Supporting Information

Predicting Hydrogen Storage in MOFs via Machine Learning

Alauddin Ahmed† and Donald J. Siegel†,§,∆,ß,*

†Mechanical Engineering Department, §Materials Science & Engineering, ∆Applied Physics Program, and ßUniversity of Michigan Energy Institute, University of Michigan, Ann Arbor, MI 48109, United States
Section S1. Details of MOF database reported earlier.¹

The database is publicly available at the HyMARC Data Hub²

Table S1. Database of MOF crystal structures, calculated crystallographic properties, and calculated usable H₂ capacities.

| Source¹               | Available in database | Zero accessible surface area | H₂ capacity evaluated empirically | H₂ capacity evaluated with GCMC |
|-----------------------|-----------------------|------------------------------|----------------------------------|----------------------------------|
| UM+CoRE+CSD17         | 15,235                | 2,950                        | 12,285                           | 12,799                           |
| Mail-Order MOFs       | 112                   | 4                            | 108                              | 112                              |
| In Silico MOFs        | 2,816                 | 154                          | 2,662                            | 466                              |
| In Silico Surface MOFs| 8,885                 | 283                          | 8,602                            | 1,058                            |
| MOF-74 Analogs        | 61                    | 0                            | 61                               | 61                               |
| ToBaCCo               | 13,512                | 214                          | 13,298                           | 2,854                            |
| Zr-MOFs               | 204                   | 0                            | 204                              | 204                              |
| NW Hypothetical MOFs  | 137,000               | 30,160                       | 106,840                          | 20,156                           |
| UO Hypothetical MOFs  | 315,615               | 32,993                       | 291,507                          | 61,247                           |
| In-house synthesized via hypothetical design | 18                   | 0                            | 18                               | 5                                |
| **Total**             | **493,458**           | **66,758**                   | **426,700**                      | **98,962**                       |
Section S2 Literature review of machine learning for gas storage in MOFs

Table S2. Summary of recent studies that use machine learning (ML) to predict gas adsorption in MOFs. ρcrys, vf, vsa, mpd, lcd represent single crystal density, void fraction, volumetric surface area, maximum pore diameter, and largest cavity diameter, respectively. R2, AUE, a represent the coefficient of determination, Average Unsigned Error, and Root-Mean-Square Error, respectively. AUC = Area Under tLASSO: Least Absolute Shrinkage and Selection Operator; MLR: Multi-Linear Regression; SVM: Support Vector Machine; DT: Tress; RF: Random Forest; NN: Nearest Neighbors; GBM: Gradient Boosting Method; RBF: Radial Bias Function; PCA: Component Analysis; ANN: Artificial Neural Network.

| Study | Gas | ML Features | ML Method | Properties Predicted | Accuracy |
|-------|-----|-------------|-----------|----------------------|----------|
| Bucior et al. (2019) | CH4 | Energetics of MOF-guest interactions | Multilinear regression with LASSO | H2: Deliverable capacity 2 and 100 bar at 77 K. CH4: Deliverable capacity between 5.8 and 65 bar at 298 K | R2 = 0.96, AUE = 1.4 - 3.4 g/L, RMSE = 3.1 - 4.4 g/L |
| Anderson et al. (2018) | CO2 | ρcrys, vf, gsa, vsa, mpd, lcd, topology | MLR, SVM, DT, RF, NN, GBM | CO2 capture | R2 = 0.601 - 0.934 |
| Pardakhti et al (2017) | CH4 | ρcrys, vf, gsa, vsa, mpd, lcd interpenetration capacity, number of interpenetration framework, 19 chemical descriptors | DT, Poison regression, SVM, and RP | Total at 35 bar and 298 K | R2 = 0.97 |
| Aghaji et al. (2016) | CO2, CO2/CH4 | vf, gsa, lcd | DT, SVM(RBF), | Working capacity for the pressure swing between 1 and 10 atm at 298 K | AUC = 0.889 to 0.953 |
| Fernandez & Barnard (2016) | CO2, N2 | ρcrys, vf, gsa, vsa, mpd, lcd | PCA, k-means clustering, archetypal analysis, DT, SVM, MLL, ANN, RF | Total at 0.1 and 0.9 bar at 298 K | ~94% |
| Ohno & Mukae (2016) | CH4 | ρcrys, vf, gsa, vsa, mpd, and lcd | GP regression, SVM regression, NN, and LR | Total at 35 bar and 298K. | R2 = 0.79 |
| Simon et al. (2015) | Xe/Kr | ρcrys, vf, vsa, pld, surface density, Voronoi energy | RF | Xe/Kr selectivity | RMSE = 2.21 for 15,000 unitless numbers between 0 and 35 | R2 not Reported |
| Sezginel et al. (2015) | CH4 | ρcrys, vf, gsa, vsa, mpd, and lcd, pld, Qse | MVL regression | Total at 298 K and pressures in 1 to 65 bar | R2 =0.3 - 0.9 |
| Fernandez et al. (2014) | CO2 | AP-RDF | SVM classification | Total at P =0.15 & 1 bar at 298 K | 94.5% (classification) |
| Fernandez et al. (2013) | CH4, CO2 | AP-RDF | PCA, MLR, and SVM regression | Total at low pressure (0.1-0.9 bar) at 298 K | ~70% - ~83% |
| Fernandez et al. (2013) | CH4 | ρcrys, vf, gsa, vsa, mpd, lcd | DT, MLR, and SVM regression | Uptake at 1, 35, and 100 bar at 298 K | ~90% at 1 bar (classification); R2 (regression) = 0.85 (35bar); R2 (regression) = 0.93 (100 bar) |
**Section S3 Description of crystallographic features**

