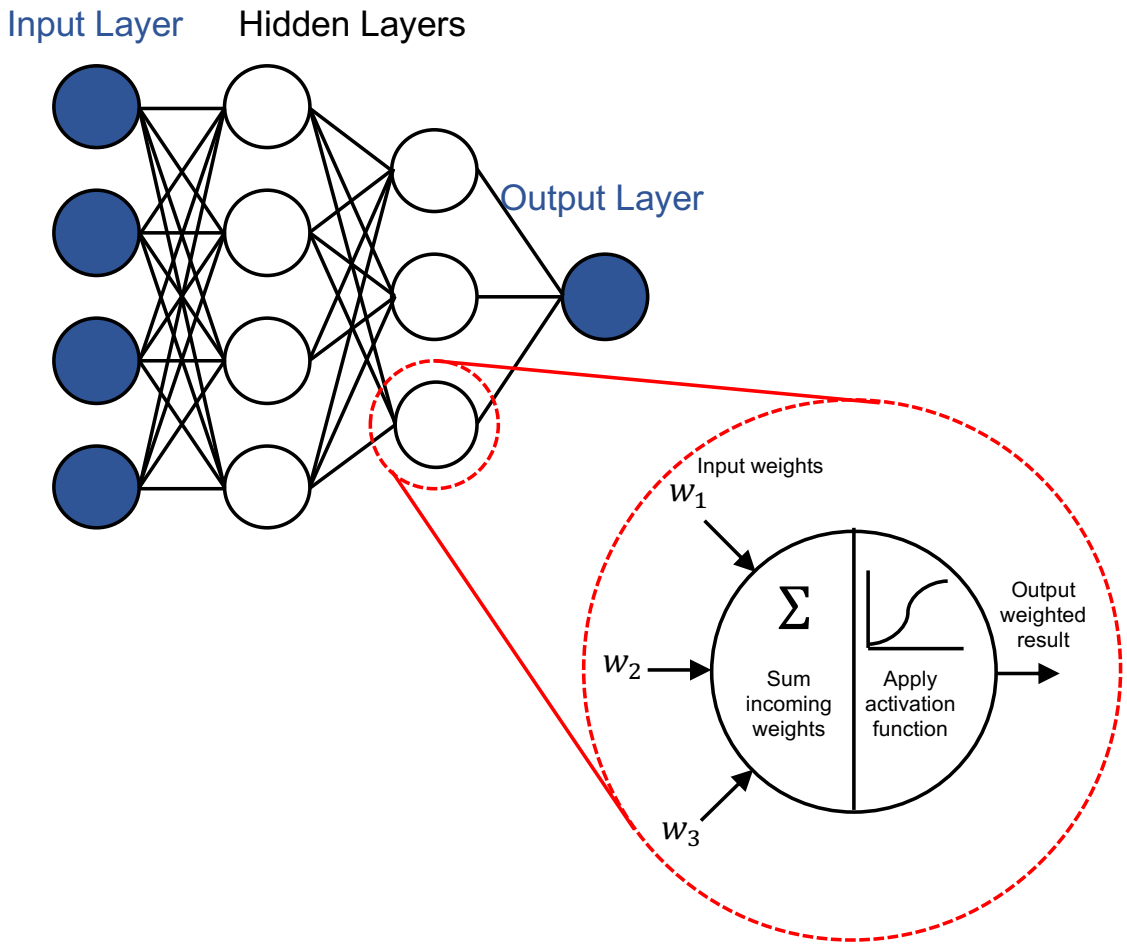

Figure 2.5 A schematic of a simple MLP, with two hidden layers (white nodes). The magnification shows an individual node, in which the input weights are combined (in this case as a sum), fed into the activation function (a sigmoid is shown), then output with the node weight.

2.2.4 Model-based Clustering

This section is reproduced from a paper submitted to the Journal of Chemical Physics entitled: R.Anderson, A. Biong, D. Gómez-Gualdrón. "Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model", reprinted with the permission of AIP Publishing.

Clustering algorithms are a class of unsupervised machine learning algorithms that seek to find groupings of a set of observations that correspond to some underlying structure. There are various clustering algorithms available, the most well-known of which is k-means clustering. A k-means clustering algorithm finds k clusters, each with a centroid in the observation space. Observations are assigned to clusters based on their distance to the cluster centroid (Euclidean distance is typically used). More specifically, (i) k initial means are randomly generated in the observation space, (ii) each observation is then assigned a group k corresponding to the nearest centroid), (iii) the initial means are updated to be the centroids of the assigned points, (iv) steps (ii) and (iii) are repeated until a convergence criteria is achieved. Given that the means are randomly initialized, running this k-means algorithm on the same data can give different results. Moreover, the user must specify k (the number of clusters) to the algorithm.

In many cases, a more robust alternative to k-means or another heuristic clustering algorithm (such as agglomerative hierarchical clustering) is model-based clustering (MBC). MBC has several advantages over "traditional" clustering methods, a few of these are: first, MBC makes soft assignments, meaning each observation is assigned a probability of being in each cluster. Second, the number of clusters is automatically calculated in most MBC implementations. Third, users can aim to fit different cluster shapes, depending on their expectation of the data structure. An MBC algorithm assumes that each observation in the data is sampled from k underlying probability distributions, a variety of underlying distributions can be used, depending on the task. Gaussian mixture modeling, a type of MBC, assumes that these distributions are all multivariate-normal. The model parameters, namely the means and covariance matrix of each multivariate normal distribution, are estimated using the Expectation Maximization (EM) algorithm, which is described in detail in ref. [174]. In most MBC implementations, the EM algorithm is initialized using the clusters obtained from agglomerative hierarchical clustering.

The covariance matrix of a multivariate-normal distribution determines the shape, volume, and orientation of the distribution. In GMM, different constraints can be applied to the distribution covariance matrices to ensure that each cluster has approximately the same volume (i.e., number of observations), shape (i.e., all the distributions may be constrained to be all spherical or elliptical), or orientation (i.e., the distributions are aligned along the same axis). Different combinations of these constraints result in different underlying models. In total, there are 14 different constraint sets (or models) that can be applied

in GMM, examples of which are shown by Srucca et al [175]. These models are assigned three-letter codes that describe whether volume, shape, and orientation are equal or variable. For example, the code "EEE" indicates equal volume, shape, and orientation, while the code "VVV" indicates variable volume, shape, and orientation. Which set of constraints to use is decided by applying all 14 models and identifying which one characterizes the data best. The Bayesian information criterion (BIC) is typically used to select the best model as well as the optimal number of clusters (a range of numbers of clusters is tested for each model) [176] although other model selection criteria could be used. The use of BIC for model selection in clustering is described in detail in ref. [176].

CHAPTER 3

AN ALGORITHM FOR THE TOPOLOGICALLY GUIDED CONSTRUCTION OF MOFS AND THE IMPORTANCE OF TOPOLOGICAL DIVERSITY

Modified from a paper published in CrystEngComm.[1]

Ryther Anderson[2], Diego Gómez-Gualdrón[2,3]

In this chapter, our algorithm (ToBaCCo) for the computational construction of MOFs according to topological blueprints is described in detail. We used ToBaCCo to build several MOF sets of various sizes for use in this thesis: from the 400 MOFs analyzed in Chapter 3 to the 50,000 MOF database screened in Chapter 7. Besides describing the ToBaCCo algorithm and its working principles, this chapter also includes an in-depth discussion of the importance of topological diversity in MOF databases used for screening. We demonstrate this importance using the properties of MOFs from seven isomorphic series. An isomorphic series is a set of MOFs that have the same node and linker but different topologies (i.e. are isomorphs). We show that textural, adsorption, and mechanical properties vary widely among isomorphs, despite their identical local chemistries. As an aside: at the time this paper was published, no database of MOFs built *in silico* contained more than 100 topologies. Our current database of 200,000 structures (all built using ToBaCCo) contains over 1,100 topologies.

3.1 Abstract

Effectively tuning the properties of porous crystals could lead to breakthroughs in areas such as molecular separations, chemical sensing, and catalysis. The most tunable kind of porous crystals are metal-organic frameworks (MOFs). Their tunability has resulted in MOFs being the focus of intensive research. However, MOF tunability also gives rise to an overwhelmingly large "MOF-space" (the conceptual space of all MOFs) that cannot be comprehensively explored experimentally. Because of this, high-throughput computational screening (HTCS) emerged as a tool to more rapidly explore MOF-space. However, the effectiveness of HTCS is tied to the ability of computational algorithms for automated crystal construction to access every point in MOF-space. To this end, here we introduce a crystallographic net rescaling algorithm to the topologically-based crystal constructor (ToBaCCo) code that allows for the automated construction of MOFs (or other porous crystals) of any topology by exploiting graph theory.

[1]Reproduced with the permission of: R. Anderson, D. Gómez-Gualdrón. "Increasing topological diversity during computational synthesis of porous crystals: how and why" *CrystEngComm*, 21, 1653-1665. Copyright 2019 Royal Society of Chemistry
[2]Chemical and Biological Engineering Department, Colorado School of Mines
[3]Corresponding author, dgomezgualdron@mines.edu

We demonstrate the capabilities of the new version of the code (ToBaCCo 3.0) by computationally "synthesizing" seven sets of isomorphic MOFs, where chemical composition remains constant within each set, but crystal topology varies across 19 previously inaccessible topologies. Furthermore, we illustrate the importance of computationally synthesizing all possible isomorphs for a given MOF composition by showing how strongly crystal topology influences simulated adsorption and mechanical properties. Specifically, we show that among crystal isomorphs i) methane deliverable capacity can vary by values as large as 150 cc(STP)/cc ii) adsorption selectivity for ethene/ethane mixtures can switch from ethene- to ethane-selective between MOF isomorphs, and iii) mechanical stability can be lost, and bulk modulus can exhibit significant variations, all solely due to a change in crystal topology.

3.2 Introduction

Porous crystals are materials whose porosity and regular architecture could be exploited in many technological areas. This is underscored by the emergence of metal-organic frameworks (MOFs) more than a decade ago as a material capturing the imagination of researchers around the world, who envisioned application of MOF crystals in areas as diverse as chemical separations [67–69, 177], energy storage [38, 40, 178, 179], sensing [60–62], medicine [63–66], environmental remediation [180, 181], and catalysis [56, 58, 59, 182]. Properties such as pores with organized arrays of functional sites [112, 183] and ultrahigh surface areas [52] are part of the appeal of MOFs, but perhaps no other property has driven research on these materials as much as their tunability [184–188]. This tunability is a result of the inherent modularity of MOFs, which form by interconnecting building blocks through coordination bonds. Therefore, new combinations of chemical properties can be built into (typically) three-dimensional networks simply by incorporating different building blocks [37, 189]. Additionally, the same combination of building blocks can potentially assemble into networks of the same chemical formula but with different underlying topologies, resulting in similar local chemistries in different spatial arrangements [35, 38, 112, 113, 124]. Given the thousands of possible organic ligands and thousands of topologies (as indicated by the Reticular Chemistry Structure Resource [79] (RCSR) database), the number of possible topology/building block combinations makes the so-called MOF space (i.e. the conceptual space that contains all plausible MOFs) overwhelmingly large. On the other hand, with so many possibilities, the existence of at least one "technology-disruptive" MOF per application seems plausible, the challenge being to find such MOFs.

A key question then is: to what extent has the MOF-space been explored? Analyses of the Cambridge Structural Database [190] (CSD) by different authors estimate the number of synthesized MOFs to be between 5,000 and 70,000 [70, 189, 191]. Either number is small relative to the vastness of the MOF-space. However, it is large enough to question why there is not yet a plethora of MOFs "disrupting" a broad array

of technologies—although it should be noted that commercial applications of MOFs are beginning to emerge [192]—and, consequently, closer examination is merited. Recent work by Smit and coworkers implies that the number of MOFs reported in the CSD may overestimate the actual number of synthesized MOFs as they found only ca. 70% of MOFs in a CSD sample to be unique [193]. Regardless of how many unique MOFs have been synthesized, however, our own analysis of the scientific literature shows that 15 well-known MOFs cover at least ca. 20% of all testing of these materials (Figure 3.1). Furthermore, for some applications, a single MOF can account for ca. 25% of all published experiments, which is the case for use of ZIF-8 in membrane separations.

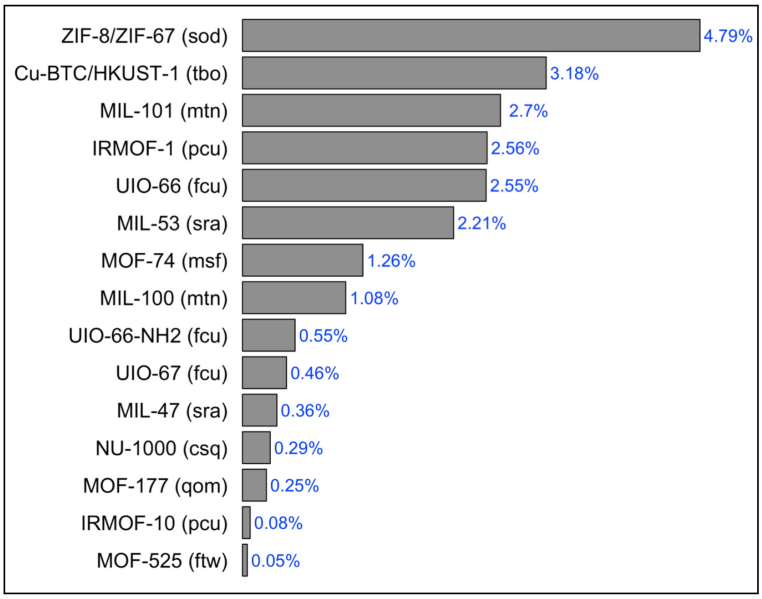

Figure 3.1 Percentage of publications corresponding to 15 of the most well-known MOFs out of 42,455 publications as determined from a literature search using ISI Web of Knowledge. Fifteen out of the thousands of MOFs reportedly synthesized to date cover 22.4% (9,489 publications) of all MOF literature. The topology of each MOF is shown in parenthesis.

So why has the actual exploration of the MOF-space been so limited? Work by Chung and coworkers [70] showed that experimentally reported surface areas for ca. 70% of MOFs in a CSD sample were significantly lower than the value for the "perfect" crystal, indicating that a dominant fraction of MOFs reported to date collapse upon porosity activation (solvent removal), which would explain why a large proportion of synthesized MOFs has not been tested. This lack of testing on so many MOFs means the tunability of these materials has been underexploited. Thus, synthesis and testing of new MOFs is needed to discover MOFs that can facilitate technological breakthroughs. However, building a synthesis protocol to obtain a single new MOF is notably time- and resource-consuming. Therefore, thoroughly canvassing the vast MOF-space via experimental synthesis and testing is unrealistic. To overcome this obstacle,

molecular simulation can be used as a "exploratory probe" that evaluates numerous plausible MOFs and identifies the most promising prospects for an application of choice [53, 163]. Allowing experimental efforts to be devoted solely to the synthesis and testing of these computationally identified structures. This idea was clearly illustrated in pioneering work by Wilmer and coworkers [37], where automated "computational synthesis" emerged as a critical component of high throughput computational screening (HTCS) of MOFs. In their approach, Wilmer et al. deconstructed experimentally synthesized MOFs into building blocks and recombined them into new structures, using a so-called "bottom-up" approach where building blocks are sequentially connected without any topological bias. Exploiting combinatorics, these authors readily obtained numerous MOFs, with the caveat that ca. 90% of the MOFs assembled into the same topology (pcu) and ca. 10% into another five topologies [194]. This limitation in the number of topologies significantly restricts the volume of MOF space which can be searched. However, to effectively detective experimental targets, molecular simulation needs to be able to access (nearly) every point in MOF-space. After realizing the need for topological diversity, subsequent computational synthesis efforts have used a topologically focused "top down" approach where building blocks are simultaneously mapped onto a topological blueprint that guides their arrangement in space [73–75, 195]. This has been implemented in codes such as AuToGraFs [74] by Addicoat and coworkers, Zeo++ [75] by Martin and Haranczyk, Weaver [195] by Schmid and coworkers, and ToBaCCo [73] by Gomez-Gualdron and coworkers. A critical step in all these algorithms is the rescaling of the blueprint, which needs to match the dimensions of the desired building blocks as closely as possible to create a plausible MOF structure, amenable to optimization by molecular simulation. The simplest scaling method is perhaps isotropic scaling, in which all blueprint unit cell parameters are expanded or contracted by the same factor, and the fractional coordinates of network vertices and edges (the midpoint between connected vertices) do not change. The codes listed above have been limited to computational synthesis of MOFs based on blueprints that rescale isotropically, hence limited to about ca. 50 topologies known as "edge-transitive" [196] (there are over 2800 topologies that are non-edge transitive [79]). In practice, this means that only "single-ligand" MOFs have been computationally synthesized, with the additional provisions is highly symmetric and has a low-aspect ratio. While, as usual, combinatorics allows obtaining large numbers of MOFs, [38] the thousands of computationally inaccessible topologies to date clearly delineate the vast regions of MOF-space that remain unexplored.

How important is MOF topology? In several instances, topology has been shown to directly or indirectly constrain the kind of properties observed in an isoreticular (same topology) series of MOFs [40, 197]. Topology has been shown to influence photophysical properties [198, 199] as well as rates and selectivities of reactions (by controlling the relative position of embedded cooperative catalytic sites) [112].

Additionally, topology may be critical to impart electrical conductivity to MOFs (as only certain combinations of connectivity and chemistry facilitate the required charge transfer mechanisms) [200, 201]. Topology may also be critical for membrane applications as it partly dictates MOF pore architecture, which in turn influences adsorption in, and diffusion through, the MOF (as recently demonstrated for Xe/Kr separation) [18]. Even for applications such as gas storage, where the primary focus is on maximizing properties such as surface area and void fraction, topology has been shown to constrain combinations of these properties [202, 203]. Therefore, motivated by the importance of topology and the desire to alleviate the topological limitations of MOF computational synthesis, we have introduced a new rescaling algorithm, which we describe here, into the publicly available ToBaCCo software. Potential computational cost limitations notwithstanding, this modification allows one in principle to access any point in MOF space. To demonstrate this capability, we illustrate the computational synthesis of a double-linker MOF of previously inaccessible topology. Then, to demonstrate the importance of MOF topology on properties relevant to applications, we computationally synthesized 133 so-called "Cu-paddlewheel MOFs" based just on seven tetratopic ligands of varied aspect ratio and 19 non-edge-transitive topologies (Figure 3.2), yielding seven sets of isomorphic MOFs, where only topology varied within each isomorphic set, but composition remained constant. Then, we proceeded to determine how topology impacted properties relevant to gas storage and separation in the 133 MOFs within each isomorphic set. Note that it is likely that only a few (or even only one) of the isomorphs within a set can be obtained experimentally, but we did not exclude any of the structures from our analysis as determining variations in physical properties between all isomorphs allows us to comprehensively examine how material properties change when crystal topology changes despite retaining identical local chemistry and composition.

If the properties of interest change dramatically with topology, then it is critical to predict which isomorph is experimentally preferred. This means resources for attempted synthesis and testing can be reserved for cases when the isomorph with desirable properties is also predicted to be more thermodynamically or kinetically accessible. While predicting the experimentally preferred isomorph(s) is beyond the scope of the work herein, note that computationally obtaining all potential isomorphs for a combination of building blocks is a crucial step if such prediction is to be made.

3.3 Computational Synthesis Methodology

With the introduction of the new rescaling algorithm, ToBaCCo retains the rationale and basic operational methodology that were detailed in earlier work [73], and remains compatible with our recently introduced molecular building block-based (MBBB) method [204] for the assignation of atomic charges to MOFs. The most significant modification is the new, more sophisticated strategy used to rescale

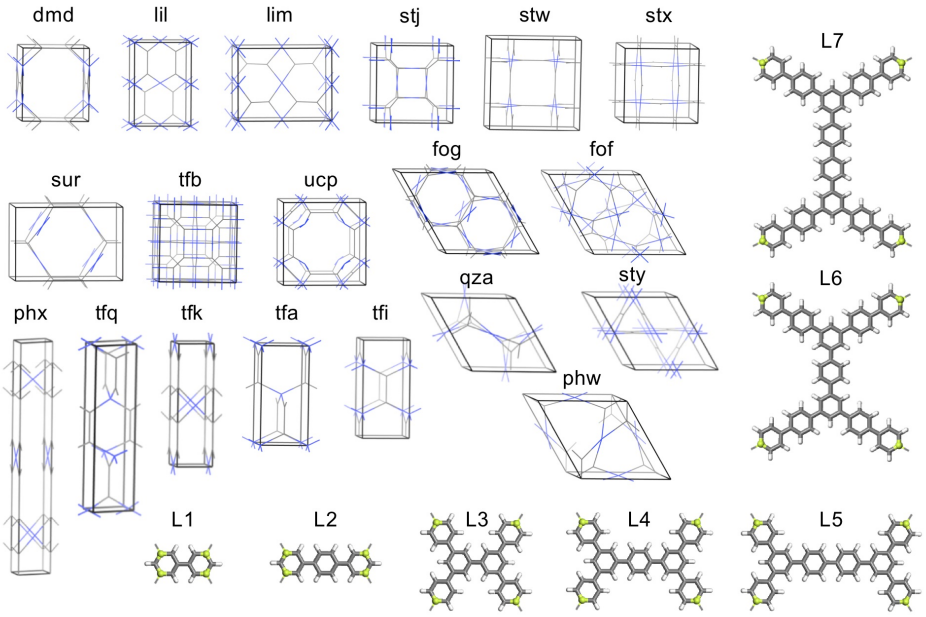

Figure 3.2 The 19 topologies (three letter codes and barycentric unit cells shown) and seven tetratopic linkers (L1-L7), which together with the Cu paddlewheel form the 133 MOFs studied in this work. Atoms colored green indicate connection sites in the linker building blocks. These sites connect to the corresponding connection sites in metal node building blocks.

topological blueprints, which allows the code to work with previously inaccessible topologies. However, other modifications have been made as well to make the code more user-friendly. For instance, users can now create blueprints more easily, as they are formatted as crystallographic information files (CIFs), a format that is familiar to most synthetic chemists (and already used for input building blocks in the original code). In addition, there is no need for explicit specifications of the symmetry of the nodes, as ToBaCCo now automatically recognizes that information from the provided node connectivity in the CIF. For user convenience, a number of blueprints (and building blocks) are already provided in the ToBaCCo Github (which also hosts a manual) including the ones used to obtain the MOFs discussed here (and we will continue to periodically add more blueprints). Before continuing, note that the methods detailed below not only allow for the construction of MOFs, but also of other porous crystal such as zeolites and covalent organic frameworks (COFs). While for generality we focus our explanation on 3-dimensional nets, ToBaCCo can also build 2-dimensional crystals as long as a suitable blueprint is provided.

The algorithm takes advantage of the fact that the crystallographic nets represented by the blueprints are a special type of mathematical graph. A graph consists of vertices $v_1, v_2, \ldots, v_n$ (where n may be infinite) connected by edges. These edges are represented as tuples of vertices, e.g. the edge $(v_1,\ v_2)$ connects vertices v_1 and v_2. A simple graph is a graph with undirected edges, i.e. (v_i, v_j) is equivalent to

(v_j, v_i) for all i and j and no edge connects a vertex to itself. A graph is called connected if any vertex can be accessed from any other vertex by traveling along edges. A net is a simple, connected graph [78]. Graphs are topological constructs that are defined only by their connectivity (edges), but they can be represented in Euclidean space as an embedding . For construction of MOFs (or other porous crystals), we are only concerned with faithful embeddings—i.e. embeddings where no edges intersect and no edges touch vertices other than those of their endpoints [78] (see Figure A.1) —in 3 dimensional Euclidean space ($\mathbb{R}^3$) where vertices are assigned coordinates as 3 component vectors and edges are represented as lines between vertices. A 3-periodic net has an embedding with infinite translational symmetry in 3 dimensions, hence it is analogous to a crystal unit cell. Based on all the above, we can state that a crystallographic net is a 3-periodic net with one or more faithful embeddings in $\mathbb{R}^3$ but no faithful embedding in 2-dimensional Euclidean space [78].

Crystallographic nets can be represented in a similar manner to MOFs (or crystals in general), with a unit cell defined by six parameters $(a, b, c, \alpha, \beta, \gamma)$ and each vertex position defined by three fractional coordinates. Given their similarities, from now on we will use the terms net, template, blueprint and crystallographic net interchangeably. Information that can be used to build numerous topological blueprints (as CIFs) can be found in the Reticular Chemistry Structure Resource (RCSR) which provides a database of crystallographic nets [79] in their barycentric embedding in Euclidean ($\mathbb{R}^3$) space. In the barycentric embedding, each vertex in the net is the centroid of its neighboring vertices, resulting in the most symmetric embedding possible [205]. However, in several instances, a given combination of a priori compatible building blocks may not match the barycentric embedding well. For instance, the local vertex geometries (defined by the set of vectors originating at the vertex and pointing along incident edges towards neighboring vertices) may deviate from the building blocks geometries (defined by the set of vectors originating at the center of the building block and pointing toward the building block connection sites).

Since building block geometry is enforced by their chemistry, to properly "fit" these building blocks into the net embedding, the latter must be readjusted so the geometry of each vertex matches the geometry of the corresponding nodular building blocks (see Figure A.2), but also so edge lengths match chemically enforced distances determined by the size of the building blocks. The original ToBaCCo could only try to readjust the blueprint through isotropically scaling the blueprint unit cell, with the net retaining its barycentric embedding. As a result, nodular building blocks that did not overlap well with the vertex geometries and edge lengths of the barycentric embedding simply could not be used. For example, it was previously impossible to map a Cu paddlewheel (square planar building block with point group D_4 which has 8 symmetry elements) onto a net whose vertex geometries in the barycentric embedding were tetrahedral (point group S_4 which has 24 symmetry elements). Also, it was previously impossible to map a

set of connecting building blocks whose length ratios (i.e. l_0/l_1 where l_i is the length of connecting building block i) differed from the length ratios of the assigned edges in the barycentric embedding (as is the case for many multiple-linker MOFs). Thus, we updated the net rescaling algorithm employed by ToBaCCo by introducing the new anisotropic scaling, which can break the original symmetries of the net and alter, in an unconstrained fashion, a provided blueprint to better suit the chemically constrained building block geometry. Note that since the original blueprint symmetry is typically lost, we found it convenient to set up ToBaCCo to accept blueprints in the "P1" space group even if they correspond to more symmetric space groups. We now broadly discuss the anisotropic scaling methodology (a more mathematically detailed discussion is provided in Appendix A).

The anisotropic scaling method introduced to TobaCCo is made possible by the results of Eon [206] and follows the spirit of the method employed by Boyd and Woo [76], but with key differences (e.g. no net augmentation) that impact the quality of the constructed MOF, which may in turn impact the success of the subsequent optimization step. In the new anisotropic scaling method, ToBaCCo alters the blueprint by adjusting all six unit cell parameters and the fractional positions of the net vertices separately, enabling the independent modification of edge lengths and vertex geometries. Inspired by algorithms to optimize crystalline solid geometries, where atom positions and crystal unit cell parameters are sequentially adjusted to minimize the potential energy of the system, we modified ToBaCCo to sequentially adjust vertex positions and net unit cell parameters to minimize the deviation between the building block and vertex geometries. While there are several ways to calculate the deviation and subsequently adjust the vertex positions and unit cell parameters, we found a method described by Eon to be particularly convenient [206]. This method uses the voltage graph representation of the net to define the vertex fractional coordinates as a function of their displacement from their barycentric positions [76, 206]. These fractional coordinates can then be transformed into Cartesian coordinates using a matrix operator built from the unit cell parameters. This allows an embedding in $\mathbb{R}^3$ to be constructed as a function of the net unit cell parameters and of the displacement of the vertices from their barycentric coordinates. Thus, the deviation between the vertex geometries of any embedding and building block geometries can be minimized (using a bounded optimization algorithm such L-BFGS-B [207, 208]) with respect to these variables. More details on voltage graphs and how ToBaCCo exploits them for crystallographic net scaling can be found in Appendix A.

The workflow of ToBaCCo (Figure 3.3) remains similar to previous versions, with some adjustments to accommodate difficulties associated with the handling of more complex blueprints (see supplementary information).[52] Briefly, for the generation of MOF libraries, ToBaCCo iterates through all provided blueprints and tries to construct the MOFs that result from all possible combinations of compatible building blocks. ToBaCCo recognizes nodular building blocks to be compatible when their number of

connection points coincides with the number of incident edges of at least one of the vertex types of the net. All connecting building blocks are compatible to all edge types in the blueprint, so given the higher complexity of blueprints that ToBaCCo can now handle, edge combinatorics can make the number of possible structures for certain blueprints rise exceedingly fast. For example, with one nodular building block, ten connecting building blocks, and a single blueprint with three types of edges there are already 310 distinct MOFs that ToBaCCo could build. It is possible, however, for the user to limit this number by specifying that connecting building blocks be assigned in the order they appear listed in the building block library. Additionally, the user can use ToBaCCo for the generation of a specific target MOF by instructing the code to use only the building blocks listed in a provided table (which also contains the specific net vertex types that the building block should be assigned to). Once ToBaCCo determines the compatible building blocks for a particular blueprint, it performs the described anisotropic scaling procedure to adjust the net geometry to the building blocks. Next, the building blocks are translated and rotated into their positions in the rescaled net, and, lastly, the corresponding MOF CIFs are generated as output. As usual, once a structure is computationally synthesized, it can be refined by a suitable molecular simulation-based optimization procedure.

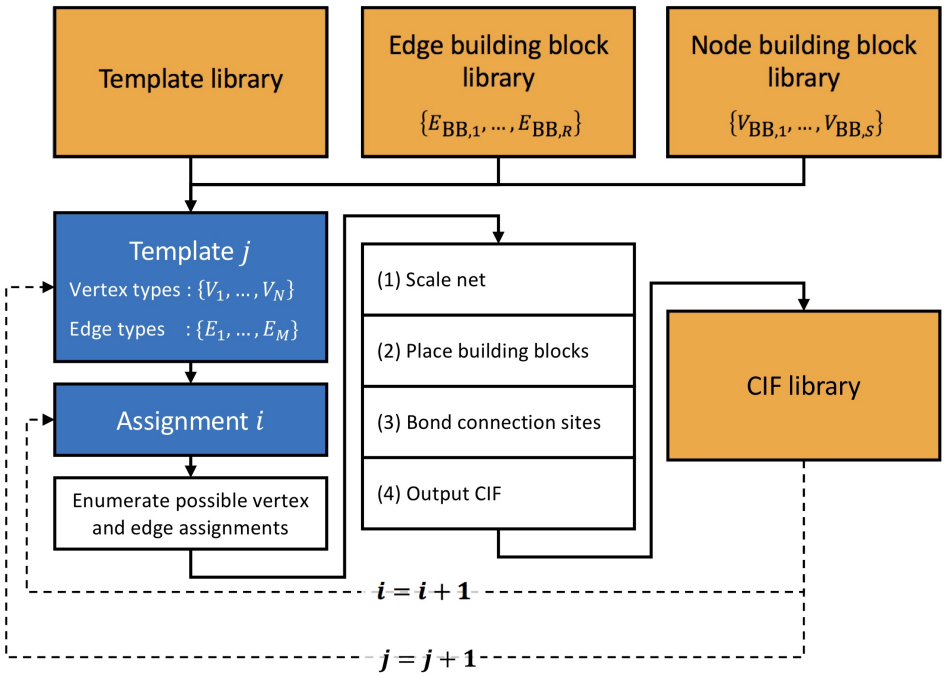

Figure 3.3 A summary of the new ToBaCCo workflow. Algorithmic details on the "scale net" step are provided in A.

To further illustrate how anisotropic scaling is useful for building topologically complex MOFs (i.e. based on non-edge transitive nets), we built two MOFs with topology **kkm** (Figure 3.4). We selected this topology for illustrative purposes as the need for anisotropic scaling is easy to explain in this net. The **kkm** net has two types of vertices (3-connected V1 and 6-connected V2) and two types of edges (E1 and E2). E1 connects 3-connected V1's to 6-connected V2's, and E2 connects 6-connected V2's to 6-connected V2's. Figure 3.4 shows how all E1's lie in the bc plane while all E2's have their only non-zero components in the c-direction. Under the constraints of isotropic scaling, the a, b, and c vector lengths would change by the same proportion, and so would the lengths of both edge types. Thus, it would not be possible to use connecting building blocks with different lengths to construct MOFs. However, the new anisotropic scaling capabilities introduced in ToBaCCo makes it possible for the code to accommodate connecting building blocks of significantly different lengths in the **kkm** blueprint. Note that the resulting MOFs (Figure 3.4) feature more than one type of organic linker. Specifically, the MOFs would be experimentally synthesized combining a ditopic and a tritopic organic linker, much like in the synthesis of the well-known material MOF-210. Computationally, however, the tritopic linker is decomposed into a 3-connected nodular building blocks (the node of the tritopic linker) that are assigned to V2's and connecting building blocks (the arms of the tritopic linker) that are assigned to E1's. Although for this example we chose MOFs with only two organic linkers to facilitate illustrations, ToBaCCo is capable of constructing MOFs with more than two linkers, and MOFs with templates in which no edges are orthogonal in the starting embedding.

3.4 General Methods

3.4.1 MOF Library Construction

To study the topological dependence of properties of MOF isomorphs, 133 Cu-paddlewheel-based MOFs featuring tetratopic linkers of different sizes and aspect ratios were constructed using our new version of ToBaCCo. For each of seven tetratopic linkers (Figure 3.2), 19 isomorphic MOFs were constructed using 19 topological blueprints (Figure 3.2). To illustrate the anisotropic scaling, the seven MOFs constructed out of the fog topological blueprint are described in more detail in Figure A.3. CIFs for all the 133 constructed MOFs are added as supplementary files. Recall that, in ToBaCCo, organic linkers are deconstructed into smaller building blocks, hence each blueprint featured two types of vertices and two types of edges: 4-connected vertices to "host" Cu-paddlewheel building blocks, 3-connected vertices to host the nodes of the linker, edges connecting 3-connected to 4-connected vertices to host the "arms" of the linker, and edges connecting 3-connected to 3-connected vertices to host the "body" of the linker. The seven linkers resulted from different combinations of the used connecting building blocks, where in some cases no edge building block was used, meaning nodular building blocks were connected directly (e.g. in

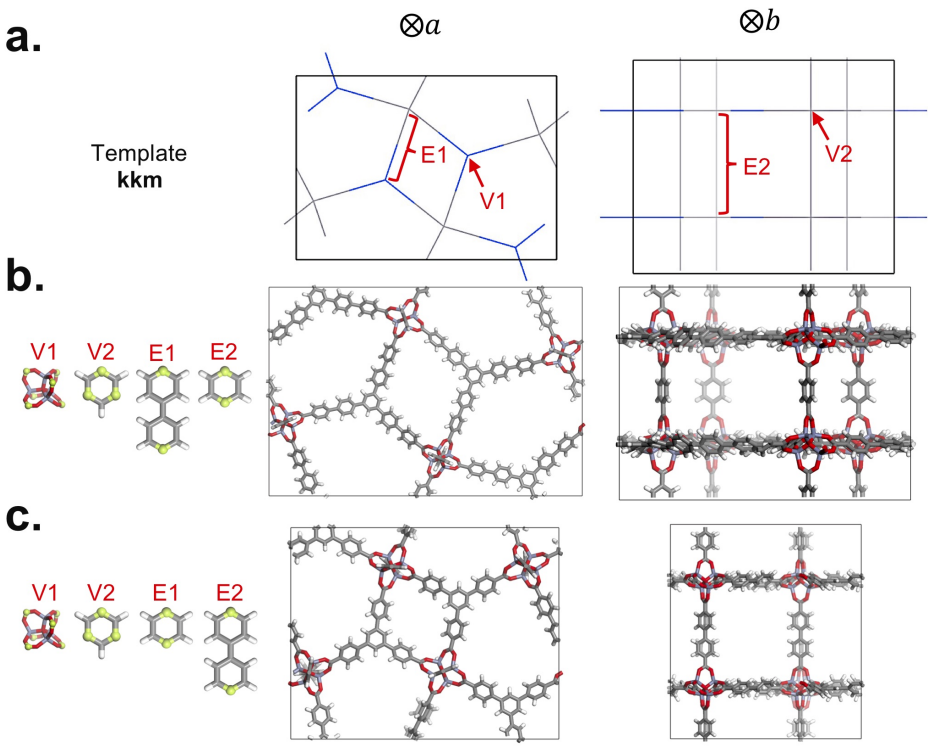

Figure 3.4 (a) The template **kkm** along the a vector (left) and the b vector (right), examples of the different vertex types (V1, V2) and edge types (E1, E2) are labelled. (b) The building block assignments used to construct the first **kkm** MOF (left), and the MOF viewed down the a and b vectors (middle and right, respectively). (c) The building block assignments used to construct the second **kkm** MOF (left), and the MOF viewed down the a and b vectors (middle and right, respectively).

L1, L2, and L3). Given the different sizes of the latter (resulting in different aspect ratios of the tetratopic linkers), the new anisotropic scaling capability of ToBaCCo played a crucial role in constructing these MOFs. CIFs created by ToBaCCo where translated into LAMMPS data files using the LAMMPS Interface code written by Smit and coworkers [209].

3.4.2 MOF Geometry Optimization

The geometry of the 133 MOFs was optimized using parameters from the UFF4MOF [82, 210] force field to describe bonded and non-bonded interactions. This force field is a modification of the Universal Force Field (UFF) [80] that was developed specifically for simulating MOF structural properties. The molecular dynamics (MD) software LAMMPS [211] was used to carry out the optimizations. The geometry optimization procedure used is as follows: first, the atomic positions were relaxed while keeping the unit cell parameters fixed with the fast inertial relaxation engine (FIRE) algorithm developed by Bitzek et al. [212] FIRE is a damped dynamics method, and thus requires a timestep (we used 1.0 fs). After this initial relaxation, a minimization cycle was performed, where in each step of the cycle the atomic positions and unit cell parameters were relaxed together, using a conjugate gradient (CG) algorithm, followed by a relaxation of only the atomic positions (again using the FIRE algorithm). This cycle was halted when the energy change from the previous step to the current step was less than 10-10 kcal/mol. The combination of the FIRE and CG algorithms was used in this cycle simply because we observed faster convergence when using the combination of algorithms than when using a single algorithm (whether FIRE or CG). Each minimization was considered converged only when the change in energy from the previous step to the current step divided by the current energy magnitude was less than 10-15 and the forces on atoms were less than 10-15 kcal mol-1 Å^{-1}. No charges were used during geometry optimizations, as recommended by Smit and coworkers when using UFF4MOF to optimize MOF crystal geometries [209]. At least two of the 133 MOFs studied here have been synthesized, corresponding to NOTT-100/MOF-505 [213, 214], and NOTT-101 [213]. For these MOFs the experimentally obtained structures agree well with the computational ones. NOTT-100 and NOTT-101 correspond to fof-L1 and fof-L2, respectively. Note that, upon optimization, some of the 133 MOFs appeared to present somewhat visible deformation/strain in their building blocks. It is likely that such MOFs are "thermodynamically inaccessible" since they are bound to be less energetically stable than other isomorphs where the strain is not visible. Other MOFs without visible strain have Cu_2 dimers too close to each other. It is likely that such MOFs are "kinetically inaccessible" since Cu_2 dimers cannot get as close during MOF crystallization due to the presence of Cu-bound solvent. We did not remove these structures from this study as to be able to comprehensively interrogate the effect of topology at constant chemistry on MOF properties. Indeed, effect of topology on

the feasibility of a structure is indirectly discussed when discussing mechanical properties.

3.4.3 Grand Canonical Monte Carlo Simulations

. Gas adsorption in each of the 133 MOFs was simulated using grand canonical Monte Carlo (GCMC) with the RASPA code [91]. A GCMC simulation consists of a series of cycles, where each cycle consists of N attempted Monte Carlo moves, where N is the maximum between the number of adsorbates in the framework and 20. A move is either adsorbate rotation, translation, insertion, or deletion (or identity change when two or more adsorbate species are present). The overall adsorbate loading is then taken to be the average of the loadings recorded after a set number of cycles. Methane adsorption simulations consisted of 5,000 initialization cycles followed by 5,000 production cycles (for calculation of averages). Ethene/ethane simulations consisted of 10,000 initialization cycles and 10,000 production cycles. All non-bonded interactions were modeled using a Lennard-Jones (LJ) potential, with the exception of ethene-Cu interactions which were modeled using a Morse potential. For methane and ethene:ethane adsorption, MOF atoms were modeled with UFF [80] parameters and adsorbates with TraPPE [215, 216] parameters, with the exception of ethene-Cu interactions which were modeled with the parameters determined by Kulkarni and Sholl [145]. Cross interactions not explicitly parametrized were modeled using Lorentz-Berthelot mixing rules. Methane deliverable capacity was calculated by subtracting the loading at 5 bar from the loading at 65 bar. Ethene selectivity was calculated via Equation 1.3.

where S is ethene selectivity, N_i is the loading of species i, and γ_i is the mole fraction of species i in the gas phase mixture. Ethene selectivity was calculated for 10:90 and 62:38 ethene:ethane mixtures at 0.1, 1, and 10 bar. The 62:38 ethene:ethane composition corresponds to that commonly observed in industrial distillation [217], while the 10:90 composition corresponds to a dilute ethene mixture. All adsorption simulations were performed at 298 K.

3.4.4 Mechanical Properties Calculations

The tensorial form of Hooke's law can be written as

$$\sigma_{ij} = \sum_{kl} c_{ijkl}\epsilon_{kl} \tag{3.1}$$

where σ_{ij} is the component of the stress tensor corresponding the ij direction (see Figure A.5), ϵ_{kl} is the component of the strain tensor corresponding to the kl direction, and c_{ijkl} are entries in C, the fourth order elasticity tensor (i, j, k, l range from 1 to 3).[80] In general, C has 81 elements. However, symmetries in stress and strain allow C to be written in Voigt notation as a 6×6 matrix with a maximum of 21 unique elements, depending on the crystal unit cell symmetry. In Voight notation $11 \rightarrow 1$, $22 \rightarrow 2$, $33 \rightarrow 3$, $23 \rightarrow 4$, $13 \rightarrow 5$, and $12 \rightarrow 6$, simplifying Hooke's law to a one index notation form [218]. Here we calculated C as a

6×6 matrix were each entry was calculated as

$$c_{ij} = \frac{\partial \sigma_i}{\partial \epsilon_j} = \frac{1}{V} \frac{\partial^2 E}{\partial \epsilon_i \partial \epsilon_j} \tag{3.2}$$

where E/V is the strain energy density. To obtain the matrix components, the MOF unit cell was compressed/expanded or sheared a small amount (1.5%) in three orthogonal directions and then atom positions were relaxed. The resulting changes in each of the components of the stress tensor were calculated, allowing Equation 2 to be evaluated. Deformation by 1.5% was selected as this value was the center of a region of values between 0.2% and 3.0% that produced nearly exactly the same elasticity tensor entries. Atom positions were relaxed after each deformation and before changes in the stress tensor components were calculated using a CG algorithm with energy and force (energies and forces where computed using UFF4MOF) stopping criteria of 10-15 and 10-15 kcal mol^{-1} Å^{-1}, respectively. This procedure was implemented using modifications of the elastic example scripts available in each LAMMPS distribution (examples of our modified scripts are provided in the supplementary files). Once C was obtained we determined whether each MOF was mechanically stable using the Born stability criteria, which states that the elastic energy for any strain must be positive (i.e. the crystal resists strain) [219]. This is equivalent to C having only positive eigenvalues.[80], [81] Thus, we classified a MOF as mechanically unstable if the smallest eigenvalue of its elastic tensor was zero or negative. To calculate the bulk modulus (K) we averaged the Voigt bulk modulus (K_V) with the Reuss bulk modulus (K_R) since the K_V and K_R define upper and lower bounds, respectively, for the true bulk modulus [220]. Details of the calculation of K_V and K_R are given in Appendix A. No charges were used during the calculation of elastic constants.

3.4.5 Textural Properties Calculations

MOF volumetric surface area (VSA), void fraction (Vf), and largest pore diameter (LPD) were calculated using zeo++ [85]. Volumetric surface area and void fraction were calculated by sampling the surface area/volume accessible to a rigid probe of defined radius, surface area was calculated with a N2 sized probe (radius of 1.82 Å) and void fraction was calculated with a helium sized probe (radius of 1.32 Å).

3.5 Results and Discussion

3.5.1 Topological Dependence for Dispersion Interactions-Driven Gas Storage

Interest in storage of fuel gases such as methane and hydrogen was probably the strongest contributor to the explosion of MOF synthesis. Methane and hydrogen interact with most chemical moieties encountered in MOFs via dispersion interactions. Thus, the ability of MOFs to store these gases at a given temperature primarily depends on how much space (quantified as void fraction) and how many interaction sites (quantified as surface area) the MOF pore structure provides. Yet, Figure 3.5 shows how changing the

topology for one combination of building blocks (i.e. composition) can engender pronounced changes in the methane deliverable capacity. Each MOF composition consists of Cu paddlewheels and tetratopic linkers (LN) in one-to-one ratio. Sizes for the seven studied linkers follow the order L1 < L2 < L3 < L4 < L5 < L6 < L7, where size is taken to be the area of the rectangle formed by the four linker connection sites.

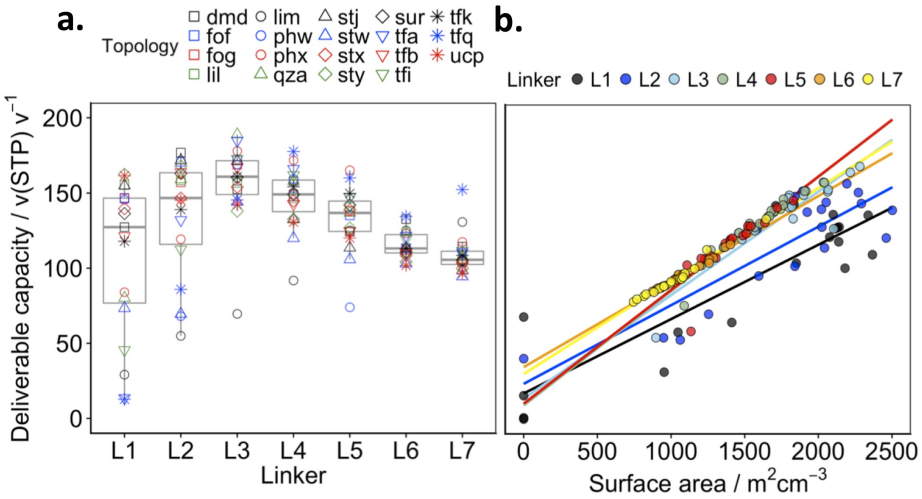

Figure 3.5 (a) Methane deliverable capacity verses linker type each of the 133 MOFs is shown as a point with color/shape corresponding to the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of deliverable capacity and surface area change with linker size. (b) Methane deliverable capacity versus volumetric surface area with the corresponding line of best fit for the MOFs of each linker type, lines and points are colored by linker type.

Figure 3.5 illustrates how methane deliverable capacity (calculated as the difference in adsorbed methane at 65 bar and 5 bar) and volumetric surface area change as a function of the constituent linker of the Cu-paddlewheel MOF. Notice that, for the same constituent linker the methane deliverable capacity can change by amounts in the 50-150 cc(STP)/cc range. Note, for instance, that the hypothesized crystal with the linker L3 assembling with the paddlewheels into the **qza** topology produces a deliverable capacity close to the 190 cc(STP)/cc, which is quite close to the maximum deliverable capacity predicted for any MOF at these conditions [54]. However, if linker L3 were to assemble into the **lim** topology, it would only present a deliverable capacity of 70 cc(STP)/cc. The reason for this topological dependence is that the volumetric surface area also present strong variability within each isomorphic set of MOFs due to changes in crystal topology. indeed, within some isomorphic sets surface area can vary by 2500 m2/cm3 range. Importantly, within each isomorphic set, the deliverable capacity correlates well with volumetric surface area (Figure 3.5). Perhaps, the correlation is somewhat less clear for the sets based on the smallest linkers (L1-L2) because void fraction limitations start to play a more important role in those cases. As a measure of the reliability of our methane simulations, note that the measured deliverable capacities of NOTT-100

and NOTT-101 have been reported elsewhere [221] to be 139 and 183 cc(STP)/cc, respectively, in good agreement with our corresponding simulated deliverables capacities (147 and 171 cc(STP)/cc, respectively).

3.5.2 Topological Dependence for Chemical Separations Driven by "Complexation" Interactions

Since olefins are known to form complexes with Cu(II) through their double bonds (which paraffins do not possess), the Cu-paddlewheel moieties encountered as building blocks of many MOFs provide these nanoporous crystals with the potential to separate olefin/paraffin mixtures such as ethene/ethane. Our GCMC simulations on ethene/ethane mixtures indicate that simply changing the topology in which one combination of building blocks assembles can lead to significant changes in adsorption characteristics, in some cases even switching the selectivity of the resulting MOF from ethene- to ethane-selective (Figure 3.6). However, the extent to which topology affects ethene selectivity depends on the crystal composition and the thermodynamic conditions of the mixture, including composition and pressure.

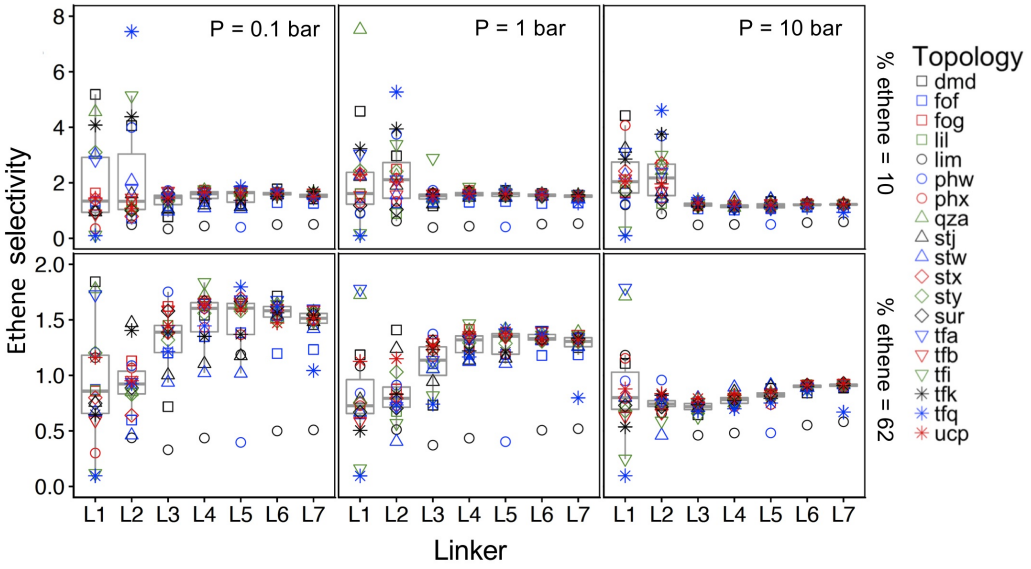

Figure 3.6 Ethene selectivity vs. MOF composition at different ethene:ethane compositions and pressures. Individual MOFs are shown as points, box and whisker plots are shown in grey for each MOF composition (isomorphic MOF set).

For the 10% ethene mixture at all pressures, the selectivity in MOFs based on the larger linkers (L4-L7) tends to be similar regardless of crystal topology. Indeed, apart from MOFs with **lim** topology, the ethene selectivity remained somewhat lower than two for all topologies. On the other hand, for the smaller linkers (L1-L3), the selectivity could reach values as high as approximately 8 or 5 depending on the pressure, but also reach values as low as almost zero (i.e. strongly ethane-selective) solely by altering the crystal

topology. For the 62% ethene mixtures, ethene selectivity is generally lower than for the 10% ethene mixtures. In fact, the selectivity at high pressure (10 bar) remained below one (i.e. ethane-selective) for all linkers regardless of topology, except for the smallest linker (L1) for which altering the topology could make the ethene selectivity increase to nearly two. At lower pressures (0.1 bar and 1 bar), ethene-selectivities higher than one are again observed, although highly depending on crystal topology, especially for the smaller the linkers (L1-L3).

It is plausible that the crystal topology strongly influences the ethene-selectivity because it alters the concentration and spatial distribution of Cu open-metal sites (OMS) within the crystal. Therefore, to understand the influence of topology better, we plotted how ethene selectivity changes with $1/\rho_{\text{oms}}$, which is the inverse of the OMS number density. We find that ethene selectivity generally increases as $1/\rho_{\text{oms}}$ (Figure 3.7), with the highest ethene selectivities for all compositions and pressures observed for the lowest values of $1/\rho_{\text{oms}}$. However, the lowest ethene selectivities are also observed for the lowest values of $1/\rho_{\text{oms}}$, which tend to occur for the smallest linkers (e.g. L1-L2). We hypothesize that this occurs because although reducing the linker size increases the concentration of OMS, the crystal topology can alter the spatial arrangement of OMS in such way as to make them inaccessible to ethene molecules; thus, preventing ethene coordination to these sites. Under these circumstances a dispersion interaction-based selectivity mechanism dominates, similar to that for Xe:Kr mixtures [18, 22, 222]. Where, provided both components can fit into the crystal pores, the selectivity for the larger component (here ethane) increases as the pores become smaller due to stronger confinement effects.

To test this hypothesis, we analyzed adsorption snapshots from our GCMC simulations (Figure 3.8). We selected two systems based on the L1 linkers, as with this linker we observed higher variability in selectivity due to crystal topology changes. The sty-L1 MOF was selected as the ethene-selective MOF representative, whereas the **lim**-L1 MOF was selected as the ethane-selective MOF representative. Ethene is clearly able to interact with OMSs in the sty MOF, but not in the **lim** MOF, because in the latter the Cu paddlewheels are "stacked" too closely to allow ethene to fit in. Thus, ethene attempting to coordinate with an OMS incurs a steric energy penalty. This paddlewheel stacking occurs in all **lim** MOFs regardless of linker, which explains why crystals in this topology are consistently "outliers" selective towards ethane. To quantify these effects in all 133 MOFs, we calculated the proportion of ethene (as a fraction of total adsorbed ethene) in the framework that was within 4 Å of any OMS (Figure A.6).

We observed a correlation between the proportion of OMS-bound ethene and selectivity, especially for the 10% ethene mixture (Figure A.6). It, therefore, becomes clear that facilitating ethene coordination to OMSs is critical to achieving high ethene selectivity. For the 10% ethene mixtures, the selectivities are higher than for the 62% ethene mixture, because in the former case the MOF crystals are able to "extract"

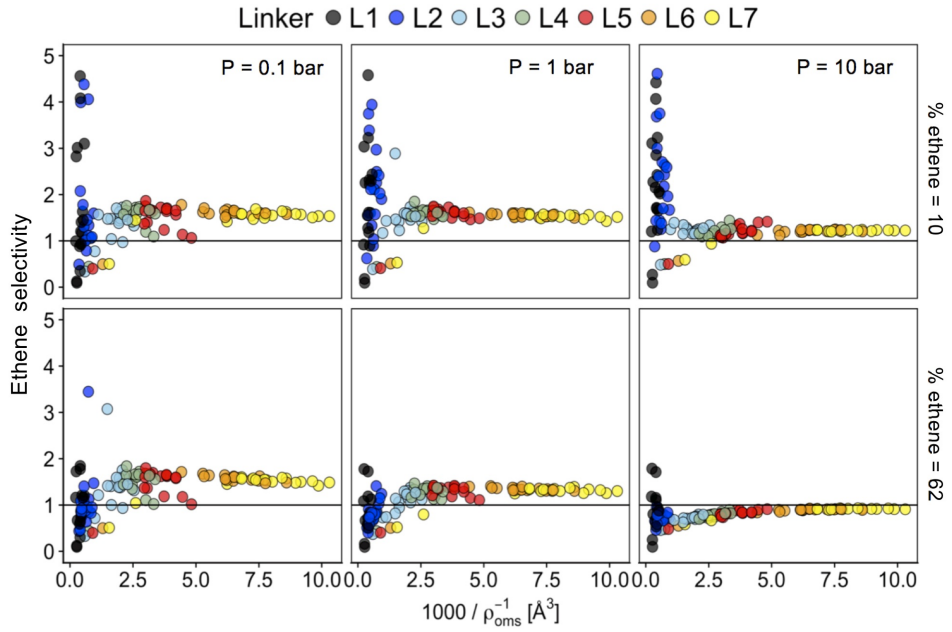

Figure 3.7 Ethene selectivity vs $1/\rho_{\text{oms}}$ for each ethene:ethane composition and pressure considered. Note that some points are excluded by the plot range, this is to better show the selectivity trend with increasing $1/\rho_{\text{oms}}$. The black lines correspond to a selectivity of one.

a larger fraction of the ethene present in the gas phase before the OMSs get "saturated." For the 62% ethene mixture there is more ethene in the gas, so the OMSs get saturated with a smaller fraction of the ethene in the gas phase. After this saturation, additional ethene can only adsorb in sites where dispersion interactions dominate the adsorption mechanism, which favors ethane adsorption. Indeed, we find the percentage of adsorbed ethene less than 4 Å from a Cu OMS to be typically lower in the 62% ethene mixture than in the 10% ethene mixture (Figure A.6).

3.5.3 Topological Dependence for MOF Mechanical Properties

For most applications one desires a MOF that not only has outstanding performance metrics, but also maintains such performance under relevant operating conditions. In most cases, this requires the MOF crystal to retain structural integrity during usage, which at least requires the crystal to be mechanically robust, and resist stress in all directions. MOFs with a relatively high bulk-modulus (which is a measure of how the crystal resist isotropic compression) have been deemed desirable to maintain structural integrity under high-pressure operating conditions, but also if pelletization is expected to be necessary to utilize the material [223, 224]. On the other hand, the mechanical stability of a MOF crystal, as evaluated by the Born stability criterion [219], could be a partial indication of its synthesizability. That is, while being mechanically stable does not guarantee that a MOF can be synthesized (or even that it is the most

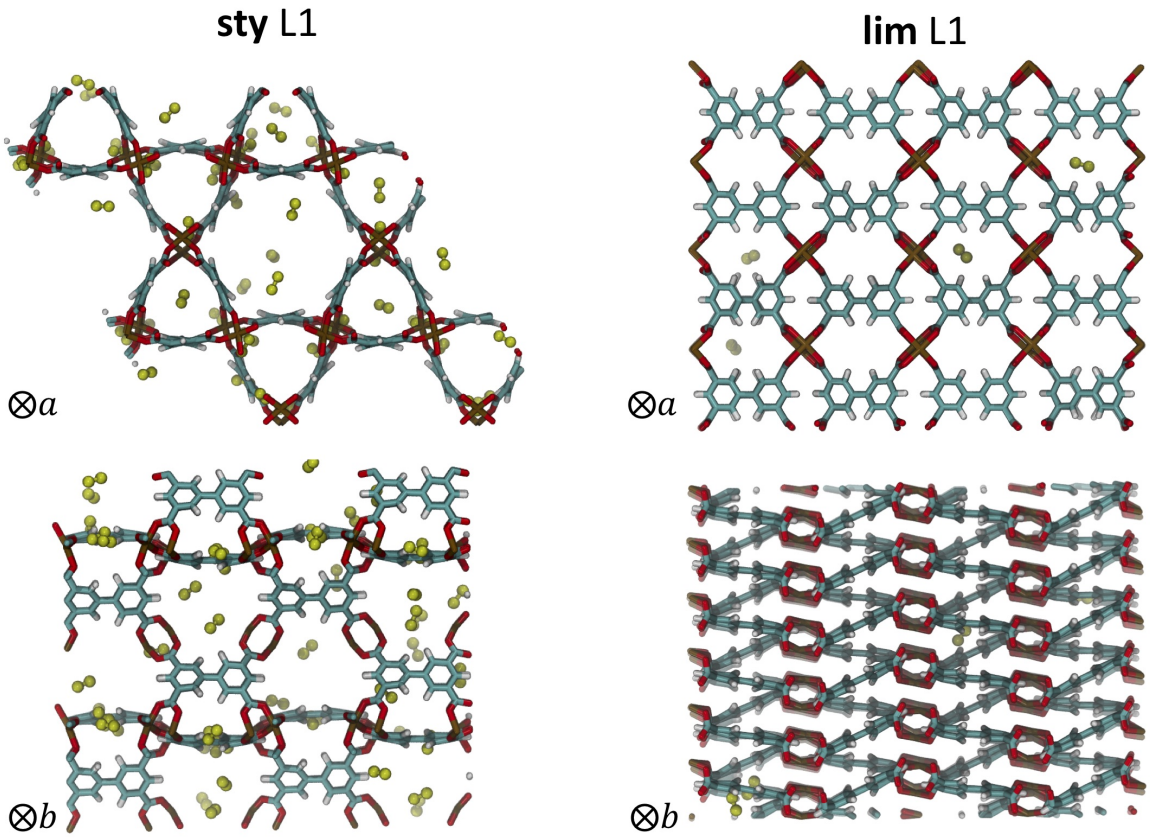

Figure 3.8 Simulation snapshots showing the location of adsorbed ethene. Left: Ethene selective ($S = 3.1$) sty-L1 MOF. Right: non-selective ($S = 0.9$) **lim**-L1 MOF. Ethene is shown in yellow (for clarity ethane is not shown). Both MOFs are shown from two perspectives: viewing down the a axis (top) and viewing down the b axis (bottom

synthetically likely isomorph), being mechanically unstable means that the MOF is unlikely to be synthesized (depending on the role of solvent during synthesis), or at least that the MOF would collapse upon pore activation.

This indicates that evaluating the Born stability criterion in large MOF libraries may provide a way to identify and filter out the unrealistic targets for synthesis and testing among computationally constructed MOF crystals. The Born stability criterion states that stable crystals must resist strain [218], which mathematically means that all the eigenvalues (also called eigenstiffnesses) of their elasticity tensor (in Voigt notation) must be positive [219]. Here, we calculated the elasticity tensor for the 133 "computationally synthesized" MOFs and classified them as mechanically stable if the smallest eigenvalue of their elasticity tensor was positive, and mechanically unstable otherwise. Our mechanical stability calculations are summarized in Figure 3.9, which, like adsorption properties, was found to be strongly dependent on crystal topology. That is, for a given linker, the corresponding MOF crystal can switch from mechanically unstable to stable (and vice versa) solely due to changes in the topology the linker assembles into (Figure 3.9). On the other hand, some topologies tend to more frequently result in mechanically stable MOFs, as was the case for the **fof**, **fog**, **tfq** and **dmd** topologies.

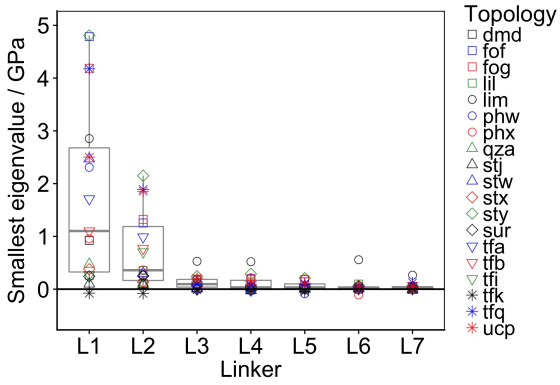

Figure 3.9 Smallest eigenvalues of the MOF elastic tensor vs. linker type. Each of the 133 MOFs is shown as a point with color/shape corresponding the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of the eigenvalues change with linker size.

Regardless of mechanical stability, we assessed the dependence of the bulk modulus on crystal topology, as the bulk modulus is often referred to in the literature when discussing MOF suitability for industrial applications. Once again, we observed significant variations due to changes in crystal topologies, as noted in the wide range of bulk modulus values for MOFs composed of the same linker (Figure 3.10). For instance, for the mechanically stable MOFs comprised of the smallest linker (L1), the bulk modulus can vary from 1 to 28 GPa. As can be seen in Figure 3.10, bulk modulus generally decreases with increasing

linker size, which means it generally decreases with increasing crystal pore size (as determined by the largest pore diameter, LPD). Therefore, the size of the linker, (which roughly correlates with the flexibility of the linker) is a crucial factor determining the ability of a MOF crystal to resist isotropic pressure. However, linker size is not the only factor as topology also play a role. Consider that MOFs with large linkers (e.g. L5-L7) can have bulk moduli similar or larger than MOFs with smaller linkers L1-L4, depending on the crystal topology. This occurs on one hand because topology can determine the pore size that results from the assembly of a given linker into a MOF, but also because topology can influence bulk moduli in similar ways as it influences mechanical stability. Namely, certain topologies may present structural motifs that lend themselves to easy compression or constrain linkers to conformations that do not appreciably resist compression. Note that a MOF having a large bulk modulus does not exclude the possibility of the MOF having soft deformation modes or even being mechanically unstable. This is because the bulk modulus relates to the resistance of the MOF crystal to only one specific strain mode (isotropic compression), whereas mechanical stability depends on all components of the elastic tensor. Therefore, discussion of MOF mechanical robustness should not be limited to the bulk modulus calculation.

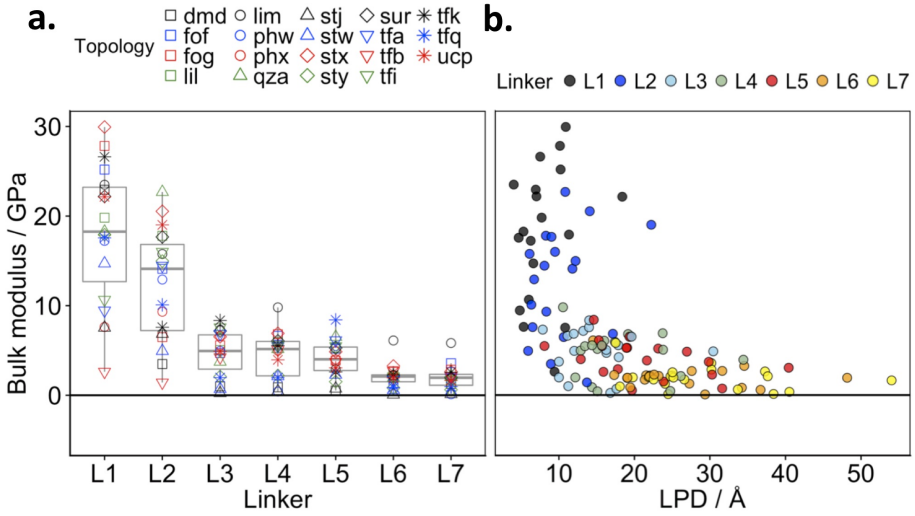

Figure 3.10 (a) Bulk modulus vs. MOF linker type. Each of the 133 MOFs is shown as a point with color/shape corresponding the legend shown. Box and whisker plots are shown behind the individual plots in grey, showing how the median and spread of the bulk moduli change with linker size. (b) Bulk modulus vs. the largest pore diameter (LPD) with points colored by linker type.

3.6 Conclusions

We have presented a critical modification to the topologically-based crystal constructor (ToBaCCo) code, which not only allows for large scale computational synthesis of topologically diverse MOF libraries,

but also for targeted construction of individual MOFs or topologically complete sets of isomorphic MOF. Additional modifications have also made ToBaCCo more user-friendly for casual users. The importance of capturing topological diversity in high throughput computational studies has been demonstrated on the basis of the strong variation of adsorption (methane deliverable capacity and ethene/ethane selectivity) and mechanical (Born stability criterion and bulk modulus) properties within isomorphic MOFs solely due to changes in crystal topology. Notably, in the properties studied here, topology seems to be most influential in MOFs with smaller linkers. Given the extremely soft deformation modes observed in many of the MOFs built with linkers L3-L7, we expect the "small-linker" MOFs to be the most experimentally relevant, at least when not considering interpenetration. In addition to enabling a more comprehensive exploration of the MOF-space, the ability to create a targeted, topologically-complete set of isomorphic MOFs could prove useful to help elucidate synthesized MOF structures by PXRD pattern comparison with computational models.

CHAPTER 4

USING MOLECULAR SIMULATION AND MACHINE LEARNING TO DESIGN MOFS FOR CO_2 CAPTURE

Modified from a paper published in Chemistry of Materials.[4]

Ryther Anderson[5], Jacob Rodgers[5], Edwin Argueta[6], Achay Biong[5], Diego Gómez-Gualdrón[5,7]

In this chapter, we investigate how machine learning algorithms can be used for the rational design of MOFs for CO_2 capture. We analyze, in-depth, the CO_2 capture capabilities of a set of MOFs based on 15 topologies and 12 linker functional groups (all with DFT calculated charges). Specifically, we use machine learning algorithms trained on GCMC data to elucidate MOF design principles based on topology, textural properties, metal species, and linker functionalization. The analyses conducted in this chapter show how useful machine learning models can be when designing MOFs (or other porous crystals) for adsorption-based separation applications. Additionally, the MOF representations used as inputs for the models trained herein provide a basis for the inputs used during the training of the larger-scale models presented in Chapters 5 and 7.

4.1 Abstract

Open framework materials such as metal-organic frameworks (MOFs) often possess structurally and chemically tailorable nanopores. This exceptional tunability has allowed for careful positioning of optimal adsorption sites within MOF pores to enable selective CO_2 physisorption, making these materials promising for energy-efficient CO_2 capture. However, given the multitude of features that can be simultaneously altered within the thousands of MOFs synthesized to date, it can be daunting to elucidate the most critical features for boosting CO_2 capture capabilities. Here we use a multiscale approach—density functional theory (DFT), grand canonical Monte Carlo (GCMC) and machine learning (ML)—to investigate the role of various pore chemical and topological features in the enhancement of CO_2 capture metrics of MOFs. To enable a thorough "sweep" of a target region of MOF structure-space, we used computational synthesis methods to create sets of MOFs encompassing all possible combinations of 16 topologies and 13 functionalized molecular building blocks. The adsorption of pure CO_2, and CO_2/H_2 and

[4]Reproduced with the permission of: R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. Gómez-Gualdrón. "Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: from molecular simulation to machine learning" *Chem. Mater.* 30, 18, 6325–6337. Copyright 2018 American Chemical Society.

[5]Chemical and Biological Engineering Department, Colorado School of Mines
[6]NuMat, Skokie IL 60077, USA
[7]Corresponding author, dgomezgualdron@mines.edu

CO_2/N_2 mixtures for the resulting 31 parent MOFs and its derivatives was then simulated and CO_2 capture metrics calculated. Functionalization with hydroxyl, thiol, cyano, amino or nitro chemistries was found to often improve CO_2 capture metrics of the parent MOFs, but the efficacy of this strategy depended strongly on the pore topology. Decision trees were trained to predict the improvement or decline of CO_2 capture metrics upon functionalization of parent MOFs, whereas five additional machine learning algorithms were trained to predict absolute metrics for all MOFs. The training of these algorithms allowed us to determine the relative importance without "human bias" of various pore chemical and structural/topological factors on the CO_2 capture capabilities of MOFs.

4.2 Introduction

Electricity generation from coal and natural gas power plants accounts for about one third of the ca. 5 billion metric tons of CO_2 emitted in the U.S. each year [225]. As energy demand continues to increase, successful strategies for on-site CO_2 capture are needed to reduce CO_2 emission rates and its related impact on atmospheric processes. Solvents have been widely studied for CO_2 capture [226–229], but solvents typically require significant energy to be regenerated [230]. On the other hand, CO_2 capture by physisorption on open framework materials (OFMs) such as metal-organic frameworks (MOFs) [189, 231] could potentially be significantly less energy-demanding [232, 233]. However, to be successful, the material must capture CO_2 with sufficient abundance and selectivity.

To this end, MOFs are particularly attractive due to their typically high surface areas [52, 234] and pore volumes [30], which offer numerous sorption sites and an abundance of void space to host CO_2 molecules. Furthermore, the modular, building block-based character of MOFs [235–238] can be exploited to carefully tune the chemistry and spatial distribution of these sites to enhance affinity and selectivity for CO_2, i.e. CO_2-friendly electrostatic pockets can be created [139, 239, 240]. But first, an accurate understanding of how pore chemistry and architecture (topology) impact CO_2 sorption under post-combustion [241–243] and/or pre-combustion [136, 238, 244], conditions is needed. However, with millions of potential MOF designs to explore [163], attaining the required level of understanding can be daunting, especially from an experimental perspective. Thus, computational alternatives aligned with the Materials Genome Initiative become attractive as means to achieving this understanding [53, 163].

Furthermore, while both the chemistry and topology of MOF pores are expected to critically impact CO_2 sorption, separating the role of each variable has been difficult experimentally, because setting up experiments at "constant" pore topology while changing chemistry (and vice versa) is not straightforward. The systematic functionalization of a reference parent linker can be challenging, and changing the functionalization of the parent linker can change the topology of the synthesized MOF [117, 245], or even

result in no MOF at all. This is reflected on the scarcity of "complete MOF series" suitable for systematic studies in the Cambridge Structural Database (CSD) [190]. Another difficulty faced by experiments to isolate the role of functionalization is that the latter can alter variables such as pore size, which, in turn, also impact CO_2 sorption [246]. Nonetheless, motivated by the amine scrubbing process to capture CO_2 [228], MOF amino-functionalization remains most popular in the search of enhanced CO_2 sorption [247–250], although synthetic approaches such as multi-functionalization (as in so-called multivariate (MTV) MOFs) have shown promise for enhancing CO_2 sorption as well [251–253]. To make the best functionality/topology choice when designing new MOFs for CO_2 capture, however, it is critical to fully understand how these two factors (and related ones) affect MOF performance, or at least fully assess their relative importance.

Recent methods to computationally synthesize MOFs while controlling topology [73–75, 195] allow for the creation of complete MOF series suitable for systematic study. Simulation data generated on these series can then be analyzed to determine the relative importance of pore chemistry and topology, and thus help determine how to tune them more effectively to enhance CO_2 capture capabilities. With the advent of high throughput computational screening (HTCS) [37, 232], vast amounts of CO_2 simulation data have been generated and analyzed to reveal general structure-property relationships [246]. However, the role of pore chemistry and topology remains to be established due to factors in earlier HTCS studies such as (*i*) limited number of topologies considered, (*ii*) limited number of functional groups considered, (*iii*) incomplete MOF series, and iv) accuracy uncertainty stemming from the description of electrostatic interactions.

Electrostatic interactions are critical for CO_2 sorption [254], so their accurate description by Coulomb interactions is necessary for meaningful simulations. However, issues with computational cost have motivated the use of inexpensive "electronegativity equalization" [255–258] methods to assign Coulomb charges to MOF atoms during HTCS. Such processes overestimate charges of, for instance, halogens and the associated electrostatic interactions with these groups [204]. Thus, in previous HTCS studies, halogenated MOFs were consistently found as top performers for CO_2 capture [246], and perhaps influenced the selection of halogenated MOFs in posterior studies [259]. Recently, several groups have made efforts to develop inexpensive (HTCS-compatible) methods to assign Coulomb charges that agree well with charges obtained from expensive density functional theory (DFT) calculations in MOF crystallographic unit cells [204, 260]. Some of us have demonstrated that the use of molecular building block-based (MBBB) charges [204] that fit the DFT calculated electrostatic potential (ESP) around individual MOF building blocks result in sorption loadings and mechanisms consistent with those obtained with charges that fit the DFT-calculated ESP in the MOF unit cell (e.g. REPEAT charges [261]).

Here, we leverage our tools to construct MOFs (ToBaCCo [73, 77]) and to calculate charges that describe well electrostatic interactions (MBBB charges [204]) to study the adsorption of CO_2 and CO_2/N_2 and CO_2/H_2 mixtures at conditions relevant to post- and pre-combustion CO_2 capture in a complete population of MOFs. Our MOF population contains all possible topology-functionality combinations for 16 topologies and 13 functionalized building blocks (Figure 4.1), which allow us to comprehensively examine the role of pore topology and chemistry on the CO_2 capture capabilities of these open framework materials. Our analysis encompasses (*i*) DFT calculations to directly interrogate the "inherent" interaction of CO_2 with the various building block chemistries, (*ii*) grand canonical Monte Carlo (GCMC) simulations to predict how these inherent interactions translate into different adsorption loadings and adsorption mechanisms, and (*iii*) training of various machine learning algorithms to predict CO_2 capture capabilities, which in turn reveal the relative importance of material properties (such as those related to pore topology and chemistry) without "human bias."

4.3 Computational Methods

4.3.1 MOF Construction

Perfect crystals for over 400 MOFs were computationally synthesized using ToBaCCo [73, 77]. For MOF construction, ToBaCCo requires a library of MOF building blocks and a set of topological blueprints onto which to map the former. The building blocks used were based on four inorganic nodes typically found in synthesized MOFs (Zn_4O [262], Cu_2 [114], $Zr_6O_4(OH)_x$ [263], and Cr_3O [264]) and 12 functionalized derivatives of the common benzoic dicarboxylate (BDC) linker (Figure 4.1) plus the un-functionalized BDC linker. The derivative linkers feature amino, bromine, cyanine, fluorine, hydroxyl, nitro, methyl, thiol, and trifluromethyl groups. Sixteen topological blueprints were used to obtained MOFs in the **acs**, **bcs**, **bcu**, **fcu**, **flu**, **ftw**, **lvt**, **nbo**, **nia**, **pcu**, **pyr**, **reo**, **rhr**, **rht**, **ssb**, and **stp** topologies. The structures generat-ed by ToBaCCo, were optimized with the Forcite module of Materials Studio using the Universal Force Field [80]. These structures were used by some of us to demonstrate the use of DFT-calculated MBBB charges in ref. [204]).

4.3.2 CO_2 Binding to Functionalized Building Blocks

DFT calculations to optimize adsorbate binding configurations around functionalized building blocks used to construct the studied MOFs were performed with the Gaussian09 code [265]. The M06-L functional [266, 267] was used to model electron exchange-correlation and the Def2SVP basis set [268] was used to construct the solution to the Kohn-Sham equations. This combination of functional and basis set has been used to study the energetics of adsorption and reactions in MOFs [269]. Counterpoise correction

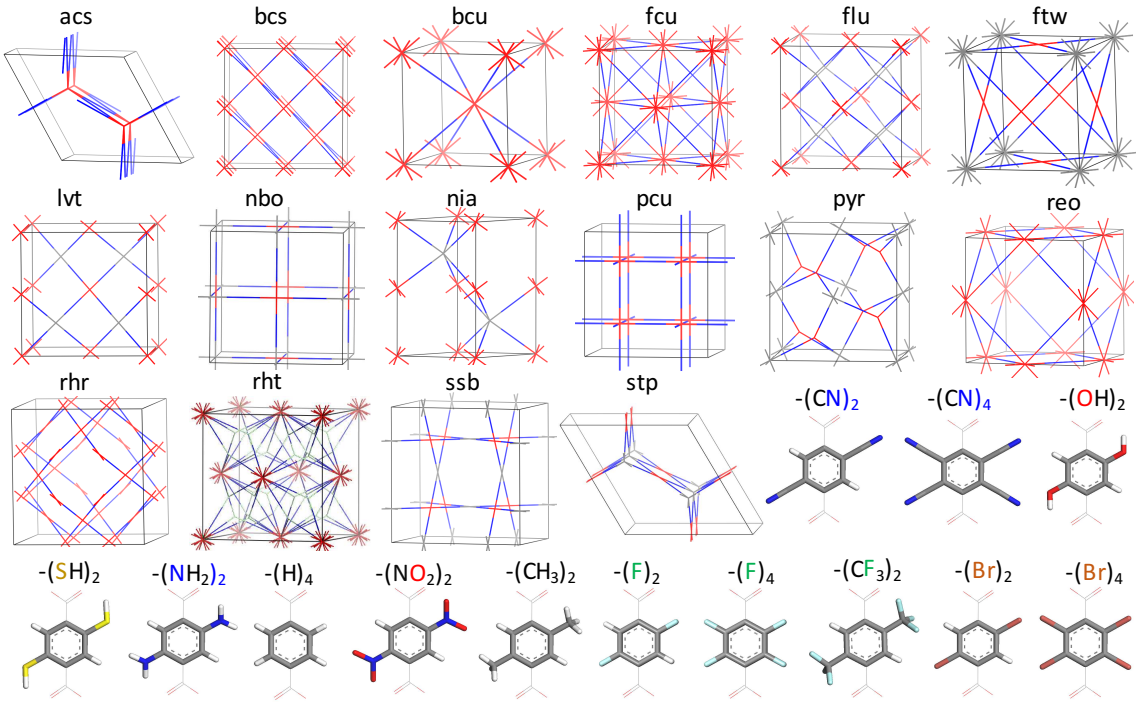

Figure 4.1 Topological blueprints and connecting molecular building blocks used in the construction of the over 400 MOF designs studied in this work. The $-(H)_4$ case corresponds to the non-functionalized linker with which parent MOFs constructed.

[270] was used to account for basis set superposition error when calculating binding energies. The binding energy, E_{binding}, for each adsorbate building block complex was obtained as:

$$E_{\text{binding}} = E_{\text{complex}} - E_{\text{adsorbate}} - E_{\text{building block}} \tag{4.1}$$

where E_{complex} is the energy of the complex, $E_{\text{adsorbate}}$ is the energy of the adsorbate and $E_{\text{building block}}$ is the energy of the building block. Several initial configurations were tried for adsorbate interactions with the functional groups (Figure B.2). Generally, the final configuration was geometrically and energetically similar regardless of the initial configuration. A similar scenario occurred for calculations exploring the interaction of adsorbates with the phenyl ring via π-type interactions (Figure B.5 to Figure B.7). All the DFT calculated interaction energies are summarized in Figure B.8.

4.3.3 Adsorption in MOFs

Adsorption loadings in MOFs at 313 K were calculated via grand canonical Monte Carlo (GCMC) simulations using the RASPA code [91]. The general CO_2 adsorption capabilities of MOFs were evaluated by simulating the adsorption of pure CO_2 at 1 bar. MOF CO_2 capture metrics relevant to pre-combustion carbon capture were evaluated by simulating adsorption for a 20:80 CO_2:H_2 mixture at 20 bar (adsorption conditions) and pure CO_2 at 1 bar (desorption conditions) as used in previous work [238]. MOF CO_2 capture metrics relevant to post-combustion carbon capture were evaluating by simulating adsorption for a 15:85 CO_2:N_2 mixture at 1 bar (adsorption conditions) and a 90:10 CO_2:N_2 mix-ture at 0.1 bar (desorption conditions). Note that the gas-phase compositions during desorption conditions is an approximation based on the assumption of high CO_2/N_2 selectivity during adsorption stage. This assumption breaks down when a given material is not selective enough, which is manifested in a negative working capacity. GCMC simulations used 3500 cycles for equilibration and 3500 for data collection. A cycle has N Monte Carlo moves (translation, rotation, insertion, deletion, and identity change), where N is the largest number between the number of adsorbates in the simulation supercell and 20.

Intermolecular interactions were described by a Lennard-Jones (LJ) plus Coulomb potential. LJ parameters and charges for CO_2 and N_2 were assigned according to the TraPPE model [271]. LJ parameters and charges for H_2 were as-signed according to the Darkrim-Levesque force field [272]. MOF LJ parameters were assigned according to the Universal Force Field [80], whereas MOF atomic charges corresponded to DFT-calculated molecular building block-based (MBBB) charges [204]. Parameters for cross-interactions were obtained using Lorentz-Berthelot mixing rules. Ewald summation was used for electrostatic interactions beyond 12.8 Å, whereas dispersion interactions beyond a 12.8 Å cutoff were neglected.

4.3.4 MOF Properties

Surface area was geometrically calculated by rolling a nitrogen-sized (σ = 3.71 Å) spherical probe on MOF atoms. The surface area corresponded to the area of the surface traced by the center of the probe [273]. Void fractions were calcu-lated using the Widom insertion method with a helium atom as probe (σ = 2.64 Å, ϵ/K = 10.9) [274]. Largest and limiting pore diameters (LPD and PLD, respectively) were calculated using zeo++ [85]. MOF Coulomb charges were DFT-calculated MBBB charges assigned as described in ref. [204]. Many descriptors such as most positive and most negative charges (MPC and MNC) in functional groups and frameworks, and highest bond dipole moment (HBDM) in functional groups are obtained by direct (automated) inspection of the MOF CIFs. The selectivity (S) of the MOF for CO_2 over N_2 or H_2 was calculated according to Equation 1.3. The CO_2 working capacity (WC) was calculated as the difference between CO_2 loading at the adsorption conditions and the CO_2 loading at the desorption conditions (see above).

4.3.5 Training of Machine Learning Models

Data analysis and visualization, as well as the training of machine learning algorithms were done using the software environment for statistical computing and graphics, R [275, 276]. All machine learning models were trained by random assignation of data points from each class into a training (80% of the data) and test set (20% of the data). That is, points were randomly chosen from within each class to ensure members of each class where present in both the training and test sets. On the training set, each algorithm was trained through a parameter grid search (where relevant). The accuracy of the model at each grid-point was evaluated using 5 times repeated 10-fold cross-validation. This grid search and cross-validation procedure was performed using the R package caret. This training procedure is summarized in Figure B.1.

The final parameters for each algorithm were those resulting in the highest accuracy. That is, those minimizing the ratio of correct classifications to incorrect classifications for classification models, and those minimizing the root mean square error (RMSE) for regression models. Once each machine learning algorithm was trained, its validity was con-firmed by evaluating its performance on the corresponding test set. In all cases, the accuracies of the algorithms on the test and training set were similar, consistent with the absence of overfitting. For regression, we trained and validated six different machine learning algorithms: multiple linear regression (MLR), support vector machine (SVM), conditional inference decision trees (DT), random forests (RF), neural networks (NN), and gradient boosting machines (GBM). The R packages used for each algorithm are shown in Appendix B.

4.4 Results and Discussion

4.4.1 Inherent Adsorbate-Adsorbent Interactions

We start by interrogating the inherent interactions of CO_2 with various building block functionalities via DFT. Optimal binding configurations for CO_2 with the functionalities of the building blocks (Figure 4.2) are consistent with interactions that are primarily electrostatic. This is indicated by how the CO_2 molecule aligns with its positively charged C atom close to the most negatively charged atom of each functionality. Meanwhile, the negatively charged O atoms of the molecule aligns with the positively charged atoms of the functionalities. This alignment pattern suggests more importance for the interactions between the positively charged C atom of CO_2 and the negatively charged atoms of the functionalities such as S, O, N (in -NH_2 and $(CN)_x$), F and Br. On the other hand, the lack of obvious alignment patterns for the corresponding optimal binding configurations for N_2 and H_2 (Figure B.3 and Figure B.4) are consistent with a negligible role of electrostatics for the adsorption of these species.

The binding strength of CO_2 with the various chemical functionalities follows the order $(NH_2)_2$ > -$(CN)_4$ > -$(SH)_2$ > -$(NO)_2$ > -$(OH)_2$ > -$(CN)_2$ > -$(F)_4$ > -$(Br)_2$ > $(CF)_3$ > -$(F)_2$ > $(CH)_3$ > -$(H)_4$ > -$(Br)_4$. This order is highly correlated with the order based on the highest bond dipole moment (HBDM) within the functional group as indicated by a Spearman ranking correlation coefficient (SRCC) of 0.73. As expected, CO_2 is generally found to interact with the functional groups more strongly than N_2 and H_2, particularly for groups with larger HBDMs. H_2 generally interacts more weakly (by approximately 2 kJ/mol on average) with the functionalities than N_2 does, so the difference in binding strength between CO_2 and H_2 is generally more pronounced (by 2 kJ/mol on average) than the difference in binding strength between CO_2 and N_2. Not surprisingly, CO_2 selectivities over H_2 are known to be higher than over N_2.

4.4.2 General Correlation of Electrostatic Interactions with CO_2 Sorption

If functional groups are the primary CO_2 binding site and the binding is dominated by electrostatics, then there should be a general correlation between DFT-calculated binding energies and predicted CO_2 loadings from GCMC sorption simulations. This indication of a multiscale link is, of course, contingent to LJ parameters and Coulomb charges in the simulations describing the relevant interactions reasonably well. We expected this to be the case with the use of our recently developed MBBB charges [204]. Again, only a general correlation is expected because the CO_2 load-ing of each individual MOF is likely to be simultaneously influenced by factors beyond inherent interactions (e.g. pore size, surface area and so forth). Therefore, to see if the correlation emerged from our simulation data, we plotted the normalized averaged CO_2 loading at 1 bar for all materials featuring the same functionalized building block versus the normalized DFT-calculated binding energy for CO_2 to that particular building block. Normalization of

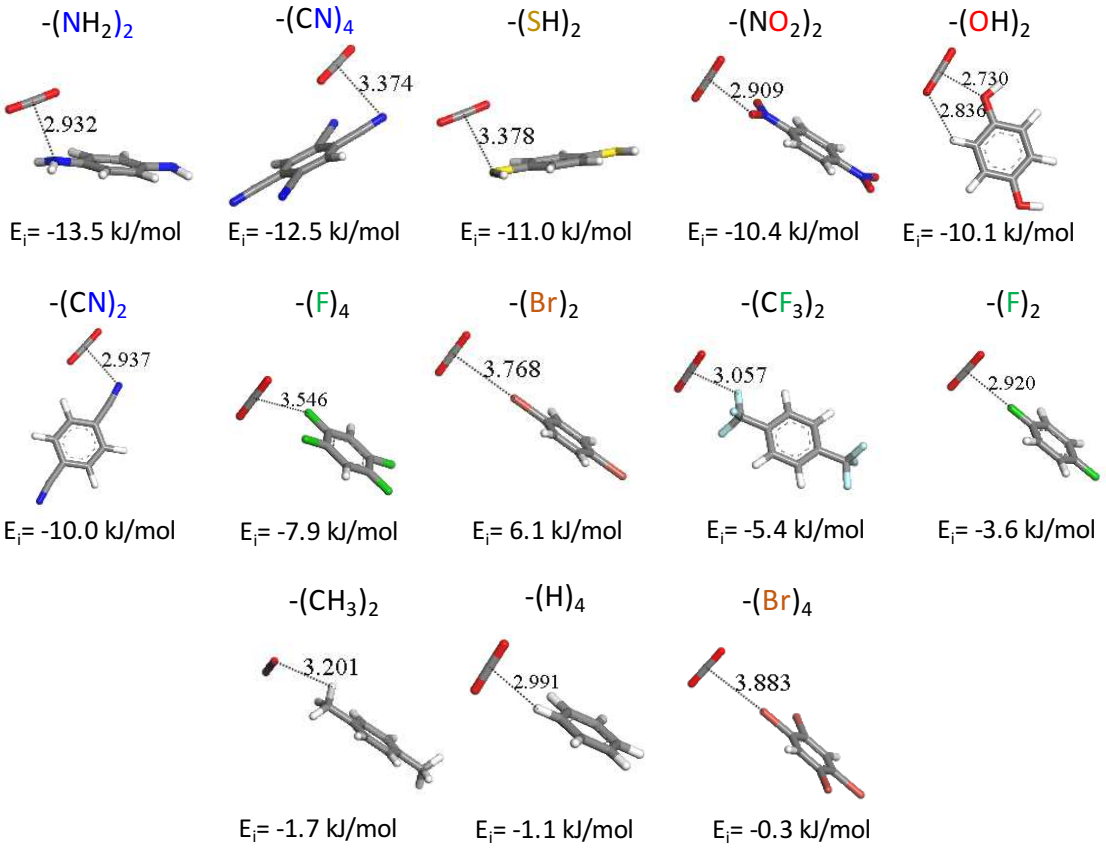

Figure 4.2 DFT-calculated optimal binding configuration between CO_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

loadings was done such that the largest and smallest loading from all simulations at same conditions would be zero and one respectively. Normalization of binding energies was done such that the strongest and weakest binding energies would. Notably, with the exception of the two outliers—which will be discussed in the next subsection—the expected correlation is evident in Figure 4.3 and indicated by a SRCC of 0.610. We obtained similar plots (Figure B.9) based on predicted CO_2 loadings for mixture adsorption, 15:85 CO_2/N_2 at 1 bar and 20:80 CO_2/H_2 at 20 bar, which reveal similar correlations (and outliers).

