**Patterns, Volume *2***

## Supplemental information

# Predicting hydrogen storage in MOFs via machine learning

**Alauddin Ahmed and Donald J. Siegel**

Table S1. **Database of MOF crystal structures, calculated crystallographic properties, and calculated usable $H_2$ capacities reported earlier.[1] This database is publicly available at the HyMARC Data Hub.[2]**

| **Source[1]** | **Available in database** | **Zero accessible surface area** | **$H_2$ capacity evaluated empirically** | **$H_2$ 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** |

**Table S2. Summary of recent studies that use machine learning (ML) to predict gas adsorption in MOFs.[3–13] $\rho_{crys}$, vf, gsa, vsa, pv, mpd, lcd, pld represent single crystal density, void fraction, gravimetric surface area, volumetric surface area, pore volume, maximum pore diameter, largest cavity diameter, and pore limiting diameter, respectively. $R^2$, AUE, and RMSE represent the coefficient of determination, Average Unsigned Error, and Root-Mean-Square Error, respectively. AUC = Area Under the Curve. LASSO: Least Absolute Shrinkage and Selection Operator; MLR: Multi-Linear Regression; SVM: Support Vector Machine; DT: Decision Tress; RF: Random Forest; NN: Nearest Neighbors; GBM: Gradient Boosting Method; RBF: Radial Bias Function; PCA: Principal Component Analysis; ANN: Artificial Neural Network.**

| Study | Gas | ML Features | ML Method | Properties Predicted | Accuracy |
|---|---|---|---|---|---|
| This work | $H_2$ | $\rho_{crys}$, gsa, vsa, vf, pv, lcd, pld | Extremely Randomized Trees | Deliverable $H_2$ storage capacity between 5-100 bar at 77 K. | UG at PS: $R^2$ = 0.997; AUE = 0.14 wt. %; RMSE = 0.18 wt. % UV at PS: $R^2$ = 0.984; AUE = 0.97 g-$H_2$ $L^{-1}$; RMSE = 1.40 g-$H_2$ $L^{-1}$ UG at TPS: $R^2$ = 0.997; AUE = 0.16 wt. %; RMSE = 0.23 wt. % UV at TPS: $R^2$ = 0.967; AUE = 1.32 g-$H_2$ $L^{-1}$; RMSE = 1.92 g-$H_2$ $L^{-1}$ |
| Anderson et al. (2019)[5] | $H_2$ | Epsilon, temperature, pressure, $\rho_{crys}$, vf, vsa, mpd, lcd, alchemical catecholate site density, unit cell volume. | Neural network | Total volumetric $H_2$ for pressures 0.1, 1, 5, 35, 65, and 100 bar at 77, 160, and 295 K | AUE = 0.75 - 2.93 g-$H_2$ $L^{-1}$ |
| Bucior et al. (2019)[2] | $H_2$, $CH_4$ | Energetics of MOF-guest interactions | Multilinear regression with LASSO | $H_2$: Deliverable capacity 2 and 100 bar at 77 K. $CH_4$: Deliverable capacity between 5.8 and 65 bar at 298 K | $R^2$ = 0.96, AUE = 1.4 - 3.4 g/L, RMSE = 3.1 - 4.4 g/L |
| Anderson et al. (2018)[3] | $CO_2$ | $\rho_{crys}$, vf, gsa, vsa, mpd, lcd, topology | MLR, SVM, DT, RF, NN, GBM | $CO_2$ capture | $R^2$ = 0.601- 0.934 |
| Pardakhti et al (2017)[6] | $CH_4$ | $\rho_{crys}$, vf, gsa, vsa, mpd, lcd interpenetration capacity, number of interpenetration framework, 19 chemical descriptors | DT, Poisson regression, SVM, and RF | Total at 35 bar and 298 K | $R^2$ = 0.97 |
| Aghaji et al. (2016)[5] | $CO_2$, $CO_2$/ $CH_4$ | 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)[6] | $CO_2$, $N_2$ | $\rho_{crys}$, 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)[9] | $CH_4$ | $\rho_{crys}$, vf, gsa, vsa, mpd, and lcd | GP regression, SVM regression, NN, and LR | Total at 35 bar and 298K. | $R^2$ = 0.79 |
| Simon e al. (2015)[8] | Xe/ Kr | $\rho_{crys}$, vf, vsa, mpd, dpd, surface density, Voronoi energy | RF | Xe/Kr selectivity | RMSE = 2.21 for 15,000 unitless numbers between 0 and 35 $R^2$ not Reported |
| Sezginel et al. (2015)[11] | $CH_4$ | $\rho_{crys}$, vf, gsa, vsa, mpd, and lcd, pld, $Q_{st}$ | MVL regression | Total at 298 K and pressures in 1 to 65 bar | $R^2$ =0.3 - 0.9 |
| Fernandez et al. (2014)[10] | $CO_2$ | AP-RDF | SVM classification | Total at P =0.15 & 1 bar at 298 K | 94.5% (classification) |
| Fernandez et al. (2013)[11] | $CH_4$, $CO_2$, $N_2$ | 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)[12] | $CH_4$ | $\rho_{crys}$, 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);$R^2$ (regression) = 0.85 (35bar);$R^2$ (regression) = 0.93 (100 bar) |

## Supplemental Experimental Procedures

### Supplemental Note S1. Grand Canonical Monte Carlo (GCMC) calculations

The pseudo-Feynman-Hibbs interatomic potential parameters of Fischer et al.[14–16] were used to model $H_2$ molecules. MOF-$H_2$ interactions were calculated using Lorentz-Berthelot[17,18] combination rules. MOFs were assumed to be rigid and were described using interatomic potential parameters from a generic[19,20] force field. The RASPA package was used to evaluate $H_2$ uptake via Grand Canonical Monte Carlo (GCMC). All calculations were carried out using a 12 Å cut-off radius with compensating long-range corrections.[21,22] GCMC calculations for a given T,P condition were performed using 1000 initial cycles followed by a 1000 cycle production run. Each cycle consisted of translation, insertion, and deletion moves with equal probabilities.[23] Further details can be found in our recent publication.[1]