Table S3. Statistics for the datasets used in this study.

| Feature          | Dataset type | Minimum | Maximum | Mean  | Median | % zero values | Skew  | Kurtosis |
|------------------|--------------|---------|---------|-------|--------|---------------|-------|----------|
| d (g cm\(^{-3}\)) | Training     | 0.03    | 5.18    | 0.76  | 0.62   | 0             | 1.84  | 5.64     |
|                  | Test         | 0.03    | 3.97    | 0.76  | 0.61   | 0             | 1.79  | 4.96     |
|                  | Unseen       | 0.04    | 4.7     | 0.84  | 0.76   | 0             | 1.37  | 3.81     |
| gsa (m\(^2\) g\(^{-1}\)) | Training     | 0       | 9750    | 3112.01 | 3516 | 10          | -0.16 | -0.80    |
|                  | Test         | 0       | 9701    | 3137.82 | 3560 | 10          | -0.16 | -0.74    |
|                  | Unseen       | 0       | 9671    | 2530.47 | 2529 | 13         | 0.16  | -0.84    |
| vsa (m\(^2\) cm\(^{-3}\)) | Training     | 0       | 3995    | 1696.35 | 1912 | 10         | -1.03 | 0.23     |
|                  | Test         | 0       | 3966    | 1703.42 | 1918 | 10         | -1.04 | 0.26     |
|                  | Unseen       | 0       | 3482    | 1473.48 | 1736 | 13         | -1.10 | 0.01     |
| vf               | Training     | 0       | 0.99    | 0.71   | 0.76   | 0             | -1.38 | 2.19     |
|                  | Test         | 0.01    | 0.99    | 0.71   | 0.76   | 0             | -1.37 | 2.18     |
|                  | Unseen       | 0       | 0.98    | 0.69   | 0.71   | 0             | -0.70 | 0.34     |
| pv (cm\(^3\) g\(^{-1}\)) | Training     | 0       | 35.73   | 1.34   | 1.23   | 0             | 6.97  | 91.45    |
|                  | Test         | 0.01    | 29.82   | 1.37   | 1.24   | 0             | 7.29  | 89.60    |
|                  | Unseen       | 0       | 24.76   | 1.18   | 0.93   | 0             | 3.22  | 30.16    |
| lcd (Å)          | Training     | 0.4     | 71.6    | 10.14  | 9.2    | 0             | 2.45  | 11.94    |
|                  | Test         | 0.4     | 66.2    | 10.21  | 9.3    | 0             | 2.49  | 11.95    |
|                  | Unseen       | 0.4     | 69.9    | 10.41  | 9.4    | 0             | 1.27  | 3.61     |
| pld (Å)          | Training     | 0       | 71.5    | 7.86   | 7.5    | 0             | 2.81  | 19.54    |
|                  | Test         | 0.1     | 57.7    | 7.91   | 7.6    | 0             | 2.84  | 18.43    |
|                  | Unseen       | 0       | 68      | 7.45   | 6.9    | 0             | 1.21  | 5.39     |

Skew and kurtosis were calculated using the scipy.stats module in the SciPy package. Skewness is calculated from the ratio of the third moment \(m_3\) and the cube of the square root of second moment \(m_2\) of a feature variable, \(skew = \mu_3/\mu_2^{3/2}\), where \(\mu_i = (\sum_{k=1}^{n_{samples}} (x[k] - \bar{x})^i/n_{samples}\) is the \(i\)-th central moment, and \(\bar{x}\) is the mean of the feature variable. Kurtosis is the fourth central moment divided by the square of the second moment: \(kurtosis = \mu_4/\mu_2^2\).
Figure S1. Distribution of 6 crystallographic features in 3 different datasets used in this study. (a) pore volume, (b) single crystal density, (c) void fraction, (d) gravimetric surface area, (e) volumetric surface area, and (f) largest cavity diameter.
Section S4 Machine learning work-flow

Figure S2. Machine learning work-flow as described in the text.

Section S5 Metrics for ML accuracy

The coefficient of determination ($R^2$), average unsigned error (AUE), root-mean-squared error (RMSE), and median absolute error (MAE) are used to assess the accuracy of the various ML models with respect to GCMC calculations. If the test/training set contains $n_{\text{samples}}$ and $y_{i,\text{gcmc}}$ is the GCMC calculated $H_2$ capacity of $i$-th sample and $y_{i,\text{ml}}$ is the corresponding ML model prediction, then $R^2$, AUE, RMSE, and MAE are defined as follows:

$$R^2(y_{\text{gcmc}}, y_{\text{ml}}) = \frac{\sum_{i=1}^{n_{\text{samples}}} (y_{i,\text{gcmc}} - \bar{y}_{\text{gcmc}})(y_{i,\text{ml}} - \bar{y}_{\text{ml}})}{\sqrt{\sum_{i=1}^{n_{\text{samples}}} (y_{i,\text{gcmc}} - \bar{y}_{\text{gcmc}})^2 \sum_{i=1}^{n_{\text{samples}}} (y_{i,\text{ml}} - \bar{y}_{\text{ml}})^2}}$$  