To confirm that the observed general correlation was tied to the description of electrostatic interactions via MBBB charges, we obtained additional parity plots relating normalized averaged CO_2 loadings and normalized DFT energies, but with the CO_2 loadings predicted with zero Coulomb charges at each MOF atom (Figure 4.3). Indeed, when zero charges are used in our adsorption simulations, the correlation notably worsens (SRCC becomes 0.022). Moreover, in further analysis, we found a stronger correlation of DFT-calculated CO_2 binding energies with the electrostatic interaction energies than with the dispersion interaction energies calculated as shown in Figure B.10. This is not to say that dispersion interactions are not important to drive CO_2 sorption (after all, in most simulations the magnitude of dispersion interactions is larger than that of electrostatic interactions), but to say that electrostatic interactions seem to be the differentiating factor across structures and critical to boost CO_2 sorption.

4.4.3 Variability of Adsorption Mechanisms

The two largest outliers in Figure 4.3 are the $-(H)_4$ and $-(CH_3)_2$ cases ($-(Br)_4$ and $-(F)_2$ can be considered outliers as well). While CO_2 interactions with these functionalities is weak, the average relative loading for all structures featuring these groups is much higher than expected based on binding energies. We thought this could be an indication of a different adsorption mechanism that comes about because the interaction with the functional group is not as relevant, meaning the CO_2 binds to other sites within the MOF. To reveal potential differences in the distribution of CO_2 inside the MOF depending on functionality, we decided to analyze probability density maps obtained by averaging adsorbate configurations during GCMC simulations.

Figure 4.4 shows the density maps for the $-(H)_4$, $-(CH_3)_2$, and $-(CN)_4$ cases of the **pcu** topology (IRMOF-1 analogues), while Figure B.11 shows other examples. The differences between the outlier cases ($-(H)_4$ and $-(CH_3)_2$) and the $-(CN)_4$ case supports the occurrence of different adsorption mechanisms depending on CO_2 binding strength. Indeed, in the $-(H)_4$, and $-(CH_3)_2$ cases of Figure 4.4, CO_2 preferentially adsorbs at the nodes of the framework. There, the linkers "converge" and the overlap of potentials create adsorption sites (π pockets) [277] stronger than the sites adjacent to the functional groups due. On the other hand, for the $-(CN)_4$ case, the sites adjacent to the functional groups seem to be

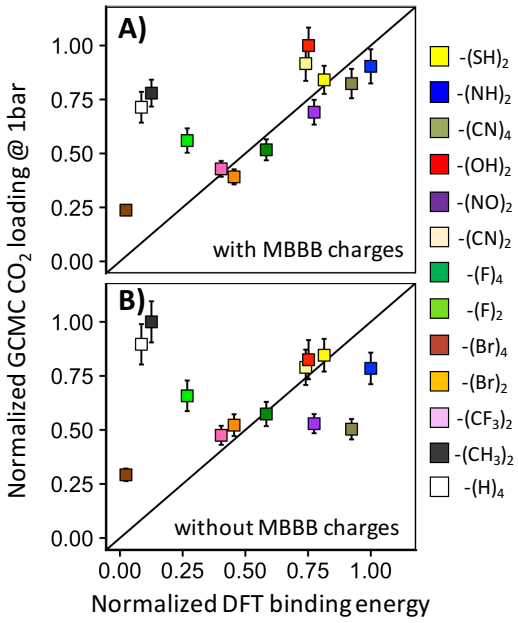

Figure 4.3 Parity plots for GCMC-calculated CO_2 loading at 1 bar vs. DFT-calculated CO_2 binding energies. The plots relate the normalized, aver-age GCMC-calculated CO_2 loading (error bars indicate standard error of the mean) for structures featuring the same functionalized building block against the normalized DFT-calculated CO_2 binding energy with the corresponding building block. (a) Simulations with MBBB charges (SRCC = 0.610, R^2 = 0.387) and (b) simulations with no charges (SRCC = 0.022, R^2 = 0.004).

stronger adsorption sites and CO_2 preferentially adsorbs at these locations. Particularly strong sorption sites seem to be those that allow CO_2 to interact with two functional groups simultaneously, where the latter scenario is facilitated or hindered depending on the topology. Note that the switching in mechanism is also consistent with the binding strength of CO_2 with the functional group relative to that of CO_2 with the phenyl ring (Figure B.8).

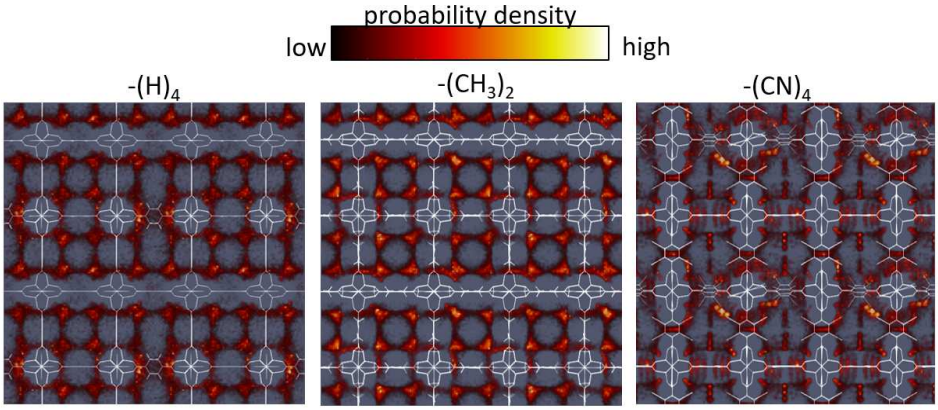

Figure 4.4 Probability density maps showing average CO_2 loading distribution in three analogous structures with different functionalization. The -$(H)_4$ and –$(CH_3)_2$ cases are outliers in Figure 4.3, while the –$(CN)_4$ is not. The maps indicate a different adsorption mechanism depending on the interaction strength between CO_2 and the functionalities.

4.4.4 On the Roles of Pore Chemistry and Topology

Now we start to scrutinize how different factors impact CO_2 capture metrics. All the discussion up to this point illustrates that the functionalization does significantly affect CO_2 adsorption. However, we reiterate that it is not the only relevant factor, and its relative importance depends on other material properties. Therefore, while in Figure 4.3 and Figure B.9 we captured the average effect of functionalization on CO_2 loadings, now we take a more detailed look on how functionalization affects different parent MOFs of various topologies. Based on CO_2:N_2 and CO_2:H_2 mixture adsorption simulation data relevant to post- and pre-combustion CO_2 capture, Figure 4.5 illustrates the improvement (or decline) of CO_2 loading and CO_2 selectivity when 31 parent MOFs were functionalized. Since functionalization increases the density of the material, for it to have a net positive effect on the reported gravimetric loadings, the interaction with CO_2 must be boosted significantly. Figure 4.5 shows that some functionalizations result in pore chemistry that typically enhances CO_2 loadings relative to the parent MOF, whereas some other functionalizations usually have the opposite effect. Specifically, with few exceptions, functionalization of the parent MOF with -$(OH)_2$, -$(SH)_2$, -$(NH_2)_2$, -$(CN)_X$, or -$(NO_2)_2$ significantly enhanced the CO_2 loading of the original structure, whereas functionalization with -$(CH)_3$, -$(F)_X$, -$(CF)_3$, -$(Br)_X$ either barely improved or notably

reduced the CO_2 loading. With "good" functionalization, CO_2 loadings of the parent MOF for the CO_2:N_2 and CO_2:H_2 mixture cases can increase as much as 1 mol/kg and 4 mol/kg, respectively. However, with "bad" functionalization, CO_2 loadings can decrease as much as 0.8 mol/kg and 6 mol/kg, respectively. The selectivity for CO_2, on the other hand, seem to generally im-prove with functionalization regardless of the chemistry introduced to the parent MOF. Selectivities were improved up to 6 and 200 units for the adsorption of CO_2:N_2 and CO_2:H_2 mixtures, respectively.

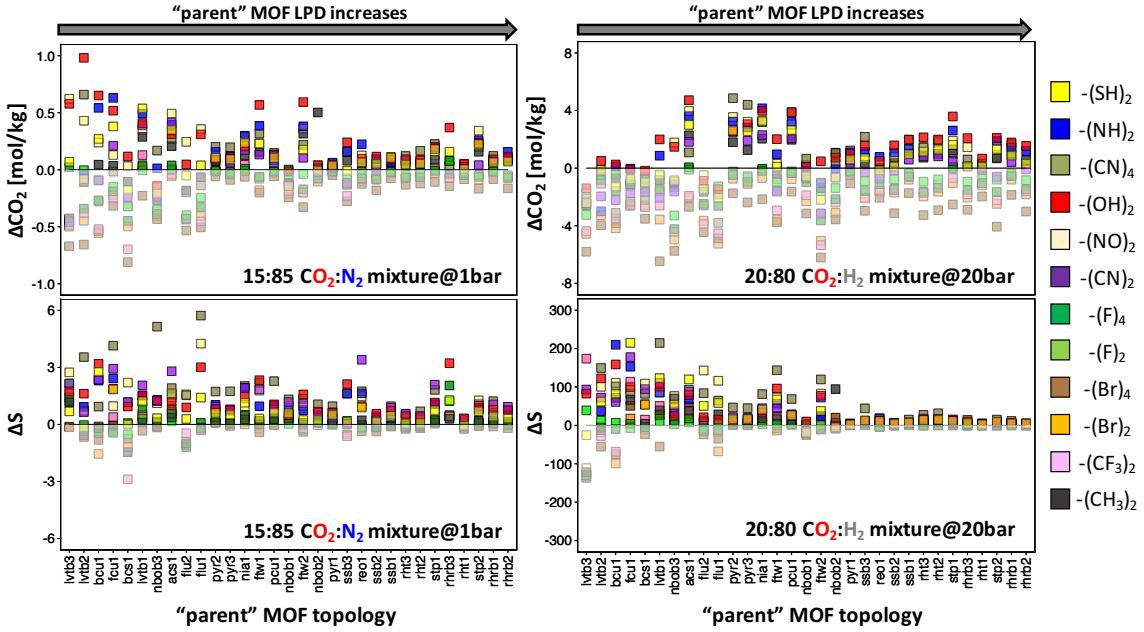

Figure 4.5 Effect of MOF functionalization (alteration of pore chemistry) of parent MOFs of different topologies on CO_2 loadings and selectivities. Results are presented for adsorption of a 15:85 CO_2:N_2 mixture at 1 bar (left) relevant to post-combustion CO_2 capture, and for a 20:80 CO_2/H_2 mixture at 20 bar (right) relevant to pre-combustion CO_2 capture. ΔS and ΔCO_2 are the changes in selectivity and CO_2 loading, respectively, due to functionalization.

While it is clear that the functionalization of a parent structure can improve its CO_2 capture capabilities, it is also clear from Figure 4.5 that the extent of improvement can significantly depend on topology. That is, not all topologies respond equally to functionalization. For instance, Figure 4.5 shows that certain parent MOFs (of certain topologies) were not susceptible to significant improvement regardless of functionalization. In some cases, such as that of CO_2 selectivity over H_2 during sorption of CO_2:H_2 mixtures, it is clear that topologies associated with relatively larger pores for the same connecting building block length (e.g. **rht**, **stp**, **rhr**, **ssb**) are the ones that experience little improvement. How-ever, for other topologies, it is not immediately apparent from Figure 4.5 why functionalization does not change the per-formance of the parental MOF as much. As an example, with functionalization of the **rht** topology,

CO_2 loading and selectivity for adsorption of CO_2:N_2 mixtures stay within 0.1 mol/kg and 0.5 units, respectively, of the parent MOF, functionalization of an **rhr** structure with larger pores, affects the parental structure more significantly. This is actually consistent with the idea that the topology "encodes" a lot of information about the structure (e.g. pore shape, pore polydispersity, etc) that may not be captured by commonplace "numerical" descriptors. This supports the use of topology as a design variable. Data complementary to Figure 4.5 are shown in Figure B.13 and Figure B.14, which show the effect of functionalization on: the same postcombustion metrics shown in Figure 4.5 when all framework charges are zero, and for other conditions, respectively.

4.4.5 Predicting the Impact of Functionalization on Altering CO_2 Capture Metrics of Parent MOFs using Decision Trees

To economize resources, an experimentalist may want to know how likely it is that a parent MOF will improve upon functionalization before attempting a complicated process to obtain the functionalized version of a parent MOF. While computational synthesis and molecular simulations can provide the answer, it is beneficial to provide experimental-ists with more accessible tools to makes such prediction simulation-free. Ideally, on the basis of simple intuitive mate-rial features (i.e. descriptors). Such a tool could also show what the most important chemical and topological/structural factors to be considered are. To illustrate this, we trained conditional inference decision trees (DTs) to predict whether CO_2 capture metrics of a parent MOF would improve with functionalization. The proposed descriptors were selected by "human intuition", but it was the training algorithm what decided which of these descriptors were important. Descriptors were proposed based on the factors with potential relevance to CO_2 capture based on the results discussed up to this point, focusing on pore chemical and topological/structural features. Specifically, as topology/structure-related factors we included the topology, void frac-tion (VF), largest pore diameter, pore limiting diameter, gravimetric surface area (GSA), and density of the parent MOF. As chemical descriptors, we included the functional group HBDM, the functional group most negative charge (MNC), the functional group most positive charge (MPC), the functional group sum of epsilons (SE), defined as $\sum_{i=1}^{N} \varepsilon_i$ where N is the number at atoms in the functional group, and ε_i is the LJ epsilon of the ith functional group atom, the functional group number density (NFG per volume), as well as the parent MOF most positive and most negative charges.

By limiting the number of descriptors, we obtained easy-to-interpret decision trees. Figure 4.6 shows the tree for CO_2 selectivity over N_2, while Figure B.15 to Figure B.18 show the DTs for CO_2 loading at 1 bar, CO_2 selectivity over H_2 and for CO_2 working capacities for the CO_2:N_2 and CO_2:H_2 mixture cases. Despite their simplicity, the trees do a good job at predict-ing whether improvement of a certain metric can

be expected (with all cross-validation and test set accuracies higher than 80%). Given that decision trees are an easily visualized "glass-box" model, it is possible to discern how the DTs make decisions and thus determine important features that control whether a parent MOF is improved or worsened by functionalization. Specifically, since the DTs follow a "greedy algorithm," descriptors nearer the root of the tree (the node for the first split) can be interpreted as being more important.

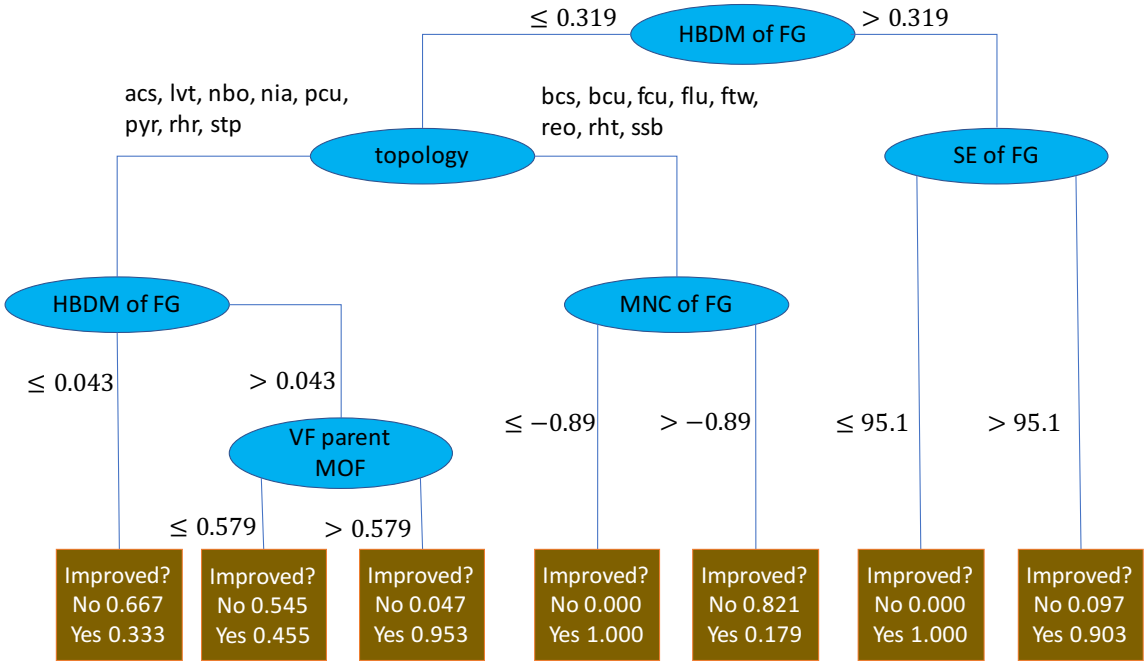

Figure 4.6 Decision tree trained to predict whether parent MOF functionalization would improve the CO_2 selectivity over N_2 of the parent MOF during exposure to a 15:85 CO_2:N_2 mixture at 1 bar. The numbers at the terminal nodes are fractions, but can also be inter-preted as the probability that a MOF falling in that node is improved or not. Variables appearing in this tree are the highest bond dipole moment (HBDM) in functional group, sum of epsilons (SE) in functional group, most negative charge (MNC) in functional group, void fraction (VF) of parent MOF, and topology.

As an example, let us consider the tree for CO_2 selectivity over N_2 in Figure 4.6. The first split of the tree is made based on the HBDM of the functional group that would be introduced to the parent. If the HBDM is larger than 0.319 eÅ, the functionalization does improve the parent in more than 90% of the cases. If the HBDM of the functional group is smaller than 0.319 e.Å, then the topology is the next most significant factor that come into play. In that case, for **bcs**, **bcu**, **fcu**, **flu**, **ftw**, **reo**, **rht** and **ssb** topologies (note that first five are based on zirconium oxoclusters), the improvement upon functionalization occurs if the functionality has an atom more negatively charged than -0.89. For **acs**, **lvt**, **nbo**, **nia**, **pcu**, **pyr**, **rhr** and **stp** topologies, the improvement on functionalization occurs in 95% of the cases when the HBDM is not so low (> 0.043 eÅ) provided that the void fraction of the parent MOF is large enough (> 0.579).

Similar analyses can be readily made with the decision trees shown in Figure B.15 to Figure B.18 for other performance metrics. We find that factors related to the changes in the pore chemistry (e.g. MNC and HBDM of the introduced functionalities) are the most relevant factor to determine whether four of the five studied CO_2 capture metrics are improved. The exception was the improvement of the CO_2 selectivity over H_2, for which the most relevant factor was surprisingly related to the structural properties (i.e. gravimetric surface area) of the parent MOF. That said, both MNC of the functional group and the parent MOF topology were descriptors appearing in all trees.

4.4.6 Correlations Between the Improvement of Different Carbon Capture Metrics

An interesting question to address is whether the improvement of one metric by functionalization implies the improvement of all other metrics, or whether trade-offs are involved. For instance, an obvious correlation to check is whether the improvement of CO_2 loading for adsorption of a CO_2:N_2 mixture implies an improvement on CO_2 selectivity over N_2. However, we also check other correlations, which are summarized in the correlation matrices in Figure 4.7. Generally speaking, the matrices illustrate that there is more correlation for the effect of the functional groups on the different CO_2 capture metrics than for the effect of the topology. This indicates that if functionalization with, for in-stance, -(OH)2 is generally good to improve the CO_2 loading at 1 bar (i.e. $-(OH)_2$ improves CO_2 loading in 90% of cases), it is also generally good to improve CO_2 selectivity over N_2 (SRCC = 0.78). This analysis also holds for the other functional group considered, whether they tend to improve or reduce performance. The rankings used to generate Figure 4.7 can be seen in Figure B.12. The strongest correlation is for the effect of functionalization on CO_2 loading at 1 bar and CO_2 working capacity for CO_2:N_2 case (SRCC = 0.98). This is in contrast with topological considerations, where the significantly weaker correlations indicate that a topology that is generally amenable to improve one metric via functionalization is not necessarily amenable to improvement in other metrics. Thus Figure 4.7 quantitatively supports qualitative observations from Figure 4.5 and suggests that MOF topology is critical as a design parameter, and thus must be selected carefully.

4.4.7 Prediction of CO_2 Capture Metrics using Machine Learning

Ultimately a synthetic chemist desires to know whether synthesizing a material with certain topological and chemical features would have outstanding CO_2 capture capabilities. Maintaining the theme of providing simulation-free tools (provided here as supplementary material) to predict CO_2 capture metrics, we now leverage our understanding of what factors strongly affect the improvement of CO_2 capture metrics to train machine learning algorithms to predict absolute metrics using simple but meaningful topology- and chemistry-related descriptors. We evaluated the performance decision trees (DTs) (analogous to those

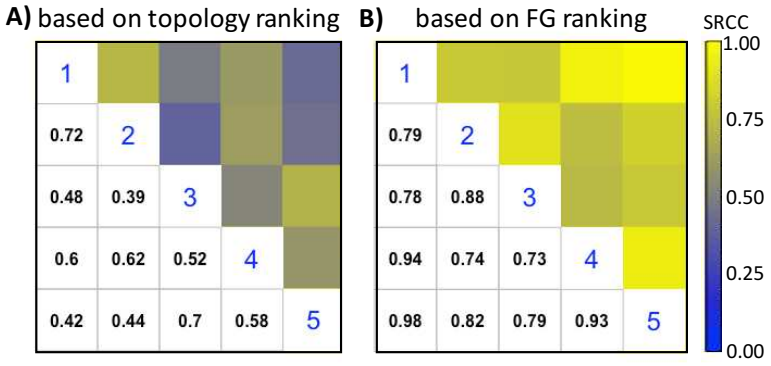

Figure 4.7 Correlations among improvements of different CO_2 capture metrics. 1 = CO_2 loading at 1 bar, 2 = CO_2 selectivity over H_2, 3 = CO_2 selectivity over N_2, 4 = CO_2 working capacity during adsorption of CO_2:H_2 mixture, 5 = CO_2 working capacity during adsorption of CO_2:N_2 mixture (a) matrix of SRCCs relating the rankings of topologies for each metric. A higher ranking means functionalization of the parent of such topology improved the metric in question more frequently. (b) Matrix of SRCCs relating the rankings of functional groups (FGs) for each metric. A higher ranking means functionalization of parent MOF with such FG improved the metric in question more frequently. Matrices are symmetric with above-diagonal positions color coded according to the color bar.

discussed earlier but now used for regression) as well as random forests [22], neural networks (NNs) [53], support vector machines (SVMs) [94], and gradient boosted machines (GBMs) [278]. To our knowledge, the latter method has not been tested to predict adsorption properties of MOFs. To train these models, we used analogous descriptors to those we used in our trained classification models. The only difference is that instead of topological/structural properties of the parent MOF, we used those of the functionalized materials. To provide a reference for the accuracy of the more robust models, we also trained simple multiple linear regression (MLR) models.

Figure 4.8 compares the predictions for CO_2 selectivity over N_2 by the different machine learning algorithms against the predictions by GCMC adsorption simulations. Analogous plots for the other metrics can be found in Figure B.19 to Figure B.22. The algorithms were trained on "standardized" variables (variables with a mean of zero and a standard deviation of unity), and the RMSE from 5 times repeated 10-fold cross-validation was calculated based on these standardized variables. This allows a fair comparison between the prediction accuracies of the different CO_2 capture metrics. On this RMSE basis, we found metrics related to the sorp-tion of CO_2:N_2 mixtures to be somewhat more difficult to predict than the others. Also on this RMSE basis, the most ac-curate predictions came from the GBM models (Figure B.23), whose predictions were more accurate than even the predictions by the NN models. Given the general appeal of NNs, it is interesting to note that the training of the GBMs was significantly less computationally demanding than the training of the NNs. Accurately ranking structures based on their performance metrics is key in "hierarchical" HTCS of large databases—where non-promising structures are

filtered out during screening rounds of increasing accuracy—so no "top-performer" materials are missed in subsequent screening rounds. The ranking ac curacy of each algorithm on the corresponding test set was determined by the SRCC, which again was closer to one for the GBM model in the case of CO_2 selectivity over N_2 (SRCC = 0.921) and for the remaining four metrics.

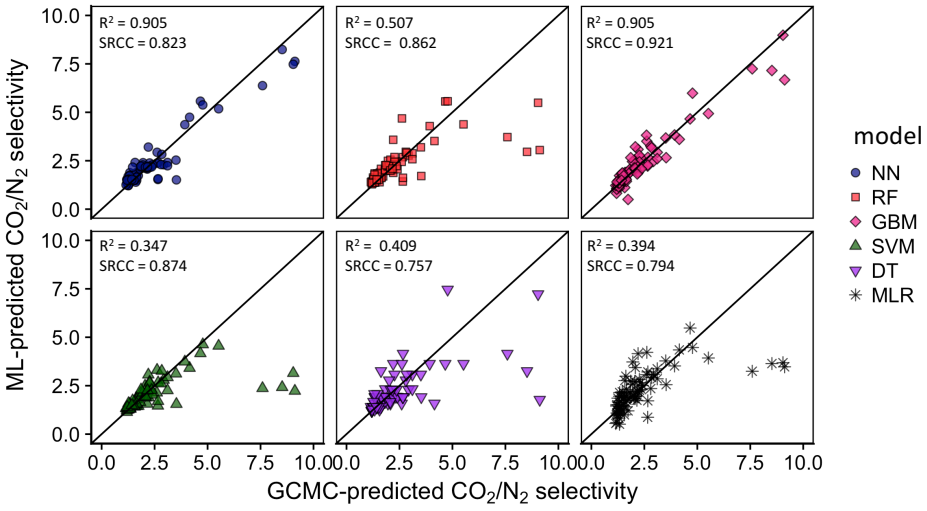

Figure 4.8 Parity plots comparing the prediction of CO_2 selectivity over N_2 by various trained machine learning algorithms versus the prediction from actual molecular (GCMC) simulations on the test set. The inset numbers indicate the R^2 and the Spearman ranking correlation coefficient (SRCC). MLR = multiple linear regression, DT = conditional inference decision trees, RF = random forests, GBM = gradient boosted machines, SVM = support vector machines, NN = artificial neural networks.

4.4.8 Relative Importance of Chemistry- and Topology-Related Factors on Absolute CO_2 Capture Metrics

The above results show that machine learning can be used to predict adsorption properties using simple, intuitive descriptors, but the focus of this work is on understanding the relative importance of properties related to pore chemistry and structural (topology) on the CO_2 capture capabilities of MOFs. To this end, from the training of the GBM, we can determine the relative importance of different proposed descriptors of the CO_2 capture capabilities of the studied MOFs without "human bias" in the analysis. We can do this by looking at the individual conditional expectation (ICE) plots, which can be used to visualize the partial and average response of the GBM model to variation in the individual descriptors used as model inputs (Figure B.24 to Figure B.28). Additionally, Figure 4.9 shows how much the different descriptors contribute to the prediction of the various CO_2 capture metrics. Among the descriptors, MPC of the functionalities or the whole framework seem to have less importance overall, and clearly less importance than the MNC of the functionalities or the whole framework. The seemingly more important

role of the negative charges of the functional groups seems consistent with the binding configurations in Figure 4.2, which show the primary interaction mechanism to be the alignment of the positively charged C atoms of CO_2 with the negatively charged atoms of the functionalities.

Different from the predictions for the improvement of the performance metrics of a parent MOF through functionalization, structural properties seem to play a bigger role to determine absolute performance metrics. For instance, the pore diameter (largest or limiting) is greatly important for all CO_2 capture metrics. Factors related to pore chemistry (e.g. MNC and HBDM in the functional group) seem to be more important for CO_2:N_2 mixtures. Going into more specifics, for adsorption of CO_2:H_2 mixtures, the most important factors determining CO_2 working capacity were the material densi-ty (23%), the pore limiting diameter (14%) and the most negative charge (14%). Meanwhile, for the CO_2 selectivity (over H_2), the most important factors were the gravimetric surface area (28%), largest pore diameter (24%) and pore topology (18%). As for adsorption of CO_2:N_2 mixtures, the most important factor determining CO_2 working capacity was the largest pore diameter (21%). Whilst for the CO_2 selectivity (over N_2), the most important factors were void fraction (21%) and the largest pore diameter (17%).

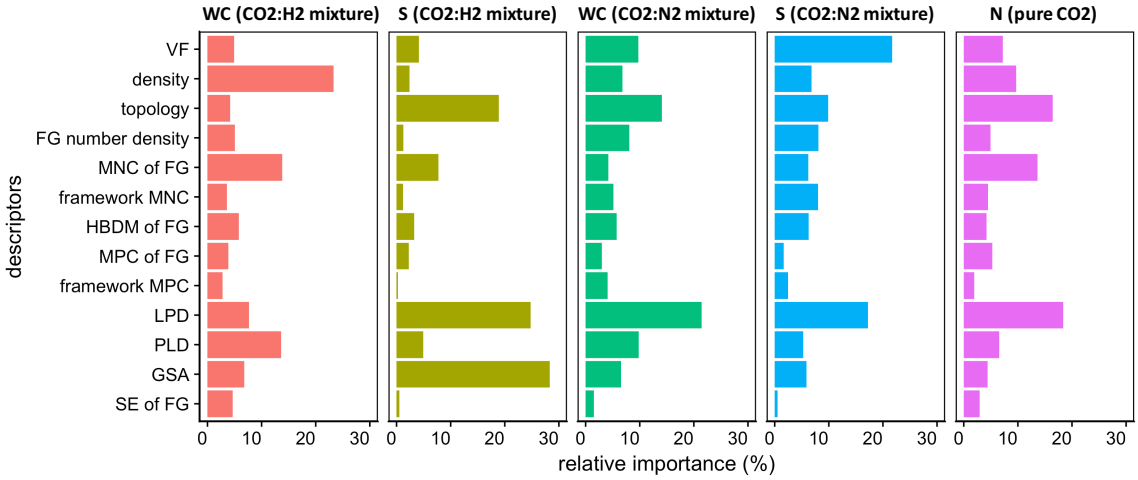

Figure 4.9 Relative importance of material descriptors obtained from GBM training. This plot is complementary to ICE plots shown in Figure B.24 to Figure B.28. S = selectivity, WC = working capacity, N = adsorption loading. FG = functional group, VF = void fraction, HDBM = highest dipole moment, MPC = most positive charge, MNC = most negative charge, LPD = largest pore diameter, PLD = limiting pore diame-ter, SE = sum of epsilons, GSA = gravimetric surface area.

4.4.9 Does Machine Learning Suggest what Chemical and Topological Properties Maximize CO_2 Capture Metrics?

As a final exercise, we used the GBM models to investigate potential optimal chemical and topological property (descriptor) combinations for outstanding CO_2 capture metrics. To do this, we used a genetic

algorithm (GA) from the R package GEoptim to maximize the predictions of all five metrics by our GBM models for each topology, albeit subject to "physical" constraints for the values of individual descriptors (see B). One consequence is that values of individual descriptors did not fall significantly outside the ranges already observed in our data set. However, we did not impose constraints on how descriptor values combine.

The complexity of the relation between metrics and descriptors and the stochastic nature of the GA method meant that different maximization runs did not produce exactly the same value for the metric or the combination of descriptor values associated with it. Thus, we averaged these values for five different runs for each topology (Figure B.29). The GA allowed us to see what property (value) combinations tended to maximize each metric, which we found to be similar for each metric across topology. Therefore, we averaged these descriptor value "patterns" across all topologies and compared these averages between metrics (Figure 4.10) to see what values the GA associates with maximized metrics. We highlight the following from this exercise: (*i*) much higher GSA values were needed to maximize working capacities than to maximize selectivities and pure CO_2 loading, (*ii*) an LPD of 9 Å and PLD of 4 Å could be "trade-off" beneficial for all metrics, (*iii*) Higher HBDMs of the functional groups were needed to maximize metrics for the CO_2:N_2 mixture case than for the CO_2:H_2 one. Note that deviations for some descriptors are significantly larger than for others. The deviations are consistent with the importance of each variable as determined by the GBM model. Essentially, the low-er the importance of a variable, the wider range of valuables it can take without drastically affecting the GA-optimized metric value.

To see how the MOFs in our data set compared to these average optimal property combinations we plotted the maximum and average property values present in our data set together with the projected "optimal" values (Figure B.30 to Figure B.34). For selectivities, the average and maximum GSAs were generally higher than the projected optimum values, with only **bcs**, **bcu**, and **fcu** topologies falling within the desired region, which is consistent with these topologies having some of the highest average selectivities. For all metrics, the average and maximum LPDs and PLDs tended to be higher than the optimal values, with **bcs**, **bcu**, **fcu**, **flu** and **ftw** topologies tending to fall closest.

4.5 Conclusions

We have exploited computational synthesis and charge calculation methods to implement molecular simulation and machine learning to elucidate the role of pore chemistry and topology on various CO_2 capture metrics relevant to post- and pre-combustion CO_2 capture of a "complete" (topology- and functionality-wise) population of MOFs. To provide a solid ground for our analysis, we demonstrated that the use of MBBB charges did establish a multiscale link between DFT calculations and GCMC simulations. We showed that by traditional analysis of adsorption simulations, simple descriptors to train

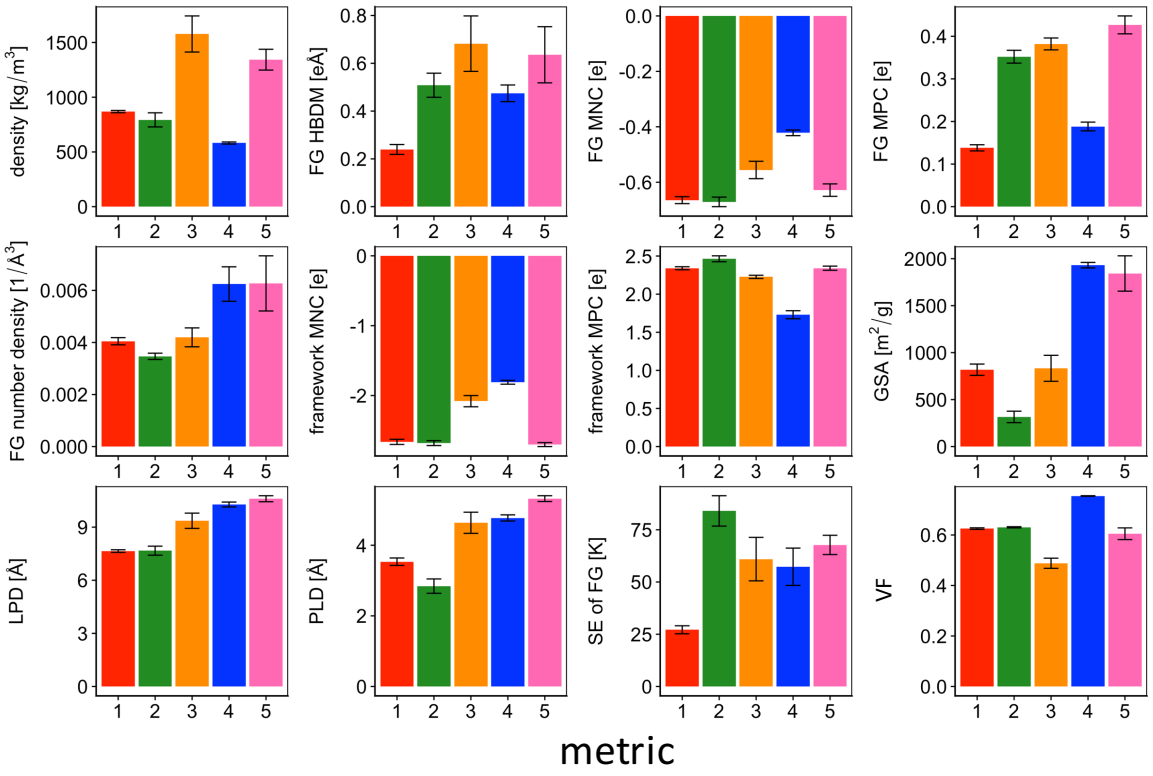

Figure 4.10 Average GA-calculated descriptor values across all topologies giving maxima for different MOF CO_2 capture metrics. Metric indexes are: 1= CO_2 loading at 1 bar (red), 2 = CO_2 selectivity over H_2 (green), 3 = CO_2 selectivity over N_2 mixture (orange), 4 = CO_2 working capacity for the CO_2:H_2 mixture case (blue), and 5: CO_2 working capacity for the CO_2:N_2 mixture case (pink). Error bars correspond to the standard deviation. Descriptors labels are the same as in Figure 4.9.

machine learning algorithms can be proposed by "human intuition." We then demonstrated that machine learning based on those simple descriptors can be an effective simulation-free tool to predict CO_2 capture metrics. This tool can be leveraged by experimentalists looking for guidance to design new improved open framework materials for CO_2 capture or by computational chemists looking to do high throughput screening of massively large databases to discover materials for CO_2 capture. Furthermore, we showed machine learning as a "human bias-free" approach to elucidate the relative importance of chemistry- and topology-related descriptors on material performance. Improvement of CO_2 capture metrics of parent MOFs was found to depend more strongly on chemistry-related descriptors. The absolute values of these metrics, however, were found to depend more strongly on topology-related descriptors. The trained algorithms should be able to predict properties of MOFs that were not included in this study as long as they combine properties that fall within the range of properties used to train the algorithms (Figure B.35). However, it should be noted that while our MOF construction strategy intended to provide a broad range of MOF properties relevant to commonly synthesized MOFs, the "universal" applicability of the trained algorithms must be carefully considered.

CHAPTER 5

PREDICTING SINGLE COMPONENT ADSORPTION ISOTHERMS FOR MULTIPLE MOLECULES IN MOFS USING A SINGLE DEEP LEARNING MODEL

Modified from a paper published in the Journal of Chemical Theory and Computation.[8]

Ryther Anderson[9], Achay Biong[9], Diego Gómez-Gualdrón[9,10]

In this chapter, we demonstrate the first example of a single machine learning model that can predict the isotherms of multiple adsorbates in diverse MOFs without modification. We accomplished this by including variables describing adsorbate properties as part of the model input. As in Chapter 4, the model described herein was trained on GCMC data. Key to the model development model was the use of alchemical adsorbates in the training set, which allowed us to generate data for adsorbates with a well-defined grid of adsorption properties. We expect that future models capable of predicting the loadings of more complex adsorbates (and multicomponent loadings) than those considered here will also be trained on data including alchemical adsorbates. Furthermore, the model developed here is later adapted (in Chapter 7) to provide single-component isotherms as inputs for ideal adsorbed solution theory (IAST) calculations, allowing for the rapid prediction of multicomponent adsorption in diverse MOFs, making it directly useful in assessing MOFs separation performance.

5.1 Abstract

Tailoring the structure and chemistry of metal-organic frameworks (MOFs) enables the manipulation of their adsorption properties to suit specific energy and environmental applications. As there are millions of possible MOFs (with tens of thousands already synthesized), molecular simulation has frequently been used to rapidly evaluate the adsorption performance of a large set of MOFs. This allows subsequent experiments to focus only on a small subset of the most promising MOFs. In many instances, however, even molecular simulation becomes prohibitively time consuming, underscoring the need for alternative screening methods, such as machine learning, to precede molecular simulation efforts. In this study, as a proof of concept, we trained a neural network—specifically, a multilayer perceptron (MLP)—as the first example of a machine learning model capable of predicting full adsorption isotherms of different molecules not included in the training of the model. To achieve this, we trained our MLP on "alchemical" species— represented only by

[8]Reproduced with the permission of: R. Anderson, A. Biong, D. Gómez-Gualdrón. "Adsorption isotherm predictions for multiple molecules in MOFs using the same deep learning model" *J. Chem. Theory Comput* 16, 2, 1271–1283. Copyright 2020 American Chemical Society.

[9]Chemical and Biological Engineering Department, Colorado School of Mines
[10]Corresponding author, dgomezgualdron@mines.edu

variables derived from their force field parameters—to predict the loadings of real adsorbates. Alchemical species used for training were small, near-spherical, and nonpolar, enabling the prediction of analogous real molecules relevant for chemical separations such as argon, krypton, xenon, methane, ethane, and nitrogen. MOFs were also represented by simple descriptors (e.g. geometric properties and chemical moieties). The trained model was shown to make accurate adsorption predictions for these six adsorbates in both hypothetical and existing MOFs. The MLP presented here is not expected to be applied "as is" to more complex adsorbates with properties not considered during its training. However, our results illustrate a new philosophy of training that can be built upon with the goal of predicting adsorption isotherms in not only a database of MOFs, but also for a database of adsorbates, and over a range of relevant operating conditions.

5.2 Introduction

Advanced porous crystals are promising materials in a number of technologies used to mitigate energy- and environment-related problems. For instance, chemical separations requiring large inputs of energy (e.g. cryogenic distillation) could instead be performed using specially tailored porous materials to retain one component selectively (and abundantly) [18, 145, 279], ultimately allowing for separation at relatively mild (i.e. non energy-intensive) conditions [3]. Porous crystals include well-known materials such as zeolites [28], as well as emerging materials such as porous organic cages (POCs) [32], covalent organic frameworks (COFs) [280], and metal-organic frameworks (MOFs) [231]. While crystal tailoring for a specific application is perhaps most readily achieved in MOFs, all these materials exhibit an exceptionally large diversity of chemistries and architectures, stemming from the use of different synthetic precursors [77, 195, 281, 282]. The number of possible synthetic precursor combinations implies an overwhelming number of possible materials, a number that would be impossible to exhaustively synthesize and experimentally test to find optimal candidates for a specific application.

Consequently, molecular simulation has been frequently used to aid the discovery of porous crystals by performing computational experiments [283]. For instance, grand canonical Monte Carlo (GCMC) simulations have been used to predict adsorption capabilities in large material databases [53, 163]. As the development of more accurate descriptions of relevant intermolecular interactions with new forcefields continues, the matching between GCMC and experiments will continue to improve [284–286]. By using GCMC, one can "narrow down" a large database of materials to a smaller set of potentially high performing materials on which to devote experimental efforts [37, 40, 238, 287]. Through this "hierarchical" approach, GCMC has led to the identification of, for instance, NOTT-101 and SBMOF-1 as high-performing MOFs for CO_2/H_2 and Xe/Kr separation, respectively [43, 238].

However, depending on the size of the database, the number and type of adsorbates involved, the operating conditions, and the number of compositions to be tested, even GCMC simulations can become prohibitively computationally intensive for comprehensive screening. This is a critical drawback if one must solely rely on GCMC for screening, especially considering that recent improvements in algorithms used to "computationally synthesize" porous crystals allow for the creation of databases of unprecedented sizes [76, 77, 288]. Therefore, building on the hierarchical screening philosophy, a computational "pre-screening" method that allows GCMC to be devoted only to the most promising materials in a database is not only desirable, but potentially necessary to maintain the efficacy of computational high-throughput screening.

Several methods have been considered for pre-screening databases, including estimation of performance metrics using analytical equations with faster-to-calculate descriptors such as Henry's constants [289–291] and surface areas [54, 191] as inputs. However, perhaps the most intriguing prospect is the use of machine learning predictions for the pre-screening stage. Some of the first efforts using machine learning to predict adsorption were presented by Woo and coworkers, who used support vector machines (SVMs) to predict methane adsorption using crystal textural properties such as void fraction, surface area, and pore size as performance descriptors [94]. The array of descriptor values used to represent a material or molecule for training a predictive algorithm is commonly referred to as a "fingerprint". Woo and coworkers also presented machine learning based predictions of CO_2 adsorption, made with more complex fingerprints (e.g. atomic property-weighted radial distribution functions) as inputs [95]. In other prominent examples, Smit and coworkers used random forests (RFs) and artificial neural networks (ANNs) to predict Xe/Kr [22] and hydrogen adsorption [53], respectively, but with a fingerprint that included some simulation-calculated, energy descriptors. Using simple descriptors instead, Fernandez and coworkers used decision trees (DTs) and SVMs to broadly classify materials for CO_2/N_2 separation (e.g. as "potentially good") [99]. Also using new, but still easily-interpretable descriptors, (e.g. metallic percentage, topology, and the chemical identities of building blocks) Srivastava and coworkers predicted methane adsorption using RFs [100], while Froudakis and coworkers predicted hydrogen and CO_2 adsorption using RFs and SVMs, respectively [292]. Previously, we also predicted loading, selectivity, and working capacity for CO_2 capture from gas mixtures with several different algorithms, finding that the highest accuracy was achieved with gradient boosting machines (GBMs) [293]

The above machine learning efforts have been constrained to generally the same approach: (*i*) GCMC is used to simulate the adsorption of a given adsorbate or adsorbate mixture for materials (e.g. MOFs) in a database at a specific operating condition, and then (*ii*) an algorithm is trained to predict the simulated adsorption data using material properties—i.e. a material fingerprint—as inputs. It is often noted that the final algorithm could be used to screen new adsorbents, which is an endeavor that may be worthwhile if a

new database emerges or the original database grows drastically. However, algorithms trained under this approach can only evaluate new materials for the combination of adsorbates and operating conditions that they were originally trained on. Clearly, this approach severely limits the scope of the predictive algorithms, especially considering that a need to explore the same database for other adsorbates (or adsorbate mixtures) and/or other operating conditions is more likely to arise than a need to explore another database.

Recently, Sholl and coworkers [294] underscored the low diversity of adsorbates so far considered in computational screening by noting that most adsorption studies on material databases focused on CO_2, CH_4 and H_2. This focus is mainly driven by interest in energy storage and carbon capture. However, the potential of advanced porous crystals extends to applications involving a much larger diversity of adsorbates. For instance, current commercial applications of MOFs involve unusual adsorbates such as 1-methylcyclopropene and boron trifluoride [192]. Other potential applications in refrigeration [186], medicine [295], protection against chemical warfare agents [296], and a myriad of chemical separations [297–299], involve many other adsorbates (e.g. CH_3OH, O_2, H_2O, H_2S). Separations relevant to the oil and gas industry can involve complex mixtures of $C_nH_mO_xN_yS_z$ adsorbates [300]. Recognizing the need for faster ways to predict adsorption for a diversity of adsorbates, Sholl and coworkers [294] tried predicting isotherms for 24 adsorbates using the Langmuir model with simulation-calculated Henry's constants and saturation loadings. Two caveats to this approach are its lack of extensibility to non-Langmuir-shaped isotherms, and the need to calculate new Henry's constants and saturation loadings for new temperatures. However, these caveats could be potentially overcome using machine learning.

In recent work [101], we found that a single multi-layer perceptron (MLP), a class of ANN, was able to predict full hydrogen isotherms and isobars, which requires predicting adsorption at temperatures and pressures not included in the training data. That is, the algorithm is required to learn the behavior of loading with changes in temperature and pressure, for a diverse range of materials (and thus isotherm/isobar shapes). In the cited work, we used inherent material properties (similar to those discussed previously), temperature (T) and pressure (P), and the relevant force field parameter describing the "chemistry" of adsorbate/adsorbent interactions as our descriptors. The success of including operational (T and P) and adsorbate dependent descriptors (force field parameters) motivated us to investigate the suitability of machine learning as a tool that could make universal adsorption predictions possible. A prerequisite for such "universal" tool is the ability to predict adsorption for molecules for which it was not originally trained.

Given that there is ongoing debate on the scope of machine learning and the best strategies to train machine learning models even when focused on a specific adsorbate or mixture, a first step toward the

development of a universal model is to study whether the same machine learning model that is used to predict the adsorption of a given adsorbate can actually be used to predict the adsorption of a different adsorbate. Accordingly, the work herein focuses on demonstrating such capability, considering the substantial increase in the complexity of the data that arises when including different adsorbates (even simple ones, as those considered here) along with different operating conditions. Additionally, an underlying theme in our work is to make the machine learning algorithm as accurate as possible while keeping model inputs brief, easily interpretable, and obtainable with minimal computational effort. To generate training data, we focused on the adsorption of 200 alchemical species (i.e. adsorbates that do not necessarily have force-field parameters that replicate the structure, chemistry or physics of any real atom or molecule) at room temperature in a relatively small, topologically and chemically diverse database of 2,400 MOFs created using our Topologically-Based Crystal Constructor (ToBaCCo) code [73, 77]. We tested the model on real adsorbates (Ar, Kr, Xe, methane, ethane, and N_2) partly chosen due to their relevance to several gas storage [191, 224, 295] and chemical separation [18, 301–304], applications. We limited the number of MOFs in our database to keep the number of simulations needed to generate the requisite data reasonable.

5.3 Data Generation

5.3.1 Database Construction

The ToBaCCo-3.0 program [73, 77] was used to "computationally synthesize" 2,400 MOFs of 50 topologies using the building blocks illustrated in Figure 5.1. This selection includes commonly observed oxometallic nodes and a diversity of non-metal chemical moieties found in existing MOFs. The selection of building blocks was made aiming to maximize the diversity of topologies and framework-adsorbate interactions in a relatively small database, which in turn helped us maintain the number of simulations needed to generate training data reasonable. Each MOF prototype constructed by ToBaCCo was structurally optimized in LAMMPS (version 31 Mar 2017) [211] using the Dreiding [83] force field. In Dreiding, functional forms and force constants for bond and angle terms are independent of atom types, but equilibrium bond lengths and angles are unavailable for some metals. In such cases, we used the crystallographic bond lengths/angles from experimental CIFs as the equilibrium bond lengths/angles. For the optimization, we used an iterative approach where in each iteration the atom coordinates were first optimized using the fast inertial relaxation engine (FIRE) algorithm developed by Bitzek et al. [212] with a timestep of 10.0 fs with the MOF cell parameters fixed. Then, the atom positions and unit cell parameters were optimized together using a conjugate gradient algorithm. For each iteration, the first and second step optimizations were stopped when the change in energy between consecutive geometries divided by the

energy of the last geometry was less than 1.0×10^{-6} and no atom experienced a force larger than 1.0×10^{-6} kcal/mol Å-1. The iterations were stopped when the energy change from the previous iteration to the current iteration was less than 1.0×10^{-6} kcal/mol.

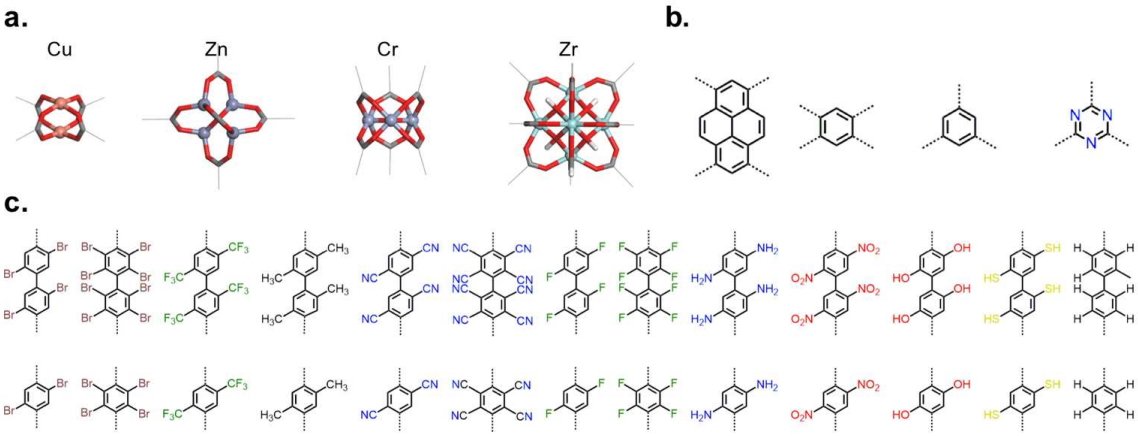

Figure 5.1 The building blocks used for MOF database construction, dashed lines indicate connections to the rest of the framework; (a) inorganic (metal-containing) nodes, which include Cu (4-connected), Zn (6-connected), Cr (6-connected), and Zr (8- and 12-connected) oxoclusters, (b) organic nodes (the central part of multitopic organic linkers) (c) connecting building blocks (the arms of multitopic organic linkers or the body of ditopic linkers).

5.3.2 Training, Validation, and Test Set Data Generation

Before any GCMC simulations were run, we randomly split our 2,400 MOFs into 1,800 training MOFs, 200 validation MOFs, and 400 test MOFs. To generate our training data, we ran GCMC simulations for 200 one-, and two-atom alchemical adsorbates at fugacities of 1, 5, 10, 50, 75, and 100 bar in the 1,800 training MOFs (fugacity was used as opposed to pressure, so we did not have to calculate critical constants for each alchemical species). To generate our validation data, we ran GCMC simulations for 200 one-, and two-atom alchemical adsorbates (all were entirely different than any adsorbate used for computing the training data) at fugacities of 2.5, 30, 60, 80, and 90 bar in the 200 validation MOFs. The LJ parameters, charges and bond-lengths used to generate all of the alchemical species considered are given in Table C.1 to Table C.4. To generate our test data, we ran GCMC simulations for 12 real adsorbates (argon, krypton, methane, xenon, nitrogen, ethane, helium, hydrogen, propane, butane, isobutane, and benzene) at fugacities of 1, 2.5, 5, 10, 25, 50, 60, 75, 80, and 100 bar in the 400 test MOFs. Note that the real adsorbates considered here were distinct from the alchemical species used to generate the training/validation data, i.e. no real adsorbate had the same LJ, charge, or bond-length parameters as the alchemical ones. GCMC simulations for six real adsorbates (argon, krypton, methane, xenon, nitrogen,

ethane) were run in 1,528 MOFs from the CoRE MOF database [70, 71] at fugacities of 1, 5, 10, 50, 75, and 100 bar for additional testing data including experimental MOFs.

LJ parameters for helium correspond to those used by Smit and coworkers [305], LJ parameters for argon were taken from Perez and coworkers [306], and LJ parameters for krypton, and xenon correspond to those used by Sikora and coworkers [222]. The parameters for these adsorbates are summarized in Table C.5. Methane, ethane, propane, n-butane, isobutane, benzene, and nitrogen, were modeled according to the TraPPE force-field developed by Siepmann and coworkers [216, 307]. Accordingly, methane, ethane, propane, n-butane/isobutane, and benzene are modeled as a one, two, three, four, and six uncharged sites, respectively, while nitrogen is modeled with three charged sites in order to reproduce electric quadrupoles. For nitrogen, the two atoms are each assigned LJ parameters and charges, while a "dummy" site at the center of mass is only assigned a charge [307]. LJ parameters and charges for hydrogen were taken from the Darkim-Levesque model [272, 308], which is also a three-site model, however, the two atoms are each assigned only a charge while a dummy site at the center of mass is assigned LJ parameters and a charge. Figure C.1 shows the generic one- and two-atom force field models used for GCMC simulations.

5.3.3 MOF Fingerprinting

We tested seven different MOF fingerprints, and found that a fingerprint combining six MOF textural properties—helium void fraction (VF), gravimetric surface area (GSA), largest pore diameter (LPD), pore limiting diameter (PLD), inverse framework density ($1/\rho$F), and the pore size standard deviation (PSSD)—together with the number density of 17 distinct MOF chemical moieties resulted in sufficiently accurate predictions. The descriptors for the fingerprint are illustrated in Figure 5.2. The range of chemistries covered by our constructed MOFs is shown in Figure 5.2, with the associated percentages indicating the frequency with which it appears in the database. The range of textural properties covered by the constructed MOFs is shown by the histograms in Figure 5.3. Our optimal fingerprint can be considered simple because it is limited to 23 easily-calculated descriptors instead of the hundreds needed when using atomic-property weighted radial distribution functions [95] or other high dimensional descriptors. VF was calculated using the Widom insertion method with helium as the probe molecule [309], while GSA was calculated by rolling a nitrogen-sized spherical probe along the framework surface [310]. Both of these calculations were implemented in RASPA-2.0 [91]. LPD and PLD were calculated using zeo++ (version 0.2.2) [85]. PSSD, as a measure of pore polydispersity, was calculated by taking the weighted standard deviation of the pore size distribution (also calculated using RASPA-2.0), where each pore diameter was weighted by the distribution value. For each MOF, the number density of a given chemical moiety was calculated by counting the number of times that moiety appeared in the unit cell and

dividing by volume of the latter.

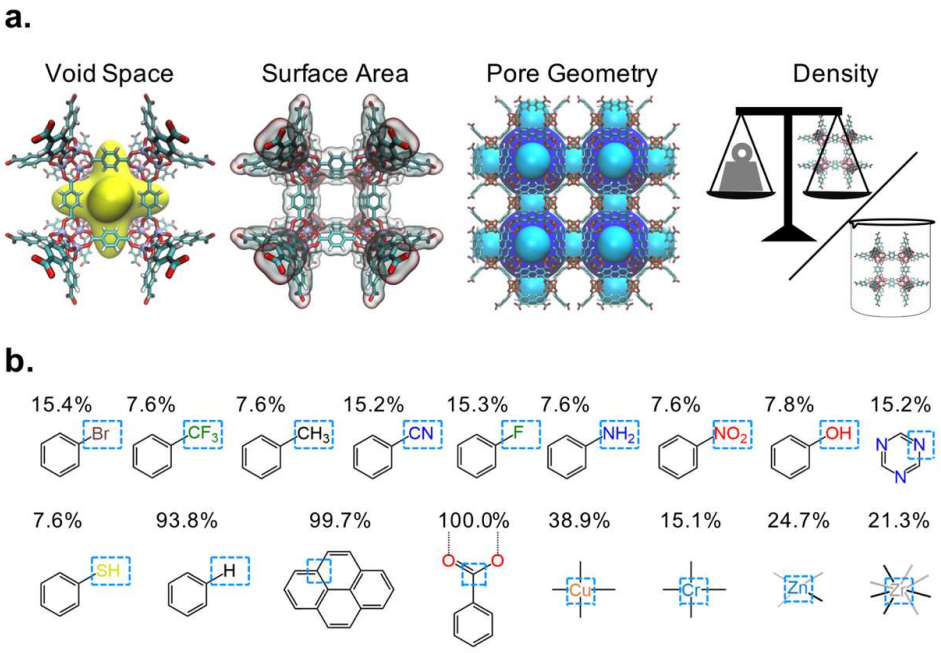

Figure 5.2 Descriptors constituting the MOF fingerprint. (a) Six textural properties: void fraction, gravimetric surface area, largest pore diameter (LPD: dark blue sphere), pore limiting diameter (PLD: light blue sphere), pore size standard deviation (PSSD), and density. (b) 17 chemical motifs (boxed), for which their respective number density in each MOF was calculated. The percentage of MOFs in the database that contain each motif is listed at the top.

5.3.4 Adsorbate Fingerprinting

Toward generalized adsorbate predictions, we set out to demonstrate that the loading of real adsorbates can be predicted using training data consisting entirely of alchemical adsorbates. Additionally, we wanted to show that an adsorbate (real or alchemical) can be represented by a fingerprint, allowing adsorbate properties to become part of the training data. As a first step, we focused on both one-site (single atom) and three-site (two atom) alchemical and real adsorbates, where three-site adsorbates had a dummy atom with only a point charge at the bond center (typical of forcefield representations of diatomic gases), where this charge may be null. As simulated adsorption loadings depend both on adsorbate adsorbate and adsorbate-framework interactions, we hypothesized that an operational adsorbate fingerprint should include descriptors related to the adsorbate features that control dispersion and electrostatic interactions. Ultimately, we used effective LJ parameters ($\epsilon_{\text{effective}}$ and $\sigma_{\text{effective}}$) for each adsorbate along with the maximum charge magnitude (which corresponds to the dummy atom charge) and the bond length (zero for single-site adsorbates), which allowed us to keep the number of descriptors in the fingerprint identical

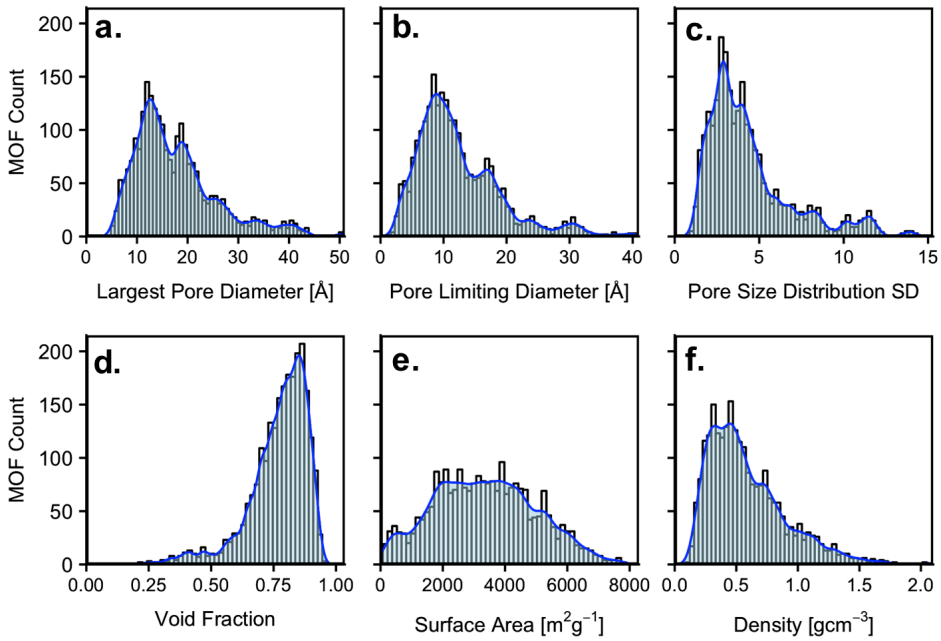

Figure 5.3 Histograms of the textural properties used as descriptors in the MOF fingerprint.

regardless of the adsorbate (a requisite for generality). For single-site adsorbates, $\epsilon_{\text{effective}}$ and $\sigma_{\text{effective}}$ are exactly the ϵ_{ii} and σ_{ii}. For two-atom adsorbates, $\epsilon_{\text{effective}}$ was the sum of the ϵ_{ii} of the different sites, and $\sigma_{\text{effective}}$ was:

$$\sigma_{\text{effective}} = (2\sigma_{ii} + r_{\text{bond}})/2 \tag{5.1}$$

which is the average of σ_{ii} and the end-to-end length of the molecule if we consider the diameter of each atom to be σ_{ii}. For the more complex adsorbates considered later (see below), $\epsilon_{\text{effective}}$ was taken to be the sum of all ϵ_{ii} values, and $\sigma_{\text{effective}}$ was taken to be the average of the shortest dimension and longest dimension (an extension of Eq. 1 to adsorbates with more than one bond). Similarly, the bond length, r_{bond}, was taken to be the longest distance between atom coordinates in the lowest energy geometry (as calculated according to the relevant force field, see above). Figure C.2 shows an example of how the fingerprint described above is calculated for both mono and diatomic adsorbates.

5.3.5 Adsorption Simulations

RASPA-2.0 [91] was used to perform all GCMC simulations. In grand canonical simulations chemical potential, volume, and temperature are kept constant. Chemical potentials were calculated directly from fugacity. Simulations consisted of 2,000 initialization cycles (no data recording) and 2,000 production cycles (data recording) for all one and two atom adsorbates. Each cycle consists of N Monte Carlo moves (translation, rotation, or insertion/deletion), where N is the highest value between 20 and the number of

adsorbates in the simulation cell. For propane, n-butane, isobutane, and benzene, simulations consisted of 15,000 initialization cycles followed by 5,000 production cycles, and configurational bias was used during insertion moves. Adsorbate-adsorbate interactions were modeled using Lennard-Jones (LJ) potentials to describe dispersion interactions and Coulomb's law to describe charge-charge interactions. Available Dreiding forcefield [83] parameters were assigned to framework atoms, otherwise UFF [80] parameters were used. Framework-adsorbate electrostatic interactions were ignored as the partial charges of the adsorbates were small. The negligible effect of this choice on loadings was shown earlier for N_2 adsorption at 77 K [310], but was tested and verified again here on our test set MOFs (Figure C.3). This test was possible by assigning atomic charges to each MOF according to our MBBB approach [204]. Lorentz-Berthelot mixing rules were used to calculate parameters for interactions between atoms not explicitly parametrized. Adsorbate force-field parameters were assigned as discussed above.

5.4 Neural Network Training

Here we trained a multilayer perceptron (MLP) to predict the adsorption data obtained from GCMC simulations. All MLPs were trained using Keras [311] through the Scikit-learn [312] Python package (all version numbers for Python packages used during training are given in Table C.6). First, before training the final MLP, we investigated different network hyperparameter configurations to determine which hyperparameter values were likely to give an accurate final model. During this procedure (called tuning) we assessed model performance using both mean absolute error (MAE) and mean absolute percentage error (MAPE) on the validation set. These errors were selected as they both have useful and relevant physical interpretations (and are what we are most concerned with minimizing when predicting loading).

Tuning was performed using a two-step procedure. First, we exhaustively investigated diverse neural network topologies from one to eight hidden layers with between 10 and 50 nodes (in increments of 10 nodes) in each layer, keeping all other hyperparameters fixed, to find a class of network topologies which generally gave the most accurate results. We found that one, two, and three hidden layer networks, while making reasonably accurate and reproducible predictions, had higher error than deeper MLPs. On the other hand, we found that many deep networks (more than five layers) had low minimum error on the validation set, but were sensitive, i.e. we observed large oscillations in the validation set error across epochs and with slight changes in network topology for these networks. Therefore, we selected a four-hidden-layer topology for our final model, as it was both highly accurate, and robust in its predictions. Second, and after settling upon this network topology, we varied other important net hyperparameters on a grid, keeping topology constant. The hyperparameters considered and their final values (Table C.7) are presented in Appendix C. The final model resulting from the above tuning procedure was then tested on

the real adsorbates in the test set MOFs, and in the CoRE MOFs (see below). Additional data, demonstrating the reproducibility of our model, is presented in Figure C.4. The architecture of this final model is shown in Figure 5.4. We reiterate that there were no shared MOFs or adsorbates between the training, validation, and test sets, during any model training. In addition, there were no shared fugacities between the training and validation set. We considered both shared and unshared fugacities between the training and test set. Every network considered was trained for a maximum of 500 epochs. Early termination with a patience of 20 epochs was employed to prevent overfitting. That is, if validation loss (MAE for the final MLP) did not improve for 20 epochs in a row, training was terminated and the lowest loss from the previous epochs was taken to be the model error.

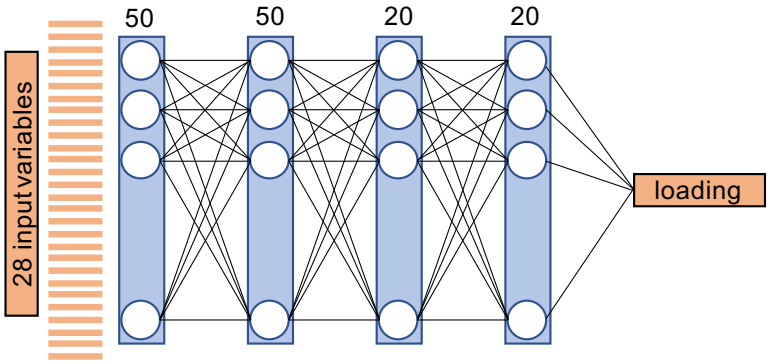

Figure 5.4 The configuration of our final model. The number of nodes in each hidden layer are shown above the corresponding layer.

5.5 Results and Discussion

5.5.1 Model Predictive Ability for Simple Adsorbates

Before discussing the overall predictive performance of our final MLP, we discuss briefly, from an intuitive perspective, some of the things that the model must learn. Figure 5.5 shows adsorption isotherms for a subset of the single-atom alchemical adsorbates (from the training set) in three MOFs. These MOFs are representative of structures with small (LPD 7 Å), intermediate (LPD 14 Å, or near the first peak in the LPD histogram shown in Figure 5.3) and large (LPD 20 Å, or near the second peak in the LPD histogram shown in Figure 5.3) pores, respectively. The adsorbates with the largest ϵ_{ii} and σ_{ii} have the highest loading at low fugacities in all three MOFs. Thus, the model thus must learn that intrinsic adsorbate-adsorbent interactions play a dominant role in controlling adsorbate loadings at low fugacity. However, the larger ϵ_{ii} and σ_{ii} are, the more easily saturation is reached as fugacity increases, giving rise to pore size limitations. For instance Figure 5.5 shows that, at high fugacity, and in small pore MOFs, a smaller molecule (σ_{ii}= 3) with a weaker interaction (ϵ_{ii} = 100), can have a higher loading than a larger

molecule (σ_{ii}= 4) with a stronger interaction (ϵ_{ii} = 150). However, this is no longer the case in the large pore MOF. The model, consequently, must learn that pore size limitations play a dominant role for larger adsorbates in smaller pores. Of course, this is just a brief glimpse into the intricacies of the interplay between MOF and adsorbate properties and fugacities (operating conditions) that determine adsorbate loading, and which the model must learn.

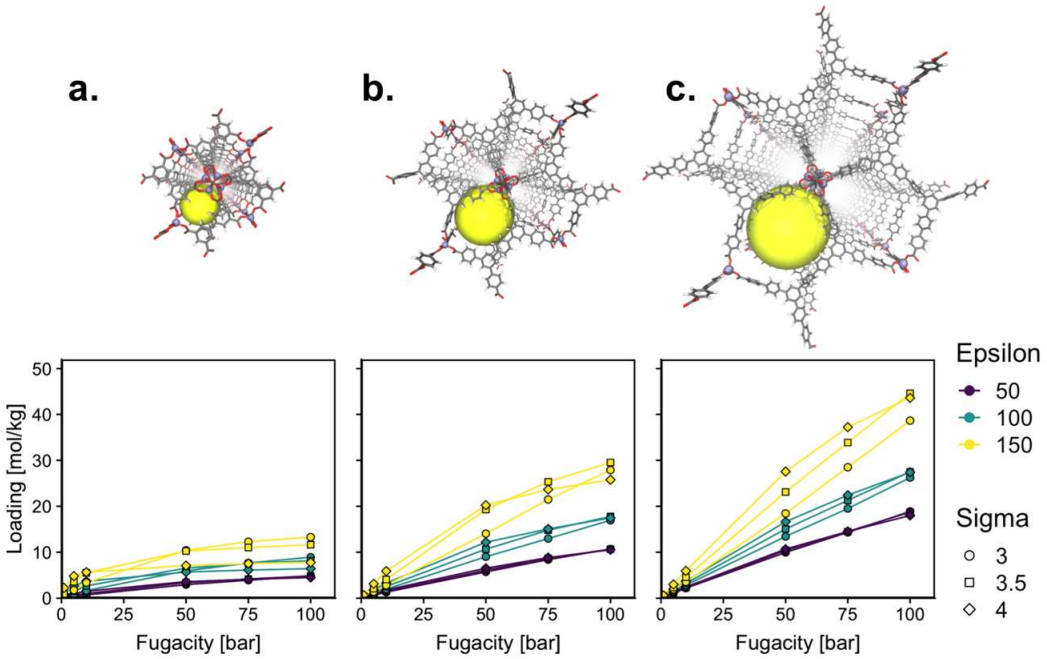

Figure 5.5 Isotherms for alchemical adsorbates considered in (a) a representative small pore MOF (LPD=6.5 Å), (b) a representative intermediate pore MOF (LPD=13.7 Å), and (c) a representative large pore MOF (LPD=19.5 Å). All the MOFs shown are of the **mcn** topology. The large pore in each case is illustrated by a yellow sphere.

Next, we consider the predictive performance of our final model on a set of real adsorbates by comparing our model predictions to GCMC calculated loading of argon, methane, krypton, xenon, ethane, and nitrogen at 10 fugacities from 1 to 100 bar in the 400 test set MOFs. We computed several measures of model performance for each adsorbate, which are presented in Table 5.1. Specifically, we consider mean absolute percentage error (MAPE), mean absolute error (MAE), and Pearson correlation (R). Perfect predictions would have zero MAPE and MAE and an R value of unity. Values of R close to one indicate a very strong linear correlation between GCMC loadings and those predicted by the MLP. As a point of comparison, our final model predicted the validation set loadings (200 alchemical adsorbates in 200 MOFs at 6 fugacities) with a MAPE of 3.0% and a MAE of 0.18 mol/kg. We note that MAPE is biased towards adsorbates with higher loadings, since a larger absolute error may still be a relatively low absolute percentage error. For example, while nitrogen and argon predictions are visibly accurate (and have the

lowest MAE values), their MAPE values are relatively high. On the other hand, MAE is biased towards adsorbates with lower loadings, since the relatively small absolute errors may be large in comparison to the actual loading value. This is why we present multiple and diverse model performance metrics, as no single metric can be used to fully assess model performance.

Table 5.1 Model performance metrics of our final model for loading predictions made on the test set MOFs.

Adsorbate	MAPE[%]	MAE [mol/kg]	R
Argon	4.4	0.17	0.99
Methane	4.6	0.24	0.99
Krypton	4.7	0.30	0.99
Xenon	4.5	0.42	0.99
Ethane	4.3	0.42	0.99
Nitrogen	4.0	0.14	0.99

Parity plots showing the MLP predicted loadings (at all 10 fugacities considered in the test set) versus the corresponding GCMC simulated loadings for each adsorbate provide a more complete picture of the predictive capabilities of the final MLP (Figure 5.6). Perfect predictions would result in all the points in these plots falling on the diagonal line. It is clear from Table 5.1 and Figure 5.6 that the final MLP predicts loadings of the six adsorbates considered here exceedingly well, as expected, given that the same model predicted the loadings of 200 similar alchemical adsorbates with similar accuracy. Parity plots, however, do not give a complete picture of the performance of a model trained to predict adsorption loading. Not only should the model predict individual loading points correctly, but it should also predict related points in the correct order and all with a similar level of accuracy. That is, the model should be able to reproduce full isotherms, which means it is accurate at each point, and also predicts the shape of the isotherm.

5.5.2 Predictive Ability for Full Isotherms

Now we proceed to illustrate the ability of our model to predict full adsorption isotherms. This is a necessary if one aims to, for instance, couple machine learning predictions for pure components with IAST theory (when applicable) to rapidly obtain mixture adsorption data to screen MOFs for chemical separation applications. Figure 5.7 compares isotherms predicted by the MLP (continuous line) and those obtained from GCMC simulations (points) for methane and ethane (plots for the other adsorbates are given in Figure C.5). To get a more accurate picture of the GCMC-simulated isotherms we ran simulations in the test set MOFs at fugacities not included in the training set (empty points in Figure 5.7).

As it is unfeasible to present the isotherm comparison for all the test cases studied here (six adsorbates in 400 MOFs), we chose to present five isotherms per adsorbate. However, to provide a fair picture of prediction accuracy, we aimed to present a range of "best to worst" cases. To do so, we first ranked all the

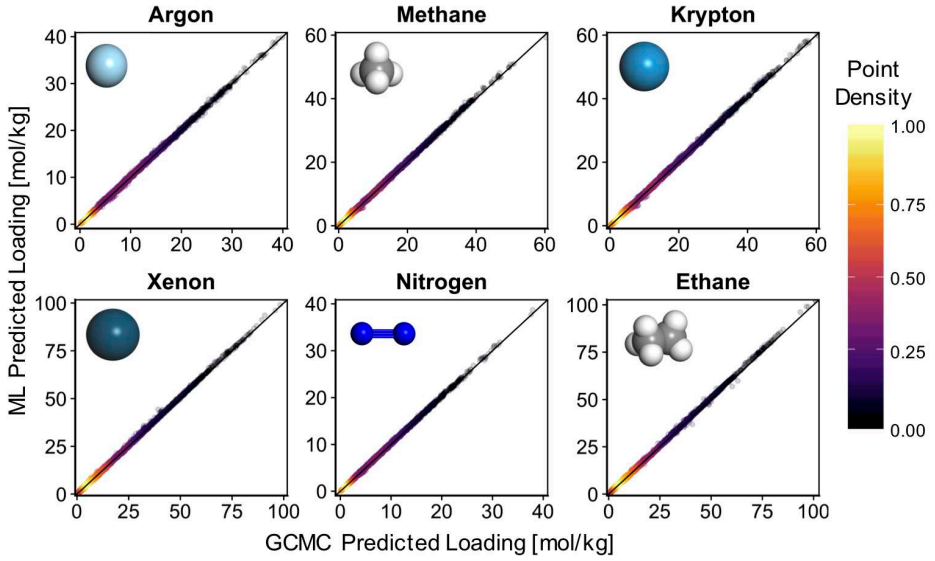

Figure 5.6 Parity plots comparing the predictions of the final MLP model for the six indicated adsorbates versus GCMC-calculated values in the 400 test set MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings).

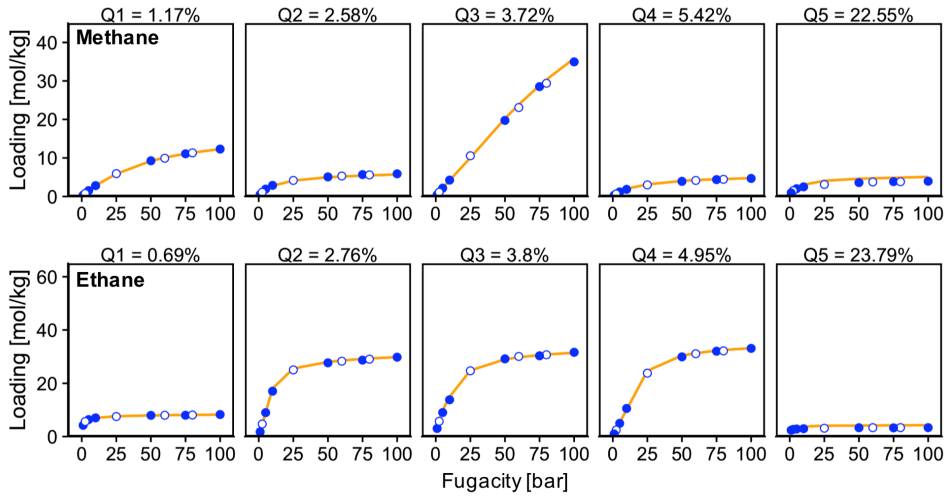

Figure 5.7 Isotherms for methane (top) and ethane (bottom), for the Q1 (0.00), Q2 (0.25), Q3 (0.50), Q4 (0.75), and Q5 (1.00) quantiles of isotherm mean percentage error (IMPE, with corresponding values shown). Points are GCMC simulated values (filled points correspond to fugacities included in training, empty points were not included in training), and orange lines show MLP predictions.

isotherms predicted by our MLP according to their isotherm mean percentage error (IMPE). For each MLP-calculated isotherm point for which we also had a GCMC-simulated value (10 fugacities for each MOF), we estimated the absolute percentage error (APE). The mean of this set of APE values was taken to be the IMPE. The IMPE values were then used to classify the MLP-predicted isotherms into five quantiles—the 0.00 (Q1), 0.25 (Q2), 0.50 (Q3), 0.75 (Q4), and 1.00 (Q5) quantiles. Thus, Q1 isotherms are representative of the best predicted isotherms according to our IMPE metric, Q5 isotherms are representative of the worst, and Q3 isotherms are representative of an "average" (or median) prediction. Figure 5.7 and Figure C.5 present one isotherm from each quantile (the one nearest to the IMPE quantile value).

Q1 isotherms are quantitively correct for all adsorbates, Q5 isotherms tend to be qualitatively correct but can deviate more significantly from GCMC-simulated values in some pressure ranges. However, as machine learning predictions are intended for use in high throughput screening of MOFs (or other porous crystals) some less than stellar predictions are acceptable as long as the vast majority of predictions are good. This is the case even for isotherms in the Q3 and Q4 quantiles. For instance, Q3 isotherms (the "average" prediction accuracy) have IMPEs ranging from 3.23% (for nitrogen) to 3.93% (for krypton). As a point of comparison, Dokur and Keskin [313] showed that that a difference of well over 10% can be observed in GCMC predicted loadings of methane and nitrogen (albeit in CO_2/N_2 and CO_2/CH_4 mixtures) when switching between using UFF and Dreiding LJ parameters for MOF atoms, and that these errors likely do not affect high throughput screening results significantly. Accordingly, the accuracy reached by the trained MLP is certainly suitable to accelerate materials discovery by utilizing it as part of a hierarchical screening strategy.

5.5.3 Transferability of Model for Predictions on MOFs with Inaccessible Pores

Some MOFs have pores that are inaccessible to certain adsorbates due to large energy barriers that separate these pores from the rest of the pore volume. Since the GCMC algorithm can insert adsorbates into such inaccessible pores, in some cases explicit "pore blocking" must be done to obtain accurate adsorption predictions. However, we did not use such blocking during the generation of training data as this would have required preparing different blocking schemes for each MOF depending on the force-field parameters for each adsorbate considered (note that which pores are inaccessible depends on adsorbate size and/or interaction energy) before running GCMC simulations and calculating textural properties. We reasoned that this procedure was unnecessarily onerous because—as proven with our use of alchemical adsorbates—the training simulation data does not have to be "realistic" to be meaningful to teach the machine learning model how adsorption depends on descriptor values. Therefore, we posited that a model

trained without pore blocking considerations would still be able to correctly predict adsorption in cases where pore blocking was necessary simply by using the new set of textural properties that results from pore blocking as input for the model. As discussed in Appendix C and illustrated in the parity plots in Figure C.6, the above hypothesis was shown to be true, further validating our training strategy.

5.5.4 Transferability of Model for Predictions on Experimental MOFs

Although we aimed to make our database as chemical and topologically diverse as possible, one may speculate as to whether this diversity is sufficient to train a model that is applicable to the space of experimentally known MOFs. One way to assess this is to test our model on a MOF database that is not subject to the biases present in our database. Thus, we tested our model predictions on 1,528 out of 12,479 structures in the CoRE MOF database, which is a collection of experimental MOF structures curated by Chung and coworkers [70, 71]. The details of our CoRE MOF selection procedure are described in Appendix C, but the majority of discarded CoRE MOFs simply had elements not encountered in the database we used for training the model. Figure 5.8, which shows predictions on the CoRE MOFs, is analogous to Figure 5.6 in that it compares our model predictions for argon, methane, krypton, xenon, nitrogen and ethane at six fugacities with GCMC simulations. Although the MAE, MAPE, and R values calculated from the model predictions (Table C.8) on the CoRE MOFs show a decrease in overall predictive ability, the parity plots in Figure 5.8 and histograms of MAPE values (Figure C.7) show similar error distributions for predictions on the CoRE MOFs and the test set for our constructed database, albeit with a longer tail towards higher MAPEs for predictions on the former.

Consistent with the errors for prediction of individual adsorption loadings, full isotherm predictions by the model on CoRE MOFs remained accurate, albeit with higher errors. For instance, for the Q3 IMPE values in the CoRE MOFs vary between 6.7% (ethane) and 11.8% (nitrogen). Analogous plots to Figure 5.7 for the CoRE MOFs are shown in Figure C.8. At this point it is important to note that a significant portion of the CoRE MOFs used in our testing have chemical motifs and building blocks not seen in our database, which is likely the source of the heightened model error. Our model naturally accounts for this to a certain extent but cannot be expected to predict the influence of all the different building blocks and chemistries extant in the CoRE MOFs. For instance, while the copper motif in our database always corresponds to Cu in a paddlewheel, we can consider any Cu atom in the CoRE MOFs to be part of the copper motif, e.g. Cu as part of an "infinite" node (as in MOF-74), or as a single-atom node, and not necessarily bound to carboxylates. Such Cu motifs will influence the adsorption differently than Cu in a paddlewheel, but likely in a manner that is still correlated to the Cu number density. More specifically, for any motif that consists of a unique element (i.e. -Br, -SH, and the metals), we count any instance of

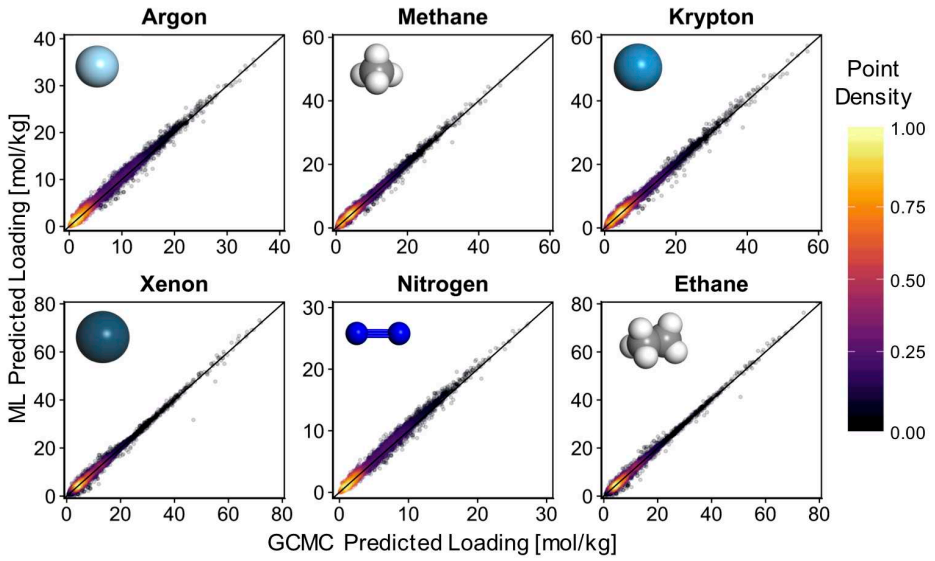

Figure 5.8 Parity plots comparing the predictions by the final MLP model for the six indicated adsorbates versus GCMC-calculated values in 1,528 CoRE MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings).

that element when computing number densities. But for motifs that correspond to a particular type of an element (i.e. aromatic carbon, carbon in the -CN functional group, etc.) we ensure that the element is actually of that type when computing number densities. Therefore, any types of certain elements present in the CoRE MOFs that are not included in our fingerprint (e.g. triple-bonded carbon) are not explicitly considered by our model, but will be accounted for to a certain extent by the MOF textural properties.

5.5.5 Material Ranking Accuracy Based on Performance Metrics

For machine learning to be effectively used in hierarchical screening, it needs to correctly rank MOFs (or whichever type of porous crystal is being studied) according to some performance metric. This way, it guarantees that the most promising MOFs are studied with more accurate methods in subsequent screening stages. Thus, here, we assess the ability of our model to identify top-performing materials. One important consideration at this point is that material performance in chemical separations and gas storage often depends on adsorption properties at more than one pressure/fugacity. For instance, the working capacity, which is the difference between adsorption loadings at a high and a low pressure is a common performance metric. Thus, to evaluate the ability of our model to rank MOFs, we decided to include the ranking based on working capacities, in particular for the 100 bar $\leftrightarrow$ 5 bar fugacity swing, we calculated for the six real adsorbates on the 400 MOF test set, and the CoRE MOF subset introduced in above.

To compare the "MLP rankings" and "GCMC rankings," we considered two approaches. In the first approach, the well-known Spearman rank correlation coefficient (S) was calculated. In the second approach, we focused on the ability of our MLP model to identify MOFs in the GCMC "top 100." This latter approach is potentially more informative as during hierarchical screening one is not necessarily concerned with capturing exact rankings with the MLP model. Rather, one would be satisfied with identifying the majority of the top performing MOFs (even if not in the correct order) for subsequent GCMC screening. Of course, S being equal to one implies that the MLP model correctly captures all MOFs in the top 100. However, the converse is not necessarily true, and the two described approaches can be considered complementary.

The values of S and the number of correctly identified top-100 MOFs are given in Table 5.2 for both loadings (at 100 bar) and working capacities (for the 100 bar $\leftrightarrow$ 5 bar fugacity swing). As an overall indication of the similarity between MLP rankings and GCMC rankings notice that the values of S are very close to one in all cases. As for the ability of the MLP model to capture MOFs in the top-100 for each adsorbate case, when our model was applied to the test MOFs, it was able to correctly identify at least 98 (97) of the top 100 MOFs based on the considered loadings (working capacities). When applied to the CoRE MOFs, our model correctly picked at least 91 (93) of the top-100 MOFs. Evidently, the MLP does a slightly better job ranking MOFs in the test set than in the CoRE MOFs. This is not surprising considered (i) the less accurate MLP predictions in the CoRE MOFs, (ii) the one order of magnitude higher number of CoRE MOFs considered here relative to the number of test set MOFs, which makes the selection of the top 100 inherently more difficult. As a complement to Table 5.2, Figure C.9 shows our model performance for selecting the top N MOFs (N ranges from 1 to 100) from both the test set and the CoRE MOF subset. From Figure C.9 we can observe that our model makes accurate selections (above 90% correct placement) even when considering N much less than 100. Based on the above results, we can conclude our MLP model is able to pick out the vast majority of high performing MOFs from diverse databases as a first screening step.

5.5.6 Testing the Limits of the Current MLP Model (Extrapolation)

In the preceding sections, we have demonstrated that it is possible to train machine learning models to predict adsorption of multiple adsorbates, which was the major goal of this work. Moreover, the trained MLP is sufficiently predictive to be applied as a first step in the screening of MOFs for various important and challenging separations involving mixtures of small nonpolar molecules such as Xe/Kr and CH_4/N_2. However, our results should be considered as a stepping-stone toward the training of "smarter" models capable of predicting adsorption for a larger diversity of molecules with different shapes, significant

Table 5.2 Comparison of MOF rankings between MLP predictions and GCMC simulations as indicated by the Spearman rank correlation coefficients (S) and the number of MOFs in the top-100 from GCMC simulation that were also found in the top-100 from MLP predictions.

Adsorbate	MOFs Correctly Placed in Top-100				S			
	Loading@100 bar [mol/kg]		Working Capacity [mol/kg]		Loading@100 bar [mol/kg]		Working Capacity [mol/kg]	
	Test MOFs	CoRE MOFs	Test MOFs	CoRE MOFs	Test MOFs	CoRE MOFs	Test MOFs	CoRE MOFs
Argon	99	92	99	93	0.999	0.991	0.999	0.987
Methane	99	95	98	95	0.999	0.994	0.999	0.980
Krypton	98	94	97	95	0.999	0.994	0.999	0.980
Xenon	99	98	99	96	0.999	0.991	0.999	0.942
Nitrogen	98	91	99	94	0.999	0.991	0.999	0.988
Ethane	99	98	99	96	0.999	0.992	0.999	0.944

flexibility, and larger polarities.

