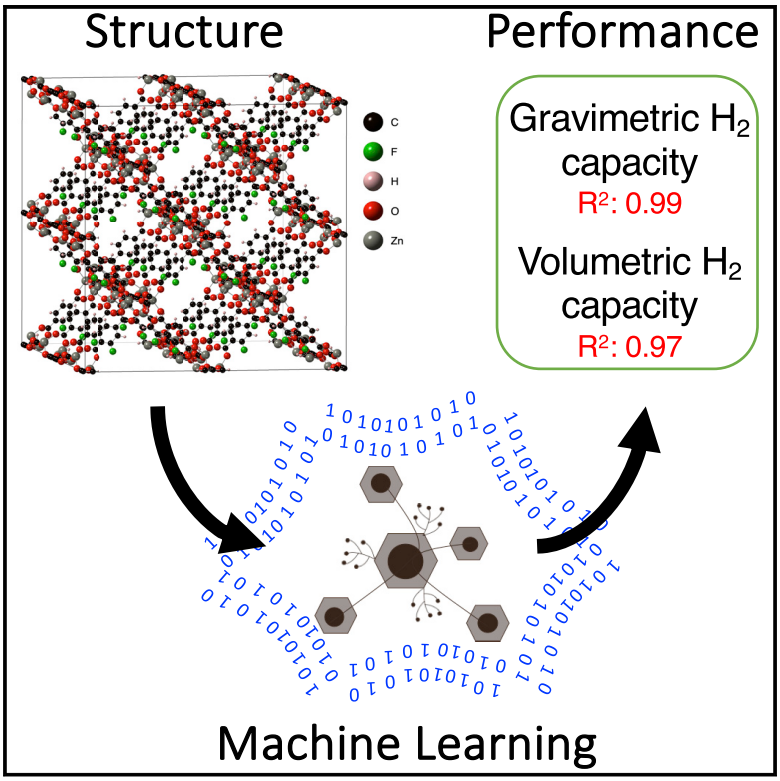
Article

Patterns

Predicting hydrogen storage in MOFs via machine learning

Graphical abstract



Authors

Alauddin Ahmed, Donald J. Siegel

Correspondence

djsiege@umich.edu

In brief

The adoption of hydrogen as a low-carbon fuel has been slowed by the low energy density of H_2 gas. Hydrogen adsorption in MOFs presents a pathway for storing hydrogen at the densities desired for mobile applications, such as fuel cell vehicles. Nevertheless, identifying a suitable MOF remains a challenge because the number of MOFs is essentially limitless. To accelerate the discovery of high-capacity hydrogen adsorbents, machine learning models are developed to predict hydrogen uptake across a diverse set of MOFs.

Highlights

- Accurate and general ML models for predicting H_2 storage in MOFs are developed
- The models require minimal input data that are easily derived from the MOF structure
- High-capacity MOFs are identified, and capacity-structure connections are revealed
- The web models (https://sorbent-ml.hymarc.org) can predict the performance of new MOFs

Ahmed & Siegel, 2021, Patterns 2, 100291
July 9, 2021
https://doi.org/10.1016/j.patter.2021.100291

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Predicting hydrogen storage in MOFs via machine learning

Alauddin Ahmed[1] and Donald J. Siegel[1,2,3,4,5,*]
[1]Mechanical Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA
[2]Materials Science & Engineering, University of Michigan, Ann Arbor, MI 48109, USA
[3]Applied Physics Program, University of Michigan, Ann Arbor, MI 48109, USA
[4]University of Michigan Energy Institute, University of Michigan, Ann Arbor, MI 48109, USA
[5]Lead contact
*Correspondence: djsiege@umich.edu
https://doi.org/10.1016/j.patter.2021.100291