### Supplemental Note S2. 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 $n_{samples}$ and $y_{i,gcmc}$ is the GCMC calculated $H_2$ capacity of *i*-th sample and $y_{i,ml}$ is the corresponding ML model prediction, then $R^2$, AUE, RMSE, and MAE are defined as follows:

$$\mathrm{R}^2\left(y_{gcmc}, y_{ml}\right) = \sqrt{\frac{\sum_{i=1}^{n_{samples}}\left(y_{i,gcmc}-y_{i,ml}\right)^2}{\sum_{i=1}^{n_{samples}}\left(y_{i,gcmc}-\overline{y_{gcmc}}\right)^2}}, \tag{1}$$

$$\mathrm{AUE}\left(y_{gcmc}, y_{ml}\right) = \frac{\sum_{i=0}^{n_{samples}-1}\left|y_{i,gcmc}-y_{i,ml}\right|}{n_{samples}} \tag{2}$$

$$\mathrm{RMSE}\left(y_{gcmc}, y_{ml}\right) = \sqrt{\frac{\sum_{i=0}^{n_{samples}-1}\left(y_{i,gcmc}-y_{i,ml}\right)^2}{n_{samples}}}, \tag{3}$$

$$\mathrm{MAE}\left(y_{gcmc}, y_{ml}\right) = median\left(\left|y_{1,gcmc}-y_{1,ml}\right|, \dots, \left|y_{n,gcmc}-y_{n,ml}\right|\right) \tag{4}$$

where. $\overline{y_{gcmc}} = \left(\sum_{i=1}^{n_{samples}} y_{i,gcmc}\right)/n_{samples}$.

Kendal $\tau$ rank correlation coefficients were calculated using the scipy.stats module[25–27] according to the definition of Kendall *τ-b*.[29–31]

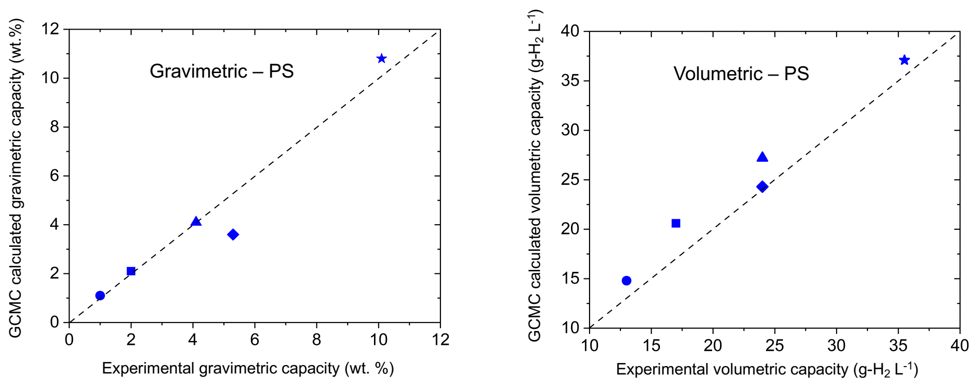


**Figure S1.** Comparison between experiments and GCMC calculations of $H_2$ capacities for a benchmark set of open-metal-site MOFs for pressure swing operation: HKUST-1 (■), NOTT-112 (◆), Cu-MOF-74 (●), NU-125 (▲), NU-100/PCN-610 (★).[1,24]

**Table S3. $H_2$ storage capacities for a benchmark set of open metal site (OMS) MOFs. Calculated capacities were predicted using the pseudo-Feynman-Hibbs interatomic potential. Measured $H_2$ storage data was compiled from García-Holley et al.[24] and from earlier work performed by the present authors.[1] 'Expt.' refers to measured capacities from the literature, 'GCMC' refers to predictions from the present study.**

| CSD Refcode | Common name | OMS density Å$^{-3}$ | Usable gravimetric capacity PS conditions (wt. %) | | Usable volumetric capacity PS conditions (g-$H_2$ $L^{-1}$) | |
|---|---|---|---|---|---|---|
| | | | Expt.[1,23] | GCMC | Expt.[1,23] | GCMC |
| FQIQCEN | HKUST-1 | $2.63 \times 10^{-3}$ | 2.0 | 2.1 | 17 | 20.6 |
| FOPFAS | NOTT-112 | $9.24 \times 10^{-4}$ | 5.3 | 3.6 | 24 | 24.3 |
| LENKIA | Cu-MOF-74 | $4.91 \times 10^{-3}$ | 1.0 | 1.1 | 13 | 14.8 |
| REWNEO | NU-125 | $1.09 \times 10^{-3}$ | 4.1 | 4.1 | 24 | 27.2 |
| HABQUY/GAGZEV | NU-100/ PCN-610 | $4.47 \times 10^{-4}$ | 10.1 | 10.8 | 35.5 | 37.1 |

**Table S4. Statistics for the datasets used in this study.** Skew and kurtosis were calculated using the scipy.stats module in the SciPy package.[25–27] 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 $\mu_i = \left(\sum_{k=1}^{n_{samples}}(x[k]-\bar{x})^i\right)/n_{samples}$ is the *i*-th central moment, and $\bar{x}$ is the mean of the feature variable.[25–27] Kurtosis is the fourth central moment divided by the square of the second moment: $kurtosis = \mu_4/\mu_2^2$. [25–28]

| 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 |

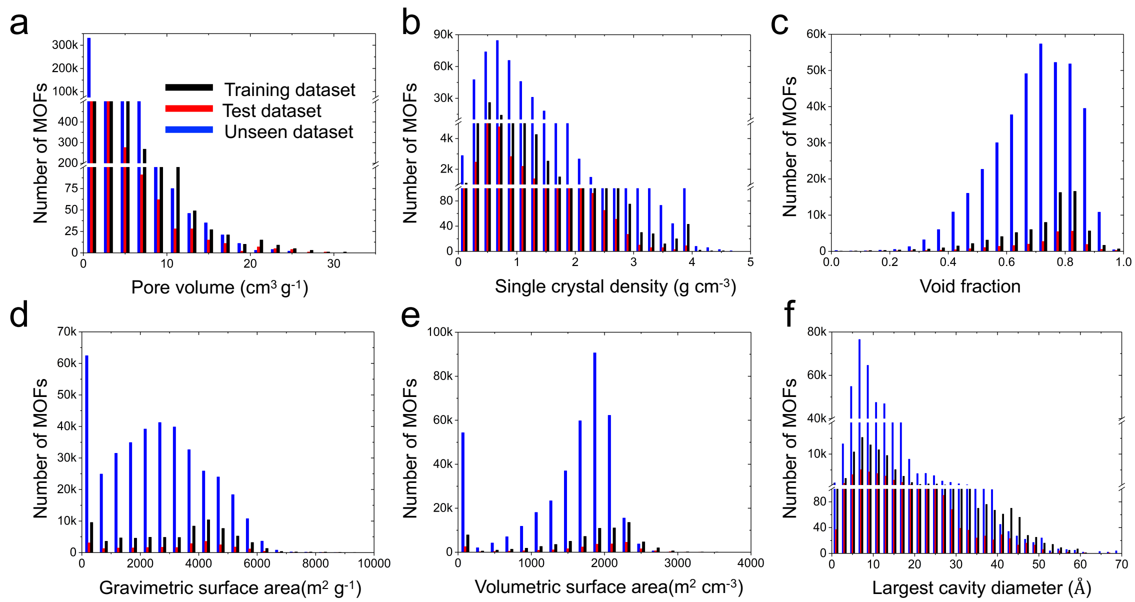


**Figure S2.** Distribution of 6 crystallographic features in 3 different datasets used in this study. (a) pore volume, (b) single crystal density, (c) void fraction, (d) gravimetric surface area, (e) volumetric surface area, and (f) largest cavity diameter.