One way to inform the development of future models able to predict adsorption for more diverse molecules, is to understand how the current model "breaks" (if it does) when applied to more "complex" adsorbates. Therefore, we applied the present MLP model to adsorbates that have properties not captured by the alchemical adsorbates used for training. The tested adsorbates were helium ($\epsilon_{\text{effective}}$ and $\sigma_{\text{effective}}$, smaller than for any alchemical species), hydrogen (LJ parameters at a non-atom site), and select hydrocarbons (propane, n-butane, isobutane and benzene), all of which have (*i*) higher $\epsilon_{\text{effective}}$ and $\sigma_{\text{effective}}$, (*ii*) more sites, and (*iii*) significantly different shapes than any alchemical species included in training. Thus, to make adsorption predictions for these adsorbates, the MLP model must extrapolate.

The parity plots in Figure 5.9 provide a visual comparison between MLP predictions and GCMC data for the six "extrapolated" adsorbates (R, MAE and MAPE values are given in Table C.9). Despite the extrapolation, predictions for helium and hydrogen remain quite good, making the present model applicable to, for instance, screen MOFs for CH_4/H_2 and CH_4/He separations. In contrast, adsorption predictions for propane, n-butane, isobutane, and benzene generally fail, but do so in an interesting fashion. Namely, MLP predictions and GCMC values are linearly correlated, but with systematic underprediction. Additionally, the linear correlations present a notable number of outliers, most of which to correspond to low loadings. Intriguingly, these "low-loading outliers" tend to center around the parity line. Further analysis of the predictions for the hydrocarbons (Figure C.10) revealed that MLP (under)predictions that linearly correlate with GCMC data correspond to cases where the loading is more than 50% of the saturation value. On the other hand, MLP predictions clustering around the parity line correspond to cases where the loading is less than 25% of the saturation value. Since a material approaches saturation as fugacity increases, the MLP model is expected to predict low fugacity isotherm

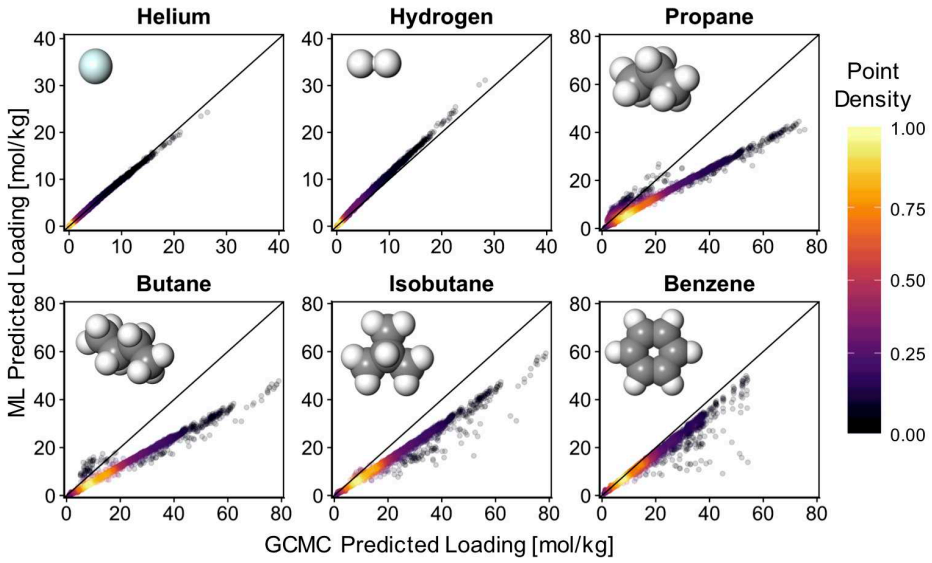

Figure 5.9 Parity plots comparing the predictions of the final MLP model for the six indicated (extrapolated) adsorbates versus GCMC-calculated values in the 400 test set MOFs. These adsorbates possess properties outside the ranges covered by the alchemical adsorbates used during model training. Points color indicates the point density in the plot (the highest density is observed at low loadings).

points better than high-fugacity ones.

What, then, did we learn from testing the model on "extrapolated" adsorbates? Our MLP predicts the adsorption of hydrogen and helium well, likely because these molecules are small, non-polar, and near-spherical (resembling the alchemical adsorbates used for model training). The fact that the MLP was able to extrapolate well for these two adsorbates suggests that the model learned some fundamentals about the physics of adsorption. However, for the hydrocarbons, the trends observed in the MLP predictions are likely related to the non-spherical shapes of these adsorbates. Considering that the MLP was trained to predict saturation loadings (which depends on how molecules pack within MOF pores) of molecules that are near-spherical, and that the highly non-spherical hydrocarbons will pack differently than these near-spherical adsorbates, the overserved systematic error is not surprising. In fact, the packing of the hydrocarbons is likely to be more efficient than our model predicts, given that ellipsoids (a shape more representative of the geometries of the considered hydrocarbons) can pack more densely than spheres of the same volume [314, 315]. In addition, the volume of a sphere with a diameter of $\sigma_{\text{effective}}$ is greater than the actual volume occupied by the hydrocarbons in the GCMC simulation (when calculating this actual volume as that occupied by overlapped spheres of diameter σ_{ii}). This is likely why the MLP systematically underestimates loadings for fugacities where the MOF is at or near saturation, but not so for fugacities where the MOF is far from saturation (where the molecules are isolated, and framework-adsorbate

interactions are more important). The systematic loading underestimation near saturation follows the order n-butane > propane > isobutane > benzene, which is consistent with the order of asymmetry of these molecules. Based on this, one would expect that future work on developing an MLP applicable to more complex molecules than those considered in the present work should include measures of adsorbate shape/asymmetry as part of the adsorbate fingerprint (perhaps along with descriptors of MOF pore shape). The introduction these additional features could "correct the slope" for the linear correlations in Figure 5.9.

5.6 Conclusions

In this chapter, we demonstrated that the same multilayer perceptron (MLP) model can be used to predict full room temperature adsorption isotherms of small, near-spherical, nonpolar, mono- and diatomic adsorbates at different pressures (fugacities). Key to accomplishing these predictive capabilities was the inclusion of thermodynamic conditions (here fugacity), and adsorbate force field parameters as model inputs, and (most importantly) the inclusion alchemical adsorbates in training set. Our MLP model, made, on average, quantitatively accurate predictions of full isotherms in MOFs and for adsorbates not included in the training set. In addition, our model shows excellent performance in ranking MOFs according to maximal loading and working capacity, the latter requiring predictions at two pressures. Our results are a first step towards the ambitious goal of universal prediction of adsorption in porous crystals, which will greatly speed up high-throughput screening of materials for adsorption applications. The next step toward universal prediction of adsorption should focus on expanding training sets to include multiple temperatures and a larger diversity of adsorbates, including large, flexible adsorbates (e.g. C_n alkanes), and adsorbates with strong electrostatic interactions with framework atoms (e.g. CO_2, water, alcohols), with the goal of correcting predictions made on highly non-spherical and flexible adsorbates. Such extension to more diverse molecules will require further development of methods for fingerprinting porous crystals and adsorbates.

CHAPTER 6

METHODS FOR THE LARGE-SCALE COMPUTATION OF MOF FREE ENERGY AND CRITERIA FOR MOF THERMODYNAMIC ACCESSIBILITY

Modified from a paper entitled published in the Chemistry of Materials.[11]

Ryther Anderson[12], Diego Gómez-Gualdrón[12,13]

In this chapter, a method for the large-scale computation of MOF free energies is described. Using this method, we can calculate the free energy of every MOF in a large database (8,500 MOFs here and 50,000 MOFs in Chapter 7). These energies can then be used to assess whether MOFs in a database are thermodynamically feasible. We outline an approximate global stability criterion by comparing the free energies of previously synthesized MOFs in our database with those of the hypothetical MOFs. Besides, we show that our calculated free energies can be used to identify the experimentally observed isomorph from an isomorphic series with a high degree of accuracy. While these stability criteria are not transferable to MOFs in general, this work is the first of its kind and will serve as a starting point for more generalized synthesizability criteria. The stability criteria outlined here are used in Chapter 7 to identify thermodynamically feasible MOFs that are also likely to be high-performing for multiple separations.

6.1 Abstract

Metal-organic frameworks (MOFs) have captivated the research community due to their modular crystal structure that can be tailored to suit diverse applications. However, identifying ideal MOFs for an application of choice is difficult due to the millions (or more) possible MOFs one could consider. Although computational screening of MOF databases has provided a fast way to evaluate MOF performance, experimental validation of the predicted "exceptional" MOFs is uncommon due to uncertainty on the synthetic likelihood of computationally constructed MOFs, hence hindering material discovery. Aiming to leverage the perspective provided by large datasets, we calculated the free energies of each MOF in a topologically diverse database of 8,500 frameworks and evaluated to what extent descriptors of MOF thermodynamic stability "discriminate" previously synthesized MOFs. Upon defining a relative free energy, $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$, that corrects for some force field artifacts specific to MOF nodes, we found that previously synthesized MOFs in our database clustered in a region below $\Delta_{\mathrm{LM}}F_{\mathrm{FL}} = 4.4$ kJ/mol per atom. This

[11]Reproduced in part with the permission of: R. Anderson, D. Gómez-Gualdrón. "Large-Scale Free Energy Calculations on a Computational Metal–Organic Frameworks Database: Toward Synthetic Likelihood Predictions" *Chem. Mater.* 32, 19, 8106–8119. Copyright 2020 American Chemical Society.

[12]Chemical and Biological Engineering Department, Colorado School of Mines
[13]Corresponding author, dgomezgualdron@mines.edu

suggests that a MOF below this $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ threshold may have a higher probability of being synthesized, although other factors may ultimately impair synthetic accessibility. For instance, when isomorphism occurs, multiple isomorphs may reside under the $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ threshold and relative stability among isomorphs come into play. From 32 isomorphic MOF series we examined in detail, we found the synthesized isomorph was the one with the lowest free energy in in 80% of cases, and in 20% of cases to be within 1 kJ/mol of the latter. These findings indicate that for a MOF to be synthetic accessible a "low" crystal free energy is necessary, albeit in some cases it may not be sufficient due to the role of other factors not considered here (e.g. solvents, modulators and kinetics).

6.2 Introduction

Nanoporous crystals, led by a class of materials known as metal-organic frameworks (MOFs), have the potential to provide technological breakthroughs in fields as diverse as catalysis [56–59], chemical separations [67–69, 238], molecular sensing [60–62], drug delivery [63–66], and energy storage [38, 40, 178, 179]. MOFs are promising in such diverse fields due to the tunability of their pore geometry and chemistry, a property which allows researchers to precisely control how molecules adsorb, diffuse, and react within the framework. This tunability is the direct result of the combinatoric nature of MOF construction, where each combination of MOF building blocks and topology defines a unique MOF, which can be thought of as a point in "MOF-space." Given the number of these combinations (i.e. the vast size of MOF space), synthesizing and experimentally testing every possible MOF is impossible. Therefore, discovering a high-performing MOF for a given target application is a major challenge that requires new, sophisticated methods for searching MOF space.

In recent years, high throughput computational screening (HTCS) has emerged as an efficient method (relative to experimental screening) to not only explore the MOF-space [37], but also for exploring other porous material spaces [316, 317]. By screening databases of experimentally reported structures (e.g. the CoRE MOF [70, 71] or the IZA zeolite [28] databases), already explored material-space can be revisited for new applications. However, screening databases of hypothetical materials allows venturing beyond the bounds of known material-space and opens the door to new material designs that could outperform existing ones [37, 222, 317, 318]. A typical HTCS study of hypothetical materials consists of the computational synthesis of a database of prototypes, followed by molecular simulations to calculate performance-defining properties for a target application. An early example of such databases of material prototypes is the Predicted Crystallography Open Database (PCOD) [319], which was primarily populated with hypothetical zeolites [320]. Ideally, some of the promising material prototypes in a database would then be selected for synthesis and experimental validation of their properties, thus accelerating material discovery. In the case

of MOFs, a few high-performing materials have been synthesized as the result of this type of HTCS study. For example, the chemically robust NU-800 for methane storage [202] or of the small-pore SBMOF-1 for Xe/Kr separation [43]. However, many HTCS studies do not result in the synthesis of a high-performing MOF, which is perhaps due to the way databases of MOF prototypes are constructed.

MOFs in these databases are built by assembling building blocks primarily on the basis of geometrical rules informed by coordination chemistry [37, 73, 74, 76, 77]. Thus, although the majority of MOF prototypes may look feasible, there is uncertainty as to whether a given prototype can actually be synthesized. Currently, synthesis decisions following HTCS are based on chemical intuition about MOF synthesizability. Given the cost of exploratory experiments, synthesis is not necessarily attempted for the best MOF prototypes. Rather, synthesis is attempted for "good" prototypes that resemble previously synthesized MOFs (e.g. a prototype within a known isoreticular series), at the expense of missing potentially higher-performing MOFs. This hinders the effectiveness of simulation-guided MOF discovery, and effectively constrains the exploration of MOF space to the vicinities of already existing MOFs. Incidentally, the tendency to constrain exploration around existing materials also pervades purely experimental work. As noted in our earlier work, despite the appeal of MOF tunability, MOF space has arguably been underexplored, with 50% of over 40,000 articles on MOFs (published at the time) focusing on only 30 MOFs [77].

Large-scale, free energy calculations on MOF databases used for HTCS, however, have not been attempted despite the potential insights that could be gleaned from the large amount of simulation data (e.g. through the emergence of well defined structure-property relationships relevant to MOF synthesizability). Here we decided to assess the viability of classical MD simulations to estimate MOF free energies in a way that is (sufficiently) accurate, yet still amenable for large-scale screening. Inspired by methods used in the pharmaceuticals field for crystal structure prediction of organic molecules,[61]–[65] we decided to apply and evaluate free energy calculation methods in a database of 8,500 MOF prototypes: the quasi-harmonic approximation (QHA) and thermodynamic integration, specifically the Frenkel-Ladd (FL) path [169, 321, 322].

Our MOF prototypes were computationally constructed using ToBaCCo-3.0 [73, 77], from 34 building blocks combinatorically assembled according to 110 topological blueprints. As is typical of this procedure [37, 38, 41, 96, 323], one expects most generated MOFs to not have been experimentally reported—regardless of whether they are actually synthesizable—but one also expects to generate some experimentally reported MOFs. Additionally, with sufficient topological diversity in the database, one expects to find numerous, non experimentally observed isomorphs of the previously synthesized MOFs. We aimed to exploit this composition of the database to assess to what extent free energies are able to

"discriminate" previously synthesized MOFs. By using the free energies of previously synthesized MOFs as a point of reference, we interrogated whether (*i*) previously synthesized MOFs share common thermodynamic stability characteristics despite varied composition and/or (*ii*) previously synthesized MOFs present the highest thermodynamic stability among MOFs of identical composition (i.e. isomorphs). In addressing these questions, we explored possibilities and limitations for the use of MOF free energy as a descriptor for synthetic likelihood of MOFs in computational databases.

6.3 Computational Methods

6.3.1 MOF Database Construction

Our 8,500-MOF database was created using ToBaCCo-3.0 [73, 77], with the building blocks shown in Figure 6.1. These building blocks were selected to create topological and chemical diversity in a relatively small number of structures, i.e. so our database spans a wide range of textural and energetic properties. In contrast to the databases we created for recent studies that focused on the prediction of adsorption properties [84, 293], the current database does not include functionalized MOFs, which allowed us to explore more topologies, linker shapes ,and coordination schemes, and increase the occurrence of isomorphism. The geometry of each MOF prototype was optimized in LAMMPS [211] according to the relevant force field: UFF [80], Dreiding [83], or UFF4MOF [82, 210]. Note, however, that since parameters for Cu in UFF are defined for tetrahedral coordination, we altered the equilibrium angle to 90° to be consistent with the Cu-paddlewheel geometry. Similarly, in Dreiding equilibrium bond lengths and angles are unavailable for some metals (Cu, Cr, Zr), so we adopted the equilibrium bonds/angles we used in our UFF implementation but otherwise retained the functional forms and force constants used for Dreiding. Note that these modifications to Dreiding and UFF have been previously validated for MOFs [209].

To optimize MOF geometries, we employed an iterative approach. First, only the atom positions were relaxed using the fast inertial relaxation engine (FIRE) algorithm [212] with a timestep of 10.0 fs. Then the atom positions and simulation cell parameters were relaxed simultaneously using a conjugate gradient algorithm. We repeated this process (relaxing only the atom positions and then the atom positions and cell parameters together) until the energy change from the previous iteration was less than 4.184×10^{-6} kJ/mol. The individual optimizations were halted when the maximum force on any atom was less than 4.184×10^{-6} kJ/(mol Å) and the energy changed by less than 1×10^{-4}% from the previous step.

6.3.2 Free Energy Calculations

All free energy calculations were performed using molecular dynamics (MD) in LAMMPS [211] with a timestep of 1 fs. The thermostat of Bussi et al. [324] (damping parameter of 100 timesteps) was used for

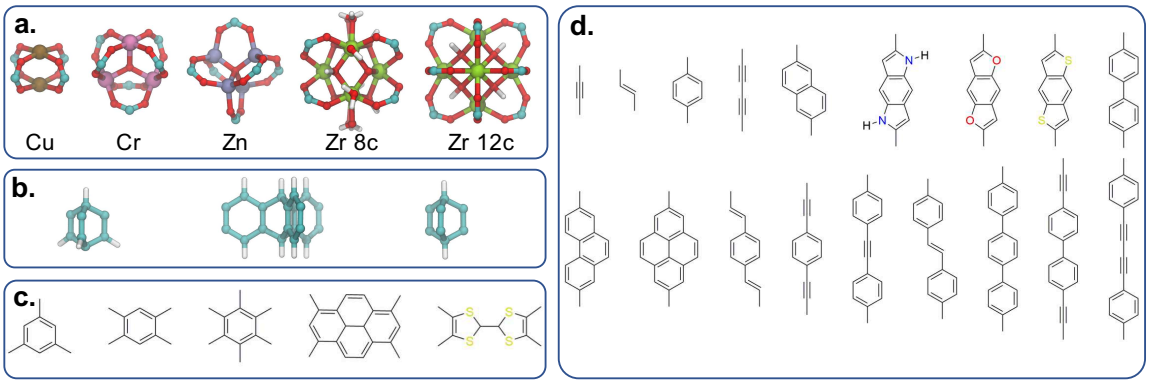

Figure 6.1 The computational building blocks used to construct our 8,500-MOF database. (a) nodular inorganic building blocks (NIBBs). (b) 3-dimensional, nodular organic building blocks (NOIBs). (c) 2-dimensional nodular organic building blocks. (d) connecting building blocks (CBBs). NOIBs approximately correspond to so-called "secondary building units" (SBUs) in MOF synthesis literature. NOIBs correspond to the node of multitopic organic linkers. CBBs approximately correspond to ditopic linkers and/or to the arms of multitopic organic linkers.

all simulations and a Nosé-Hoover barostat (damping parameter of 1000 timesteps) was used for NPT simulations. We calculated quasi-harmonic approximation (QHA) free energies using:

$$F_{\mathrm{QHA}} = U_0 + \frac{1}{2}\int_{\nu=0}^{\nu_{\max}} h\nu D\left(\nu\right)\mathrm{d}v + k_B T \int_{\nu=0}^{\nu_{\max}} D\left(\nu\right)\ln\left[1-\exp\left(\frac{-h\nu}{k_B T}\right)\right]\mathrm{d}\nu \tag{6.1}$$

where U_0 is the potential energy at 0 K for the equilibrium cell shape, ν are vibrational frequencies, and $D\left(\nu\right)$ is the vibrational (phonon) density of states (PDOS) [126, 325, 326]. $D\left(\nu\right)$ is normalized such that $\int D(\nu)\mathrm{d}v = 3N$ where N is the number of atoms in the system [327]. To apply Equation 6.1 we needed the PDOS for each MOF, which we obtained by Fourier-transforming an averaged, mass-weighted velocity trajectory (i.e. the mass-weighted velocity of each atom at each timestep) [327, 328]. First, we calculated the equilibrium simulation cell shape in the NPT ensemble at zero pressure by averaging each parameter over 100 ps after 100 ps of equilibration. The velocity of each MOF atom was then output every 5 timesteps for 10 ps after 50 ps of equilibration in the NVT ensemble (at the equilibrium cell shape). These velocities were used to compute the PDOS for each MOF. More details are provided in Appendix D.

We calculated the Frenkel-Ladd (FL) path free energies using the method of Freitas et al. which is described in detail in ref. [169]. Briefly, the FL path is a thermodynamic integration method in which the system is switched from an Einstein crystal (where each atom is constrained to its defined equilibrium position by a harmonic potential) to a force field representation. During thermodynamic integration, an initial system is switched to a final system according to a parametrical Hamiltonian

$$H\left(\lambda\right) = \lambda H_f + \left(1-\lambda\right)H_i \tag{6.2}$$

where H_i is the initial Hamiltonian, H_f is the final Hamiltonian, and λ is the switching parameter. For $\lambda = 0$, $H(\lambda)$ is H_i and for $\lambda = 1$, $H(\lambda)$ is H_f [164, 169, 321]. The difference in free energy (ΔF) between the initial and final states is equal to the work required to switch between the Hamiltonians:

$$\Delta F = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{6.3}$$

where angled brackets denote an ensemble average at each value of λ. Since the free energy of an Einstein crystal F_{EC} is known analytically, an absolute free energy of the solid is given by:

$$F_{\mathrm{FL}} = \Delta F + F_{\mathrm{EC}} \tag{6.4}$$

where ΔF is the work required to switch from the Einstein crystal to the force field representation, and where

$$F_{\mathrm{EC}} = \ 3k_BT \sum_{i=1}^{N} \ln\left(\frac{\hbar\omega_i}{k_BT}\right) \tag{6.5}$$

with ω_i being the harmonic frequency of atom i in an N atom system. One way to calculate the ΔF in 6.3 is by performing a series of equilibrium MD simulations for each value of λ, during which $\partial H/\partial\lambda$ is calculated as an ensemble average. These averages can then be numerically integrated from $\lambda = 0$ to $\lambda = 1$ to yield ΔF. However, Freitas et al. showed, following work by de Koning [329], that ΔF can be calculated as the average work of two non-equilibrium switching processes [169]. First, the system is switched from the Einstein crystal to the forcefield representation (i.e. λ is changed from 0 to 1) during the course of a single non-equilibrium MD trajectory. Then, the system is switched back to the Einstein crystal after equilibration with the forcefield Hamiltonian. The forward work is denoted $W^{\mathrm{irr}}_{i\to f}$ and the reverse work is denoted $W^{\mathrm{irr}}_{f\to i}$, with the superscript indicating that the forward/reverse switching processes are irreversible. For sufficiently long switching trajectories, the heat dissipated during switching is small [168] and the heat dissipated during the forward and reverse processes is of the same magnitude and can be cancelled (see ref. [169] for a detailed justification), meaning:

$$\Delta F = \frac{1}{2}(W^{\mathrm{irr}}_{i\to f} - \ W^{\mathrm{irr}}_{f\to i}) \tag{6.6}$$

Accordingly, from Equations 6.4, 6.5, and 6.6 one can write:

$$F_{\mathrm{FL}} = \ \frac{1}{2}(W^{irr}_{i\to f} - \ W^{\mathrm{irr}}_{f\to i}) + 3k_BT \sum_{i=1}^{N} \ln\left(\frac{\hbar\omega_i}{k_BT}\right) \tag{6.7}$$

Typically, there are additional terms that correct for fixing the center of mass during the simulation, however we found this term to be significantly smaller than the statistical error for our systems (see Appendix D) [164, 169]. In principle, $W^{irr}_{i\to f}$ and $W^{\mathrm{irr}}_{f\to i}$ should be averages calculated from several replicate simulations. However, we found that a single simulation with long enough switching time was sufficiently

precise (again, see Appendix D).

Practically, we implemented the FL path free energy calculations in a single MD simulation. First, the equilibrium simulation cell parameters were calculated with the same method used for our PDOS calculations (see above). We then switched to the NVT ensemble (as the FL path is only valid at fixed volume) using the equilibrium cell shape. Next, we calculated the mean-squared displacement (MSD) of each atom type (according to the force field) as an ensemble average over 50 ps after 50 ps of equilibration (MSDs are used in Equation 6.8, below). We then ran the switching procedure, with switching time of 400 ps, using the LAMMPS command fix ti/spring [169]. Details on the convergence of free energy with switching time are provided in Appendix D and Figure D.1. When using fix ti/spring the system is switched from the force field representation to the Einstein crystal representation, equilibrated with the Einstein crystal Hamiltonian (in our case for 30 ps), then switched back, which is the reverse of the traditional FL path (described above). However, we can simply reverse the sign of the calculated work to obtain the ΔF used in Equation 6.4.

The "strength" of the springs for each atom in the Einstein crystal is defined by force constants. Here, we calculated a different force constant (α_j) for each atom type j in the simulation (according to the relevant force field), using:

$$\alpha_j = \frac{3k_BT}{\langle \Delta r^2 \rangle_j} \tag{6.8}$$

where $\langle \Delta r^2 \rangle_j$ is the mean squared displacement (MSD) of atom type j (calculated as described above). This ensures that the MSDs for the real and Einstein crystals are similar, which makes the switching process more efficient [169]. However, we enforced a lower bound on α_j of 41.84 kJ/(mol Å^2) (see detailed justification in Figure D.2). If α_j is too small, Equation 6.3 can be difficult to evaluate due to divergence in the integrand from overlapping atoms [164]. In addition, if an atom travels too far from its Einstein crystal equilibrium position LAMMPS may lose that atom when building neighbor lists.

6.4 Results and Discussion

6.4.1 Comparison of MOF Free Energy Calculation Methods

The quasi-harmonic approximation (QHA) assumes that each MOF vibrational mode can be reasonably modeled with a harmonic potential. By contrast, no such assumption is made in the Frenkel-Ladd (FL) path method. In principle, then, one can take the FL free energies to be more accurate. However, should QHA and FL free energies differ by a sufficiently small amount, one would prefer to calculate QHA free energies in MOF databases as they require shorter simulations (Figure D.3). Thus, we start our discussion by comparing the QHA and FL path free energies (both calculated using UFF4MOF) in our 8,500 MOF

database.

Figure 6.2 is a parity plot comparing QHA (y-axis) and FL (x-axis) free energies. All free energies in the present work are calculated at 300 K and reported on a "per atom" basis unless otherwise specified. If the energies calculated using the two methods were equal, all points would lie on the line y=x (the parity line). Based on Figure 6.2, there is systematic deviation between FL path and QHA free energies, which is consistent with observations by Pastorino and Gamba as well as Stoessel and Nowak when comparing QHA and FL free energies of sulfur crystals and biomolecules, respectively [326, 330]. Note that purely systematic deviation in QHA free energies would not matter when using free energy as a metric of relative thermodynamic stability, as free energy differences between individual structures would be maintained. Such a scenario would be reflected by a perfectly linear correlation between QHA and FL free energies.

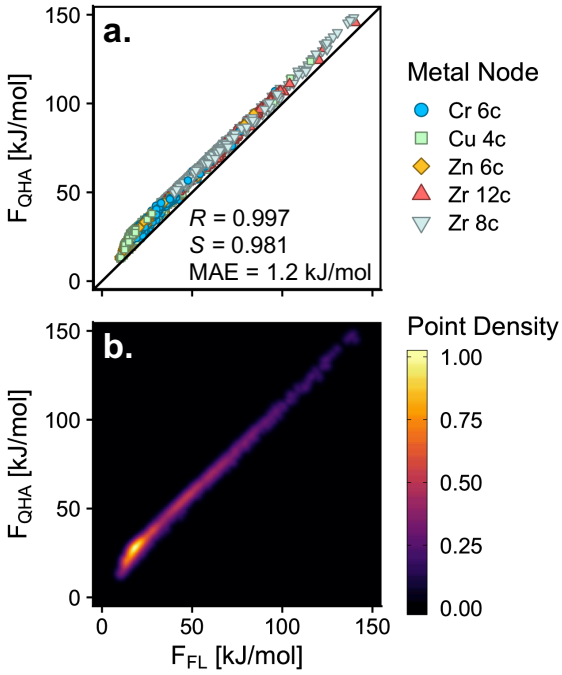

Figure 6.2 Comparison between free energies calculated with the QHA (F_{QHA}) versus free energies calculated with via the FL path (F_{FL}). (a) Parity plot with point shape and color corresponding to the type of MOF inorganic node. (b) Parity plot illustrating the point density according to the color bar. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE) calculated for adjusted QHA free energies.

To quantify the correlation between QHA and FL free energies, we report the Pearson (R) and Spearman rank (S) correlation coefficients. An R value of one indicates a perfectly linear correlation between QHA and FL free energies, whereas an S value of one indicates that the stability ranking of MOFs based on QHA or FL path free energies would be identical. We found both R and S to be close to one

(Figure 6.2, $R = 0.997$, $S = 0.981$). These R and S values suggest that one would observe similar free energy vs. MOF property relationships with either the QHA or the FL method (i.e. broad trends would be preserved when switching between methods). However, the reproduction of broad trends has no bearing on whether QHA and FL path free energies are equally valid criteria for assessing the relative thermodynamic stability of MOFs.

Thus, we calculated the absolute deviation between the QHA and FL free energies, after removing the apparent systematic error from the QHA values, which we did by subtracting the y-intercept of a linear model fit to the data shown in Figure 6.2(9.0 kJ/mol). Note that this subtraction does not affect differences among QHA free energies. However, it enables us to better evaluate the ability of the less expensive QHA method to correctly assess relative thermodynamic stability of MOFs. Once this adjustment was made, we found the absolute errors to span the 0-7.5 kJ/mol per atom range (Figure D.4), with the mean absolute error (MAE) being 1.2 kJ/mol per atom. We will see in below that (*i*) previously synthesized MOFs encountered in our database cluster below a 4.4 kJ/mol per atom free energy threshold, and (*ii*) free energy differences between MOF isomorphs are typically small (with a mean difference of 3 kJ/mol and many differences of less than 1 kJ/mol). On these grounds, one can see that the use of QHA free energies to assess thermodynamic stability (and perhaps synthetic likelihood) of MOFs is potentially problematic.

Based on the above, in subsequent sections our discussion will be based on FL free energies. Before proceeding, however, we think it valuable to briefly examine the origins of the differences between QHA and FL free energies in MOFs. MOFs present vibrations over a wide range of frequencies (0-3336 cm^{-1})—typically with a gap around 2300 cm^{-1}—as illustrated by representative phonon density of states (PDOS) in Figure 6.3 and Figure D.5. By dividing the overall frequency interval evenly into five ranges, we present in the histograms in Figure 6.3b to what extent vibrations in each range collectively contribute to the QHA free energy of MOFs. Figure 6.3 shows that high (¿ 2669 cm^{-1}), mid-high (2001-2669 cm^{-1}), mid (1334-2001 cm^{-1}), mid-low (667-1334 cm^{-1}) and low (¡ 667 cm^{-1}) frequency ranges typically contribute 25%, 7%, 35%, 30%, and 3%, respectively, to the QHA free energy. The exact collective contribution of vibrations in each range, however, varies with MOF composition.

Informed by the 95th percentile threshold in each histogram, we classified MOFs whose high-frequency (mid-low-frequency) vibrations contribute more than 31% (32%) to the QHA free-energy as "high-frequency" ("low-frequency") MOFs (details in D). As reported in earlier works [331–334], high-frequency vibrations correspond to bond stretching modes involving (light) H atoms. Mid-frequency vibrations mostly correspond to bond stretching modes not involving H or (heavy) metal atoms, and in-plane-angle-bending modes in aromatic rings. Mid low frequency vibrations mostly correspond to out of-plane-angle-bending modes in aromatic rings, and stretching modes involving metal atoms (e.g. in M-O

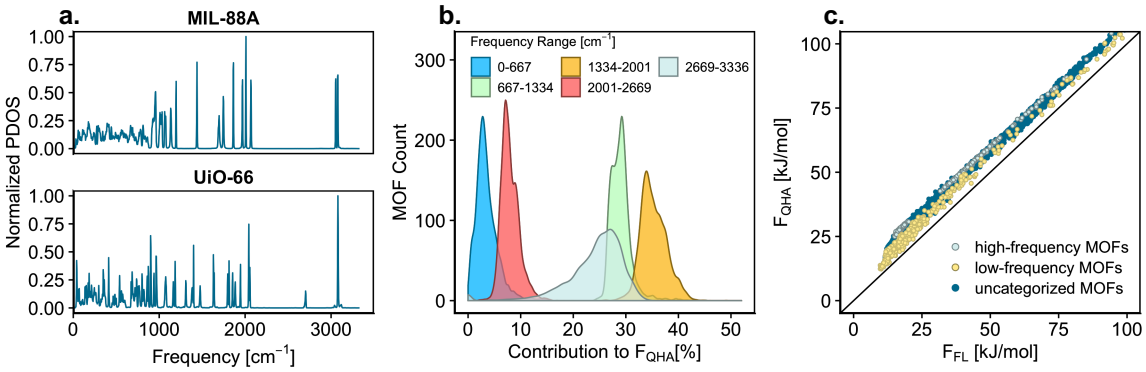

Figure 6.3 (a) Phonon density of states (PDOS) for two previously synthesized MOFs (names are shown above each plot). (b) Histograms of the contribution of frequencies in five ranges to the total QHA free energy (F_{QHA}) for each MOF (e.g. for most MOFs, frequencies in the 0-667 cm^{-1} range contribute less than 10% to F_{QHA}). (c) Parity plot comparing F_{QHA} and F_{FL} (FL free energies), with point colors indicating low-frequency, high-frequency, and uncategorized MOFs.

coordination bonds). Low frequencies mostly correspond to collective motions, including twisting of organic linkers and rocking of inorganic nodes. The PDOS of high-frequency MOFs in our database are dominated by C-H stretching modes, whereas low-frequency MOFs present very few (or none) of these modes (Figure D.5).

The location of high- and low-frequency MOFs in the parity plot in Figure 6.3 reveals that the QHA free energy error is more significant for high-frequency MOFs (albeit more "systematic"). This is perhaps surprising as high frequency MOFs are probably more "harmonic" and thus expected to be more amenable to QHA. To understand this observation, consider that the QHA free energy (Equation 6.1) can be considered as the lattice energy U "corrected" by vibrational terms: the zero-point energy (ZPE) and the temperature dependent logarithmic term to U. At 300 K, ZPE becomes the dominant correction for frequencies above 334 cm^{-1}. Given that U is less than the FL path free energy, the error between QHA and FL free energies must be primarily due to the ZPE, which is consistent with the QHA free energies of "high frequency" MOFs presenting larger errors (Figure 6.3) and the strong correlation between the error and ZPE (Figure D.6). Note, however, that the error tends to be more systematic (and hence less consequential for relative thermodynamic stability) in high-frequency MOFs than in low-frequency ones. This discussion is extended somewhat in the appendix to this chapter, see also Figure D.7, and the details on temperature dependence for QHA and FL free energies shown in Figure D.8 and Figure D.9. Furthermore, a discussion of using mass-wieghted vs. non-mass-weighted velocities is presented with additional data in Figure D.10 and Figure D.11. However, a deeper discussion onto this subject is beyond the scope of this contribution.

6.4.2 Comparison Between Free Energies Calculated with Different Force Fields

A question that emerges in any classical simulation study is the sensitivity of calculations to force field selection. Up to this point, we have focused on free energies calculated using UFF4MOF, because this update to UFF by Addicoat et al. [82, 210] was specifically designed for MOFs. While several MOF-specific force fields (of varied generalizability) exist [335–337], here we focus on a comparison of UFF4MOF with the generic force fields Dreiding [83] and UFF [80]. This was motivated by the tendency of HCTS studies to use these generic force fields (given their essentially complete transferability), as well as by the work of Boyd et al. [209] who reported that Dreiding and UFF perform reasonably well when predicting MOF properties such as bulk modulus and thermal expansion coefficients.

Parity plots comparing FL free energies among force fields are presented in Figure 6.4. The closest agreement in free energies is between UFF and UFF4MOF, especially for MOFs with Cu (paddlewheel) nodes (Figure D.12). The largest deviations are observed for MOFs with Cr, Zn, and especially Zr nodes, which can be rationalized by considering that UFF4MOF only modifies the UFF parameters that involve metals (e.g. M-O bonds, O-M-O, M-O-C angles, etc.) [82, 210], and there are more of these bond/angles in the Cr, Zn, and Zr MOFs. The largest disagreement in free energies is between Dreiding and UFF4MOF, which is expected, given that Dreiding uses different functional forms and force constants than UFF4MOF for all interactions. The "outliers" in Figure 6.4 correspond to MOFs whose lattice parameters change significantly when switching from UFF4MOF/UFF to Dreiding.

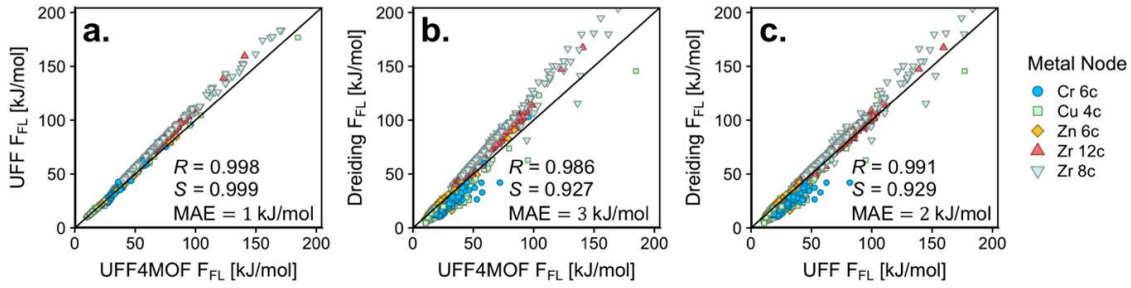

Figure 6.4 Parity plots comparing FL free energies (F_{FL}) calculated with different force fields. (a) UFF vs. UFF4MOF. (b) Dreiding vs. UFF4MOF. (c) Dreiding vs. UFF. Point shape and color correspond to the type of metal node in each MOF plotted. Different point features indicate the type of metal node in the MOF. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE).

Although the point spread in Figure 6.4 seems large, density plots (Figure D.12) show that most points remain near the parity line in each plot. Thus, a more comprehensive comparison is given by the values of R, S and MAE shown in Figure 6.4. In all cases R and S are close to one, indicating that similar free

energy vs. MOF property relationships would be obtained with UFF4MOF, UFF, or Dreiding force field. However, consistent with our discussion comparing QHA and FL free energies (MAE 1 kJ/mol) we must conclude that the use of Dreiding (MAE 3 kJ/mol) instead of UFF4MOF to assess the relative thermodynamic stability of MOFs may be problematic. Similarly, UFF free energies may not be sufficiently accurate (MAE 1 kJ/mol) when dealing with small differences in thermodynamic stability. From this point on, then, we primarily base our discussion on results using UFF4MOF, although we will point (where appropriate) differences and similarities in results were one to use Dreiding or UFF.

6.4.3 Correlations Between Free Energy and MOF Properties

Prior studies [338–341] postulated criteria for zeolite synthetically "feasibility," with varied degree of success. Inspired by these works, we decided to explore whether a similar criterion can emerge for MOFs, and assess to what extent such criterion succeeds or fails. However, acknowledging the numerous factors that may facilitate or hamper the synthesis of a particular MOF, in the context of HTCS, one probably can only talk of a synthetic likelihood criterion. An effective criterion would enable a preliminary step in HTCS studies as to eliminate MOFs with low synthetic likelihood from a computational database. Thus, computer resources can be allocated only to MOFs that have a higher chance of being synthesized. Even in instances where calculations on non-synthesizable MOFs could still be valuable (e.g. for machine learning), the criterion could be used to downsize a database that is too large for costly simulations.

Given that MOFs have diverse compositions, postulating a synthetic likelihood criterion is perhaps less straightforward than for zeolites. For example, while Li et al. identified a criterion for zeolite stability based on the relationship between T-site to T-site distances and O-site to O-site distances [338], such a criterion is only possible because all zeolites are arrangements of TO4 tetrahedra. An analogous criterion certainly does not exist for MOFs in general. On the other hand, Simperler et al. [341] identified a region in lattice energy vs. density plots where previously synthesized zeolites tended to cluster, hence postulated a "synthetic feasibility factor" based on how close (or far) a hypothetical zeolite is to (or from) this region. However, note that synthesis of one "unfeasible" zeolite has been recently accomplished [342], which highlights that factors like these should be interpreted from a probabilistic perspective. Nonetheless, it is easier to imagine an extension of this lattice stability based approach to MOFs. Thus, we examined the correlation between MOF free energies and MOF chemical/textural properties to see whether previously synthesized MOFs exhibited a clear relationship with any of these properties.

First, we needed to identify the MOFs in our database that had been previously synthesized. To accomplish this, we calculated the MOFid, a systematic MOF identifier developed by Bucior et al [343], for each MOF our database. We then compared the MOFids in our database to those of the CoRE MOFs (a

database of previously synthesized MOFs) [71]. Since the MOFid for a specific MOF is unique [343], if a MOFid from our database was also present in the CoRE MOFs, that means that the corresponding MOF in our database has been previously synthesized. Through this procedure, we found 54 previously synthesized MOFs in our database. For comparison, consider that Bucior et al.[343] found 50 and 18 previously synthesized MOFs from 10,159- and 1614-MOF subsets of earlier computationally constructed databases by Gómez-Gualdrón and coworkers [38, 73] and Wilmer and coworkers [37].

After identifying the existing MOFs in our database, we proceeded to plot FL free energies against MOF properties (Figure 6.5, additional relationships are shown in Figure D.13). Contrary to zeolites, we found that previously synthesized MOFs do not cluster in a single region. Rather, they tended to cluster in different regions depending on the composition of their metal nodes. For instance, the Zr MOFs occupy higher free energy regions than other MOFs. We found that this occupation of high free energy regions (Zr-MOFs ¿ Cr-MOFs ¿ Zn-MOFs Cu-MOFs) correlates with the strain energy (per atom) of the corresponding isolated nodes (Table D.1). The relative energies of the isolated nodes also explains the different behavior of free energy versus the metal/linker atom ratio for different MOF compositions. To calculate this ratio, any Cu, Cr, Zn, or Zr atom was counted as a metal, and any atom bonded to a carbon was counted as a linker atom (a sufficient criteria for the linkers in our database, but not for MOFs in general). For instance, in Zr-MOFs, free energy increases rapidly with the metal/linker atom ratio as adding more Zr-nodes will always increase potential energy. (Qualitatively similar trends are obtained when using UFF or Dreiding as shown in Figure D.14 and Figure D.15).

When examining the different correlations separately, we can observe that the free energies of previously synthesized MOFs tend towards the lower (free energy) boundaries of the corresponding region. This indicates that among MOFs of the same type, those with the lower free energies tend to have demonstrated synthetically accessibility. However, developing separate criteria for different MOF types is inconvenient and less informative than a single criterion. The clearest pattern for previously synthesized MOFs emerges from Figure 6.5c, where a linear relationship (for each MOF type) between the free energy of synthesized MOFs and metal /linker atom ratio is apparent for each MOF type, allowing us to fit four separate linear models to these free energies, as illustrated Figure 6.5d-g. Thus, other synthetic factors aside, it is reasonable to assume that MOFs on and under the corresponding fit line—which can be seen as a thermodynamic stability threshold—have higher probability of being synthesized.

The data shown in Figure 6.5c can be transformed by subtracting the relevant linear model from each MOF free energy. We refer to the free energy transformed in this manner as $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$. The results of this transformation are presented in Figure 6.6, where all the previously synthesized MOFs in our database now cluster within a small $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ region. The maximum $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ among experimentally observed MOFs in

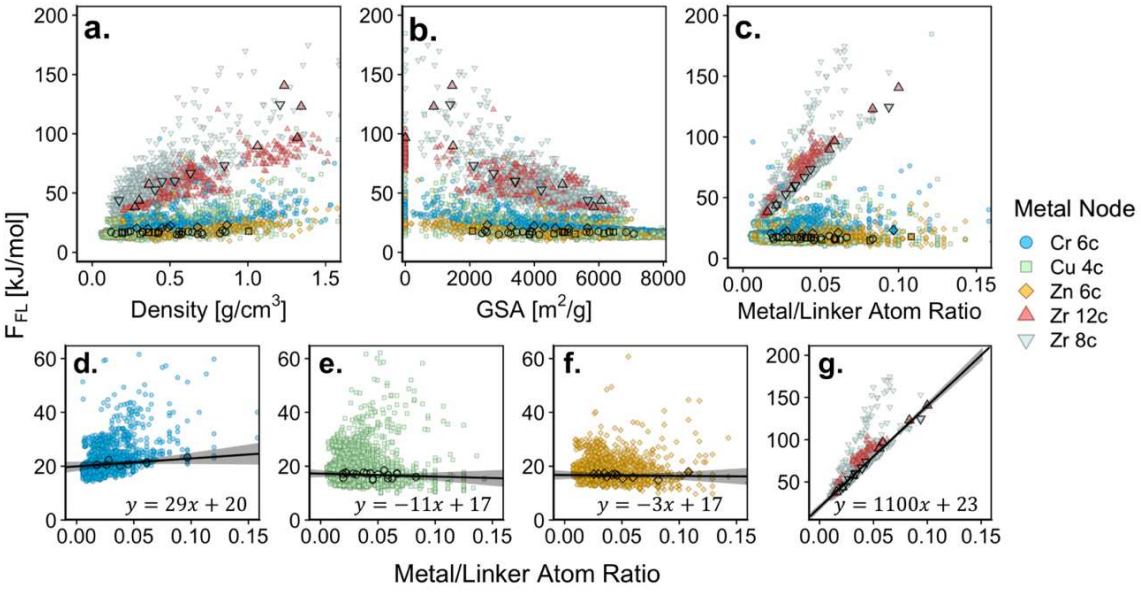

Figure 6.5 FL free energies (F_{FL}) versus MOF properties. (a) F_{FL} vs. MOF density. b) F_{FL} vs. MOF gravimetric surface area (GSA). (c) F_{FL} vs. metal to linker atoms ratio. (d-f) breakdown of F_{FL} vs. metal to linker atoms ratio plot in (c) by metal node. (d) 6c Cr MOFs. (e) 4c Cu MOFs. (f) 6c Zn MOFs. (g) 8/12c Zr MOFs. Point features indicate the inorganic node in the MOF. Points outlined in black correspond to previously synthesized MOFs found in our 8,500-MOF database. Black lines in plots (d-f) are linear fits (equation shown) for the free energies of previously synthesized MOFs of each inorganic node type, and shaded regions correspond to the 95% confidence interval of each linear fit.

our database is 4.4 kJ/mol. If one were to consider all MOFs with $\Delta_{LM}F_{FL}$ higher than this value as having insufficient thermodynamic stability, one could deem them to be synthetically unlikely. Thus, one could eliminate 1,613 MOFs from our 8,500-MOF database (or 1,502 and 707 MOFs when postulating an analogous threshold on the basis of calculations with UFF and Dreiding, respectively, Table D.2 and Figure D.16). We visually inspected many of the MOFs categorized as synthetically unlikely, and a large portion of these are not visibly strained (Figure D.17). This indicates that (*i*) not only is visual inspection inefficient, but also insufficient to identify MOFs with low thermodynamic stability, and (*ii*) a geometry-based criterion for MOF synthetic likelihood is not evident.

Note that the exact $\Delta_{LM}F_{FL}$ threshold presented here is specific to free energies calculated with UFF4MOF at 300 K. However, following the strategy outlined here, analogous criteria can emerge for other forcefields (as it occurred for UFF and Dreiding, Figure D.16) and temperatures. However, we do not expect pronounced changes in the general threshold across the typical temperature range for MOF synthesis (around 300-400 K). The reason is that within this range, we observed only small variations (at the database scale) of MOF free energy with temperature (F Figure D.18)—which is not to say that the synthetic accessibility of a particular MOF is not susceptible to temperature changes. Finally, it is important to note that probably there are MOFs below the $\Delta_{LM}F_{FL}$ threshold that are synthetically inaccessible, as there will be instances where high thermodynamic stability will not be sufficient to warrant successful synthesis (as other factors come into play). However, the absence of experimentally reported MOFs above the postulated thermodynamically stability threshold suggests that "low" crystal free energy is a prerequisite for synthetic accessibility.

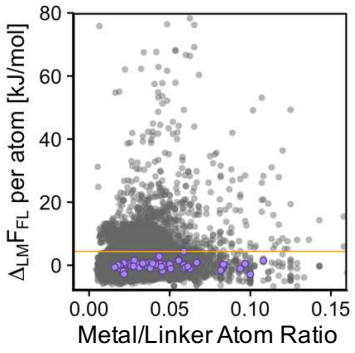

Figure 6.6 Relative FL free energy ($\Delta_{LM}F_{FL}$) versus metal to linker atoms ratio of each MOF. $\Delta_{LM}F_{FL}$ is obtained by subtracting the corresponding free energy from the linear fits in Figure 6.5d-g from F_{FL}. The horizontal line shows the bound that the 54 previously synthesized MOFs (purple points) in our 8,500-MOF database are all found below.

6.4.4 Consequences of MOF Isomorphism on the Application of the Overall Thermodynamic Stability Criterion

Given the occurrence of isomorphism in MOFs, it is possible that a MOF that is considered synthetically likely (as per the threshold defined above), may not actually be synthetically accessible simply because a more stable isomorph exists that is preferentially formed during synthesis. Crucially, previous work [77, 112] has shown that key MOF properties (e.g. mechanical stability, adsorption selectivity, and catalytic activity) can vary dramatically among isomorphic MOFs. Therefore, to inform synthetic attempts following HTCS, it is important to predict the most stable MOF within every isomorphic family in a database. Additionally, by only retaining the most stable isomorphs (by eliminating less stable isomorphs, even if they fall below the $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ threshold) one could further reduce the size a computational MOF database. Thus, we now proceed to assess thermodynamic stability within various MOFs isomorphic series to determine (*i*) how close in stability isomorphs typically are, (*ii*) how often multiple isomorphs fall below the $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ threshold, and (*iii*) how often the experimentally observed isomorph corresponds to the most stable isomorph within its series.

To this end, we first found which MOFs in our databases had isomorphs. To accomplish this, we identified all the unique linkers in our database by their canonical SMILES string, and then considered any MOFs with the same node and linker(s) to be isomorphs. We found that the free energy differences among isomorphs can frequently be well under 1 kJ/mol per atom (Figure 6.7), which is less than the width of the stability region in Figure 6.6, indicating that when a MOF is found to lie below the $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$= 4.4 kJ/mol threshold, it is extremely common to find many of its isomorphs (when existing) to also lie below the aforementioned threshold (Figure D.19). Thus, it follows that to examine synthetic likelihood for "low" free-energy MOFs belonging to isomorphic series, one may need to also consider their stability relative to their isomorphs.

To directly assess whether free energies can accurately differentiate between an experimentally observed MOF and its non-observed isomorphs, we identified all the isomorphs for each previously synthesized MOF in our database. This yielded 32 isomorphic series in our database that also included at least one existing MOF (not all previously synthesized MOFs in our database had isomorphs). The linkers observed among the MOFs in these series are shown in Figure D.20. Two of the 32 series contained multiple (two) experimentally observed isomorphs, as discussed in Appendix D along with other case-studies. For each of the 32 series, we determined whether the experimentally observed isomorph had the lowest free energy at 300 K. This was the case for 26 (81%) of the isomorphic series (Table D.3), or 24 (80%) if we do not consider the series with two experimental MOFs. Changes in isomorph prediction accuracies with temperature are given in Table D.4. When using UFF and Dreiding, respectively, 23 and 18 experimentally

observed isomorphs had the lowest free energy within their series).

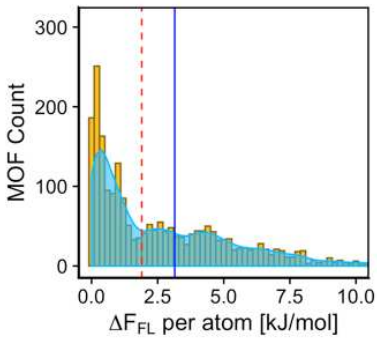

Figure 6.7 A histogram of the free energy differences ΔF_{FL} per atom observed between isomorphs in our 8,500 MOF database. The solid, blue line indicates the mean difference, and the dashed, red line indicates the median difference.

In Figure 6.8. we show one of the 32 studied isomorphic series, where the free energy of the experimentally observed NOTT-100 (**fof** topology) [213] is less than 0.5 kJ/mol per atom lower than three other possible isomorphs. Note that in contrast to the whole database, by definition we can say that all MOFs (experimentally observed and non-observed) in the subset of these 32 isomorphic series have had their synthesis attempted. Therefore, the abovementioned accuracy to predict the experimentally observed MOF can be used as a preliminary estimate to gauge the effectiveness of thermodynamic stability (as noted by free energies) as a criterion for MOF synthetic likelihood.

Underscoring the importance of "low" free energy for a MOF to be synthetically accessible, note that although for six isomorphic series the experimentally observed member did not have the lowest free energy, the free energy of the experimentally isomorph was (at most) within 2.5% of the lowest one in the series (within 2.0 and 2.1% when using UFF and Dreiding, respectively). The infrequent discrepancies between free energy rankings and the experimentally reported isomorph may be due to (*i*) synthetic factors (e.g. solvents, modulators, and different synthesis temperatures for different MOFs) not accounted for by our free energy calculation, (*ii*) experimental oversight where the lowest free energy isomorph has formed during synthesis but not reported—isomorphs that are close in free energy may coexist such as the NU-1000 and NU-901 cases discussed in the appendix to this chapter—or (*iii*) simply a prediction error inherent to either the forcefield or our free energy calculation method. Based on (*ii*) and (*iii*), one may consider the use of a tolerance before deciding to discard a MOF isomorph as synthetically unlikely. Although the effect of solvation on absolute free energies have been previously noted to be small [126], given the small free energy difference among isomorphs, it seems that examining the effect of solvents may be important to modulate the order of stability of isomorphic MOFs under synthesis conditions. However,

considering the diversity of solvents used in MOF synthesis, and the computational cost associated with including solvent and calculating free energies of solvation, such task may be impractical for HTCS at this point (although it merits study in smaller, rationally-selected MOF subsets).

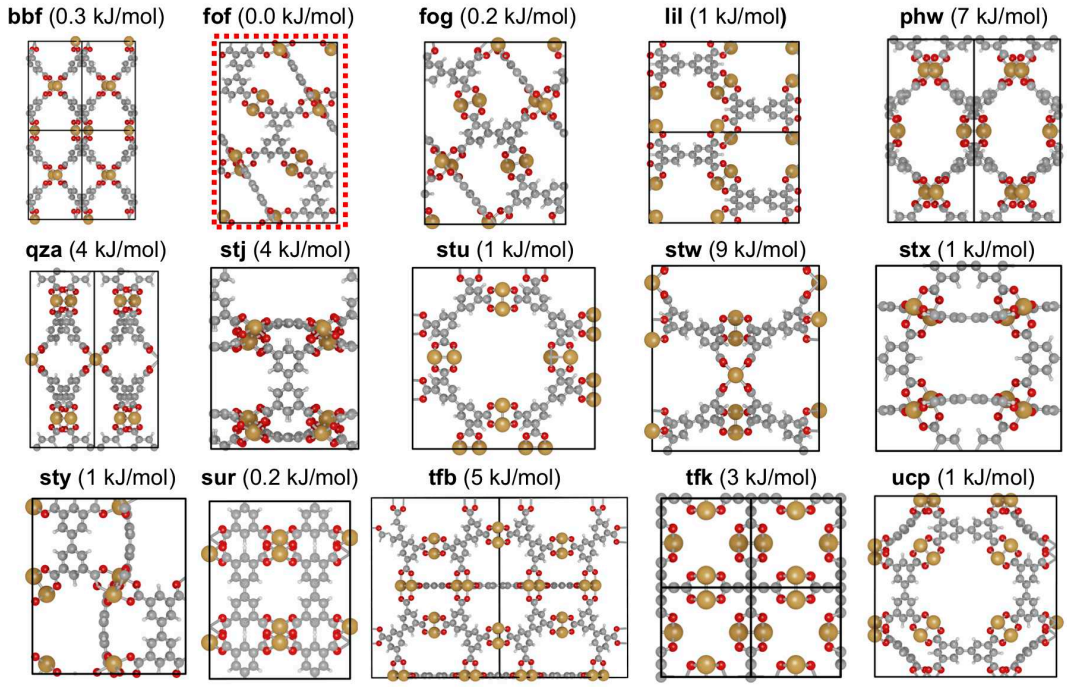

Figure 6.8 The experimentally observed NOTT-100 (outlined by the dashed line) along with all their isomorphs found in our database. Isomorphs are labeled by their topology, with their free energy per atom relative to the NOTT-100 (**fof** topology) presented in parenthesis.

6.4.5 Can Relative MOF Thermodynamic Stability be Assessed with Alternative Metrics?

While not 100% effective, the FL free energies of "activated" (i.e. solvent free) MOFs have shown considerable potential to inform the synthetic likelihood of MOF prototypes. Thus, free energies could be used to eliminate some synthetically unlikely MOFs from databases (reducing their size and increasing the computational efficiency of HTCS), and/or to inform synthesis attempts (e.g. when presented with two "promising" MOF prototypes, an experimentalist may decide to attempt synthesis for the one with the lowest free energy, provided it does not have a isomorph with lower free-energy). However, for this approach to work, free energies would need to be calculated for all MOFs in a database. Hence, an important question to ask is whether more affordable metrics could be used as effectively. Although we have not envisioned a geometry-based metric at this point, we decided to briefly inspect whether metrics such as potential energy at 0 K (strain energy, U_{0K}) and at a finite temperature (here, U_{300K}) could be as effective as the FL free energy as a stability metric. This was partly motivated by the strong correlation

between free, potential and strain energy (Table 6.1 and Figure D.21), which presents a MAE smaller than 0.8 kJ/mol.

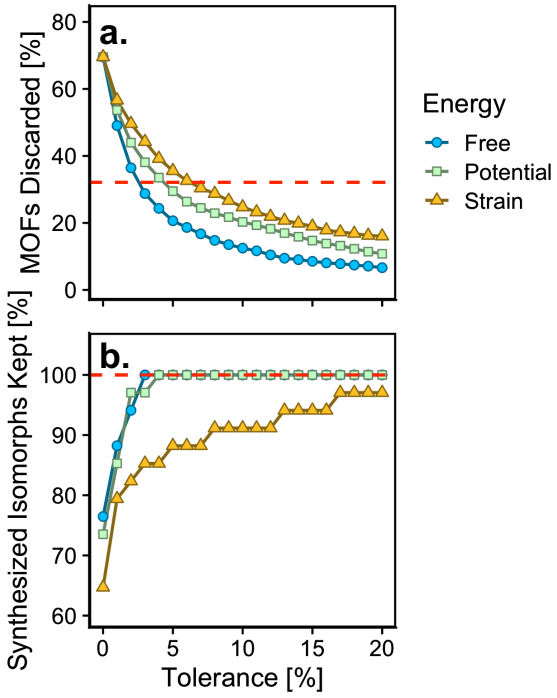

Figure 6.9 (a) Percentage of MOFs that can be discarded because they are synthetically inaccessible as a function of tolerance, the dashed line corresponds to the percent of MOFs discarded at the tolerance for free energy (2.5%) that retains all of the experimentally observed isomorphs in our database. (b) Percentage of synthetically accessible isomorphs that are kept as a function of tolerance, the dashed line corresponds to all 34 isomorphs. The tolerance is defined as being within a certain percent of the lowest free, potential or strain energy predicted within an isomorphic series.

For the 300 K temperature considered here, it turns out that using potential or strain energy in an analogous manner to how we used free energy data in Figure 6.5 to obtain Figure 6.6, yields similar results. That is, stability regions for MOFs could also be identified on the basis of strain or potential energy (Figure D.22), and which could be used to discard a roughly similar number of MOFs deemed synthetically unlikely. However, as noted earlier, a more challenging task (since smaller free energy differences are involved) is to be able to discard low-energy MOFs that have a more stable isomorph.

From the discussion above we already know that on the basis of having the lowest free energy, we would correctly predict the experimentally observed isomorph for 26 (out of 32 isomorphic series) to be synthetically accessible, and would need a 2.5% tolerance to classify all experimentally observed isomorphs as synthetically accessible (at the expense of including false positives). If we were to use potential or strain energy as a proxy for free energy, we would correctly predict the experimentally observed isomorph for 25

Table 6.1 Correlation metrics between different energies calculated for all MOFs in our database. R and S are the Pearson and Spearman Ranking correlation coefficients, respectively, and (MAE) is the mean absolute error.

Correlation	R	S	MAE (kJ/mol)
F_{FL} vs. F_{QHA}	0.997	0.981	1.2
F_{FL} vs. $U_{300\mathrm{K}}$	0.999	0.992	0.5
F_{FL} vs. $U_{0\mathrm{K}}$	0.998	0.984	0.8

and 22 isomorphic series, respectively. Thus, using potential (strain) energy results in one (four) more false negatives than free energy. A tolerance of 3.3% would be needed to classify all the experimentally observed isomorphs in our database as synthetically accessible using potential energy, slightly higher than the tolerance required using free energy. However, a dramatically higher tolerance (93.0%) would be required when using strain energy. Therefore, using strain energy to identify synthetically accessible isomorphs would certainly result in significantly more false positives.

Of course, it would be possible to lower the tolerance when using strain or potential energy as a proxy for free energy, if one were to accept the occurrence of false negatives. As the balance between false negatives and false positives is a recurring subject in efforts to accelerate HTCS, we calculated the percentage of synthetically accessible isomorphs that would be kept in our database subset as a function of tolerance (Figure 6.9b), as well as the percentage of MOFs that can be discarded as they have a more stable isomorph (Figure 6.9a). A loose tolerance would result in fewer MOFs being discarded, fewer false negatives, and more false positives. A tight tolerance allows for more MOFs to be discarded, more false negatives, and fewer false positives. At the tolerance required to retain all the synthetically accessible isomorphs in our database (i.e. an estimate of the tolerance required to eliminate the false negatives), the free energy metric enables the elimination of the largest number of MOFs (i.e. minimizes the number of false positives).

Based on the above observations, we conclude that free and potential energy would definitively outperform strain energy for finding stable isomorphs. Both free and potential energy require low tolerances (2.5 and 3.3%, respectively) to remove false negatives in our data, and would likely perform similarly in the vast majority of cases. Note that the discussion here has focused on calculations at 300 K. However, if one were to calculate free and potential energies at different temperatures, the number of correctly identified experimentally observed isomorphs may change (see Table D.4). Interestingly, for the 32 isomorphic series studied here, calculations at 300 K performed best at predicting the experimentally observed isomorphs among tested temperatures (100 K, 300 K, 350 K, 400 K).

6.5 Conclusions

In this contribution, we demonstrated the first large-scale calculation of MOF free energies, which we performed on a topologically diverse, computationally constructed 8500-MOF database. Based on our results, we recommended using the Frenkel-Ladd (FL) method over the quasi-harmonic approximation for molecular dynamics free energy calculations in diverse MOFs. While we discussed our results based on calculations with the (MOF-specific) UFF4MOF forcefield, generic forcefields UFF and Dreiding performed surprisingly well at replicating trends in free energy observed with UFF4MOF. Thermodynamic stability of the "activated" (i.e. solvent-free) MOF crystals at 300 K was found to correlate strongly (albeit not perfectly) with MOF synthesizability based on the following key observations: (*i*) the 54 previously synthesized MOFs identified in our database all clustered below a free energy threshold of 4.4 kJ/mol, and (*ii*) in 80% of 32 studied isomorphic series, the experimentally observed isomorph had the lowest free energy within its series. As no MOFs with "high" free energy were found to be experimentally reported, "low" free energy seems a necessary, albeit sometimes insufficient, condition for a MOF to be synthetically accessible. Instances where low free energy is insufficient to warrant synthetic accessibility may be due to solvents, modulators, and kinetic considerations that were not inspected here, but that merit closer inspection in the future. Nonetheless, high throughput calculation of MOF thermodynamic stability seems plausible and a key first step towards gauging the synthetic likelihood of MOFs in computational databases.

CHAPTER 7

DEEP LEARNING COMBINED WITH IAST TO SCREEN THERMODYNAMICALLY FEASIBLE MOFS FOR ADSORPTION-BASED SEPARATION OF MULTIPLE BINARY MIXTURES

Modified from a paper submitted to the Journal of Chemical Physics.[14]

Ryther Anderson[15], Diego Gómez-Gualdrón[15,16]

In this chapter, the deep learning model developed in Chapter 4 is used to generate single component isotherms for use in ideal adsorbed solution theory (IAST) calculations. This hybrid machine learning/theory model allows us to rapidly predict the adsorption loadings of diverse binary gas mixtures at multiple conditions. The ability of our combined model to predict separation performance across hundreds of operating conditions allows us to find, using unsupervised machine learning, conditions that are expected to have high-performing MOFs with similar properties. Combining our rapid screening capability with the free energy computation methods and the criteria for thermodynamic accessibility outlined in Chapter 6 yields a powerful MOF-discovery pipeline. Several MOFs that are predicted to be high performing for the separation of two industrially relevant gas mixtures *and* that are predicted to be thermodynamically accessible (but do not yet exist) are shown herein. Moreover, we detected several existing MOFs in our database with similarly high performance.

7.1 Abstract

The structures of metal-organic frameworks (MOFs) can be tuned to reproducibly create adsorption properties that enable the use of these materials in fixed-adsorption beds for non-thermal separations. However, with millions of possible MOF structures, the challenge is to find the MOF with the best adsorption properties to separate a given mixture. Thus, computational, rather than experimental, screening is necessary to identify promising MOF structures that merit further examination, a process traditionally done using molecular simulation. However, even molecular simulation can become intractable when screening an expansive MOF database for their separation properties at more than a few composition, temperature, and pressure combinations. Here, we illustrate progress towards an alternative computational framework that can efficiently identify the highest-performing MOFs for separating various gas mixtures at a variety of conditions and at a fraction of the computational cost of molecular simulation. This framework uses a "multipurpose" multilayer perceptron (MLP) model that can predict single

[14]Reprinted with the permission of AIP Publishing

[15]Chemical and Biological Engineering Department, Colorado School of Mines
[16]Corresponding author, dgomezgualdron@mines.edu

component adsorption of various small adsorbates which, upon coupling with ideal adsorbed solution theory (IAST), can predict binary adsorption for mixtures such as Xe/Kr, CH_2/CH6, N_2/CH_2 and Ar/Kr at multiple compositions and pressures. For this MLP+IAST framework to work with sufficient accuracy we found it critical for the MLP to make accurate predictions at low pressures (0.01-0.1 bar). After training a model with this capability, we found that MOFs in the 95th and 90th percentiles of separation performance determined from MLP+IAST calculations were 65% and 87%, respectively, the same as MOFs in the simulation-predicted 95th percentile across several mixtures at diverse conditions (on average). After validating our MLP+IAST framework, we used a clustering algorithm to identify "privileged" MOFs that are high performing for multiple separations at multiple conditions. As an example, we focused on MOFs that were high performing for the industrially relevant separations 80/20 Xe/Kr at 1 bar and 80/20 N_2/CH_2 at 5 bar. Finally, we used the MOF free energies (calculated on our entire database) to identify privileged MOFs that were also likely synthetically accessible, at least from a thermodynamic perspective.

7.2 Introduction

Approximately half of the energy consumed by the U.S. chemical industry is used to separate mixtures [3]. This is partly due to the dominant role of distillation, a thermal separation process, which accounts for nearly half of the energy used in chemical separations [3]. Therefore, substitution of distillation with non-thermal methods to separate molecules could greatly save energy and reduce CO_2 emissions [3]. Using fixed adsorption beds is one way to separate molecules non-thermally. However, one of the challenges for widespread implementation of these methods is the need for new porous materials that can carry the separation of interest economically. Nanoporous crystals are well suited for adsorption-based separations because their highly ordered structures allow for the rational tuning of their pore sizes, shapes, and chemistries to reproducibly optimize separation properties.

Among nanoporous crystals, metal-organic frameworks (MOFs) are particularly promising (not only for separations) due to their large "design space," whose size arises from the millions of possible MOF variations that there can be [53]. However, this also makes identifying the best possible MOFs for a given application an overwhelming task experimentally. Consequently, over the past decade, high throughput computational screening (HTCS) has emerged as a powerful strategy to identify promising MOFs. HTCS has been especially prominent for adsorption-based gas storage. Some of the best MOFs for hydrogen and methane storage were identified, synthesized, and tested following HTCS—using grand canonical Monte Carlo (GCMC) simulations [37, 38, 40–42]. Additionally, the extensive data from these screenings revealed structure-property relationships (with unprecedented clarity) and apparent "ceilings" for MOF hydrogen and methane adsorption (deliverable) capacities [41, 53, 54]. In principle, a similar screening strategy can

be applied to screen MOFs for adsorption-based separations, but there are additional challenges associated with screening for separation performance. For example, while gas storage depends on single-component adsorption at established T, P values, chemical separation depends on the adsorption of mixtures, whose composition, and T, P values may vary widely depending on the mixture source. Therefore, the best MOF for a given separation could be different depending on the exact operating conditions, likely making coupled MOF-process optimization necessary.

To date, MOF screenings for separation have tended to focus on one or two operating conditions at a time. For instance, Sikora et al. used GCMC to screen a database of 137,000 MOFs for separation of Kr/Xe mixtures at compositions relevant to separation from air (80% Xe, 20% Kr) for three pressures (1, 5, and 10 bar) at 273 K [222]. On the other hand, a separate Monte Carlo (specifically using the Widom insertion method) screening of 125,000 MOFs was conducted by Banerjee et al. for separation of Kr/Xe mixtures at conditions relevant to used nuclear fuel reprocessing, i.e., 400 ppm Xe and 40 ppm Kr at 1 bar and 298 K [43]. Note that these Monte Carlo-based adsorption calculations are relatively affordable—due to the monoatomic, nonpolar gases involved— but even then the computational cost would rapidly become prohibitive if one were to repeatedly screen hundreds of thousands of MOFs for different combinations of composition, T, and P in a single study.

The problem of computational cost is even more serious for other separations of interest, where the relevant Monte Carlo simulations are inherently more expensive. One example is separations involving long hydrocarbons. For instance, the increased cost of simulating the adsorption these types of molecules led to Chung et al. performing (configurational bias) GCMC simulations only for 318 MOFs (at one T, P and composition) as part of hierarchical screening of 5000 MOFs for separation of hexane and heptane isomers [290]. Similar issues are encountered when considering separations involving water, where several works [39, 289, 344, 345] have devised clever hierarchical screening strategies to minimize the number of MOFs for which adsorption of water-containing mixtures is directly simulated.

The above indicates that the development of alternative methods that can assess the separation ability of MOFs faster than GCMC simulations could revolutionize MOF screening for separations. Accordingly, there is an ever growing interest in developing machine learning (ML) models, trained on GCMC data, to predict MOF adsorption-based separation capabilities [22, 96, 97]. As examples, in early efforts, Simon et al. trained random forests (RFs) to predict selectivity for 80/20 molar Kr/Xe mixtures at 298 K and 1 bar [22]. In more recent efforts, Dureckova et al. used tree-based models to predict selectivity and working capacity for separation of 40/60 CO_2/H_2 mixtures for one operating conditions (a 40 bar $\leftrightarrow$ 1 bar swing at 313 K), training a different model for each metric [96]. Burner et al. also trained separate artificial neural networks (ANN) to predict selectivity and working capacity for 15/85 CO_2/N_2 mixtures at 1 bar and 298

K [97].

The caveat is that ML models trained as described above have suffered from a lack of "reusability." Briefly, the ML models are trained from GCMC simulation to predict separation metrics for a given MOF database for specific components, compositions, and operating conditions. However, since the separation properties of the given MOF database were already known from the "training" GCMC data, the motivation to reuse the ML model is nullified and the computational time "investment" to train the model is not recovered. In contrast, a given ML model could find widespread application if it were transferable to components, compositions, and operating conditions on which it was not trained. Thus, most recently, studies using "transfer learning" and "multipurpose" models are emerging [101–103, 346].

Anderson et al. [101] showed that an ANN could predict H_2 adsorption in MOFs at T, P values different from those included in the training data. Sun et al. demonstrated a similar capability for an ANN to predict sorption of pentanediol/water on zeolite MFI, and also illustrated the reuse of some ANN layers on a new model that predicts sorption when changing either the adsorbent or one of the adsorbates [103]. Ma et al. reused some layers of an ANN trained to predict H_2 sorption in a new ANN trained to predict CH_2 sorption [346]. Using genetic algorithm regression and adsorption data for multiple molecules in a subset of previously synthesized MOFs, Gharagheizi et al. obtained an equation capable of predicting isotherms for a diversity of $C_xH_yS_uO_vN_w$ molecules [102]. These authors used this equation to screen MOFs for many chemical separations. However, since the equation did not predict mixture adsorption, adsorption selectivities were approximated as ratios of single-component adsorption loadings at the corresponding single component vapor pressure.

In our own efforts to develop "reusable" models, we trained a deep learning, multipurpose model (a multilayer perceptron or MLP) capable of predicting single-component adsorption for a diversity of small, nonpolar adsorbates across different MOFs and pressures [84]. The key to the development of this model is the inclusion of adsorbate properties and fugacity (along with MOF properties) as model inputs, allowing the model to learn the underlying physics of adsorption, agnostic to the chemistry of the adsorbates, and the use of adsorption data for alchemical species to increase the diversity of the training data. In fact, no data for real molecules was used in training. The caveat is that despite its versatility, much like the multipurpose model by Gharagheizi et al. [102], our model could not directly predict mixture adsorption data.

Intriguingly, however, if one were to use single component isotherm predictions from ML models as inputs for established thermodynamic theories such as Ideal Adsorbed Solution Theory (IAST), one would be able to generate mixture adsorption data for any composition at unprecedented speed. The potential issue is that while the difficulty in measuring mixture adsorption has made IAST popular with

experimentalists, the accuracy of IAST is a subject of debate [347–350]. Indeed, there are assumptions in IAST (such as interactions of all mixture components with the adsorbent being identical) that one would expect to break under the selective adsorption scenario that one pursues for adsorption-based separations. However, considering the transferability of IAST, and the numerous demonstrations of its sufficient accuracy for calculating multicomponent loadings of small molecules in MOFs and other porous crystals (albeit with some error) [347–349, 351, 352], combining IAST with a machine learning model that can predict full isotherms for multiple adsorbates has would allow us to (almost) instantly screen MOFs for hundreds of separation conditions. This motivated to test and demonstrate the efficacy of this approach.

To achieve this goal, we leveraged a herein updated version of our previously trained multipurpose MLP model [84], which predicts single-component isotherms for components of a variety of mixtures relevant to chemical separations, and our ability to directly verify the accuracy of IAST via GCMC simulations of mixture adsorption. We chose to focus on the accuracy of the MLP+IAST framework to screen a 51,520 MOF database for their ability to separate the binary mixtures Ar/Kr, Kr/Xe, CH_2/C_2H_6, and N_2/CH_2 at different molar compositions and pressures. The components in these mixtures present different degrees of similarity, providing different scenarios under which to test IAST validity. Additionally, these mixtures are industrially/experimentally relevant. As noted earlier, Kr/Xe separation is relevant to Xe extraction from air [5] and used nuclear fuel reprocessing [5, 18]. N_2/CH_2 separation is relevant to upgrading natural gas, as many deposits have ¿10% N_2 content [155, 353], which needs to be reduced below 5% before use. N_2 selective adsorbents are generally desirable for natural gas upgrading since the majority component is typically CH_2. Such adsorbents rely on a sieving mechanism to (ideally) entirely exclude CH_2 [11]. However, CH_2 selective adsorbents are useful to purify N_2-rich streams from a previous cryogenic distillation [156] or other methane-poor sources, such as land fill gas [158] (to prevent the atmospheric emission of CH_2). CH_2/C_2H_6 separation is also relevant to natural gas processing [159]. Natural gas often contains large amounts of ethane, which is a precursor for the synthesis of ethene, an important industrial chemical [7]. Finally, Ar/Kr separation is relevant to separating noble gases from air [5, 348], and has niche applications in radioisotope dating, as isotopes of Kr and Ar are excellent environmental tracers, mainly due to their inertness [354, 355].

To assess MOF separation performance, here we calculated the hybrid selectivity-capacity metric $\Delta Q_{B/A}$ defined by Krishna [23], which aims to anticipate relative MOF performance in breakthrough experiments normally done during the lab testing phase of MOFs for separation. Thus, based on this metric, we specifically investigated the ability of MLP+IAST to i) correctly capture top-performing MOFs within our database for direct material discovery, and ii) faithfully reproduce structure-performance relationships that can guide rational material design. Subsequently, using a clustering algorithm, we

examined the similarity among top-performing MOFs for different separations and different operating conditions, thus revisiting the concept of "privileged" adsorbents introduced earlier by Tang et al. [294] Finally, we complemented our MOF performance predictions with large-scale free energy calculations, using our recently developed computational framework [356], to identify which MOFs among top performers are more synthetically likely, at least from a thermodynamic point of view.

7.3 Computational Methods

7.3.1 Database Construction

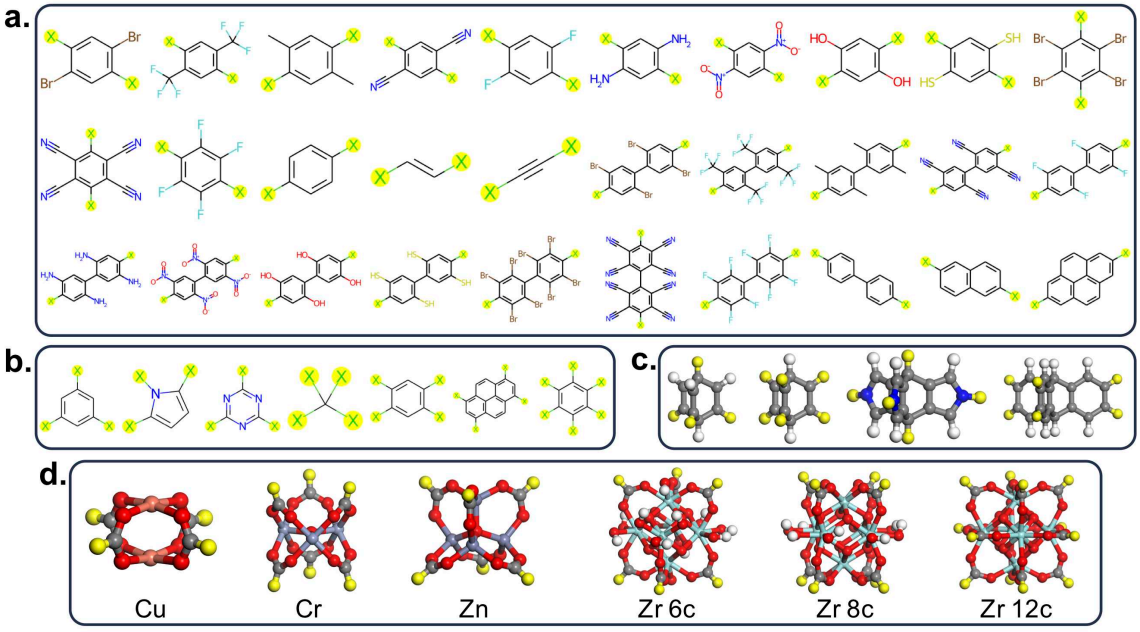

Figure 7.1 The computational MOF building blocks used to construct our 51,520 MOF database. (a) Connecting (so-called "edge") building blocks. (b) Planar organic nodular building blocks. (c) Non-planar organic nodular building blocks. (d) Nodular inorganic building blocks. Connection sites "x", where bonds are formed with other building blocks, are highlighted in yellow.

We used our graph-theory based ToBaCCo-3.0 code [73, 77] to build a 51,520 MOF database featuring MOFs in 424 different topologies. The building blocks for MOF construction are shown Figure 7.1. The geometry of each MOF built by ToBaCCo was optimized using LAMMPS (version 31 March 2017) [211] with interactions described by the UFF4MOF forcefield [82, 210].We used an iterative approach for the optimizations, where in each iteration the atom coordinates were first optimized keeping the MOF lattice parameters fixed, and then the atom positions and lattice parameters were optimized together. Iterations continued until the energy per atom changed by less than 1.0×10^{-4} kcal/mol. The individual optimizations were considered converged when no atom experienced a force large than 1.0×10^{-6}

kcal/(mol Å^{-1}) and the energy between one geometry and the next changed by less than 1.0 $\times$ 10^{-4} %. The fast inertial relaxation engine algorithm of Bitzek at al. [212] was used when optimizing only the atom positions and a steepest descent algorithm was used when optimizing the atom positions and lattice parameters together.

7.3.2 Adsorption Simulations

Adsorption loadings were simulated using GCMC as implemented in RASPA-2.0 [91]. In these simulations, the chemical potential (μ), volume (V) and temperature (T) are kept constant, while the number of molecules (N) varies. To obtain single-component adsorption loadings used to train our neural network (see below), we directly used adsorbate fugacity as input, which RASPA then used to calculate chemical potential. To obtain mixture adsorption data to evaluate our MLP+IAST model, target pressures were first converted to fugacities using the Peng-Robinson equation of state, which were then used as input for the simulations. Single component GCMC simulations used 2,000 equilibration cycles, followed by 2,000 production cycles to calculate ensemble averages. Binary mixture GCMC simulations used 3,000 equilibration cycles followed by 3,000 production cycles. Adsorbate-adsorbate interactions were modeled using Coulomb and Lennard-Jones (LJ) interactions, while adsorbate-framework interactions were modeled using only LJ interactions. Lorentz-Berthelot mixing rules were used. Long-range electrostatic interactions between adsorbates were modeled using Ewald summation with a precision of 1.0×10^{-5}, while long-range LJ interactions were neglected beyond a cutoff of 12.8 Å. The LJ parameters for all real adsorbates where the same as those used in Chapter 5 (see Appendix C). The LJ parameters for MOF atoms were assigned from the Dreiding forcefield [83] if available, otherwise UFF [80] parameters were used.

7.3.3 MLP+IAST Model Construction

We retrained our previously published single-adsorption-predicting multi-layer perceptron (MLP) model [84], which was originally trained with adsorption loadings at fugacities between 1 and 100 bar, to now include lower fugacities (as here IAST required adsorption data starting at sub-atmospheric pressures). We used broadly the same training framework detailed in our previous work[23] with a few key differences (noted here). Specifically, while we used adsorption loadings of the same 200 alchemical adsorbates on the same 2,400 MOFs, we now included fugacities as low as 0.01 bar (see Table E.1 for the training, validation, and testing fugacities used). We also employed the same 200-MOF validation set, the same adsorbate features as MLP inputs, and similar MOF features. However, while we used the same six MOF textural properties, namely void fraction (VF), gravimetric surface area (GSA), largest pore diameter (LPD), pore-limiting diameter (PLD), and pore size standard deviation (PSSD), as in ref. [84], we updated the

chemical features to use the UFF4MOF atom types found in our 51,520-MOF databases. We use this typing scheme as it gives more implicit information about the chemical environment of the MOF atoms than the Dreiding/UFF atom types. The MOF atom types considered (with a brief description) are listed in Table E.2. LPDs and PLDs for every MOF were calculated using zeo++ (version 0.2.2) [85]. The GSAs, VFs, and pore size distributions were calculated using RASPA-2.0 with a N_2 sized probe for surface area calculations and a helium probe for VF calculations. Each chemical motif number density was calculated by dividing the number of each motif type present in a given MOF by the MOF unit cell volume in $Å^3$.

Another difference with our previous MLP training was that here we used log of fugacity as input instead of fugacity itself. We used the TensorFlow Python library to build and train our MLP. All MLP training details and hyperparameters are given in Appendix E and Table E.3. The learning curve for our MLP is shown in Figure E.1 and a comparison of our MLP validation and test predictions is shown in Figure E.2. Co-adsorption loadings were calculated from predicted single component isotherms using the pyIAST Python package written by Simon et al. [15] IAST requires single-component adsorption isotherms of each component of the mixture as input to estimate mixture adsorption. Traditionally, a model (e.g. a Langmuir model) with an analytical integral is used to provide single component loadings at any required pressure, but it is also possible to linearly interpolate single-component isotherms and calculate integrals numerically [15, 348]. This latter method has the advantage of not making assumptions about the shapes of the isotherms but requires sufficient "resolution" in adsorption data for linear interpolation and integral calculation to be accurate. We use our retrained MLP-model to do just that, providing single-component isotherms consisting of adsorption loadings at 73 pressures between 0.01 and 100 bar. Given that our MLP works with (logs of) fugacity as input, pressures used in IAST calculations were converted to fugacities using the Peng-Robinson equation of state before being input to the MLP. All code versions used for our MLP training and MLP+IAST calculations are given in Table E.4.

7.3.4 Model-Based Clustering

For the clustering study discussed in Section 3.4, we used a model-based clustering approach as implemented in the R package mclust [175] (version 5.4.7 with R version 3.6.2). In model-based clustering the data are assumed to be sampled from a mixture of probability distributions, typically from the same parametric family. The mclust package is specifically for Gaussian Mixture Modeling (GMM) [175], in which each observation (data point) is assumed to be sampled from a multivariate-normal distribution. Modeling the data with k distributions corresponds to categorizing the data in k clusters. One of the main advantages of model-based over k-means or hierarchical clustering is that each observation can be assigned a probability of belonging to each cluster, i.e., the cluster assignment is "soft", which allows for more

refined analysis of the data structure. More details on model-based clustering and our use of the mclust algorithm are given in Chapter 2.

7.3.5 Free Energy Calculations

We calculated the free energies used below using the Frenkel-Ladd (FL) path method [321]. The FL path is a thermodynamic integration approach used to calculate the absolute free energies of crystalline solids (in our case MOFs). Essentially, the free energy difference between an Einstein crystal (EC) representation of a MOF (in which each atom is tethered to its equilibrium position with a harmonic restraint and does not interact with the other framework atoms) to a flexible forcefield representation of the MOF is calculated using thermodynamic integration. Then, since the free energy of an EC is known analytically, the absolute free energy of the MOF is obtained as the sum of the calculated free energy difference and the free energy of the EC. The specifics of this calculation are given Chapter 6. We used the same approach as described therein for all free energy calculations.

7.4 Results and Discussion

7.4.1 Predicting Single-Component Loadings at Low Fugacities

There are two major reasons why an MLP+IAST framework could potentially give significantly different adsorption loadings than direct GCMC simulation of mixture adsorption: (*i*) shortcomings in the IAST framework and (*ii*) inaccurate isotherms used as input for IAST. Our previously trained MLP was capable of accurately predicting single-component adsorption loadings of a variety of small-gas adsorbates in MOFs at fugacities between 1 and 100 bar [84]. But, numerous small gas separations depend on mixture adsorption in the 1 – 10 bar range which, if it is to be predicted with IAST, requires accurate single-component loadings for pressures around the 0.01-0.1 bar range to be accurate This is because IAST calculations heavily depend on the shape of the single component isotherm, especially at (relatively) low pressures. Therefore, we retrained our previous MLP model to make accurate predictions at fugacities as low as 0.01 bar, which required adding adsorption data at low fugacities to the original 1-100 bar data.

To test the prediction capabilities of the new MLP model, we compared MLP predictions for six adsorbates in 1,000 test MOFs at 15 fugacities with the results from GCMC simulations. We started by comparing individual loading predictions for 15 different fugacities between 0.02 and 100 bar as shown in the parity plots in Figure 7.2 The plots are presented in log scale to emphasize the comparison for the new predictions at low fugacities, although it should be noted that we retained similar prediction accuracy at high fugacities as in our previous MLP. Four prediction performance metrics were considered for the data in Figure 7.2: the median absolute error (MdAE) median absolute percentage error (MdAPE), the Pearson

correlation coefficient (R), and Spearman rank correlation coefficient (S), all calculated for each fugacity (see Figure E.3).

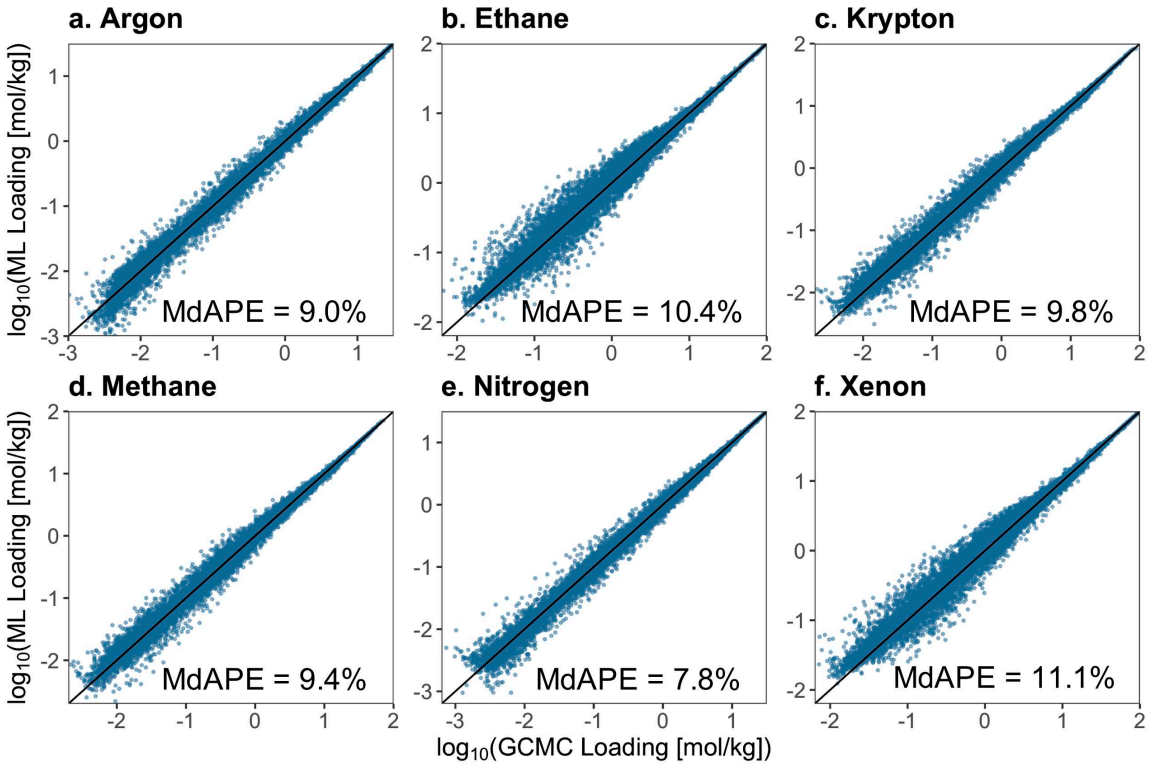

Figure 7.2 Machine learning (ML) predictions versus GCMC predictions for single-component loadings of the indicated adsorbates in our 1,000 test MOFs. Loading values are compared for each MOF corresponding to 15 different fugacities between 0.02 and 100 bar. The MdAPE calculated across all fugacities is shown for each plot.