THE BIGGER PICTURE The efficient storage of hydrogen fuel remains a barrier to the adoption of fuel cell vehicles. Although many storage technologies have been proposed, adsorptive storage in metal-organic frameworks (MOFs) holds promise due to the low operating pressures, fast kinetics, reversibility, and high gravimetric densities typical of MOFs. Nevertheless, the volumetric storage densities of known MOFs are generally low; hence, new MOFs with improved volumetric performance are desired. Identifying optimal MOFs remains a challenge, however, because relatively few MOFs have been characterized experimentally, and the building-block structure of MOFs suggests that the number of possible materials is limitless. To accelerate the discovery process, this study develops machine learning models that predict the hydrogen capacity of MOFs. The models identify promising materials, clarify structure-property relations, and can be used—on the web or through an API—to predict the performance of new MOFs.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

The H_2 capacities of a diverse set of 918,734 metal-organic frameworks (MOFs) sourced from 19 databases is predicted via machine learning (ML). Using only 7 structural features as input, ML identifies 8,282 MOFs with the potential to exceed the capacities of state-of-the-art materials. The identified MOFs are predominantly hypothetical compounds having low densities (<0.31 g cm^{-3}) in combination with high surface areas (>5,300 m^2 g^{-1}), void fractions (~0.90), and pore volumes (>3.3 cm^3 g^{-1}). The relative importance of the input features are characterized, and dependencies on the ML algorithm and training set size are quantified. The most important features for predicting H_2 uptake are pore volume (for gravimetric capacity) and void fraction (for volumetric capacity). The ML models are available on the web, allowing for rapid and accurate predictions of the hydrogen capacities of MOFs from limited structural data; the simplest models require only a single crystallographic feature.

INTRODUCTION

Hydrogen (H_2) is considered to be a future automotive fuel.[1–6] This potential reflects its high specific energy compared with competing fuels, such as natural gas and gasoline, and the ability of H_2 to be produced renewably and consumed without CO_2 emissions.[2,7] Nevertheless, the adoption of hydrogen in mobile applications, such as fuel cell (FC) vehicles has been limited by its low volumetric energy density.[2,6,7] Consequently, the design of low-cost H_2 storage systems that overcome these volumetric limitations has been the focus of recent research.[4,8–12] At present, FC vehicles employ storage systems based on gaseous H_2 compressed to pressures up to 700 bar.[13] This approach is costly and can incur limitations in driving range.[7,11,13,14]

Storage based on adsorption in porous hosts is an alternative to high-pressure compression.[15] Due to their high gravimetric densities, fast kinetics, and reversibility, metal-organic frameworks (MOFs) have emerged as one of the most promising

Table 1. Summary of recent studies that use machine learning to predict H_2 adsorption in MOFs

Study	ML features	ML method	Properties predicted	Accuracy
Anderson et al.[43]	epsilon, temperature, pressure, ρ_{crys}, vf, vsa, mpd, lcd, alchemical catecholate site density, unit cell volume	neural network[76]	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.[80]	energetics of MOF-guest interactions	multilinear regression with LASSO[76]	deliverable H_2 storage capacity between 2 and 100 bar at 77 K	R^2 = 0.96; AUE = 1.4–3.4 g-H_2 L^{-1}; RMSE = 3.1–4.4 g-H_2 L^{-1}
Borboudakis et al.[63]	92 binary features based on linker, metal cluster, and 12 functional groups	ridge linear regression and support vector machine with polynomial/Gaussian kernel[76–78]	total H_2 storage capacity at 1 bar and 77 K	AUE = 0.47 (ridge regression), 0.50 (SVM) g-H_2 g^{-1}-MOF
Thornton et al.[61]	adsorption energy, ρ_{crys}, vf, gsa, vsa, lcd	neural network[76]	net H_2 capacity for pressure swing between 1 and 100 bar at 77 and 298 K	R^2 = 0.88; RMSE = 3.6 g-H_2 L^{-1}

ρ_{crys}, vf, vsa, mpd, lcd represent single-crystal density, void fraction, volumetric surface area, maximum pore diameter, and largest cavity diameter, respectively. R^2, AUE, and RMSE represent the coefficient of determination, average unsigned error, and root-mean-square error, respectively.

classes of hydrogen sorbents.[2,7] MOFs are crystalline materials formed by the self-assembly of inorganic metal clusters and organic linkers.[16–22] By virtue of their building-block structure and the large number of potential components, the number of MOFs is potentially limitless.[21–25] Further modifications to MOF chemistry can be achieved by introducing functional groups, substituting different metals, and by mixing metals and/or linkers.[26–28]