(1)

$$\text{AUE}(y_{\text{gcmc}}, y_{\text{ml}}) = \frac{1}{n_{\text{samples}}-1} \sum_{i=1}^{n_{\text{samples}}-1} |y_{i,\text{gcmc}} - y_{i,\text{ml}}|$$  

(2)

$$\text{RMSE}(y_{\text{gcmc}}, y_{\text{ml}}) = \sqrt{\frac{1}{n_{\text{samples}}-1} \sum_{i=1}^{n_{\text{samples}}-1} (y_{i,\text{gcmc}} - y_{i,\text{ml}})^2}$$  

(3)

$$\text{MAE}(y_{\text{gcmc}}, y_{\text{ml}}) = \text{median}(|y_{1,\text{gcmc}} - y_{1,\text{ml}}|, \ldots, |y_{n,\text{gcmc}} - y_{n,\text{ml}}|)$$  

(4)

where $\bar{y}_{\text{gcmc}} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} y_{i,\text{gcmc}}$.

Kendal $\tau$ rank correlation coefficients were calculated using the scipy.stats module\textsuperscript{13–15} according to the definition of Kendall $\tau$-b.\textsuperscript{17–19}
Section S6 Training set sizes

Table S4. Training set sizes.

Table S5. Performance of ML models in predicting usable gravimetric capacities under pressure swing conditions. R², AUE, RSME, and MAE represent the coefficient of determination, average unsigned error, root-mean-squared error, and median absolute error, respectively.

| ML model                          | Model abbreviation | Feature scaling method | R²     | AUE (wt. %) | RMSE (wt. %) | Kendal τ | EV  | MAE   |
|-----------------------------------|--------------------|------------------------|--------|-------------|--------------|-----------|-----|-------|
| Ada Boost                         | AB                 | unscaled               | 0.975  | 0.476       | 0.332        | 0.910     | 0.976| 0.410 |
| Bagging with Decision Tree        | B/DT               | unscaled               | 0.997  | 0.141       | 0.037        | 0.959     | 0.997| 0.110 |
| Bagging with Random Forest        | B/RF               | unscaled               | 0.997  | 0.141       | 0.037        | 0.959     | 0.997| 0.110 |
| Boosted Decision Trees            | BDT                | unscaled               | 0.997  | 0.136       | 0.037        | 0.963     | 0.997| 0.100 |
| Decision Trees                    | DT                 | unscaled               | 0.995  | 0.180       | 0.065        | 0.949     | 0.995| 0.100 |
| Extremely Randomized Trees        | ERT                | unscaled               | 0.997  | 0.136       | 0.034        | 0.961     | 0.997| 0.104 |
| Gradient Boosting                 | GB                 | unscaled               | 0.997  | 0.158       | 0.045        | 0.955     | 0.997| 0.123 |
| K-Nearest Neighbors               | K-NN               | unscaled               | 0.983  | 0.346       | 0.226        | 0.900     | 0.983| 0.260 |
| Linear Regression                 | LR                 | unscaled               | 0.987  | 0.307       | 0.170        | 0.915     | 0.987| 0.241 |
| Nu-Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K       | minmax scale           | 0.986  | 0.235       | 0.187        | 0.958     | 0.987| 0.173 |
| Random Forest                     | RF                 | unscaled               | 0.997  | 0.141       | 0.037        | 0.959     | 0.997| 0.110 |
| Ridge Regression                  | RR                 | unscaled               | 0.987  | 0.307       | 0.170        | 0.915     | 0.987| 0.241 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K          | minmax scale           | 0.986  | 0.236       | 0.187        | 0.958     | 0.987| 0.174 |
| Support Vector Machine with Linear Kernel | SVM/L-K            | minmax scale           | 0.986  | 0.306       | 0.187        | 0.920     | 0.986| 0.224 |

Section S7 Performance comparison for ML algorithms

Table S6. Performance of ML models in predicting usable volumetric capacities under pressure swing condition. R², AUE, RSME, and MAE represent the coefficient of determination, average unsigned error, root-mean-squared error, and median absolute error, respectively.