We used median errors instead of mean errors because the error distributions herein have long tails, making the medians a better measure of central tendency for the error. Furthermore, for small loadings, percentage errors can be very large even for accurate predictions, e.g., predicting 0.01 mol/kg as 0.02 mol/kg would correspond to a 100% error, despite being an accurate prediction. Thus, MdAPE is a more reliable metric than mean absolute percentage error (MAPE) in cases (such as this one) where predictions can span orders of magnitude. As for the correlation coefficients, R values near unity indicate a strong linear relationship between the MLP- and GCMC-predicted values, whereas S values near unity indicate that the ranking of MOFs using MLP predictions is largely similar to that obtained using GCMC data.

Our MLP model MdAE increased with fugacity for all adsorbates, the MdAPE error decreased, and the correlation coefficients got closer to unity. This makes intuitive sense as at low fugacities (0.02-0.08 bar) loadings tend to be small ($<$ 1 mol/kg) and with low variance (e.g., the spread of loading values across MOFs is smaller at 0.02 bar than at 1 bar). This means that even relatively small absolute errors can

result in large percentage errors and significant deviations from linearity/rank correspondence. Nonetheless, for each fugacity (including low fugacities) the MdAEs for each adsorbate are all considerably less (generally by order of magnitude) than the corresponding median GCMC loadings. Accordingly, we were satisfied with the predictions of the retrained MLP model. Notably, this illustrates the advantage of multipurpose models. In our case, the ability of our original MLP to already predict adsorption at different fugacities made it so that to make the model transferable to lower fugacities we could still reuse old training data only needing to add data for new fugacities (we expect an analogous scenario if were to add data for, say, new types of adsorbates).

Crucially, while an MLP model that can accurately predict loadings at individual pressures for a variety of adsorbates is useful, to be able to couple the model with IAST, full isotherms must also be accurately predicted (especially capturing the isotherm curvature). Thus, to understand how well our model predicts full isotherms we calculated the isotherm mean absolute percentage error (IMAPE) for each of our 1000 test MOFs. IMAPE is defined as the mean of the absolute percentage errors (MAPE) of each point in an isotherm (in this case the 15-point isotherms we used to test our model). Contrary to the performance metrics for individual predictions, we use means instead of medians for the isotherm percentage errors because in this case we are not interested in the central tendency of the errors in each isotherm, but rather we are seeking an aggregate measure of the quality of the "fit" of each isotherm. The latter can be interpreted as the distance between the simulated isotherm and the corresponding predicted isotherm.

Next, for each adsorbate, we calculated the 0.00 (Q_1), 0.25 (Q_2), 0.50 (Q_3), 0.75 (Q_4), and 1.00 (Q_5) quantiles of IMAPE and plotted the isotherms of the MOFs nearest each quantile (Figure 7.3). The Q_1 isotherm can be thought of as the "best" prediction, the Q_5 as the worst, and the Q_3 as an average (median) prediction (according to the IMAPE values). From Figure 7.3 we can observe our model generally reproduces isotherm shapes well, without noise or unphysical behavior (i.e., loading decreasing with fugacity). Even the "worst" predictions tend to be qualitatively correct, albeit with large systematic errors. Figure E.4 shows the same data as Figure 7.3 but on a log-log scale, which is better for observing deviations for small loading values.

At this point, it is worth emphasizing that the prediction performance metrics discussed so far are for 1,000 MOFs that were not present in the training or validation set and have many organic building blocks and hundreds of topologies that were not present in the training set. Thus, we expect our MLP model will achieve similar accuracy when applied to other MOFs with a diversity of chemistries and topologies that were not present in the current work (although in our experience ML predictions seem to be relatively more sensitive to the chemistry of the inorganic nodes).

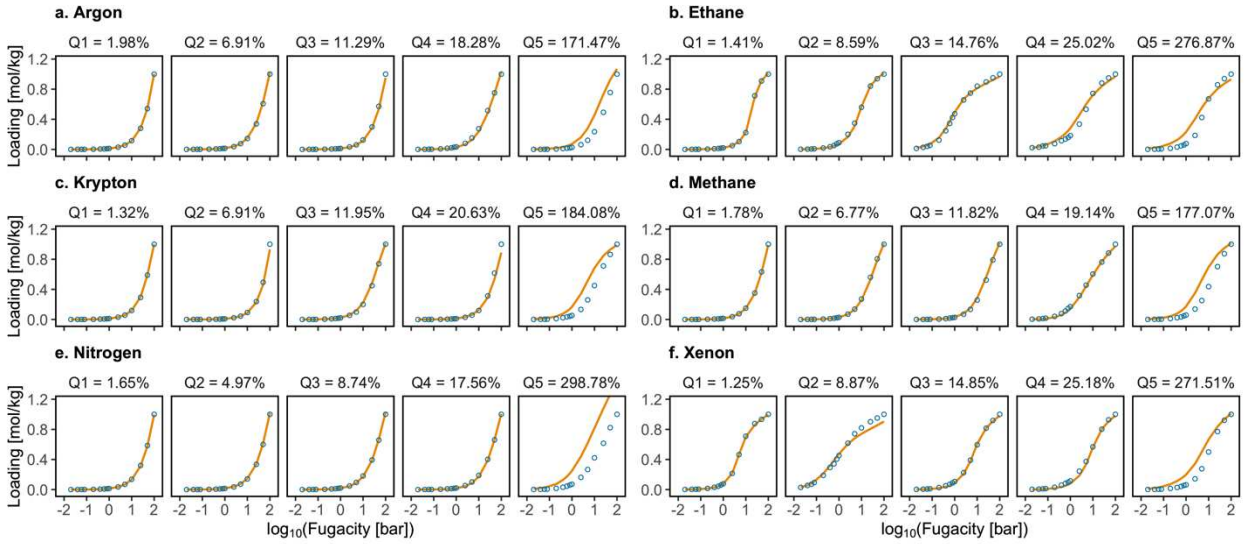

Figure 7.3 Comparison of our MLP-predicted isotherms for the six indicated adsorbates to GCMC simulated isotherms for the 0.00 (Q_1), 0.25 (Q_2), 0.50 (Q_3), 0.75 (Q_4), and 1.00 (Q_5) quantiles of the IMAPE. The open, blue points are GCMC simulated values and the orange lines are MLP predictions.

7.4.2 Combining Neural Network Predictions with IAST

With a promising MLP model completed, we then proceeded to evaluate the efficacy of combining it with IAST to screen MOFs for separation processes using fixed adsorption beds. For such application we are concerned with both the selectivity and adsorption capacity for the strongly adsorbed component. Screening materials for either metric individually often results in contradicting rankings as it has become clear that materials with the highest selectivities tend to have low capacities (and vice versa) [19, 23]. Consequently, screening is best done using a combined selectivity-capacity metric. Here we use the separation potential ΔQ, a hybrid metric recently defined by Krishna [23], see Equations 1.4, 1.5, and 1.6. A detailed derivation of ΔQ is given in ref. [23]. To assess the accuracy of MLP+IAST, then, we compared MLP+IAST predictions of loadings and $\Delta Q$for four binary gas mixtures (Ar/Kr, Kr/Xe, CH_2/C_2H_6, and N_2/CH_2) at four compositions (20/80, 40/60, 60/40, and 80/20) and at 10 pressures from 1 to 10 bar with the corresponding metric calculated from binary adsorption GCMC simulations in the same 1,000 test MOFs we focused on in the previous section. From this point on we refer to a pressure, composition and mixture combination as a "condition," in which case in this section we considered a total of 160 conditions for our tests. Parity plots comparing the binary loadings and ΔQs predicted by our MLP+IAST model with the corresponding GCMC results are shown in Figure 7.4 for all the conditions considered. Further, we show four error metrics for each condition in Figure E.5 to Figure E.8. These metrics include the previously described MdAPE, R, and S, as well as an additional ranking metric described below. The

MdAPEs for each condition were all (generally, well) below 25% and R and S values were all above 0.80 for both loading and ΔQpredictions.

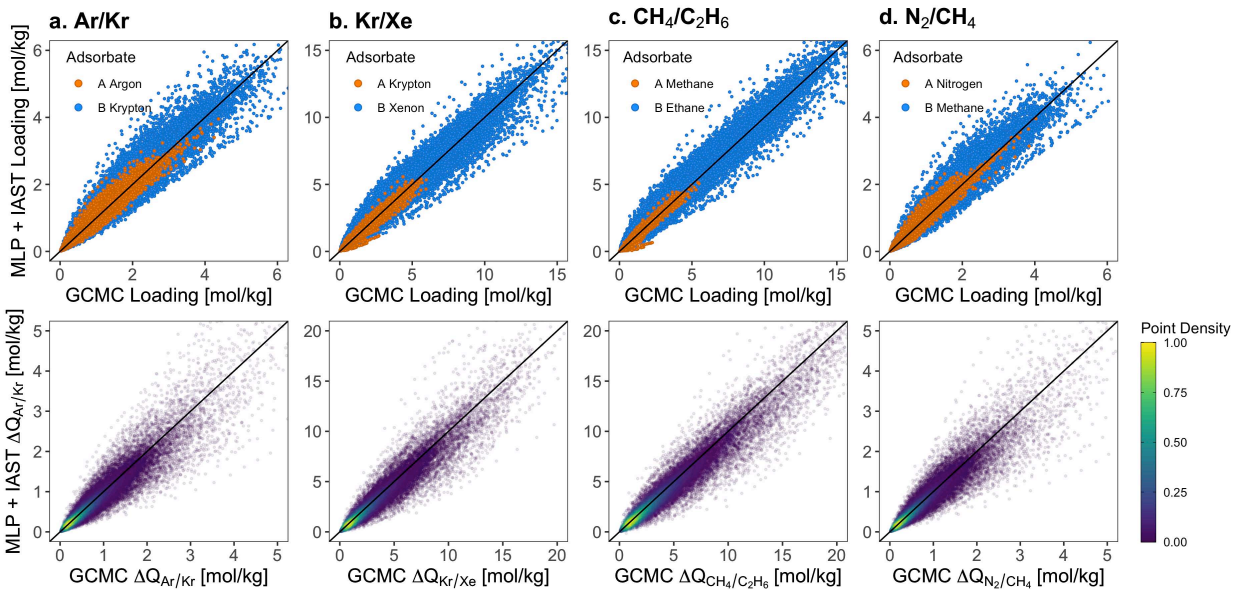

Figure 7.4 MLP+IAST predictions of the loadings of each component (top) and $\Delta Q_{\mathrm{B/A}}$ (bottom) verses GCMC predictions for the indicated binary mixtures. The loading plots are colored according to the component and the $\Delta Q_{\mathrm{B/A}}$ plots are colored by point density. These plots show values for all 10 pressures (1 to 10 bar) and mixture compositions (20/80, 40/60, 60/40, 80/20) considered.

Ultimately, for HTCS studies, provided that ΔQpredictions are "reasonable," one is arguably more interested in correctly capturing the highest performing MOFs so they can be suggested for synthesis and experimental testing. While to some extent S can gauge whether this is true for our models, we find that it is more informative to directly compare the top N MOFs predicted by our MLP+IAST framework with the top M MOFs predicted by GCMC for increasing values of N (with $N \geq M$, and N and M being integers). Specifically, we considered how many GCMC-predicted top 50 ΔQMOFs (i.e. $M = 50$ corresponding to the 95th percentile) were identified when considering the MLP+IAST-predicted top N MOFs for each condition. This equates to 8,000 top performing MOFs (50 for each of the 160 studied condition). These results are summarized in Figure 7.5.

When $N = 50$ we found that between 60 and 70% of GCMC-predicted top 50 MOFs were identified (depending on the condition). By increasing N so $N = 100$ (i.e. considering the 90th percentile), we found that between 80 and 95% of the GCMC predicted top 50 MOFs were identified, with an average (mean) of 87.5% correctness across conditions. Thus, in screening MOFs for chemical separations, one can expect (on average) to capture 87.5% of the MOFs in the actual 95th percentile (top 50 for 1000 MOFs) if one considers MOFs in the 90th percentile of ΔQvalues predicted by MLP+IAST. In other words, 90% of

MOFs in a database could be discarded by our MLP+IAST framework, if this were to be used in a first step of a hierarchical screening study.

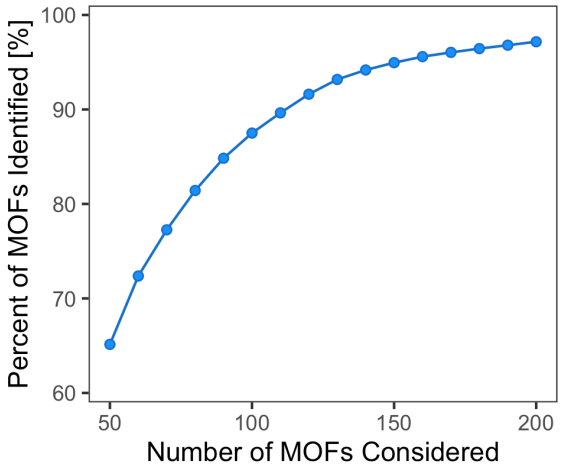

Figure 7.5 The percent of GCMC-predicted top 8000 ΔQMOFs (y-axis) identified (50 top MOFs for each of the 160 conditions considered) when considering the top N MLP+IAST-predicted MOFs for each condition (x-axis).

7.4.3 Predicting Structure-Performance Relationships

While direct discovery of top-performing MOFs is the central goal of HTCS, it is not the only goal. Structure-performance relationships are an invaluable secondary result of HCTS, which can provide "clear-cut" design rules to guide rational material design experiments. The clarity of the derived rules hinges on a large amount of data from which structure-performance relationships are visualized, even to the point of revealing potential physical limits for a given performance metric [37, 38, 53, 54, 163]. Thus, a predictive model that replaces GCMC simulations should—in addition to being quantitively accurate enough for use in hierarchical screening—correctly reproduce these relationships. Here, we tested that ability for our MLP+IAST framework to do this. For various conditions, we plotted MLP+IAST and GCMC ΔQvalues versus MOF properties on the same graph. Additionally, we compared the mean and standard deviation of $\Delta Q$for MOFs in the 95th percentile for each condition using MLP+IAST and GCMC.

As an example, Figure 7.6 shows MLP+IAST and GCMC $\Delta Q_{\mathrm{Kr/Xe}}$ versus MOF void fraction (VF) for various pressures and mixture compositions. Clearly, the MLP+IAST predictions closely reproduce the GCMC-predicted relationships for these conditions. Figure E.9 to Figure E.11 show analogous examples for the rest of gas mixtures we considered in this work. We also found that the mean and standard deviations of VF, LPD, PLD, GSA, and density for MOFs in the 95th percentile of MLP+IAST ΔQs matched the analogous values for GCMC ΔQs very closely for all our considered conditions (see Figure E.12

to Figure E.16), indicating that our MLP+IAST model reproduces the relationships of $\Delta Q$with other MOF properties with high fidelity. Accordingly, one would draw the same conclusions about which MOF properties generally yield high performance regardless of whether one uses our MLP+IAST model or GCMC simulations to screen a MOF database. Notice that both the spread and central tendency of the properties of top MOFs are informative because relationships with sharp peaks (as in the $\Delta Q_{\mathrm{Kr/Xe}}$ verses VF relationship) potentially indicate that the MOF property under consideration is more important for high performance than another property with a broad peak in the structure-performance relationship.

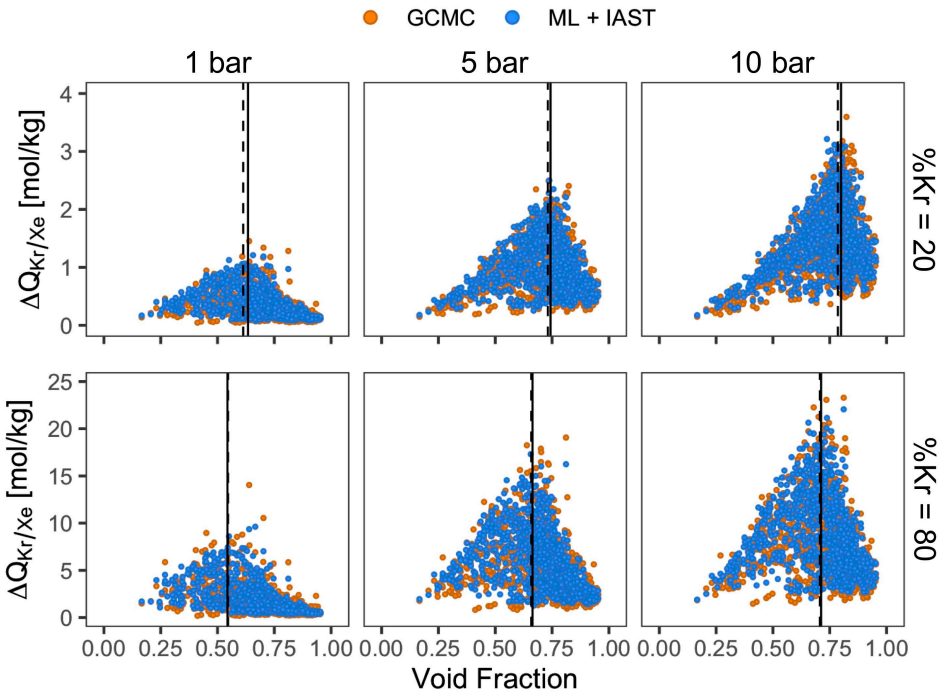

Figure 7.6 Plots comparing the relationship of $\Delta Q_{\mathrm{Kr/Xe}}$ versus MOF void fraction as predicted by GCMC (orange) and our MLP+IAST model (blue). The mean VF of MOFs in the 95th percentile of $\Delta Q_{\mathrm{Kr/Xe}}$ is shown by the solid line for GCMC $\Delta Q_{\mathrm{Kr/Xe}}$, and by the dashed line for MLP+IAST $\Delta Q_{\mathrm{Kr/Xe}}$.

7.4.4 Identifying for What Separations there may be Privileged MOFs

After validating the predictive ability of our MLP+IAST model in the previous sections, in this final section we use our model in an example screening study. We highlight some of the unique results that can be garnered by harnessing the transferability of our model to a wide variety of conditions. We screened a database of 51,520 MOFs for separation performance (assessed with ΔQ) for four gas mixtures (Ar/Kr, Kr/Xe, CH_2/C_2H_6, and N_2/CH_2) at two compositions (80/20 and 20/80) and five pressures (1.0, 2.5, 5.0, 7.5 and 10.0 bar). This equates to 2,060,800 data points, which would take 2,000,000 CPU hours to

calculate using GCMC (assuming 1 CPU hour per simulation). In contrast, we were able the generate these data in 12.6 CPU hours using our MLP+IAST framework. We decided to use this data to identify MOFs that are expected to be "privileged" and be high performing for multiple conditions (recall we define a condition as a pressure, composition, and mixture combination). We accomplished this by calculating the mean properties of MOFs in the 95th percentile of $\Delta Q$for each condition and then clustered the conditions based on these mean properties.

Conditions that are in the same (or similar) clusters are expected to have top-performing MOFs with similar properties, and the property distributions of each cluster yield information useful for designing materials that are useful for multiple separations. Specifically, we considered the MOF largest pore diameter (LPD), pore limiting diameter (PLD), gravimetric surface area (GSA), void fraction (VF), density, and pore size standard deviation (PSSD) as the clustering variables and used a model-based clustering algorithm as implemented in the mclust R package. We used these variables as we expect the ΔQvalues for the conditions we considered to be primarily influenced by MOF textural properties and because these variables have underlying distributions that can be modeled.

The results of the clustering are summarized in Figure 7.7 which show the clustered conditions projected onto the first three directions resulting from the dimension reduction algorithm of Scrucca (these directions account for the majority of the structure in the clustered data) [357]. This algorithm finds a set of linear combinations of the features considered during clustering that results in maximal cluster separation. Figure 7.7 also shows the boundaries between the classification regions defined for each cluster, colored according to the boundary uncertainty (darker areas indicate higher uncertainty). For example, we can observe from Figure 7.7that clusters 2 and 3 have some close points, and that two conditions lie on the boundary between clusters 2 and 3. The conditions 80/20 Kr/Xe at 1 bar and 80/20 N_2/CH_2 at 5 bar are both in cluster 3. This means that MOFs with high ΔQs for these separations are likely to be similar, a significant result as both these conditions are relevant to important separation conditions. In other words, this type of analysis can be used to identify privileged MOFs, i.e. MOFs that are high performing for multiple separations and/or multiple T, P combinations. An additional plot justifying this conclusion using an alternative method is shown in Figure E.17.

7.4.5 Using MOF Thermodynamic Feasibility Calculations in HTCS

After identifying potentially privileged MOFs using the methods outlined in the previous section, we can proceed to use a free energy threshold to identify MOFs more likely to be synthesizable on account of having similar free energy per atom as other previously synthesized MOFs [356]. While the exact criterion we outlined in ref. [356] may not necessarily be transferable to all MOFs in general, it does apply to all the

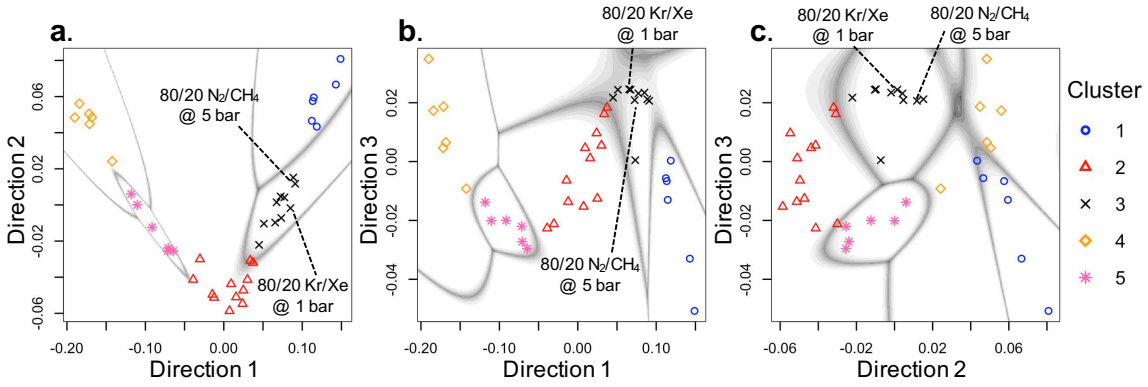

Figure 7.7 Clusters of conditions as clustered by their mean high-performing MOF properties. Each cluster is shown as a different shape and color. The clusters are projected onto directions found from a cluster-based dimensionality reduction algorithm that seeks to find linear combinations of the mean MOF properties that results in maximal cluster separation. Here the first three of four directions are shown in (a), (b), and (c). These three directions account for the majority of the structure in the clustered data. Uncertainty boundaries between the clusters are shown in grayscale, the darker the area the higher uncertainty. The specific conditions 80/20 N_2/CH_2 at 5 bar and 80/20 Kr/Xe at 1 bar are directly labeled.

structures in our 51,520 MOF database as they are based on the same type of inorganic nodes. This single, global criterion, however, is not the complete picture, as if there is isomorphism (same nodes and linkers, but a different topology) observed for a particular MOF composition, the isomorph with the lowest, calculated free energy has around 80% likelihood (based on 32 isomorphic families in ref. [356]) of being the one observed experimentally.

Having calculated the ΔQs across our database using our MLP+IAST model for 80/20 Kr/Xe mixtures at 1 bar and 80/20 N_2/CH_2 mixtures at 5 bar, we can proceeded to identifying MOFs that are high-performing for both separations (recalling the results of the clustering analysis in the previous section), while simultaneously considering the thermodynamic feasibility of these high-performing MOFs. To accomplish this, we calculated the free energies of each structure in our database using the Frenkel-Ladd path method. Following findings from our previous work we considered MOFs with free energy below an approximately 4 kJ/mol per atom threshold thermodynamically feasible.

Using the free energy criteria described above, in Figure 7.8 we show $\Delta Q_{\mathrm{Kr/Xe}}$ versus $\Delta Q_{\mathrm{N_2/CH_4}}$, with MOFs that are predicted to be simultaneously thermodynamically feasible and in the 90th percentile of both ΔQs, these MOFs correspond to the "star" points. We considered the 90th percentiles of our MLP+IAST predictions as our results from earlier suggest that this would allow us to capture the vast majority of the 95th percentile of MOFs that would be identified using GCMC. Using this data, we were able to identify several promising structures that were ranked highly for both separations and were thermodynamically accessible. We used the average of the Kr/Xe and N_2/CH_2 rankings to obtain a single

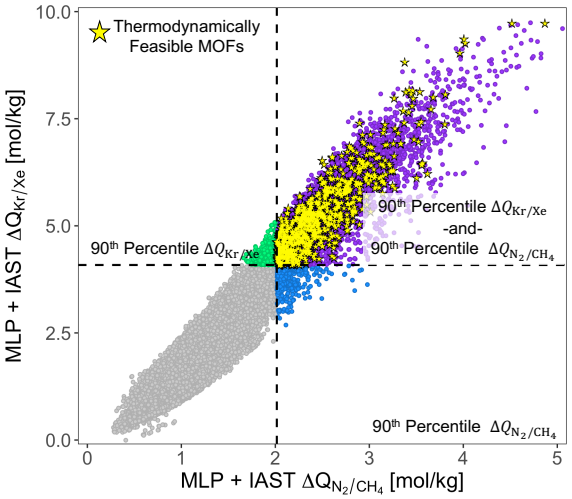

Figure 7.8 $\Delta Q_{\mathrm{Kr/Xe}}$ for an 80/20 composition at 1 bar verses $\Delta Q_{\mathrm{N_2/CH_4}}$ for an 80/20 composition at 5 bar, both values were predicted using our MLP+IAST model. Cutoffs for to be in the 90th percentile for both ΔQs are shown and thermodynamically accessible MOFs in both regions are highlighted as yellow stars.

ranking for both separations. Six of the thermodynamically feasible MOFs with high rankings for both separations are shown in Figure 7.9 . Reassuringly for the validity of our criterion, we were to connect these structures with previously synthesized MOFs.

The MOF in Figure 7.9a (average rank = 197.5) of **edq** topology is similar to the existing MOF Co-ABTC synthesized by Duan et al (note that **edq** topology MOFs are often reported as having soc topology) [358]. Various isoreticular versions of this edq MOF, with similarly sized linkers, also have high performance for both separations with average ranks as high as 33. The MOFs in Figure 7.9b (average rank = 16.5) and Figure 7.9c (average rank = 102) are isomorphs that were determined to both be thermodynamically accessible (their free energies were with 1% of each other) and are isoreticular with the existing MOF-150 (**pyr**) [359] and MOF-177 (**qom**) [30], but with a contracted linker leading to smaller pores. Conveniently, both thermodynamically feasible isomorphs are ranked highly in this case. The MOFs in Figure 7.9d (average rank = 13) andFigure 7.9e (average rank = 9) are also isomorphic and both considered thermodynamically feasible. The **jjt** structure is isoreticular with the existing MOF JUC-101 [360], but with a smaller linker. Finally, the MOF inFigure 7.9f (average rank = 99) is an existing MOF previously synthesized by Chen et al.[361] We also found that the analogs of the existing MOFs NOTT-100 [213] and UTSA-20 [362] in our database have reasonably high performance with average ranks of 835 and 605, respectively (note that these are ranks out of over 50,000 MOFs).

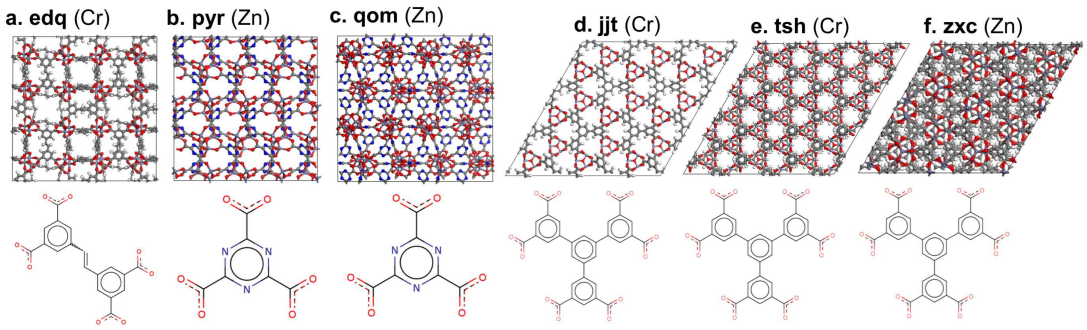

Figure 7.9 Six high-performing and thermodynamically feasible MOFs identified from our screening study. Topologies are shown in bold with the metal in parenthesis. The corresponding linker molecules (carboxylate form) are shown below each MOF.

7.5 Conclusions

We have demonstrated the combining of Ideal Adsorbed Solution Theory (IAST) with multilayer perceptron (MLP) predictions of single component isotherms to predict binary adsorption data in MOFs. Given that our MLP model is transferable across multiple adsorbates, and IAST is inherently general, the MLP+IAST framework is transferable (without modification) to many binary gas mixtures and operating conditions. This allows for the rapid screening of MOFs for performance in a variety of fixed bed adsorption applications. We showed that our MLP+IAST framework can (*i*) effectively capture the majority of high performing MOFs (e.g., the 95th percentile) with calculations that are many orders of magnitude faster than GCMC simulations, (*ii*) faithfully reproduce the same structure-performance relationships that would result from GCMC simulations, and (*iii*) screen MOFs for many different chemical separations, enabling us to identify MOFs that are simultaneously high performing for multiple separations/conditions, i.e. privileged MOFs. We demonstrated that these privileged MOFs can be efficiently identified using clustering algorithms. Finally, by combining our MLP+IAST results with large-scale free energy calculations we identify structures that are both high-performing and likely to be synthesizable from a thermodynamic standpoint. Overall, the theoretical framework presented in this work illustrates a computational discovery "pipeline" where MOFs are built in-silico, their separations properties are rapidly predicted for conditions of interest (screening a diversity of conditions in this step would allow one to identify privileged MOFs), and promising synthetic targets are proposed based on their expected performance and on their synthetic likelihood.

CHAPTER 8
GENERAL CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

8.1 General Conclusions

In this thesis, we developed methods for the discovery and design of metal-organic frameworks for chemical separations. We have developed an algorithm for the computational construction of MOFs according to arbitrary topological templates, general MOF design rules for CO_2 capture using machine learning algorithms, increasingly universal models for the prediction of adsorption loading in MOFs, and have investigated the thermodynamic stability of MOFs on a large scale. Throughout, we have used machine learning not only to increase the diversity of applications available for high-throughput screening but also to gain fundamental insight into MOF adsorption mechanisms on multiple scales. The insights on CO_2 adsorption mechanisms described in Chapter 4 were a direct result of our ability to exhaustively enumerate MOF structures with a variety of topologies and linker chemistries using our ToBaCCo program. These mechanistic insights, in turn, aided in the development of the larger-scale models described in Chapters 5 and 6 (that are more directly relevant to broadening the applications available for large-scale screening). These models can be used to rapidly screen MOFs for a wide variety of separations, but can also be used to derive generalized design principles from a "big-data" perspective. For example, the results of the clustering analysis performed in Chapter 7 suggest that many separation applications are likely to have similar high-performing MOFs, allowing researchers to target MOF platforms for use as separation agents in multiple separations. This result was made possible only by our ability to rapidly screen over 50,000 MOFs for 40 different separations. Additionally, we outlined methods for the large-scale calculation of MOF free energy in Chapter 6, allowing us to identify criteria, from a thermodynamic perspective, for MOF synthetic likelihood. By combining large-scale free energy computations with predictions of MOF separation performance, we also gain insight into what high-performing (hypothetical) MOFs are feasible synthetic targets. This result is a first step towards determining what hypothetical MOFs can be synthesized, and, while the specific criteria are likely to be updated, our initial criteria are logical starting points. Besides, the Frenkel-Ladd path method is promising for any future, large-scale thermodynamic stability calculations.

Moreover, our results and algorithms are not only applicable to MOFs. In many cases, they are directly transferable to other classes of porous crystals. ToBaCCo, for example, is capable of building covalent organic frameworks (COFs), porous organic cages (POCs), zeolites, various 2-dimensional materials, and polymers according to continuous random networks. Our adsorption-predicting models are directly

transferable to crystalline organic materials without metals (such as COFs and POCs), and the methodology used to train these models could be extended to arbitrary porous crystals. Finally, while the free energy criteria derived in Chapter 6 are directly applicable only to certain classes of MOFs, the free energy calculation method (our implementation of the Frenkel-Ladd path) can be used to calculate the free energy of essentially any crystalline material.

8.2 Recommendations for Future Work

8.2.1 Further Development of Machine Learning Models for Predicting Adsorption

One can imagine extending the models described in Chapters 5 and 7 to more complex adsorbates using an analogous training methodology. Large, flexible alchemical adsorbates (with or without large partial charges) representative of hydrocarbons, alcohols, and other $C_xH_yS_uO_vN_w$, can be used to generate training data for single component isotherm predictions. These single component isotherms could then be used as input for ideal adsorbed solution theory (IAST). IAST is likely to be less accurate for these more complex adsorbates, so a multicomponent adsorption prediction algorithm could also be trained to replace IAST. The variables used to describe adsorbates will need to be updated to include representations of adsorbate flexibility, shape, and the distribution of strong binding sites (i.e. portions of the adsorbate that are expected to interact strongly with framework atoms). Updated MOF descriptors may also be necessary, not only to increase the model accuracy but to also make it easier to use. For example, our current models require the computation of MOF textural properties. While these are rapid calculations, the data manipulation required to compile the descriptors of each MOF in a large database is non-trivial. An algorithm that takes a crystallographic information file (CIF) as input, a common format for storing MOF structures as element-labeled (x, y, z) coordinates, is more likely to see widespread use. This would require the algorithm to internally convert the MOF labeled (x, y, z) coordinates to a format compatible with a machine learning model. This suggests the use of convolutional neural networks (CNNs) that can take the MOF atom coordinates directly as input and automatically learn representations of MOF pore geometry and chemistry.

We generated our isotherm-predicting model using an exhaustive grid of adsorbate properties and pressures. The total number of training data points when using this approach is $n_a \times n_p \times n_M$, where n_a is the number of adsorbates, n_p is the number of pressures, and n_M is the number of MOFs. We trained the final model described in Chapter 7 on over 5 million data points (over 5 million individual GCMC simulations). This was only possible due to our access to a relatively large number of processors and the simplicity of the adsorbates we considered. For small, nonpolar adsorbates, relatively short GCMC simulations yield accurate results. For larger, polar adsorbates, adsorbate-framework electrostatic

interactions, and adsorbate flexibility must be considered (nonetheless, generating the training data took months of continuous computing on hundreds of processors). This results in GCMC simulations where each cycle is slower (since energy calculations are more computationally intensive) and more cycles must be run to achieve equilibration and convergence (since the adsorbates can adopt more configurations). Finally, exhaustive enumeration of adsorbate properties for larger, polar adsorbates would result in significantly more alchemical adsorbates than we used when training our models. Therefore, more efficient training data generation schemes must be used. One can imagine an iterative approach in which data is generated for a set of alchemical adsorbates drawn from an exhaustively enumerated "pool". An algorithm is trained using this data and the prediction performance is checked. Another set of training data can then be generated (making sure the overlap with the previous set is minimal) and used to continue the model training. This process can be repeated until a target prediction error is achieved or until the model predictions no longer improve significantly with more training data.

8.2.2 Towards Universal MOF Synthesizability Criteria

The thermodynamic accessibility criteria derived in Chapter 6 are not transferable to all MOF structures. Rather, they only apply to MOFs with carboxylate linkers and the same five metal clusters (Cu, Cr, Zn, and two Zr nodes) present in our MOF database. Furthermore, the transformation applied to the MOF per atom free energies to derive the global stability criteria relied on the presence of previously synthesized MOFs in the database. Therefore, future research should focus, in part, on deriving broadly applicable stability criteria. One approach is to create a forcefield that does not produce artificially high strain energies in MOF nodes, making direct comparisons between absolute MOF free energies possible. Given the structural and chemical diversity of MOFs, deriving a "universal" flexible forcefield for MOFs would be a daunting task. However, deriving a forcefield that applies to several commonly encountered metal nodes and is compatible with existing transferable potentials for MOF organic linkers would be useful for future MOF energy calculations (and for simulations of flexible MOFs in general). Another promising approach is to calculate the free energies of MOFs relative to their building blocks (i.e. calculate the effective free energy of formation). Specifically, a MOF could be decomposed into a set of computationally convenient building blocks (not necessarily the experimental constituents, see Figure 1.9), the free energies (or perhaps just the potential energies) for these individual building blocks can then be calculated and subtracted from the MOF absolute free energy as calculated via the Frenkel-Ladd path. This removes any contribution to the MOF free energy that is due to inherent strain in the building blocks. To be applied on a large scale, this method requires algorithms for accurately decomposing MOFs into useful building blocks without user intervention. Another possible approach for calculating an effective

MOF free energy of formation is to simulate the formation of a MOF from its building blocks in a single simulation, and then calculate the free energy difference between the formed and unformed MOF. This does not require the formation mechanism to be physical. Rather, bonds between alchemical MOF building blocks could be switched on during a nonequilibrium molecular dynamics (MD) simulation as a function of a switching parameter λ (see Section 2.1.5), allowing the free energy difference between the formed and unformed MOF to be calculated with thermodynamic integration.

Given that MOFs are solvothermally synthesized, the free energy of solvation is likely an important quantity in predicting MOF stability. The energies of empty frameworks have been used successfully in this thesis and in other studies to predict the relative stabilities of MOF isomorphs, however, there are situations in which the solvent used during synthesis changes the observed MOF phase. Free energies of solvation can be obtained using thermodynamic integration simulations, in which the interactions (Lennard-Jones and Coulomb) between the solvent molecules and the MOF are switched on as a function of a switching parameter λ. The free energies of solvation obtained in this manner are expected to heavily depend on the loading of solvent molecules in the MOF, this loading can be accurately calculated at experimentally relevant conditions using GCMC combined with NPT MD simulations (allowing the MOF to change conformation, shape, and size to accommodate solvent molecules). Additionally, the methods described above for computing the effective MOF free energies of formation could be done in the presence of a solvent.

A final key consideration is that thermodynamically stable MOFs may be kinetically inaccessible due to large energy barriers in the MOF formation mechanism. And, while the formation mechanisms of several MOF structures have been analyzed using various biased molecular dynamics approaches (see Section 1.3.5), the general time-dependent formation mechanisms of MOFs are still unknown [107–111]. A promising way forward is to investigate MOF formation mechanisms using idealized model systems in which MOF metal clusters and linkers are represented as coarse-grained models. Mechanistic insights (i.e. the locations of free energy barriers in the MOF formation process) garnered from these idealized model systems could then be applied to atomistic MOFs to obtain more quantitatively accurate free energy profiles of MOF formation. This method, as opposed to studying the formation mechanisms of individual MOFs, has the potential to identify general mechanistic "rules" for MOF formation, since a large variety of idealized MOF analogs could be considered. Regardless of the method used, kinetics must be taken into account to develop a truly universal set of MOF synthesizability criteria.

REFERENCES

[1] Key World Energy Statistics 2020 – Analysis - IEA. URL https://www.iea.org/reports/key-world-energy-statistics-2020.

[2] U.S. energy facts explained - consumption and production - U.S. Energy Information Administration (EIA). URL https://www.eia.gov/energyexplained/us-energy-facts/.

[3] David S Sholl and Ryan P Lively. Seven Chemical Separations to Change the World. *Nature*, 532: 435–437, 2016.

[4] Oak Ridge National Laboratory. Materials for separation technologies: energy and emission reduction opportunities. Technical report, 2005.

[5] Frank G. Kerry. *Industrial gas handbook: Gas separation and purification*. CRC Press, 1 2007. ISBN 9781420008265. doi: 10.1201/9781420008265. URL https://www.taylorfrancis.com/books/industrial-gas-handbook-frank-kerry/10.1201/9781420008265.

[6] Jennifer Wilcox. Cryogenic Distillation and Air Separation BT - Carbon Capture. pages 219–229. Springer New York, New York, NY, 2012. ISBN 978-1-4614-2215-0. doi: 10.1007/978-1-4614-2215-0{_}6. URL https://doi.org/10.1007/978-1-4614-2215-0_6.

[7] Tao Ren, Martin Patel, and Kornelis Blok. Olefins from conventional and heavy feedstocks: Energy use in steam cracking and alternative processes. *Energy*, 31(4):425–451, 2006. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2005.04.001. URL https://www.sciencedirect.com/science/article/pii/S0360544205000745.

[8] Jennifer Wilcox. Adsorption BT - Carbon Capture. pages 115–175. Springer New York, New York, NY, 2012. ISBN 978-1-4614-2215-0. doi: 10.1007/978-1-4614-2215-0{_}4. URL https://doi.org/10.1007/978-1-4614-2215-0_4.

[9] Ralph T. Yang. *Adsorbents: Fundamentals and Applications*. John Wiley & Sons, Ltd, Hoboken, 2003.

[10] Soumya Mukherjee, Amrit Kumar, and Michael J Zaworotko. 2 - Metal-organic framework based carbon capture and purification technologies for clean environment. pages 5–61. Elsevier, 2019. ISBN 978-0-12-814633-0. doi: https://doi.org/10.1016/B978-0-12-814633-0.00003-X. URL https://www.sciencedirect.com/science/article/pii/B978012814633000003X.

[11] Rajamani Krishna. Metrics for Evaluation and Screening of Metal-Organic Frameworks for Applications in Mixture Separations. *ACS Omega*, 5(28):16987–17004, 2020. ISSN 24701343. doi: 10.1021/acsomega.0c02218.

[12] Nobuyuki Shigaki, Yasuhiro Mogi, Takashi Haraoka, and Eiji Furuya. Measurements and calculations of the equilibrium adsorption amounts of CO2–N2, CO–N2, and CO2–CO mixed gases on 13X zeolite. *SN Applied Sciences*, 2(3):488, 2020. ISSN 2523-3971. doi: 10.1007/s42452-020-2298-y. URL https://doi.org/10.1007/s42452-020-2298-y.

[13] A L Myers and J M Prausnitz. Thermodynamics of mixed-gas adsorption. *AIChE Journal*, 11(1): 121–127, 1 1965. ISSN 0001-1541. doi: https://doi.org/10.1002/aic.690110125. URL https://doi.org/10.1002/aic.690110125.

[14] Krista S Walton and David S Sholl. Predicting multicomponent adsorption: 50 years of the ideal adsorbed solution theory. *AIChE Journal*, 61(9):2757–2762, 9 2015. ISSN 0001-1541. doi: https://doi.org/10.1002/aic.14878. URL https://doi.org/10.1002/aic.14878.

[15] Cory M. Simon, Berend Smit, and Maciej Haranczyk. PyIAST: Ideal adsorbed solution theory (IAST) Python package. *Computer Physics Communications*, 200(January 2020):364–380, 2016. ISSN 00104655. doi: 10.1016/j.cpc.2015.11.016.

[16] Anne Goj, David S Sholl, E Demet Akten, and Daniela Kohen. Atomistic Simulations of CO2 and N2 Adsorption in Silica Zeolites: The Impact of Pore Size and Shape. *The Journal of Physical Chemistry B*, 106(33):8367–8375, 8 2002. ISSN 1520-6106. doi: 10.1021/jp025895b. URL https://doi.org/10.1021/jp025895b.

[17] François Xavier Coudert and Alain H. Fuchs. Computational characterization and prediction of metal-organic framework properties. *Coordination Chemistry Reviews*, 307:211–236, 2016. ISSN 00108545. doi: 10.1016/j.ccr.2015.08.001.

[18] Ryther Anderson, Benjamin Schweitzer, Ting Wu, Moises A Carreon, and Diego A Gómez-Gualdrón. Molecular Simulation Insights on Xe/Kr Separation in a Set of Nanoporous Crystalline Membranes. *ACS Applied Materials & Interfaces*, 10(1):582–592, 1 2018. ISSN 1944-8244. doi: 10.1021/acsami.7b14791. URL https://doi.org/10.1021/acsami.7b14791.

[19] Benjamin J Sikora, Christopher E Wilmer, Michael L Greenfield, and Randall Q Snurr. Thermodynamic analysis of Xe/Kr selectivity in over 137 000 hypothetical metal–organic frameworks www.rsc.org/BiomaterialsScience Thermodynamic analysis of Xe/Kr selectivity in over 137 000 hypothetical metal–organic frameworks. 3(1):2177–2396, 2012. ISSN 2047-4830. doi: 10.1039/c2sc01097f.

[20] Linjiang Chen, Paul S Reiss, Samantha Y Chong, Daniel Holden, Kim E Jelfs, Tom Hasell, Marc A Little, Adam Kewley, Michael E Briggs, Andrew Stephenson, K Mark Thomas, Jayne A Armstrong, Jon Bell, Jose Busto, Raymond Noel, Jian Liu, Denis M Strachan, Praveen K Thallapally, and Andrew I Cooper. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nature Materials*, 13(10):954–960, 2014. ISSN 1476-4660. doi: 10.1038/nmat4035. URL https://doi.org/10.1038/nmat4035.

[21] Seda Keskin. Adsorption, Diffusion, and Separation of CH4/H2 Mixtures in Covalent Organic Frameworks: Molecular Simulations and Theoretical Predictions. *The Journal of Physical Chemistry C*, 116(2):1772–1779, 1 2012. ISSN 1932-7447. doi: 10.1021/jp209804x. URL https://doi.org/10.1021/jp209804x.

[22] Cory M Simon, Rocio Mercado, Sondre K Schnell, Berend Smit, and Maciej Haranczyk. What Are the Best Materials To Separate a Xenon/Krypton Mixture? *Chemistry of Materials*, 27(12): 4459–4475, 6 2015. ISSN 0897-4756. doi: 10.1021/acs.chemmater.5b01475. URL https://doi.org/10.1021/acs.chemmater.5b01475.

[23] Rajamani Krishna. Screening metal–organic frameworks for mixture separations in fixed-bed adsorbers using a combined selectivity/capacity metric. *RSC Advances*, 7(57):35724–35737, 2017. doi: 10.1039/C7RA07363A. URL http://dx.doi.org/10.1039/C7RA07363A.

[24] Karson T. Leperi, Yongchul G. Chung, Fengqi You, and Randall Q. Snurr. Development of a General Evaluation Metric for Rapid Screening of Adsorbent Materials for Postcombustion CO2 Capture. *ACS Sustainable Chemistry and Engineering*, 7(13):11529–11539, 7 2019. ISSN 21680485. doi: 10.1021/acssuschemeng.9b01418. URL https://doi.org/10.1021/acssuschemeng.9b01418.

[25] Frank Notaro, Joseph Timothy Mullhaupt, Frederick Wells Wells, and Mark William Ackley. Adsorption process and system using multilayer adsorbent beds. 3(19), 1997. URL https://patents.google.com/patent/US5810909.

[26] Salil U Rege and Ralph T Yang. A simple parameter for selecting an adsorbent for gas separation by pressure swing adsorption. *Separation Science and Technology*, 36(15):3355–3365, 11 2001. ISSN 0149-6395. doi: 10.1081/SS-100107907. URL https://www.tandfonline.com/doi/abs/10.1081/SS-100107907.

[27] Gerhard D Pirngruber, Lomig Hamon, Sandrine Bourrelly, Philip L Llewellyn, Estelle Lenoir, Vincent Guillerm, Christian Serre, and Thomas Devic. A Method for Screening the Potential of MOFs as CO2 Adsorbents in Pressure Swing Adsorption Processes. *ChemSusChem*, 5(4):762–776, 4 2012. ISSN 1864-5631. doi: https://doi.org/10.1002/cssc.201100716. URL https://doi.org/10.1002/cssc.201100716.

[28] Ch. Baerlocher and L.B. McCusker. Database of Zeolite Structures. URL http://www.iza-structure.org/databases/.

[29] Vladislav A Blatov, Olga A Blatova, Frits Daeyaert, and Michael W Deem. Nanoporous materials with predicted zeolite topologies. *RSC Advances*, 10(30):17760–17767, 2020. doi: 10.1039/D0RA01888K. URL http://dx.doi.org/10.1039/D0RA01888K.

[30] Hiroyasu Furukawa, Nakeun Ko, Yong Bok Go, Naoki Aratani, Sang Beom Choi, Eunwoo Choi, A Özgür Yazaydin, Randall Q Snurr, Michael O'Keeffe, Jaheon Kim, and Omar M Yaghi. Ultrahigh Porosity in Metal-Organic Frameworks. *Science*, 329(5990):424 LP – 428, 7 2010. doi: 10.1126/science.1192160. URL http://science.sciencemag.org/content/329/5990/424.abstract.

[31] Xiao Feng, Xuesong Ding, and Donglin Jiang. Covalent organic frameworks. *Chemical Society Reviews*, 41(18):6010–6022, 2012. ISSN 0306-0012. doi: 10.1039/C2CS35157A. URL http://dx.doi.org/10.1039/C2CS35157A.

[32] Tomokazu Tozawa, James T A Jones, Shashikala I Swamy, Shan Jiang, Dave J Adams, Stephen Shakespeare, Rob Clowes, Darren Bradshaw, Tom Hasell, Samantha Y Chong, Chiu Tang, Stephen Thompson, Julia Parker, Abbie Trewin, John Bacsa, Alexandra M Z Slawin, Alexander Steiner, and Andrew I Cooper. Porous organic cages. *Nature Materials*, 8:973, 10 2009. URL https://doi.org/10.1038/nmat2545http://10.0.4.14/nmat2545https://www.nature.com/articles/nmat2545#supplementary-information.

[33] Richard L Martin, Cory M Simon, Berend Smit, and Maciej Haranczyk. In silico Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *Journal of the American Chemical Society*, 136(13):5006–5022, 4 2014. ISSN 0002-7863. doi: 10.1021/ja4123939. URL https://doi.org/10.1021/ja4123939.

[34] Michael O'Keeffe and Omar M. Yaghi. Deconstructing the Crystal Structures of Metal–Organic Frameworks and Related Materials into Their Underlying Nets. *Chemical Reviews*, 112(2):675–702, 2 2012. ISSN 0009-2665. doi: 10.1021/cr200205j. URL https://doi.org/10.1021/cr200205j.

[35] Mian Li, Dan Li, Michael O'Keeffe, and Omar M Yaghi. Topological Analysis of Metal–Organic Frameworks with Polytopic Linkers and/or Multiple Building Units and the Minimal Transitivity Principle. *Chemical Reviews*, 114(2):1343–1370, 1 2014. ISSN 0009-2665. doi: 10.1021/cr400392k. URL https://doi.org/10.1021/cr400392k.

[36] Nathan W Ockwig, Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M Yaghi. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Accounts of Chemical Research*, 38(3):176–182, 3 2005. ISSN 0001-4842. doi: 10.1021/ar020022l. URL https://doi.org/10.1021/ar020022l.

[37] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chemistry*, 4:83, 11 2011. URL https://doi.org/10.1038/nchem.1192http://10.0.4.14/nchem.1192https://www.nature.com/articles/nchem.1192#supplementary-information.

[38] Diego A Gómez-Gualdrón, Yamil J Colón, Xu Zhang, Timothy C Wang, Yu-Sheng Chen, Joseph T Hupp, Taner Yildirim, Omar K Farha, Jian Zhang, and Randall Q Snurr. Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy & Environmental Science*, 9(10):3279–3289, 2016. ISSN 1754-5692. doi: 10.1039/C6EE02104B. URL http://dx.doi.org/10.1039/C6EE02104B.

[39] Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gladysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, Jeffrey A Reimer, Jorge A R Navarro, Tom K Woo, Susana Garcia, Kyriakos C Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature*, 576(7786):253–256, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1798-7. URL https://doi.org/10.1038/s41586-019-1798-7.

[40] Diego A Gomez-Gualdron, Oleksii V Gutov, Vaiva Krungleviciute, Bhaskarjyoti Borah, Joseph E Mondloch, Joseph T Hupp, Taner Yildirim, Omar K Farha, and Randall Q Snurr. Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chemistry of Materials*, 26(19):5632–5639, 10 2014. ISSN 0897-4756. doi: 10.1021/cm502304e. URL https://doi.org/10.1021/cm502304e.

[41] Alauddin Ahmed, Saona Seth, Justin Purewal, Antek G Wong-Foy, Mike Veenstra, Adam J Matzger, and Donald J Siegel. Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nature Communications*, 10(1):1568, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09365-w. URL https://doi.org/10.1038/s41467-019-09365-w.

[42] Zhijie Chen, Penghao Li, Ryther Anderson, Xingjie Wang, Xuan Zhang, Lee Robison, Louis R Redfern, Shinya Moribe, Timur Islamoglu, Diego A Gómez-Gualdrón, Taner Yildirim, J Fraser Stoddart, and Omar K Farha. Balancing volumetric and gravimetric uptake in highly porous materials for clean energy. *Science*, 368(6488):297 LP – 303, 4 2020. doi: 10.1126/science.aaz8881. URL http://science.sciencemag.org/content/368/6488/297.abstract.

[43] Debasis Banerjee, Cory M Simon, Anna M Plonka, Radha K Motkuri, Jian Liu, Xianyin Chen, Berend Smit, John B Parise, Maciej Haranczyk, and Praveen K Thallapally. Metal–organic framework with optimally selective xenon adsorption and separation. *Nature Communications*, 7: ncomms11831, 6 2016. URL https://doi.org/10.1038/ncomms11831http://10.0.4.14/ncomms11831https://www.nature.com/articles/ncomms11831#supplementary-information.

[44] Celia Castillo-Blas and Felipe Gándara. Metal-organic Frameworks Incorporating Multiple Metal Elements. *Israel Journal of Chemistry*, 58(9-10):1036–1043, 10 2018. ISSN 0021-2148. doi: https://doi.org/10.1002/ijch.201800085. URL https://doi.org/10.1002/ijch.201800085.

[45] Richard F D'Vries, Susana Álvarez-García, Natalia Snejko, Luisa E Bausá, Enrique Gutiérrez-Puebla, Alicia de Andrés, and M Ángeles Monge. Multimetal rare earth MOFs for lighting and thermometry: tailoring color and optimal temperature range through enhanced disulfobenzoic triplet phosphorescence. *Journal of Materials Chemistry C*, 1(39):6316–6324, 2013. ISSN 2050-7526. doi: 10.1039/C3TC30858H. URL http://dx.doi.org/10.1039/C3TC30858H.

[46] Jilan Long, Kui Shen, Liang Chen, and Yingwei Li. Multimetal-MOF-derived transition metal alloy NPs embedded in an N-doped carbon matrix: highly active catalysts for hydrogenation reactions. *Journal of Materials Chemistry A*, 4(26):10254–10262, 2016. ISSN 2050-7488. doi: 10.1039/C6TA00157B. URL http://dx.doi.org/10.1039/C6TA00157B.

[47] Jacek Klinowski, Filipe A Almeida Paz, Patrícia Silva, and João Rocha. Microwave-Assisted Synthesis of Metal–Organic Frameworks. *Dalton Transactions*, 40(2):321–330, 2011. ISSN 1477-9226. doi: 10.1039/C0DT00708K. URL http://dx.doi.org/10.1039/C0DT00708K.

[48] Nazmul Abedin Khan and Sung Hwa Jhung. Synthesis of metal-organic frameworks (MOFs) with microwave or ultrasound: Rapid reaction, phase-selectivity, and size reduction. *Coordination Chemistry Reviews*, 285:11–23, 2015. ISSN 0010-8545. doi: https://doi.org/10.1016/j.ccr.2014.10.008. URL https://www.sciencedirect.com/science/article/pii/S0010854514002811.

[49] David J Tranchemontagne, Joseph R Hunt, and Omar M Yaghi. Room temperature synthesis of metal-organic frameworks: MOF-5, MOF-74, MOF-177, MOF-199, and IRMOF-0. *Tetrahedron*, 64 (36):8553–8557, 2008. ISSN 0040-4020. doi: https://doi.org/10.1016/j.tet.2008.06.036. URL https://www.sciencedirect.com/science/article/pii/S0040402008011344.

[50] Manuel Sánchez-Sánchez, Negash Getachew, Kenya Díaz, Manuel Díaz-García, Yonas Chebude, and Isabel Díaz. Synthesis of metal–organic frameworks in water at room temperature: salts as linker sources. *Green Chemistry*, 17(3):1500–1509, 2015. ISSN 1463-9262. doi: 10.1039/C4GC01861C. URL http://dx.doi.org/10.1039/C4GC01861C.

[51] Matthew R DeStefano, Timur Islamoglu, Sergio J Garibay, Joseph T Hupp, and Omar K Farha. Room-Temperature Synthesis of UiO-66 and Thermal Modulation of Densities of Defect Sites. *Chemistry of Materials*, 29(3):1357–1361, 2 2017. ISSN 0897-4756. doi: 10.1021/acs.chemmater.6b05115. URL https://doi.org/10.1021/acs.chemmater.6b05115.

[52] Omar K Farha, Ibrahim Eryazici, Nak Cheon Jeong, Brad G Hauser, Christopher E Wilmer, Amy A Sarjeant, Randall Q Snurr, SonBinh T Nguyen, A Özgür Yazaydın, and Joseph T Hupp. Metal–Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit? *Journal of the American Chemical Society*, 134(36):15016–15021, 9 2012. ISSN 0002-7863. doi: 10.1021/ja3055639. URL https://doi.org/10.1021/ja3055639.

[53] Aaron W Thornton, Cory M Simon, Jihan Kim, Ohmin Kwon, Kathryn S Deeg, Kristina Konstas, Steven J Pas, Matthew R Hill, David A Winkler, Maciej Haranczyk, and Berend Smit. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chemistry of Materials*, 29(7):2844–2854, 4 2017. ISSN 0897-4756. doi: 10.1021/acs.chemmater.6b04933. URL https://doi.org/10.1021/acs.chemmater.6b04933.

[54] Diego A Gómez-Gualdrón, Christopher E Wilmer, Omar K Farha, Joseph T Hupp, and Randall Q Snurr. Exploring the Limits of Methane Storage and Delivery in Nanoporous Materials. *The Journal of Physical Chemistry C*, 118(13):6941–6951, 4 2014. ISSN 1932-7447. doi: 10.1021/jp502359q. URL https://doi.org/10.1021/jp502359q.

[55] Jordan K. Pommerenck, Cory M. Simon, and David Roundy. An upper bound to gas storage and delivery via pressure-swing adsorption in porous materials. 6 2020. URL https://arxiv.org/abs/2006.01886.

[56] Xinfa Wei, Shun Wang, Zile Hua, Lisong Chen, and Jianlin Shi. Metal–Organic Framework Nanosheet Electrocatalysts for Efficient H2 Production from Methanol Solution: Methanol-Assisted Water Splitting or Methanol Reforming? *ACS Applied Materials & Interfaces*, 10(30):25422–25428, 8 2018. ISSN 1944-8244. doi: 10.1021/acsami.8b06948. URL https://doi.org/10.1021/acsami.8b06948.

[57] Benjamin Schweitzer, Chloe Archuleta, Bomsaerah Seong, Ryther Anderson, and Diego A Gómez-Gualdrón. Electronic effects due to organic linker-metal surface interactions: implications on screening of MOF-encapsulated catalysts. *Physical Chemistry Chemical Physics*, 22(4):2475–2487, 2020. ISSN 1463-9076. doi: 10.1039/C9CP05380H. URL http://dx.doi.org/10.1039/C9CP05380H.

[58] Jingjing Duan, Sheng Chen, and Chuan Zhao. Ultrathin metal-organic framework array for efficient electrocatalytic water splitting. *Nature Communications*, 8(1):15341, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15341. URL https://doi.org/10.1038/ncomms15341.

[59] Dong Yang, Manuel A Ortuño, Varinia Bernales, Christopher J Cramer, Laura Gagliardi, and Bruce C Gates. Structure and Dynamics of Zr6O8 Metal–Organic Framework Node Surfaces Probed with Ethanol Dehydration as a Catalytic Test Reaction. *Journal of the American Chemical Society*, 140(10):3751–3759, 3 2018. ISSN 0002-7863. doi: 10.1021/jacs.7b13330. URL https://doi.org/10.1021/jacs.7b13330.

[60] Zhigang Xie, Liqing Ma, Kathryn E deKrafft, Athena Jin, and Wenbin Lin. Porous Phosphorescent Coordination Polymers for Oxygen Sensing. *Journal of the American Chemical Society*, 132(3): 922–923, 1 2010. ISSN 0002-7863. doi: 10.1021/ja909629f. URL https://doi.org/10.1021/ja909629f.

[61] Anjian Lan, Kunhao Li, Haohan Wu, David H. Olson, Thomas J. Emge, Woosoek Ki, Maochun Hong, and Jing Li. A Luminescent Microporous Metal–Organic Framework for the Fast and Reversible Detection of High Explosives. *Angewandte Chemie International Edition*, 48(13):2334–2338, 3 2009. ISSN 1433-7851. doi: 10.1002/anie.200804853. URL https://doi.org/10.1002/anie.200804853.

[62] Jenna A Gustafson and Christopher E Wilmer. Computational Design of Metal–Organic Framework Arrays for Gas Sensing: Influence of Array Size and Composition on Sensor Performance. *The Journal of Physical Chemistry C*, 121(11):6033–6038, 3 2017. ISSN 1932-7447. doi: 10.1021/acs.jpcc.6b09740. URL https://doi.org/10.1021/acs.jpcc.6b09740.

[63] Qing-Lin Li, Jia-Ping Wang, Wei-Cong Liu, Xiao-Yi Zhuang, Jian-Qiang Liu, Gui-Ling Fan, Bao-Hong Li, Wei-Na Lin, and Jian-Hui Man. A new (4,8)-connected topological MOF as potential drug delivery. *Inorganic Chemistry Communications*, 55:8–10, 2015. ISSN 1387-7003. doi: https://doi.org/10.1016/j.inoche.2015.02.023. URL http://www.sciencedirect.com/science/article/pii/S1387700315000751.

[64] Kristina Roth Stefaniak, Charity C Epley, Joshua J Novak, Margaret L McAndrew, Hannah D Cornell, Jie Zhu, Dylan K McDaniel, Jennifer L Davis, Irving C Allen, Amanda J Morris, and Tijana Z Grove. Photo-triggered release of 5-fluorouracil from a MOF drug delivery vehicle. *Chemical Communications*, 54(55):7617–7620, 2018. ISSN 1359-7345. doi: 10.1039/C8CC01601A. URL http://dx.doi.org/10.1039/C8CC01601A.

[65] Karel J Hartlieb, Daniel P Ferris, James M Holcroft, Irawati Kandela, Charlotte L Stern, Majed S Nassar, Youssry Y Botros, and J Fraser Stoddart. Encapsulation of Ibuprofen in CD-MOF and Related Bioavailability Studies. *Molecular Pharmaceutics*, 14(5):1831–1839, 5 2017. ISSN 1543-8384. doi: 10.1021/acs.molpharmaceut.7b00168. URL https://doi.org/10.1021/acs.molpharmaceut.7b00168.

[66] Xuerui Chen, Rongliang Tong, Zheqi Shi, Beng Yang, Hua Liu, Shiping Ding, Xu Wang, Qunfang Lei, Jian Wu, and Wenjun Fang. MOF Nanoparticles with Encapsulated Autophagy Inhibitor in Controlled Drug Delivery System for Antitumor. *ACS Applied Materials & Interfaces*, 10(3): 2328–2337, 1 2018. ISSN 1944-8244. doi: 10.1021/acsami.7b16522. URL https://doi.org/10.1021/acsami.7b16522.

[67] Zoey R Herm, Eric D Bloch, and Jeffrey R Long. Hydrocarbon Separations in Metal–Organic Frameworks. *Chemistry of Materials*, 26(1):323–338, 1 2014. ISSN 0897-4756. doi: 10.1021/cm402897c. URL https://doi.org/10.1021/cm402897c.

[68] Zoey R Herm, Brian M Wiers, Jarad A Mason, Jasper M van Baten, Matthew R Hudson, Pawel Zajdel, Craig M Brown, Norberto Masciocchi, Rajamani Krishna, and Jeffrey R Long. Separation of Hexane Isomers in a Metal-Organic Framework with Triangular Channels. *Science*, 340(6135):960 LP – 964, 5 2013. doi: 10.1126/science.1234071. URL http://science.sciencemag.org/content/340/6135/960.abstract.

[69] Surendar R Venna and Moises A Carreon. Highly Permeable Zeolite Imidazolate Framework-8 Membranes for CO2/CH4 Separation. *Journal of the American Chemical Society*, 132(1):76–78, 1 2010. ISSN 0002-7863. doi: 10.1021/ja909263x. URL https://doi.org/10.1021/ja909263x.

[70] Yongchul G. Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J. Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K. Farha, David S. Sholl, and Randall Q. Snurr. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chemistry of Materials*, 26(21):6185–6192, 2014. ISSN 15205002. doi: 10.1021/cm502594j.

[71] Yongchul G Chung, Emmanuel Haldoupis, Benjamin J Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S Camp, Ben Slater, J Ilja Siepmann, David S Sholl, and Randall Q Snurr. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*, 11 2019. ISSN 0021-9568. doi: 10.1021/acs.jced.9b00835. URL https://doi.org/10.1021/acs.jced.9b00835.

[72] Yamil J Colón and Randall Q Snurr. High-throughput computational screening of metal–organic frameworks. *Chemical Society Reviews*, 43(16):5735–5749, 2014. ISSN 0306-0012. doi: 10.1039/C4CS00070F. URL http://dx.doi.org/10.1039/C4CS00070F.

[73] Yamil J Colón, Diego A Gómez-Gualdrón, and Randall Q Snurr. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Crystal Growth & Design*, 17(11):5801–5810, 11 2017. ISSN 1528-7483. doi: 10.1021/acs.cgd.7b00848. URL https://doi.org/10.1021/acs.cgd.7b00848.

[74] Matthew A Addicoat, Damien E Coupry, and Thomas Heine. AuToGraFS: Automatic Topological Generator for Framework Structures. *The Journal of Physical Chemistry A*, 118(40):9607–9614, 10 2014. ISSN 1089-5639. doi: 10.1021/jp507643v. URL https://doi.org/10.1021/jp507643v.

[75] Richard Luis Martin and Maciej Haranczyk. Construction and Characterization of Structure Models of Crystalline Porous Polymers. *Crystal Growth & Design*, 14(5):2431–2440, 5 2014. ISSN 1528-7483. doi: 10.1021/cg500158c. URL https://doi.org/10.1021/cg500158c.

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

[77] Ryther Anderson and Diego A Gómez-Gualdrón. Increasing topological diversity during computational "synthesis" of porous crystals: how and why. *CrystEngComm*, 21(10):1653–1665, 2019. doi: 10.1039/C8CE01637B. URL http://dx.doi.org/10.1039/C8CE01637B.

[78] Olaf Delgado-Friedrichs and Michael O'Keeffe. Crystal nets as graphs: Terminology and definitions. *Journal of Solid State Chemistry*, 178(8):2480–2485, 2005. ISSN 0022-4596. doi: https://doi.org/10.1016/j.jssc.2005.06.011. URL https://www.sciencedirect.com/science/article/pii/S002245960500263X.

[79] Michael O'Keeffe, Maxim A Peskov, Stuart J Ramsden, and Omar M Yaghi. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Accounts of Chemical Research*, 41(12):1782–1789, 12 2008. ISSN 0001-4842. doi: 10.1021/ar800124u. URL https://doi.org/10.1021/ar800124u.

[80] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, 1992. ISSN 15205126. doi: 10.1021/ja00051a040.

[81] Matthew A. Addicoat and Thomas Heine. Extension of the Universal Force Field for Metal–Organic Frameworks. *Journal of Chemical Theory and Computation*, 12(10):5215–5225, 10 2014. doi: 10.1021/acs.jctc.6b00664.

[82] Damien E Coupry, Matthew A Addicoat, and Thomas Heine. Extension of the Universal Force Field for Metal–Organic Frameworks. *Journal of Chemical Theory and Computation*, 12(10):5215–5225, 10 2016. ISSN 1549-9618. doi: 10.1021/acs.jctc.6b00664. URL https://doi.org/10.1021/acs.jctc.6b00664.

[83] Stephen L Mayo, Barry D Olafson, and William A Goddard. DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, 94(26):8897–8909, 12 1990. ISSN 0022-3654. doi: 10.1021/j100389a010. URL https://doi.org/10.1021/j100389a010.

[84] Ryther Anderson, Achay Biong, and Diego A Gómez-Gualdrón. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *Journal of Chemical Theory and Computation*, 1 2020. ISSN 1549-9618. doi: 10.1021/acs.jctc.9b00940. URL https://doi.org/10.1021/acs.jctc.9b00940.

[85] Thomas F Willems, Chris H Rycroft, Michaeel Kazi, Juan C Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012. ISSN 1387-1811. doi: https://doi.org/10.1016/j.micromeso.2011.08.020. URL http://www.sciencedirect.com/science/article/pii/S1387181111003738.

[86] Marielle Pinheiro, Richard L Martin, Chris H Rycroft, Andrew Jones, Enrique Iglesia, and Maciej Haranczyk. Characterization and comparison of pore landscapes in crystalline porous materials. *Journal of Molecular Graphics and Modelling*, 44:208–219, 2013. ISSN 1093-3263. doi: https://doi.org/10.1016/j.jmgm.2013.05.007. URL https://www.sciencedirect.com/science/article/pii/S109332631300096X.

[87] A L Myers and P A Monson. Adsorption in Porous Materials at High Pressure: Theory and Experiment. *Langmuir*, 18(26):10261–10273, 12 2002. ISSN 0743-7463. doi: 10.1021/la026399h. URL https://doi.org/10.1021/la026399h.

[88] Tina Düren, Franck Millange, Gérard Férey, Krista S Walton, and Randall Q Snurr. Calculating Geometric Surface Areas as a Characterization Tool for MetalOrganic Frameworks. *The Journal of Physical Chemistry C*, 111(42):15350–15356, 10 2007. ISSN 1932-7447. doi: 10.1021/jp074723h. URL https://doi.org/10.1021/jp074723h.

[89] Tina Düren, Youn-Sang Bae, and Randall Q Snurr. Using molecular simulation to characterise metal–organic frameworks for adsorption applications. *Chemical Society Reviews*, 38(5):1237–1247, 2009. ISSN 0306-0012. doi: 10.1039/B803498M. URL http://dx.doi.org/10.1039/B803498M.

[90] Lev D Gelb and K E Gubbins. Pore Size Distributions in Porous Glasses: A Computer Simulation Study. *Langmuir*, 15(2):305–308, 1 1999. ISSN 0743-7463. doi: 10.1021/la9808418. URL https://doi.org/10.1021/la9808418.

[91] David Dubbeldam, Sofía Calero, Donald E Ellis, and Randall Q Snurr. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*, 42(2): 81–101, 1 2016. ISSN 0892-7022. doi: 10.1080/08927022.2015.1010082. URL https://doi.org/10.1080/08927022.2015.1010082.

[92] Kyle A Baseden and Jesse W Tye. Introduction to Density Functional Theory: Calculations by Hand on the Helium Atom. *Journal of Chemical Education*, 91(12):2116–2123, 12 2014. ISSN 0021-9584. doi: 10.1021/ed5004788. URL https://doi.org/10.1021/ed5004788.

[93] Hanjun Fang, Hakan Demir, Preeti Kamakoti, and David S Sholl. Recent developments in first-principles force fields for molecules in nanoporous materials. *Journal of Materials Chemistry A*, 2(2):274–291, 2014. ISSN 2050-7488. doi: 10.1039/C3TA13073H. URL http://dx.doi.org/10.1039/C3TA13073H.

[94] Michael Fernandez, Tom K Woo, Christopher E Wilmer, and Randall Q Snurr. Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks. *The Journal of Physical Chemistry C*, 117(15):7681–7689, 4 2013. ISSN 1932-7447. doi: 10.1021/jp4006422. URL https://doi.org/10.1021/jp4006422.

[95] Michael Fernandez, Nicholas R Trefiak, and Tom K Woo. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity. *The Journal of Physical Chemistry C*, 117(27):14095–14105, 7 2013. ISSN 1932-7447. doi: 10.1021/jp404287t. URL https://doi.org/10.1021/jp404287t.

[96] Hana Dureckova, Mykhaylo Krykunov, Mohammad Zein Aghaji, and Tom K Woo. Robust Machine Learning Models for Predicting High CO2 Working Capacity and CO2/H2 Selectivity of Gas Adsorption in Metal Organic Frameworks for Precombustion Carbon Capture. *The Journal of Physical Chemistry C*, 123(7):4133–4139, 2 2019. ISSN 1932-7447. doi: 10.1021/acs.jpcc.8b10644. URL https://doi.org/10.1021/acs.jpcc.8b10644.

[97] Jake Burner, Ludwig Schwiedrzik, Mykhaylo Krykunov, Jun Luo, Peter G Boyd, and Tom K Woo. High-Performing Deep Learning Regression Models for Predicting Low-Pressure CO2 Adsorption Properties of Metal–Organic Frameworks. *The Journal of Physical Chemistry C*, 124(51): 27996–28005, 12 2020. ISSN 1932-7447. doi: 10.1021/acs.jpcc.0c06334. URL https://doi.org/10.1021/acs.jpcc.0c06334.

[98] Benjamin J Bucior, N Scott Bobbitt, Timur Islamoglu, Subhadip Goswami, Arun Gopalan, Taner Yildirim, Omar K Farha, Neda Bagheri, and Randall Q Snurr. Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Molecular Systems Design & Engineering*, 4 (1):162–174, 2019. doi: 10.1039/C8ME00050F. URL http://dx.doi.org/10.1039/C8ME00050F.

[99] Michael Fernandez, Peter G Boyd, Thomas D Daff, Mohammad Zein Aghaji, and Tom K Woo. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO2 Capture. *The Journal of Physical Chemistry Letters*, 5(17):3056–3060, 9 2014. ISSN 1948-7185. doi: 10.1021/jz501331m. URL https://doi.org/10.1021/jz501331m.

[100] Maryam Pardakhti, Ehsan Moharreri, David Wanik, Steven L Suib, and Ranjan Srivastava. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Combinatorial Science*, 19(10): 640–645, 10 2017. ISSN 2156-8952. doi: 10.1021/acscombsci.7b00056. URL https://doi.org/10.1021/acscombsci.7b00056.

[101] Grace Anderson, Benjamin Schweitzer, Ryther Anderson, and Diego A. Gómez-Gualdrón. Attainable Volumetric Targets for Adsorption-Based Hydrogen Storage in Porous Crystals: Molecular Simulation and Machine Learning. *The Journal of Physical Chemistry C*, 123(1):120–130, 1 2019. ISSN 1932-7447. doi: 10.1021/acs.jpcc.8b09420. URL https://doi.org/10.1021/acs.jpcc.8b09420.

[102] Farhad Gharagheizi, Dai Tang, and David S Sholl. Selecting Adsorbents to Separate Diverse Near-Azeotropic Chemicals. *The Journal of Physical Chemistry C*, 124(6):3664–3670, 2 2020. ISSN 1932-7447. doi: 10.1021/acs.jpcc.9b10955. URL https://doi.org/10.1021/acs.jpcc.9b10955.

[103] Yangzesheng Sun, Robert F DeJaco, and J Ilja Siepmann. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. *Chemical Science*, 10(16):4377–4388, 2019. ISSN 2041-6520. doi: 10.1039/C8SC05340E. URL http://dx.doi.org/10.1039/C8SC05340E.

[104] Mary J Van Vleet, Tingting Weng, Xinyi Li, and J R Schmidt. In Situ, Time-Resolved, and Mechanistic Studies of Metal–Organic Framework Nucleation and Growth. *Chemical Reviews*, 118(7): 3681–3721, 4 2018. ISSN 0009-2665. doi: 10.1021/acs.chemrev.7b00582. URL https://doi.org/10.1021/acs.chemrev.7b00582.

[105] Caroline Mellot Draznieks, John M Newsam, Alan M Gorman, Clive M Freeman, and Gérard Férey. De Novo Prediction of Inorganic Structures Developed through Automated Assembly of Secondary Building Units (AASBU Method). *Angewandte Chemie International Edition*, 39(13):2270–2275, 7 2000. ISSN 1433-7851. doi: 10.1002/1521-3773(20000703)39:13⟨2270::AID-ANIE2270⟩3.0.CO;2-A. URL https://doi.org/10.1002/1521-3773(20000703)39:13%3C2270::AID-ANIE2270%3E3.0.COhttp://2-a.

[106] Caroline Mellot-Draznieks, Julien Dutour, and Gérard Férey. Hybrid Organic–Inorganic Frameworks: Routes for Computational Design and Structure Prediction. *Angewandte Chemie International Edition*, 43(46):6290–6296, 11 2004. ISSN 1433-7851. doi: 10.1002/anie.200454251. URL https://doi.org/10.1002/anie.200454251.

[107] Makoto Yoneya, Seiji Tsuzuki, and Masaru Aoyagi. Simulation of metal–organic framework self-assembly. *Physical Chemistry Chemical Physics*, 17(14):8649–8652, 2015. ISSN 1463-9076. doi: 10.1039/C5CP00379B. URL http://dx.doi.org/10.1039/C5CP00379B.

[108] Debasmita Biswal and Peter G Kusalik. Molecular simulations of self-assembly processes in metal-organic frameworks: Model dependence. *The Journal of Chemical Physics*, 147(4):44702, 7 2017. ISSN 0021-9606. doi: 10.1063/1.4994700. URL https://aip.scitation.org/doi/abs/10.1063/1.4994700.

[109] Debasmita Biswal and Peter G Kusalik. Probing Molecular Mechanisms of Self-Assembly in Metal–Organic Frameworks. *ACS Nano*, 11(1):258–268, 1 2017. ISSN 1936-0851. doi: 10.1021/acsnano.6b05444. URL https://doi.org/10.1021/acsnano.6b05444.

[110] Loukas Kollias, David C Cantu, Marcus A Tubbs, Roger Rousseau, Vassiliki-Alexandra Glezakou, and Matteo Salvalaglio. Molecular Level Understanding of the Free Energy Landscape in Early Stages of Metal–Organic Framework Nucleation. *Journal of the American Chemical Society*, 141(14): 6073–6081, 4 2019. ISSN 0002-7863. doi: 10.1021/jacs.9b01829. URL https://doi.org/10.1021/jacs.9b01829.

[111] Yamil J Colón, Ashley Z Guo, Lucas W Antony, Kyle Q Hoffmann, and Juan J de Pablo. Free energy of metal-organic framework self-assembly. *The Journal of Chemical Physics*, 150(10):104502, 3 2019. ISSN 0021-9606. doi: 10.1063/1.5063588. URL https://doi.org/10.1063/1.5063588.

[112] Pravas Deria, Diego A Gómez-Gualdrón, Idan Hod, Randall Q Snurr, Joseph T Hupp, and Omar K Farha. Framework-Topology-Dependent Catalytic Activity of Zirconium-Based (Porphinato)zinc(II) MOFs. *Journal of the American Chemical Society*, 138(43):14449–14457, 11 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b09113. URL https://doi.org/10.1021/jacs.6b09113.

[113] Saeed Amirjalayer, Maxim Tafipolsky, and Rochus Schmid. Exploring Network Topologies of Copper Paddle Wheel Based Metal–Organic Frameworks with a First-Principles Derived Force Field. *The Journal of Physical Chemistry C*, 115(31):15133–15139, 8 2011. ISSN 1932-7447. doi: 10.1021/jp200123g. URL https://doi.org/10.1021/jp200123g.

[114] Stephen S.-Y. Chui, Samuel M.-F. Lo, Jonathan P H Charmant, A Guy Orpen, and Ian D Williams. A Chemically Functionalizable Nanoporous Material [Cu3(TMA)2(H2O)3]n. *Science*, 283(5405):1148 LP – 1150, 2 1999. doi: 10.1126/science.283.5405.1148. URL http://science.sciencemag.org/content/283/5405/1148.abstract.

[115] Banglin Chen, M Eddaoudi, S T Hyde, M O'Keeffe, and O M Yaghi. Interwoven Metal-Organic Framework on a Periodic Minimal Surface with Extra-Large Pores. *Science*, 291(5506):1021 LP – 1023, 2 2001. doi: 10.1126/science.1056598. URL http://science.sciencemag.org/content/291/5506/1021.abstract.

[116] Sarawoot Impeng, Ruel Cedeno, Johannes P Dürholt, Rochus Schmid, and Sareeya Bureekaew. Computational Structure Prediction of (4,4)-Connected Copper Paddle-wheel-based MOFs: Influence of Ligand Functionalization on the Topological Preference. *Crystal Growth & Design*, 18(5): 2699–2706, 5 2018. ISSN 1528-7483. doi: 10.1021/acs.cgd.8b00238. URL https://doi.org/10.1021/acs.cgd.8b00238.

[117] Yang Cai, Ambarish R Kulkarni, You-Gui Huang, David S Sholl, and Krista S Walton. Control of Metal–Organic Framework Crystal Topology by Ligand Functionalization: Functionalized HKUST-1 Derivatives. *Crystal Growth & Design*, 14(11):6122–6128, 11 2014. ISSN 1528-7483. doi: 10.1021/cg501285q. URL https://doi.org/10.1021/cg501285q.

[118] Mihails Arhangelskis, Athanassios D Katsenis, Novendra Novendra, Zamirbek Akimbekov, Dayaker Gandrath, Joseph M Marrett, Ghada Ayoub, Andrew J Morris, Omar K Farha, Tomislav Friščić, and Alexandra Navrotsky. Theoretical Prediction and Experimental Evaluation of Topological Landscape and Thermodynamic Stability of a Fluorinated Zeolitic Imidazolate Framework. *Chemistry of Materials*, 31(10):3777–3783, 5 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b00994. URL https://doi.org/10.1021/acs.chemmater.9b00994.

[119] Zhijie Chen, Sylvia L Hanna, Louis R Redfern, Dalal Alezi, Timur Islamoglu, and Omar K Farha. Reticular chemistry in the rational synthesis of functional zirconium cluster-based MOFs. *Coordination Chemistry Reviews*, 386:32–49, 2019. ISSN 0010-8545. doi: https://doi.org/10.1016/j.ccr.2019.01.017. URL http://www.sciencedirect.com/science/article/pii/S0010854518305514.

[120] Timur Islamoglu, Ken-ichi Otake, Peng Li, Cassandra T Buru, Aaron W Peters, Isil Akpinar, Sergio J Garibay, and Omar K Farha. Revisiting the structural homogeneity of NU-1000, a Zr-based metal–organic framework. *CrystEngComm*, 20(39):5913–5918, 2018. doi: 10.1039/C8CE00455B. URL http://dx.doi.org/10.1039/C8CE00455B.

[121] Thomas E Webber, Wei-Guang Liu, Sai Puneet Desai, Connie C Lu, Donald G Truhlar, and R Lee Penn. Role of a Modulator in the Synthesis of Phase-Pure NU-1000. *ACS Applied Materials & Interfaces*, 9(45):39342–39346, 11 2017. ISSN 1944-8244. doi: 10.1021/acsami.7b11348. URL https://doi.org/10.1021/acsami.7b11348.

[122] Sergio J Garibay, Ivan Iordanov, Timur Islamoglu, Jared B DeCoste, and Omar K Farha. Synthesis and functionalization of phase-pure NU-901 for enhanced CO2 adsorption: the influence of a zirconium salt and modulator on the topology and phase purity. *CrystEngComm*, 20(44):7066–7070, 2018. doi: 10.1039/C8CE01454J. URL http://dx.doi.org/10.1039/C8CE01454J.

[123] Peng Wang, Li Luo, Jian Fan, Gao-Chao Lv, You Song, and Wei-Yin Sun. Syntheses, structures, sorption and magnetic properties of copper(II) frameworks with varied topologies. *Microporous and Mesoporous Materials*, 175:116–124, 2013. ISSN 1387-1811. doi: https://doi.org/10.1016/j.micromeso.2013.03.021. URL http://www.sciencedirect.com/science/article/pii/S1387181113001492.

[124] Paul M Forster, Andrea R Burbank, Carine Livage, Gérard Férey, and Anthony K Cheetham. The role of temperature in the synthesis of hybrid inorganic–organic materials: the example of cobalt succinates. *Chemical Communications*, (4):368–369, 2004. ISSN 1359-7345. doi: 10.1039/B311156C. URL http://dx.doi.org/10.1039/B311156C.

[125] Yin-Xia Sun and Wei-Yin Sun. Influence of temperature on metal-organic frameworks. *Chinese Chemical Letters*, 25(6):823–828, 2014. ISSN 1001-8417. doi: https://doi.org/10.1016/j.cclet.2014.04.032. URL http://www.sciencedirect.com/science/article/pii/S100184171400206X.

[126] Jason A. Gee and David S. Sholl. Characterization of the Thermodynamic Stability of Solvated Metal–Organic Framework Polymorphs Using Molecular Simulations. *The Journal of Physical Chemistry C*, 117(40):20636–20642, 10 2013. ISSN 1932-7447. doi: 10.1021/jp405774e. URL https://doi.org/10.1021/jp405774e.

[127] Johannes Hoja, Anthony M Reilly, and Alexandre Tkatchenko. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. *WIREs Computational Molecular Science*, 7(1):e1294, 1 2017. ISSN 1759-0876. doi: 10.1002/wcms.1294. URL https://doi.org/10.1002/wcms.1294.

[128] Anthony M Reilly and Alexandre Tkatchenko. Role of Dispersion Interactions in the Polymorphism and Entropic Stabilization of the Aspirin Crystal. *Physical Review Letters*, 113(5):55701, 7 2014. doi: 10.1103/PhysRevLett.113.055701. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.055701.

[129] Jonas Nyman and Graeme M Day. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm*, 17(28):5154–5165, 2015. doi: 10.1039/C5CE00045A. URL http://dx.doi.org/10.1039/C5CE00045A.

[130] Mariana Rossi, Piero Gasparotto, and Michele Ceriotti. Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol. *Physical Review Letters*, 117(11):115702, 9 2016. doi: 10.1103/PhysRevLett.117.115702. URL https://link.aps.org/doi/10.1103/PhysRevLett.117.115702.

[131] Elia Schneider, Leslie Vogt, and Mark E Tuckerman. Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallographica Section B*, 72(4):542–550, 8 2016. doi: 10.1107/S2052520616007873. URL https://doi.org/10.1107/S2052520616007873.

[132] Anthony M Reilly and Alexandre Tkatchenko. Seamless and Accurate Modeling of Organic Molecular Materials. *The Journal of Physical Chemistry Letters*, 4(6):1028–1033, 3 2013. doi: 10.1021/jz400226x. URL https://doi.org/10.1021/jz400226x.

[133] David H Case, Josh E Campbell, Peter J Bygrave, and Graeme M Day. Convergence Properties of Crystal Structure Prediction by Quasi-Random Sampling. *Journal of Chemical Theory and Computation*, 12(2):910–924, 2 2016. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b01112. URL https://doi.org/10.1021/acs.jctc.5b01112.

[134] Farren Curtis, Xiayue Li, Timothy Rose, Álvaro Vázquez-Mayagoitia, Saswata Bhattacharya, Luca M Ghiringhelli, and Noa Marom. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation*, 14(4):2246–2264, 4 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b01152. URL https://doi.org/10.1021/acs.jctc.7b01152.

[135] Xiang Zhao, Yanxiang Wang, Dong-Sheng Li, Xianhui Bu, and Pingyun Feng. Metal–Organic Frameworks for Separation. *Advanced Materials*, 30(37):1705189, 9 2018. ISSN 0935-9648. doi: https://doi.org/10.1002/adma.201705189. URL https://doi.org/10.1002/adma.201705189.

[136] Zoey R Herm, Joseph A Swisher, Berend Smit, Rajamani Krishna, and Jeffrey R Long. MetalOrganic Frameworks as Adsorbents for Hydrogen Purification and Precombustion Carbon Dioxide Capture. *Journal of the American Chemical Society*, 133(15):5664–5667, 4 2011. ISSN 0002-7863. doi: 10.1021/ja111411q. URL https://doi.org/10.1021/ja111411q.

[137] Stephen R Caskey, Antek G Wong-Foy, and Adam J Matzger. Dramatic Tuning of Carbon Dioxide Uptake via Metal Substitution in a Coordination Polymer with Cylindrical Pores. *Journal of the American Chemical Society*, 130(33):10870–10871, 8 2008. ISSN 0002-7863. doi: 10.1021/ja8036096. URL https://doi.org/10.1021/ja8036096.

[138] Quan-Guo Zhai, Xianhui Bu, Chengyu Mao, Xiang Zhao, and Pingyun Feng. Systematic and Dramatic Tuning on Gas Sorption Performance in Heterometallic Metal–Organic Frameworks. *Journal of the American Chemical Society*, 138(8):2524–2527, 3 2016. ISSN 0002-7863. doi: 10.1021/jacs.5b13491. URL https://doi.org/10.1021/jacs.5b13491.

[139] Jian-Rong Li, Jiamei Yu, Weigang Lu, Lin-Bing Sun, Julian Sculley, Perla B Balbuena, and Hong-Cai Zhou. Porous materials with pre-designed single-molecule traps for CO2 selective adsorption. *Nature Communications*, 4(1):1538, 2013. ISSN 2041-1723. doi: 10.1038/ncomms2552. URL https://doi.org/10.1038/ncomms2552.

[140] Bin Wang, Hongliang Huang, Xiu-Liang Lv, Yabo Xie, Ming Li, and Jian-Rong Li. Tuning CO2 Selective Adsorption over N2 and CH4 in UiO-67 Analogues through Ligand Functionalization. *Inorganic Chemistry*, 53(17):9254–9259, 9 2014. ISSN 0020-1669. doi: 10.1021/ic5013473. URL https://doi.org/10.1021/ic5013473.

[141] Ji Woong Yoon, You-Kyong Seo, Young Kyu Hwang, Jong-San Chang, Hervé Leclerc, Stefan Wuttke, Philippe Bazin, Alexandre Vimont, Marco Daturi, Emily Bloch, Philip L. Llewellyn, Christian Serre, Patricia Horcajada, Jean-Marc Grenèche, Alirio E. Rodrigues, and Gérard Férey. Controlled Reducibility of a Metal–Organic Framework with Coordinatively Unsaturated Sites for Preferential Gas Sorption. *Angewandte Chemie International Edition*, 49(34):5949–5952, 8 2010. ISSN 1433-7851. doi: https://doi.org/10.1002/anie.201001230. URL https://doi.org/10.1002/anie.201001230.

[142] Zongbi Bao, Sufian Alnemrat, Liang Yu, Igor Vasiliev, Qilong Ren, Xiuyang Lu, and Shuguang Deng. Adsorption of Ethane, Ethylene, Propane, and Propylene on a Magnesium-Based Metal–Organic Framework. *Langmuir*, 27(22):13554–13562, 11 2011. ISSN 0743-7463. doi: 10.1021/la2030473. URL https://doi.org/10.1021/la2030473.

[143] Youn-sang Bae, Chang Yeon Lee, Ki Chul Kim, Omar K Farha, Peter Nickias, Joseph T Hupp, Sonbinh T Nguyen, and Randall Q Snurr. High Propene / Propane Selectivity in Isostructural Metal – Organic Frameworks with High Densities of Open Metal Sites ** Angewandte. pages 1893–1896, 2012. doi: 10.1002/ange.201107534.