Despite these many possibilities, a relatively small fraction of MOFs have been synthesized.[29,30] While the crystal structures of these "real" MOFs are available in the Cambridge Structural Database (CSD),[29,30] many exhibit disorder, missing atoms, or have negligible porosity; consequently, these materials are not immediately amenable to assessment via computational modeling.[29,31–35]

One way to bypass these complications is through computational design. To date, nearly a million "hypothetical" MOFs have been reported,[1,36–46] and it is reasonable to expect that many more materials will be proposed.[47–51] High-throughput screening using Grand Canonical Monte Carlo (GCMC)[52–56] has been successful in identifying promising candidates with superior gas storage capacities on sub-sets of these catalogs.[36,38,39,46,50,57–60] Nevertheless, given the large number of possibilities, a systematic search across all of these materials is challenging even with high-throughput techniques.[1,61] Furthermore, differences in the implementation (i.e., use of different temperature/pressure conditions or interatomic potentials) can complicate comparisons between screening studies. Thus, more efficient and consistent screening approaches are desirable for predicting the gas storage properties of MOFs in existing and future databases.

Machine learning (ML) could provide a path forward.[62–65] For ML to be helpful, access to high-quality training data is essential. Unfortunately, training on experimental H_2 storage data in MOFs is non-trivial[1,2,6,66–68]: experimental uptake data are generally restricted to a relatively small number of MOFs, and can depend sensitively upon the experimental conditions and the purity of the sample.[2,67,69] Employing a dataset based on a consistent set of computational predictions may be a better choice.[62,63]

Earlier work has demonstrated that accurate isotherms for H_2 uptake in MOFs can be predicted using the pseudo-Feynman-Hibbs potential (to describe H_2) combined with general interatomic potentials to describe the MOF.[1,2,6,68] This approach was used to screen a database of 5,309 real MOFs, from which IRMOF-20 was identified and experimentally demonstrated to have a favorable balance of high gravimetric and volumetric H_2 density.[2] In a follow-on study, a larger database of 495,305 MOFs was compiled from several publicly available databases (see Table S1 for details).[1,29,31,33,36–40,45] Following a pre-screen based on crystallographic properties and empirical correlations, the H_2 capacities of a subset of 43,777 MOFs were evaluated using GCMC. Three additional MOFs—SNU-70, UMCM-9, and PCN-610/NU-100—were identified and shown experimentally to out-perform the leading MOF candidate, IRMOF-20.[1]

The database of MOF properties[70] generated in these previous studies presents an opportunity to develop ML models that can predict H_2 uptake across even larger MOF datasets.[1,70] Table 1 summarizes previous ML studies of H_2 storage in MOFs. (Reports employing ML for other adsorbates, such as CH_4,[71,72] CO_2,[73,74] and N_2[73,74] are summarized in Table S2.) To the best of our knowledge, ML was first used to predict H_2 uptake in compounds from the Nanoporous Materials Genome.[75] A neural network (NN)[76] was used to predict usable capacities on a test set of ~1,000 compounds, including MOFs.[61] In the same year, Borboudakis et al.[63] predicated H_2 capacities in 100

Table 2. MOF datasets employed in this study

Source	Database identity	No. of MOFs
Goldsmith et al.,[31] Chung et al.,[33] Moghadam et al.,[29] Groom et al.[30]	real MOFs: UM[31]+CoRE[33]+CSD[29,30]	15,235
Chung et al.[34]	CoRE 2019[34]	14,142
Moghadam et al.,[29] Groom et al.[30]	[a]CSD 2017 additional[29,30]	48,696
Martin et al.[38]	mail-order[38]	112
Bao et al.[46]	*in silico* deliverable[46]	2,816
Bao et al.[39]	*in silico* surface[39]	8,885
Witman et al.[40]	MOF-74 analogs[40]	61
Colón et al.[59]	ToBaCCo[59]	13,512
Gomez-Gualdron et al.[45]	Zr-MOFs[45]	204
Wilmer et al.[36]	Northwestern[36]	137,000
Aghaji et al.,[37] Boyd et al.[85,86]	[b]Univ. of Ottawa[37,85,86]	317,462
Lan et al.[81]	BJT MOFs[81]	303,793
Chung et al.[41,87]	[c]R-WLLFHS[41,87]	51,163
Li et al.[82]	MTV[82]	11,555
Anderson et al.[42]	CSM-2018-I[42]	117
Anderson et al.[43]	CSM-2018-II[43]	32
Anderson et al.[44]	CSM-2019-I[44]	99
Ahmed et al.[1]	in-house[1]	18
	total	918,734