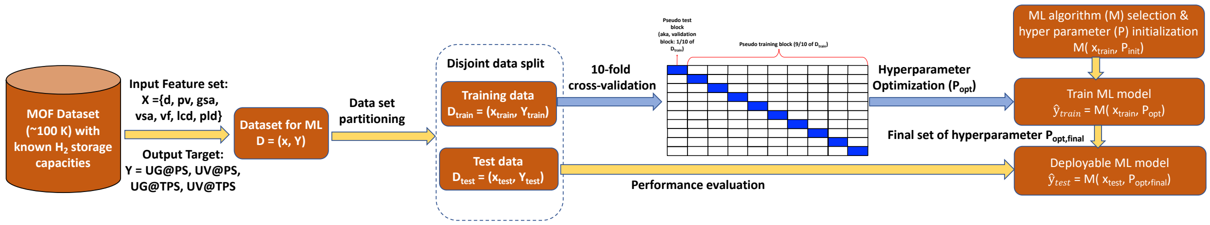


**Figure S3.** Machine learning work-flow.

**Table S5. Training set sizes.**

100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200,300,400,500,600,700,800,900,1000, 1200, 1300,1400,1500,1600,1700,1800,1900,2000,3000,4000,5000,6000,7000,8000,9000,10000,11000,12000, 13000,14000,15000,16000,17000,18000,19000,20000,21000,22000,23000,24000,25000,26000,27000, 28000,29000,30000,31000,32000,33000,34000,35000,36000,37000,38000,39000,40000,41000,42000, 43000,44000,45000,46000,47000,48000,49000,50000,51000,52000,53000,54000,55000,56000,57000, 58000,59000,60000,61000,62000,63000,64000,65000,66000,67000,68000,69000,70000,71000,72000, 73000,74000

**Table S6. Performance of ML models in predicting usable gravimetric capacities under pressure swing conditions. $R^2$, 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 | $R^2$ | AUE (wt. %) | RMSE (wt. %) | Kendal $\tau$ | 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 |

**Table S7. Performance of ML models in predicting usable volumetric capacities under pressure swing conditions. $R^2$, 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 | $R^2$ | AUE (g -$H_2$ $L^{-1}$) | RMSE (g -$H_2$ $L^{-1}$) | Kendal $\tau$ | EV | MAE |
|---|---|---|---|---|---|---|---|---|
| Ada Boost | AB | unscaled | 0.936 | 2.258 | 7.732 | 0.873 | 0.938 | 1.983 |
| Bagging with Decision Tree | B/DT | unscaled | 0.982 | 1.011 | 2.133 | 0.918 | 0.982 | 0.720 |
| Bagging with Random Forest | B/RF | unscaled | 0.983 | 0.997 | 2.048 | 0.919 | 0.983 | 0.710 |
| Boosted Decision Trees | BDT | unscaled | 0.983 | 0.979 | 2.104 | 0.922 | 0.983 | 0.700 |
| Decision Trees | DT | unscaled | 0.971 | 1.298 | 3.568 | 0.895 | 0.971 | 0.900 |
| Extremely Randomized Trees | ERT | unscaled | 0.984 | 0.967 | 1.960 | 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 | 0.829 | 0.917 | 1.981 |
| Nu-Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K | minmax scale | 0.949 | 1.899 | 6.137 | 0.858 | 0.951 | 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 | 10.046 | 0.829 | 0.917 | 1.980 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K | minmax scale | 0.951 | 1.836 | 5.957 | 0.863 | 0.954 | 1.468 |
| Support Vector Machine with Linear Kernel | SVM/L-K | minmax scale | 0.910 | 2.398 | 10.905 | 0.846 | 0.913 | 1.902 |

**Table S8. Performance of ML models in predicting usable gravimetric capacities under temperature+pressure swing conditions. $R^2$, 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 | $R^2$ | AUE (wt. %) | RMSE (wt. %) | Kendal $\tau$ | EV | MAE |
|---|---|---|---|---|---|---|---|---|
| Ada Boost | AB | unscaled | 0.970 | 0.557 | 0.497 | 0.939 | 0.970 | 0.459 |
| Bagging with Decision Tree | B/DT | unscaled | 0.997 | 0.172 | 0.055 | 0.962 | 0.997 | 0.130 |
| Bagging with Random Forest | B/RF | unscaled | 0.997 | 0.171 | 0.054 | 0.961 | 0.997 | 0.130 |
| Boosted Decision Trees | BDT | unscaled | 0.997 | 0.165 | 0.051 | 0.963 | 0.997 | 0.127 |
| Decision Trees | DT | unscaled | 0.994 | 0.223 | 0.095 | 0.951 | 0.994 | 0.200 |
| Extremely Randomized Trees | ERT | unscaled | 0.997 | 0.163 | 0.053 | 0.966 | 0.997 | 0.100 |
| Gradient Boosting | GB | unscaled | 0.996 | 0.199 | 0.068 | 0.956 | 0.996 | 0.158 |
| K-Nearest Neighbors | K-NN | unscaled | 0.993 | 0.250 | 0.117 | 0.943 | 0.993 | 0.200 |
| Linear Regression | LR | unscaled | 0.992 | 0.266 | 0.131 | 0.947 | 0.992 | 0.208 |
| Nu-Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K | minmax scale | 0.991 | 0.285 | 0.155 | 0.952 | 0.991 | 0.217 |
| Random Forest | RF | unscaled | 0.997 | 0.173 | 0.056 | 0.961 | 0.997 | 0.130 |
| Ridge Regression | RR | unscaled | 0.992 | 0.266 | 0.131 | 0.947 | 0.992 | 0.208 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K | minmax scale | 0.991 | 0.283 | 0.155 | 0.952 | 0.991 | 0.215 |
| Support Vector Machine with Linear Kernel | SVM/L-K | minmax scale | 0.968 | 0.451 | 0.535 | 0.948 | 0.973 | 0.345 |