[144] Sihai Yang, Anibal J Ramirez-Cuesta, Ruth Newby, Victoria Garcia-Sakai, Pascal Manuel, Samantha K Callear, Stuart I Campbell, Chiu C Tang, and Martin Schröder. Supramolecular binding and separation of hydrocarbons within a functionalized porous metal–organic framework. *Nature Chemistry*, 7(2):121–129, 2015. ISSN 1755-4349. doi: 10.1038/nchem.2114. URL https://doi.org/10.1038/nchem.2114.

[145] Ambarish R Kulkarni and David S Sholl. Screening of Copper Open Metal Site MOFs for Olefin/Paraffin Separations Using DFT-Derived Force Fields. *The Journal of Physical Chemistry C*, 120(40):23044–23054, 10 2016. ISSN 1932-7447. doi: 10.1021/acs.jpcc.6b07493. URL https://doi.org/10.1021/acs.jpcc.6b07493.

[146] Michael Fischer, José R B Gomes, Michael Fröba, and Miguel Jorge. Modeling Adsorption in Metal–Organic Frameworks with Open Metal Sites: Propane/Propylene Separations. *Langmuir*, 28 (22):8537–8549, 6 2012. ISSN 0743-7463. doi: 10.1021/la301215y. URL https://doi.org/10.1021/la301215y.

[147] Miguel Jorge, Michael Fischer, José R B Gomes, Christophe Siquet, João C Santos, and Alírio E Rodrigues. Accurate Model for Predicting Adsorption of Olefins and Paraffins on MOFs with Open Metal Sites. *Industrial & Engineering Chemistry Research*, 53(40):15475–15487, 10 2014. ISSN 0888-5885. doi: 10.1021/ie500310c. URL https://doi.org/10.1021/ie500310c.

[148] Cigdem Altintas and Seda Keskin. Computational screening of MOFs for C2H6/C2H4 and C2H6/CH4 separations. *Chemical Engineering Science*, 139:49–60, 2016. ISSN 0009-2509. doi: https://doi.org/10.1016/j.ces.2015.09.019. URL https://www.sciencedirect.com/science/article/pii/S0009250915006417.

[149] Jian-Rong Li, Julian Sculley, and Hong-Cai Zhou. Metal–Organic Frameworks for Separations. *Chemical Reviews*, 112(2):869–932, 2 2012. ISSN 0009-2665. doi: 10.1021/cr200190s. URL https://doi.org/10.1021/cr200190s.

[150] Jian Liu, Praveen K. Thallapally, and Denis Strachan. Metal-organic frameworks for removal of Xe and Kr from nuclear fuel reprocessing plants. *Langmuir*, 28(31):11584–11589, 2012. ISSN 07437463. doi: 10.1021/la301870n.

[151] U Mueller, M Schubert, F Teich, H Puetter, K Schierle-Arndt, and J Pastré. Metal–organic frameworks—prospective industrial applications. *Journal of Materials Chemistry*, 16(7):626–636, 2006. ISSN 0959-9428. doi: 10.1039/B511962F. URL http://dx.doi.org/10.1039/B511962F.

[152] John J Perry, Stephanie L Teich-McGoldrick, Scott T Meek, Jeffery A Greathouse, Maciej Haranczyk, and Mark D Allendorf. Noble Gas Adsorption in Metal–Organic Frameworks Containing Open Metal Sites. *The Journal of Physical Chemistry C*, 118(22):11685–11698, 6 2014. ISSN 1932-7447. doi: 10.1021/jp501495f. URL https://doi.org/10.1021/jp501495f.

[153] Ting Wu, Xuhui Feng, Sameh K Elsaidi, Praveen K Thallapally, and Moises A Carreon. Zeolitic Imidazolate Framework-8 (ZIF-8) Membranes for Kr/Xe Separation. *Industrial & Engineering Chemistry Research*, 56(6):1682–1686, 2 2017. ISSN 0888-5885. doi: 10.1021/acs.iecr.6b04868. URL https://doi.org/10.1021/acs.iecr.6b04868.

[154] Alternative Fuels Data Center: Natural Gas Benefits. URL https://afdc.energy.gov/fuels/natural_gas_benefits.html.

[155] R H Hugman, E H Vidas, and P S Springer. Chemical composition of discovered and undiscovered natural gas in the Lower-48 United States. Project Summary. Final report, 1 November 1988-31 March 1990, 11 1990.

[156] Hermes R. Sant Anna, Amaro G. Barreto, Frederico W. Tavares, and Jaílton F. do Nascimento. Methane/nitrogen separation through pressure swing adsorption process from nitrogen-rich streams. *Chemical Engineering and Processing: Process Intensification*, 103:70–79, 2016. ISSN 02552701. doi: 10.1016/j.cep.2015.11.002. URL http://dx.doi.org/10.1016/j.cep.2015.11.002.

[157] ATSDR - Landfill Gas Primer - Chapter 2: Landfill Gas Basics. URL https://www.atsdr.cdc.gov/hac/landfill/html/ch2.html.

[158] Liangying Li, Lifeng Yang, Jiawei Wang, Zhiguo Zhang, Qiwei Yang, Yiwen Yang, Qilong Ren, and Zongbi Bao. Highly efficient separation of methane from nitrogen on a squarate-based metal-organic framework. *AIChE Journal*, 64(10):3681–3689, 2018. ISSN 15475905. doi: 10.1002/aic.16335.

[159] Jiangtao Jia, Lei Wang, Fuxing Sun, Xiaofei Jing, Zheng Bian, Lianxun Gao, Rajamani Krishna, and Guangshan Zhu. The Adsorption and Simulated Separation of Light Hydrocarbons in Isoreticular Metal–Organic Frameworks Based on Dendritic Ligands with Different Aliphatic Side Chains. *Chemistry – A European Journal*, 20(29):9073–9080, 7 2014. ISSN 0947-6539. doi: https://doi.org/10.1002/chem.201304962. URL https://doi.org/10.1002/chem.201304962.

[160] Ji Woong Yoon, Hyunju Chang, Seung-Joon Lee, Young Kyu Hwang, Do-Young Hong, Su-Kyung Lee, Ji Sun Lee, Seunghun Jang, Tae-Ung Yoon, Kijeong Kwac, Yousung Jung, Renjith S Pillai, Florian Faucher, Alexandre Vimont, Marco Daturi, Gérard Férey, Christian Serre, Guillaume Maurin, Youn-Sang Bae, and Jong-San Chang. Selective nitrogen capture by porous hybrid materials containing accessible transition metal ion sites. *Nature Materials*, 16(5):526–531, 2017. ISSN 1476-4660. doi: 10.1038/nmat4825. URL https://doi.org/10.1038/nmat4825.

[161] Yadava Krishnan Ponraj and Bhaskarjyoti Borah. Separation of methane from ethane and propane by selective adsorption and diffusion in MOF Cu-BTC: A molecular simulation study. *Journal of Molecular Graphics and Modelling*, 97:107574, 2020. ISSN 1093-3263. doi: https://doi.org/10.1016/j.jmgm.2020.107574. URL https://www.sciencedirect.com/science/article/pii/S1093326319310083.

[162] Yadava Krishnan Ponraj and Bhaskarjyoti Borah. High-Throughput Computational Screening of Metal–Organic Frameworks for the Separation of Methane from Ethane and Propane. *The Journal of Physical Chemistry C*, 125(3):1839–1854, 1 2021. ISSN 1932-7447. doi: 10.1021/acs.jpcc.0c09117. URL https://doi.org/10.1021/acs.jpcc.0c09117.

[163] Cory M Simon, Jihan Kim, Diego A Gomez-Gualdron, Jeffrey S Camp, Yongchul G Chung, Richard L Martin, Rocio Mercado, Michael W Deem, Dan Gunter, Maciej Haranczyk, David S Sholl, Randall Q Snurr, and Berend Smit. The materials genome in action: identifying the performance limits for methane storage. *Energy & Environmental Science*, 8(4):1190–1199, 2015. ISSN 1754-5692. doi: 10.1039/C4EE03515A. URL http://dx.doi.org/10.1039/C4EE03515A.

[164] Berend Frenkel, Daan; Smith. *Understanding Molecular Simulations From Algorithms to Applications*. Academic Press, 2002.

[165] Argon. URL https://webbook.nist.gov/cgi/cbook.cgi?ID=C7440371&Mask=1E9F.

[166] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 6 1953. ISSN 0021-9606. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

[167] Abdulnour Y Toukmaji and John A Board. Ewald summation techniques in perspective: a survey. *Computer Physics Communications*, 95(2):73–92, 1996. ISSN 0010-4655. doi: https://doi.org/10.1016/0010-4655(96)00016-1. URL https://www.sciencedirect.com/science/article/pii/0010465596000161.

[168] Masakatsu Watanabe and William P Reinhardt. Direct dynamical calculation of entropy and free energy by adiabatic switching. *Physical Review Letters*, 65(26):3301–3304, 12 1990. doi: 10.1103/PhysRevLett.65.3301. URL https://link.aps.org/doi/10.1103/PhysRevLett.65.3301.

[169] Rodrigo Freitas, Mark Asta, and Maurice de Koning. Nonequilibrium free-energy calculation of solids using LAMMPS. *Computational Materials Science*, 112:333–341, 2016. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2015.10.050. URL http://www.sciencedirect.com/science/article/pii/S0927025615007089.

[170] P Hohenberg and W Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, 11 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

[171] W Kohn and L J Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 11 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[172] Carolin Strobl, J. Malley, and Gerhard Tutz. Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*, 14(4):323–348, 2009. doi: 10.1037/a0016973.An.

[173] Raul Rojas. *Neural Networks: A Systematic Introduction*. Springer-Verlag, Berlin, 1996.

[174] Abhijit Dasgupta and Adrian E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302, 1998. ISSN 1537274X. doi: 10.1080/01621459.1998.10474110.

[175] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8(1):289–317, 2016. ISSN 20734859. doi: 10.32614/rj-2016-021.

[176] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. ISSN 01621459. doi: 10.1198/016214502760047131.

[177] Yongchul G Chung, Diego A Gómez-Gualdrón, Peng Li, Karson T Leperi, Pravas Deria, Hongda Zhang, Nicolaas A Vermeulen, J Fraser Stoddart, Fengqi You, Joseph T Hupp, Omar K Farha, and Randall Q Snurr. In silico discovery of metal-organic frameworks for precombustion CO₂ capture using a genetic algorithm. *Science Advances*, 2(10):e1600909, 10 2016. doi: 10.1126/sciadv.1600909. URL http://advances.sciencemag.org/content/2/10/e1600909.abstract.

[178] Jianming Zheng, Jian Tian, Dangxin Wu, Meng Gu, Wu Xu, Chongmin Wang, Fei Gao, Mark H Engelhard, Ji-Guang Zhang, Jun Liu, and Jie Xiao. Lewis Acid–Base Interactions between Polysulfides and Metal Organic Framework in Lithium Sulfur Batteries. *Nano Letters*, 14(5): 2345–2352, 5 2014. ISSN 1530-6984. doi: 10.1021/nl404721h. URL https://doi.org/10.1021/nl404721h.

[179] G de Combarieu, M Morcrette, F Millange, N Guillou, J Cabana, C P Grey, I Margiolaki, G Férey, and J.-M. Tarascon. Influence of the Benzoquinone Sorption on the Structure and Electrochemical Performance of the MIL-53(Fe) Hybrid Porous Material in a Lithium-Ion Battery. *Chemistry of Materials*, 21(8):1602–1611, 4 2009. ISSN 0897-4756. doi: 10.1021/cm8032324. URL https://doi.org/10.1021/cm8032324.

[180] Ashlee J. Howarth, Yangyang Liu, Peng Li, Zhanyong Li, Timothy C. Wang, Joseph T. Hupp, and Omar K. Farha. Chemical, thermal and mechanical stabilities of metal-organic frameworks. *Nature Reviews Materials*, 1(15018):1–15, 2016. ISSN 20588437. doi: 10.1038/natrevmats.2015.18.

[181] Xiaoli Zhao, Shuangliu Liu, Zhi Tang, Hongyun Niu, Yaqi Cai, Wei Meng, Fengchang Wu, and John P Giesy. Synthesis of magnetic metal-organic framework (MOF) for efficient removal of organic dyes from water. *Scientific Reports*, 5(1):11849, 2015. ISSN 2045-2322. doi: 10.1038/srep11849. URL https://doi.org/10.1038/srep11849.

[182] Diego A Gomez-Gualdron, Sean T Dix, Rachel B Getman, and Randall Q Snurr. A modelling approach for MOF-encapsulated metal catalysts and application to n-butane oxidation. *Physical Chemistry Chemical Physics*, 17(41):27596–27608, 2015. ISSN 1463-9076. doi: 10.1039/C5CP04705F. URL http://dx.doi.org/10.1039/C5CP04705F.

[183] Mohammad R Momeni, Dale R Pahls, Dong Yang, Timothy C Wang, Omar K Farha, Joseph T Hupp, Christopher J Cramer, Laura Gagliardi, and Bruce C Gates. Correction to "Metal–Organic Framework Nodes as Nearly Ideal Supports for Molecular Catalysts: NU-1000- and UiO-66-Supported Iridium Complexes". *Journal of the American Chemical Society*, 139(50):18406, 12 2017. ISSN 0002-7863. doi: 10.1021/jacs.7b02871. URL https://doi.org/10.1021/jacs.7b02871.

[184] Hiroyasu Furukawa, Kyle E Cordova, Michael O'Keeffe, and Omar M Yaghi. The Chemistry and Applications of Metal-Organic Frameworks. *Science*, 341(6149):1230444, 8 2013. doi: 10.1126/science.1230444. URL http://science.sciencemag.org/content/341/6149/1230444.abstract.

[185] Chi-Kai Lin, Dan Zhao, Wen-Yang Gao, Zhenzhen Yang, Jingyun Ye, Tao Xu, Qingfeng Ge, Shengqian Ma, and Di-Jia Liu. Tunability of Band Gaps in Metal–Organic Frameworks. *Inorganic Chemistry*, 51(16):9039–9044, 8 2012. ISSN 0020-1669. doi: 10.1021/ic301189m. URL https://doi.org/10.1021/ic301189m.

[186] Adam J Rieth, Ashley M Wright, Sameer Rao, Hyunho Kim, Alina D LaPotin, Evelyn N Wang, and Mircea Dincă. Tunable Metal–Organic Frameworks Enable High-Efficiency Cascaded Adsorption Heat Pumps. *Journal of the American Chemical Society*, 140(50):17591–17596, 12 2018. ISSN 0002-7863. doi: 10.1021/jacs.8b09655. URL https://doi.org/10.1021/jacs.8b09655.

[187] Cheng Wang, Demin Liu, and Wenbin Lin. Metal–Organic Frameworks as A Tunable Platform for Designing Functional Molecular Materials. *Journal of the American Chemical Society*, 135(36): 13222–13234, 9 2013. ISSN 0002-7863. doi: 10.1021/ja308229p. URL https://doi.org/10.1021/ja308229p.

[188] Hui Wu, Yong Shen Chua, Vaiva Krungleviciute, Madhusudan Tyagi, Ping Chen, Taner Yildirim, and Wei Zhou. Unusual and highly tunable missing-linker defects in zirconium metal-organic framework UiO-66 and their important effects on gas adsorption. *Journal of the American Chemical Society*, 135(28):10525–10532, 2013. ISSN 00027863. doi: 10.1021/ja404514r.

[189] Peyman Z Moghadam, Aurelia Li, Seth B Wiggin, Andi Tao, Andrew G P Maloney, Peter A Wood, Suzanna C Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chemistry of Materials*, 29(7):2618–2625, 4 2017. ISSN 0897-4756. doi: 10.1021/acs.chemmater.7b00441. URL https://doi.org/10.1021/acs.chemmater.7b00441.

[190] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The Cambridge Structural Database. *Acta Crystallographica Section B*, 72(2):171–179, 4 2016. doi: 10.1107/S2052520616003954. URL https://doi.org/10.1107/S2052520616003954.

[191] Jacob Goldsmith, Antek G Wong-Foy, Michael J Cafarella, and Donald J Siegel. Theoretical Limits of Hydrogen Storage in Metal–Organic Frameworks: Opportunities and Trade-Offs. *Chemistry of Materials*, 25(16):3373–3382, 8 2013. ISSN 0897-4756. doi: 10.1021/cm401978e. URL https://doi.org/10.1021/cm401978e.

[192] Alex Scott. Round two for MOF commercialization, 2019. URL https://cen.acs.org/articles/95/i24/Round-two-MOF-commercialization.html.

[193] Senja Barthel, Eugeny V. Alexandrov, Davide M. Proserpio, and Berend Smit. Distinguishing Metal-Organic Frameworks. *Crystal Growth and Design*, 18(3):1738–1747, 2018. ISSN 15287505. doi: 10.1021/acs.cgd.7b01663.

[194] Benjamin J. Sikora, Randy Winnegar, Davide M. Proserpio, and Randall Q. Snurr. Textural properties of a large collection of computationally constructed MOFs and zeolites. *Microporous and Mesoporous Materials*, 186:207–213, 2014. ISSN 13871811. doi: 10.1016/j.micromeso.2013.11.041.

[195] S Bureekaew and R Schmid. Hypothetical 3D-periodic covalent organic frameworks: exploring the possibilities by a first principles derived force field. *Crystengcomm*, 15(8):1551–1562, 2013. ISSN 1466-8033. doi: Doi10.1039/C2ce26473k.

[196] Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Three-periodic nets and tilings: Edge-transitive binodal structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 62(5):350–355, 2006. ISSN 01087673. doi: 10.1107/S0108767306022707.

[197] Richard Luis Martin and Maciej Haranczyk. Exploring frontiers of high surface area metal–organic frameworks. *Chemical Science*, 4(4):1781, 3 2013. doi: 10.1039/c3sc00033h.

[198] Jierui Yu, JaeHong Park, Andrea Van Wyk, Garry Rumbles, and Pravas Deria. Excited-State Electronic Properties in Zr-Based Metal–Organic Frameworks as a Function of a Topological Network. *Journal of the American Chemical Society*, 140(33):10488–10496, 8 2018. ISSN 0002-7863. doi: 10.1021/jacs.8b04980.

[199] Pravas Deria, Jierui Yu, Rajesh P. Balaraman, Jamil Mashni, and Sandra N. White. Topology-dependent emissive properties of zirconium-based porphyrin MOFs. *Chemical Communications*, 52(88):13031–13034, 10 2016. doi: 10.1039/C6CC07343C.

[200] Lei Sun, Michael G Campbell, and Mircea Dincă. Electrically Conductive Porous Metal–Organic Frameworks. *Angewandte Chemie International Edition*, 55(11):3566–3579, 1 2016. ISSN 1433-7851. doi: 10.1002/anie.201506219.

[201] Jin-Hu Dou, Lei Sun, Yicong Ge, Wenbin Li, Christopher H Hendon, Ju Li, Sheraz Gul, Junko Yano, Eric A Stach, and Mircea Dincă. Signature of Metallic Behavior in the Metal–Organic Frameworks M3(hexaiminobenzene)2 (M = Ni, Cu). *Journal of the American Chemical Society*, 139(39): 13608–13611, 10 2017. ISSN 0002-7863. doi: 10.1021/jacs.7b07234.

[202] Joseph E. Mondloch, Taner Yildirim, Joseph T. Hupp, Oleksii V. Gutov, Omar K. Farha, Vaiva Krungleviciute, Randall Q. Snurr, Diego A. Gomez-Gualdron, Bhaskarjyoti Borah, Oleksii V. Gutov, Vaiva Krungleviciute, Bhaskarjyoti Borah, Joseph E. Mondloch, Joseph T. Hupp, Taner Yildirim, Omar K. Farha, and Randall Q. Snurr. Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chemistry of Materials*, 26 (19):5632–5639, 10 2014. ISSN 0897-4756. doi: 10.1021/cm502304e. URL https://doi.org/10.1021/cm502304ehttps://pubs.acs.org/sharingguidelines.

[203] Richard Luis Martin and Maciej Haranczyk. Insights into Multi-Objective Design of Metal–Organic Frameworks. *Crystal Growth & Design*, 13(10):4208–4212, 10 2013. ISSN 1528-7483. doi: 10.1021/cg401240f.

[204] Edwin Argueta, Jeena Shaji, Arun Gopalan, Peilin Liao, Randall Q Snurr, and Diego A Gómez-Gualdrón. Molecular Building Block-Based Electronic Charges for High-Throughput Screening of Metal–Organic Frameworks for Adsorption Applications. *Journal of Chemical Theory and Computation*, 14(1):365–376, 1 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00841. URL https://doi.org/10.1021/acs.jctc.7b00841.

[205] Olaf Delgado-Friedrichs and Michael O'Keeffe. Identification of and symmetry computation for crystal nets. *Acta Crystallographica Section A: Foundations of Crystallography*, 59(4):351–360, 2003. ISSN 01087673. doi: 10.1107/S0108767303012017.

[206] Jean Guillaume Eon. Euclidian embeddings of periodic nets: Definition of a topologically induced complete set of geometric descriptors for crystal structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 67(1):68–86, 2011. ISSN 01087673. doi: 10.1107/S0108767310042832.

[207] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 12 1997. ISSN 00983500. doi: 10.1145/279232.279236.

[208] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 9 1995. doi: 10.1137/0916069.

[209] Peter G Boyd, Seyed Mohamad Moosavi, Matthew Witman, and Berend Smit. Force-Field Prediction of Materials Properties in Metal-Organic Frameworks. *The Journal of Physical Chemistry Letters*, 8(2):357–363, 1 2017. doi: 10.1021/acs.jpclett.6b02532. URL https://doi.org/10.1021/acs.jpclett.6b02532.

[210] Matthew A Addicoat, Nina Vankova, Ismot Farjana Akter, and Thomas Heine. Extension of the Universal Force Field to Metal–Organic Frameworks. *Journal of Chemical Theory and Computation*, 10(2):880–891, 2 2014. ISSN 1549-9618. doi: 10.1021/ct400952t. URL https://doi.org/10.1021/ct400952t.

[211] Steve Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995. ISSN 0021-9991. doi: https://doi.org/10.1006/jcph.1995.1039. URL http://www.sciencedirect.com/science/article/pii/S002199918571039X.

[212] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural Relaxation Made Simple. *Physical Review Letters*, 97(17):170201, 10 2006. doi: 10.1103/PhysRevLett.97.170201. URL https://link.aps.org/doi/10.1103/PhysRevLett.97.170201.

[213] Xiang Lin, Irvin Telepeni, Alexander J Blake, Anne Dailly, Craig M Brown, Jason M Simmons, Marco Zoppi, Gavin S Walker, K Mark Thomas, Timothy J Mays, Peter Hubberstey, Neil R Champness, and Martin Schröder. High Capacity Hydrogen Adsorption in Cu(II) Tetracarboxylate Framework Materials: The Role of Pore Size, Ligand Functionalization, and Exposed Metal Sites. *Journal of the American Chemical Society*, 131(6):2159–2171, 2 2009. ISSN 0002-7863. doi: 10.1021/ja806624j. URL https://doi.org/10.1021/ja806624j.

[214] Banglin Chen, Nathan W Ockwig, Andrew R Millward, Damacio S Contreras, and Omar M Yaghi. High H2 Adsorption in a Microporous Metal–Organic Framework with Open Metal Sites. *Angewandte Chemie International Edition*, 44(30):4745–4749, 7 2005. ISSN 1433-7851. doi: 10.1002/anie.200462787. URL https://doi.org/10.1002/anie.200462787.

[215] Collin D Wick, Marcus G Martin, and J Ilja Siepmann. Transferable Potentials for Phase Equilibria. 4. United-Atom Description of Linear and Branched Alkenes and Alkylbenzenes. *The Journal of Physical Chemistry B*, 104(33):8008–8016, 8 2000. ISSN 1520-6106. doi: 10.1021/jp001044x.

[216] Marcus G Martin and J Ilja Siepmann. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J Phys Chem B*, 102:2569–2577, 1998.

[217] Meha Rungta, Liren Xu, and William J. Koros. Carbon molecular sieve dense film membranes derived from Matrimid® for ethylene/ethane separation. *Carbon*, 50(4):1488–1502, 4 2012. doi: 10.1016/J.CARBON.2011.11.019.

[218] Aurélie U. Ortiz, A. Boutin, Alain H. Fuchs, and François Xavier Coudert. Metal-organic frameworks with wine-rack motif: What determines their flexibility and elastic properties? *Journal of Chemical Physics*, 138(17), 2013. ISSN 00219606. doi: 10.1063/1.4802770.

[219] J W Morris and C R Krenn. The internal stability of an elastic solid. *Philosophical Magazine A*, 80 (12):2827–2840, 12 2000. ISSN 0141-8610. doi: 10.1080/01418610008223897. URL https://doi.org/10.1080/01418610008223897.

[220] R. Hill. The elastic behaviour of a crystalline aggregate. *Proceedings of the Physical Society. Section A*, 65(5):349–354, 1952. ISSN 03701298. doi: 10.1088/0370-1298/65/5/307.

[221] Yabing He, Wei Zhou, Guodong Qian, and Banglin Chen. Methane storage in metal-organic frameworks. *Chemical Society Reviews*, 43(16):5657–5678, 2014. ISSN 14604744. doi: 10.1039/c4cs00032c.

[222] Benjamin J Sikora, Christopher E Wilmer, Michael L Greenfield, and Randall Q Snurr. Thermodynamic analysis of Xe/Kr selectivity in over 137,000 hypothetical metal–organic frameworks. *Chemical Science*, 3(7):2217–2223, 2012. ISSN 2041-6520. doi: 10.1039/C2SC01097F. URL http://dx.doi.org/10.1039/C2SC01097F.

[223] Tian Tian, Zhixin Zeng, Diana Vulpe, Mirian E Casco, Giorgio Divitini, Paul A Midgley, Joaquin Silvestre-Albero, Jin-Chong Tan, Peyman Z Moghadam, and David Fairen-Jimenez. A sol–gel monolithic metal–organic framework with enhanced methane uptake. *Nature Materials*, 17:174, 12 2017. URL https://doi.org/10.1038/nmat5050http://10.0.4.14/nmat5050https://www.nature.com/articles/nmat5050#supplementary-information.

[224] Yang Peng, Vaiva Krungleviciute, Ibrahim Eryazici, Joseph T Hupp, Omar K Farha, and Taner Yildirim. Methane Storage in Metal–Organic Frameworks: Current Records, Surprise Findings, and Challenges. *Journal of the American Chemical Society*, 135(32):11887–11894, 8 2013. ISSN 0002-7863. doi: 10.1021/ja4045289. URL https://doi.org/10.1021/ja4045289.

[225] EIA. U.S. Energy-Related Carbon Dioxide Emissions. 2019. URL https://www.eia.gov/environment/emissions/carbon/.

[226] Eni Oko, Meihong Wang, and Atuman S Joel. Current status and future development of solvent-based carbon capture. *International Journal of Coal Science & Technology*, 4(1):5–14, 2017. ISSN 2198-7823. doi: 10.1007/s40789-017-0159-0.

[227] David J Heldebrant, Phillip K Koech, Vassiliki-Alexandra Glezakou, Roger Rousseau, Deepika Malhotra, and David C Cantu. Water-Lean Solvents for Post-Combustion CO2 Capture: Fundamentals, Uncertainties, Opportunities, and Outlook. *Chemical Reviews*, 117(14):9594–9624, 7 2017. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00768.

[228] Patricia Luis. Use of monoethanolamine (MEA) for CO2 capture in a global scenario: Consequences and alternatives. *Desalination*, 380:93–99, 2 2016. ISSN 0011-9164. doi: 10.1016/J.DESAL.2015.08.004.

[229] Bihong Lv, Bingsong Guo, Zuoming Zhou, and Guohua Jing. Mechanisms of CO2 Capture into Monoethanolamine Solution with Different CO2 Loading during the Absorption/Desorption Processes. *Environmental Science & Technology*, 49(17):10728–10735, 9 2015. ISSN 0013-936X. doi: 10.1021/acs.est.5b02356.

[230] D.C Department of Energy: Washington. Cost and Performance Baseline for Fossil Energy Plants Volume 1: Bituminous Coal and Natural Gas to Electricity, Revision 3; DOE/NETL-2010/1397;.

[231] Hiroyasu Furukawa, Kyle E Cordova, Michael O'Keeffe, and Omar M Yaghi. The Chemistry and Applications of Metal-Organic Frameworks. *Science*, 341(6149):974–980, 8 2013.

[232] Li-Chiang Lin, Adam H Berger, Richard L Martin, Jihan Kim, Joseph A Swisher, Kuldeep Jariwala, Chris H Rycroft, Abhoyjit S Bhown, Michael W Deem, Maciej Haranczyk, and Berend Smit. In silico screening of carbon-capture materials. *Nature Materials*, 11:633, 5 2012.

[233] Shyamapada Nandi, Sean Collins, Debanjan Chakraborty, Debasis Banerjee, Praveen K Thallapally, Tom K Woo, and Ramanathan Vaidhyanathan. Ultralow Parasitic Energy for Postcombustion CO2 Capture Realized in a Nickel Isonicotinate Metal–Organic Framework with Excellent Moisture Stability. *Journal of the American Chemical Society*, 139(5):1734–1737, 2 2017. ISSN 0002-7863. doi: 10.1021/jacs.6b10455.

[234] Timothy C Wang, Wojciech Bury, Diego A Gómez-Gualdrón, Nicolaas A Vermeulen, Joseph E Mondloch, Pravas Deria, Kainan Zhang, Peyman Z Moghadam, Amy A Sarjeant, Randall Q Snurr, J Fraser Stoddart, Joseph T Hupp, and Omar K Farha. Ultrahigh Surface Area Zirconium MOFs and Insights into the Applicability of the BET Theory. *Journal of the American Chemical Society*, 137(10):3585–3591, 3 2015. ISSN 0002-7863. doi: 10.1021/ja512973b. URL https://doi.org/10.1021/ja512973b.

[235] Mohamed Eddaoudi, Jaheon Kim, Nathaniel Rosi, David Vodak, Joseph Wachter, Michael O'Keeffe, and Omar M Yaghi. Systematic Design of Pore Size and Functionality in Isoreticular MOFs and Their Application in Methane Storage. *Science*, 295(5554):469 LP – 472, 1 2002. doi: 10.1126/science.1067208. URL http://science.sciencemag.org/content/295/5554/469.abstract.

[236] Yabing He, Bin Li, Michael O'Keeffe, and Banglin Chen. Multifunctional metal–organic frameworks constructed from meta-benzenedicarboxylate units. *Chem. Soc. Rev.*, 43(16):5618–5656, 4 2014. ISSN 0306-0012. doi: 10.1039/C4CS00041B.

[237] Sean P Collins, Thomas D Daff, Sarah S Piotrkowski, and Tom K Woo. Materials design by evolutionary optimization of functional groups in metal-organic frameworks. *Science Advances*, 2(11): e1600954, 11 2016. doi: 10.1126/sciadv.1600954. URL http://advances.sciencemag.org/content/2/11/e1600954.abstract.

[238] Yongchul G Chung, Diego A Gómez-Gualdrón, Peng Li, Karson T Leperi, Pravas Deria, Hongda Zhang, Nicolaas A Vermeulen, J Fraser Stoddart, Fengqi You, Joseph T Hupp, Omar K Farha, and Randall Q Snurr. In silico discovery of metal-organic frameworks for precombustion CO₂ capture using a genetic algorithm. *Science Advances*, 2(10):e1600909, 10 2016. doi: 10.1126/sciadv.1600909. URL http://advances.sciencemag.org/content/2/10/e1600909.abstract.

[239] Pravas Deria, Song Li, Hongda Zhang, Randall Q. Snurr, Joseph T. Hupp, and Omar K. Farha. A MOF platform for incorporation of complementary organic motifs for CO2 binding. *Chemical Communications*, 51(62):12478–12481, 7 2015. doi: 10.1039/C5CC04808G.

[240] Mario Wriedt, Julian P. Sculley, Andrey A. Yakovenko, Yuguang Ma, Gregory J. Halder, Perla B. Balbuena, and Hong-Cai Zhou. Low-Energy Selective Capture of Carbon Dioxide by a Pre-designed Elastic Single-Molecule Trap. *Angewandte Chemie International Edition*, 51(39):9804–9808, 9 2012. ISSN 14337851. doi: 10.1002/anie.201202992.

[241] Karim Adil, Prashant M. Bhatt, Youssef Belmabkhout, Sk Md Towsif Abtab, Hao Jiang, Ayalew H. Assen, Arijit Mallick, Amandine Cadiau, Jamal Aqil, and Mohamed Eddaoudi. Valuing Metal-Organic Frameworks for Postcombustion Carbon Capture: A Benchmark Study for Evaluating Physical Adsorbents. *Advanced Materials*, 29(39):1702953, 10 2017. ISSN 09359648. doi: 10.1002/adma.201702953.

[242] Jarad A. Mason, Kenji Sumida, Zoey R. Herm, Rajamani Krishna, and Jeffrey. R. Long. Evaluating metal–organic frameworks for post-combustion carbon dioxide capture via temperature swing adsorption. *Energy & Environmental Science*, 4(8):3030, 8 2011. ISSN 1754-5692. doi: 10.1039/c1ee01720a.

[243] Arunkumar Samanta, An Zhao, George K H Shimizu, Partha Sarkar, and Rajender Gupta. Post-Combustion CO2 Capture Using Solid Sorbents: A Review. *Industrial & Engineering Chemistry Research*, 51(4):1438–1463, 2 2012. ISSN 0888-5885. doi: 10.1021/ie200686q.

[244] Nathalie Casas, Johanna Schell, Richard Blom, and Marco Mazzotti. MOF and UiO-67/MCM-41 adsorbents for pre-combustion CO2 capture by PSA: Breakthrough experiments and process design. *Separation and Purification Technology*, 112:34–48, 7 2013. ISSN 1383-5866. doi: 10.1016/J.SEPPUR.2013.03.042.

[245] Yue-Biao Zhang, Hiroyasu Furukawa, Nakeun Ko, Weixuan Nie, Hye Jeong Park, Satoshi Okajima, Kyle E Cordova, Hexiang Deng, Jaheon Kim, and Omar M Yaghi. Introduction of Functionality, Selection of Topology, and Enhancement of Gas Adsorption in Multivariate Metal–Organic Framework-177. *Journal of the American Chemical Society*, 137(7):2641–2650, 2 2015. ISSN 0002-7863. doi: 10.1021/ja512311a.

[246] Christopher E Wilmer, Omar K Farha, Youn-Sang Bae, Joseph T Hupp, and Randall Q Snurr. Structure-property relationships of porous materials for carbon dioxide separation and capture. *Energy & Environmental Science*, 5(12):9849–9856, 2012. ISSN 1754-5692. doi: 10.1039/C2EE23201D.

[247] Bjørnar Arstad, Helmer Fjellvåg, Kjell Ove Kongshaug, Ole Swang, and Richard Blom. Amine functionalised metal organic frameworks (MOFs) as adsorbents for carbon dioxide. *Adsorption*, 14 (6):755–762, 2008. ISSN 1572-8757. doi: 10.1007/s10450-008-9137-6.

[248] Robinson W Flaig, Thomas M Osborn Popp, Alejandro M Fracaroli, Eugene A Kapustin, Markus J Kalmutzki, Rashid M Altamimi, Farhad Fathieh, Jeffrey A Reimer, and Omar M Yaghi. The Chemistry of CO2 Capture in an Amine-Functionalized Metal–Organic Framework under Dry and Humid Conditions. *Journal of the American Chemical Society*, 139(35):12125–12128, 9 2017. ISSN 0002-7863. doi: 10.1021/jacs.7b06382.

[249] Xingrui Wang, Huiquan Li, and Xin-Juan Hou. Amine-Functionalized Metal Organic Framework as a Highly Selective Adsorbent for CO2 over CO. *The Journal of Physical Chemistry C*, 116(37): 19814–19821, 9 2012. ISSN 1932-7447. doi: 10.1021/jp3052938.

[250] Oguarabau Benson, Ivan da Silva, Stephen P Argent, Rafel Cabot, Mathew Savage, Harry G W Godfrey, Yong Yan, Stewart F Parker, Pascal Manuel, Matthew J Lennox, Tamoghna Mitra, Timothy L Easun, William Lewis, Alexander J Blake, Elena Besley, Sihai Yang, and Martin Schröder. Amides Do Not Always Work: Observation of Guest Binding in an Amide-Functionalized Porous Metal–Organic Framework. *Journal of the American Chemical Society*, 138(45):14828–14831, 11 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b08059.

[251] Hexiang Deng, Christian J Doonan, Hiroyasu Furukawa, Ricardo B Ferreira, John Towne, Carolyn B Knobler, Bo Wang, and Omar M Yaghi. Multiple Functional Groups of Varying Ratios in Metal-Organic Frameworks. *Science*, 327(5967):846 LP – 850, 2 2010.

[252] Michael L Drummond, Thomas R Cundari, and Angela K Wilson. Cooperative Carbon Capture Capabilities in Multivariate MOFs Decorated with Amino Acid Side Chains: A Computational Study. *The Journal of Physical Chemistry C*, 117(28):14717–14722, 7 2013. ISSN 1932-7447. doi: 10.1021/jp4053538.

[253] Zhigang Hu, Abhishek Gami, Yuxiang Wang, and Dan Zhao. A Triphasic Modulated Hydrothermal Approach for the Synthesis of Multivariate Metal-Organic Frameworks with Hydrophobic Moieties for Highly Efficient Moisture-Resistant CO2 Capture. *Advanced Sustainable Systems*, 1(11):1700092, 11 2017. doi: 10.1002/adsu.201700092.

[254] Edder J García, Javier Pérez-Pellitero, Christian Jallut, and Gerhard D Pirngruber. How to Optimize the Electrostatic Interaction between a Solid Adsorbent and CO2. *The Journal of Physical Chemistry C*, 118(18):9458–9467, 5 2014. ISSN 1932-7447. doi: 10.1021/jp500209v.

[255] Anthony K Rappe and William A Goddard III. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem*, 95:3358–3363, 1991.

[256] Christopher E Wilmer, Ki Chul Kim, and Randall Q Snurr. An Extended Charge Equilibration Method. *The Journal of Physical Chemistry Letters*, 3(17):2506–2511, 9 2012. ISSN 1948-7185. doi: 10.1021/jz3008485.

[257] Eugene S Kadantsev, Peter G Boyd, Thomas D Daff, and Tom K Woo. Fast and Accurate Electrostatics in Metal Organic Frameworks with a Robust Charge Equilibration Parameterization for High-Throughput Virtual Screening of Gas Adsorption. *The Journal of Physical Chemistry Letters*, 4(18):3056–3061, 9 2013. ISSN 1948-7185. doi: 10.1021/jz401479k.

[258] Brad A Wells, Caspar De Bruin-Dickason, and Alan L Chaffee. Charge Equilibration Based on Atomic Ionization in Metal–Organic Frameworks. *The Journal of Physical Chemistry C*, 119(1): 456–466, 1 2015. ISSN 1932-7447. doi: 10.1021/jp510415h.

[259] Song Li, Yongchul G Chung, Cory M Simon, and Randall Q Snurr. High-Throughput Computational Screening of Multivariate Metal–Organic Frameworks (MTV-MOFs) for CO2 Capture. *The Journal of Physical Chemistry Letters*, 8(24):6135–6141, 12 2017. ISSN 1948-7185. doi: 10.1021/acs.jpclett.7b02700.

[260] Qing Xu and Chongli Zhong. A General Approach for Estimating Framework Charges in MetalOrganic Frameworks. *The Journal of Physical Chemistry C*, 114(11):5035–5042, 3 2010. ISSN 1932-7447. doi: 10.1021/jp910522h.

[261] Carlos Campañá, Bastien Mussard, and Tom K Woo. Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional. *Journal of Chemical Theory and Computation*, 5(10):2866–2878, 10 2009. ISSN 1549-9618. doi: 10.1021/ct9003405.

[262] Hailian Li, Mohamed Eddaoudi, M O'Keeffe, and O.M. Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(November):276–279, 1999. ISSN 00280836. doi: 10.1038/46248.

[263] Karl Petter Lillerud, Jasmina Hafizovic Cavka, Carlo Lamberti, Nathalie Guillou, Silvia Bordiga, Søren Jakobsen, and Unni Olsbye. A New Zirconium Inorganic Building Brick Forming Metal Organic Frameworks with Exceptional Stability. *Journal of the American Chemical Society*, 130(42): 13850–13851, 2008. ISSN 0002-7863. doi: 10.1021/ja8057953. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84926173614&doi=10.1016%2Fj.ica.2015.03.019&partnerID=40&md5=9df71383a8211e33c25fee459990b26f%0Ahttps://www.scopus.com/inward/record.uri?eid=2-s2.0-85013395216&doi=10.1021%2Fjacs.6b12885&partnerID=40&md5.

[264] G Férey, C Mellot-Draznieks, C Serre, F Millange, J Dutour, S Surblé, and I Margiolaki. A Chromium Terephthalate-Based Solid with Unusually Large Pore Volumes and Surface Area. *Science*, 309(5743):2040 LP – 2042, 9 2005.

[265] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian09 Revision A.02, 2016. Gaussian Inc. Wallingford CT.

[266] Yan Zhao and Donald G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other function. *Theoretical Chemistry Accounts*, 120(1-3):215–241, 2008. ISSN 1432881X. doi: 10.1007/s00214-007-0310-x.

[267] Yan Zhao and Donald G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *The Journal of Chemical Physics*, 125(19):194101, 11 2006. ISSN 0021-9606. doi: 10.1063/1.2370993.

[268] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–305, 2005. ISSN 1463-9076. doi: 10.1039/b508541a.

[269] Varinia Bernales, Manuel A Ortuño, Donald G Truhlar, Christopher J Cramer, and Laura Gagliardi. Computational Design of Functionalized Metal–Organic Framework Nodes for Catalysis. *ACS Central Science*, 4(1):5–19, 1 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00500.

[270] S F Boys and F Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics*, 19(4):553–566, 10 1970. ISSN 0026-8976. doi: 10.1080/00268977000101561.

[271] Ganesh Kamath, Nusrat Lubna, and Jeffrey J. Potoff. Effect of partial charge parametrization on the fluid phase behavior of hydrogen sulfide. *The Journal of Chemical Physics*, 123(12):124505, 9 2005. ISSN 0021-9606. doi: 10.1063/1.2049278.

[272] Farida Darkrim and Dominique Levesque. Monte Carlo simulations of hydrogen adsorption in single-walled carbon nanotubes. *The Journal of Chemical Physics*, 109(12):4981–4984, 9 1998. ISSN 0021-9606. doi: 10.1063/1.477109. URL https://doi.org/10.1063/1.477109.

[273] Diego A Gómez-Gualdrón, Peyman Z Moghadam, Joseph T Hupp, Omar K Farha, and Randall Q Snurr. Application of Consistency Criteria To Calculate BET Areas of Micro- And Mesoporous Metal-Organic Frameworks. *Journal of the American Chemical Society*, 138(1):215–24, 2016. ISSN 1520-5126. doi: 10.1021/jacs.5b10266. URL http://www.ncbi.nlm.nih.gov/pubmed/26651496.

[274] Orhan Talu and Alan L. Myers. Molecular simulation of adsorption: Gibbs dividing surface and comparison with experiment. *AIChE Journal*, 47(5):1160–1168, 5 2001. ISSN 00011541. doi: 10.1002/aic.690470521.

[275] R Core Team. R: A language and environment for statistical computing., 2017.

[276] D. Lattice Sarkar. *Multivariate Data Visualization with R*. Springer, New York, 2008.

[277] Shoushun Chen, Bryan E G Lucier, Paul D Boyle, and Yining Huang. Understanding The Fascinating Origins of CO2 Adsorption and Dynamics in MOFs. *Chemistry of Materials*, 28(16): 5829–5846, 8 2016. ISSN 0897-4756. doi: 10.1021/acs.chemmater.6b02239.

[278] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29:1189–1232, 2001. doi: 10.2307/2699986.

[279] Frederik Vermoortele, Michael Maes, Peyman Z Moghadam, Matthew J Lennox, Florence Ragon, Mohammed Boulhout, Shyam Biswas, Katrien G M Laurier, Isabelle Beurroies, Renaud Denoyel, Maarten Roeffaers, Norbert Stock, Tina Düren, Christian Serre, and Dirk E De Vos. p-Xylene-Selective Metal–Organic Frameworks: A Case of Topology-Directed Selectivity. *Journal of the American Chemical Society*, 133(46):18526–18529, 11 2011. ISSN 0002-7863. doi: 10.1021/ja207287h. URL https://doi.org/10.1021/ja207287h.

[280] Adrien P Côté, Annabelle I Benin, Nathan W Ockwig, Michael O'Keeffe, Adam J Matzger, and Omar M Yaghi. Porous, Crystalline, Covalent Organic Frameworks. *Science*, 310(5751):1166 LP – 1170, 11 2005. doi: 10.1126/science.1120411. URL http://science.sciencemag.org/content/310/5751/1166.abstract.

[281] Lukas Turcani, Rebecca L Greenaway, and Kim E Jelfs. Machine Learning for Organic Cage Property Prediction. *Chemistry of Materials*, 31(3):714–727, 2 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.8b03572. URL https://doi.org/10.1021/acs.chemmater.8b03572.

[282] David J Earl and Michael W Deem. Toward a Database of Hypothetical Zeolite Structures. *Industrial & Engineering Chemistry Research*, 45(16):5449–5454, 8 2006. ISSN 0888-5885. doi: 10.1021/ie0510728. URL https://doi.org/10.1021/ie0510728.

[283] Peter G Boyd, Yongjin Lee, and Berend Smit. Computational development of the nanoporous materials genome. *Nature Reviews Materials*, 2:17037, 7 2017. URL https://doi.org/10.1038/natrevmats.2017.37http://10.0.4.14/natrevmats.2017.37.

[284] Douglas Franz, Katherine A Forrest, Tony Pham, and Brian Space. Accurate H2 Sorption Modeling in the rht-MOF NOTT-112 Using Explicit Polarization. *Crystal Growth & Design*, 16(10):6024–6032, 10 2016. ISSN 1528-7483. doi: 10.1021/acs.cgd.6b01058. URL https://doi.org/10.1021/acs.cgd.6b01058.

[285] Christopher Campbell, José R B Gomes, Michael Fischer, and Miguel Jorge. New Model for Predicting Adsorption of Polar Molecules in Metal–Organic Frameworks with Unsaturated Metal Sites. *The Journal of Physical Chemistry Letters*, 9(12):3544–3553, 6 2018. ISSN 1948-7185. doi: 10.1021/acs.jpclett.8b00967. URL https://doi.org/10.1021/acs.jpclett.8b00967.

[286] Li-Chiang Lin, Kyuho Lee, Laura Gagliardi, Jeffrey B Neaton, and Berend Smit. Force-Field Development from Electronic Structure Calculations with Periodic Boundary Conditions: Applications to Gaseous Adsorption and Transport in Metal–Organic Frameworks. *Journal of Chemical Theory and Computation*, 10(4):1477–1488, 4 2014. ISSN 1549-9618. doi: 10.1021/ct500094w. URL https://doi.org/10.1021/ct500094w.

[287] Peyman Z Moghadam, Timur Islamoglu, Subhadip Goswami, Jason Exley, Marcus Fantham, Clemens F Kaminski, Randall Q Snurr, Omar K Farha, and David Fairen-Jimenez. Computer-aided discovery of a metal–organic framework with superior oxygen uptake. *Nature Communications*, 9(1): 1378, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03892-8. URL https://doi.org/10.1038/s41467-018-03892-8.

[288] Julian Keupp and Rochus Schmid. TopoFF: MOF structure prediction using specifically optimized blueprints. *Faraday Discussions*, 211(0):79–101, 2018. ISSN 1359-6640. doi: 10.1039/C8FD00051D. URL http://dx.doi.org/10.1039/C8FD00051D.

[289] Song Li, Yongchul G Chung, and Randall Q Snurr. High-Throughput Screening of Metal–Organic Frameworks for CO2 Capture in the Presence of Water. *Langmuir*, 32(40):10368–10376, 10 2016. ISSN 0743-7463. doi: 10.1021/acs.langmuir.6b02803. URL https://doi.org/10.1021/acs.langmuir.6b02803.

[290] Yongchul G Chung, Peng Bai, Maciej Haranczyk, Karson T Leperi, Peng Li, Hongda Zhang, Timothy C Wang, Tim Duerinck, Fengqi You, Joseph T Hupp, Omar K Farha, J Ilja Siepmann, and Randall Q Snurr. Computational Screening of Nanoporous Materials for Hexane and Heptane Isomer Separation. *Chemistry of Materials*, 29(15):6315–6328, 8 2017. ISSN 0897-4756. doi: 10.1021/acs.chemmater.7b01565. URL https://doi.org/10.1021/acs.chemmater.7b01565.

[291] Peng Bai, Mi Young Jeon, Limin Ren, Chris Knight, Michael W Deem, Michael Tsapatsis, and J Ilja Siepmann. Discovery of optimal zeolites for challenging separations and chemical transformations using predictive materials modeling. *Nature Communications*, 6:5912, 1 2015. URL https://doi.org/10.1038/ncomms6912http://10.0.4.14/ncomms6912https://www.nature.com/articles/ncomms6912#supplementary-information.

[292] Giorgos Borboudakis, Taxiarchis Stergiannakos, Maria Frysali, Emmanuel Klontzas, Ioannis Tsamardinos, and George E Froudakis. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Computational Materials*, 3(1):40, 2017. ISSN 2057-3960. doi: 10.1038/s41524-017-0045-8. URL https://doi.org/10.1038/s41524-017-0045-8.

[293] Ryther Anderson, Jacob Rodgers, Edwin Argueta, Achay Biong, and Diego A Gómez-Gualdrón. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chemistry of Materials*, 30(18):6325–6337, 9 2018. ISSN 0897-4756. doi: 10.1021/acs.chemmater.8b02257. URL https://doi.org/10.1021/acs.chemmater.8b02257.

[294] Dai Tang, Ying Wu, Ross J Verploegh, and David S Sholl. Efficiently Exploring Adsorption Space to Identify Privileged Adsorbents for Chemical Separations of a Diverse Set of Molecules. *ChemSusChem*, 11(9):1567–1575, 5 2018. ISSN 1864-5631. doi: 10.1002/cssc.201702289. URL https://doi.org/10.1002/cssc.201702289.

[295] Jared B DeCoste, Mitchell H Weston, Patrick E Fuller, Trenton M Tovar, Gregory W Peterson, M Douglas LeVan, and Omar K Farha. Metal–Organic Frameworks for Oxygen Storage. *Angewandte Chemie International Edition*, 53(51):14092–14095, 12 2014. ISSN 1433-7851. doi: 10.1002/anie.201408464. URL https://doi.org/10.1002/anie.201408464.

[296] Ismael Matito-Martos, Peyman Z Moghadam, Aurelia Li, Valentina Colombo, Jorge A R Navarro, Sofia Calero, and David Fairen-Jimenez. Discovery of an Optimal Porous Crystalline Material for the Capture of Chemical Warfare Agents. *Chemistry of Materials*, 30(14):4571–4579, 7 2018. ISSN 0897-4756. doi: 10.1021/acs.chemmater.8b00843. URL https://doi.org/10.1021/acs.chemmater.8b00843.

[297] Marta Mon, Rosaria Bruno, Jesus Ferrando-Soria, Donatella Armentano, and Emilio Pardo. Metal–organic framework technologies for water remediation: towards a sustainable ecosystem. *Journal of Materials Chemistry A*, 6(12):4912–4947, 2018. ISSN 2050-7488. doi: 10.1039/C8TA00264A. URL http://dx.doi.org/10.1039/C8TA00264A.

[298] Brandon R Barnett, Miguel I Gonzalez, and Jeffrey R Long. Recent Progress Towards Light Hydrocarbon Separations Using Metal–Organic Frameworks. *Trends in Chemistry*, 1(2):159–171, 2019. ISSN 2589-5974. doi: https://doi.org/10.1016/j.trechm.2019.02.012. URL http://www.sciencedirect.com/science/article/pii/S2589597419300255.

[299] Arwyn Evans, Ryan Luebke, and Camille Petit. The use of metal–organic frameworks for CO purification. *Journal of Materials Chemistry A*, 6(23):10570–10594, 2018. ISSN 2050-7488. doi: 10.1039/C8TA02059K. URL http://dx.doi.org/10.1039/C8TA02059K.

[300] Alan G Marshall and Ryan P Rodgers. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Accounts of Chemical Research*, 37(1):53–59, 1 2004. ISSN 0001-4842. doi: 10.1021/ar020177t. URL https://doi.org/10.1021/ar020177t.

[301] Nathan W Ockwig and Tina M Nenoff. Membranes for Hydrogen Separation. *Chemical Reviews*, 107 (4):4078–4110, 4 2007. ISSN 0009-2665. doi: 10.1021/cr078108l. URL https://doi.org/10.1021/cr078108lhttps://doi.org/10.1021/cr0501792.

[302] Nathan W Ockwig and Tina M Nenoff. Membranes for Hydrogen Separation. *Chemical Reviews*, 110 (4):2573–2574, 4 2010. ISSN 0009-2665. doi: 10.1021/cr078108l. URL https://doi.org/10.1021/cr078108l.

[303] Carlos A Grande, Simone Cavenati, Francisco A Da Silva, and Alírio E Rodrigues. Carbon Molecular Sieves for Hydrocarbon Separations by Adsorption. *Industrial & Engineering Chemistry Research*, 44 (18):7218–7227, 8 2005. ISSN 0888-5885. doi: 10.1021/ie050376r. URL https://doi.org/10.1021/ie050376r.

[304] H C Cheng and F B Hill. Separation of helium-methane mixtures by pressure swing adsorption. *AIChE Journal*, 31(1):95–102, 1 1985. ISSN 0001-1541. doi: 10.1002/aic.690310111. URL https://doi.org/10.1002/aic.690310111.

[305] Daniele Ongari, Peter G Boyd, Senja Barthel, Matthew Witman, Maciej Haranczyk, and Berend Smit. Accurate Characterization of the Pore Volume in Microporous Crystalline Materials. *Langmuir*, 33(51):14529–14538, 12 2017. ISSN 0743-7463. doi: 10.1021/acs.langmuir.7b01682. URL https://doi.org/10.1021/acs.langmuir.7b01682.

[306] E García-Pérez, J B Parra, C O Ania, D Dubbeldam, T J H Vlugt, J M Castillo, P J Merkling, and S Calero. Unraveling the Argon Adsorption Processes in MFI-Type Zeolite. *The Journal of Physical Chemistry C*, 112(27):9976–9979, 7 2008. ISSN 1932-7447. doi: 10.1021/jp803753h. URL https://doi.org/10.1021/jp803753h.

[307] Jeffrey J Potoff and J Ilja Siepmann. Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE Journal*, 47(7):1676–1682, 7 2001. ISSN 0001-1541. doi: 10.1002/aic.690470719. URL https://doi.org/10.1002/aic.690470719.

[308] D Levesque, A Gicquel, F Lamari Darkrim, and S Beyaz Kayiran. Monte Carlo simulations of hydrogen storage in carbon nanotubes. *Journal of Physics: Condensed Matter*, 14(40):9285–9293, 2002. ISSN 0953-8984. doi: 10.1088/0953-8984/14/40/318. URL http://dx.doi.org/10.1088/0953-8984/14/40/318.

[309] B Widom. Some Topics in the Theory of Fluids. *The Journal of Chemical Physics*, 39(11):2808–2812, 12 1963. ISSN 0021-9606. doi: 10.1063/1.1734110. URL https://doi.org/10.1063/1.1734110.

[310] Youn-Sang Bae, A Özgür Yazaydın, and Randall Q Snurr. Evaluation of the BET Method for Determining Surface Areas of MOFs and Zeolites that Contain Ultra-Micropores. *Langmuir*, 26(8): 5475–5483, 4 2010. ISSN 0743-7463. doi: 10.1021/la100449z. URL https://doi.org/10.1021/la100449z.

[311] Francois Chollet. Keras. *GitHub repository*, 2015. URL https://github.com/fchollet/keras.

[312] Thomas Klikauer. Scikit-learn: Machine Learning in Python. *TripleC*, 14(1):260–264, 2016. ISSN 1726670X. doi: 10.1007/s13398-014-0173-7.2. URL http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490.

[313] Derya Dokur and Seda Keskin. Effects of Force Field Selection on the Computational Ranking of MOFs for CO2 Separations. *Industrial & Engineering Chemistry Research*, 57(6):2298–2309, 2 2018. ISSN 0888-5885. doi: 10.1021/acs.iecr.7b04792. URL https://doi.org/10.1021/acs.iecr.7b04792.

[314] Weining Man, Aleksandar Donev, Frank H. Stillinger, Matthew T. Sullivan, William B. Russel, David Heeger, Souheil Inati, Salvatore Torquato, and P. M. Chaikin. Experiments on random packings of ellipsoids. *Physical Review Letters*, 94(19):1–4, 2005. ISSN 00319007. doi: 10.1103/PhysRevLett.94.198001.

[315] Aleksandar Donev, Frank H. Stillinger, P. M. Chaikin, and Salvatore Torquato. Unusually dense crystal packings of ellipsoids. *Physical Review Letters*, 92(25 I):1–4, 2004. ISSN 00319007. doi: 10.1103/PhysRevLett.92.255506.

[316] Tongan Yan, Youshi Lan, Minman Tong, and Chongli Zhong. Screening and Design of Covalent Organic Framework Membranes for CO2/CH4 Separation. *ACS Sustainable Chemistry & Engineering*, 7(1):1220–1227, 1 2019. doi: 10.1021/acssuschemeng.8b04858. URL https://doi.org/10.1021/acssuschemeng.8b04858.

[317] Jihan Kim, Mahmoud Abouelnasr, Li-Chiang Lin, and Berend Smit. Large-Scale Screening of Zeolite Structures for CO2 Membrane Separations. *Journal of the American Chemical Society*, 135(20): 7545–7552, 5 2013. ISSN 0002-7863. doi: 10.1021/ja400267g. URL https://doi.org/10.1021/ja400267g.

[318] Richard L. Martin, Thomas F. Willems, Li Chiang Lin, Jihan Kim, Joseph A. Swisher, Berend Smit, and MacIej Haranczyk. Similarity-driven discovery of zeolite materials for adsorption-based separations. *ChemPhysChem*, 13(16):3595–3597, 2012. ISSN 14394235. doi: 10.1002/cphc.201200554.

[319] Saulius Grazulis, Daniel Chateigner, Robert T Downs, A F T Yokochi, Miguel Quiros, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography Open Database - an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4): 726–729, 8 2009. ISSN 0021-8898. URL https://doi.org/10.1107/S0021889809016690.

[320] Michael W Deem, Ramdas Pophale, Phillip A Cheeseman, and David J Earl. Computational Discovery of New Zeolite-Like Materials. *The Journal of Physical Chemistry C*, 113(51):21353–21360, 12 2009. ISSN 1932-7447. doi: 10.1021/jp906984z. URL https://doi.org/10.1021/jp906984z.

[321] Daan Frenkel and Anthony J C Ladd. New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. *The Journal of Chemical Physics*, 81(7):3188–3193, 10 1984. ISSN 0021-9606. doi: 10.1063/1.448024. URL https://doi.org/10.1063/1.448024.

[322] J L Aragones, C Valeriani, and C Vega. Note: Free energy calculations for atomic solids through the Einstein crystal/molecule methodology using GROMACS and LAMMPS. *The Journal of Chemical Physics*, 137(14):146101, 10 2012. ISSN 0021-9606. doi: 10.1063/1.4758700. URL https://doi.org/10.1063/1.4758700.

[323] Ryther Anderson, Bomsaerah Seong, Zöe Peterson, Molly Stevanak, Moises A Carreon, and Diego A Gómez-Gualdrón. Exploiting hydrophobicity and hydrophilicity in nanopores as a design principle for "smart" MOF microtanks for methane storage. *Molecular Systems Design & Engineering*, 5(1): 166–176, 2020. doi: 10.1039/C9ME00072K. URL http://dx.doi.org/10.1039/C9ME00072K.

[324] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):14101, 1 2007. ISSN 0021-9606. doi: 10.1063/1.2408420. URL https://doi.org/10.1063/1.2408420.

[325] Manolis Vasileiadis. Calculation of the free energy of crystalline solids. *Imperial College London*, (November), 2013. URL https://spiral.imperial.ac.uk/bitstream/10044/1/12638/3/Vasileiadis-M-2013-PhD-Thesis.pdf.

[326] C Pastorino and Z Gamba. Free-energy calculations of elemental sulphur crystals via molecular dynamics simulations. *The Journal of Chemical Physics*, 119(4):2147–2154, 7 2003. ISSN 0021-9606. doi: 10.1063/1.1582840. URL https://doi.org/10.1063/1.1582840.

[327] Shiang Tai Lin, Mario Blanco, and William A. Goddard. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *Journal of Chemical Physics*, 119(22):11792–11805, 2003. ISSN 00219606. doi: 10.1063/1.1624057.

[328] Vito Dario Camiola and Valentina Tozzini. Collective Mode Mining from Molecular Dynamics Simulations: A Comparative Approach. *International Journal of Computational Methods*, 15(3):1–18, 2018. ISSN 17936969. doi: 10.1142/S0219876218501086.

[329] Maurice de Koning. Optimizing the driving function for nonequilibrium free-energy calculations in the linear regime: A variational approach. *The Journal of Chemical Physics*, 122(10):104106, 3 2005. ISSN 0021-9606. doi: 10.1063/1.1860556. URL https://doi.org/10.1063/1.1860556.

[330] Peter Nowak and James P. Stoessel. Absolute Free Energies in Biomolecular Systems. *Macromolecules*, 23(7):1961–1965, 1990. ISSN 15205835. doi: 10.1021/ma00209a014.

[331] Alexander E J Hoffman, Louis Vanduyfhuys, Irena Nevjestić, Jelle Wieme, Sven M J Rogge, Hannes Depauw, Pascal Van Der Voort, Henk Vrielinck, and Veronique Van Speybroeck. Elucidating the Vibrational Fingerprint of the Flexible Metal–Organic Framework MIL-53(Al) Using a Combined Experimental/Computational Approach. *The Journal of Physical Chemistry C*, 122(5):2734–2746, 2 2018. ISSN 1932-7447. doi: 10.1021/acs.jpcc.7b11031. URL https://doi.org/10.1021/acs.jpcc.7b11031.

[332] Alexander Krylov, Alexander Vtyurin, Petko Petkov, Irena Senkovska, Mariia Maliuta, Volodymyr Bon, Thomas Heine, Stefan Kaskel, and Evgenia Slyusareva. Raman spectroscopy studies of the terahertz vibrational modes of a DUT-8 (Ni) metal–organic framework. *Physical Chemistry Chemical Physics*, 19(47):32099–32104, 2017. ISSN 1463-9076. doi: 10.1039/C7CP06225G. URL http://dx.doi.org/10.1039/C7CP06225G.

[333] Tomas Kamencek, Natalia Bedoya-Martínez, and Egbert Zojer. Understanding phonon properties in isoreticular metal-organic frameworks from first principles. *Physical Review Materials*, 3(11):116003, 11 2019. ISSN 24759953. doi: 10.1103/PhysRevMaterials.3.116003. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.3.116003.

[334] Bartolomeo Civalleri, Francesco Napoli, Yves Noël, Carla Roetti, and Roberto Dovesi. Ab-initio prediction of materials properties with CRYSTAL: MOF-5 as a case study. *CrystEngComm*, 8(5): 364–371, 2006. doi: 10.1039/B603150C. URL http://dx.doi.org/10.1039/B603150C.

[335] Johannes P Dürholt, Guillaume Fraux, François-Xavier Coudert, and Rochus Schmid. Ab Initio Derived Force Fields for Zeolitic Imidazolate Frameworks: MOF-FF for ZIFs. *Journal of Chemical Theory and Computation*, 15(4):2420–2432, 4 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01041. URL https://doi.org/10.1021/acs.jctc.8b01041.

[336] Jessica K Bristow, Davide Tiana, and Aron Walsh. Transferable Force Field for Metal–Organic Frameworks from First-Principles: BTW-FF. *Journal of Chemical Theory and Computation*, 10(10): 4644–4652, 10 2014. ISSN 1549-9618. doi: 10.1021/ct500515h. URL https://doi.org/10.1021/ct500515h.

[337] Sareeya Bureekaew, Saeed Amirjalayer, Maxim Tafipolsky, Christian Spickermann, Tapta Kanchan Roy, and Rochus Schmid. MOF-FF – A flexible first-principles derived force field for metal-organic frameworks. *physica status solidi (b)*, 250(6):1128–1141, 6 2013. ISSN 0370-1972. doi: 10.1002/pssb.201248460. URL https://doi.org/10.1002/pssb.201248460.

[338] Yi Li, Jihong Yu, and Ruren Xu. Criteria for Zeolite Frameworks Realizable for Target Synthesis. *Angewandte Chemie International Edition*, 52(6):1673–1677, 2 2013. ISSN 1433-7851. doi: 10.1002/anie.201206340. URL https://doi.org/10.1002/anie.201206340.

[339] Colby J Dawson, Vitaliy Kapko, Michael F Thorpe, Martin D Foster, and Michael M J Treacy. Flexibility As an Indicator of Feasibility of Zeolite Frameworks. *The Journal of Physical Chemistry C*, 116(30):16175–16181, 8 2012. ISSN 1932-7447. doi: 10.1021/jp2107473. URL https://doi.org/10.1021/jp2107473.

[340] Yuriy G Bushuev and German Sastre. Feasibility of Pure Silica Zeolites. *The Journal of Physical Chemistry C*, 114(45):19157–19168, 11 2010. ISSN 1932-7447. doi: 10.1021/jp107296e. URL https://doi.org/10.1021/jp107296e.

[341] Alexandra Simperler, Martin D. Foster, Olaf Delgado Friedrichs, Robert G. Bell, Filipe A. Almeida Paz, and Jacek Klinowski. Hypothetical binodal zeolitic frameworks. *Acta Crystallographica Section B: Structural Science*, 61(3):263–279, 2005. ISSN 01087681. doi: 10.1107/S0108768105013340.

[342] Michal Mazur, Paul S Wheatley, Marta Navarro, Wieslaw J Roth, Miroslav Položij, Alvaro Mayoral, Pavla Eliášová, Petr Nachtigall, Jiří Čejka, and Russell E Morris. Synthesis of 'unfeasible' zeolites. *Nature Chemistry*, 8(1):58–62, 2016. ISSN 1755-4349. doi: 10.1038/nchem.2374. URL https://doi.org/10.1038/nchem.2374.

[343] Benjamin J Bucior, Andrew S Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E Ziebel, Omar K Farha, Joseph T Hupp, J Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q Snurr. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design*, 19(11):6682–6697, 11 2019. ISSN 1528-7483. doi: 10.1021/acs.cgd.9b01050. URL https://doi.org/10.1021/acs.cgd.9b01050.

[344] Hilal Daglar and Seda Keskin. Computational Screening of Metal–Organic Frameworks for Membrane-Based CO2/N2/H2O Separations: Best Materials for Flue Gas Separation. *The Journal of Physical Chemistry C*, 122(30):17347–17357, 8 2018. ISSN 1932-7447. doi: 10.1021/acs.jpcc.8b05416. URL https://doi.org/10.1021/acs.jpcc.8b05416.

[345] Justyna Rogacka, Agnieszka Seremak, Azahara Luna-Triguero, Filip Formalik, Ismael Matito-Martos, Lucyna Firlej, Sofia Calero, and Bogdan Kuchta. High-throughput screening of metal – Organic frameworks for CO2 and CH4 separation in the presence of water. *Chemical Engineering Journal*, 403:126392, 2021. ISSN 1385-8947. doi: https://doi.org/10.1016/j.cej.2020.126392. URL https://www.sciencedirect.com/science/article/pii/S1385894720325201.

[346] Ruimin Ma, Yamil J Colón, and Tengfei Luo. Transfer Learning Study of Gas Adsorption in Metal–Organic Frameworks. *ACS Applied Materials & Interfaces*, 12(30):34041–34048, 7 2020. ISSN 1944-8244. doi: 10.1021/acsami.0c06858. URL https://doi.org/10.1021/acsami.0c06858.

[347] Naomi F Cessford, Nigel A Seaton, and Tina Düren. Evaluation of Ideal Adsorbed Solution Theory as a Tool for the Design of Metal–Organic Framework Materials. *Industrial & Engineering Chemistry Research*, 51(13):4911–4921, 4 2012. ISSN 0888-5885. doi: 10.1021/ie202219w. URL https://doi.org/10.1021/ie202219w.

[348] Timothy Van Heest, Stephanie L Teich-McGoldrick, Jeffery A Greathouse, Mark D Allendorf, and David S Sholl. Identification of Metal–Organic Framework Materials for Adsorption Separation of Rare Gases: Applicability of Ideal Adsorbed Solution Theory (IAST) and Effects of Inaccessible Framework Regions. *The Journal of Physical Chemistry C*, 116(24):13183–13195, 6 2012. ISSN 1932-7447. doi: 10.1021/jp302808j. URL https://doi.org/10.1021/jp302808j.

[349] Yeliz Gurdal and Seda Keskin. Predicting Noble Gas Separation Performance of Metal Organic Frameworks Using Theoretical Correlations. *The Journal of Physical Chemistry C*, 117(10): 5229–5241, 3 2013. ISSN 1932-7447. doi: 10.1021/jp312838v. URL https://doi.org/10.1021/jp312838v.

[350] Sylwester Furmaniak, Stanisław Koter, Artur P Terzyk, Piotr A Gauden, Piotr Kowalczyk, and Gerhard Rychlicki. New insights into the ideal adsorbed solution theory. *Physical Chemistry Chemical Physics*, 17(11):7232–7247, 2015. ISSN 1463-9076. doi: 10.1039/C4CP05498A. URL http://dx.doi.org/10.1039/C4CP05498A.

[351] Allison N Dickey, A Özgür Yazaydın, Richard R Willis, and Randall Q Snurr. Screening CO2/N2 selectivity in metal-organic frameworks using Monte Carlo simulations and ideal adsorbed solution theory. *The Canadian Journal of Chemical Engineering*, 90(4):825–832, 8 2012. ISSN 0008-4034. doi: https://doi.org/10.1002/cjce.20700. URL https://doi.org/10.1002/cjce.20700.

[352] R Krishna and J M van Baten. Using molecular simulations for screening of zeolites for separation of CO2/CH4 mixtures. *Chemical Engineering Journal*, 133(1):121–131, 2007. ISSN 1385-8947. doi: https://doi.org/10.1016/j.cej.2007.02.011. URL https://www.sciencedirect.com/science/article/pii/S1385894707001106.

[353] Jiangliang Hu, Tianjun Sun, Xiaowei Liu, Ya Guo, and Shudong Wang. Separation of CH4/N2 mixtures in metal–organic frameworks with 1D micro-channels. *RSC Advances*, 6(68):64039–64046, 2016. doi: 10.1039/C6RA12280A. URL http://dx.doi.org/10.1039/C6RA12280A.

[354] Sandrine Péron, Sujoy Mukhopadhyay, and Michael Huh. A new dual stainless steel cryogenic trap for efficient separation of krypton from argon and xenon. *Journal of Analytical Atomic Spectrometry*, 35(11):2663–2671, 2020. ISSN 0267-9477. doi: 10.1039/D0JA00052C. URL http://dx.doi.org/10.1039/D0JA00052C.

[355] Xi-Ze Dong, Florian Ritterbusch, Yan-Qing Chu, Ji-Qiang Gu, Shui-Ming Hu, Wei Jiang, Zheng-Tian Lu, Guo-Min Yang, and Lei Zhao. Dual Separation of Krypton and Argon from Environmental Samples for Radioisotope Dating. *Analytical Chemistry*, 91(21):13576–13581, 11 2019. ISSN 0003-2700. doi: 10.1021/acs.analchem.9b02716. URL https://doi.org/10.1021/acs.analchem.9b02716.

[356] Ryther Anderson and Diego A Gómez-Gualdrón. Large-Scale Free Energy Calculations on a Computational Metal–Organic Frameworks Database: Toward Synthetic Likelihood Predictions. *Chemistry of Materials*, 32(19):8106–8119, 10 2020. ISSN 0897-4756. doi: 10.1021/acs.chemmater.0c00744. URL https://doi.org/10.1021/acs.chemmater.0c00744.

[357] Luca Scrucca. Dimension reduction for model-based clustering. *Statistics and Computing*, 20(4): 471–484, 2010. ISSN 1573-1375. doi: 10.1007/s11222-009-9138-7. URL https://doi.org/10.1007/s11222-009-9138-7.

[358] Xing Duan, Wei Zheng, Ben Yu, and Zhenguo Ji. A microporous metal–organic framework with soc topology for adsorption and separation selectivity of C2H2/CO2. *Chemical Papers*, 73(9):2371–2375, 2019. ISSN 1336-9075. doi: 10.1007/s11696-019-00794-x. URL https://doi.org/10.1007/s11696-019-00794-x.

[359] Hee K Chae, Jaheon Kim, Olaf Delgado Friedrichs, Michael O'Keeffe, and Omar M Yaghi. Design of Frameworks with Mixed Triangular and Octahedral Building Blocks Exemplified by the Structure of [Zn4O(TCA)2] Having the Pyrite Topology. *Angewandte Chemie International Edition*, 42(33): 3907–3909, 8 2003. ISSN 1433-7851. doi: https://doi.org/10.1002/anie.200351546. URL https://doi.org/10.1002/anie.200351546.

[360] Jiangtao Jia, Fuxing Sun, Tsolmon Borjigin, Hao Ren, Tingting Zhang, Zheng Bian, Lianxun Gao, and Guangshan Zhu. Highly porous and robust ionic MOFs with nia topology constructed by connecting an octahedral ligand and a trigonal prismatic metal cluster. *Chemical Communications*, 48(48):6010–6012, 2012. ISSN 1359-7345. doi: 10.1039/C2CC32347H. URL http://dx.doi.org/10.1039/C2CC32347H.

[361] Zhenxia Chen, Shengchang Xiang, Tengbiao Liao, Yongtai Yang, Yu-Sheng Chen, Yaming Zhou, Dongyuan Zhao, and Banglin Chen. A New Multidentate Hexacarboxylic Acid for the Construction of Porous MetalOrganic Frameworks of Diverse Structures and Porosities. *Crystal Growth & Design*, 10(6):2775–2779, 6 2010. ISSN 1528-7483. doi: 10.1021/cg100316s. URL https://doi.org/10.1021/cg100316s.

[362] Zhiyong Guo, Hui Wu, Gadipelli Srinivas, Yaming Zhou, Shengchang Xiang, Zhenxia Chen, Yongtai Yang, Wei Zhou, Michael O'Keeffe, and Banglin Chen. A Metal–Organic Framework with Optimized Open Metal Sites and Pore Spaces for High Methane Storage at Room Temperature. *Angewandte Chemie International Edition*, 50(14):3178–3181, 3 2011. ISSN 1433-7851. doi: 10.1002/anie.201007583. URL https://doi.org/10.1002/anie.201007583.

[363] Telikepalli Kavitha, Christian Liebchen, Kurt Mehlhorn, Dimitrios Michail, Romeo Rizzi, Torsten Ueckerdt, and Katharina A. Zweig. Cycle bases in graphs characterization, algorithms, complexity, and applications. *Computer Science Review*, 3(4):199–243, 11 2009. doi: 10.1016/J.COSREV.2009.08.001.

[364] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 6 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163.

[365] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006. ISSN 10618600. doi: 10.1198/106186006X133933.

[366] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. ISSN 09603174. doi: 10.1023/B:STCO.0000035301.49549.88.

[367] Richard P. Lippmann. An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*, 4 (2):4–22, 1987. ISSN 07407467. doi: 10.1109/MASSP.1987.1165576.

[368] Amit Gupta and Monica Lam. The weight decay backpropagation for generalizations with missing values. *Annals of Operations Research*, 78:165–187, 1998. ISSN 02545330. doi: 10.1023/A:1018945915940.

[369] Greg Ridgeway. Generalized Boosted Models: A guide to the gbm package. *Compute*, 1(4):1–12, 2007. ISSN 14679752. doi: 10.1111/j.1467-9752.1996.tb00390.x.

[370] Greg Ridgeway. The state of boosting. *Computing Science and Statistics*, 31:172–181, 1999. doi: citeulike-article-id:7678637.

[371] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 1 2015. ISSN 1061-8600. doi: 10.1080/10618600.2014.907095.

[372] David Ardia, Kris Boudt, Peter Carl, Katharine M Mullen, and Brian G Peterson. Differential Evolution with DEoptim An Application to Non-Convex Portfolio Optimiza- tion. *The R Journal*, 3 (1):27–34, 2011.

[373] Sanggyu Chong, Günther Thiele, and Jihan Kim. Excavating hidden adsorption sites in metal-organic frameworks using rational defect engineering. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-01478-4. URL http://dx.doi.org/10.1038/s41467-017-01478-4.

[374] J M Polson, E Trizac, S Pronk, and D Frenkel. Finite-size corrections to the free energies of crystalline solids. *The Journal of Chemical Physics*, 112(12):5339–5342, 3 2000. ISSN 0021-9606. doi: 10.1063/1.481102. URL https://doi.org/10.1063/1.481102.

[375] Jessica K Bristow, Jonathan M Skelton, Katrine L Svane, Aron Walsh, and Julian D Gale. A general forcefield for accurate phonon properties of metal–organic frameworks. *Physical Chemistry Chemical Physics*, 18(42):29316–29329, 2016. ISSN 1463-9076. doi: 10.1039/C6CP05106E. URL http://dx.doi.org/10.1039/C6CP05106E.

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 3

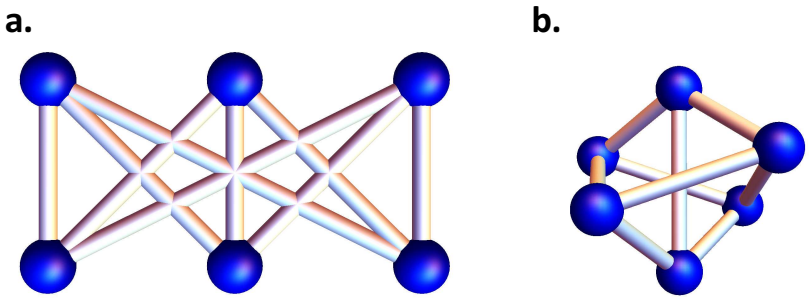

Figure A.1 Example of faithful vs. unfaithful embedding of a graph. (a) The $K_{3,3}$ graph (known as the utility graph) unfaithfully embedded in $\mathbb{R}^2$. Notice that edges cross as they lie in the same plane, (b) The $K_{3,3}$ graph faithfully embedded in $\mathbb{R}^3$. Notice that edges do not cross.

A.1 The ToBaCCo Net Scaling Algorithm

Once ToBaCCo has made the assignment for the molecular building blocks to each vertex and edge type in the template, the template needs to be rescaled to fit the geometry of specific building blocks as well as possible. To quantify the fit, an objective function defined by the total deviation between the template and the building blocks is specified. To optimize the fit, the objective function is to be minimized, resulting in a rescaled template that overlaps the chosen building blocks as closely as possible. The voltage graph (or labeled quotient graph) representation of a crystallographic net makes defining this deviation relatively straightforward. However, let us first discuss what a voltage graph is.1–3 To make the explanation more accessible, we often make reference to how the crystallographic **qtz** net was converted to a voltage graph (Figure A.4).

To construct the **qtz** voltage graph, we take the unit cell of the **qtz** net in Euclidian space as the reference (Figure A.4a). First, we note that V1 (or any V1 translational equivalent) is connected to four other vertices: i) two vertices that are translationally equivalent to V2, ii) V3, and iii) one vertex that is translationally equivalent to V3. A similar analysis can be made for V2 and V3. Then, to construct the voltage graph, we first define quotient groups, where the members of each group are translationally equivalent vertices. Each quotient group is represented by a vertex in the voltage graph, and each edge of

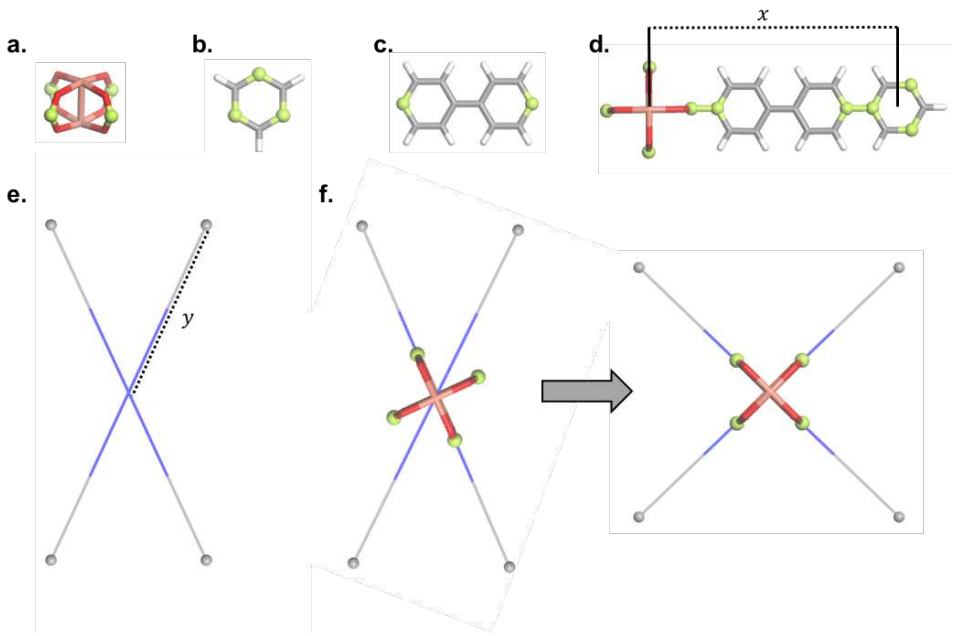

Figure A.2 Building blocks and illustration of the need for anisotropic rescaling to properly map a building block onto a node in a topological blueprint. Connection sites are highlighted in yellow for all shown building blocks. (a) 4-connected Cu-Paddlewheel (a nodular building block), (b) 3-connected phenyl (a nodular building block), (c) a 2-connected diphenyl (a connecting building block), and (d) how building blocks connect with each other during MOF computational synthesis (the anticipated value of "x" is used in rescaling procedures). (e) The vertex geometry of 4-connected nodes in the barycentric representation of the **sty** topology. The vertex itself is represented in blue and the connected vertices are presented as gray points (the value of "y" is used in rescaling procedure). (f) how anisotropic rescaling of the net allows the node geometry to better fit the Cu-paddlewheel, which in turn allows the paddlewheel to be properly mapped onto the node. (note that the rescaling is such that "y" becomes equal to "x").

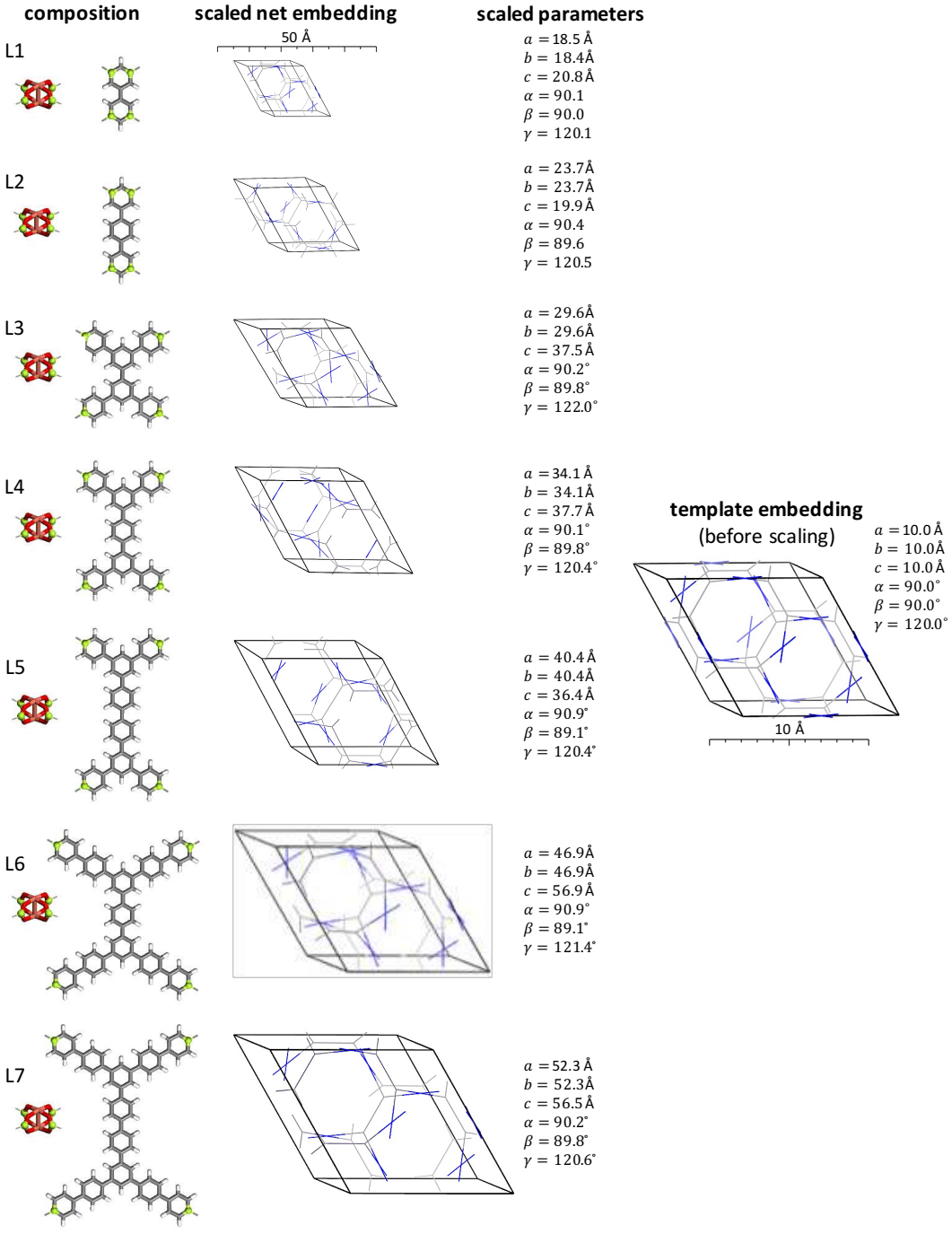

Figure A.3 Example of how ToBaCCo anisotropically rescaled the same topological blueprint to build seven MOFs based on linkers of different size and aspect ratio. The original barycentric embedding of the **fog** topology (chosen as example) is shown to the right. Inorganic nodes and organic linkers used in MOF construction are shown to the left. The anisotropically rescaled net embeddings (all shown to scale) with the corresponding unit cell parameters are shown at the center. Green atoms represent connection sites in the building blocks.

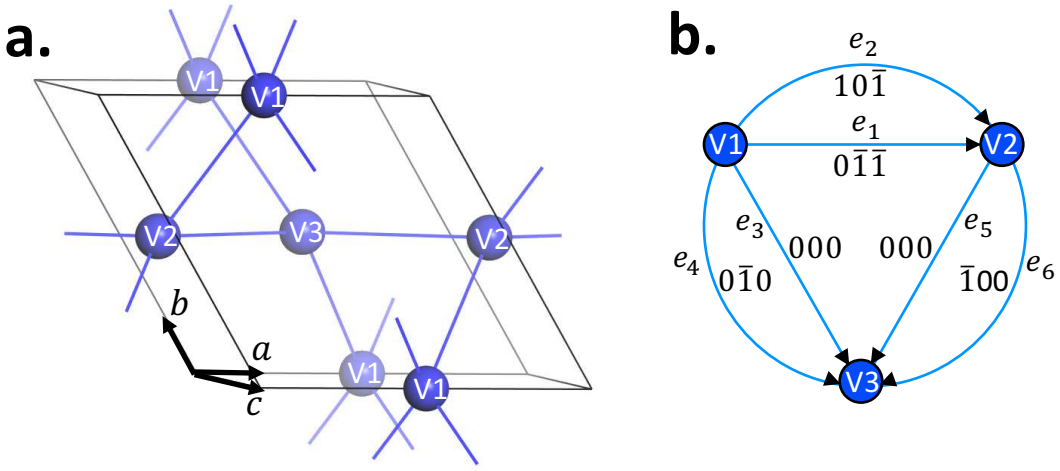

Figure A.4 (a) The **qtz** net unit cell, which has three types of vertices, in its barycentric embedding. (b) The voltage graph with edges labelled e_i. The voltage graph vertices are labelled by the unit cell vertices which are in the quotient groups they represent.

the voltage graph (Figure A.4b) represents a connection between a member of a quotient group and a member of another quotient group (or sometimes the same quotient group). For instance, for the **qtz** net, the V1 quotient group is the group constituted by the vertex V1 and all vertices translationally equivalent to V1.

Since all members of the V1 quotient group are connected to two equivalents of V2, there will be two edges (e_1 and e_2) connecting V1 and V2 in the voltage graph. To draw e_1 we say "V1 is connected to the vertex obtained by translating V2 by the vector (a, 0, -c)" and hence draw the edge e_1 pointing from V1 to V2, and then label the edge with the translation vector that relates V2 (in the unit cell) to the vertex that V1 is connected to. This translation vector is the "voltage" of the edge and it is conventionally represented by the vector coefficients with an overbar used to signify negatives. In this case, $(a, 0b, -c)$ is represented as $10\bar{1}$. Note that e_1 could have been drawn pointing from V2 to V1, in which case the translation vector labeling the edge would have opposite sign. The **qtz** voltage graph in Figure A.4b was obtained by following this procedure for all vertices. Note that a net is uniquely represented by its voltage graph, i.e. no other net has the same voltage graph [78]. However, a net can have multiple voltage graphs, as the edge directions and vertex labels are not fixed. Voltage graphs are useful in creating embeddings of nets in Euclidean space, so to understand this representation better, let us discuss a few concepts first.

A vector space consists of a group of vectors S and a field of scalars F such that the sum of any two vectors in S is also a vector in S and that the multiplication of any vector in S by any scalar in F is also a vector in S. The quintessential example of a vector space is $\mathbb{R}^3$, where vectors are 3-tuples of real numbers

and scalars are real numbers. However, vector spaces can be more abstract, because, for instance, the vector addition operation does not have to be defined as the "traditional" addition we are used to when adding vectors in $\mathbb{R}^3$. Rather, any operation between two vectors in S that produces another vector in S can be defined as "addition." Moreover, the field of scalars F need not be infinite. For instance, the field could be constituted by just 1, -1, and 0.

The basis of a vector space S is defined as any set of linearly independent vectors (excluding the zero vector), where every vector in S can be formed as a linear combination of this set. For example, the vectors $(1,0,0),(0,1,0),(0,0,1)$ form a well-known basis for $\mathbb{R}^3$. Finally, the dimension of the vector space is defined as number of vectors required to complete a basis. With these standard definitions in mind, we now proceed to discuss some of the algebraic properties of voltage graphs.