[a]A subset of the CSD 2017 MOF dataset[29,30] whose crystallographic properties were found to exhibit extremely low values (e.g., GSA ~0) in a previous study.
[b]A recent version of this database is available publicly;[85,86] however, this study employs an earlier version[37] that was shared privately.
[c]A curated subset of the Northwestern[36] database.

MOFs using 92 binary features related to a MOF's linker, metal cluster, and functional group(s). Ridge linear regression (RR)[76–78] and support vector machine (SVM)[76,79] algorithms were used to predict gravimetric capacity. Later, Bucior et al[80] predicted the H_2 capacities of 54,776 MOFs extracted from the CSD using multilinear regression (MLR).[76] The models were trained using the energetics of H_2-MOF interactions and the usable volumetric capacities predicted by GCMC. More recently, ML was used to predict H_2 storage capacities in 105 hypothetical MOFs constructed from 17 different topologies, 4 distinct metal clusters, and 5 unique organic linkers.[43] NN[76] models employing 11 features were trained to predict total volumetric uptake at various temperatures and pressures.[43]

Expanding upon these previous reports, this study applies ML to explore a large database of 918,734 known and proposed MOFs. The database was assembled from a diverse collection of publicly available MOF repositories,[1,29,31,33,34,36–45,81,82] and allows for a wide-ranging and consistent assessment of H_2 uptake in MOFs.

Here, the extremely randomized trees (ERT)[76,83] algorithm was identified as the most accurate ML model for predicting H_2 uptake. A training set comprising 24,674 MOFs was sufficient to enable accurate predictions of usable capacities across 820,039 unseen compounds.[70] These predictions were made using a small set of seven crystallographic features as input: single-crystal density, pore volume, gravimetric and volumetric surface area, void fraction, largest cavity diameter, and pore limiting diameter. Importantly, ML identified 8,282 MOFs—8,187 appropriate for pressure swing (PS) operation and 95 for temperature-PS (TPS) use—with the potential to exceed both the gravimetric and volumetric capacities of state-of-the-art materials. These compounds are comprised predominantly of hypothetical MOFs, and exhibit low densities (<0.31 g cm^{-3}) in combination with high surface areas (>5,300 m^2 g^{-1}), void fractions (~0.90), and pore volumes (>3.3 cm^3 g^{-1}). In addition to identifying high-capacity MOFs, the relative importance of the input features is quantified; dependencies on the ML algorithm and training set size and are also assessed. The most important features for predicting H_2 uptake are pore volume (for gravimetric capacity) and void fraction (for volumetric capacity). A simplified model using only two input features is demonstrated to predict capacities with high accuracy—within 0.2 wt % and 1.4 g-H_2 L^{-1} of more expensive Monte Carlo calculations. The ML models are available for use via the web,[84] allowing for rapid and accurate predictions of hydrogen capacities with only a small amount of structural data required as input.

Methods

MOF database

A database of crystal structures for 918,734 MOFs was created by combining 19 existing databases.[1,29,31,33,34,36–45,81,82] Table 2 summarizes the source databases and the number of MOFs contained in each. Out of these 19 databases, only the UM,[31] CSD,[29,30] and CoRE[33,34] databases contain data on MOFs that have been previously synthesized. (MOFs listed in these datasets are referred to as "real" MOFs.) The remaining databases contain data for proposed, or "hypothetical", MOFs. The seven crystallographic properties for all MOFs in the database were calculated using the zeo++ code[25,47] with a probe radius of 1.86 Å. These data are available at the HyMARC data hub.[70] Additional details can be found in our previous work.[1] These properties include: single-crystal density (d), pore volume (pv), gravimetric surface area (gsa), volumetric surface area (vsa), void fraction (vf), largest cavity diameter (lcd), and pore limiting diameter (pld).