**Table S9. Performance of ML models in predicting usable volumetric capacities under temperature+pressure swing condition. $R^2$, 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 | $R^2$ | AUE (wt. %) | RMSE (wt. %) | Kendal $\tau$ | EV | MAE |
|---|---|---|---|---|---|---|---|---|
| Ada Boost | AB | unscaled | 0.911 | 2.387 | 9.954 | 0.752 | 0.912 | 1.877 |
| Bagging with Decision Tree | B/DT | unscaled | 0.963 | 1.381 | 4.147 | 0.809 | 0.963 | 0.940 |
| Bagging with Random Forest | B/RF | unscaled | 0.964 | 1.380 | 4.042 | 0.809 | 0.964 | 0.940 |
| Boosted Decision Trees | BDT | unscaled | 0.965 | 1.322 | 3.887 | 0.819 | 0.965 | 0.900 |
| Decision Trees | DT | unscaled | 0.936 | 1.812 | 7.150 | 0.755 | 0.936 | 1.200 |
| Extremely Randomized Trees | ERT | unscaled | 0.967 | 1.320 | 3.700 | 0.819 | 0.967 | 0.912 |
| Gradient Boosting | GB | unscaled | 0.955 | 1.572 | 4.953 | 0.785 | 0.955 | 1.126 |
| K-Nearest Neighbors | K-NN | unscaled | 0.926 | 2.036 | 8.202 | 0.710 | 0.926 | 1.460 |
| Linear Regression | LR | unscaled | 0.913 | 2.048 | 9.691 | 0.764 | 0.913 | 1.329 |
| Nu-Support Vector Machine with Radial Basis Function (RBF) Kernel | Nu-SVM/RBF-K | minmax scale | 0.913 | 2.033 | 9.656 | 0.767 | 0.915 | 1.310 |
| Random Forest | RF | unscaled | 0.963 | 1.383 | 4.169 | 0.809 | 0.963 | 0.940 |
| Ridge Regression | RR | unscaled | 0.913 | 2.049 | 9.692 | 0.764 | 0.913 | 1.331 |
| Support Vector Machine Radial Basis Function (RBF) Kernel | SVM/RBF-K | minmax scale | 0.913 | 2.029 | 9.641 | 0.768 | 0.915 | 1.307 |
| Support Vector Machine with Linear Kernel | SVM/L-K | minmax scale | 0.907 | 2.117 | 10.404 | 0.767 | 0.911 | 1.390 |

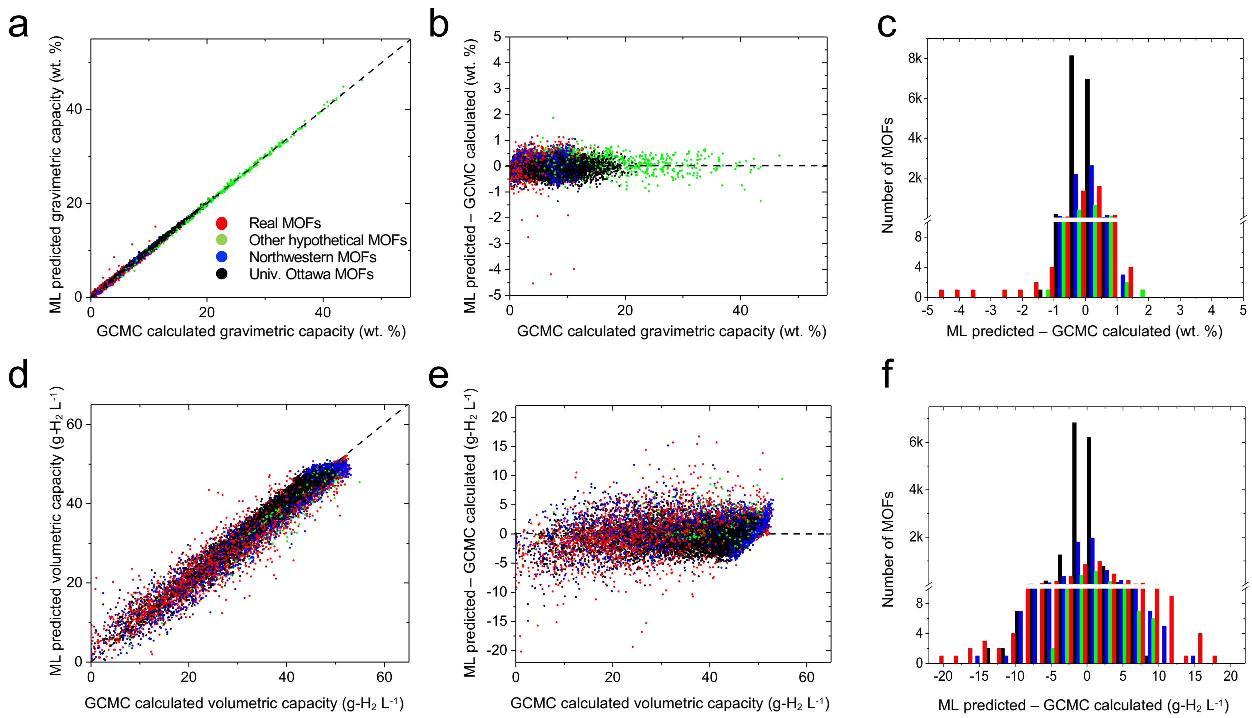


**Figure S4.** Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable $H_2$ capacities in MOFs. Data is collected under TPS conditions on a test set of 24,674 MOFs. Different colors represent different categories of MOFs. Top **(a-c)** and bottom **(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.