| ML model                          | Model abbreviation | Feature scaling method | R²     | AUE (g. Hr.L⁻¹) | RMSE (g. Hr.L⁻¹) | Kendal τ | EV   | MAE   |
|-----------------------------------|--------------------|------------------------|--------|-----------------|-----------------|----------|------|-------|
| Ada Boost                         | AB                 | unscaled               | 0.936  | 2.218           | 7.732           | 0.873    | 0.918| 1.983 |
| Bagging with Decision Tree        | B/DT               | unscaled               | 0.982  | 1.011           | 2.333           | 0.918    | 0.982| 0.720 |
| Bagging with Random Forest        | B/RF               | unscaled               | 0.983  | 0.997           | 2.846           | 0.919    | 0.981| 0.710 |
| Boosted Decision Trees            | BDT                | unscaled               | 0.983  | 0.979           | 2.104           | 0.922    | 0.981| 0.700 |
| Decision Trees                    | DT                 | unscaled               | 0.971  | 1.298           | 3.566           | 0.995    | 0.971| 0.900 |
| Extremely Randomized Trees        | ERT                | unscaled               | 0.984  | 0.967           | 1.606           | 0.922    | 0.984| 0.692 |
| Gradient Boosting                 | GB                 | unscaled               | 0.980  | 1.104           | 2.454           | 0.911    | 0.980| 0.829 |
| K-Nearest Neighbors               | K-NN               | unscaled               | 0.913  | 2.378           | 10.517          | 0.794    | 0.913| 1.760 |
| Linear Regression                 | LR                 | unscaled               | 0.917  | 2.403           | 10.045          | 0.820    | 0.917| 1.961 |
| Nu-Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K       | minmax scale           | 0.949  | 1.891           | 6.337           | 0.858    | 0.931| 1.549 |
| Random Forest                     | RF                 | unscaled               | 0.982  | 1.011           | 2.156           | 0.918    | 0.982| 0.720 |
| Ridge Regression                  | RR                 | unscaled               | 0.917  | 2.404           | 10.066          | 0.820    | 0.917| 1.980 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K          | minmax scale           | 0.951  | 1.858           | 5.957           | 0.863    | 0.954| 1.468 |
| Support Vector Machine with Linear Kernel | SVM/L-K            | minmax scale           | 0.910  | 2.398           | 10.005          | 0.846    | 0.911| 1.982 |
Table S7. Performance of ML models in predicting usable gravimetric capacities under temperature+pressure swing condition. R², AUE, RSME, and MAE represent the coefficient of determination, average unsigned error, root-mean-squared error, and median absolute error, respectively.

| ML model                                      | Model abbreviation | Feature scaling method | R² | AUE (wt. %) | RSME (wt. %) | Kendal τ | EV  | MAE |
|-----------------------------------------------|--------------------|------------------------|----|-------------|--------------|----------|-----|-----|
| Ada Boost                                     | AB                 | unscaled               | 0.970 | 0.357 | 0.497 | 0.039 | 0.070 | 0.459 |
| Bagging with Decision Tree                    | B/DT               | unscaled               | 0.997 | 0.172 | 0.055 | 0.062 | 0.097 | 0.150 |
| Bagging with Random Forest                    | B/RF               | unscaled               | 0.997 | 0.171 | 0.054 | 0.061 | 0.097 | 0.150 |
| Boosted Decision Trees                        | BDT                | unscaled               | 0.997 | 0.165 | 0.051 | 0.063 | 0.097 | 0.127 |
| Decision Trees                                | DT                 | unscaled               | 0.994 | 0.223 | 0.095 | 0.051 | 0.094 | 0.200 |
| Extremely Randomized Trees                    | ERT                | unscaled               | 0.997 | 0.163 | 0.053 | 0.068 | 0.097 | 0.100 |
| Gradient Boosting                             | GB                 | unscaled               | 0.986 | 0.199 | 0.088 | 0.056 | 0.096 | 0.118 |
| K Nearest Neighbors                           | K-NN               | unscaled               | 0.995 | 0.250 | 0.117 | 0.063 | 0.095 | 0.200 |
| Linear Regression                             | LR                 | unscaled               | 0.992 | 0.266 | 0.111 | 0.047 | 0.092 | 0.208 |
| Nu Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K      | minmax scale           | 0.991 | 0.285 | 0.155 | 0.052 | 0.091 | 0.217 |
| Random Forest                                 | RF                 | unscaled               | 0.997 | 0.173 | 0.058 | 0.061 | 0.097 | 0.150 |
| Ridge Regression                              | RR                 | unscaled               | 0.992 | 0.266 | 0.131 | 0.047 | 0.092 | 0.208 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K        | minmax scale           | 0.991 | 0.283 | 0.155 | 0.052 | 0.091 | 0.215 |
| Support Vector Machine with Linear Kernel     | SVM/L-K            | minmax scale           | 0.988 | 0.451 | 0.525 | 0.040 | 0.073 | 0.141 |

Table S8. Performance of ML models in predicting usable volumetric capacities under temperature+pressure swing condition. R², AUE, RSME, and MAE represent the coefficient of determination, average unsigned error, root-mean-squared error, and median absolute error, respectively.