The edge space of a voltage graph is defined as the vector space where the field of scalars is the set $-1,0,1$. In this space, vectors are subsets of the edge set. For instance, for the **qtz** voltage graph, the edge set is $\{e_1,e_2,e_3,e_4,e_5,e_6\}$, and some subsets of the edge set are $\{e_1,e_2\},\{e_3,e_4\},\{e_5\},\{e_2,e_4,e_6\}$, and other such combinations. Accordingly, a convenient basis for any edge space (heretofore called the canonical basis) is the set of edges of the graph (represented as subsets with one element), meaning the edgespace of a graph has dimension N_e, where N_e is the number of edges. For the qtz voltage graph, this basis is$\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5\},\{e_6\}\}$. The "addition" operation between edge space vectors e_1 and e_2 is defined as $e_1 \triangle e_2$. Where the $\triangle$ symbol represents the "symmetric difference" operation, which returns all the elements that are in one set but not in the other. For instance, for the **qtz** graph, $\{e_1,e_3\}\triangle\{e_2,e_3\}$ is equal to $\{e_1,e_2\}$, which is also a member of the edge space.

The cycle space of a voltage graph is a subspace of the edge space whose vectors are cycles, where each cycle is represented as the subset of edges that comprise the cycle. A cycle is a path that traverses from vertex to vertex along edges, and starts and ends on the same vertex. For instance, for the **qtz** voltage graph a cycle could be $\{e_1,e_5,e_4\}$ and another cycle could be $\{e_1,e_2\}$. Note that, consistent with the definition of vector space, the "symmetric difference" operation between these two cycles produces another cycle: $\{e_2,e_5,e_4\}$. A basis of the cycle space is called the cycle basis, and has $N_e - N_v + 1$ elements (for a connected graph) [206, 363] where N_v is the number of vertices. For instance, a basis for the **qtz** voltage graph has four (6 - 3 + 1) elements. A basis could be $\{\{e_1,\ e_2\},\{e_3,e_4\},\{e_5,e_6\},\{e_2,e_4,e_6\}\}$.

The co-boundary of a subset of vertices of a graph is the set of edges connecting vertices in the subset to vertices not in the subset [206]. For instance, the co-boundary of the subset of vertices V1,V2 of the qtz voltage graph is the set $\{e_1,e_2,e_5,e_6\}$. The co-cycle space of a graph is a subspace of the edge-space whose vectors are co-boundaries [206]. The co-cycle space of a graph has dimension $N_v - 1$, and the set of co-boundaries of the first $N_v - 1$ individual vertices in a graph can be a basis for the co-cycle space.[1] For

instance, for the qtz voltage graph a basis for the co-cycle space could be $\{\{e_1, e_2, e_3, e_4\}, \{e_1, e_2, e_5, e_6\}\}$, which are the co-boundaries of V1 and V2. Notice that the sum of $N_e - N_v + 1$ (the dimension of the cycle space) and $N_v - 1$ (the dimension of the co-cycle space) is N_e (the dimension of the edge space). The cycle and co-cycles bases are both comprised of linearly independent vectors, furthermore, since the edges incident on a subset of vertices cannot form a cycle, cycles and co-boundaries are inherently linearly independent. Thus, since the size of the cycle basis combined with the co-cycle basis is also the dimension of the edge space, the cycle-co-cycle basis combination is an alternative basis of the edge space.

A crystallographic net can be embedded (mapped) into $\mathbb{R}^3$ by assigning a 3-component vector to each edge of the voltage graph. One only needs to define a matrix operator that maps the set of edges (i.e. the canonical basis of the edge space) to a set of 3-component vectors. Such a representation would already, in principle, allow the geometric deviation between the embedding and building blocks to be calculated, hence allowing an objective function of this deviation to be defined. However, an equivalent basis, that of the cycle basis combined with the co-cycle basis (from now on referred to as the cycle-co-cycle basis) of the voltage graph, can be mapped to a set of 3-vectors using so-called lattice and co-lattice vectors. This later mapping provides a more convenient definition for the problem of scaling a crystallographic net, as will subsequently be shown. A lattice vector is a translation vector that "transforms" points in the crystallographic unit cell to an equivalent point in a repeat unit. Accordingly, the sum of voltages (net voltage) in a cycle of a voltage graph must be a lattice vector. The net voltage must correspond to a lattice vector because a cycle in a voltage graph defines a path from a point in the net unit cell, to a translationally equivalent point in a repeat unit (recalling that each node in the voltage graph is a group of equivalent vertices). A co-lattice vector is similar to a lattice vector, except defined on a co-boundary and can be thought of as the sum of the 3-vectors corresponding to the edges in the co-boundary for an embedding in Euclidean space [206]. Keep in mind that lattice (co-lattice) vectors are the Euclidian space equivalents of cycles (co-boundaries), which means they can be added just like vectors in $\mathbb{R}^3$ can. Then, because the cycle-co-cycle basis forms a basis for the edge space, the corresponding lattice and co-lattice can be used to define a function that maps the cycle-co-cycle basis to the edge vectors of an embedding in Euclidean space. Using this mapping, a representation of a net in Cartesian coordinates can be constructed from the cycles and co-boundaries of the voltage graph and the corresponding lattice and co-lattice vectors.

If a crystallographic net can be embedded directly into Euclidean space by mapping its edges to 3-vectors, then why bother with cycles and co-boundaries? The reason is to exploit the fact that the co-lattice vector of all vertices in any barycentric embedding must be equal to the zero vector, because in barycentric embeddings all vertices are the center of mass of their neighboring vertices. Therefore, a non-zero co-lattice vector directly quantifies the magnitude and direction of a vertex deviation from its

original barycentric placement. Consequently, the co-lattice vectors of a net embedding can be used to quantify its total deviation from its barycentric embedding. Since we are concerned with finding lower symmetry embeddings that match building block geometries, the lattice vectors are extremely useful geometric measures, because they can be used as variables in an objective function that allows the fractional coordinates of vertices to deviate from their barycentric positions to better fit building block geometry. However, since a voltage-graph co-cycle space only forms an $N_v - 1$-dimensional subspace of the edge space, a cycle basis must be used together with co-cycle basis to form a complete edge space basis and, thus, a complete representation of the edge geometry of the net of interest. To clarify the above concepts, the entire process of constructing an objective function from the voltage graph representation of the qtz net is shown next, and the corresponding details in ToBaCCo explained.

In ToBaCCo, vectors of edges, such as cycles and co-boundaries, are represented as ordered N_e-tuples (recall that we have defined N_e as the number of edges in a graph). For example, if edges e_i, e_j, and e_k form a cycle in some voltage graph, the tuple representing that cycle would have a 1 at positions i, j, and k and 0 elsewhere. If any edge is traversed backwards in the cycle, a negative 1 is used at that edge position. The exact numbering (or ordering) of the edges is irrelevant as long as it is kept consistent for every cycle. Since the **qtz** unit cell has six edges, cycles and co-boundaries in the **qtz** voltage graph are represented as 6-tuples. We construct the cycle-co-cycle basis by first constructing the cycle and co-cycles bases separately and then combining them. In ToBaCCo, the cycle basis is constructed by first finding the spanning tree of the voltage graph of interest. The spanning tree of a graph G is the subgraph which connects all the nodes of G with the minimum possible number of edges and that contains no cycles [363]. An element of the cycle basis is then found by adding an edge not present in the spanning tree back into spanning tree to form a new subgraph, which conveniently has exactly one cycle (the new cycle-basis member). Repeating this process for each edge not present in the spanning tree yields a cycle basis [363]. For the **qtz** voltage graph, however, a cycle basis can be found easily by inspection. ToBaCCo gives

$$\{(1,-1,0,0,0,0),(0,0,-1,1,0,0),(1,0,0,-1,1,0),(1,0,0,-1,0,1)\}$$

for the edge ordering shown in Figure A.4. Note that there are four cycles in the basis, as we expect. Next, a co-cycle basis must be constructed. In ToBaCCo, a co-cycle space is built with the co-boundaries of the first $N_v - 1$ vertices (although the co-boundaries of any $N_v - 1$ vertices would work just as well). For qtz ToBaCCo gives the co-cycle basis

$$\{(1,1,1,1,0,0),(-1,-1,0,0,1,1)\}$$

with vitally, the same edge ordering as was used for constructing the cycle basis. Next, we calculate the lattice vectors corresponding to the cycles in the cycle basis (i.e. the cycle net voltage):

$$\{\bar{1}\bar{1}0,\ 0\bar{1}0,\ 00\bar{1},\ \bar{1}0\bar{1}\}.$$

Then, the co-lattice vectors are defined as variables (two 3-vectors, or six variables total):

$$\{x_1x_2x_3, x_4x_5x_6\}.$$

In the barycentric embedding, all the x_i are 0. Using the cycle-co-cycle basis a matrix B^* is built:

$$B^* = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 1 \end{pmatrix},$$

where the first $N_e - N_v + 1$ (4) rows are the cycle basis vectors and the next $N_v - 1$ (2) rows are the co-cycle basis vectors [206]. The inverse of B^* is then used to map the lattice and co-lattice vectors to a Euclidean representation of the voltage graph edges. The lattice and co-lattice vectors are encoded in the matrix α.

$$\alpha(x_1, x_2, x_3, x_4, x_5, x_6) = \begin{pmatrix} -1 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & -1 \\ x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \end{pmatrix},$$

where the first $N_e - N_v + 1$ (4) rows are the lattice vectors corresponding to the first $N_e - N_v + 1$ rows of B^* (the exact ordering does not matter so long as it is consistent with B^*), and the next $N_v - 1$ (3) rows are the co-lattice vectors corresponding to the last $N_v - 1$ rows of B^*. Put more succinctly, the matrix

$$\Omega = B^{*-1}\alpha(0,0,0,0,0,0) = \begin{pmatrix} -1/2 & -1/2 & -1/3 \\ 1/2 & 1/2 & -1/3 \\ 0 & 1/2 & 1/3 \\ 0 & -1/2 & 1/3 \\ 1/2 & 0 & -1/3 \\ -1/2 & 0 & -1/3 \end{pmatrix}$$

gives the vectors, in fractional coordinates, corresponding to edges of the barycentric embedding (all $x_i = 0$) of **qtz** [206]. The full reason why this is the correct mapping between the cycle-co-cycle basis and the edge vectors is beyond the scope of this contribution, but is described in detail by Eon [206]. Briefly, B^* maps the edge vectors in Ω to their lattice/co-lattice representation in α (being the change of basis matrix between the canonical edge space basis and the cycle-co-cylce basis), and, thus, B^{*-1} performs the inverse mapping.

The matrix Ω, then, specifies an embedding in fractional coordinates that is a function of the co-lattice vectors. However, we are also interested in how the template unit cell parameters can be used to adjust vertex geometry to fit building blocks. Therefore, we compare net geometry to building block geometry in

Cartesian coordinates, as conversion between fractional and Cartesian coordinates requires a matrix operator Z, that is built using all six lattice parameters. More precisely, we compare the pairwise dot-products (calculated in Cartesian coordinates) of the net edges to the connection site vector angles and lengths of the assigned building blocks, as suggested by Boyd and Woo [76]. Using dot products instead of coordinates negates the need for the many coordinate frame shifts and rotations needed to directly compare vertex and building block coordinates [76]. An $N_e \times N_e$ matrix of the pairwise edge dot-products in Cartesian coordinates (where the ij entry corresponds to the dot product between e_i and e_j) is formed from the product P=$\Omega Z \Omega^T$ where

$$Z = \begin{pmatrix} a^2 & (ab)\cos(\gamma) & (ac)\cos(\beta) \\ (ab)\cos(\gamma) & b^2 & (bc)\cos(\alpha) \\ (ac)\cos(\beta) & (bc)\cos(\alpha) & c^2 \end{pmatrix},$$

and $a, b, c, \alpha, \beta, \gamma$ are the template unit cell parameters [76, 206] (the product $\Omega\Omega^T$ would give the pairwise dot products in fractional coordinates). Thus, for

qtz:https://www.overleaf.com/project/602feabf3f77840ced2ce3fe

$$P\left(a,b,c,\alpha,\beta,\gamma,x_1,x_2,x_3,x_4,x_5,x_6\right) = B^{*-1}\alpha\left(x_1,x_2,x_3,x_4,x_5,x_6\right) Z \left(B^{*-1}\alpha\left(x_1,x_2,x_3,x_4,x_5,x_6\right)\right)^T$$

The entries in P can then be compared to the corresponding building block dot products. This requires first assigning an building block to each edge and vertex in the voltage graph, as described in Section S1.4. In ToBaCCo, this assignment is then translated into a direction and length for each edge which must be matched in the final scaled template. That is, if vertices V1 and V2 in the **qtz** net are assigned a nodular building block with distance d_1 between its center-of-mass and connection sites, and the edge between them is assigned a linker of length d_2, edges e_1 and e_2 must have length $d_1 + d_2 + 2d_3$, where d_3 is the desired bond length between nodular and connecting building blocks. In addition, the angle between e_1 and e_2 must match the angle between the node building block connection site vectors assigned to e_1 and e_2, respectively. Connection site vectors are assigned to edges by translating and rotating normalized node building blocks into the position of their assigned vertex before any scaling occurs, each connection site vector is then assigned the edge it is nearest after this preliminary placement. Assigning edges in this way ensures that the building block and vertex geometries are aligned as closely as possible before scaling begins. After the matrix P is constructed and the all edges assigned their desired lengths and directions, the final objective function can be defined.

For each nodular building block we have the terms

$$D_{ij} = L_i L_j \cos\theta_{ij}$$

where L_i is the required distance between the assigned vertex and the vertex connected to it via e_i, L_j is the analogous distance for e_j, and θ_{ij} is the required angle between e_i and e_j, all as enforced by the linker/node building block geometry. These terms allow the objective function to be defined (for **qtz**) as

$$O\left(a,b,c,\alpha,\beta,\gamma,x_1,x_2,x_3,x_4,x_5,x_6\right) = \sum_{i=1}^{N_e}\sum_{j=1}^{N_e}\left(D_{ij}-P_{ij}\right)^2 ,$$

where P_{ij} is the ijth entry in P . In general, O is a function of the unit cell parameters and $3(N_v - 1)$ co-lattice vector variables (rather than the six shown above). In ToBaCCo, this objective function is minimized using a bounded L-BFGS-B algorithm [207, 208]. The variable starting values are taken to be those of the barycentric embedding.

The scaling method presented here and introduced to ToBaCCo follows the spirit of the method presented by Boyd and Woo.[2] However, these authors "augment" their voltage graphs before computing the pairwise edge dot-products by adding several "artificial" 2-connected vertices between the actual vertices. Since each new vertex increases the number of co-lattice vectors by one (see ref. [76] for more details), this has the effect of adding more degrees of freedom in α and thus to the objective function. Their rationale behind augmentation is that the extra vertices can help better accommodate building block geometry, especially in cases where the initial net vertex geometry is markedly different with respect to the building block geometry.

We briefly considered the above approach, by adding five extra 2-connected vertices between all voltage graph vertices. However, we found that this augmentation procedure often had the effect of creating undesirable angles between nodular and connecting building blocks while leaving the template unit cell parameters essentially unchanged during scaling. The reason this happened is that, with the extra vertices, there are multiples ways to minimize the objective function. As it turns out, the rescaling algorithm often prefers to minimize the "deviation" between a vertex and the corresponding building block geometry by unphysically "bending" the edges around vertex in question, instead of by altering the template unit cell parameters. Although unphysical bending at connection sites could potentially be removed during geometry optimization, in our experience, the relaxation process can be significantly influenced by starting geometries which have large local deviations from equilibrium. Therefore, in our modification of ToBaCCo, we decided that nets should not be augmented (i.e. should not have extra vertices added) as to instead force the unit cell parameters to expand and contract more to suit the building block geometry. Once the template is scaled, the building blocks are translated into their fractional positions in the scaled unit cell and rotated to maximally overlap with their assigned edges. These transformations are accomplished using a script directly from the Biopython project [364], which finds the translation vector and rotation matrices which minimize the root-mean-squared-deviation (RMSD) between two arrays of vectors.

A.2 Calculation of MOF Crystal Mechanical Properties

As stated in the main text Hooke's law in tensorial form is given by Equation 3.1, reproduced below for convenience:

$$\sigma_{ij} = \sum_{kl} c_{ijkl}\epsilon_{kl}$$

where σ_{ij} and ϵ_{kl} are components of the stress and strain tensor, respectively.

The indices i, j, k, l range between three directions (see Figure A.5). In general, the stress tensor can be written as a 3×3 symmetric matrix:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix};$$

meaning there are six unique elements. The strain tensor is written in analogously:

$$\begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix},$$

and is again symmetric (i.e. it has six unique elements). In Voigt notation (where indices are mapped thus: $11 \to 1$, $22 \to 2$, $33 \to 3$, $23 \to 4$, $13 \to 5$, and $12 \to 6$), the stress and strain tensors are written as 6-vectors:

$$\begin{pmatrix} \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 & \sigma_6 \end{pmatrix}^T \quad , \quad \begin{pmatrix} \epsilon_1 & \epsilon_2 & \epsilon_3 & \epsilon_4 & \epsilon_5 & \epsilon_6 \end{pmatrix}^T$$

and the elasticity tensor (C) can, thus, be written as a 6×6 matrix:

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56} \\ c_{61} & c_{62} & c_{63} & c_{64} & c_{65} & c_{66} \end{pmatrix}$$

with each entry calculated via Equation 3.2 in the main text, reproduced below for convenience

$$c_{ij} = \frac{\partial \sigma_i}{\partial \epsilon_j} = \frac{1}{V}\frac{\partial^2 E}{\partial \epsilon_i \partial \epsilon_j}$$

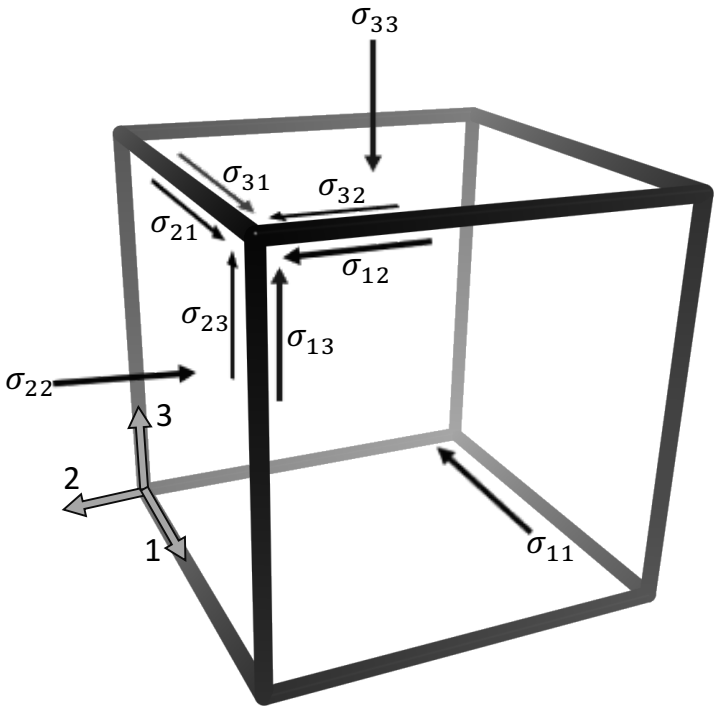

Figure A.5 Components of the stress tensor in each direction.

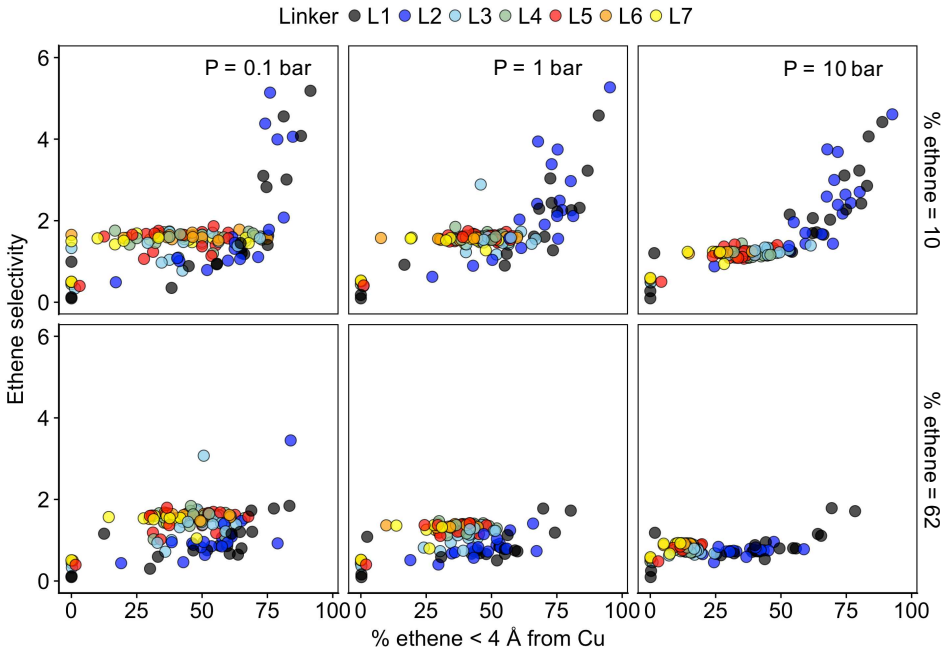

Figure A.6 Plots of ethene selectivity versus the percentage of ethene (out of the total adsorbed ethene) less than 4 Å from a Cu OMS for each combination of pressure and mixture composition considered.

APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER 4

B.1 Machine Learning Model Training Details

The purpose of training a machine learning algorithm is to improve the accuracy with which the output values Y_i are predicted based on the corresponding values X_{ij} associated with them. However, for the algorithm to be useful outside the training set, it is critical to prevent overfitting. A situation that it is typically reflected in achieving small prediction errors for observations in the training set, but then encountering large prediction errors for observations outside the training set. In general, to prevent overfitting we applied the procedure described in Figure B.1.

Each model was constructed on a unique training set and validated on the correspondingly unique test set. Training sets consisted of 80% of the data points randomly selected from within each class, i.e. class ratios were maintained in the training set. For the classification models, classes to select from were already defined: 0 (worsening) and 1 (improvement) for the respective CO_2 capture metric. For regression models five classes were defined for each CO_2 capture metric corresponding to the 0-20, 20-40, 40-60, 60-80, 80-100 percentile ranges. Such selection was used to ensure each model was trained on a set representing arbitrary data as closely as possible. The training of the machine learning algorithms was done exclusively in R, using publicly available packages. Important details of the machine learning algorithms and how the training was performed is discussed below.

Multiple Linear Regression (MLR). We used the built-in lm package in the basic R distribution to fit all MLR models.

Conditional inference decision trees (DTs). We used recursive partitioning, which is a common method to construct binary decision trees. The method works by splitting the feature (descriptor) space into hypercube areas, then assigning each area to a value of the response function (e.g. 0 or 1 for the case of a binary response function). The assignment groups together observations with identical or similar responses, hence attempting to minimize the "impurity" of each partition. We used a "greedy" algorithm. Thus, for the root node (first split), the algorithm chooses the variable that results in the lowest impurity to perform the split and the data divided into two partitions accordingly. The two new partitions are then split again with a similar procedure. The recursive process is allowed to continue until a stopping criterion is met. Typical stopping criteria include: (*i*) the leaf node is pure, (*ii*) the lowest acceptable of number of observations in the leaf node has been reached, and (*iii*) and subsequent split would not change the impurity beyond an acceptable minimum [172].

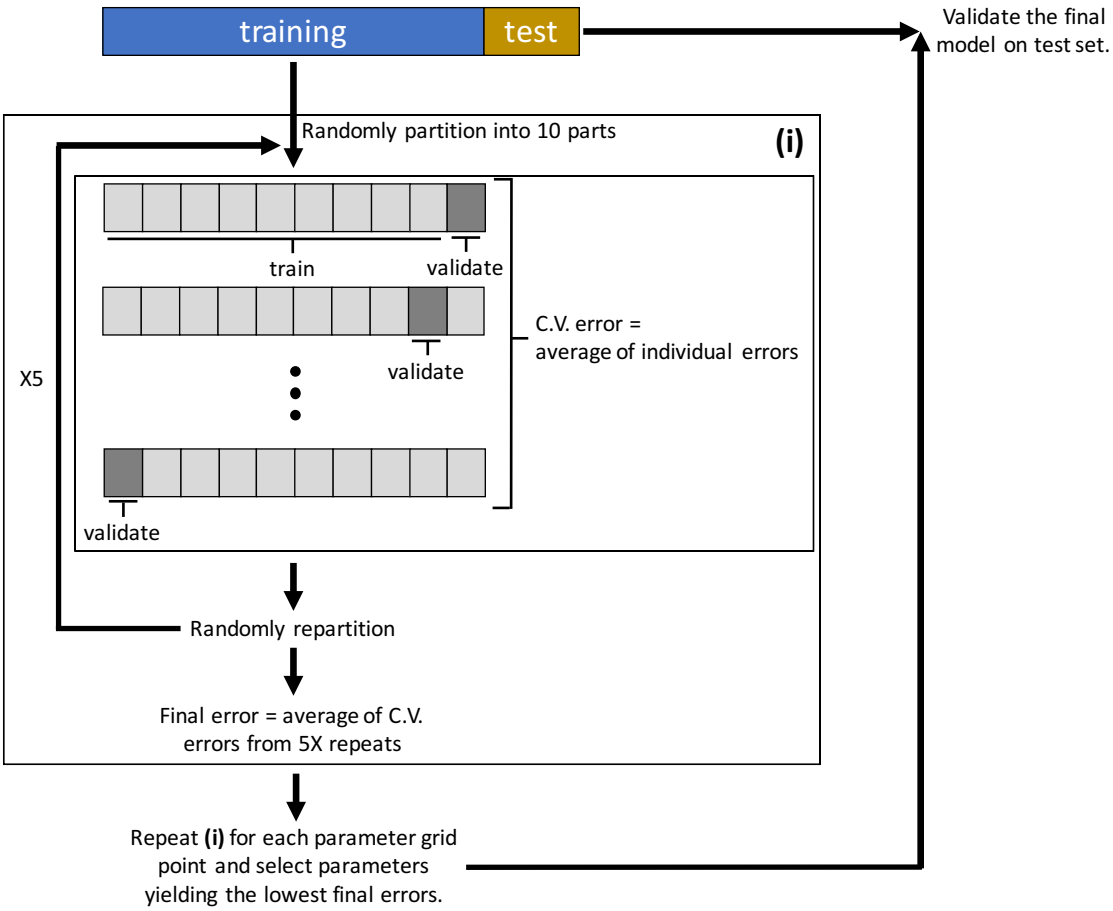

Figure B.1 Schematics for the process we employed for training and validating the classification and regression models. Parameter grids were used for the NN, RF, GBM, DT, and SVM models, but not for the MLR models (which do not have model parameters). We used 5-times-repeated 10-fold cross validation. C.V. = cross-validation.

We used the partykit package to construct all decision trees. This package employs the CTree algorithm of Hothorn et al. [365] The CTree algorithm addresses the issues of overfitting and selection bias towards variables with many possible splits, often associated with traditional decision trees, by employing statistically motivated splitting and stopping criteria. Recursive partitioning was stopped when the null hypothesis that the response variable is independent from the explanatory variables could not be rejected [365]. Our conditional inference trees were tuned on the parameter $1 - p$, where p is the p-value used to "reject" or "not reject" the null hypothesis. To avoid overly large trees only $1 - p$ values greater than 0.5 were considered.

Conditional inference tree random forests (RFs). We also used the partykit package to build all RF models. Each tree in a random forest was built using a randomly selected subset of the training data, without replacement (i.e. members of the current subset were not used again in subsequent subsets), a method referred to as sub-setting or sample splitting.[1] To further increase diversity among the trees in the forest, each split in each tree is tested on a randomly selected subset of the variables [172]. The number of variables in this random selection is a parameter of the algorithm, which we tuned for each RF model using a grid search. The model RMSE was evaluated using 5-times-repeated 10-fold cross-validation at each point, and the model parameters producing the lowest RMSE were used in the final model. The number of trees (ntree) grown in the forest can also be tuned. Since, in all our models, little variation in RMSE for ntree greater than 500 was observed, we maintained ntree constant at 500.

Support vector machines (SVMs). We used the kernlab package for SVM regression. During SVM regression, we first mapped the input vectors to a feature/descriptor space of higher dimension (hyperspace) using a kernel function, Φ. The algorithm looks to define the hypersurface that best approximates the response (output) variables for the given input vectors by minimizing a cost function of the error. Input vectors closer than a threshold ε to the hypersurface do not count in the cost function [366]. For the kernel function Φ, we chose a radial basis function kernel, which is one a number of typical kernel functions subject to certain constraints as described by Smola and Schölkopf [366].

We used regularization to avoid overfitting the training data. The regularization parameter is a parameter tuned to allow larger errors with less bias, tuning the trade-off between a complex function with little error on the training set and a less complex function that may have less bias. Typically, different regularization parameters are tested, with the value giving the best results selected through a validation procedure [366]. Here we tuned the regularization parameter with a grid search, and the value resulting in the lowest RMSE computed from 5-times-repeated 10-fold cross-validation was selected.

Artificial neural networks (NNs). We used the RSNNS package to build all NN models. The output of a NN is the approximate function value for a particular set of inputs, essentially corresponding to

a sum of weighting factors multiplied by input values. The original variable values are often transformed by an activation function, which outputs a function of the weighted sum of variables to ensure a small change in a variable will result in a small change in the predicted value [367]. As activation function, we used the logistic function (or sigmoid function), which is a common choice. We used weight-decayed error backpropagation to minimize the cost function. Backpropagation is a method for computing the gradient of the cost function with respect to the weights. It proceeds "back to front," where the gradient is computed first in the last layer of nodes, then on the second to last, and so forth. The computed gradient is used in a gradient descent algorithm to adjust the weighting factors to minimize the cost function. In weight-decayed backpropagation, each weighting factor is additionally reduced by a portion d of its old value (we used $d = 1 \times 10^{-3}$). This process drives small weights, and weights not increased by back propagation to zero, thus reducing the complexity of the network an avoiding overfitting [368]. We tuned our NN models on a parameter grid in which the number of nodes in a "3 hidden-layer" perceptron were varied. We tested nine through 14 nodes for the first hidden layer, and 0 through 14 nodes for the second and third hidden layers. The structure with the lowest RMSE computed via 5-times-repeated 10-fold cross validation was used.

Gradient boosted machines (GBMs). The package gbm was used to build all GBM models. We employed the gradient descent algorithm described by Friedman [278] and implemented in the aforementioned package, which we briefly describe here [369, 370]. First, the algorithm initializes $\hat{F}$ as a constant (for i in 1, ..., T). Where $\hat{F}$ is the regression function estimate (i.e. model). Second, it defines $z_i = \partial\Psi(y_i, F(x_i))/\partial, F(x_i)$ s.t. $F(x_i) = \hat{F}(x_i)$, where Ψ is the cost function, y_i is the value being fit in the training set, x_i are the observations being used to predict y_i, and $F(x_i)$ is the regression function, set equal to the regression function estimate $\hat{F}(x_i)$. Third, it randomly selects some portion P of the training set observations to form a set S and fits a regression tree $f(x|x_i \in S)$ with K terminal nodes predicting z_i from the variables x_i. Here x indicates an arbitrary set of features. Fourth, it computes the gradient descent step size ρ_k as

$$\rho_k = \text{argmin}_\rho \sum_{x_i \in S_k} \Psi(y_i, F(x_i) + \rho)$$

where S_k is the set of all x_i in terminal node k. Fifth, it updates the estimate $\hat{F}(x)$ to

$$\hat{F}(x) \leftarrow \hat{F}(x) + \lambda\rho_{k(x)}$$

where $k(x)$ indicates the terminal node x would be assigned to. T, P, K, and λ are all parameters of the algorithm; T is the number of trees built, P is the proportion of the data used to build the tree at each step, K is the number of terminal nodes in the trees, and λ is the learning rate. We tuned these parameters on a grid with λ and P constant at 0.05 and 0.5 respectively. The method proposed by

Friedman was used to compute relative variable importance from the GBM models [278].

ICE plots. We used the package ICEbox to create the individual conditional inference (ICE) plots shown later in this appendix. ICEbox follows the ICE plot implementation of Goldstein et al. [371]. Individual conditional inferences (ICE) plots show the model response to each observation of a descriptor, as well as the average response over all observations [371]. Thus, ICE plots provide some indication of variable importance. Descriptors with a large partial effect (i.e. significant changes over their range) from individual variables or on average are more important to the model than otherwise. For example, in Figure B.25 (the ICE plot for CO_2 selectivity over H_2) the model response varies significantly over the range of GSA, which comparing this to Figure 4.9 in the main text, is the most important descriptor for CO_2 selectivity over H_2. Note that variable importance in Figure 4.9 was not computed from the ICE plots, but rather from the method described by Friedman [278], however the two methods are complimentary. Note that in all ICE plots shown the model response is a scaled variable (i.e. unitless).

Genetic algorithm optimization. The differential evolution optimization algorithm implemented in R by Ardia et al. [372] was used to maximize our GBM model predictions subject to certain constraints:

1. framework MPC $\geq$ framework MNC
2. All charge absolute values < 3
3. LPD $\geq$ PLD
4. $2.5 \leq$ LPD ≤ 50
5. $1.5 \leq$ PLD ≤ 50
6. $0 <$ VF < 1
7. $0 <$ GSA < 6000
8. $0 <$ density < 3000
9. $0 <$ number density FG < 0.05
10. $0 <$ SE of FG < 200
11. Topology remains unchanged

Topology was held constant to ensure physical solutions (as it is not a continuous variable). We averaged the results of five maximizations for each metric, as given the complexity of our objective function, we observed some variability between the obtained maxima and the corresponding variable

patterns between runs. 1000 algorithm iterations were used for each optimization, with default cross-over probability and differential weighting factor of 0.5 and 0.8 respectively.

B.2 Supplementary Figures

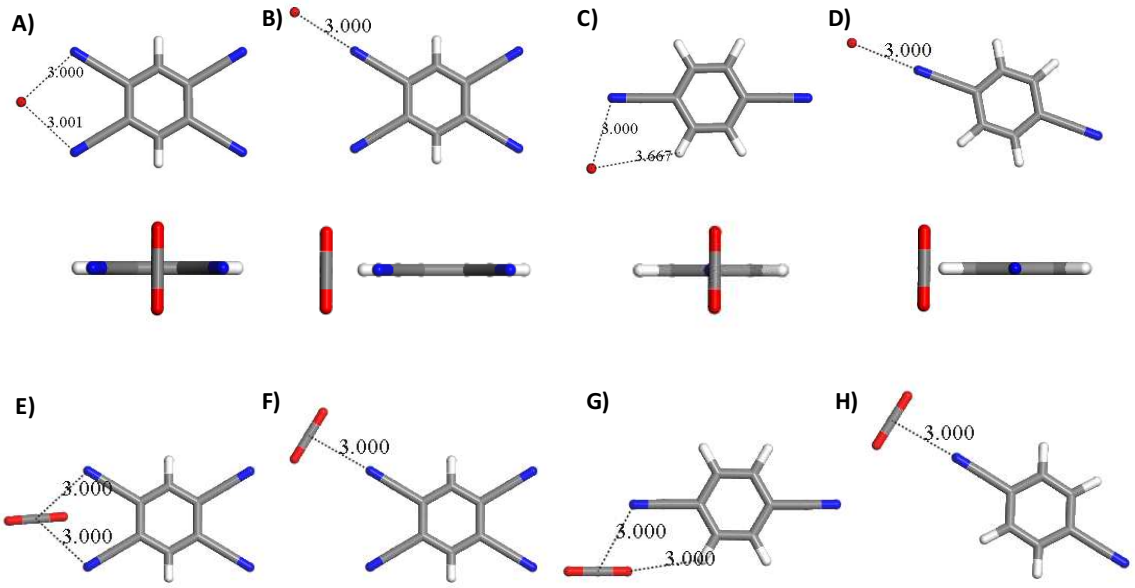

Figure B.2 Examples of initial configurations attempted when exploring the interaction of adsorbates with functional groups in MOF building blocks.

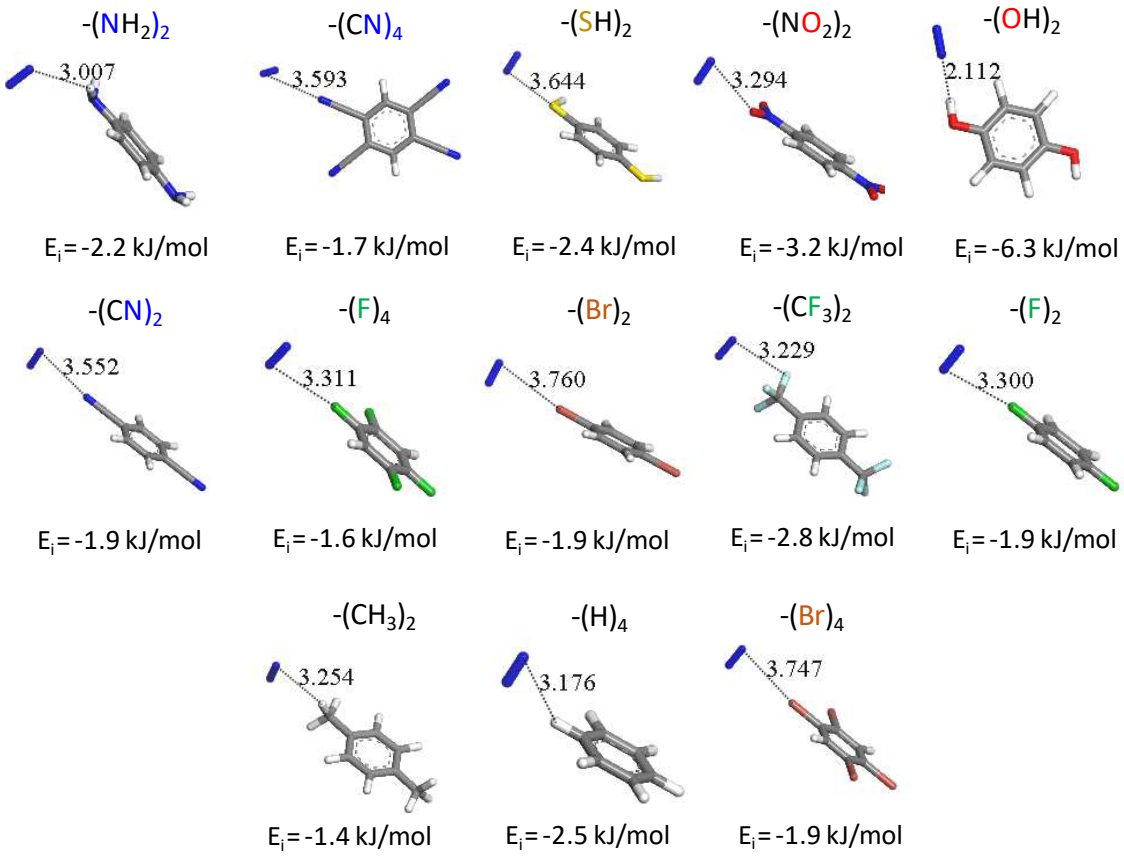

Figure B.3 DFT-calculated optimal binding configuration between N_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

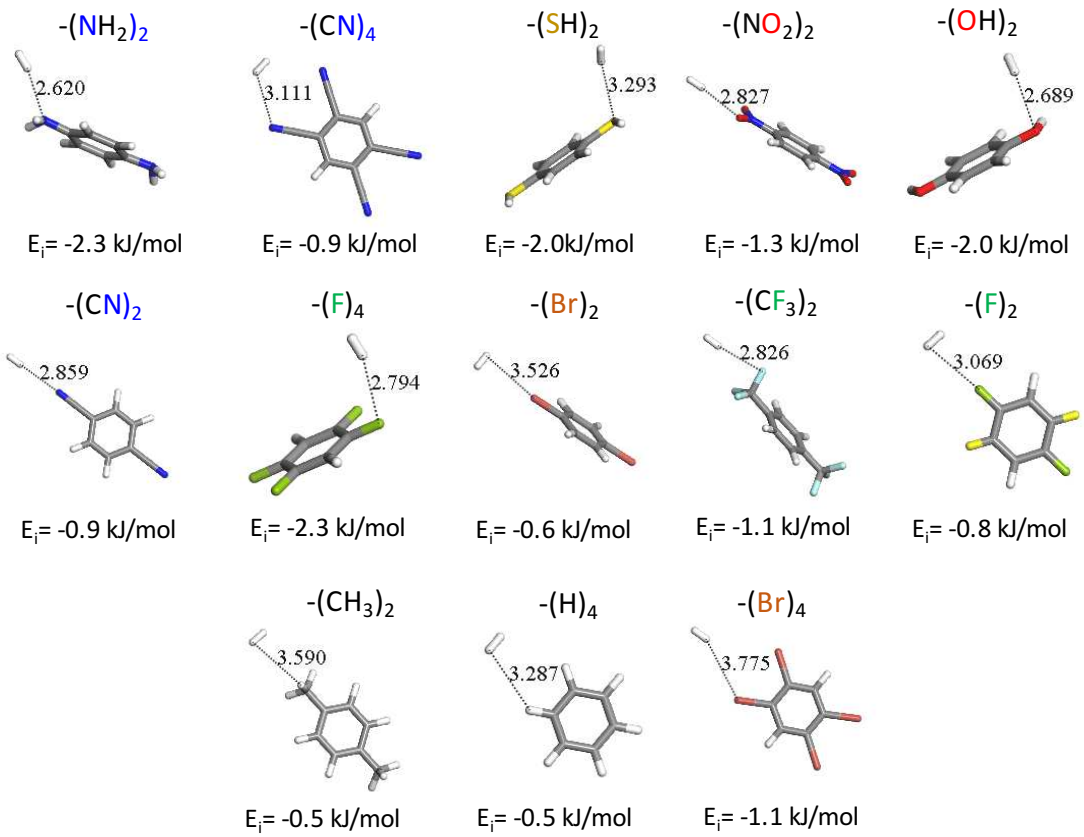

Figure B.4 DFT-calculated optimal binding configuration between H_2 molecules and the functional groups of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

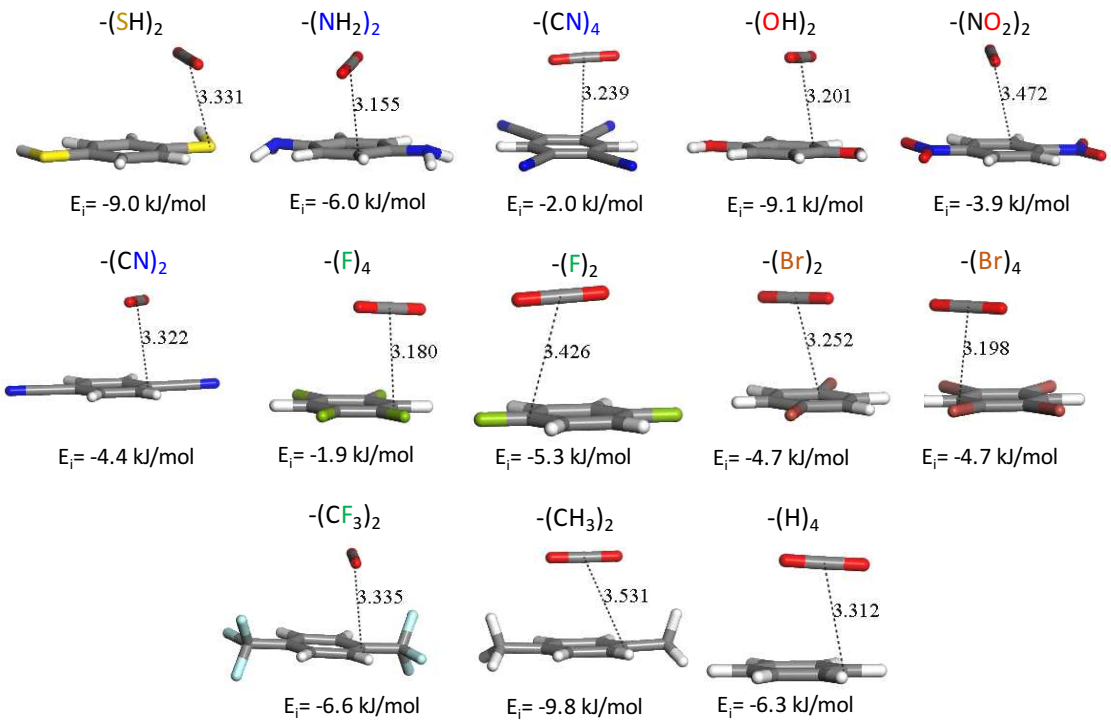

Figure B.5 DFT-calculated binding configuration between CO_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

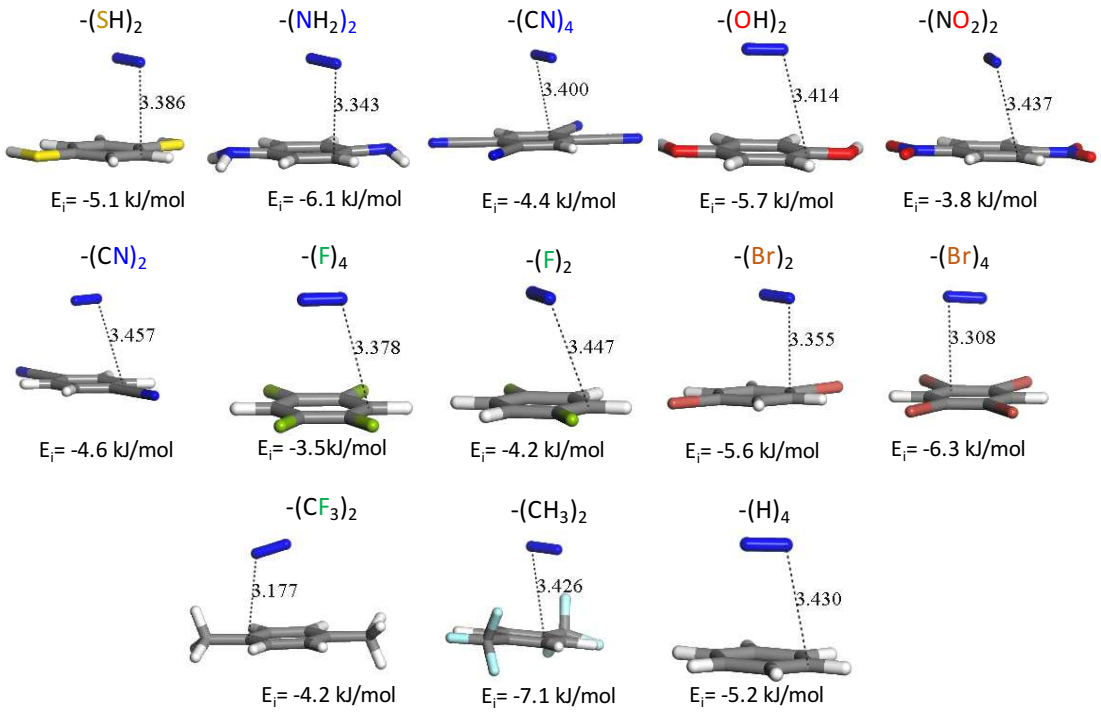

Figure B.6 DFT-calculated binding configuration between N_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

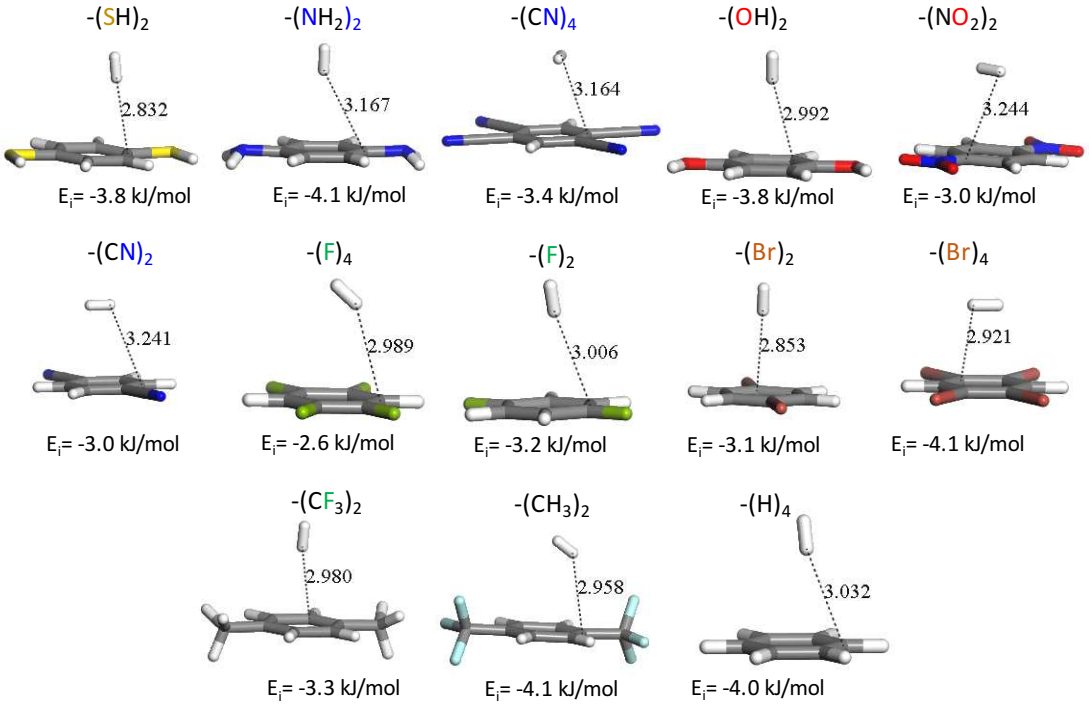

Figure B.7 DFT-calculated binding configuration between H_2 molecules and the phenyl rings of MOF connecting building blocks. Corresponding DFT-calculated binding energies are listed below each configuration. Interaction distances (Å) are indicated with dotted lines.

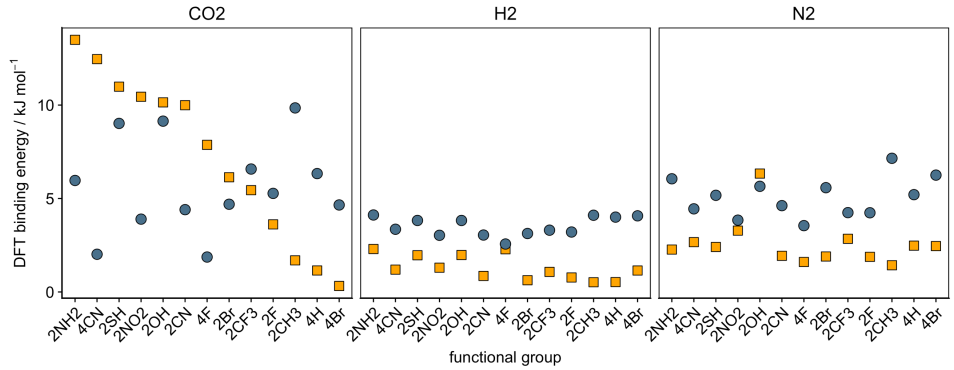

Figure B.8 DFT-calculated binding energies for the interaction of the adsorbates (CO_2, N_2, and H_2) with the functional groups (squares), and with the phenyl ring (circles).

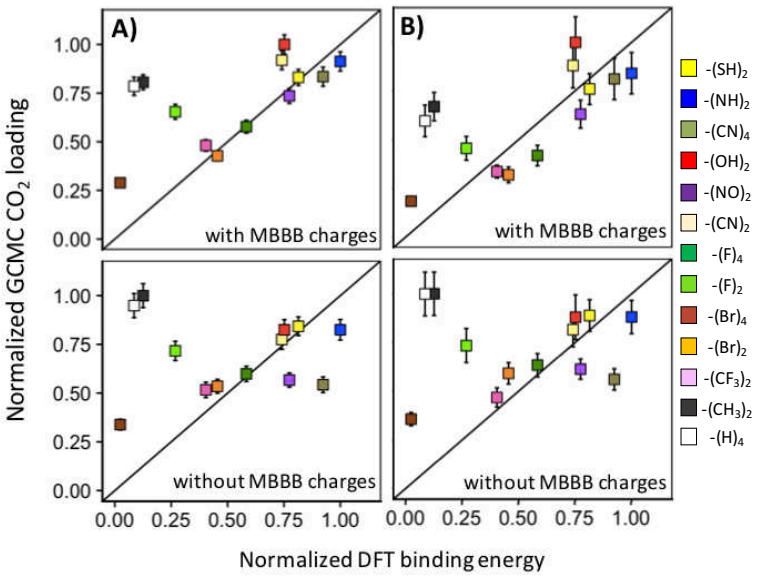

Figure B.9 Parity plots for GCMC-simulated CO_2 loadings and DFT-calculated binding energies.(a) CO_2 loading in a 20:80 CO_2:H_2 mixture at 20 bar, with (top) and without (bottom) MBBB charges and (b) CO_2 loading in a 15:85 CO_2:N_2 mixture at 1 bar, with (top) and without (bottom) MBBB charges. Error bars correspond to the standard error for the averaged CO_2 loadings.

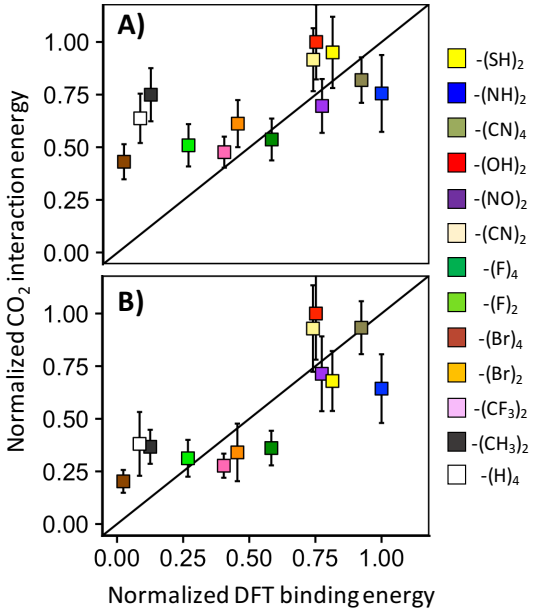

Figure B.10 Parity plots for the GCMC-calculated CO_2 – framework interactions during simulated CO_2 at 1 bar and DFT-calculated CO_2 binding energies for A) dispersion interactions (R2 = 0.411, SRCC = 0.681) and B) electrostatic interactions (R2 = 0.623, SRCC = 0.710).

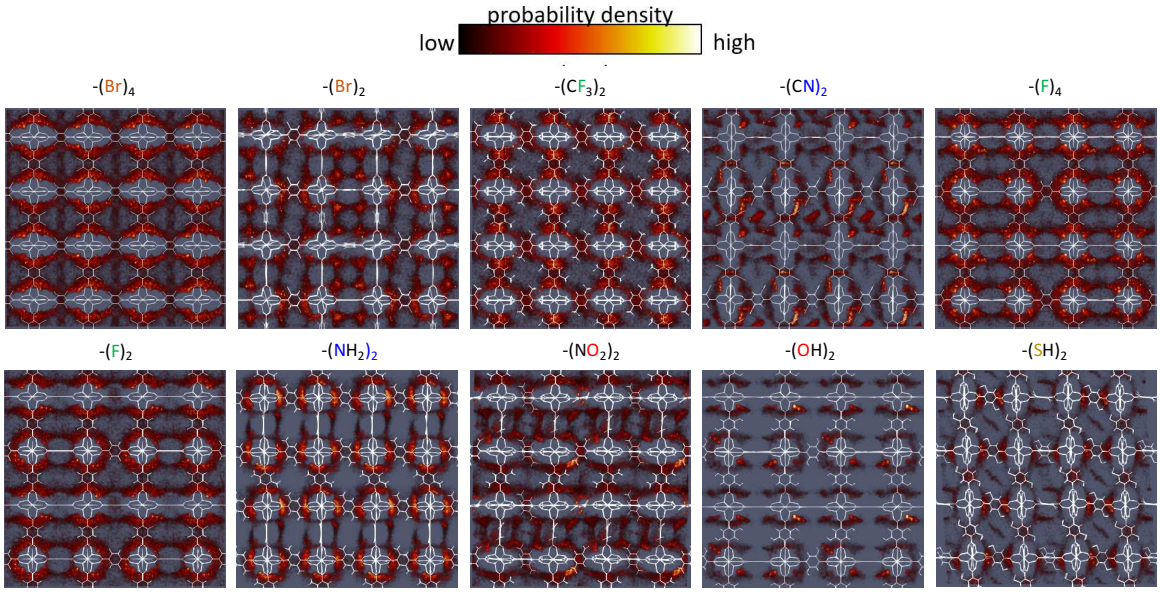

Figure B.11 CO_2 probability density maps showing the average CO_2 loading distribution in ten analogous structures for the remaining functionalizations not included in Figure 4.4.

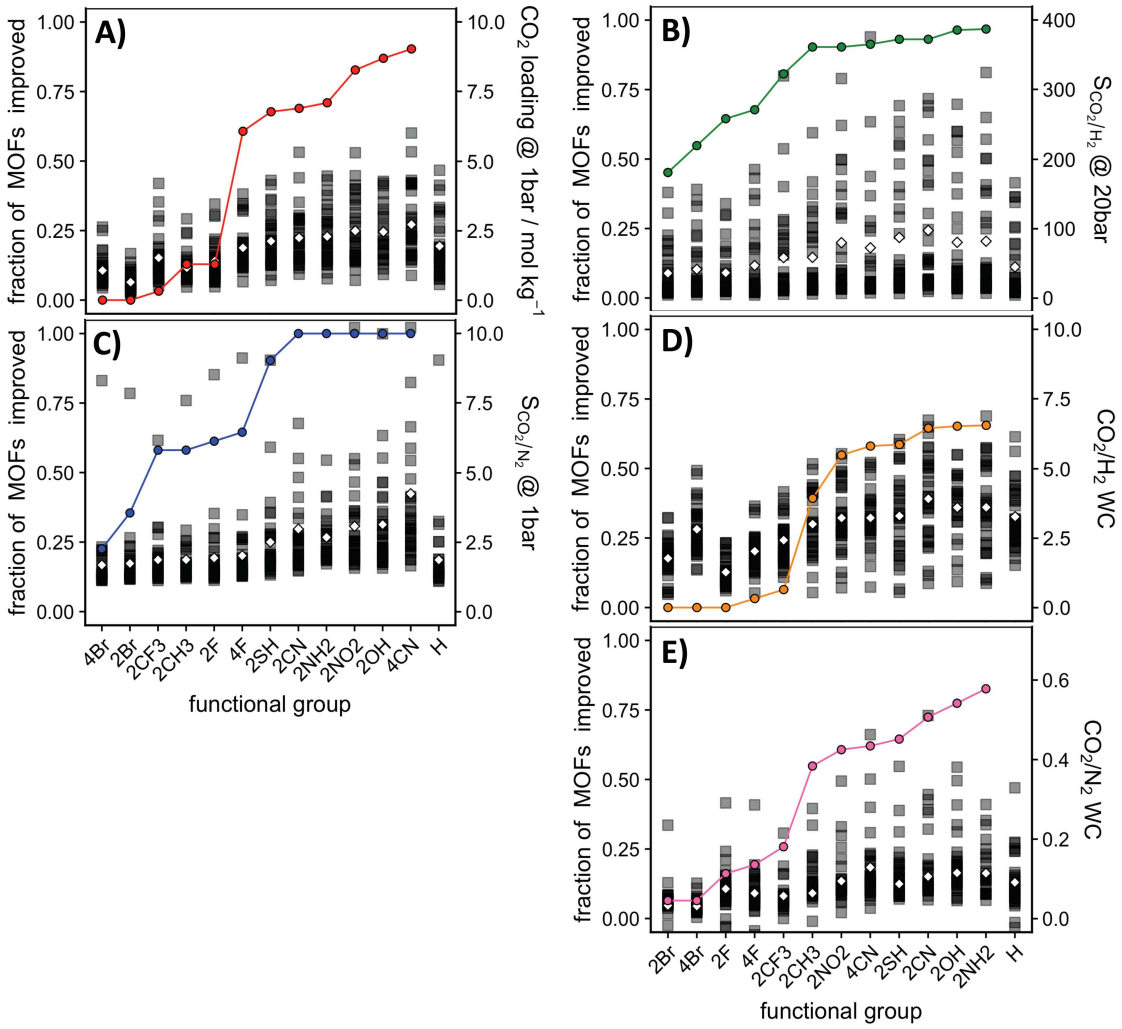

Figure B.12 Circles denote fraction of MOFs improved by each functional group (left axis), black squares are the values of each MOF for the variable on the right axis, the white diamonds are the averages of these values for (a) CO_2 loading at 1 bar, (b) CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar, (c) CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar, (d) CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar, (e) CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar.

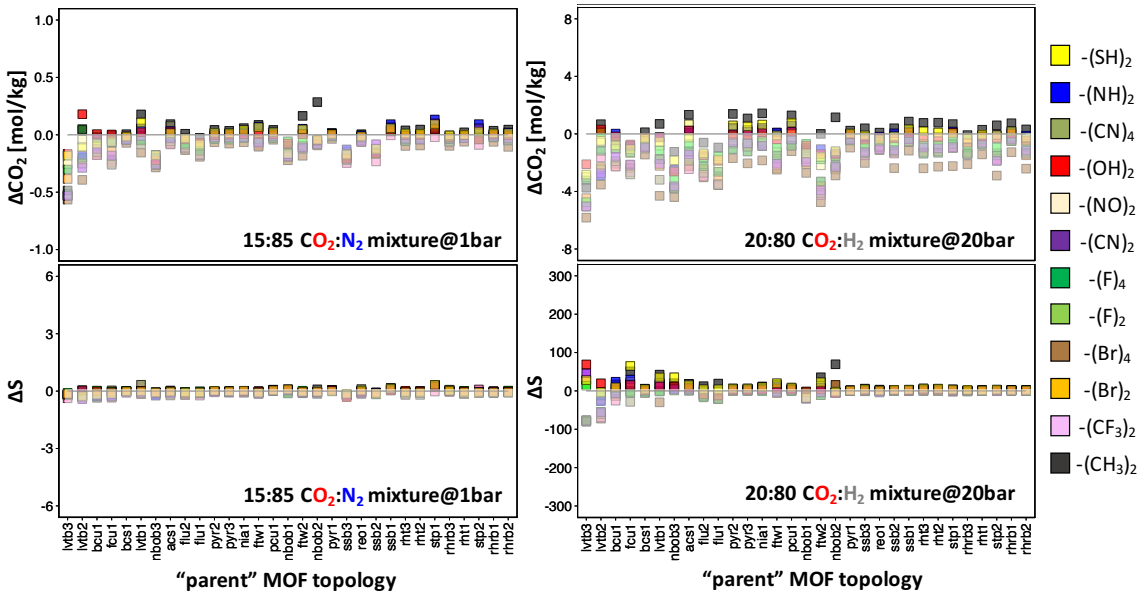

Figure B.13 Effect of MOF functionalization (pore chemistry) on the CO_2 loading and CO_2 selectivity of "parent" MOFs of different topologies calculated with "zero" MOF atom charges. Results are presented for adsorption of a 15:85 CO_2:N_2 mixture at 1 bar (left, post-combustion carbon capture) and a 20:80 CO_2/H_2 mixture at 20 bar (right, pre-combustion carbon capture). As a reference, compare this with Figure 4.5 in the main text.

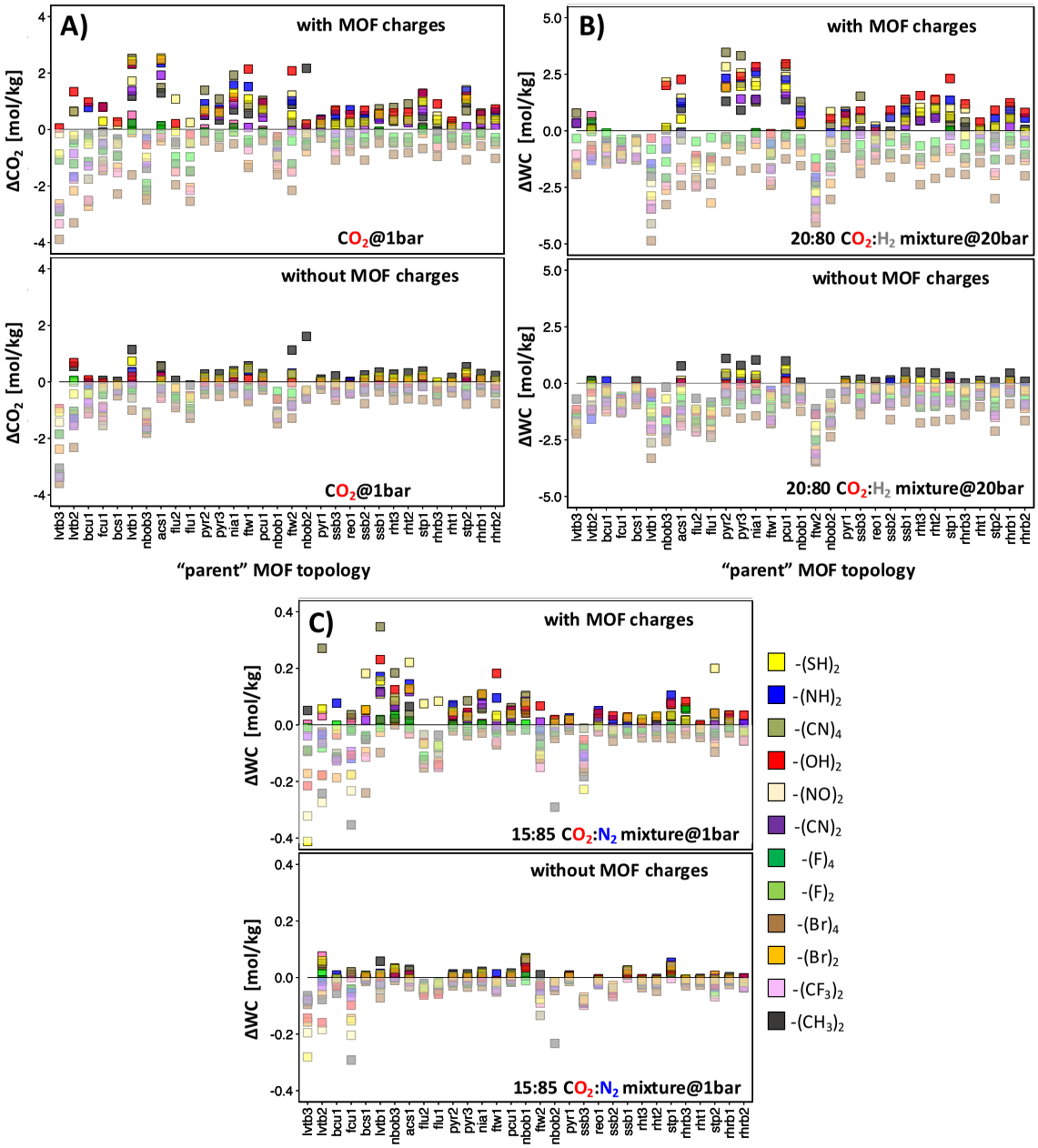

Figure B.14 Effect of MOF functionalization (pore chemistry) on the CO_2 capture metrics of "parent" MOFs of different topologies. (a) CO_2 loading at 1 bar (b) CO_2 working capacity for adsorption of CO_2:H_2 mixture. (c) CO_2 working capacity for adsorption of CO_2:N_2 mixture.

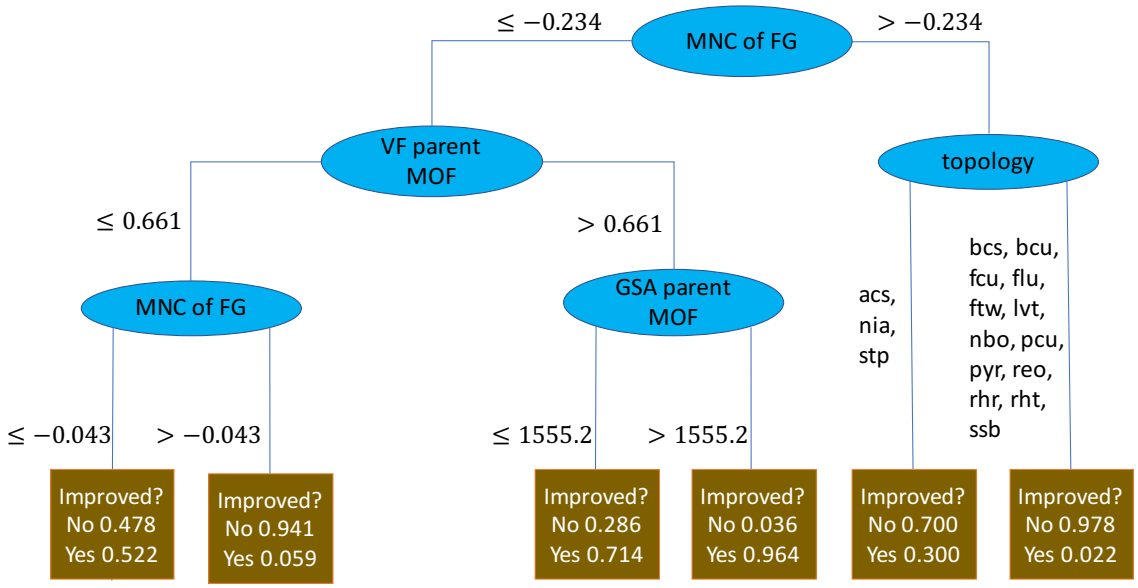

Figure B.15 Decision tree trained to predict whether parent MOF functionalization would improve the CO_2 loading of the parent MOF during exposure to pure CO_2 at 1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) in the functional group; topology, void fraction (VF), and gravimetric surface area (GSA) of the parent MOF.

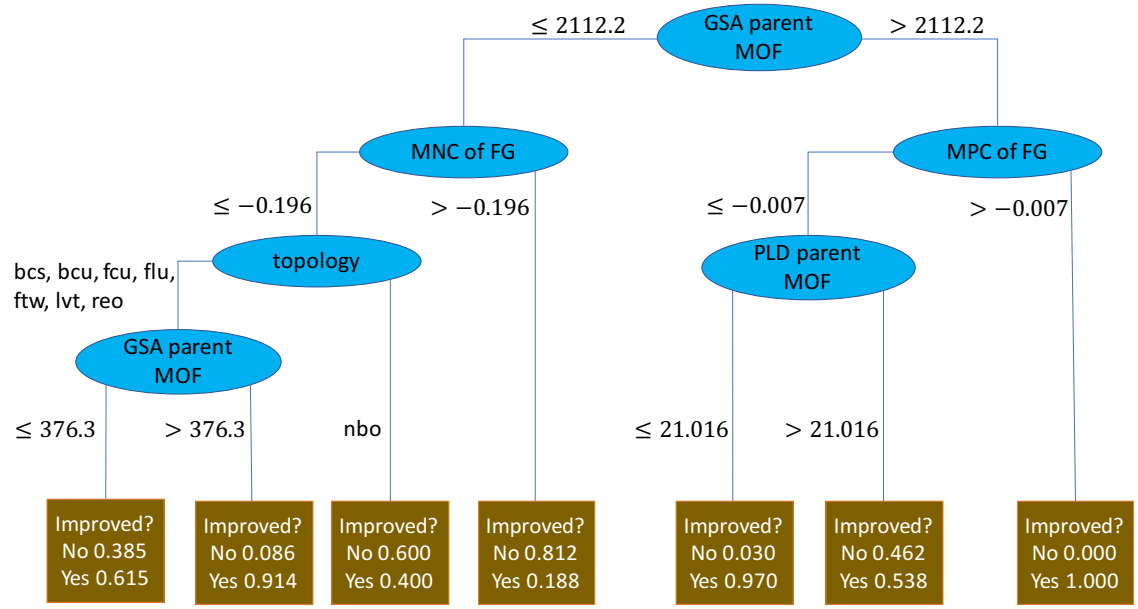

Figure B.16 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 selectivity over H_2 of the parent MOF during exposure to a 20:80 CO_2:H_2 mixture at 20 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) and most positive charge (MPC) in the functional group; topology, largest pore diameter (LPD), and gravimetric surface area (GSA) of the parent MOF.

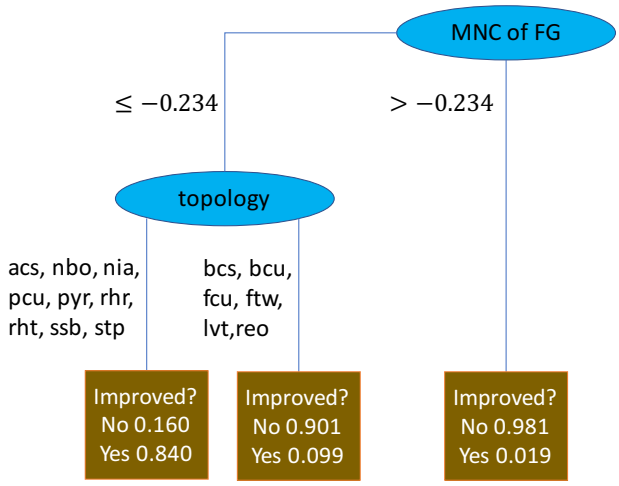

Figure B.17 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 working capacity the parent MOF for adsorption conditions corresponding to a 20:80 CO_2:H_2 mixture at 20 bar and desorption conditions corresponding to pure CO_2 at 1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) in the functional group and the topology of the parent MOF.

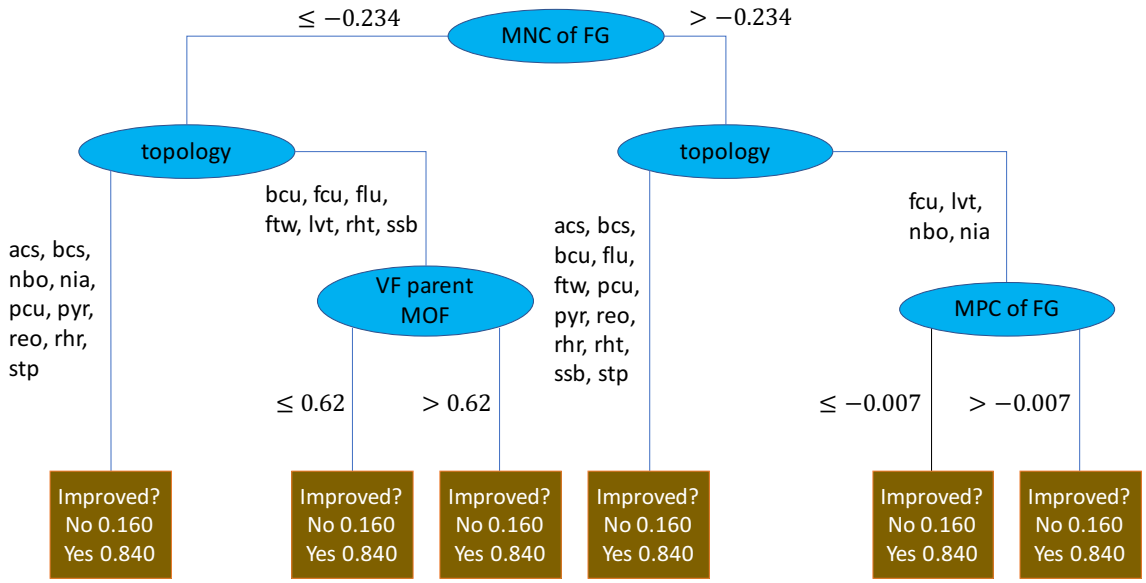

Figure B.18 Decision tree trained to predict whether parent MOF functionalization would improve the MOF CO_2 working capacity of the parent MOF for adsorption conditions corresponding to a 15:85 CO_2:N_2 mixture at 1 bar and desorption conditions corresponding to a 90:10 CO_2:N_2 mixture at 0.1 bar. "Yes" indicates improvement, "No" indicates no improvement. The numbers next to the predictions indicate the probability the categorization is correct. Variables appearing in this tree are the most negative charge (MNC) and most positive charge (MPC) in the functional group; the topology and void fraction (VF) of the parent MOF.

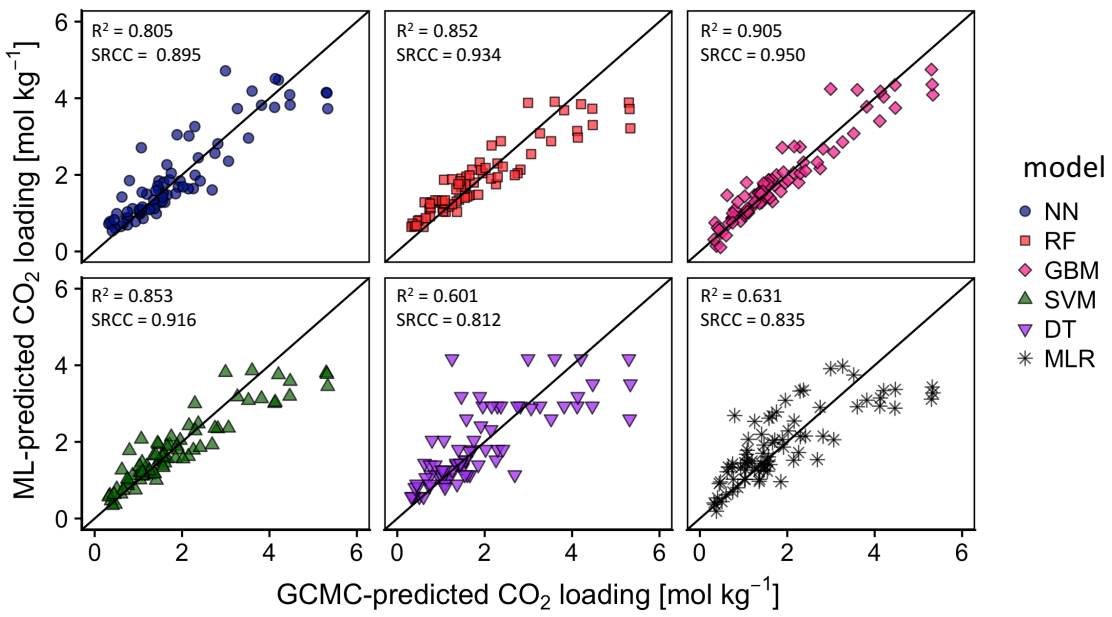

Figure B.19 Parity plots comparing the predictions on the test set of CO_2 loading at 1 bar by each machine learning algorithm to those obtained by GCMC simulations. GBM emerged as the most accurate algorithm.

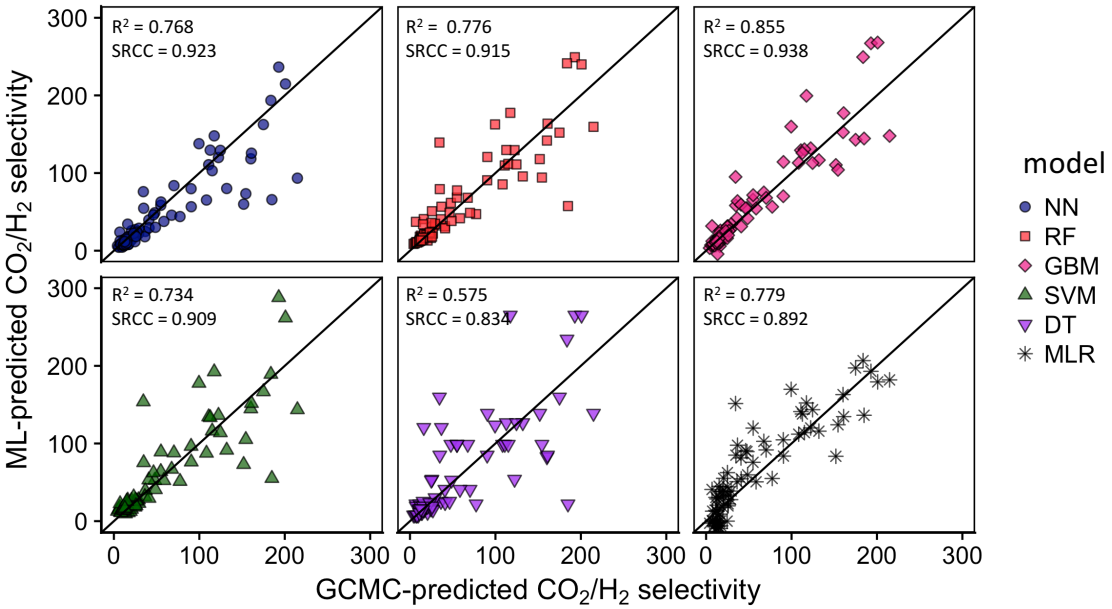

Figure B.20 Parity plots comparing the predictions on the test set of CO_2 selectivity for adsorption of 20:80 CO_2:H_2 mixture at 20 bar by each machine learning algorithm to those obtained by GCMC simulation. GBM emerged as the most accurate algorithm. The three largest GBM prediction outliers are all **fcu** MOFs with functionalization -$(CN)_2$, -$(CN)_4$, and, -$(OH)_2$.

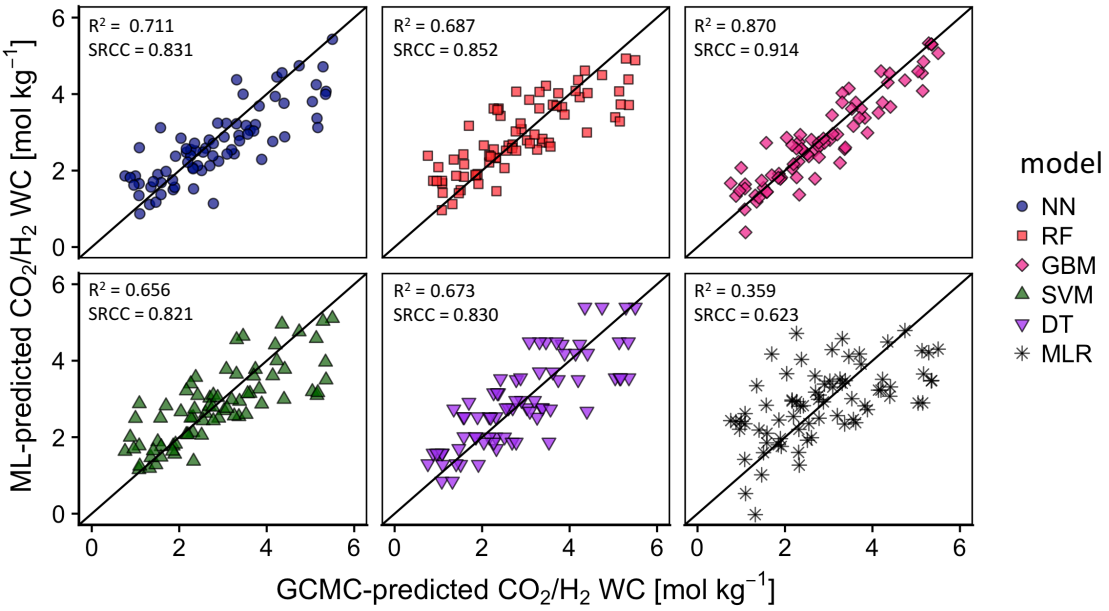

Figure B.21 Parity plots comparing the predictions on the test set for CO_2 working capacity in the CO_2:H_2 mixture case by each machine learning algorithm, showing the GBM model to be the most accurate.

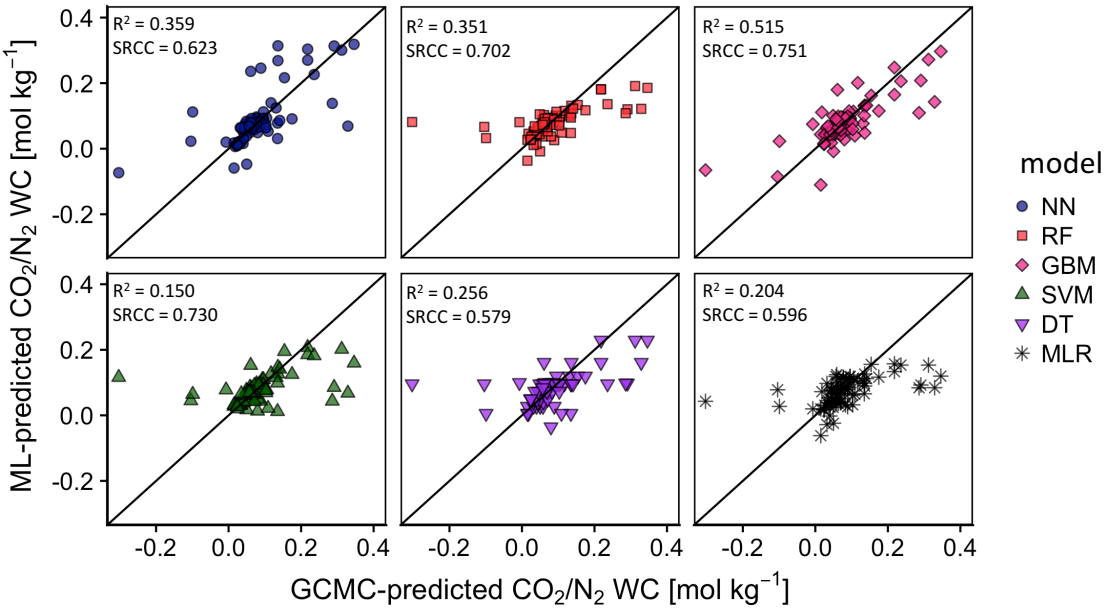

Figure B.22 Parity plots for regression model predictions on the test set of CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Again, the GBM model is the most accurate on the test set. The top 5 largest GBM prediction outliers (for largest to smallest) are acs1 MOF with $-(CN)_2$ functionalization, bcs1 MOF with $-(H)_4$, **bcs1** MOF with $-(NO_2)_2$, **lvtb1** MOF with $-(CN)_4$, and **lvtb3** MOF with $-(SH)_2$. Working capacity was calculated by setting the gas-phase compositions during desorption conditions to a 90:10 CO_2:N_2 mixture, which is an approximation based on the assumption of high CO_2/N_2 selectivity during adsorption stage. A negative working capacity is an indication that such assumption is incorrect for the specific material.

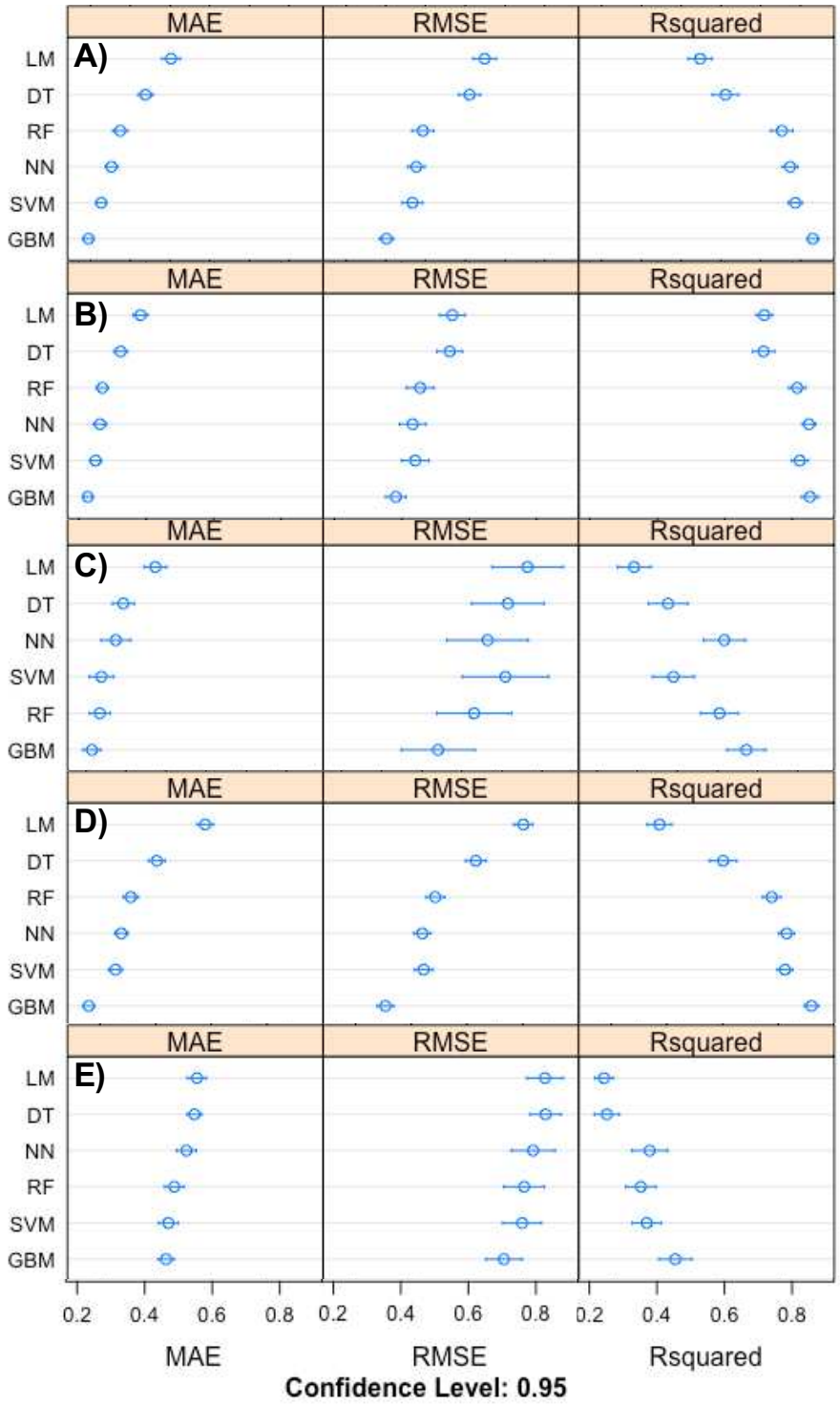

Figure B.23 Cross-validation metrics: mean absolute error (MAE), root mean squared error (RMSE), and R2 ("Rsquared"), from regression model training for (a) CO_2 loading at 1 bar, (b) CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar, (c) CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar, (d) CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar, (e) CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Error bars correspond to the 95% confidence interval.