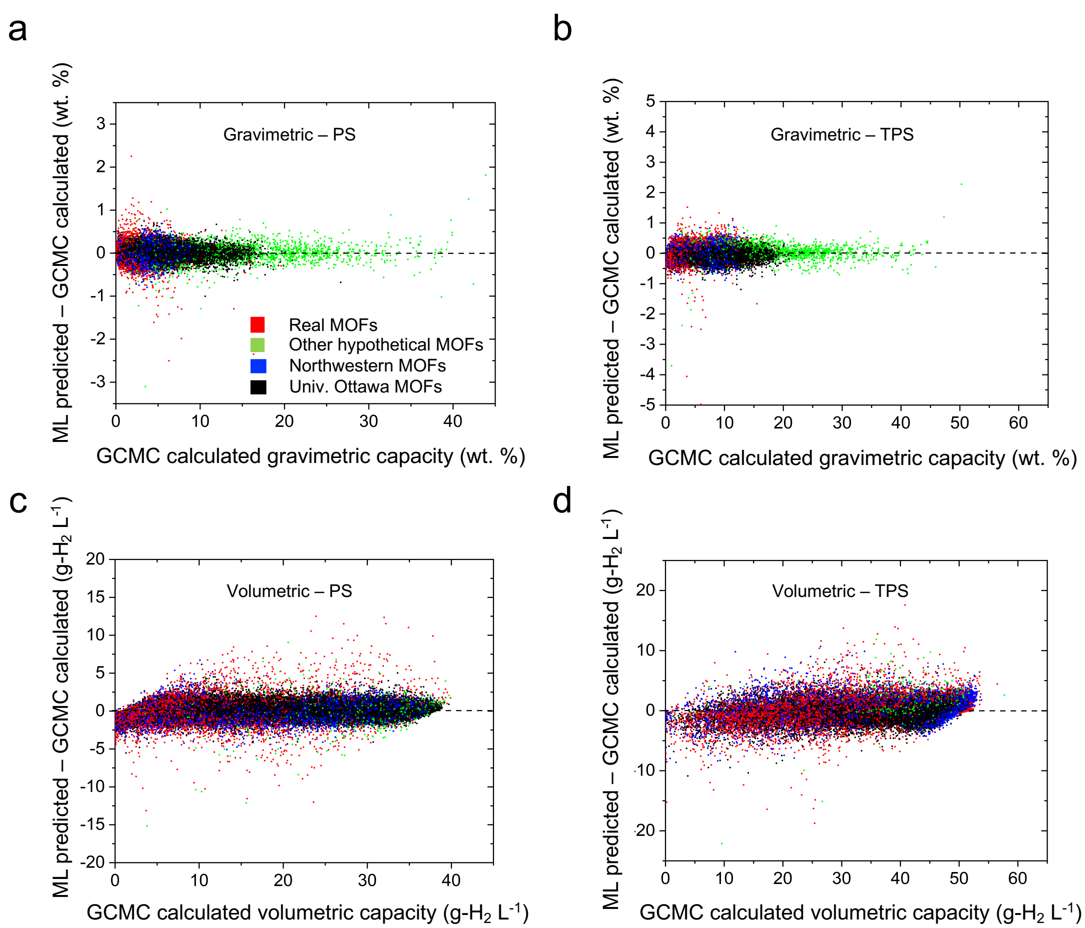


**Figure S5.** Difference between ML and GCMC as a function of GCMC capacity for the training set of 74,201 MOFs. Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable $H_2$ capacities in MOFs. Data is collected under PS **(a, c)** and TPS **(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.

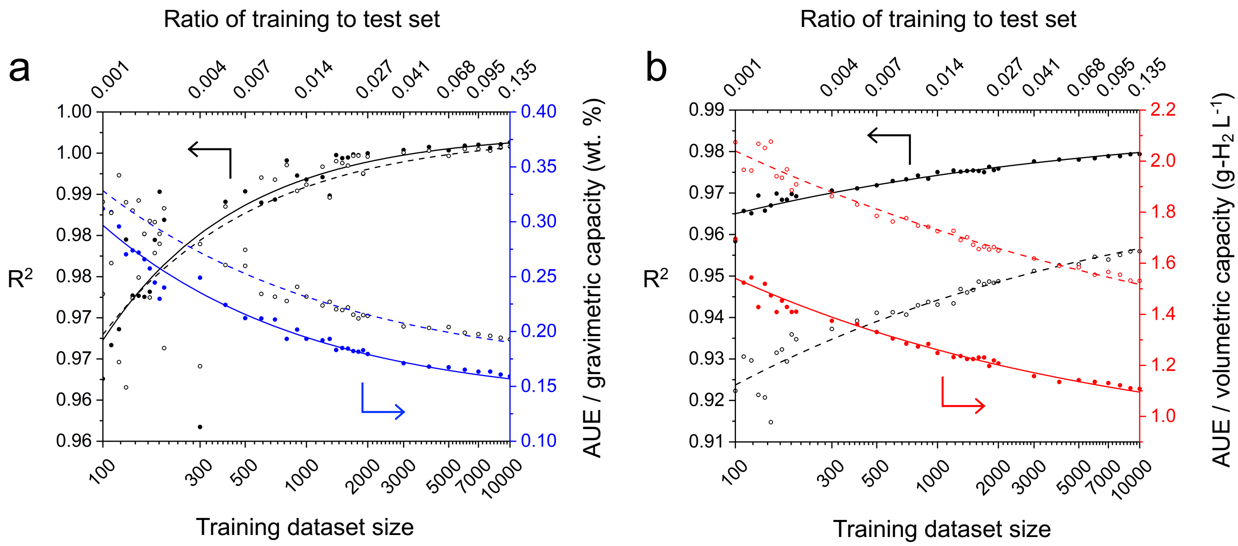


**Figure S6.** Performance of Extremely Randomized Trees ML models as a function of training set size and the ratio of training to test set size. (a) Usable gravimetric and (b) volumetric $H_2$ capacity. 100 different training sets ranging in size between 100 and 74,021 MOFs were examined. A common set of 24,674 MOFs was used for testing. Performance is quantified using $R^2$ (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 S10. Parameters of the power-law fit, $\varepsilon(m) = \alpha m^{\beta} + \gamma$, where m is the size of the training dataset and $\varepsilon$ represents the metric of accuracy (here average unsigned error or AUE). α, β, and $\gamma$ are the power-law coefficient, exponent, and constant, respectively.**

| Condition | β (scaling factor) | α (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 |