A previous study examined a subset of the present database, wherein the hydrogen uptake in 495,305 MOFs was estimated using the Chahine rule.[1,2,70] Subsequently, usable uptake in a portion of this subset comprising 43,777 MOFs predicted to be promising based on the Chahine rule was evaluated using GCMC. This GCMC-evaluated dataset contained a mix of real and hypothetical MOFs: 15,235 real MOFs were sourced from the UM,[31] CoRE,[33] and CSD,[29,30] and 28,542 hypothetical MOFs were extracted from the mail-order,[38] *in silico* deliverable,[46] *in silico* surface,[39] MOF-74 analogs,[40] ToBaCCo,[59] Zr-MOFs,[45] Northwestern,[36] University of Ottawa,[37,85,86] and in-house[1] hypothetical MOF databases (see Ahmed et al.[1] or Table S1 for details).[1,29,31,33,36–40] Hydrogen uptake isotherms for two operating conditions were predicted: for an isothermal PS at T = 77 K between 5 and 100 bar, and for a combined

Table 3. Machine learning regression algorithms employed in this work

Machine learning algorithm	Abbreviation
Extremely randomized trees[76,83,103,104]	ERT
Boosted decision trees[76,92,102–104]	BDT
Bagging with decision trees[76,90,93,103,104]	B/DT
Random forest[76,90,94,103,104]	RF
Bagging with random forest[76,93,94,103,104]	B/RF
Gradient boosting[76,92,95,102–104]	GB
Decision trees[76,90,103,104]	DT
Nu-support vector machine with radial basis function (RBF) kernel[76,79,90,96,98,103,104]	Nu-SVM/RBF-K
Support vector machine RBF kernel[76,79,90,97,98,103,104]	SVM/RBF-K
Support vector machine with linear kernel[76,79,96,99,103,104]	SVM/L-K
Linear regression[76–78,99,100,103,104]	LR
Ridge regression[76–78,99,100,103,104]	RR
K-nearest neighbors[76,90,101,103,104]	K-NN
AdaBoost[76,92,102–104]	AB

TPS between 77 K/100 bar (filled state) and 160 K/5 bar (empty state). UG and UV capacities were then calculated based on the isotherm data.

In addition to the 43,777 MOFs examined in Ahmed et al.,[1] in this study GCMC isotherms were evaluated for an additional 54,918 MOFs (see Ahmed et al.[1] and Table S1 for further details). These additional MOFs were selected at random from the 495,305-entry HyMARC database and therefore represent a more diverse sampling of the MOF property space. To this dataset, 423,429 additional compounds were added from 7 additional datasets: BJT (Beijing, Jiangsu, Tianjin) MOFs,[81] R-WLLFHS,[41,87] MTV,[82] CSM-2018-I,[42] CSM-2018-II,[43] and CSM-2019-I,[44] and selected MOFs from the CSD 2017 dataset.[29,30] Subsequently, the capacities of the MOFs from these additional datasets were predicted by the ML models without retraining (i.e., no MOFs from these datasets were used for training or testing, and none of their isotherms were evaluated in advance with GCMC). In total, the dataset employed in this study contains H_2 uptake data for 98,695 MOFs[70] and crystallographic property data for 918,734 MOFs.

The present dataset includes approximately 74,000 MOFs having open metal sites (OMS), comprising roughly 8% of the total dataset. As the interatomic potential used in our GCMC calculations is not tuned to capture the unique aspects of the H_2-OMS interaction, it is possible that the calculated capacities for this class of MOFs will be less accurate. Figure S1 and Table S3 compare experiments and the present GCMC calculations of H_2 capacities across a benchmark set of OMS MOFs discussed by Garcia-Holley et al.[88] and in our previous work.[1] These data show that GCMC calculations using the pseudo-Feynman-Hibbs potential are in good agreement with experimental data for these OMS MOFs. The good agreement between theory and experiments is a consequence of the low temperature operating conditions used in our study, combined with the relatively low density of OMS in these MOFs.