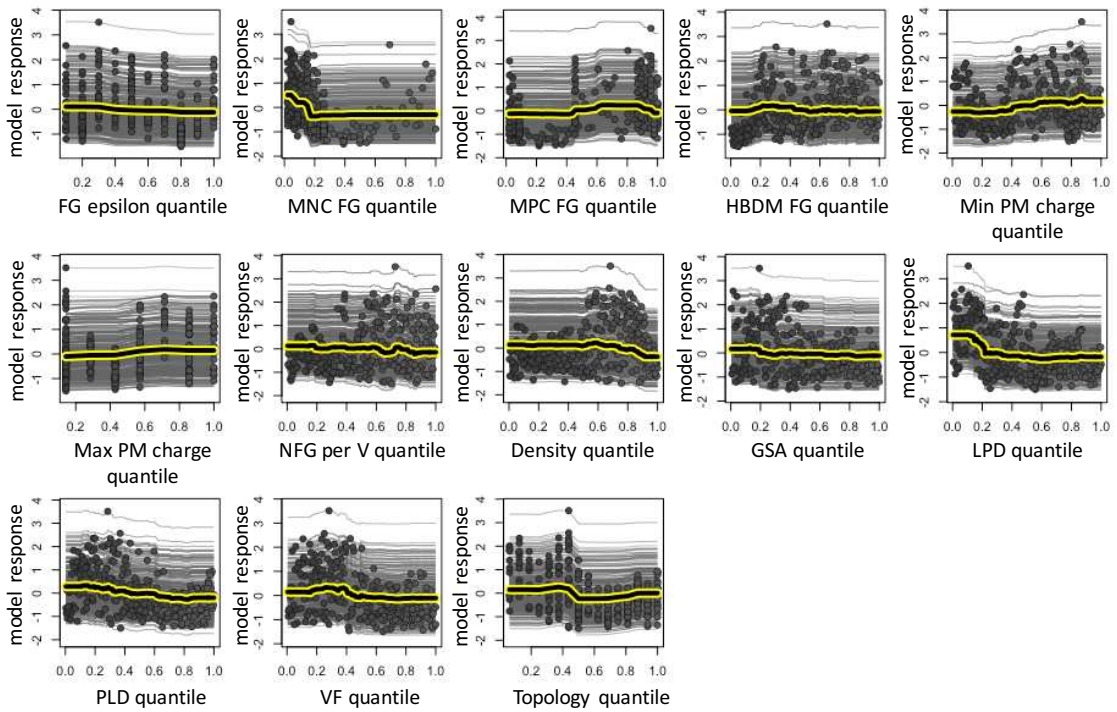

Figure B.24 ICE plots for each variable from the GBM model of CO_2 loading at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations.

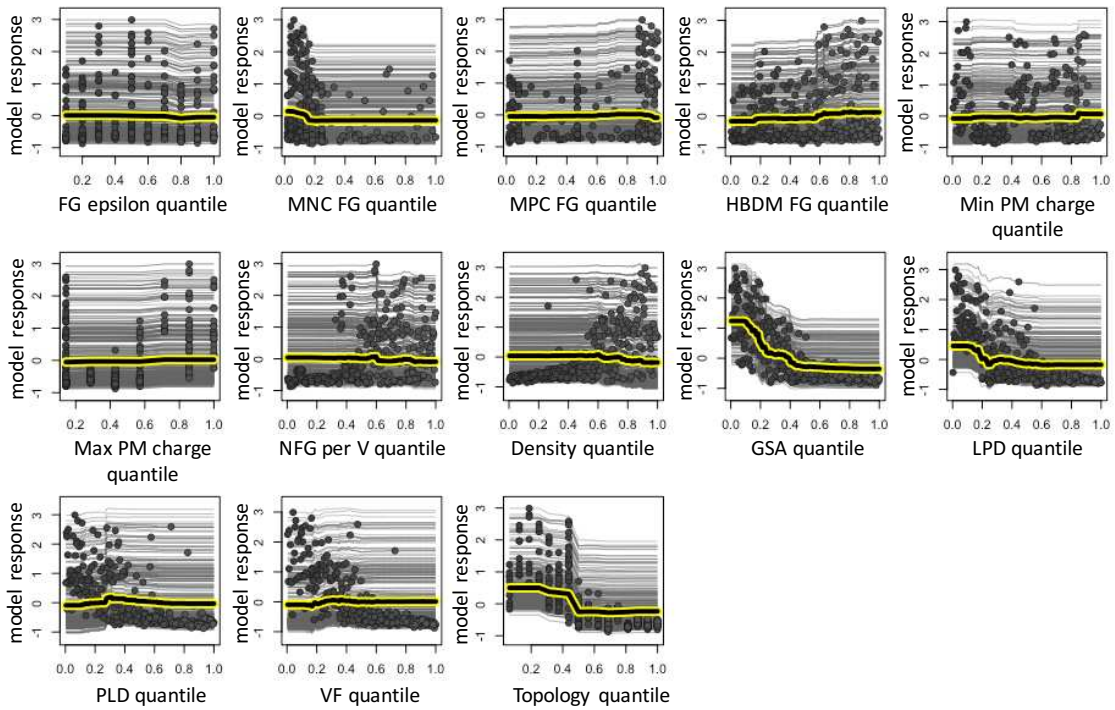

Figure B.25 ICE plots for each variable from the GBM model of CO_2 selectivity in a 20:80 CO_2:H_2 mixture at 20 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations.

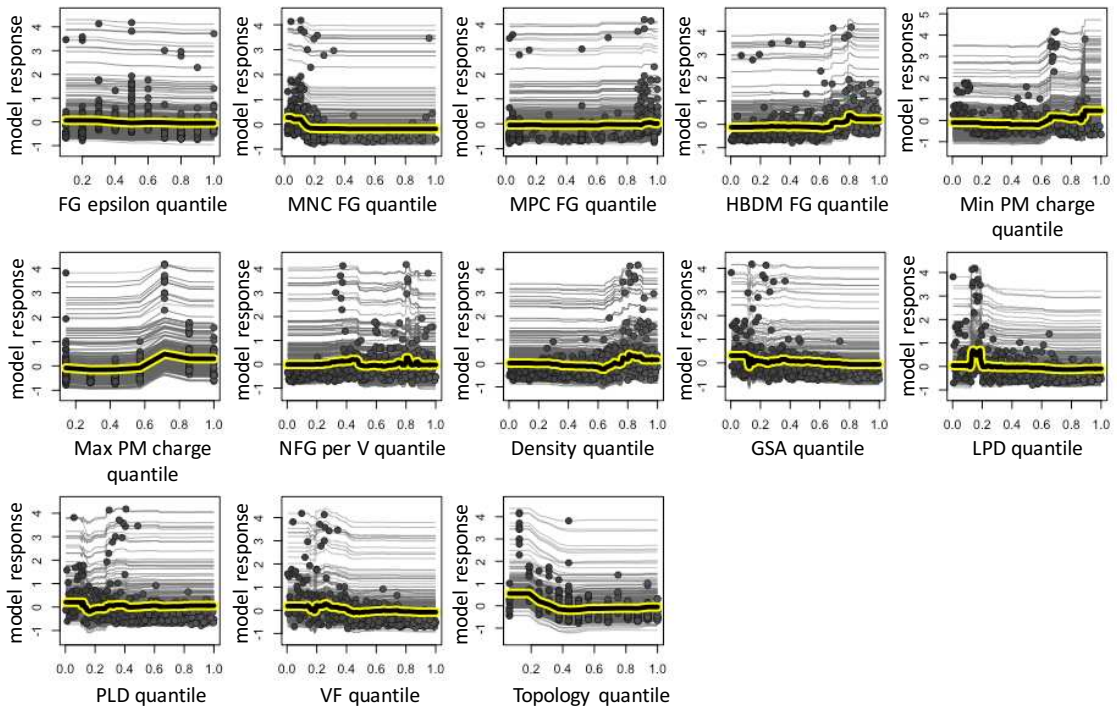

Figure B.26 ICE plots for each variable from the GBM model of CO_2 selectivity in a 15:85 CO_2:N_2 mixture at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations.

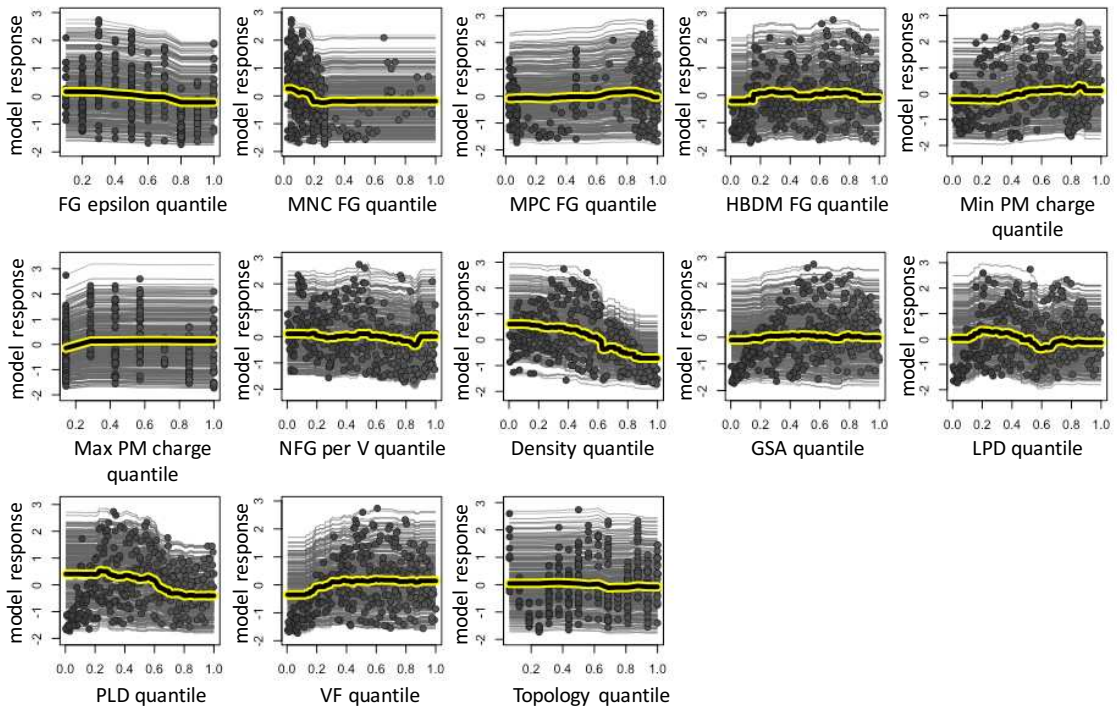

Figure B.27 ICE plots for each variable from the GBM model of CO_2 working capacity in a 20:80 CO_2:H_2 mixture at 20 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations.

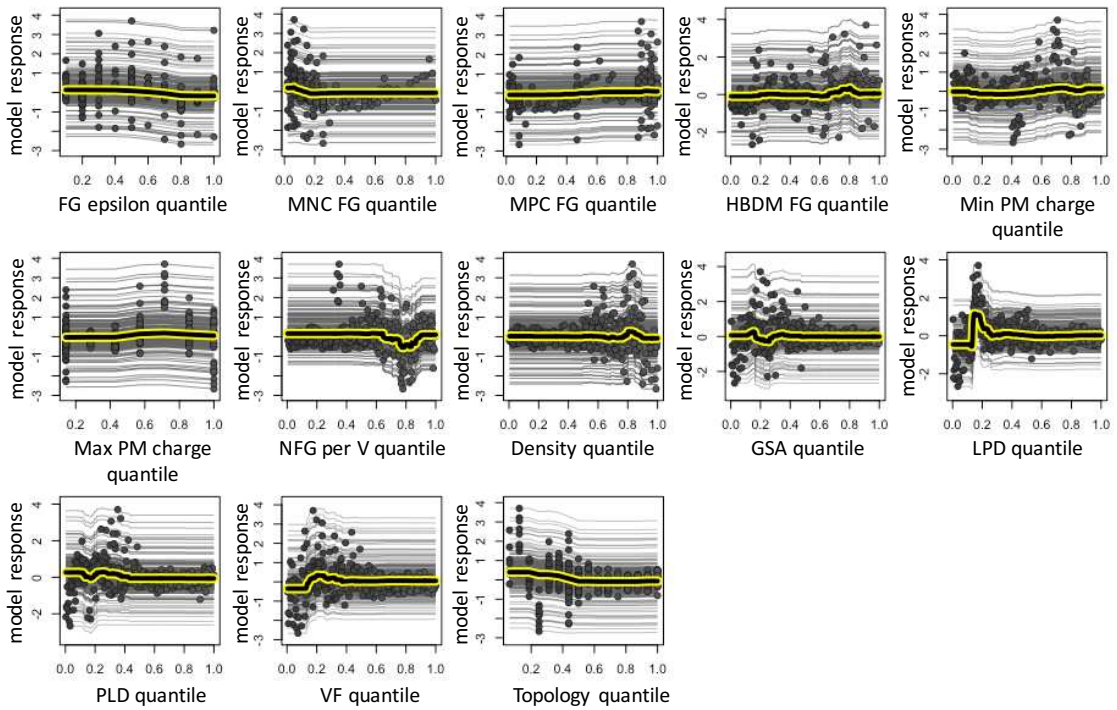

Figure B.28 ICE plots for each variable from the GBM model of CO_2 working capacity in a 15:85 CO_2:N_2 mixture at 1 bar. Thin gray lines and gray points are the model fit to individual observations and the individual observations, respectively. The yellow-outlined line is the average of the individual expectations.

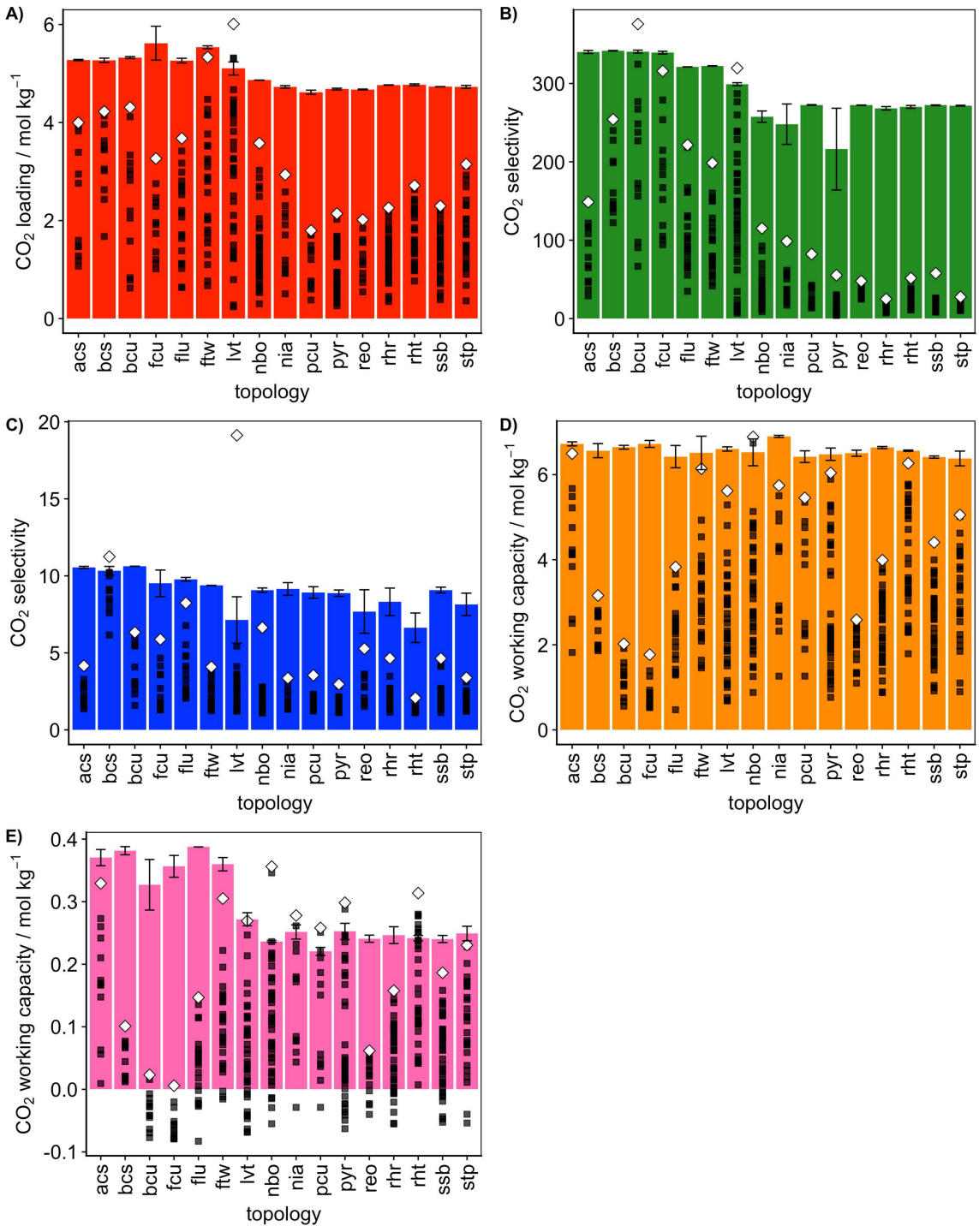

Figure B.29 Genetic algorithm maximization values (bars) along with GCMC data (black squares). The maximum values from GCMC simulations are white diamonds. (a) CO_2 loading at 1 bar, (b) CO_2 selectivity over H_2, (c) CO_2 selectivity over N_2, (d) CO_2 working capacity for the CO_2:H_2 mixture case, (e) CO_2 working capacity for the CO_2:N_2 mixture case. Error bars are the standard error of the mean across five maximizations.

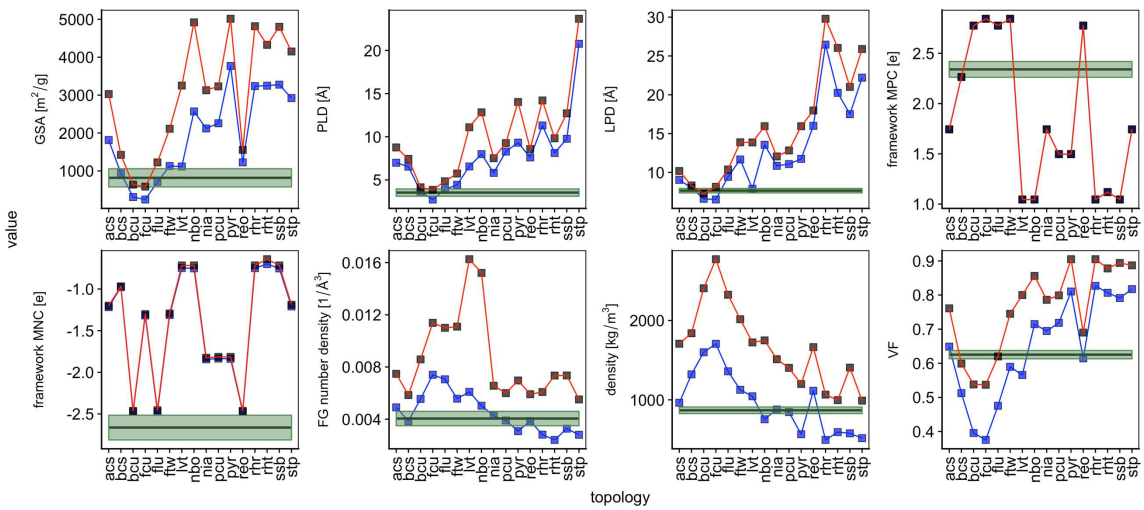

Figure B.30 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 loading at 1 bar (black lines). Green regions indicate the standard deviation of the optimum values.

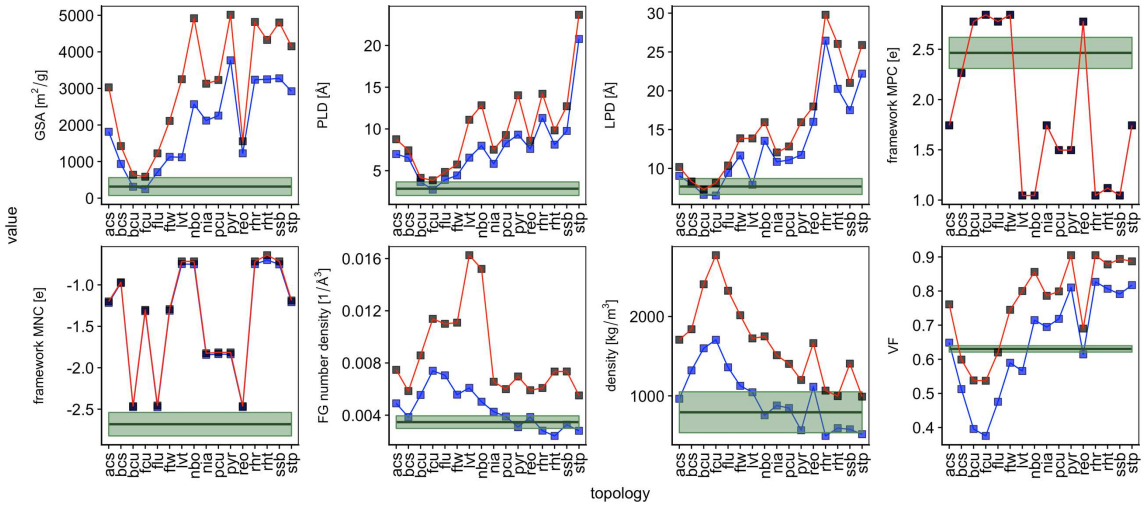

Figure B.31 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 selectivity over H_2 (black lines). Green regions indicate the standard deviation of the optimum values.

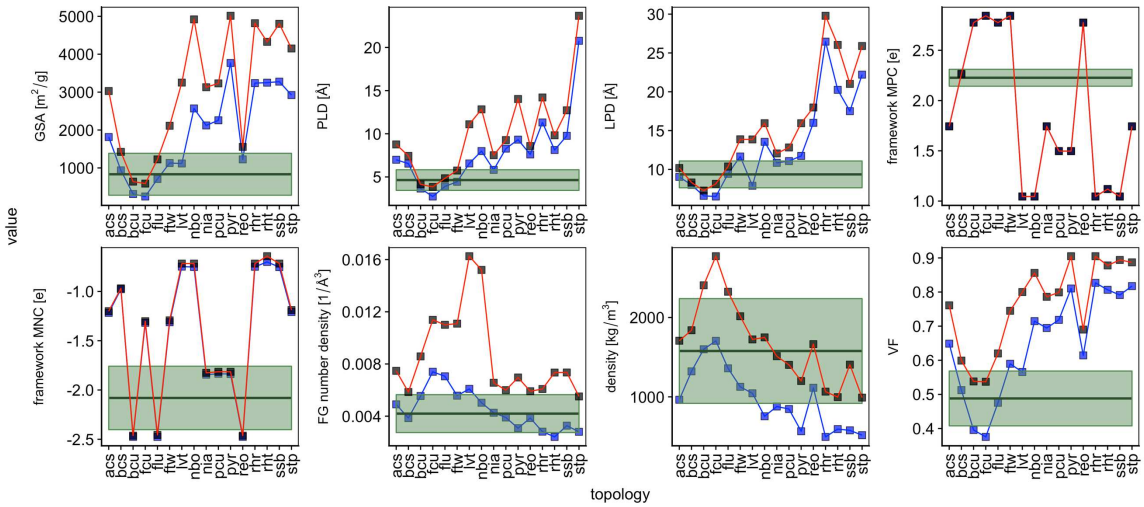

Figure B.32 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 selectivity over N_2 mixture (black lines). Green regions indicate the standard deviation of the optimum values.

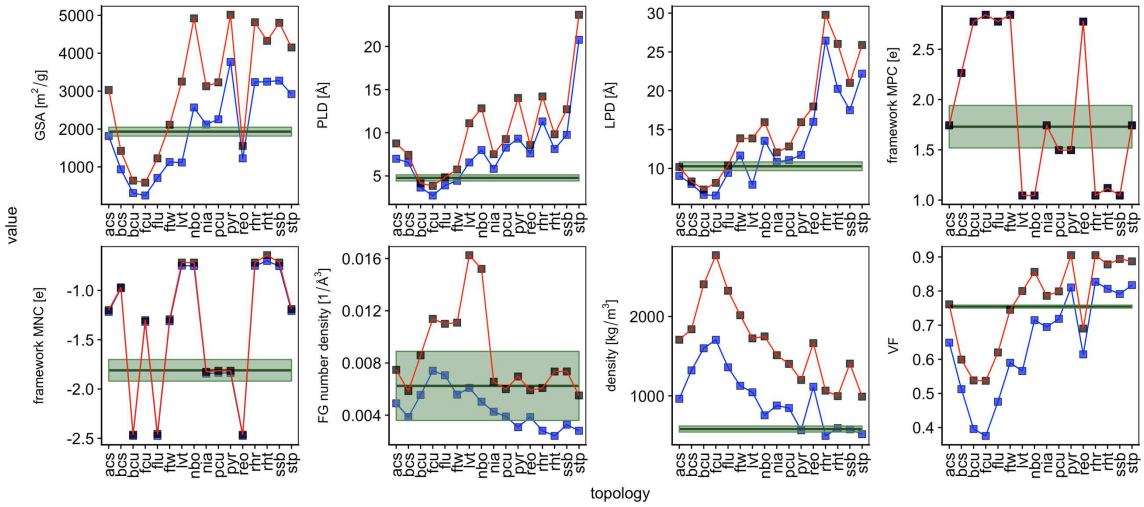

Figure B.33 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 working capacity for the 20:80 CO_2:H_2 mixture case (black lines). Green regions indicate the standard deviation of the optimum values.

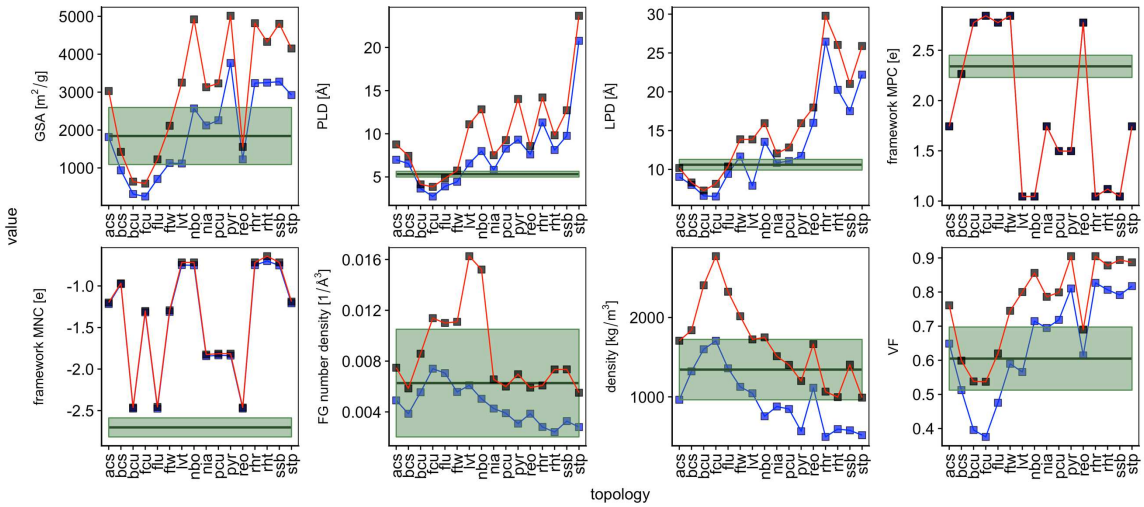

Figure B.34 Maximum (red squares) and average (blue squares) values compared to optimum values of the variables depending on parent MOF topology and building blocks (i.e. functional group properties excluded) predicted by maximization of CO_2 working capacity for the CO_2:N_2 mixture case (black lines). Green regions indicate the standard deviation of the optimum values.

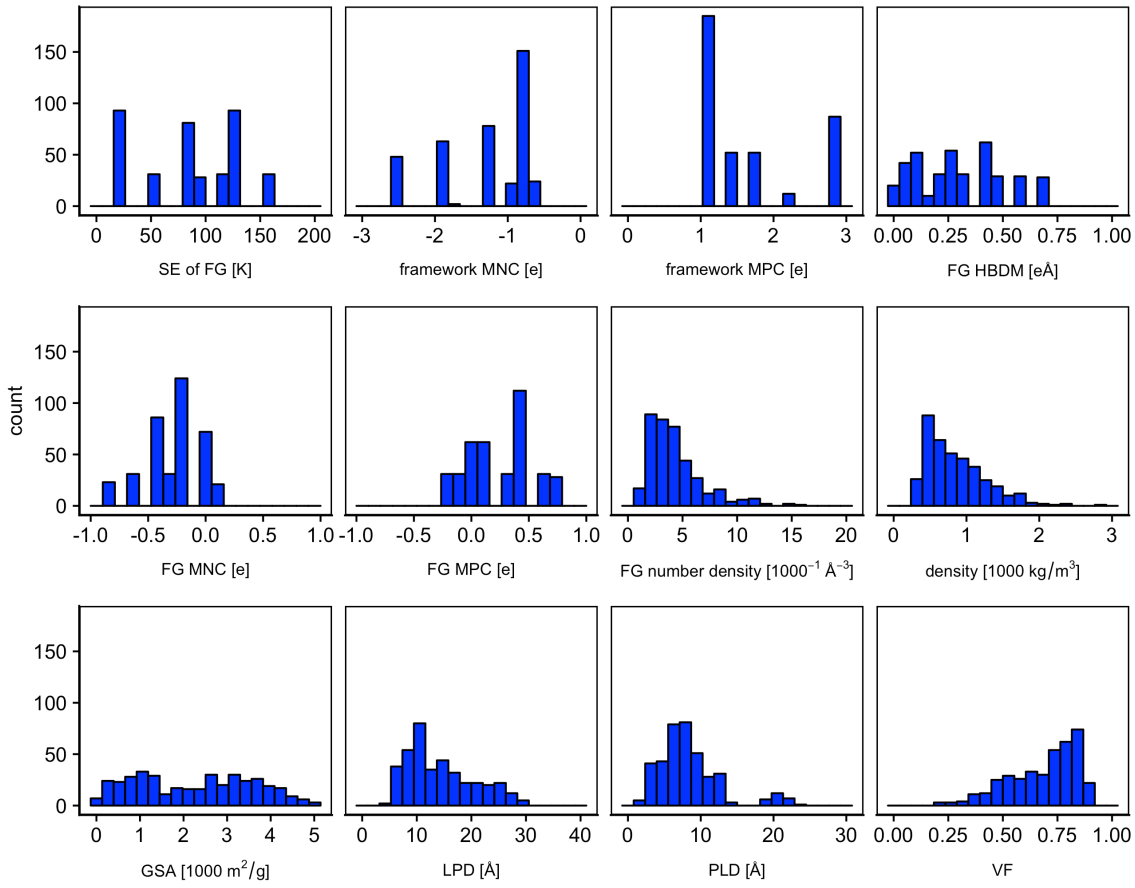

Figure B.35 Histograms of properties encountered in the MOFs constructed for this study, and which were used to train the machine learning algorithms.

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 5

C.1 Adsorbate Force Field Parameters and Fingerprinting

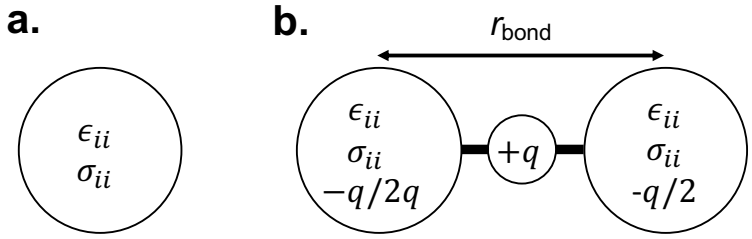

Figure C.1 (a) Shows the force-field model used for single atom adsorbates, where only Lennard-Jones parameters are assigned to these adsorbates. (b) Shows the force-field model used for two-atom adsorbates, where each atom is assigned LJ parameters and a charge, the atoms are separated by a distance rbond, and a charge-balancing pseudo atom is placed at the bond center (note that q can be zero).

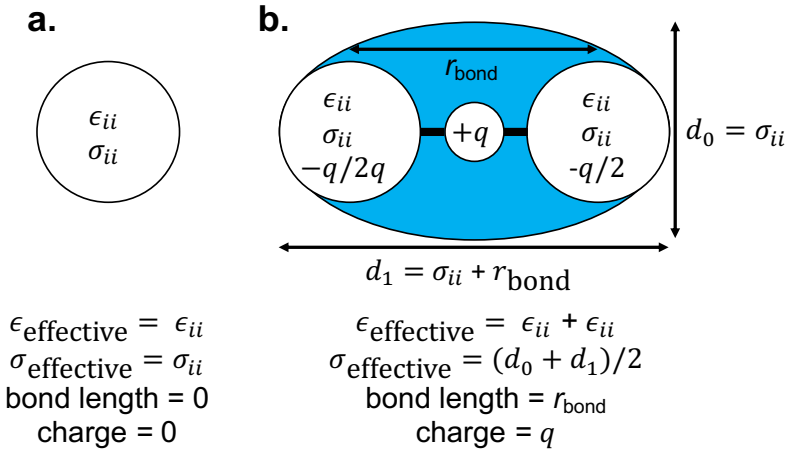

Figure C.2 The adsorbate fingerprints used as input for the MLP. a. Shows the fingerprint for single atom adsorbates and b. shows the fingerprint for two atom adsorbates, where $\sigma_{\text{effective}}$ is taken to be the average of the adsorbate end-to-end and top-to-bottom distances.

C.2 Assessing the Importance of Framework-Adsorbate Electrostatic Interactions

To determine how much loading changes when considering framework-electrostatic interactions we calculated the loadings of N2 in our test set MOFs at the same 10 pressures considered in the main text (between 1 and 100 bar). Note that N2 has slightly higher magnitude partial charges than any alchemical adsorbate, so we expect the electrostatic interactions of N2 to be the strongest of any adsorbate considered

Table C.1 Parameters for single-site alchemical adsorbates used for *training*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]
50	3.0
100	3.5
150	4.0
200	4.5
250	-

Table C.2 Parameters for single-site alchemical adsorbates used for *validation*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]
60	3.10
75	3.25
125	4.75
175	4.25
225	-

Table C.3 Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) used for *training*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]	Charge [e]	Bond Length [Å]
15	3.0	0.0	1.0
30	3.7	0.2	1.3
45	4.5	0.5	1.6
60	-	0.9	-
95	-	-	-

Table C.4 Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) used for *validation*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]	Charge [e]	Bond Length [Å]
25	3.25	0.0	1.1
35	3.50	0.3	1.3
50	4.00	0.6	1.5
70	-	0.9	-
80	-	-	-

Table C.5 Lennard-Jones parameters for single site real adsorbates.

Adsorbate	ϵ_{ii} [K]	σ_{ii} [Å]
Helium	10.9	2.64
Argon	124.1	3.42
Methane	148.0	3.73
Krypton	166.4	3.64
Xenon	221.0	4.10

in this study. Figure C.3 is a parity plot showing loading calculated with adsorbate-framework electrostatic interactions vs. loading without these interactions. The loadings calculated with framework charges is slightly higher (by about 0.07 mol/kg, on average) than those calculated without considering framework charges. The MAE and MAPE corresponding to Figure C.3 are 0.18 mol/kg and 5.5%, respectively, which is similar to the model error on the test set, so we expect our model predictions to be similarly accurate with/without including framework charges. In fact, our final model overestimates N2 loading calculated without framework charges (see Figure 5.6 in the main text) by 0.02 mol/kg on average for the test set MOFs, so the error associated with including framework charges would be mitigated slightly by error cancellation.

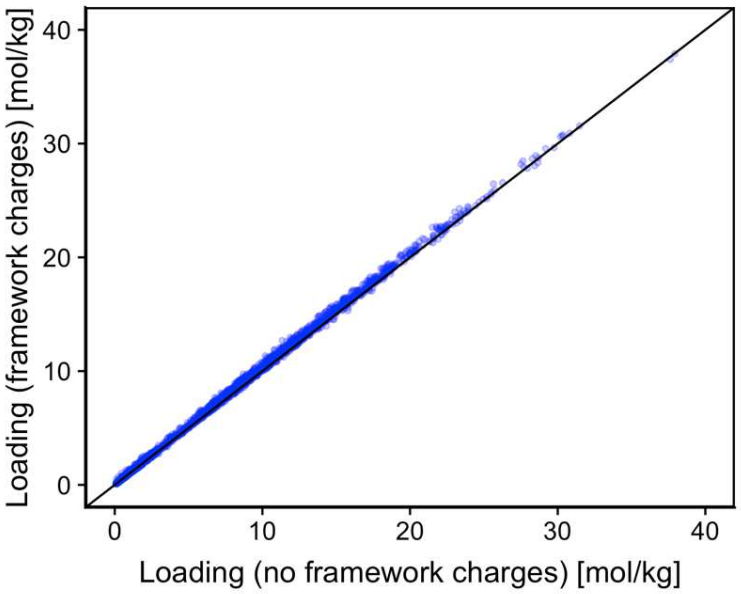

Figure C.3 A comparison of N2 loading in the 400 test set MOFs calculated with (y-axis) and without (x-axis) framework charges.

C.3 Model Training Details

Python 2.7 was used for all training tasks.

Table C.6 Version information for Python packages used for MLP training.

Package	Version
sklearn	0.20.3
numpy	1.16.4
keras	2.2.4
tensorflow	1.14.0
joblib	0.12.1

Table C.7 The optimal hyperparameter values for our final MLP found using a grid search. Topology is denoted by $N_1 - \cdots - N_M$, where N_i is the number of nodes in each hidden layer.

Hyperparamter	Value
Topology	50-50-20-20
Activation	sigmoid
Optimizer	Adam
Initializer	Glorot normal
Learning Rate	0.0001
Batch Size	50
Loss	MAE

C.4 Additional Prediction Data

To show that our neural network can make accurate predictions of GCMC loadings calculated in structures with blocking spheres, we used zeo++ to calculate the blocking spheres for a methane sized probe in our 400 MOF test set and calculated the textural properties and loading using the blocked pore structures. While zeo++ uses only MOF/adsorbate geometry to calculate inaccessible regions of pore volume [305], we note that there are other methods that consider MOF-adsorbate interaction energies, such as the flood-fill algorithm employed by Chong et al., in which pores with energy barriers for entry higher than 15kBT are blocked (in this case by artificially increasing the adsorption energy in regions inside the inaccessible pores) [373]. However, our model can be used with any pore- blocking scheme, so long as the textural properties input to the model are consistent with that blocking scheme. When fed textural properties calculated without blocking spheres and asked to predict loadings from GCMC with blocking spheres our model makes a number of egregious overpredictions, as one might expect (see Figure C.6). Note that the vast majority of MOFs in the test set do not require blocking spheres, and thus, most predictions remained the same as those shown in Figure 5.6 But, when fed textural properties calculated with blocking spheres, all these overpredictions are corrected. Therefore, it is not necessary to include blocking spheres in model training, so long as the MOF fingerprint contains variables which can be corrected for the presence of blocking spheres.

We eliminated MOFs from the CoRE MOFs [70, 71] according to the following procedure. First, 238 unphysical MOFs with close contacts (any atoms closer than 0.70 Å) were removed. Next, we eliminated any MOFs with elements not included in our database (9286 MOFs). Next, MOFs with textural properties outside the range covered by our training MOFs were eliminated (931 MOFs). This essentially meant imposing a lower bound on the VF, GSA, LPD, PLD, and inverse density, as no CoRE MOFs surpassed the upper bounds of these properties in our training set. Next, any MOFs without a carboxylate moiety were eliminated (496 MOFs). These MOFs were eliminated as our model was trained only on MOFs that

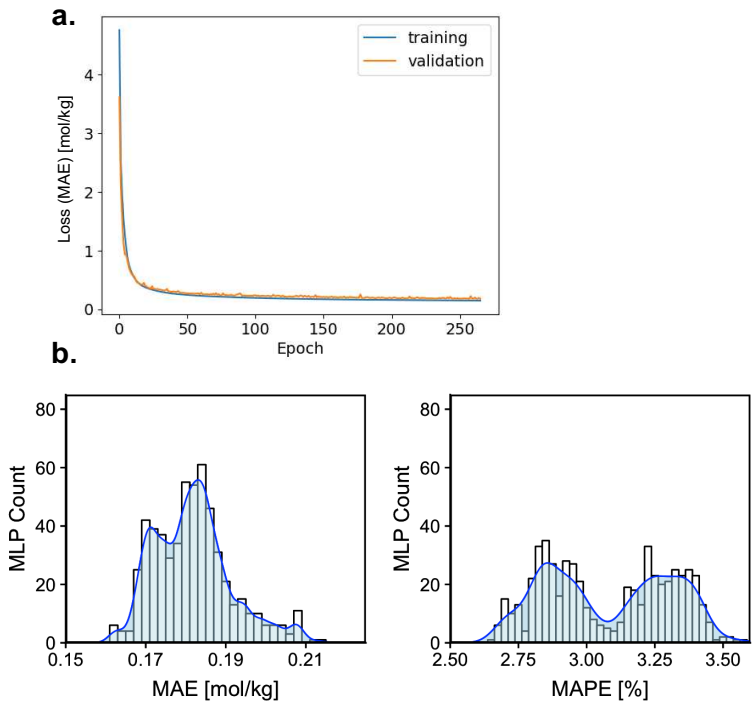

Figure C.4 (a) Shows the learning curve of our final model, demonstrating that our training set is representative of the validation set (and vice-versa) and that the model was not overfit by fitting over too many epochs at the learning rate used. (b) shows the error variability in 500 replications of our final model trained with different random seeds.

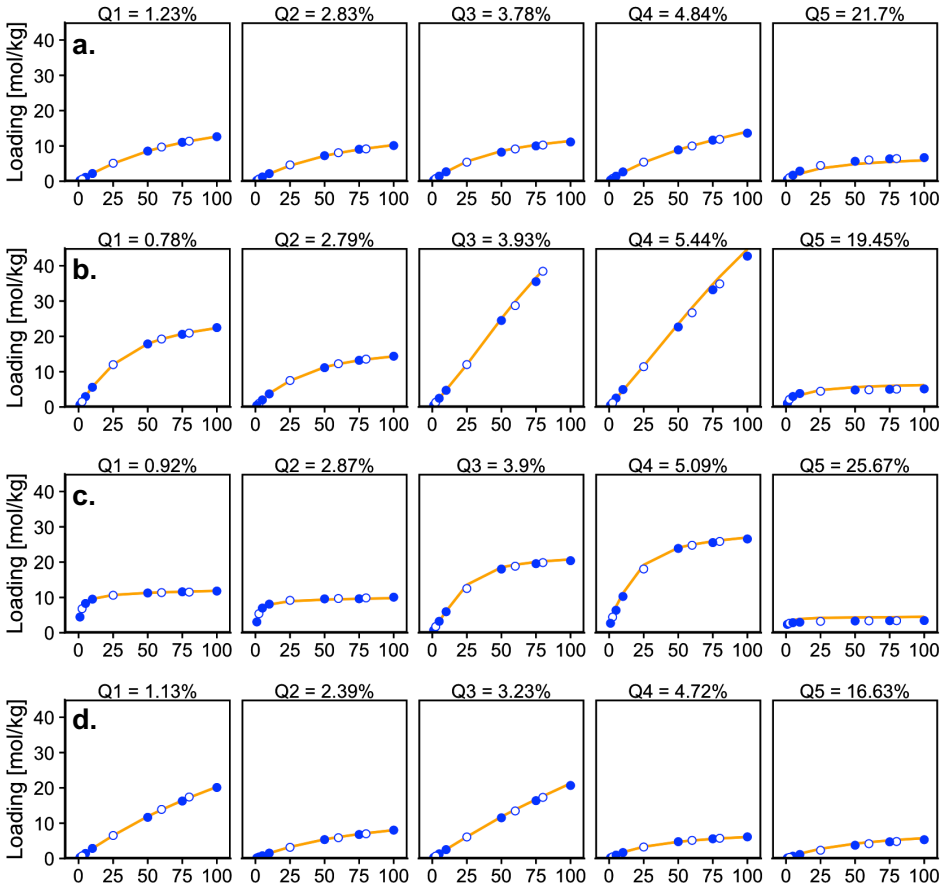

Figure C.5 Plots analogous to Figure 5.7 in the main text for predictions on the remaining simple adsorbates, with adsorbates: (a) argon, (b) krypton, (c) xenon, and (d) nitrogen. Fugacities corresponding to filled, blue points were included in training, empty points correspond to pressures not included in training, orange lines correspond to model predictions.

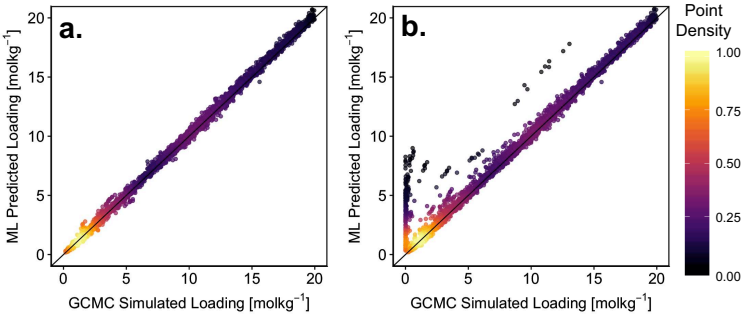

Figure C.6 Model predictions for methane loading vs GCMC simulated loading with blocking spheres for the 400 MOF test set.(a) Shows predictions using MOF textural properties corrected for the presence of blocking spheres,(b) shows predictions using uncorrected MOF textural properties. Note that this plot only shows loadings in the range that changed due to the presence of blocking spheres, i.e. over a lower loading range as the blocking spheres only affected the loading in certain small-pore MOFs.

possessed this moiety. We made sure that many MOFs with non-oxocluster nodes remained after this final elimination. Given our method of elimination, not only are there metal nodes not present in our training data in the remaining 1528 MOFs, but there are also MOFs with different chemical motifs and topologies.

Table C.8 Model performance metrics of our final model for loading predictions made on the CoRE MOFs.

Adsorbate	MAPE[%]	MAE [mol/kg]	R
Argon	16.3	0.52	0.993
Methane	14.2	0.52	0.994
Krypton	14.5	0.61	0.994
Xenon	10.6	0.63	0.994
Ethane	9.5	0.58	0.995
Nitrogen	14.7	0.38	0.995

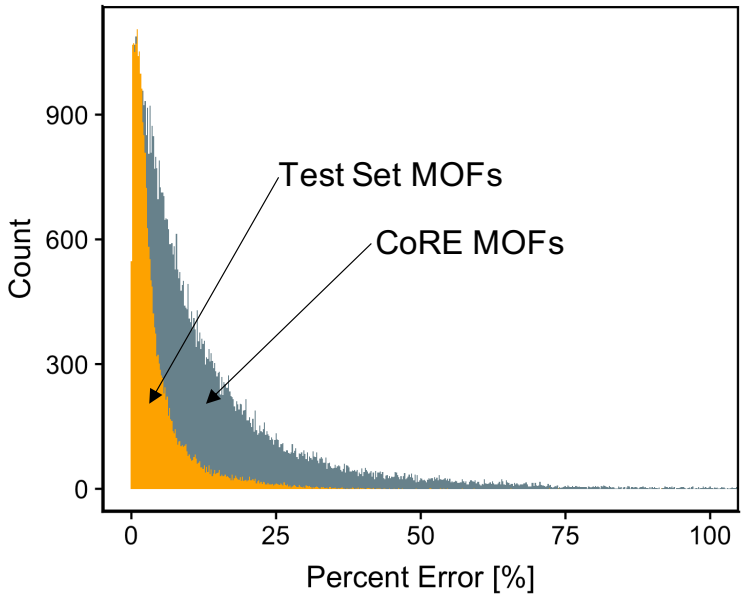

Figure C.7 Showing the distribution of percent error for loadings predictions in the test set MOFs (indicated, orange) and the CoRE MOFs (indicated, blue). These distributions include predictions made on all adsorbates at all fugacities considered.

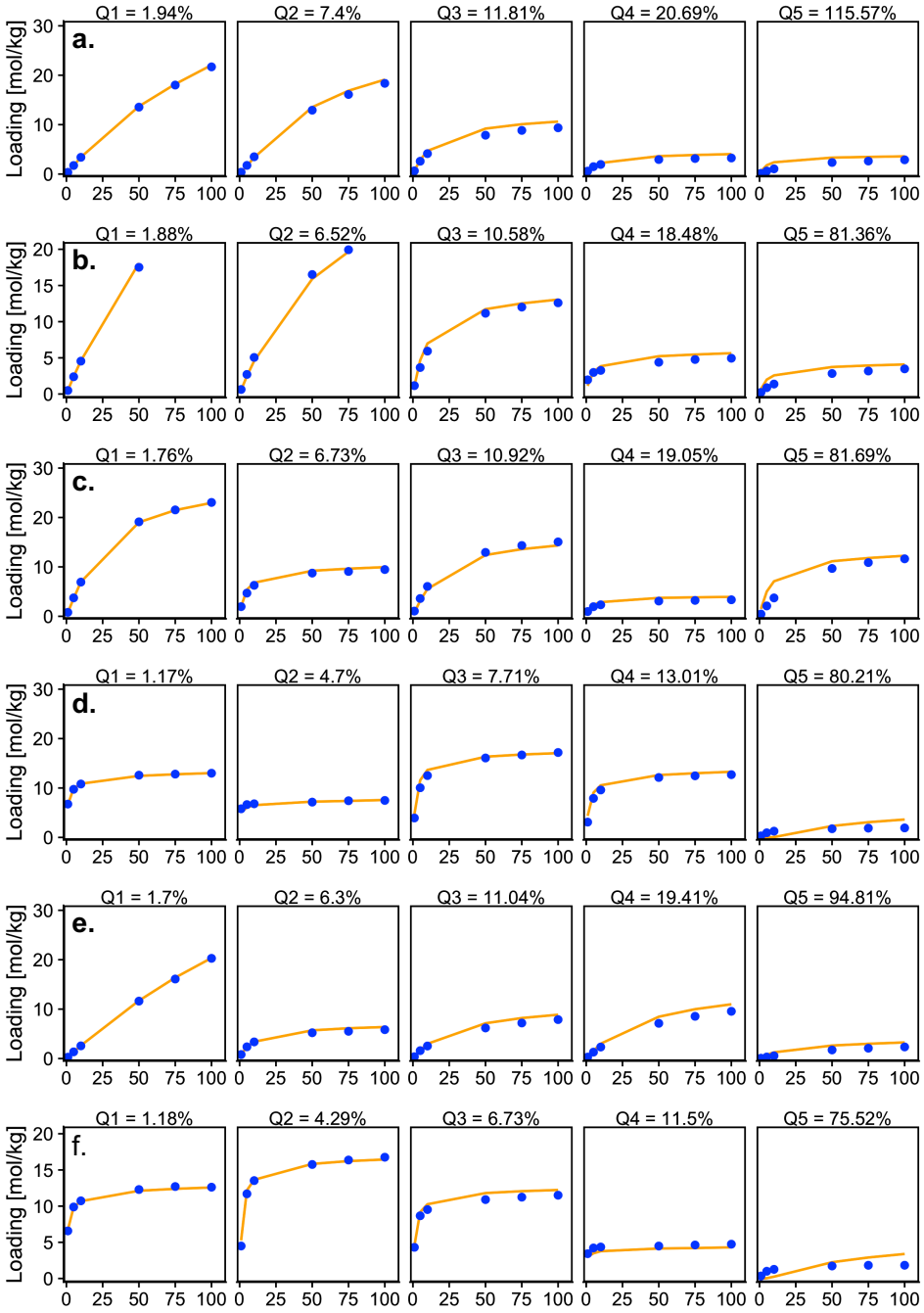

Figure C.8 Plots analogous to Figure 5.7 in the main text for CoRE MOF predictions, with adsorbates: (a) argon, (b) methane, (c) krypton, (d) xenon, (e) nitrogen, and (f) ethane. Filled, blue points correspond to GCMC simulated values, orange lines correspond to model predictions. Only the first four and five points (respectively) of the methane Q1 and Q2 isotherms are shown as higher-pressure points surpass the y-axis scale.

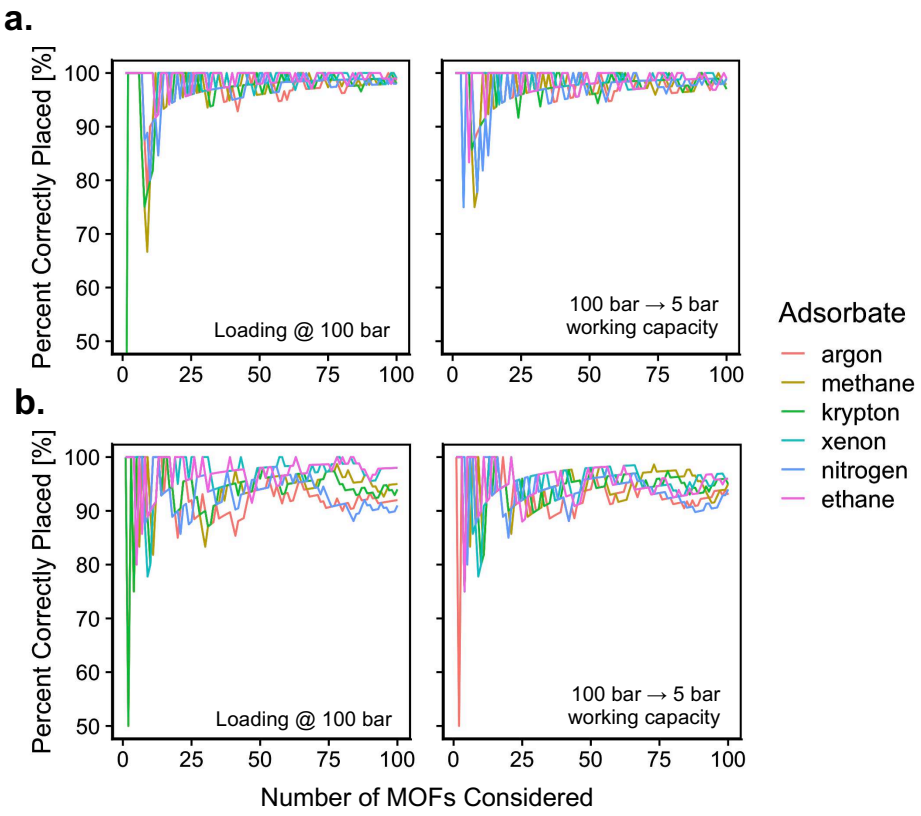

Figure C.9 The percent of MOFs correctly placed by the MLP (y-axis) out the top N MOFs where N is the value on the x-axis for all six adsorbates considered for (a) the 400 test set MOFs, and for (b) the CoRE MOF subset.

Table C.9 Model performance metrics of our final model for loading predictions made on adsorbates with properties outside of range spanned by the training set (extrapolative predictions).

Adsorbate	MAPE[%]	MAE [mol/kg]	R
Helium	8.4	0.11	0.998
Hydrogen	19.6	0.52	0.998
Propane	38.1	6.67	0.985
Butane	42.0	7.12	0.989
Isobutane	27.6	4.89	0.990
Benzene	12.6	2.14	0.976

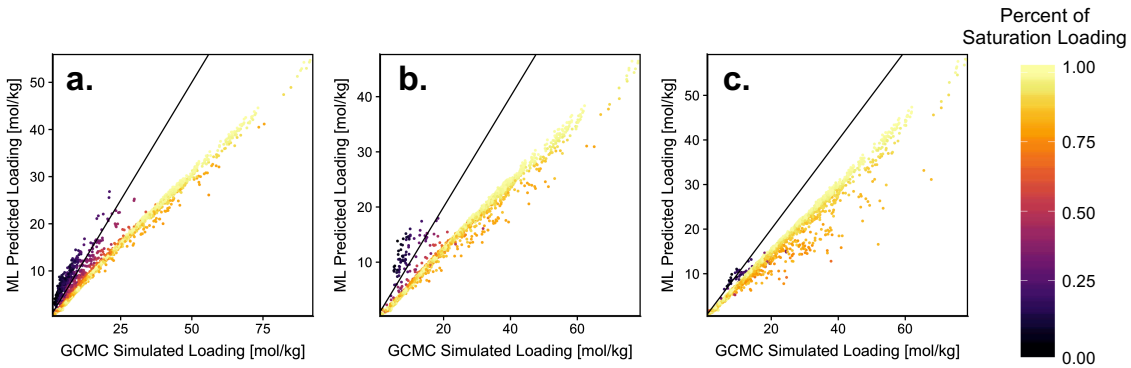

Figure C.10 Parity plots comparing the predictions of the final MLP model for (a) propane, (b) n-butane, and (c) isobutane versus GCMC-calculated values. Points are colored by how close they are to saturation loading (1.00 being at saturation).

APPENDIX D

SUPPORTING INFORMATION FOR CHAPTER 6

D.1 Convergence of the Non-Equilibrium Switching Method

During the application of the Frenkel-Ladd path, the convergence of the switching method implemented in LAMMPS by Freitas et al. depends on both the size/shape of the simulation cell and the switching time, where the latter is the time it takes to transition from the force field Hamiltonian to the Einstein crystal Hamiltonian, and vice versa [169]. Thus, it is necessary to test the convergence of the resulting free energy with respect to these two parameters. These tests were performed using UFF4MOF [82, 210] parameters for the force field Hamiltonian.

To test convergence with system size (here quantified by the number of atoms), we randomly selected 500 MOF prototypes from our database and calculated their free energy using 10 system sizes for each MOF varying (approximately) between 1,000 and 10,000 atoms. To create different system sizes for a given MOF, we replicated each MOF unit cell in different crystallographic directions to produce a larger supercell of approximately the desired number of atoms. We aimed to make the supercell dimensions as close to each other as possible, because supercells that are significantly larger along one crystallographic direction could lead to numerical instabilities during MD simulations (especially in cases where unit cell angles dramatically deviate from 90°) and could potentially introduce significant shape effects to the calculation. Accordingly, our unit cell replication strategy was as follows: (*i*) we determined the minimum number of replications required to reach the desired number of atoms, (*ii*) we determined whether this number of replications allowed for a supercell with no dimension more than twice the shortest dimension, (*iii*) if so, we used this supercell, if not, we incremented the number of replications by one and returned to step (*ii*). To summarize, starting from the number of replications required to reach the desired minimum number of atoms, we found the minimum number of replications that could produce a supercell with no dimension longer than twice the shortest dimension. For example, if the number of replications required to reach the desired minimum number of atoms is three, and the unit cell is 30 Å in each direction, a $3 \times 1 \times 1$ supercell would have dimensions $90 \times 30 \times 30$ Å, which does not fulfill the criteria that the longest dimension is no more than twice the shortest dimension. However, if we increment the number of replications to four, we can have a $2 \times 2 \times 1$ unit cell with dimensions $60 \times 60 \times 30$ Å, which does fulfill our criteria.

We plotted the deviation in free energy for each MOF system size relative to a corresponding 15,000-atom supercell (Figure D.1b). On average, at least for the replication strategy that we used, we found negligible size-dependence for the free energy calculation as the MOF system size varied between

1,000 and 10,000 atoms (note that the average deviations from the 15,000-atom simulations was dramatically less than 1% for each system size). However, we did observe a slight systematic error until the system size reached 5000 atoms at minimum. Hence, we used supercells that included at least 5,000 atoms. Since we checked size convergence before time convergence, we performed each switch for 1,000 ps, which we estimated to be a safe choice to ensure time convergence (as confirmed below).

Once a minimum system size criterion was established, we tested convergence with switching time on 500 randomly selected MOFs from our database (the same subset used for size-convergence testing) and calculated their free energy using 12 switching times between 10 and 500 ps. We then plotted the deviation in free energy for each MOF for each switching time relative to a 1,000 ps simulation (Figure D.1a). We concluded that a switching time of 400 ps was likely sufficient for high throughput screening, given that the average deviation from the 1000 ps simulation was 0.6% (with a maximum of 4% error).

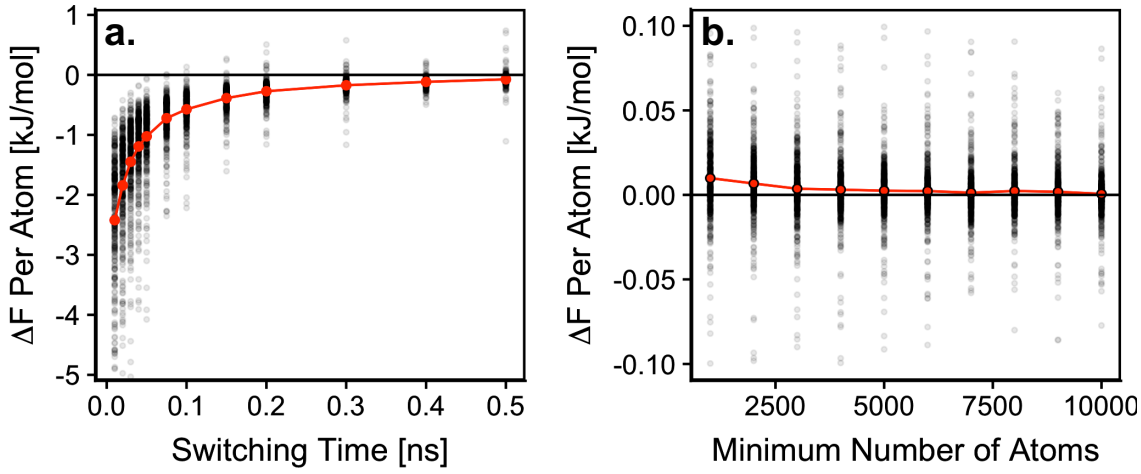

Figure D.1 FL path free energy convergence with (a) switching time and (b) minimum number of atoms in the simulation supercell.

D.2 Determining Einstein Crystal Force Constants

For each MOF, we calculated different force constants α_i for each atom type (as per the force field) according to Equation refeq:C68, However, any α_i that would be less than 10.0 kcal/mol Å^2 was instead set to 10.0 kcal/mol Å^2. We did this because we found that, when some force constants were too small, corresponding atoms could eventually migrate too far from equilibrium under the Einstein crystal Hamiltonian, causing them to be "lost" from bond, angle, or dihedral lists, resulting in a LAMMPS error. In addition, smaller force constants increase the probability of atom overlap for values of the switching parameter close to one. Note that although bonds/angles/dihedrals are not contributing to the energy of

the system when under the Einstein crystal Hamiltonian, LAMMPS must still keep track of them to initiate the reverse switching process. We found that these errors could be minimized by enforcing the 10.0 kcal/mol Å^2 lower bound for α_i. In principle, given a sufficiently long simulation time, any force constant(s) could be used during the FL path method. However, we use Equation 6.8 to estimate the constants to keep the "real" Hamiltonian and Einstein crystal Hamiltonian dynamics similar, which results in less energy dissipation during the switching process, and thus makes the process more efficient (requiring less switching time) [169]. Accordingly, we did not expect the setting of the lower bound to significantly affect our results, provided long enough switching times were used.

To confirm our expectations, we calculated the FL path free energies in 500 randomly selected MOFs with and without the lower bound on α_i. The results of this test are presented in Figure D.2, which is a parity plot of the free energies without any restriction on the force constants versus free energies with the 10.0 kcal/mol Å^2 lower bound. The MAE calculated from the data in Figure D.2 of 0.22 kJ/mol is only slightly higher than the mean statistical error of our FL path implementation (0.12 kJ/mol with a standard deviation of 0.13 kJ/mol). However, there are a few outliers for exceptionally "mobile" MOFs (i.e. MOFs with large MSDs). Comparing the median absolute error calculated from the data in Figure D.2 (0.11) with the median statistical error (0.09) shows more robustly that that the lower bound on α_i will not significantly influence our results.

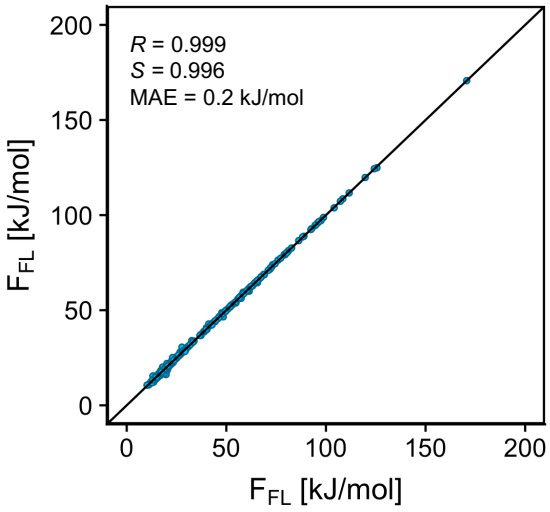

Figure D.2 FL free energies calculated without any restriction on the force constants versus those calculated with a force constant lower bound of 10 kcal/mol Å^2. R and S are the Pearson and Spearman ranking correlation coefficients, respectively, and MAE is the mean absolute error.

D.3 Estimating the Error Associated with our FL Path Implementation

In addition to the above tests, we also used our randomly selected 500 MOFs to estimate the error of our implementation of the FL path method. To do this, we ran 10 replicate simulations for each MOF and calculated the span of free energies (the difference of the maximum and the minimum free energies among the replicates) covered for each MOF. The average span is 0.12 kJ/mol and the median span is 0.09 kJ/mol with a standard deviation of 0.13 kJ/mol. The average (median) percent difference between the minimum and maximum free energies in 0.6%. (0.5%). We also calculated the standard deviation of the 10 replicates for each MOF, the average of these standard deviations across the 500 MOFs was 0.04 kJ/mol. However, given that our reported energies are not the average of replicate simulations, the span statistics are a more robust estimate of the error for our calculations.

D.4 Fixed Center of Mass Corrections

Corrections are often made to Equation 6.8 in the main text to account for the finite size of the system and constraining of the crystal center of mass during Frenkel-Ladd path simulations, thus:

$$F_{\mathrm{FL}} = \frac{1}{2}(W_{i\to f}^{irr} - \ W_{f\to i}^{\mathrm{irr}}) + 3k_BT\sum_{i=1}^{N}\ln\left(\frac{\hbar\omega_i}{k_BT}\right) + k_BT\ln\left(\frac{n_r}{V}\left(2\pi k_BT\sum_{i}^{N}\frac{{\mu_i}^2}{\alpha_i}\right)^{\frac{3}{2}}\right) \tag{D.1}$$

where $\mu_i = m_i/\sum_i m_i$, n_r is the number of unit cell replications used in the simulation, and V is the equilibrium volume. The third term on the RHS is the correction term [164, 374]. Note that n_r is the number of primitive cells in the system if we consider each unit cell to have P1 symmetry (which is really the only practical assumption to make for a large database of computational MOFs). We found that this correction was both small compared to the statistical error of the method and similar across all MOFs (see above). Therefore, we concluded that the correction would negligibly affect absolute and relative free energies. However, for systems with fewer replications this may not be the case since the term decreases with increasing n_r.

D.5 Comparison of Calculation Times for QHA Free Energy and FL Path Free Energy

Both the quasi-harmonic approximation (QHA) and Frenkel-Ladd (FL) path require a molecular dynamics (MD) simulation to collect data, which is then integrated to produce the final free energy. The QHA methods requires equilibration in the NPT ensemble (200 ps, note that this equilibration is the same for the QHA and FL path methods), followed by a short (60 ps) production run in the NVT ensemble during which the velocities of each atom are output. These velocities are used to calculate the velocity autocorrelation function (VACF). The frequency of velocity output determines the maximum vibrational frequency available in the resulting phonon density of states (PDOS). Outputting velocities every 5 fs

allows for a maximum vibrational frequency of 100 THz, which is sufficient for most MOFs. The FL path method requires NPT equilibration (200 ps), then the calculation of the MSDs for each atom type in the NVT ensemble (100 ps), and finally the NVT switching procedure which consists of the forward/backward switching (400 ps each) and equilibration with the force field Hamiltonian and with the Einstein crystal Hamiltonian (30 ps each). Thus, the MD simulation time for calculated QHA free energies are much shorter than for the FL path. Indeed, the total required simulation time for QHA and FL methods are 260 ps and 1160 ps , respectively. However, the QHA method requires a significantly more expensive integration step, because the velocities from the velocity trajectories (saved in an approximately 500 MB file) must be averaged and Fourier transformed (this takes minutes of computer time). On the other hand, the data that is integrated in the FL path method can be directly output during the MD simulation, so the integration requires only a few seconds of computer time.

From the above considerations, it is perhaps not obvious which free energy method is computationally faster. Thus, we recorded the total computer time required to produce each MOF free energy in our database. These results are shown in Figure D.3. Note that we used Python scripts to calculate the final free energies from the data output from the MD simulations for both the QHA and the FL path, so (especially the QHA) calculations could probably be sped up by using a compiled programming language. We can see from Figure D.3 that QHA is approximately three times faster to implement that the FL path. The average time for QHA calculations is 167 core minutes and the average time for FL path calculations is 501 core minutes. Note that we define core minutes as the product of calculation time (in minutes) and the number of cores used for the calculation. We used 12 cores for each MD simulation, and a single core for each data processing step.

D.6 Additional Frenkel-Ladd Path/Quasi-Harmonic Approximation Comparison Data

We classified MOFs in our database as "low-frequency" or "high-frequency" depending on the extent that higher and lower frequency vibrations contribute to their total QHA vibrational free energy. One could naively classify a MOF as high-frequency only if over 50% of its energy came from relatively high-frequency (¿ 2669 cm^{-1}) vibrations, and as low-frequency if over 50% of its energy came from, say, mid-low-frequency (667-1334 cm^{-1}) vibrations. However, it would be difficult to directly compare correlations between QHA and FL free energies using this categorization scheme, because there would be different numbers of MOFs in each category—in particular, there would be fewer high-frequency MOFs.

Instead, we built our classification scheme based on the histograms shown in Figure 6.3b in the main text. Specifically, we classified a MOF as high-frequency if it was in the 95th percentile as per the "2669-3336 cm 1" histogram in Figure 6.3b. In other words, if a MOF was in the 95th percentile of

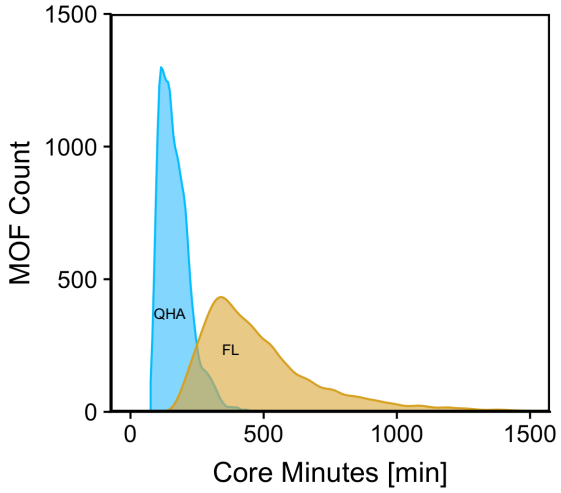

Figure D.3 Comparison of total calculation times in core minutes (defined as the product of calculation time and the number of cores used for the calculation) for QHA (blue) and FL (orange) free energies in our MOF database.

high-frequency contributions to its vibrational free energy we classified it as high-frequency. The cutoff to be in this 95th percentile is 31%, meaning that if a MOF had more than 31% of its vibrational free energy coming from high-frequency vibrations, we classified it as a "high-frequency" MOF. Analogously, we classified a MOF as low-frequency if it was in the 95th percentile of the "667-1134 cm^{-1}" distribution. In other words, if a MOF was in the 95th percentile of mid-low-frequency contributions to its vibrational free energy we classified it as "low-frequency" MOFs. The cutoff to be in this 95th percentile is 32%, meaning that if a MOF had more than 32% of its vibrational free energy coming from mid low-frequency vibrations, we classified it as low-frequency.

As stated in the main text the observed deviation between the QHA free energies and the FL free energies is likely due to overestimation of the vibrational energy under the harmonic approximation, and that this overestimation is higher for MOFs with more high energy modes. This is consistent with Figure D.6 where the deviation between QHA and FL energies correlates with the ZPE correction, which is larger the more high frequency modes a MOF has. Interestingly, while the errors are larger for high-frequency MOFs, the correlation is stronger for (which makes the error more predictable). Conversely, while the errors are smaller for low-frequency MOFs, the correlation is weaker (which makes the error less predictable). On the other hand, if the QHA free energies are calculated using vibrational spectra without the highest frequency C-H/O-H vibrations, we would expect to see significantly less deviation between QHA and FL free energies. Our expectation is confirmed by Figure D.7, which shows how the QHA versus FL path free energy relationship becomes significantly more linear when integrating each PDOS only to

2669 cm^{-1}.

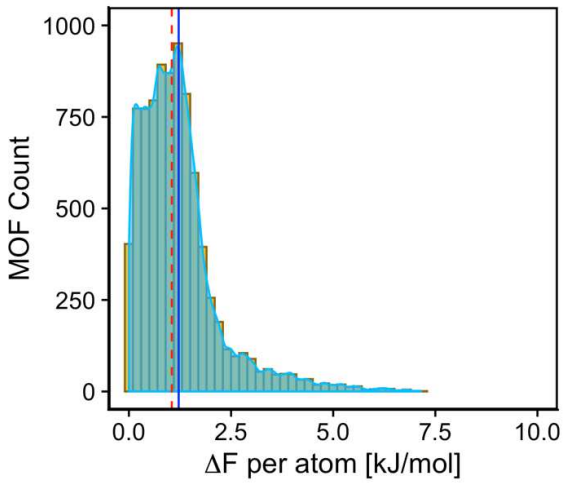

Figure D.4 Histogram of the absolute deviation between QHA and FL free energies (ΔF) after removing the apparent QHA systematic error. The dashed and solid vertical lines indicate the median and absolute deviations, respectively.

D.7 Temperature Behavior of QHA and FL Free Energies

The QHA accounts for the temperature dependence of free energy in three ways: (*i*) the equilibrium lattice shape changes with temperature, resulting in changes to U_0 and the PDOS, (*ii*) the occupancy of vibrational states changes with temperature, which alters the PDOS, and (*iii*) the explicit temperature dependence of Eq. 1 in the main text. Generally, the changes to U and the PDOS are small across the range of temperatures we consider, meaning the temperature dependence of QHA free energies depends primarily on the third term of Equation 6.1, which decreases with temperature. Consequently, the QHA free energies of the MOFs shown in Figure D.8 increase with temperature up to around 100 K and then begin to decrease, whereas the FL free energies only begin decreasing after around 350 K, depending on the MOF. This means that we expect higher systematic deviation between QHA and FL path free energies at lower temperatures, which is the case when comparing 100 K calculations to 300 K calculations, as can be observed from Figure D.9.

D.8 Mass Weighting Velocities to Calculate Phonon Density of States

Phonon density of states (PDOS) can be calculated using velocities or mass-weighted velocities. For homogeneous systems atom-wise both methods yield the same results. For heterogenous systems, however, only the PDOS calculated with mass-weighted VACFs can be compared directly to those obtained from normal mode analysis under the harmonic approximation [328], which is a method frequently employed to

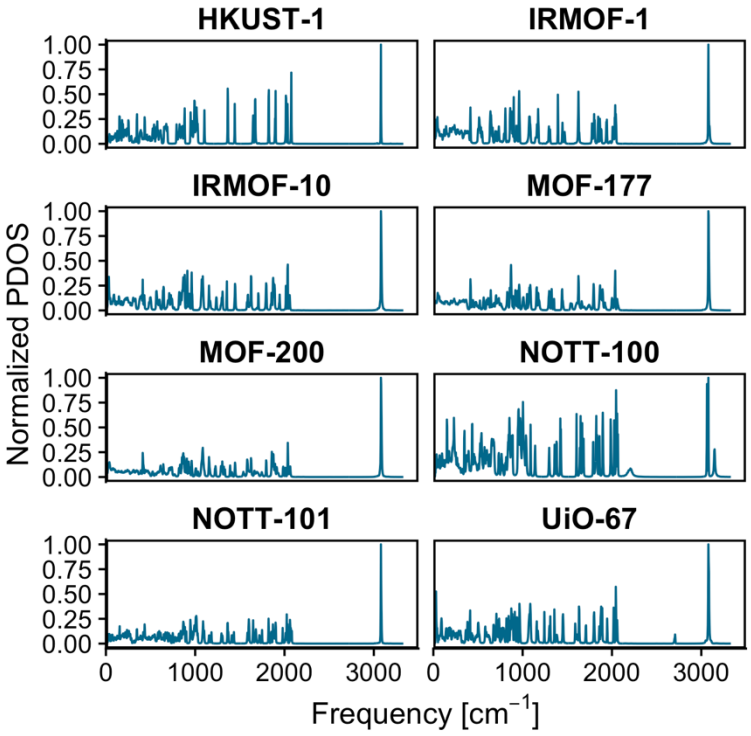

Figure D.5 Phonon density of states (PDOS) for select previously synthesized MOFs. Each PDOS was individually normalized to vary between 0 and 1 before plotting.

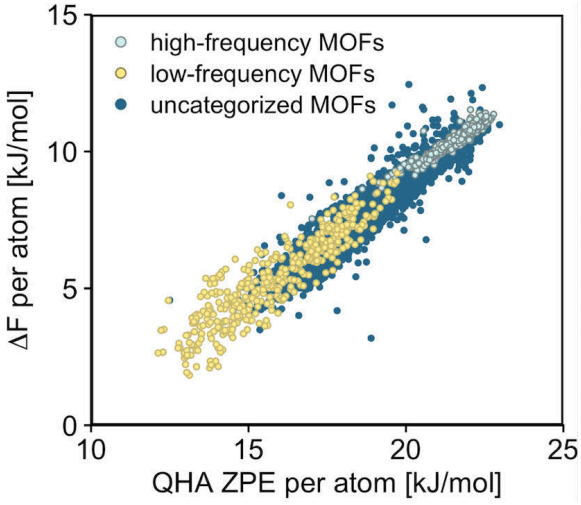

Figure D.6 Absolute Error for QHA free energy relative to FL free energy in all MOFs in our database. High-frequency and low-frequency MOFs are highlighted with different color. Note that larger errors but more "systematic" errors occur for high-frequency MOFs, while smaller but less "systematic" errors occur for low-frequency MOFs.

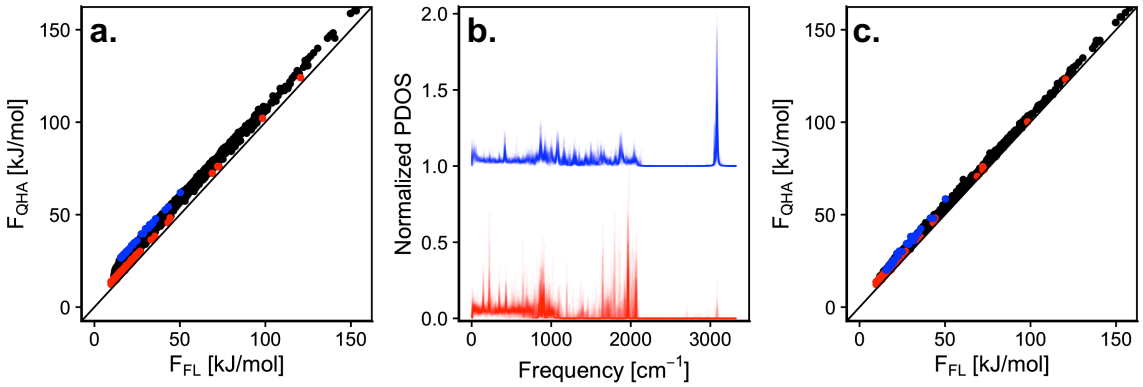

Figure D.7 (a) QHA free energies versus FL free energies for each MOF in our database. The 100 points farthest from the parity line are colored blue, while the 100 closest to the parity line are colored red. (b) Superimposed PDOS for each of the points shown in (a) with corresponding color. (c) QHA free energies versus FL free energies when using modes with frequency up to 2669 cm^{-1} in the QHA calculations.

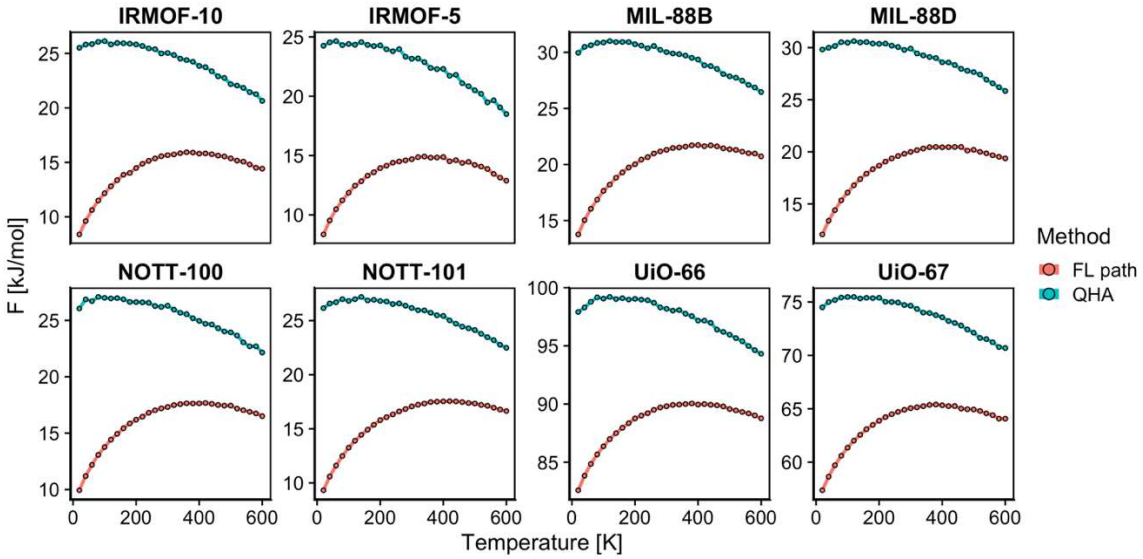

Figure D.8 QHA and FL free energies over a range of temperatures between 20 and 600 K for the indicated MOFs.

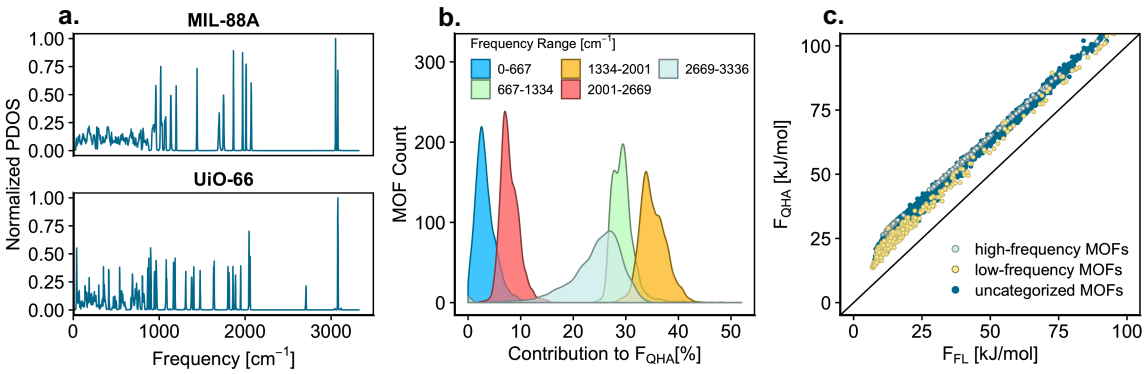

Figure D.9 Analogous to Figure 6.3 in the main text, but for QHA and FL free energies calculated at 100 K.

calculate MOF PDOS using DFT [333, 375]. This is because (*i*) the peak positions observed in the PDOS obtained from mass-weighted velocities correspond to the eigenvalues of the dynamical matrix under the harmonic approximation and (*ii*) the PDOS peak height varies only due to degeneracy in the observed frequencies when using mass-weighted velocities [328]. For a more detailed of the effect of mass-weighting when computing vibrational spectra from MD simulations the reader is encouraged to read ref. [328].

We have found that using non-mass-weighted VACFs to compute our PDOS results in both higher systematic and unsystematic error for the QHA versus FL free energy relationship. This is because, without mass weighting, the peaks associated with C-H/O-H bond stretching (i.e. those above 3000 cm-1) tend to be significantly higher than the lower-frequency peaks, resulting in higher vibrational energy calculated from the QHA. Figure D.10 is analogous to Figure 6.3 from the main text but with QHA free energies calculated using non-mass-weighted velocities. Notably, vibrations in the highest defined frequency range (2669-3336 cm^{-1}) contribute significantly more, on-average, to MOF vibrational energy (Figure D.11b) and there is a clearly defined group of MOFs closest to the parity line. This group corresponds to MOFs with no hydrogen atoms, as can be seen from Figure D.11.

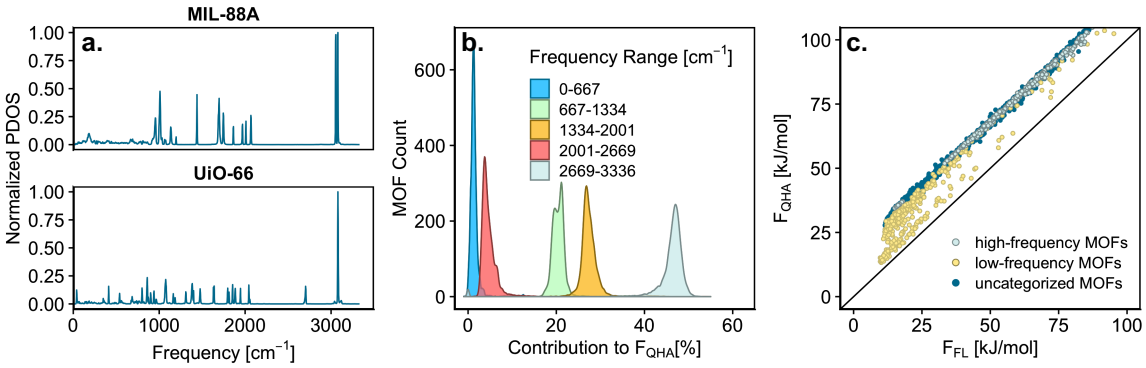

Figure D.10 Analogous to Figure 6.3 in in the main text but for PDOS and QHA free energies calculated using non-mass-weighted velocities.

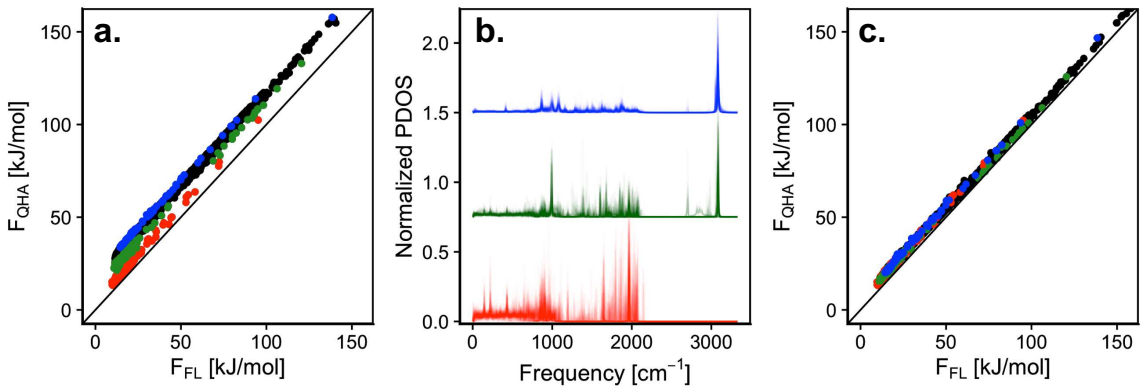

Figure D.11 Analogous to Figure D.6 but for QHA free energies calculated with non-mass-weighted velocities. (a) Shows three groups of MOFs each colored according to their distance from the parity line, blue corresponds to the 100 MOFs farthest from the parity line, green to the 100 MOFs nearest the average distance from the parity line, and red the 100 MOFs closest to the parity line. (b) shows the superimposed PDOS for all 100 MOFs in each category, colored accordingly, and (c) is analogous to (a) but with the QHA free energy calculated only using vibrations with frequencies less than 2669 cm^{-1}.

D.9 Supplementary Tables and Figures

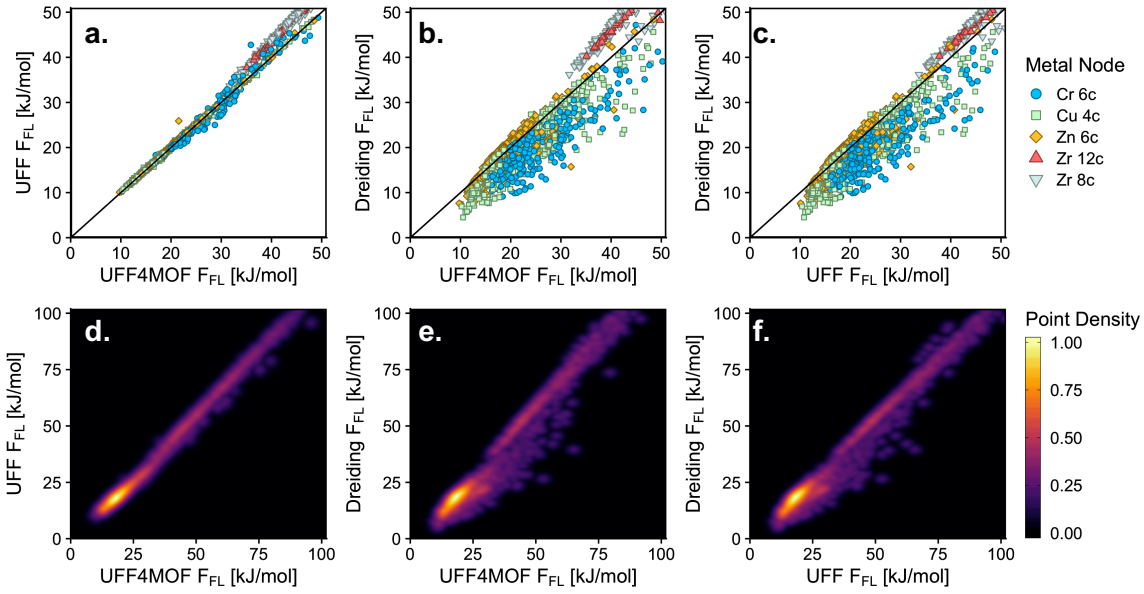

Figure D.12 (a-c) Parity plots comparing FL free energies (F_{FL}) calculated with different force fields. (a) UFF vs. UFF4MOF. (b) Dreiding vs. UFF4MOF. (c) Dreiding vs. UFF. Point shape and color correspond to the type of metal node in each MOF plotted. (d-f) Parity plots comparing FL free energies (F_{FL}) calculated with different force fields, illustrating the point density. (d) UFF vs. UFF4MOF. (e) Dreiding vs. UFF4MOF. (f) Dreiding vs. UFF. This figure is made using the same data as Figure 6.4 in the main text.

Table D.1 Strain energy for each metal node present in our database. These energies were calculated by optimizing the isolated metal node in LAMMPS using UFF4MOF. The carboxylate groups on each node where capped with hydrogen.