| ML model                                      | Model abbreviation | Feature scaling method | R² | AUE (wt. %) | RSME (wt. %) | Kendal τ | EV  | MAE |
|-----------------------------------------------|--------------------|------------------------|----|-------------|--------------|----------|-----|-----|
| Ada Boost                                     | AB                 | unscaled               | 0.911 | 2.187 | 9.834 | 0.752 | 0.012 | 1.077 |
| Bagging with Decision Tree                    | B/DT               | unscaled               | 0.965 | 1.381 | 4.147 | 0.809 | 0.066 | 0.540 |
| Bagging with Random Forest                    | B/RF               | unscaled               | 0.964 | 1.380 | 4.042 | 0.809 | 0.066 | 0.540 |
| Boosted Decision Trees                        | BDT                | unscaled               | 0.965 | 1.322 | 3.887 | 0.819 | 0.066 | 0.500 |
| Decision Trees                                | DT                 | unscaled               | 0.930 | 1.912 | 7.150 | 0.755 | 0.058 | 1.200 |
| Extremely Randomized Trees                    | ERT                | unscaled               | 0.967 | 1.320 | 3.700 | 0.839 | 0.067 | 0.912 |
| Gradient Boosting                             | GB                 | unscaled               | 0.955 | 1.572 | 4.953 | 0.875 | 0.055 | 1.126 |
| K Nearest Neighbors                           | K-NN               | unscaled               | 0.926 | 2.056 | 8.202 | 0.750 | 0.028 | 1.460 |
| Linear Regression                             | LR                 | unscaled               | 0.913 | 2.048 | 9.691 | 0.784 | 0.013 | 1.529 |
| Nu Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K      | minmax scale           | 0.913 | 2.053 | 9.656 | 0.787 | 0.015 | 1.510 |
| Random Forest                                 | RF                 | unscaled               | 0.965 | 1.303 | 4.109 | 0.809 | 0.066 | 0.540 |
| Ridge Regression                              | RR                 | unscaled               | 0.913 | 2.049 | 9.692 | 0.784 | 0.013 | 1.551 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K        | minmax scale           | 0.913 | 2.059 | 9.441 | 0.788 | 0.015 | 1.507 |
| Support Vector Machine with Linear Kernel     | SVM/L-K            | minmax scale           | 0.987 | 2.117 | 10.494 | 0.787 | 0.011 | 1.300 |
Section S8 Performance of ML models under TPS conditions

Figure S3. Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable H$_2$ capacities in MOFs. Data is collected under TPS conditions on a test set of 24,674 MOFs. Different colors represent different categories of MOFs. Top (a-c) and bottom (d-f) panels illustrate performance for usable gravimetric and volumetric capacities, respectively. (a, d): Agreement between ML and GCMC predictions. (b, e): Difference between ML and GCMC as a function of GCMC capacity. (c, f) Distribution of differences in predictions between ML and GCMC.
Section S9 Difference between ML and GCMC as a function of GCMC capacity for the training set

Figure S4. Difference between ML and GCMC as a function of GCMC capacity for the training set of 74,201 MOFs. Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable H\textsubscript{2} capacities in MOFs. Data is collected under PS (a, c) and TPS (b, d). Different colors represent different categories of MOFs. Top (a, b) and bottom (c, d) panels illustrate performance for usable gravimetric and volumetric capacities, respectively.
Section S10 Effect of training set size on ML model accuracies

Figure S5. Performance of Extremely Randomized Trees ML models for predicting usable (a) gravimetric and (b) volumetric H₂ capacity as a function of training set size (up to a dataset size of 10,000 MOFs) and the ratio of training to test set size. 100 different training sets ranging in size between 100 and 74,021 MOFs were examined. A common set of 24,674 MOFs was used for testing. Performance is quantified using R² (left axis, black) and the average unsigned error, AUE (right axis, blue and red for UG and UV, respectively). Lines represent a power-law fit to the data.

Table S9. Parameters of the power-law fit, ε(m) = \(\alpha m^\beta + \gamma\), where m is the size of the training dataset and \(\varepsilon\) represents the accuracy (here average unsigned error or AUE). \(\alpha\), \(\beta\), and \(\gamma\) are the power-law coefficient, exponent, and constant, respectively:

| Condition | \(\beta\) (scaling factor) | \(\alpha\) (coefficient) | \(\gamma\) (constant) |
|-----------|----------------------------|--------------------------|-----------------------|
| UG - PS   | -0.43                      | 1.19                     | 0.13                  |
| UG - TPS  | -0.37                      | 0.92                     | 0.16                  |
| UV - PS   | -0.23                      | 1.96                     | 0.85                  |
| UV - TPS  | -0.16                      | 2.10                     | 1.04                  |
Section 11 Univariate Feature Importance\textsuperscript{20,21}

| Feature | Gravimetric – PS | Volumetric – PS | Gravimetric – TPS | Volumetric – TPS |
|---------|-----------------|-----------------|-----------------|-----------------|
| d       | 7               | 7               | 7               | 7               |
| gsa     | 2               | 5               | 6               | 6               |
| lcd     | 4               | 3               | 5               | 5               |
| pld     | 1               | 3               | 3               | 3               |
| pv      | 1               | 1               | 1               | 1               |
| vf      | 1               | 1               | 1               | 1               |
| vsa     | 1               | 1               | 1               | 1               |

**Figure S6.** Relative importance of seven features in predicting H\textsubscript{2} storage in MOFs. Features are ranked 1 (most important) through 7 (least important). Four different methods were used: Pearson’s correlation coefficient (r), Breiman and Friedman’s tree-based algorithm as implemented in Scikit-learn, and the permutation importance method as implemented in rfpimp package. (a) usable gravimetric and (b) volumetric capacities for PS conditions. (c) usable gravimetric and (d) volumetric capacities for TPS conditions.
Section 12. GCMC verification of ML predictions

Table S10. MOFs predicted by ML to have high capacities under PS condition and whose performance was subsequently verified with GCMC. Here NW and UO represent Northwestern University and University of Ottawa databases.