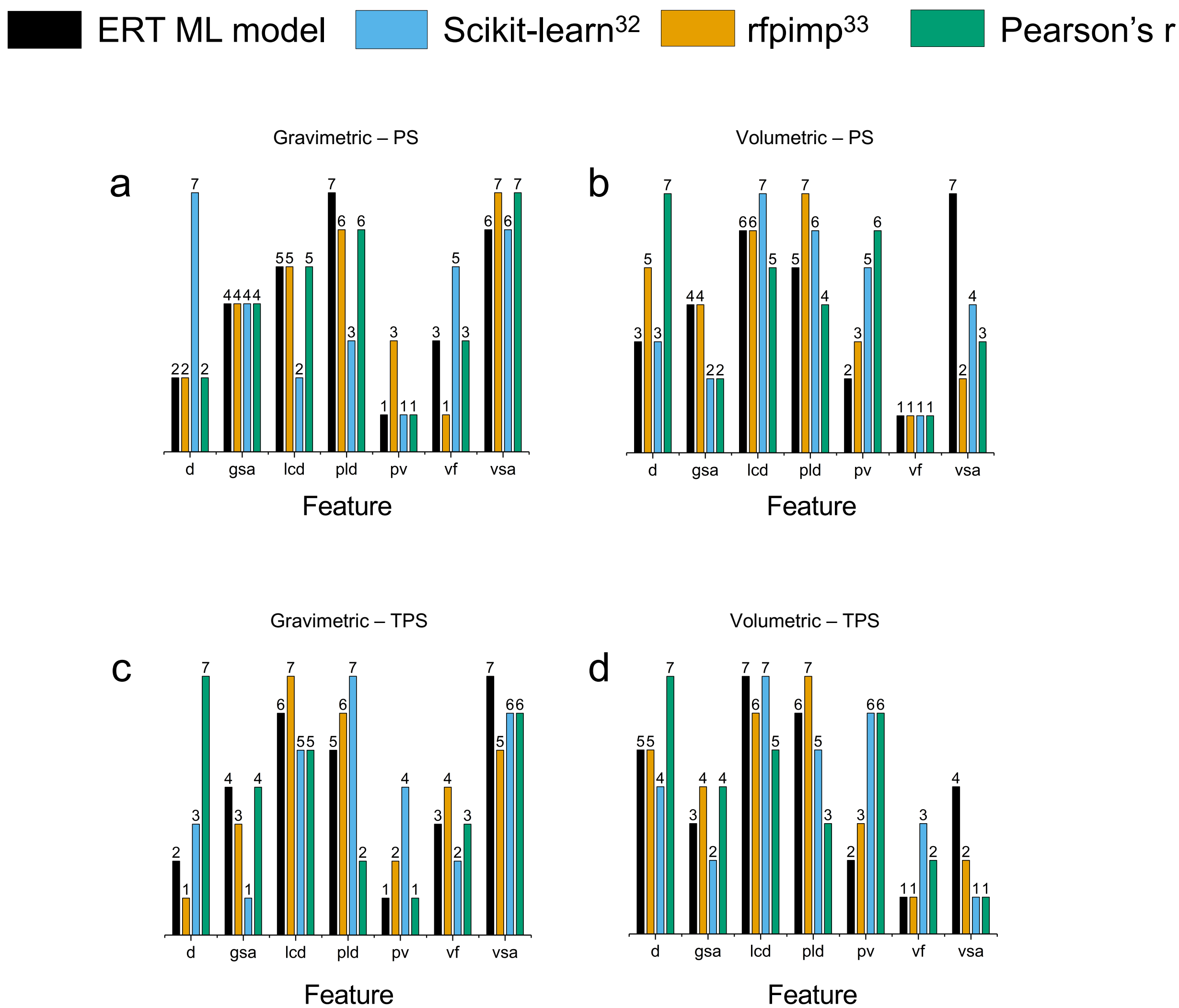


**Figure S7.** Relative importance of seven features in predicting $H_2$ storage in MOFs. [32,33] Features are ranked 1(most important) through 7 (least important). Four different methods were used: Pearson's correlation coefficient (r), Breiman and Fried-man's tree-based algorithm as implemented in 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.

**Table S11. Machine learning models generated for various combinations of features**

**Table S12. 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 | Pore volume (cm$^3$ g$^{-1}$) | Largest cavity diameter (Å) | Pore limiting diameter (Å) | Usable gravimetric capacity (wt. %) | | Usable volumetric capacity (g-$H_2$ L$^{-1}$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | GCMC | ML | GCMC | ML |
| mof_7642 | ToBaCCo | 0.30 | 5561 | 1695 | 0.89 | 2.93 | 12.8 | 11.8 | 11.1 | 10.3 | 40.5 | 37.4 |
| mof_7690 | ToBaCCo | 0.30 | 5715 | 1706 | 0.89 | 2.98 | 12.8 | 12.0 | 11.3 | 10.4 | 40.3 | 37.3 |
| mof_7594 | ToBaCCo | 0.40 | 5070 | 2031 | 0.86 | 2.15 | 11.2 | 9.7 | 8.6 | 7.9 | 39.9 | 37.0 |
| mof_7210 | ToBaCCo | 0.29 | 5936 | 1730 | 0.89 | 3.04 | 13.4 | 11.7 | 11.4 | 10.5 | 39.8 | 37.1 |
| mof_7738 | ToBaCCo | 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 |
| str_m3_o19_o19_f0_nbo.sym.1.out | UO | 0.31 | 5073 | 1583 | 0.90 | 2.88 | 17.7 | 12.9 | 10.8 | 10.1 | 39.7 | 37.1 |
| hypotheticalMOF_5037315_i_1_j_20_k_12_m_1 | NW | 0.31 | 5818 | 1787 | 0.88 | 2.86 | 16.0 | 11.0 | 10.9 | 10.0 | 39.7 | 37.0 |
| hypotheticalMOF_5037467_i_1_j_20_k_12_m_8 | NW | 0.31 | 5860 | 1800 | 0.88 | 2.85 | 16.0 | 11.0 | 10.9 | 10.0 | 39.7 | 37.0 |
| str_m3_o5_o20_f0_nbo.sym.1.out | UO | 0.39 | 4772 | 1882 | 0.87 | 2.22 | 14.1 | 9.6 | 8.7 | 8.1 | 39.7 | 37.2 |
| 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_5055561_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 | 5858 | 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 |