Inorganic Node	Strain Energy (kJ/mol)	Strain Energy Per Atom (kJ/mol)
Cu 4c	20	1
Cr 6c	126	5
Zn 6c	30	1
Zr 8c	1,986	28
Zr 12c	2,112	32

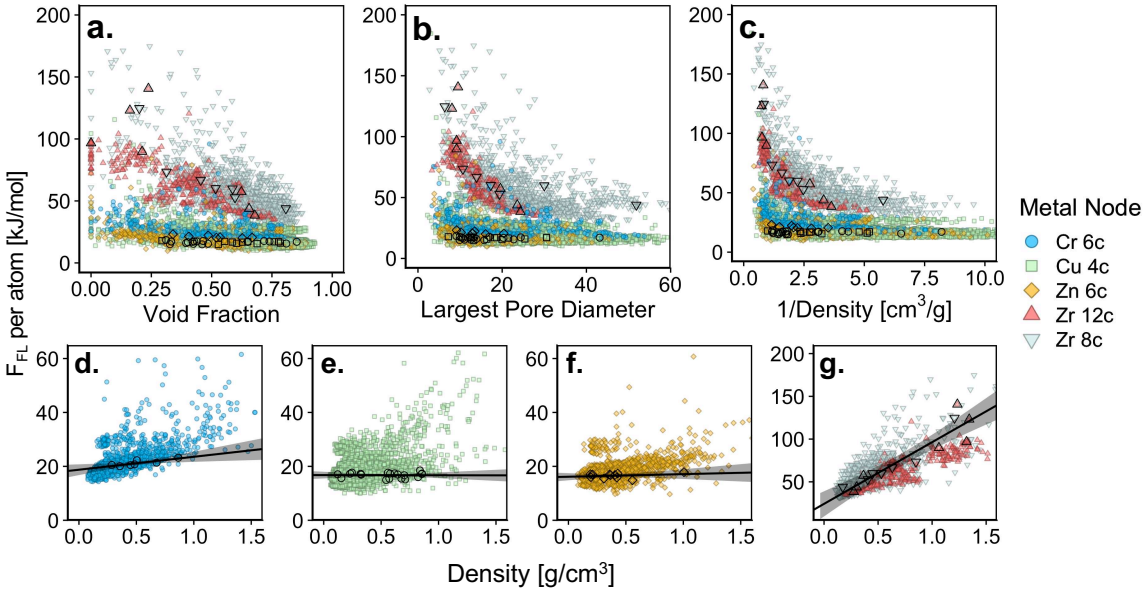

Figure D.13 FL free energies (F_{FL}) versus MOF properties. (a) F_{FL} vs. MOF void fraction. (b) F_{FL} vs. MOF largest pore diameter. (c) F_{FL} vs. inverse density. (d-f) breakdown of F_{FL} vs. density plot (Figure 6.5a in the main text) by metal node. d) 6c Cr MOFs. e) 4c Cu MOFs. (f) 6c Zn MOFs. (g) 8/12c Zr MOFs. Point features indicate the inorganic node in the MOF. Points outlined in black correspond to previously synthesized MOFs found in our 8,500-MOF database. Black lines in plots (d-f) are linear fits for the free energies of previously synthesized MOFs of each inorganic node type, and shaded regions correspond to the 95% confidence interval of each linear fit.

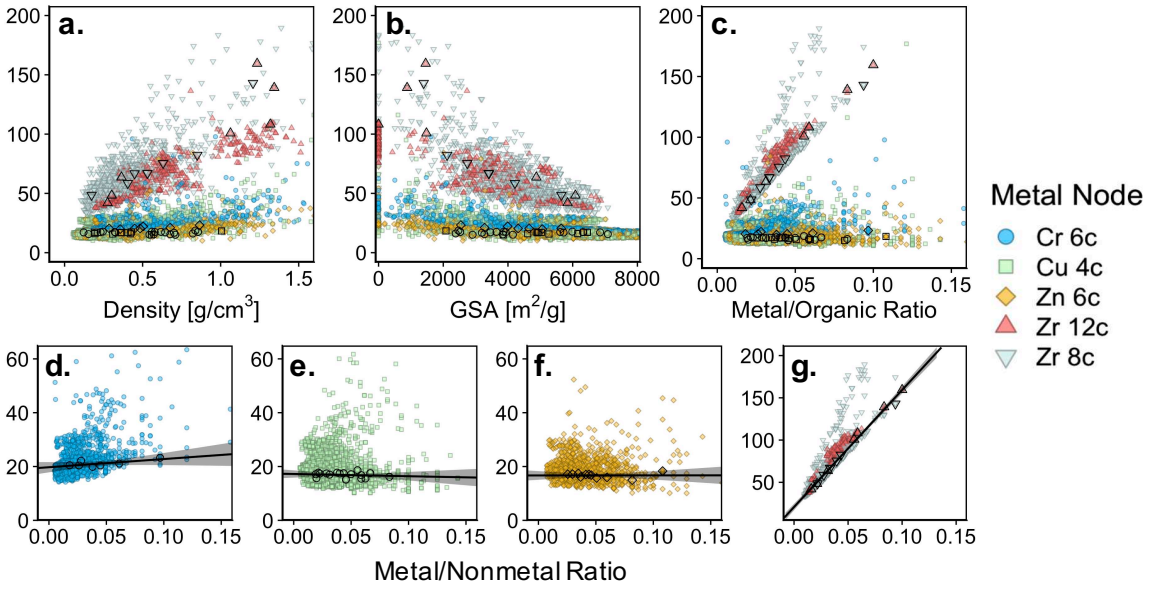

Figure D.14 Analogous to Figure 6.5 but for free energies calculated using the UFF.

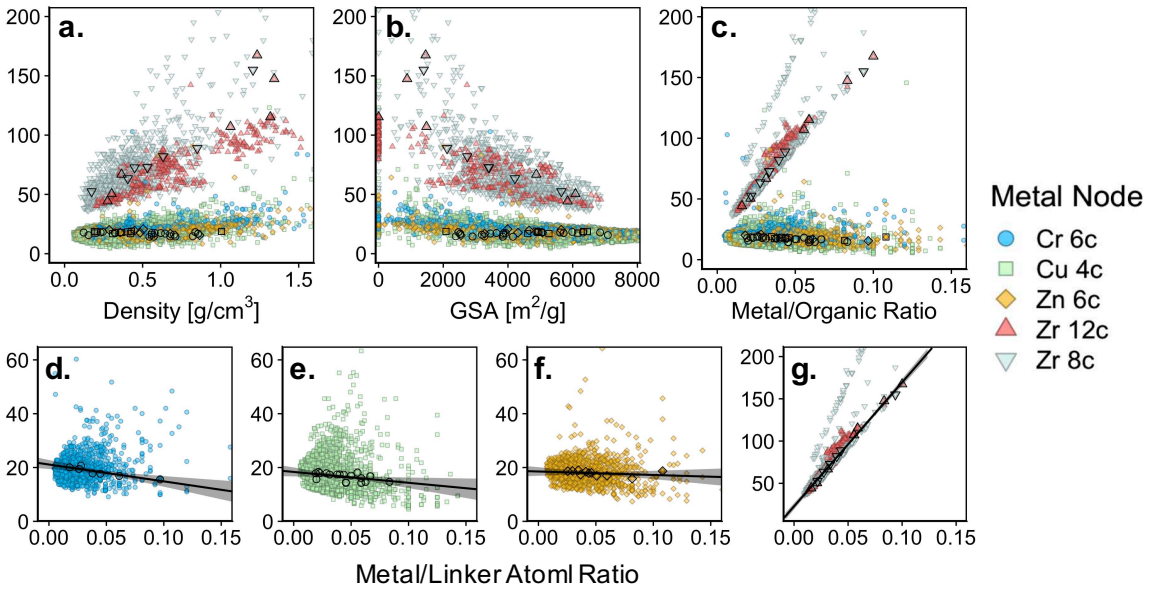

Figure D.15 Analogous to Figure 6.5 but for free energies calculated using the Dreiding forcefield.

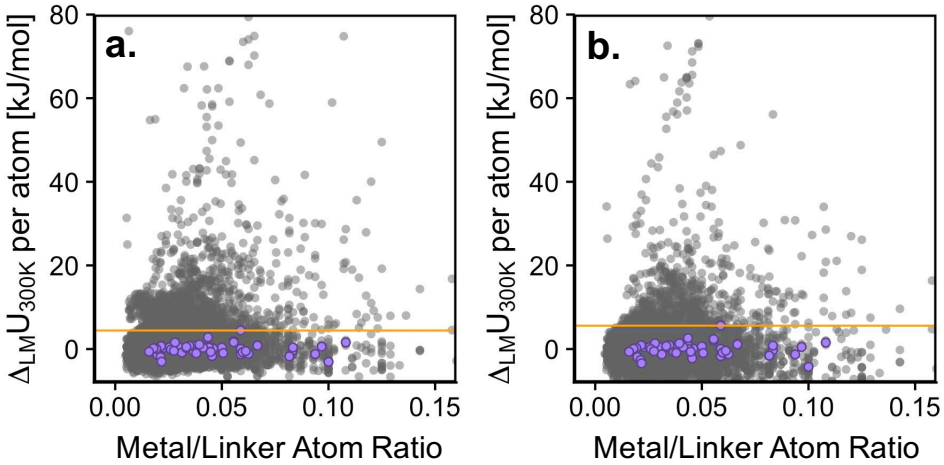

Figure D.16 Analogous to Figure 6.6 but for for UFF (a) free energies and Dreiding (b) free energies.

Table D.2 The number of MOFs with $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ higher than the highest $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ among the previously synthesized MOFs in our database for both an absolute cutoff and a cutoff with a 20% tolerance for UFF4MOF, UFF, and Dreiding free energies.

Force field	Number of MOFs	
	Absolute cutoff	20% tolerance
UFF4MOF	1,613	1,356
UFF	1,502	1,257
Dreiding	770	616

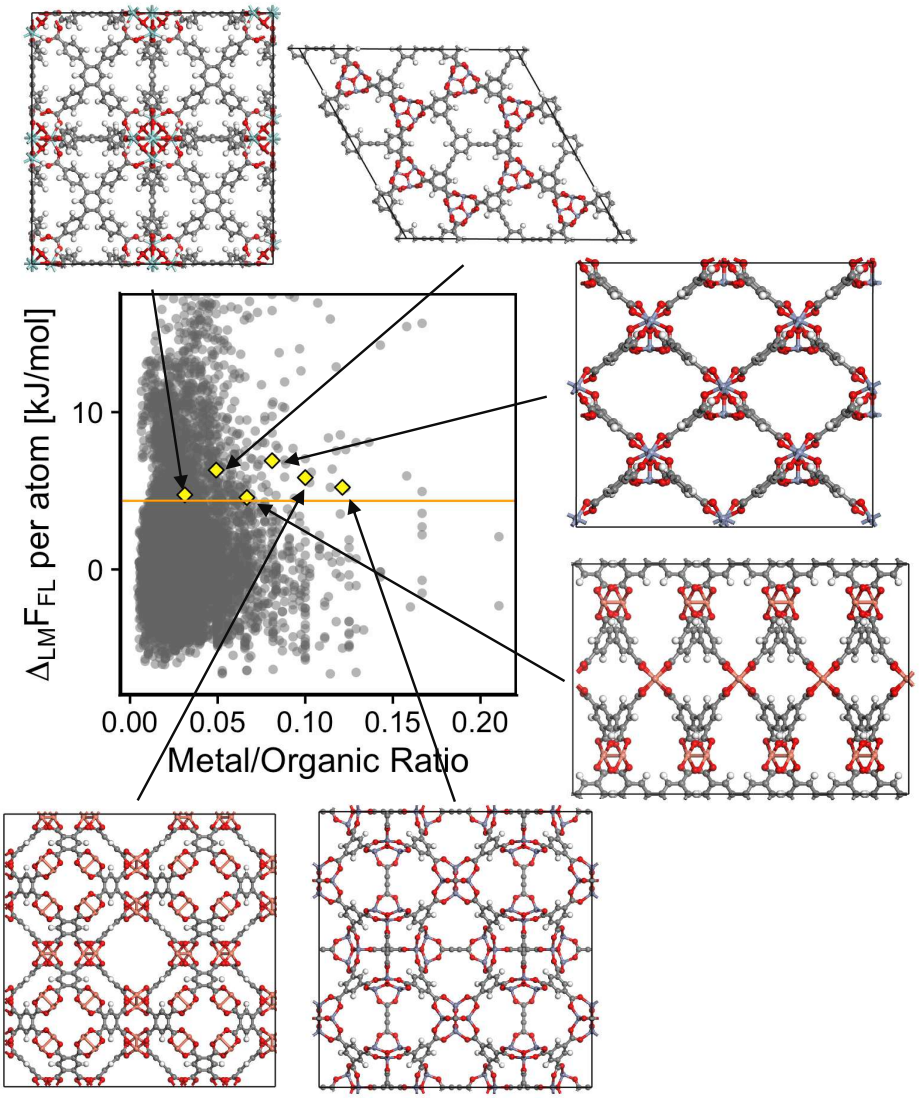

Figure D.17 Example MOFs selected from outside the "synthesizability" region of the plot of relative free energy ($\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$) versus MOF metal to linker atoms ratio. $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ is obtained by subtracting the free energy from the linear models shown in Figure 6.5d-g from the corresponding absolute from F_{FL}. The orange line delimits the region where the $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ of the 54 previously synthesized MOFs (purple points) in our 8,500-MOF database are found below.

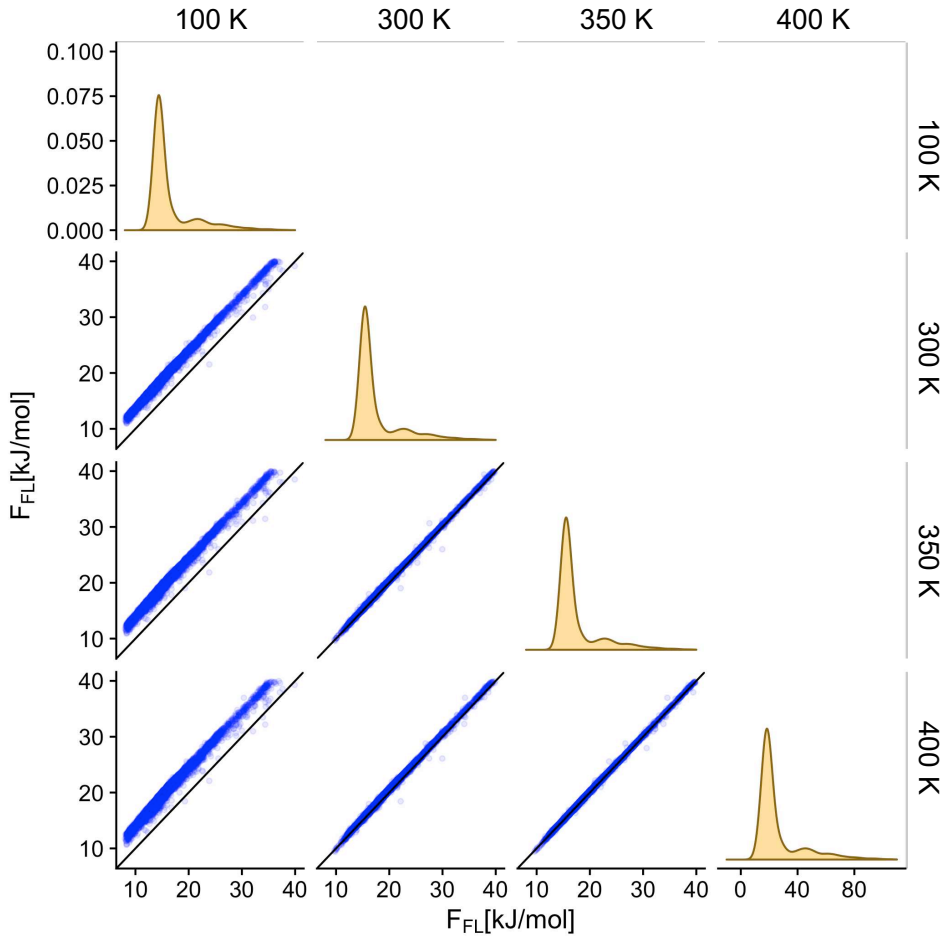

Figure D.18 Parity plots of FL path free energy calculated at 100, 300, 350, and 400 K, point densities for each temperature are shown on the leading diagonal. Given the close correlation of free energies at 300, 350, and 400 K Figure 6.5 and Figure 6.6 in the main text will not appreciably change across these temperatures.

Table D.3 List of the experimentally observed MOF isomorphs for the 32 isomorphic MOF series discussed in Chapter 6. The experimentally observed MOF is described by the "Linker" (as per Figure D.20), "Metal" and "Observed topology" columns. The remaining columns indicate the topology that it is predicted on the basis of having lowest free energy (F_{FL}), potential energy ($U_{300\mathrm{K}}$), and strain energy ($U_{0\mathrm{K}}$). Cases where the predicted topology differ from the experimentally observed one include a percent value in parenthesis, which indicates the deviation between the F_{FL} (or $U_{300\mathrm{K}}$ or $U_{0\mathrm{K}}$) of the predicted and experimentally observed topologies [356]

Linker	Metal	Observed topology	F_{FL} predicted topology	$U_{300\mathrm{K}}$ predicted topology	$U_{0\mathrm{K}}$ predicted topology
L3	Cr	acs	acs	acs	acs
L2	Cr	acs	acs	acs	acs
L5	Cr	acs	acs	acs	acs
L4	Cr	acs	acs	acs	acs
L1	Zr	bcu	bcu	bcu	bcu
L22	Zr	csq/scu	csq	csq	csq
L23	Zr	csq	csq	csq	csq
L15	Cu	fof	fof	fof	fof
L14	Cu	fof	fof	fof	fof
L13	Cu	fof	fof	fof	fof
L19	Cu	fof	stx (2.5%)	stx (1.2%)	stx (12.6%)
L17	Cu	fof	stx (0.4%)	stx (0.3%)	stx (16.9%)
L16	Cu	fof	fof	fof	bbf (7.4%)
L12	Cu	fof	fof	fof	fog (0.2%)
L25	Cr	jjt	jjt	jjt	jjt
L12	Zr	kle	kle	kle	kle
L3	Zn	pcu	pcu	pcu	pcu
L7	Zn	pcu	pcu	pcu	pcu
L5	Zn	pcu	pcu	pcu	pcu
L4	Zn	pcu	pcu	pcu	pcu
L6	Zn	pcu	pcu	pcu	pcu
L9	Cu	pto-c	pto	fjh (0.7%)	pto
L10	Cu	pto-c	pto	pto	pto
L9	Zn	qom	pyr(0.5%)	pyr (1.2%)	pyr (3.0%)
L11	Zn	qom	apo (0.4%)	pyr (1.1%)	pyr (0.5%)
L8	Zn	rtl	pyr (1.1%)	pyr (3.3%)	pyr (4.7%)
L20	Zr	scu	csq (0.4%)	csq (0.3%)	csq (1.2%)
L18	Cu	stx/fof	stx	stx	stx
L21	Cu	sur	sur	sur	bbf (2.5%)
L11	Cu	tbo	tbo	tbo	fjh (8.1%)
L8	Cu	tbo	tbo	tbo	tbo
L23	Cu	zyg	zyg	zyg	zyg

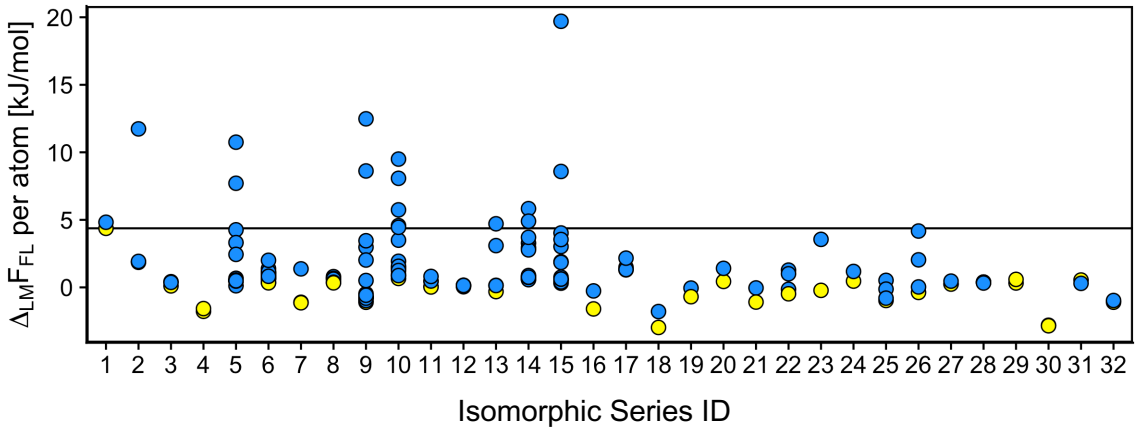

Figure D.19 The spread of $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ values within the 32 studied isomorphic MOF series. The yellow points correspond to the experimentally observed isomorph for each series. The horizontal, red line corresponds to the upper boundary in $\Delta_{\mathrm{LM}}F_{\mathrm{FL}}$ above which MOFs have low synthetic likelihood. Accordingly, note that many MOFs that would be considered stable may not be experimentally observed, simply because a more stable isomorph is more stable.

Table D.4 Experimental isomorph identification metrics for free and potential energy at different temperatures (strain energy metrics do not change with temperature).

Temperature [K]	# of experimentally observed isomorphs corresponding to the lowest energy isomorph		# of experimentally observed isomorphs with energy within 1% of the lowest energy isomorph		Max percent difference observed between experimentally observed isomorph and the isomorph with lowest energy	
	Free	Potential	Free	Potential	Free	Potential
100	23	22	29	27	2.9%	4.1%
300	26	25	29	28	2.5%	3.3%
350	26	24	30	31	2.0%	3.0%
400	23	21	30	29	2.0%	4.1%

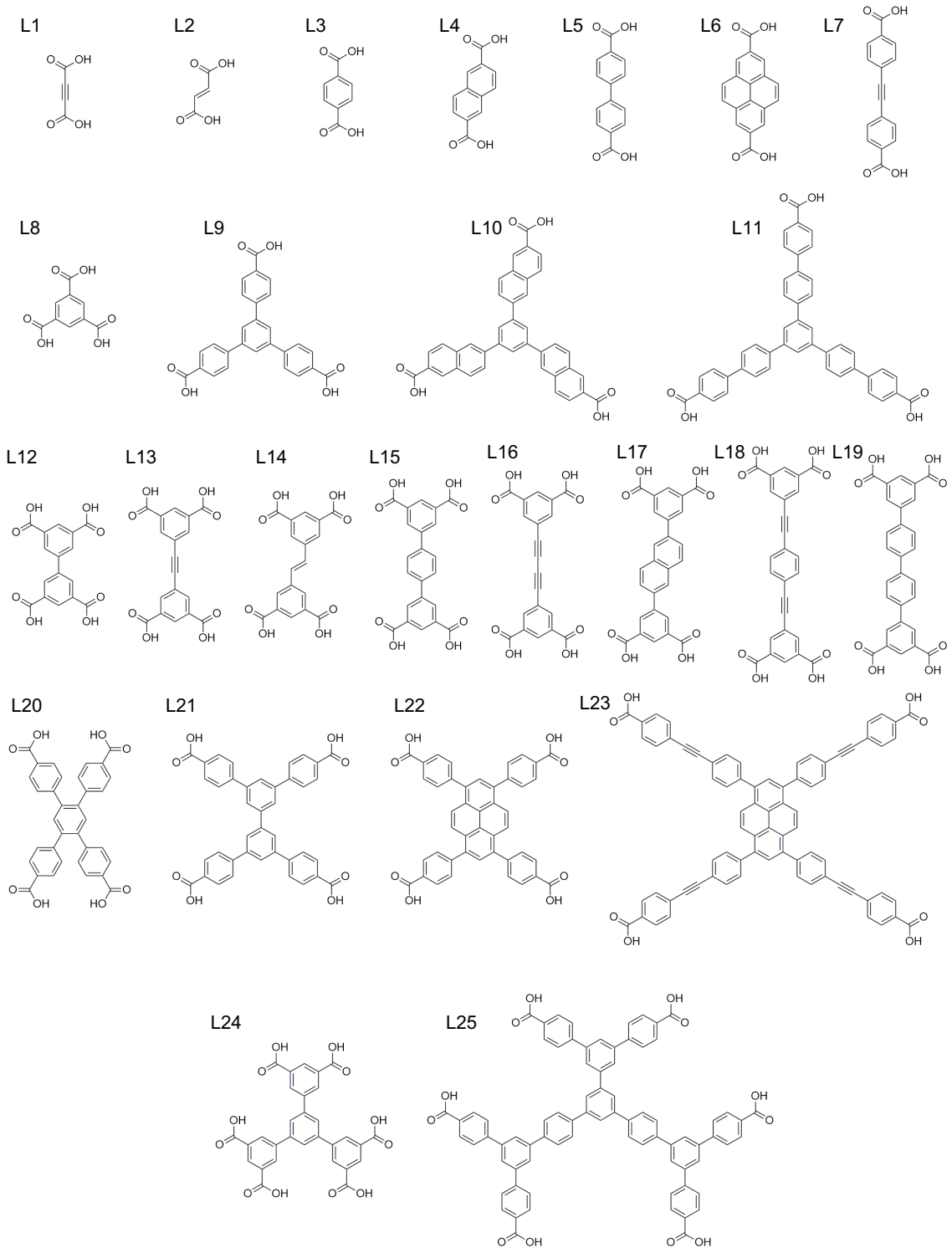

Figure D.20 The linkers observed in the experimentally observed isomorphs for the 32 isomorphic MOF series discussed in Chapter 6.

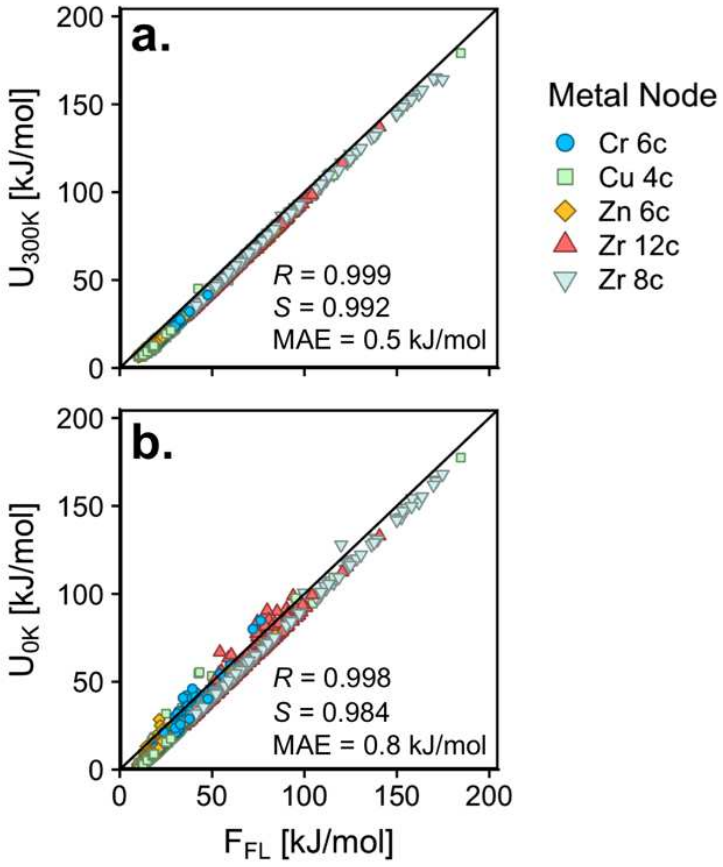

Figure D.21 Parity plots comparing potential energy at 300 K ($U_{300\mathrm{K}}$) and 0 K ($U_{0\mathrm{K}}$, strain energy) with free energy F_{FL} (at 300 K) (a) $U_{300\mathrm{K}}$ vs $\mathrm{F_{FL}}$. (b) $U_{0\mathrm{K}}$ vs $\mathrm{F_{FL}}$. Point features indicate the type of metal node. Insets are the Pearson (R) and Spearman Ranking (S) correlation coefficients, and the mean absolute error (MAE).

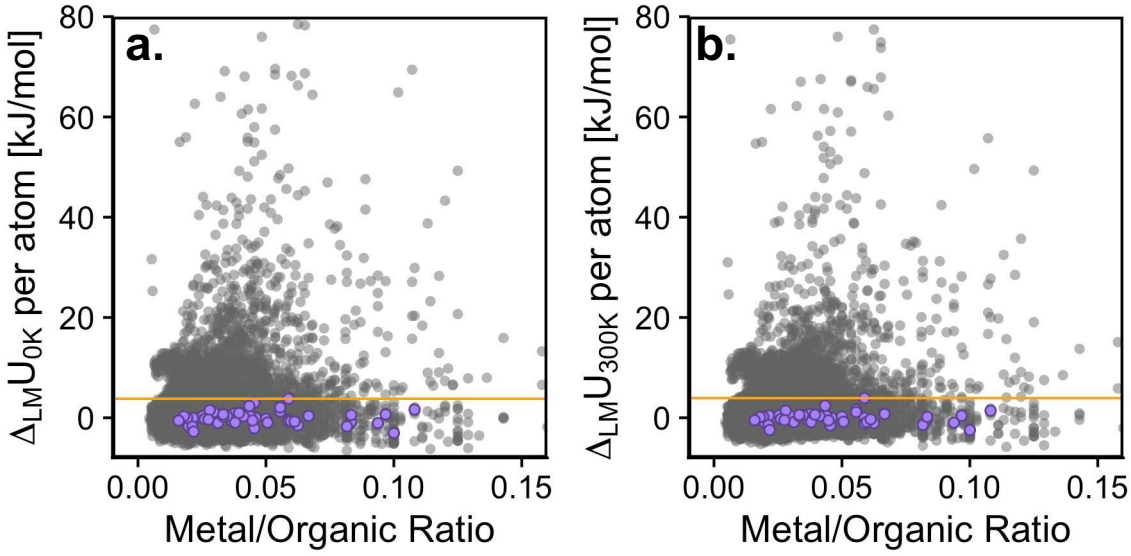

Figure D.22 The region of synthetic likelihood outlined in Chapter 6 but using (a) strain energy and (b) potential energy at 300 K

APPENDIX E

SUPPORTING INFORMATION FOR CHAPTER 7

E.1 Model Training Details

We used the same alchemical adsorbates and adsorbate descriptors used in Chapter 5 for the training and validation of our MLP. These alchemical adsorbates use the same "template" adsorbate models shown in Figure C.1 but have their parameters (ϵ_{ii}, σ_{ii}, q, and r_{bond}) assigned on a grid. Please refer to Chapter 5 and Appendix C for complete details on the alchemical adsorbate forcefield parameters used to generate GCMC training/validation data and the adsorbate descriptors used as input to our MLP. We used 14 fugacities for training, 13 fugacities for validation and 15 fugacities for testing for all MOF/adsorbate combinations; these fugacities are listed in Table E.1.

We trained our MLP for predicting single component isotherms using the TensorFlow Python package. We used the same topology for our MLP as in Chapter 5. The hyperparameters explicitly set during training are given in Table E.3, for any hyperparameters not listed in Table S5 the defaults were used. All code versions used are listed in Table E.4 (note that the hyperparameter defaults may change between versions). The learning curve for our MLP is shown in Figure E.1.

All descriptors input to the model were scaled to vary between 0 and 1 (min/max scaling) according to the range of values present in the training set. Early termination with a patience of 20 epochs was used to avoid overfitting. We used an initial learning rate of 2×10^{-4}, which was halved every time the validation loss did not improve for five consecutive epochs (up to a minimum learning rate of 1×10^{-5}. A maximum of 500 epochs was allowed (this maximum was never reached).

Table E.1 The fugacities used for GCMC training, validation, and test data for our MLP trained to predict single component isotherms.

Training Fugacities [bar]	Validation Fugacities [bar]	Testing Fugacities [bar]
0.01	0.02	0.02
0.025	0.04	0.04
0.05	0.06	0.06
0.075	0.08	0.08
0.10	0.2	0.2
0.25	0.4	0.4
0.50	0.6	0.6
0.75	0.8	0.8
1.0	2.5	1.0
5.0	30	2.5
10	60	5.0
50	80	10
75	90	25
100	-	50
-	-	100

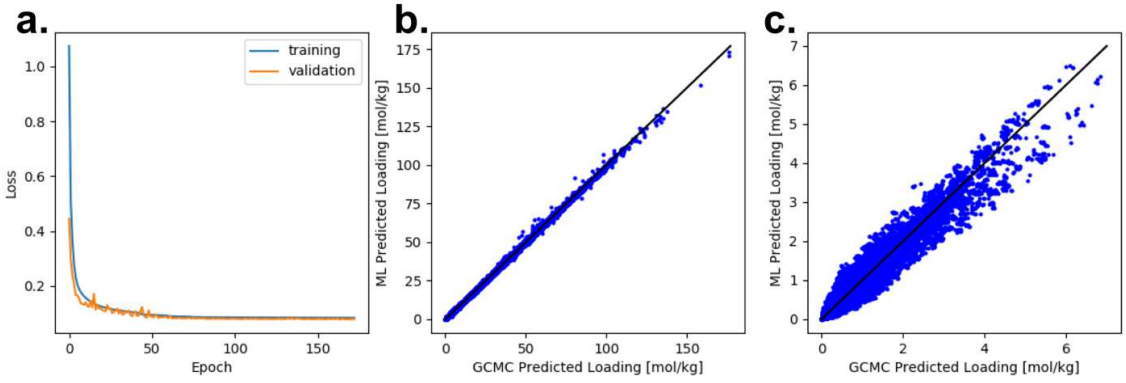

Figure E.1 (a) Learning curve for our MLP trained to predict single component isotherms. (b) Parity plot of our model predictions versus GCMC values for our validation set (all points shown). (c) is a zoomed-in version of (b), only showing predicted loadings up to 7 mol/kg.

Table E.2 The MOF fingerprint used to train our MLP for predicting single-component isotherms. The first 20 descriptors shown are MOF chemical motif number densities. Each motif is described using the atom type nomenclature used by UFF. To obtain the number density of the motif, the count of each atom type was normalized by the MOF volume in $Å^3$. The last six MOF descriptors listed are MOF textural properties.

MOF Descriptor	Brief Description
Br	the element bromine
C_1	sp carbon
C_2/C_R	sp2 or aromatic carbon (we grouped these together)
C_3	sp3 carbon
Cr6f3	UFF4MOF type of Cr for the cluster we consider
Cu4+2	UFF4MOF type of Cu for the cluster we consider
F_	the element fluorine
H_	the element hydrogen
N_1	sp nitrogen
N_2/N_R	sp2 or aromatic nitrogen (we grouped these together)
N_3	sp3 nitrogen
O_1	sp oxygen
O_2_M	carboxylate oxygen bound to a metal
O_2_z	UFF4MOF type for the Cr cluster oxygen (bound to three Cr)
O_3	sp3 oxygen
O_3_M	sp3 oxygen bound to a metal
O_3_f	UFF4MOF type for the Zr/Zn cluster oxygen (bound to three Zr or four Zn)
S_3	sp3 sulfur
Zn3f2	UFF4MOF type of Zn for the cluster we consider
Zr8f4	UFF4MOF type of Zr for the cluster we consider
LPD	Largest Pore Diameter [Å]
PLD	Pore Limiting Diameter [Å]
Gravimetric SA	Gravimetric Surface Area [m^2/g]
VF	Void Fraction
PSSD	Pore Size Standard Deviation
Density	Density g/cm^3

Table E.3 The hyperparameters used during the training of our MLP for predicting single component isotherms. Topology is denoted by $N_{\ldots-N_M}$, where Ni is the number of nodes in each hidden layer.

Hyperparameter	Value
Topology	50-50-20-20
Activation Function	sigmoid
Optimizer	Adam
Learning Rate	0.0002
Batch Size	50
Loss	Mean Absolute Error (MAE)

Table E.4 Versions for all the Python packages used for our MLP training and IAST calculations.

Package	Version
python	3.7.7
numpy	1.19.2
pyiast	1.4.3
scikit-learn	0.19.1
tensorflow	1.14.0

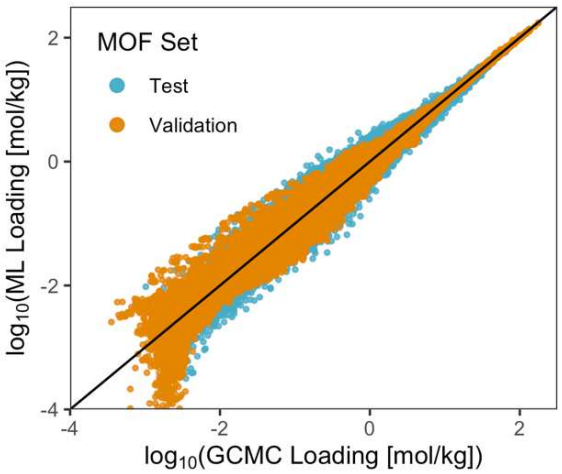

Figure E.2 Comparison of our MLP predictions of the validation and test sets on a log-log scale.

E.2 Supplementary Figures

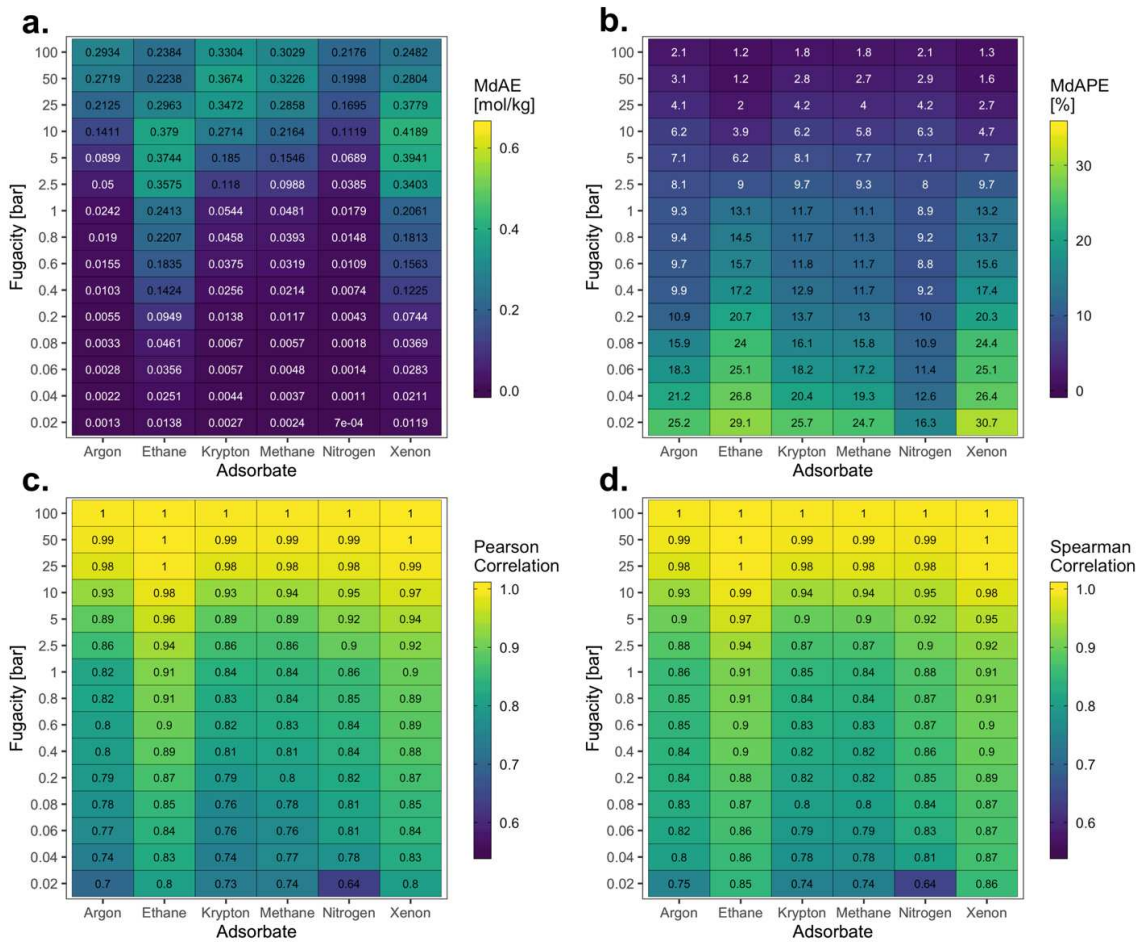

Figure E.3 Error metrics for all the adsorbate/fugacity combinations for our single component isotherm testing data. Each error metric was calculated for all 1,000 test set MOFs. (a) Median Absolute Errors (MdAEs), (b) Median Absolute Percentage Errors (MdAPEs), (c) Pearson correlation coefficients, and (d) Spearman rank correlation coefficients.

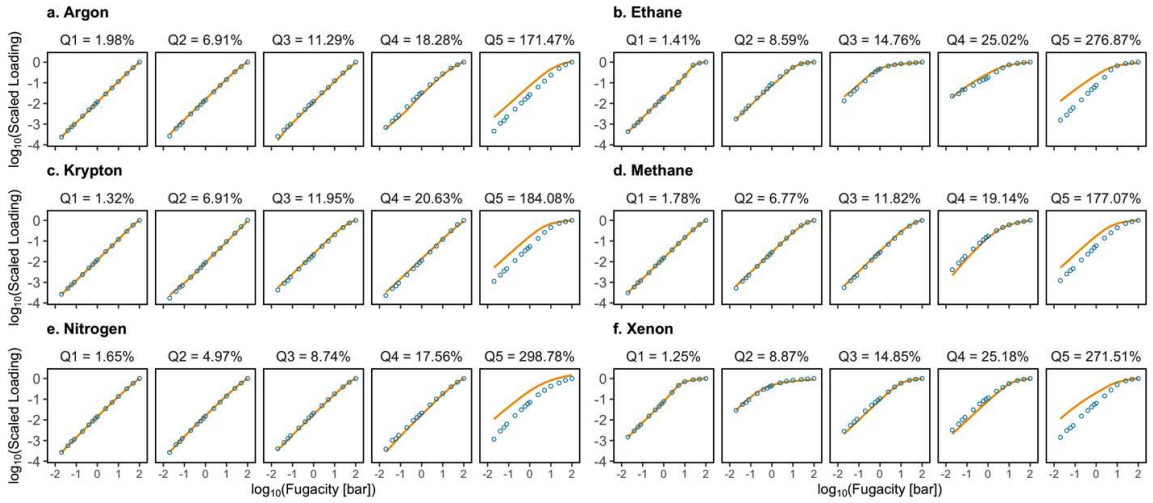

Figure E.4 Comparison of our MLP-predicted isotherms for the six indicated adsorbates to GCMC simulated isotherms on a log-log scale for the 0.00 (Q_1), 0.25 (Q_2), 0.50 (Q_3), 0.75 (Q_4), and 1.00 (Q_5) quantiles of the IMAPE. The open, blue points are GCMC simulated values and the orange lines are MLP predictions.

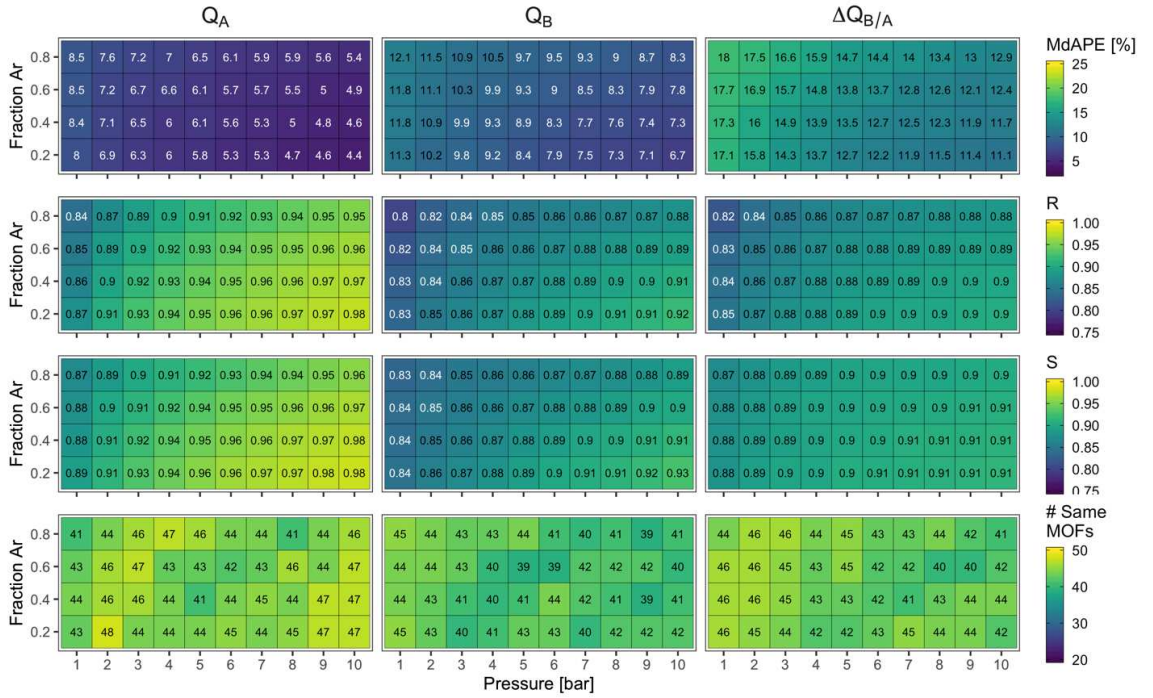

Figure E.5 Error metrics for Ar/Kr binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{B/A}$.

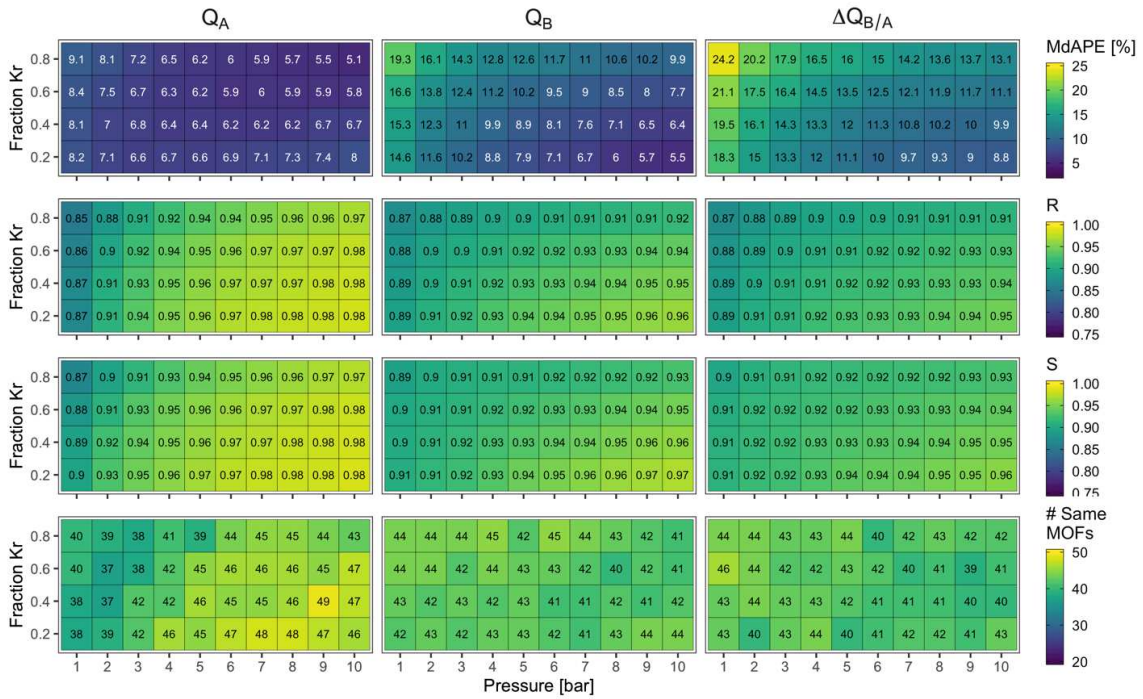

Figure E.6 Error metrics for Kr/Xe binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{B/A}$.

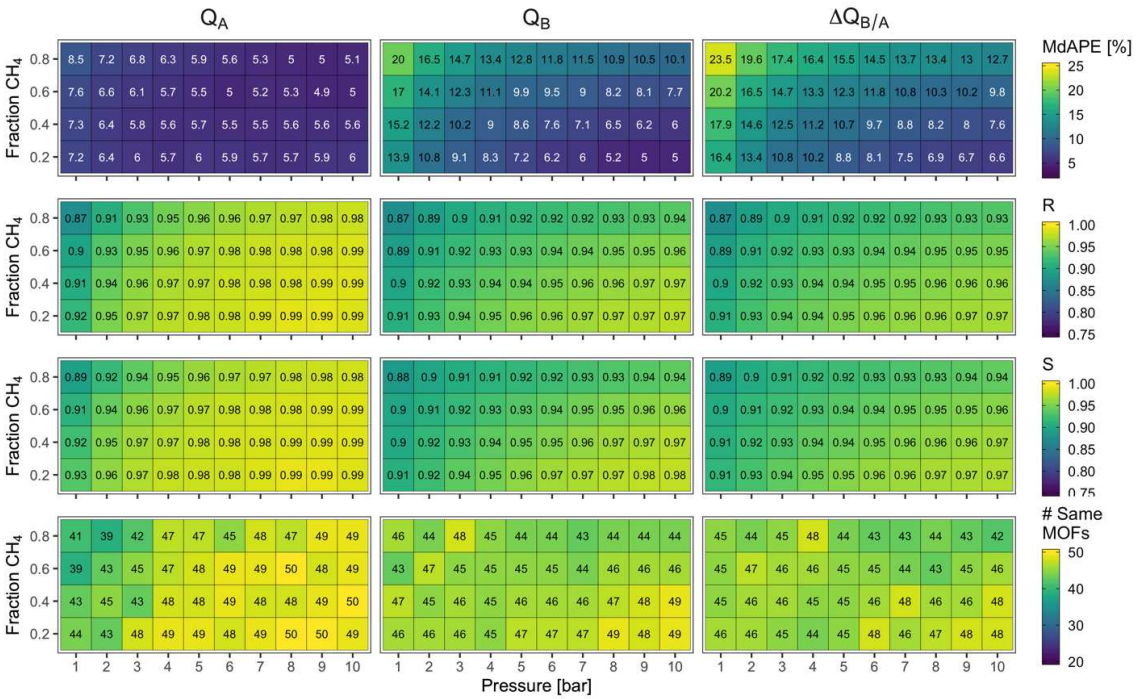

Figure E.7 Error metrics for CH_4/C_2H_6 binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{B/A}$.

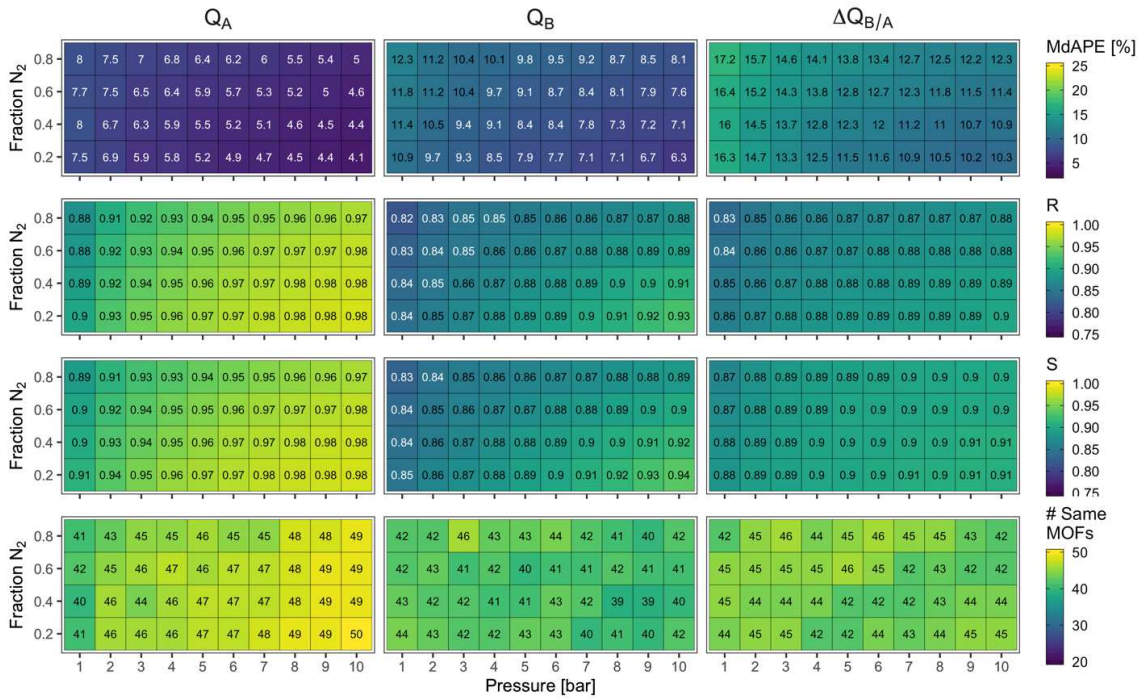

Figure E.8 Error metrics for N_2/CH_4 binary adsorption data. The MdAPE, Pearson correlation (R), Spearman correlation (S), and the number of top-50 GCMC predicted MOFs observed in the top-100 MLP+IAST predicted MOFs are shown for the loading of component A (Q_A), the loading of component B (Q_B), and $\Delta Q_{B/A}$.

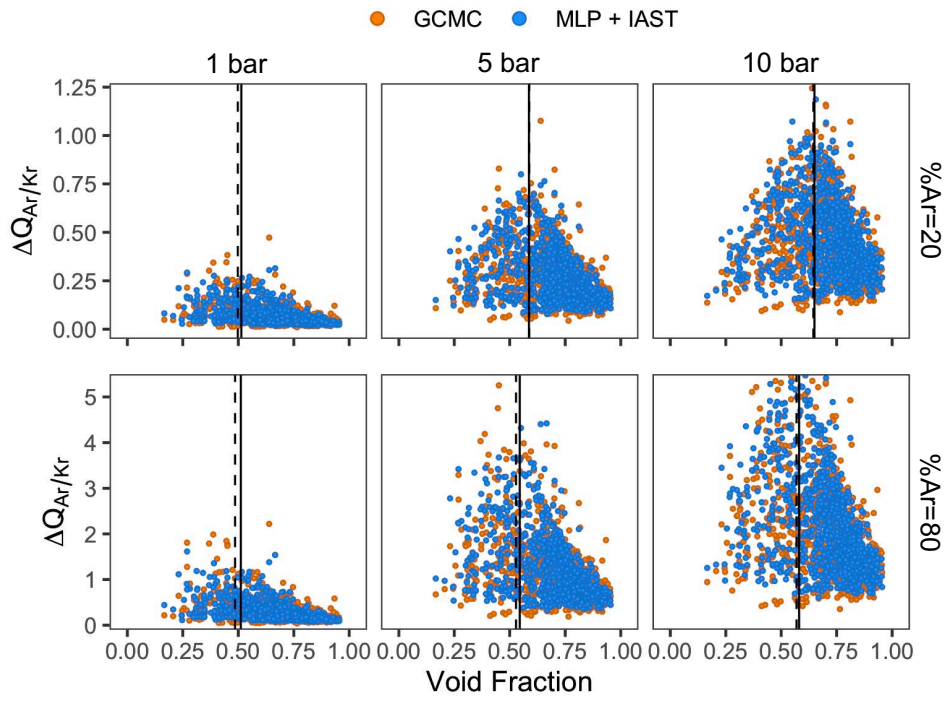

Figure E.9 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for Ar/Kr. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line.

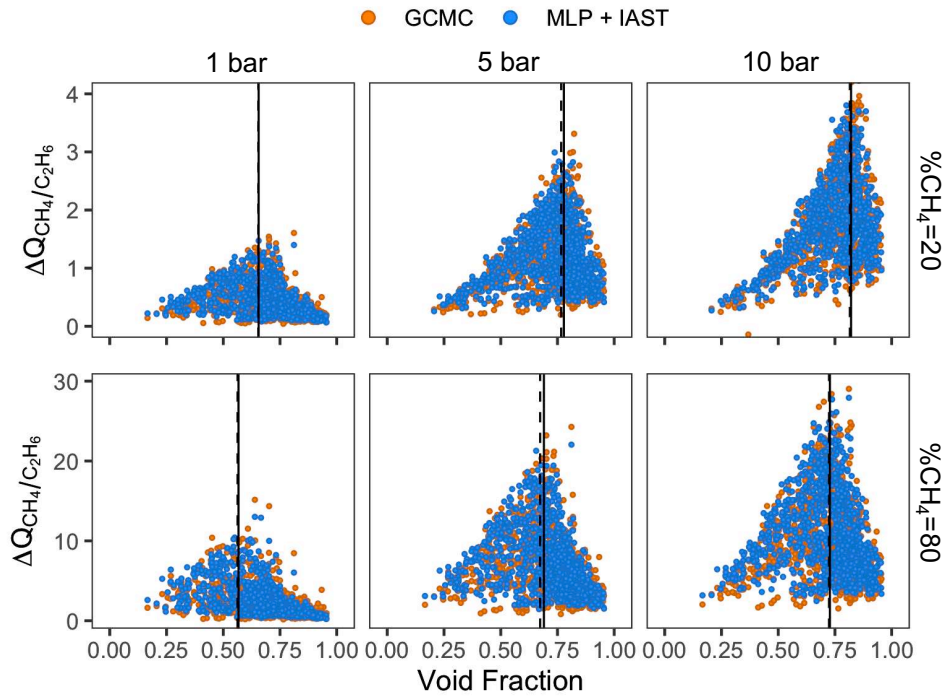

Figure E.10 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for CH2/C_2H_6. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line.

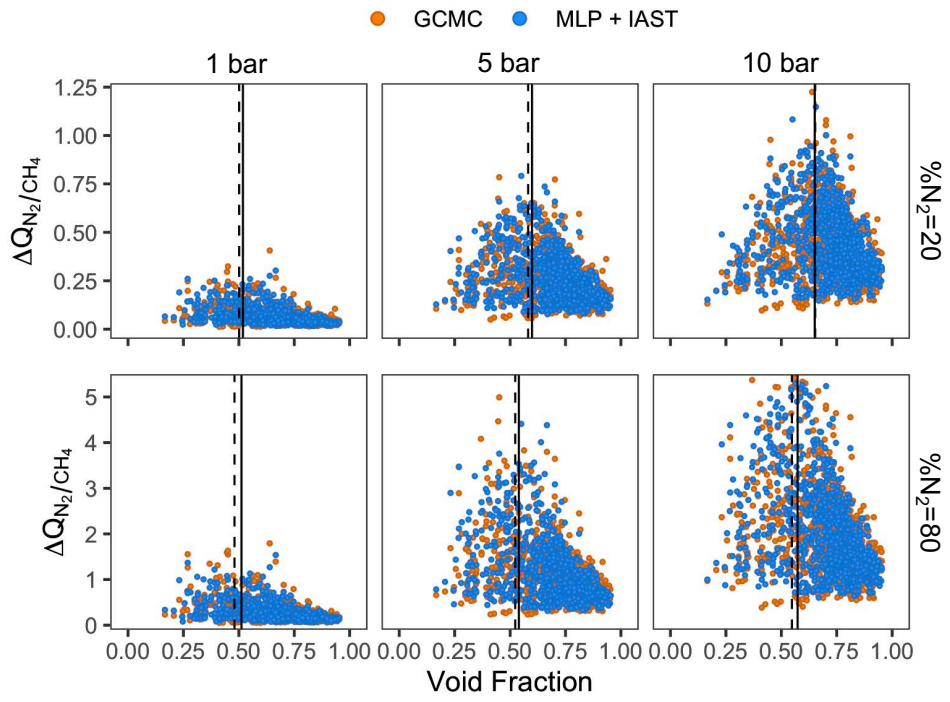

Figure E.11 MLP+IAST and GCMC ΔQs versus MOF void fraction at the indicated conditions for N_2/CH_4. The mean VF of MOFs in the 95th percentile of GCMC predicted ΔQ is shown by the solid line. The mean VF of MOFs in the 95th percentile of MLP+IAST predicted ΔQ is shown by the dashed line.

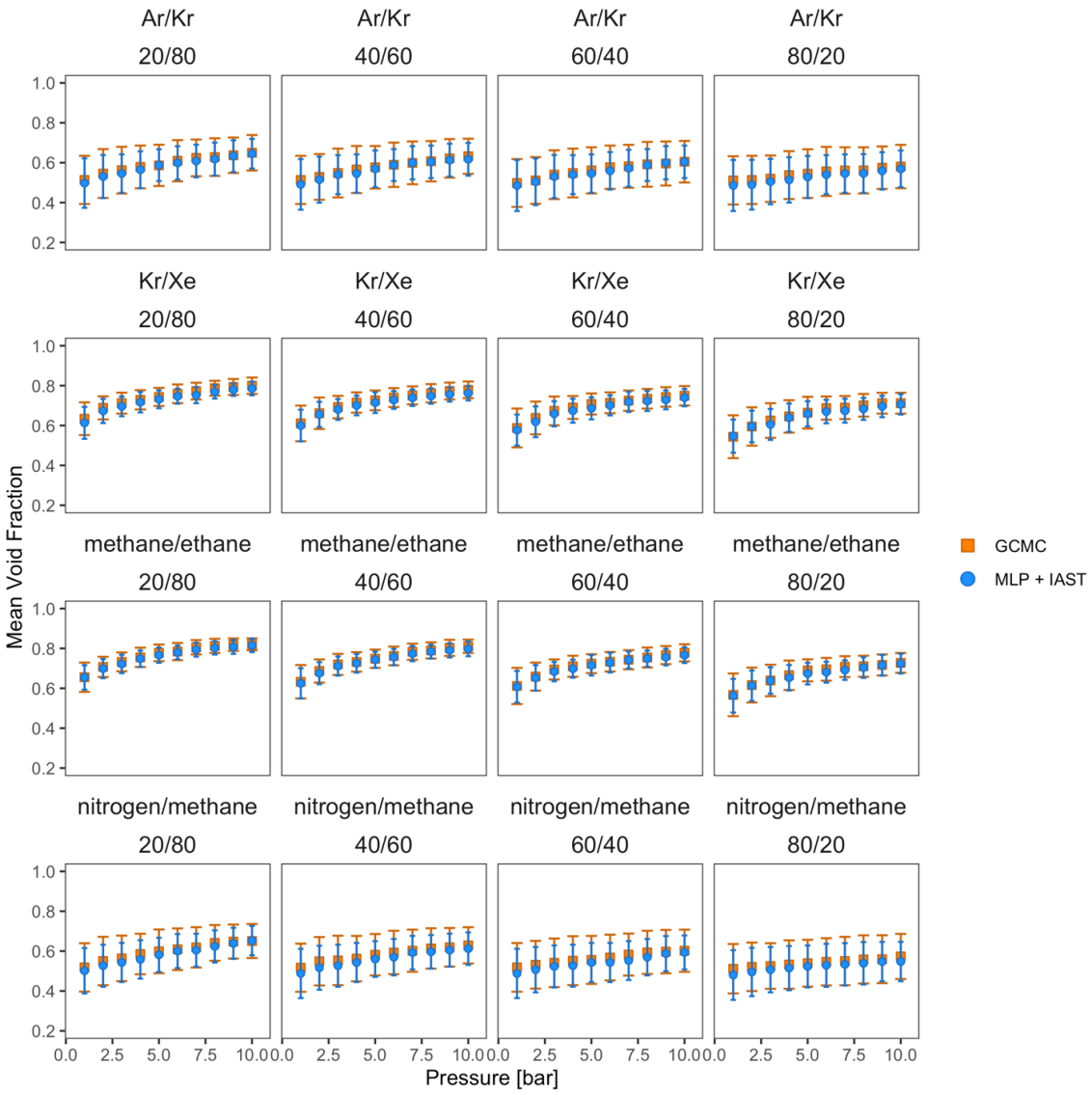

Figure E.12 The means (points) and standard deviations (bars) of the VFs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations.

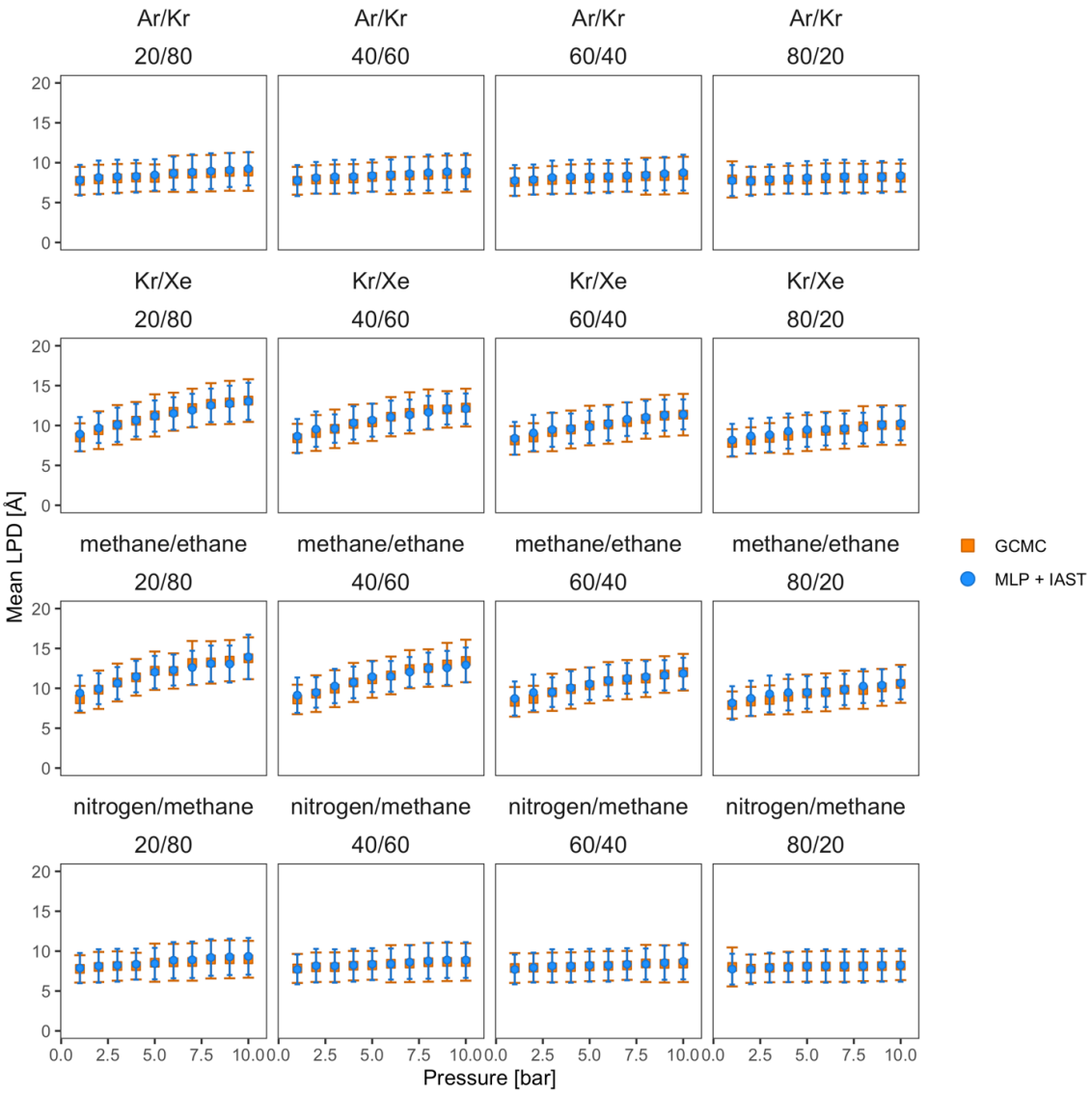

Figure E.13 The means (points) and standard deviations (bars) of the LPDs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations.

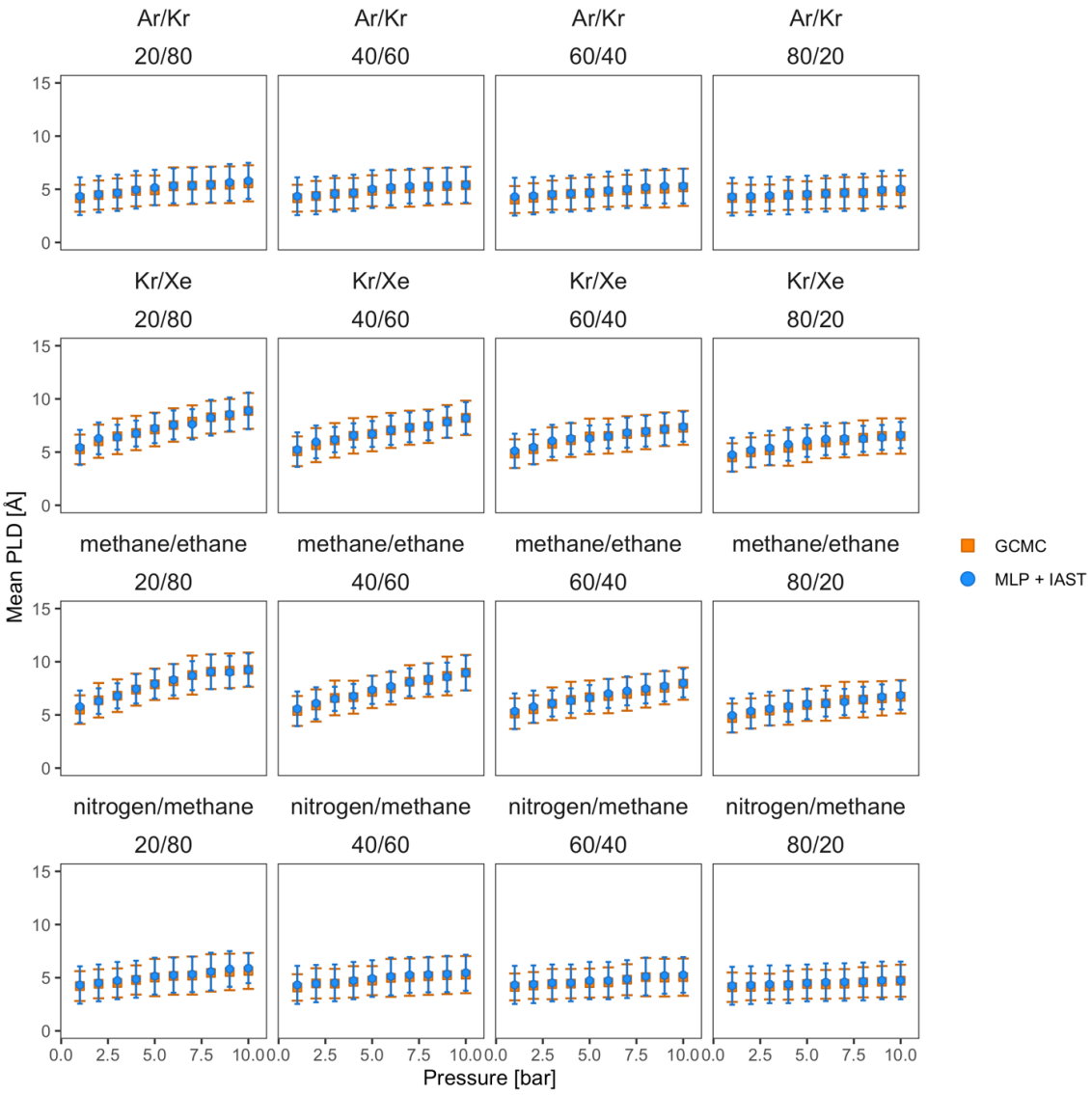

Figure E.14 The means (points) and standard deviations (bars) of the PLDs of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations.

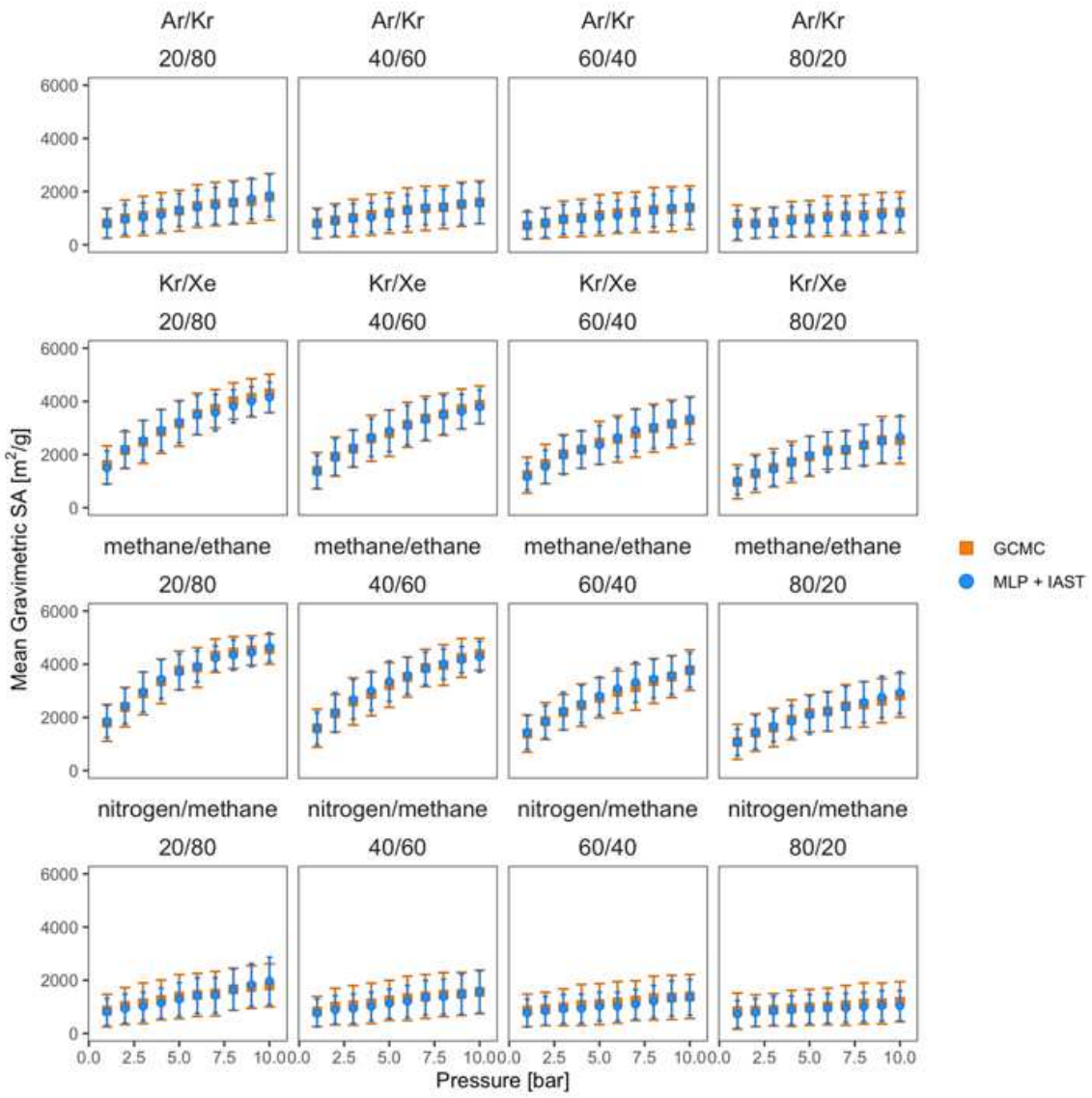

Figure E.15 The means (points) and standard deviations (bars) of the GSAs MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations.

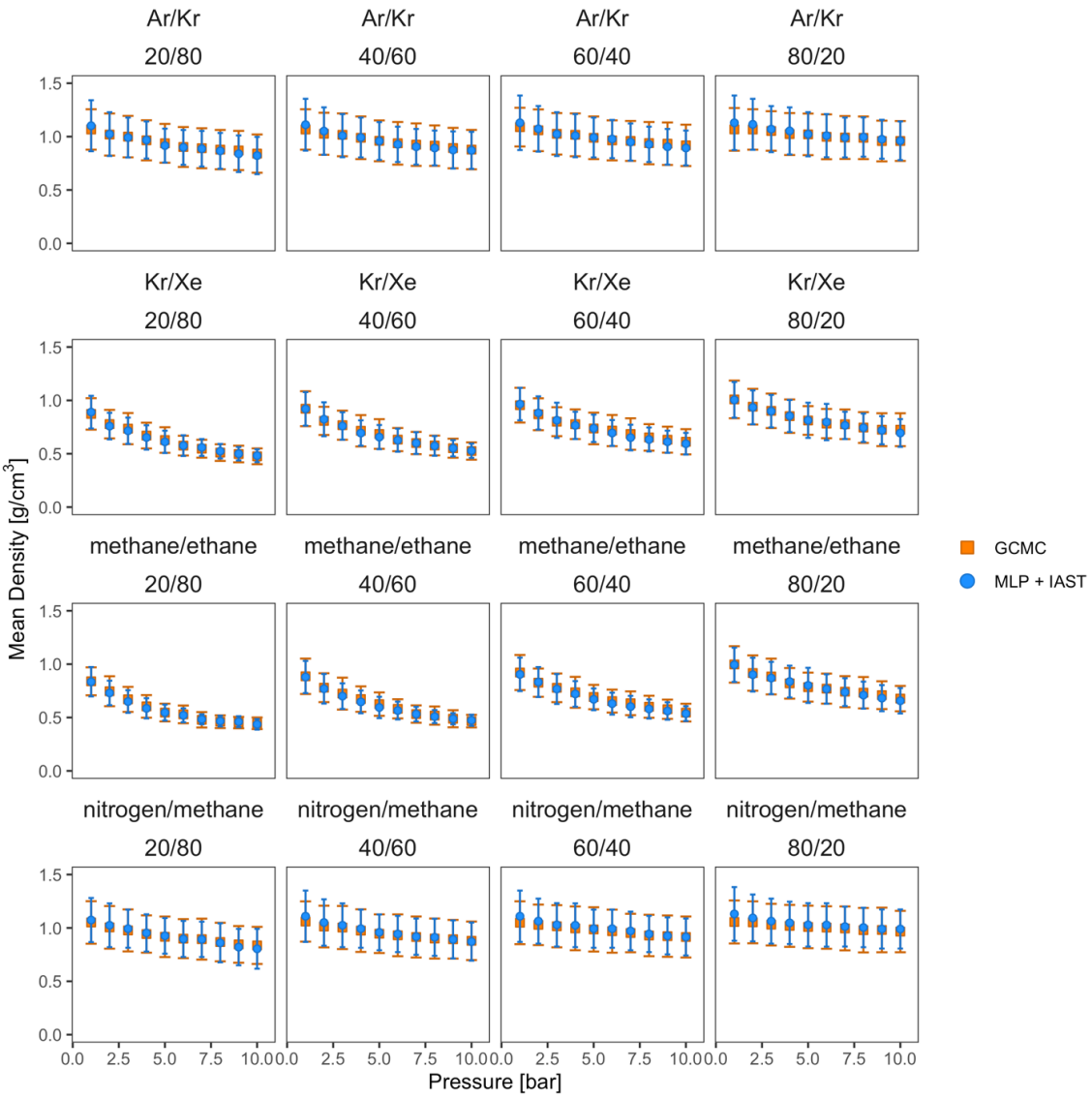

Figure E.16 The means (points) and standard deviations (bars) of the densities of MOFs in the 95th percentile of GCMC-predicted ΔQs compared with the analogous values for predictions from our MLP+IAST model for all composition, mixture, and pressure combinations.

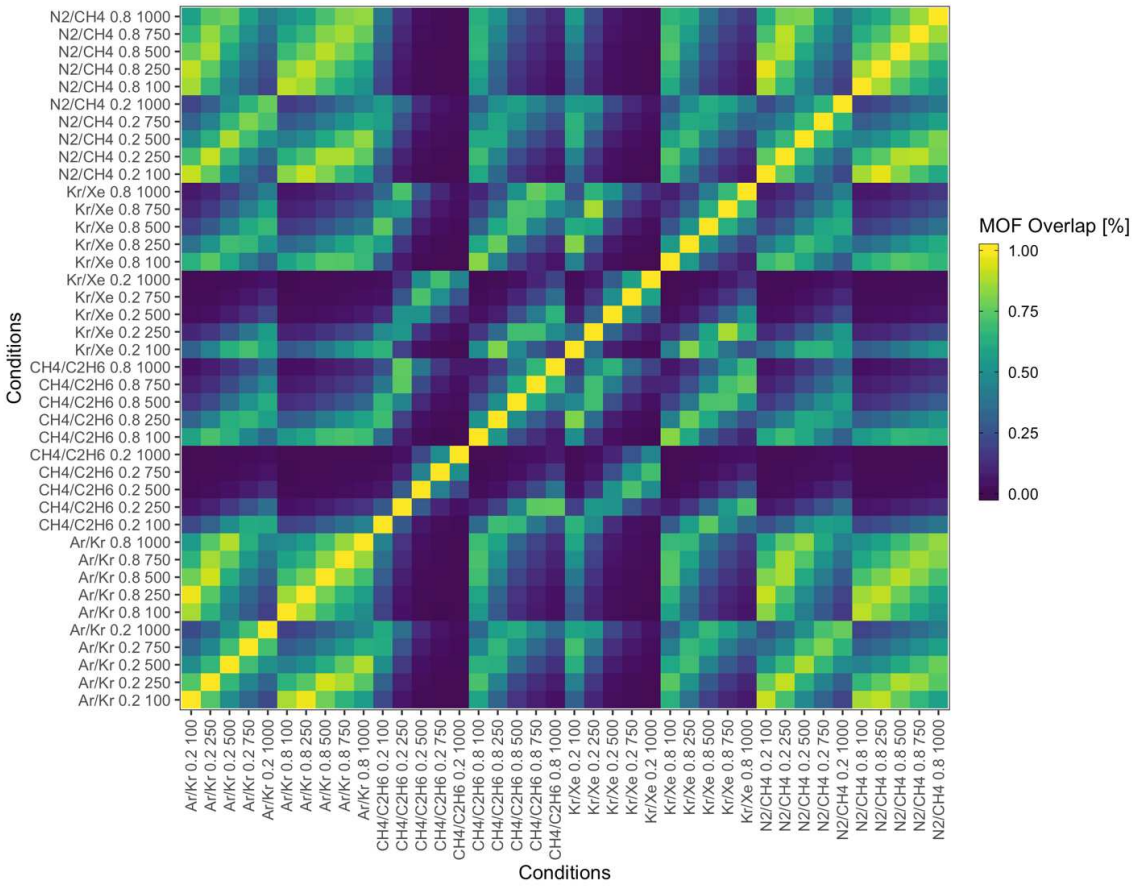

Figure E.17 Pairwise parentage overlaps of high-performing MOFs (MOFs in the 95th percentile of ΔQ) for the indicated conditions. Conditions are shown as space-separated strings of the mixture, fraction of the weakly adsorbed component, and pressure in kPa.

APPENDIX F

A LIST OF THE AUTHOR'S CONTRIBUTIONS

F.1 First Authored Publications

1. Deep learning combined with IAST to screen thermodynamically feasible MOFs for adsorption-based separation of multiple binary mixtures. Ryther Anderson, Diego A Gómez-Gualdrón, under review at *J. Chem. Phys* (**2021**).

2. Balancing volumetric and gravimetric uptake in highly porous materials for clean energy. Zhijie Chen[17], Penghao Li[17], Ryther Anderson[17], Xingjie Wang, Xuan Zhang, Lee Robison, Louis R Redfern, Shinya Moribe, Timur Islamoglu, Diego A Gómez-Gualdrón, Taner Yildirim, J Fraser Stoddart, Omar K Farha. *Science* (**2020**) DOI: 10.1126/science.aaz8881.

3. Large-Scale Free Energy Calculations on a Computational MOF Database: Toward Synthetic Likelihood Predictions. Ryther Anderson, Diego A Gómez-Gualdrón. *Chemistry of Materials* (**2020**) DOI: 10.1021/acs.chemmater.0c00744.

4. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. Ryther Anderson, Achay Biong, Diego A Gómez-Gualdrón. *J. Chem. Theory Comput.* (**2020**) DOI: 10.1021/acs.jctc.9b00940.

5. Exploiting hydrophobicity and hydrophilicity in nanopores as a design principle for "smart" MOF microtanks for methane storage. Ryther Anderson, Bomsaerah Seong, Zöe Peterson, Molly Stevanak, Moises A Carreon, Diego A Gómez-Gualdrón. *Mol. Syst. Des. Eng.* (**2020**) DOI: 10.1039/C9ME00072K.

6. Increasing topological diversity during computational "synthesis" of porous crystals: how and why. Ryther Anderson, Diego A Gómez-Gualdrón. *Cryst. Eng. Comm.* (**2019**) DOI: 10.1039/C8CE01637B.

7. Role of Pore Chemistry and Topology in the CO_2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. Ryther Anderson, Jacob Rodgers, Edwin Argueta, Achay Biong, Diego A Gómez-Gualdrón. *Chemistry of Materials* (**2018**) DOI: 10.1021/acs.chemmater.8b02257.

[17] authors with equal contribution

8. Molecular simulation insights on Xe/Kr separation in a set of nanoporous crystalline membranes. Ryther Anderson, Benjamin Schweitzer, Ting Wu, Moises A Carreon, Diego A Gómez-Gualdrón. *ACS App. Mat. Inter.* (**2018**). DOI: 10.1021/acsami.7b14791.

F.2 Coauthored Publications

1. Vacancy Healing as a Desorption Tool: Oxygen Triggered Removal of Stored Ammonia from NiO1–x/MOR Validated by Experiments and Simulations. James M Crawford, Ryther Anderson, Ryan J Gasvoda, Nolan C Kovach, Courtney S Smoljan, Jacek B Jasinski, Brian G Trewyn, Sumit Agarwal, Diego A Gómez-Gualdrón, Moises A Carreon. *ACS Applied Energy Materials* (**2020**). DOI: 10.1021/acsaem.0c00634.

2. Improving Energy Transfer within Metal–Organic Frameworks by Aligning Linker Transition Dipoles along Framework Axis. Jierui Yu, Ryther Anderson, Xinlin Li, Wenqian Xu, Subhadip Goswami, Sreehari Surendran Rajasree, Karan Maindan, Diego A Gómez-Gualdrón, Pravas Deria. *JACS* (**2020**) DOI: 10.1021/jacs.0c03949.

3. Electronic effects due to organic linker-metal surface interactions: implications on screening of MOF-encapsulated catalysts. Benjamin Schweitzer, Chloe Archuleta, Bomsaerah Seong, Ryther Anderson, Diego A Gómez-Gualdrón. *Phys. Chem. Chem. Phys.* (**2020**) DOI: doi.org/10.1039/C9CP05380H.

4. Dissociation, Dissolution, and Diffusion of Nitrogen on VxFey and VxCry Alloy Membranes Studied by First Principles. Peter Psarras, Ryther Anderson, Jennifer Wilcox, Diego A Gómez-Gualdrón. *J. Phys. Chem. C* (**2019**) DOI: 10.1021/acs.jpcc.9b09273.

5. Modular Synthesis of Highly Porous Zr-MOFs Assembled from Simple Building Blocks for Oxygen Storage. Jiafei Lyu, Xuan Zhang, Zhijie Chen, Ryther Anderson, Xingjie Wang, Megan C Wasson, Peng Bai, Xianghai Guo, Timur Islamoglu, Diego A Gómez-Gualdrón, Omar K Farha. *ACS App. Mat. Inter.* (**2019**) DOI: 10.1021/acsami.9b14439.

6. Material consequences of hydrogen dissolution in palladium alloys observed from first principles. Peter Psarras, Ryther Anderson, Diego A Gómez-Gualdrón, Jennifer Wilcox. *J. Phys. Chem. C.* (**2019**) DOI: 10.1021/acs.jpcc.9b03685.

7. Attainable volumetric targets for adsorption-based hydrogen storage in porous crystals: molecular simulation and machine learning. Grace Anderson, Benjamin Schweitzer, Ryther Anderson, Diego A Gómez-Gualdrón. *J. Phys. Chem. C* (**2018**) DOI: 10.1021/acs.jpcc.8b09420.

8. Hierarchically engineered mesoporous metal-organic frameworks toward cell-free immobilized enzyme systems. Peng Li, Qishui Chen, Timothy C Wang, Nicolaas A Vermeulen, B Layla Mehdi, Alice Dohnalkova, Nigel D Browning, Dengke Shen, Ryther Anderson, Diego A Gómez-Gualdrón, Furkan M Cetin, Jacek Jagiello, Abdullah M Asiri, J Fraser Stoddart, Omar K Farha. *Chem* (**2018**) DOI: 10.1016/j.chempr.2018.03.001.

9. Time dependent structural evolution of porous organic cage CC3. Jolie Lucero, Sameh K Elsaidi, Ryther Anderson, Ting Wu, Diego A Gómez-Gualdrón, Praveen K Thallapally, Moises A Carreon. *Cryst. Growth Des.* (**2018**) DOI: 10.1021/acs.cgd.7b01405.

F.3 Software

1. ToBaCCo (Topologically Based Crystal Constructor) - I am the primary developer of version 3.0, a major overhaul of the 1.0 and 2.0 versions authored by Yamil Colón, Diego Gomez-Gualdrón, and Edwin Argueta. Andrew Rosen contributed tools to the 3.0 version which allowed us to greatly expand our database of topologies. The current version of ToBaCCo can be found at https://github.com/tobacco-mofs/tobacco_3.0 at the time of writing.

2. cif2lammps - I am the developer of this Python program used to convert Crystallographic Information Files (CIFs) to LAMMPS inputs. This code was developed to batch-convert MOFs and other crystals (including adsorbed molecules) stored as CIFs to LAMMPS data and input files according to various forcefields. cif2lammps is what made the high-throughput LAMMPS calculations included in this thesis possible. The current version of cif2lammps can be found at https://github.com/rytheranderson/cif2lammps at the time of writing.

3. materials_interface - I am the developer of this suite of Python scripts that can be used to enumerate the adsorption configurations of small molecules in bulk metals or on metal surfaces (generally for use as initial inputs for DFT calculations). The current version of materials_interface can be found at https://github.com/rytheranderson/materials_interface at the time of writing.

4. XRD_matching - I am the developer of this Python program used to match the peak positions of simulated X-ray diffraction patterns with experimental patterns by varying the simulated structure lattice parameters. This code can be used to produce initial guesses of complex structures (such as MOFs) for use in Rietveld or Pawley refinement. The current version of XRD_matching can be found at https://github.com/rytheranderson/XRD_matching at the time of writing.

APPENDIX G

COPYRIGHT PERMISSIONS

The papers used in Chapters 3, 4, 5, and 6 have been previously published, as indicated at the beginning of those chapters. Other sections in the introduction also use parts of these papers. Chapter 7 was reproduced from an article submitted to a peer-reviewed journal (but not yet published). The American Chemical Society (ACS), Royal Society of Chemistry (RSC), and America Institute of Physics (AIP), which published these papers (Chapter 7 is under review in an AIP journal), allow authors to reuse their papers in a thesis or dissertation without explicit permission. Relevant text outlining these policies from the ACS, RSC, and AIP websites are provided in the Supplementary Copyright Permissions document included with this thesis. This document also includes written permission (as emails) from the coauthors of the papers in Chapters 4 and 6 to use these papers in my thesis. This document is organized by chapter, with all relevant documents included for each chapter in the order they appear here. The papers used in Chapters 3, 5, and 7 were authored only by myself and my advisor, and therefore do not require coauthor permission.