| Name                           | Source | Density (g cm$^{-3}$) | Gravimetric surface area (m$^2$ g$^{-1}$) | Volumetric surface area (m$^2$ cm$^{-3}$) | Void fraction | Void volume (cm$^3$ g$^{-1}$) | Largest cavity diameter (Å) | Pore limiting diameter (Å) | Usable gravimetric capacity (wt. %) | Usable volumetric capacity (g H$_2$ L$^{-1}$) |
|--------------------------------|--------|-----------------------|-------------------------------------------|------------------------------------------|---------------|-------------------------------|----------------------------|-------------------------------|--------------------------------|----------------------------------------|
| mof_7642                       | ToBaCo | 0.30                  | 5561                                      | 1695                                     | 0.89          | 2.93                          | 12.8                       | 11.8                          | 11.1                            | 10.5                                   | 40.5                                   | 37.4                                   |
| mof_7690                       | ToBaCo | 0.30                  | 5715                                      | 1706                                     | 0.89          | 2.98                          | 12.8                       | 12.0                          | 11.3                            | 10.4                                   | 40.3                                   | 37.3                                   |
| mof_7594                       | ToBaCo | 0.40                  | 5070                                      | 2031                                     | 0.86          | 2.15                          | 11.2                       | 9.7                           | 8.6                             | 7.9                                    | 39.9                                   | 37.0                                   |
| mof_7210                       | ToBaCo | 0.29                  | 5936                                      | 1730                                     | 0.89          | 3.04                          | 13.4                       | 11.7                          | 11.4                            | 10.5                                   | 39.8                                   | 37.1                                   |
| mof_7738                       | ToBaCo | 0.25                  | 6054                                      | 1502                                     | 0.90          | 3.64                          | 14.5                       | 13.5                          | 13.0                            | 12.0                                   | 39.7                                   | 37.0                                   |
| hypotheticalMOF_5045702_i_1_j_24_k_20_m_2 | NW     | 0.31                  | 5926                                      | 1820                                     | 0.88          | 2.87                          | 16.0                       | 11.0                          | 10.9                            | 10.1                                   | 39.7                                   | 37.2                                   |
| hypotheticalMOF_5037315_i_1_j_20_k_12_m_1 | NW     | 0.31                  | 5073                                      | 1583                                     | 0.90          | 2.88                          | 17.7                       | 12.9                          | 10.8                            | 10.1                                   | 39.7                                   | 37.1                                   |
| hypotheticalMOF_5037467_i_1_j_20_k_12_m_8 | NW     | 0.31                  | 5600                                      | 1800                                     | 0.88          | 2.85                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.7                                   | 37.0                                   |
| hypotheticalMOF_5037563_i_1_j_20_k_12_m_13 | NW     | 0.31                   | 5897                                    | 1811                                     | 0.88          | 2.87                          | 16.1                       | 11.0                          | 10.9                            | 10.1                                   | 39.7                                   | 37.2                                   |
| hypotheticalMOF_5038404_i_1_j_20_k_20_m_15 | NW     | 0.31                   | 5870                                    | 1803                                     | 0.88          | 2.87                          | 16.0                       | 11.0                          | 10.9                            | 10.1                                   | 39.7                                   | 37.2                                   |
| hypotheticalMOF_5037379_i_1_j_20_k_12_m_4 | NW     | 0.31                   | 5818                                    | 1787                                     | 0.88          | 2.86                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5037407_i_1_j_20_k_12_m_5 | NW     | 0.31                   | 5818                                    | 1787                                     | 0.88          | 2.86                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5037479_i_1_j_20_k_12_m_9 | NW     | 0.31                   | 5818                                    | 1787                                     | 0.88          | 2.86                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5035561_i_1_j_28_k_20_m_11 | NW     | 0.31                   | 5874                                    | 1804                                     | 0.88          | 2.87                          | 16.0                       | 11.0                          | 10.9                            | 10.1                                   | 39.6                                   | 37.2                                   |
| hypotheticalMOF_5037439_i_1_j_20_k_12_m_7 | NW     | 0.31                   | 5838                                    | 1799                                     | 0.88          | 2.85                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5037499_i_1_j_20_k_12_m_10 | NW     | 0.31                   | 5854                                    | 1798                                     | 0.88          | 2.85                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5037531_i_1_j_20_k_12_m_11 | NW     | 0.31                   | 5818                                    | 1787                                     | 0.88          | 2.86                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.0                                   |
| hypotheticalMOF_5037523_i_1_j_20_k_12_m_11 | NW     | 0.31                   | 5857                                    | 1799                                     | 0.88          | 2.86                          | 16.0                       | 11.0                          | 10.9                            | 10.0                                   | 39.6                                   | 37.1                                   |
Figure S7. Comparison of GCMC calculations with ML predictions for the 21,700 highest-capacity MOFs predicted by ML for PS conditions. Top (a-c) and bottom (d-f) panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (a, d) show the correlation between GCMC and ML capacities; the diagonal lines indicate perfect correlations. Middle panels (b, e) show the difference between GCMC and ML, where the horizontal lines represent a zero difference. Right panels (c, f) show the distribution of differences from plots b and e.
Table S11. MOFs predicted by ML to have high capacities under TPS condition and whose performance was subsequently verified with GCMC. Here UO represents University of Ottawa database.