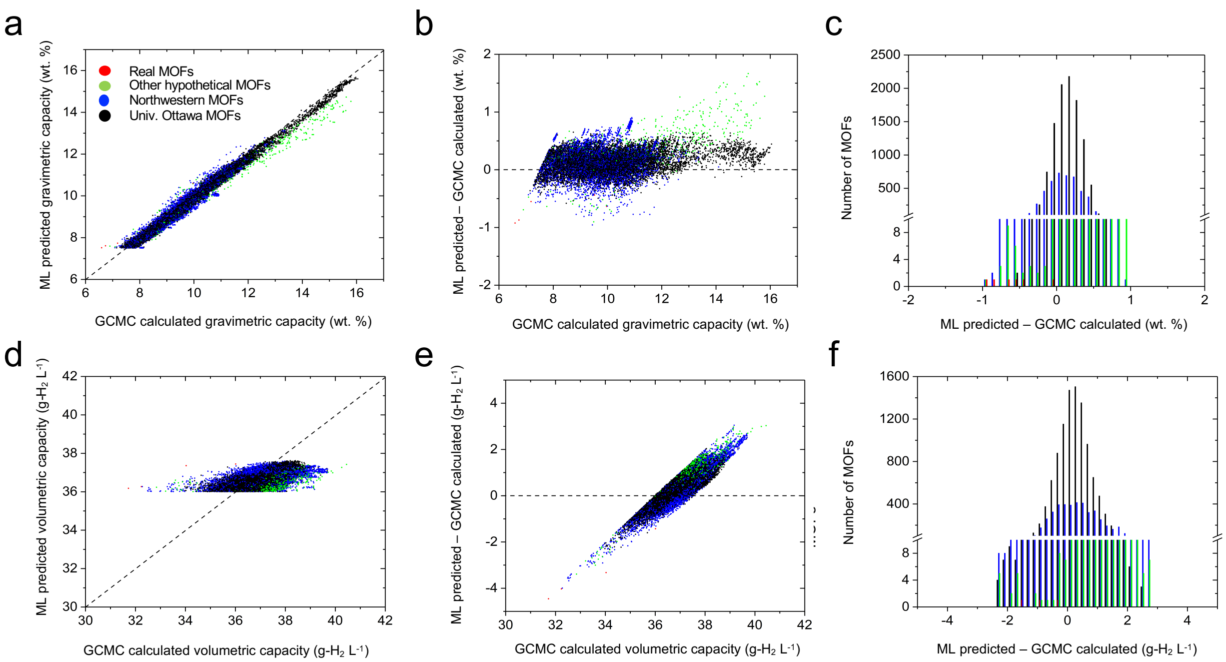


**Figure S8.** Comparison of GCMC calculations with ML predictions for the 21,700 highest-capacity MOFs predicted by ML for PS conditions. Top (**a**-**c**) and bottom (**d-f)** panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (**a**, **d**) show 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 S13. MOFs predicted by ML to have high capacities under TPS condition and whose performance was subsequently verified with GCMC. Here UO represents the 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 | Pore volume (cm$^3$ g$^{-1}$) | Largest cavity diameter (Å) | Pore limiting diameter (Å) | Usable gravimetric capacity (wt. %) | | Usable volumetric capacity (g-$H_2$ L$^{-1}$) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | GCMC | ML | GCMC | ML |
| str_m1_o1_o11_f0_pcu.sym.102.out | UO | 0.45 | 4352 | 1974 | 0.84 | 1.84 | 12.9 | 10.1 | 10.4 | 9.7 | 53.1 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.117.out | UO | 0.47 | 4162 | 1977 | 0.83 | 1.74 | 12.8 | 9.9 | 9.9 | 9.0 | 52.8 | 48.0 |
| str_m1_o1_o11_f0_pcu.sym.121.out | UO | 0.47 | 4263 | 2006 | 0.83 | 1.76 | 12.1 | 10.2 | 10.0 | 9.4 | 52.7 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.13.out | UO | 0.46 | 4326 | 2005 | 0.83 | 1.79 | 12.7 | 9.9 | 10.1 | 9.3 | 52.6 | 48.0 |
| str_m1_o1_o11_f0_pcu.sym.159.out | UO | 0.58 | 3703 | 2138 | 0.80 | 1.38 | 10.4 | 8.6 | 8.3 | 7.6 | 52.6 | 48.5 |
| str_m1_o1_o11_f0_pcu.sym.200.out | UO | 0.45 | 4359 | 1978 | 0.84 | 1.84 | 12.9 | 10.1 | 10.3 | 9.6 | 52.6 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.212.out | UO | 0.60 | 3417 | 2035 | 0.83 | 1.39 | 12.0 | 10.1 | 8.1 | 7.5 | 52.5 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.51.out | UO | 0.46 | 4330 | 2007 | 0.83 | 1.79 | 11.9 | 9.9 | 10.1 | 9.3 | 52.5 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.71.out | UO | 0.45 | 4436 | 1980 | 0.84 | 1.87 | 13.0 | 10.9 | 10.4 | 9.7 | 52.5 | 48.1 |
| str_m1_o1_o11_f0_pcu.sym.89.out | UO | 0.58 | 3507 | 2043 | 0.83 | 1.42 | 12.4 | 9.8 | 8.2 | 7.7 | 52.5 | 48.1 |
| str_m1_o1_o17_f0_pcu.sym.1.out | UO | 0.46 | 4283 | 1985 | 0.83 | 1.79 | 11.9 | 9.9 | 10.1 | 9.4 | 52.5 | 48.3 |
| str_m1_o1_o17_f0_pcu.sym.104.out | UO | 0.46 | 4439 | 2032 | 0.83 | 1.82 | 12.5 | 11.0 | 10.2 | 9.6 | 52.4 | 48.2 |
| str_m1_o1_o17_f0_pcu.sym.129.out | UO | 0.60 | 3585 | 2157 | 0.83 | 1.37 | 14.6 | 9.2 | 7.9 | 7.6 | 52.3 | 48.2 |
| str_m1_o1_o17_f0_pcu.sym.132.out | UO | 0.60 | 3438 | 2048 | 0.83 | 1.39 | 12.7 | 10.8 | 8.0 | 7.8 | 52.3 | 48.3 |
| str_m1_o1_o17_f0_pcu.sym.28.out | UO | 0.57 | 3732 | 2117 | 0.80 | 1.41 | 13.1 | 10.9 | 8.4 | 7.8 | 52.2 | 48.1 |
| str_m1_o1_o2_f0_pcu.sym.1.out | UO | 0.56 | 3615 | 2011 | 0.83 | 1.49 | 13.1 | 10.8 | 8.5 | 7.9 | 52.2 | 48.4 |
| str_m1_o1_o2_f0_pcu.sym.101.out | UO | 0.56 | 3549 | 1978 | 0.84 | 1.50 | 12.9 | 10.7 | 8.5 | 7.7 | 52.1 | 48.1 |
| str_m1_o1_o2_f0_pcu.sym.11.out | UO | 0.44 | 4487 | 1986 | 0.84 | 1.89 | 12.4 | 10.3 | 10.4 | 9.7 | 52.0 | 48.2 |
| str_m1_o1_o2_f0_pcu.sym.15.out | UO | 0.41 | 4983 | 2054 | 0.84 | 2.04 | 12.7 | 9.1 | 11.1 | 10.3 | 52.0 | 48.1 |
| str_m1_o1_o2_f0_pcu.sym.2.out | UO | 0.47 | 4179 | 1977 | 0.83 | 1.75 | 11.9 | 9.8 | 9.8 | 9.0 | 52.0 | 48.0 |
| **MOF-5** | | | | | | | | | **7.8** | | **51.9** | |