ML models

The No Free Lunch Theorem[89] implies that the optimal choice of ML algorithm is problem specific. The differing performance of the algorithms summarized in Tables 1 and S2 is consistent with this notion. Identifying the best algorithm for a given dataset requires comparing multiple ML methods, each with optimized hyperparameters. Unfortunately, few comparisons of ML methods for gas adsorption exist; although dozens of ML algorithms are available,[76–79,83,90–104] only RR,[76–78] MLR,[76] SVM,[76,79] and NN[76] have been examined for predicting H_2 storage.[43,61,63,80,103] This study casts a wider net by comparatively assessing 14 ML algorithms (Table 3).[76–79,83,90–104]

The crystallographic properties of MOFs are known to correlate with H_2 capacities.[2,31,88,105–108] The ML models developed here exploit these correlations by adopting only crystallographic properties as input features. Moreover, the number of features was restricted to a small set comprising seven properties: d, pv, gsa, vsa, vf, lcd, and pld. These are the same properties employed in our previous work.[1,2,47,109] Figure S2 shows the distribution of crystallographic properties for the training, test, and unseen datasets. Also, Table S4 summarizes five descriptive (minimum, maximum, mean, median, and percent of 0's) and two distribution statistics (skew and kurtosis) of all crystallographic features for the training, test, and unseen datasets. (The details regarding these statistics and the definitions of skew and kurtosis can be found in Table S4.) The maxima and minima of the features in the training set establish the validity ranges of the ML models developed here.

The goal of the ML models is to predict four output properties: UG and UV for each of PS and TPS operating conditions. This was accomplished by developing separate ML models for each of the four targeted capacities. Figure S3 illustrates the overall work flow.

The existing dataset of 98,695 MOFs (for which both crystallographic and capacity data are available)[70] was initially split into training and test sets of 74,201 and 24,674 MOFs, respectively, after shuffling the entire dataset.[104] ML algorithms[90,93,94,104] (Table 3) were implemented using the Scikit-learn library.[104] Both scaled and unscaled features were used in training ML models. Ten-fold cross-validation was used to optimize the hyperparameters of each model. The performance of the ML algorithms was assessed by comparing the predicted H_2 capacities with the capacity predicted by GCMC for the MOFs in the test set. The metrics used for the performance assessment of ML models were the R^2, AUE, RMSE, MAE, and τ. Additional details regarding these calculations can be found in supplemental note S2 of the supplemental information.

Dataset size

An obstacle to wider adoption of ML in materials science is the availability of sufficient quantities of high-quality training data.[110,111] Unfortunately, it is not yet clear how much data are needed to construct a useful ML model for a given system. Fernandez et al.[72] found that a reasonable balance between accuracy ($R^2 \sim 0.85$–0.93) and computational expense for predicting methane storage in MOFs was achieved for a training set containing data on 10,000 MOFs with 3 features. In contrast, Fanourgakis et al.[112] showed that a much smaller training set of ~1,000 MOFs was sufficient to predict methane uptake when using six crystallographic features and four fictitious features. The