| Name                          | Source | Density (g cm\(^{-3}\)) | Gravimetric surface area (m \(^2\) g\(^{-1}\)) | Volumetric surface area (m \(^2\) cm\(^{-3}\)) | Void fraction | Porosity volume (g cm\(^{-3}\)) | Largest cavity diameter (Å) | Pore limiting diameter (Å) | Usable gravimetric capacity (wt.%) | Usable volumetric capacity (g H\(_2\) L\(^{-1}\)) |
|-------------------------------|--------|--------------------------|-----------------------------------------------|-----------------------------------------------|---------------|-------------------------------|-------------------------------|-------------------------------|----------------------------------|----------------------------------|
| str_m1_o11_f0_pcu.sym.102.out | UO     | 0.45                     | 4332                                           | 1974                                           | 0.84          | 1.84                          | 12.9                          | 10.1                          | 10.4                             | 9.7                              |
| str_m1_o11_f0_pcu.sym.117.out | UO     | 0.47                     | 4162                                           | 1977                                           | 0.83          | 1.74                          | 12.8                          | 9.9                           | 9.9                              | 9.0                              |
| str_m1_o11_f0_pcu.sym.121.out | UO     | 0.47                     | 4263                                           | 2006                                           | 0.83          | 1.76                          | 12.1                          | 10.2                          | 10.0                             | 9.4                              |
| str_m1_o11_f0_pcu.sym.13.out  | UO     | 0.46                     | 4326                                           | 2005                                           | 0.83          | 1.79                          | 12.7                          | 9.9                           | 10.1                             | 9.3                              |
| str_m1_o11_f0_pcu.sym.159.out | UO     | 0.58                     | 3703                                           | 2138                                           | 0.80          | 1.38                          | 10.4                          | 8.6                           | 8.3                              | 7.6                              |
| str_m1_o11_f0_pcu.sym.200.out | UO     | 0.45                     | 4339                                           | 1978                                           | 0.84          | 1.84                          | 12.9                          | 10.1                          | 10.3                             | 9.6                              |
| str_m1_o11_f0_pcu.sym.212.out | UO     | 0.60                     | 3417                                           | 2035                                           | 0.83          | 1.39                          | 12.0                          | 10.1                          | 8.1                              | 7.5                              |
| str_m1_o11_f0_pcu.sym.51.out  | UO     | 0.46                     | 4330                                           | 2007                                           | 0.83          | 1.79                          | 11.9                          | 9.9                           | 10.1                             | 9.3                              |
| str_m1_o11_f0_pcu.sym.71.out  | UO     | 0.45                     | 4436                                           | 1980                                           | 0.84          | 1.87                          | 13.0                          | 10.9                          | 10.4                             | 9.7                              |
| str_m1_o11_f0_pcu.sym.89.out  | UO     | 0.58                     | 3507                                           | 2043                                           | 0.83          | 1.42                          | 12.4                          | 9.8                           | 8.2                              | 7.7                              |
| str_m1_o17_f0_pcu.sym.1.out   | UO     | 0.46                     | 4283                                           | 1985                                           | 0.83          | 1.79                          | 11.9                          | 9.9                           | 10.1                             | 9.4                              |
| str_m1_o17_f0_pcu.sym.104.out | UO     | 0.46                     | 4439                                           | 2032                                           | 0.83          | 1.82                          | 12.5                          | 11.0                          | 10.2                             | 9.6                              |
| str_m1_o17_f0_pcu.sym.129.out | UO     | 0.60                     | 3585                                           | 2157                                           | 0.83          | 1.37                          | 14.6                          | 9.2                           | 7.9                              | 7.6                              |
| str_m1_o17_f0_pcu.sym.132.out | UO     | 0.60                     | 3438                                           | 2048                                           | 0.83          | 1.39                          | 12.7                          | 10.8                          | 8.0                              | 7.8                              |
| str_m1_o17_f0_pcu.sym.28.out  | UO     | 0.57                     | 3732                                           | 2117                                           | 0.80          | 1.41                          | 13.1                          | 10.9                          | 8.4                              | 7.8                              |
| str_m1_o2_f0_pcu.sym.1.out    | UO     | 0.56                     | 3615                                           | 2011                                           | 0.83          | 1.49                          | 13.1                          | 10.8                          | 8.5                              | 7.9                              |
| str_m1_o2_f0_pcu.sym.101.out  | UO     | 0.56                     | 3549                                           | 1978                                           | 0.84          | 1.50                          | 12.9                          | 10.7                          | 8.5                              | 7.7                              |
| str_m1_o2_f0_pcu.sym.11.out   | UO     | 0.44                     | 4487                                           | 1986                                           | 0.84          | 1.89                          | 12.4                          | 10.3                          | 10.4                             | 9.7                              |
| str_m1_o2_f0_pcu.sym.15.out   | UO     | 0.41                     | 4983                                           | 2054                                           | 0.84          | 2.04                          | 12.7                          | 9.1                           | 11.1                             | 10.3                             |
| str_m1_o2_f0_pcu.sym.2.out    | UO     | 0.47                     | 4179                                           | 1977                                           | 0.83          | 1.75                          | 11.9                          | 9.8                           | 9.8                              | 9.0                              |

**MOF-5**: 7.8     $51.9$
Figure S8. Comparison of GCMC calculations with ML predictions for the 7,901 highest-capacity MOFs predicted by ML for TPS conditions. Top (a-c) and bottom (d-f) panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (a, d) show the correlation between GCMC and ML capacities; the diagonal lines indicate perfect correlations. Middle panels (b, e) show the difference between GCMC and ML, where the horizontal lines represent a zero difference. Right panels (c, f) show the distribution of differences from plots b and e.