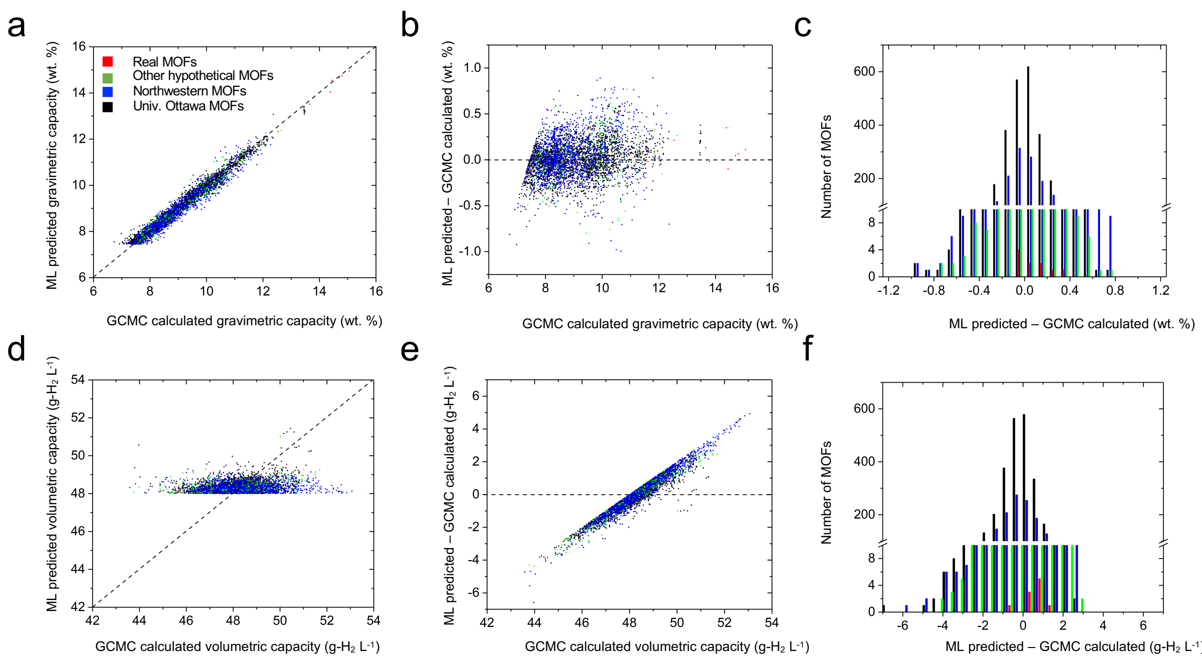


**Figure S9.** Comparison of GCMC calculations with ML predictions for the 7,901 highest-capacity MOFs predicted by ML for TPS conditions. Top (**a-c**) and bottom (**d-f)** panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (**a**, **d**) show the 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 S14. Comparison between ML-predicated and GCMC-calculated $H_2$ capacities in unseen MOFs for PS and TPS conditions.**

| Metric | Pressure swing | | Temperature + pressure swing | |
|---|---|---|---|---|
| | UG (wt. %) | UV (g-$H_2$ $L^{-1}$) | UG (wt. %) | UV (g-$H_2$ $L^{-1}$) |
| Largest overprediction with respect to GCMC | 1.67 | 3.36 | 0.94 | 4.93 |
| Largest underprediction with respect to GCMC | -0.96 | -4.46 | -1.0 | -6.59 |
| Average unsigned error with respect to GCMC | 0.24 | 0.66 | 0.24 | 1.28 |
| Standard deviation with respect to GCMC | 0.20 | 0.53 | 0.17 | 0.99 |

## Supplemental 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*, 1568.
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 CO 2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. Chem. Mater *30*, 11.
5. Anderson, G., Schweitzer, B., Anderson, R., and Gómez-Gualdrón, D.A. (2019). Attainable Volumetric Targets for Adsorption-Based Hydrogen Storage in Porous Crystals: Molecular Simulation and Machine Learning. J. Phys. Chem. C *123*, 120–130.
6. 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.
7. 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 CO 2 Working Capacity and CO2/CH4 Selectivity for Methane Purification. 4505–4511.
8. Fernandez, M., and Barnard, A.S. (2016). Geometrical Properties Can Predict CO $_2$ and N $_2$ Adsorption Performance of Metal–Organic Frameworks (MOFs) at Low Pressure. ACS Comb. Sci. *18*, 243–252.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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 CO 2 Capture. J. Phys. Chem. Lett. *5*, 3056–3060.
14. Fischer, M., Hoffmann, F., and Fröba, M. (2009). Preferred hydrogen adsorption sites in various MOFs-A comparative computational study. ChemPhysChem *10*, 2647–2657.
15. Feynman, R.P., and Hibbs, A.R. (1965). Quantum mechanics and path integrals (McGraw-Hill).
16. Ahmed, A., Liu, Y., Purewal, J., Tran, L.D., Veenstra, M., Wong-Foy, A., Matzger, A., and Siegel, D. (2017). Balancing Gravimetric and Volumetric Hydrogen Density in MOFs. Energy Environ. Sci. *10*, 2459–2471.
17. Lorentz, H.A. (1881). Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. Ann. Phys. *248*, 127–136.
18. Sandler, S.I. (2006). Chemical, biochemical, and engineering thermodynamics 4th ed. (Wiley).
19. Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard, W.A., and Skiff, W.M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J. Am. Chem. Soc. *114*, 10024–10035.
20. Mayo, S.L., Olafson, B.D., and Goddard III, W.A. (1990). DREIDING: A Generic Force Field for Molecular Simulations. J . Phys. Chem *94*, 8897–8909.
21. Allen, M.P., and Tildesley, D.J. (1989). Computer simulation of liquids (Oxford University Press).
22. Sadus, R.J. (1999). Molecular simulation of fluids: theory, algorithms, and object-orientation. (Elsevier).
23. Dubbeldam, D., Calero, S., Ellis, D.E., and Snurr, R.Q. (2016). RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. Mol. Simul. *42*, 81–101.
24. García-Holley, P., Schweitzer, B., Islamoglu, T., Liu, Y., Lin, L., Rodriguez, S., Weston, M.H., Hupp, J.T., Gómez-Gualdrón, D.A., Yildirim, T., et al. (2018). Benchmark Study of Hydrogen Storage in Metal–Organic Frameworks under Temperature and Pressure Swing Conditions. ACS Energy Lett., 748–754.
25. Zwillinger, D., Kokoska, S., Raton, B., New, L., and Washington, Y. (2000). standard probability and Statistics tables and formulae CRC.
26. Oliphant, T.E. (2007). Python for Scientific Computing. Comput. Sci. Eng. *9*, 10–20.
27. Millman, K.J., and Aivazis, M. (2011). Python for Scientists and Engineers. Comput. Sci. Eng. *13*, 9–12.
28. Abramowitz, M., and Stegun, I.A. (1965). Handbook of mathematical functions, with formulas, graphs, and mathematical tables, (Dover Publications).
29. Kendall, M.G. (1945). The Treatment of Ties in Ranking Problems. Biometrika *33*, 239–251.
30. Kendall, M.G. (1938). A New Measure of Rank Correlation. Biometrika *30*, 81–93.
31. Press, W.H. (2007). Numerical recipes : the art of scientific computing (Cambridge University Press).
32. 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.
33. Parrt, T., and Turgutlu, K. rfpimp 1.3.4, https://github.com/parrt/random-forest-importances.