Table S12. Differences between ML-predicted and GCMC-calculated H₂ storage capacities of unseen MOFs at PS and conditions. Overprediction and underprediction mean ML predicted values are greater and smaller than those of GCMC calculated actual values, respectively.

| Statistics                  | Pressure swing | Temperature + pressure swing |
|-----------------------------|----------------|-----------------------------|
|                             | UG (wt. %)     | UV (g-H₂ L⁻¹)               | UG (wt. %)     | UV (g-H₂ L⁻¹)               |
| Largest overprediction      | 1.67           | 3.36                        | 0.94           | 4.93                        |
| Largest underprediction     | -0.96          | -4.46                       | -1.0           | -6.59                       |
| Average unsigned error      | 0.24           | 0.66                        | 0.24           | 1.28                        |
| Standard deviation          | 0.20           | 0.53                        | 0.17           | 0.99                        |
References

1. Ahmed, A., Seth, S., Purewal, J., Wong-Foy, A.G., Veenstra, M., Matzger, A.J., and Siegel, D.J. (2019). Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. Nat. Commun. 10.

2. Ahmed, A., and Siegel, D.J. HyMARC Sorbent Machine Learning Model: Predicting the hydrogen storage capacity of metal-organic frameworks via machine learning. https://sorbent-ml.hymarc.org/.

3. Bucior, B.J., Bobbitt, N.S., Islamoglu, T., Goswami, S., Gopalan, A., Yildirim, T., Farha, O.K., Bagheri, N., and Snurr, R.Q. Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. Mol. Syst. Des. Eng. 2018. DOI 10.1039/c8me00050f.

4. Anderson, R., Rodgers, J., Argueta, E., Biong, A., and Go, D.A. (2018). Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. Chem. Mater 30, 11.

5. Pardakhti, M., Moharreri, E., Wanik, D., Suib, S.L., and Srivastava, R. (2017). Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). ACS Comb. Sci. 19, 640–645.

6. Aghaji, M.Z., Fernandez, M., Boyd, P.G., Daff, T.D., and Woo, T.K. (2016). Quantitative Structure–Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selectivity for Methane Purification. 4505–4511.

7. Fernandez, M., and Barnard, A.S. (2016). Geometrical Properties Can Predict CO2 and N2 Adsorption Performance of Metal–Organic Frameworks (MOFs) at Low Pressure. ACS Comb. Sci. 18, 243–252.

8. Ohno, H., and Mukae, Y. (2016). Machine Learning Approach for Prediction and Search: Application to Methane Storage in a Metal–Organic Framework. J. Phys. Chem. C 120, 23963–23968.

9. Simon, C.M., Kim, J., Gomez-Gualdron, D.A., Camp, J.S., Chung, Y.G., Martin, R.L., Mercado, R., Deem, M.W., Gunter, D., Haranczyk, M., et al. (2015). The materials genome in action: identifying the performance limits for methane storage. Energy Environ. Sci. 8, 1190–1199.

10. Sezginel, K.B., Uzun, A., and Keskin, S. (2015). Multivariable linear models of structural parameters to predict methane uptake in metal–organic frameworks. Chem. Eng. Sci. 124, 125–134.

11. Fernandez, M., Woo, T.K., Wilmer, C.E., and Snurr, R.Q. (2013). Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks. J. Phys. Chem. C 117, 7681–7689.

12. Fernandez, M., Boyd, P.G., Daff, T.D., Aghaji, M.Z., and Woo, T.K. (2014). Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO2 Capture. J. Phys. Chem. Lett. 5, 3056–3060.

13. Zwilinger, D., Kokoska, S., Raton, B., New, L., and Washington, Y. (2000). standard probability and Statistics tables and formulae CRC.

14. Oliphant, T.E. (2007). Python for Scientific Computing. Comput. Sci. Eng. 9, 10–20.

15. Millman, K.J., and Aivazis, M. (2011). Python for Scientists and Engineers. Comput. Sci. Eng. 13, 9–12.

16. Abramowitz, M., and Stegun, I.A. (1965). Handbook of mathematical functions, with formulas, graphs, and mathematical tables, (Dover Publications).

17. Kendall, M.G. (1945). The Treatment of Ties in Ranking Problems. Biometrika 33, 239–251.

18. Kendall, M.G. (1938). A New Measure of Rank Correlation. Biometrika 30, 81–93.

19. Press, W.H. (2007). Numerical recipes : the art of scientific computing (Cambridge University Press).

20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
21. Parrt, T., and Turgutlu, K. rfpimp 1.3.4, https://github.com/parrt/random-forest-importances.