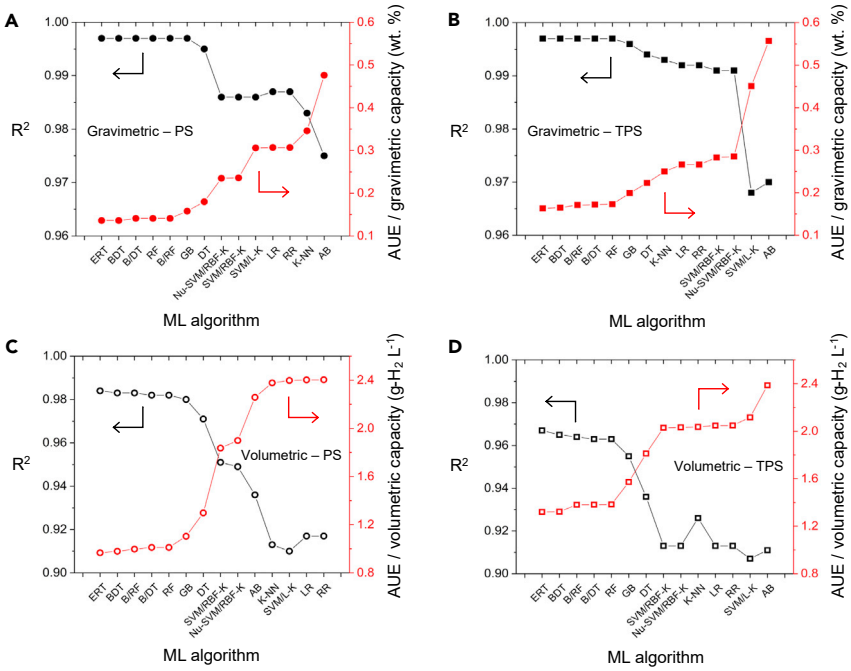


Figure 1. Comparison of ML algorithms for predicting hydrogen uptake in MOFs

(A and C) Left and (B and D) right panels report performance for PS and TPS conditions, respectively. (A and B) Top and (C and D) bottom panels report performance for usable gravimetric and volumetric capacities, respectively. The abbreviations for the ML methods are defined in Table 3.

different training set sizes required in these previous studies arise from the differing numbers and types of features used.

This study explores this issue further by systematically examining the effect of training set size, and the training set to test set ratio, on ML accuracy. For each of the four targeted capacity outputs, 100 independent ML models were developed by varying the size of the training set between 100 and 74,000 MOFs (see Table S5 for a list of the training set sizes). The four best-performing ERT ML algorithms identified earlier were used with 10-fold cross-validation. The resulting models were assessed using a common test set of 24,674 MOFs.

Feature importance/selection

The well-known Chahine rule proposes a linear correlation between gravimetric surface area and excess gravimetric H_2 capacity in adsorbents.[113,114] Nevertheless, the Chahine rule overpredicts H_2 capacities for MOFs with high surface areas,[114] and has not been extended to predict usable capacities.[1,2,6] Hence, a model for predicting H_2 uptake that is more general than the Chahine rule, yet requires limited input data, would be very helpful. In principle, ML could be used to generate such a predictive model if the features that are the most important for predicting H_2 uptake could be identified. Along these lines, Pardakhti et al. reported improved accuracy in predicting CH_4 adsorption when using a combination of (7) crystallographic and (19) chemical features.[71] Recently, Moosavi et al. explored feature importance in predicting the synthesis of MOFs.[115]

This study determines the minimum number and optimal combination of crystallographic features necessary to achieve a specified accuracy in predicting H_2 uptake. The relative importance of the input features was assessed for all possible univariate and multivariate feature combinations using ERT ML models. The number of multivariate feature combinations, M, is given by: $M(n_{tot}, n_{sub}) = \frac{n_{tot}!}{n_{sub}!(n_{tot}-n_{sub})!}$, where $n_{tot} = 7$ is the total number of available features, and $1 \leq n_{sub} \leq 7$ is the number of features used as input to a given ML model. A total of 127 feature combinations are possible. ML models were developed for each of these feature combinations for each of the 4 output capacities, resulting in a total of 508 distinct ML models. All models were trained using a dataset of 74,021 MOFs and tested on a common set of 24,674 MOFs. Ten-fold cross-validation was used for tuning and validating the models using only the training set. Univariate feature importance was further assessed using (1) Pearson's correlation coefficient (r),[116–118] (2), Breiman and Friedman's tree-based algorithm as implemented in Scikit-learn,[90,104] and (3) the permutation importance method as implemented in rfpimp package.[119] Additional details regarding these methods can be found in Figure S7.

RESULTS

Evaluating ML algorithms

Tables S6–S9 illustrate the effect of several feature scaling methods on the performance of the ML algorithms examined here. Only the SVM family of models (SVM/L-K, SVM/RBF-K, and Nu-SVM/RBF-K)[76,90,96,98,99,104] were impacted by the choice of scaling method.

Figure 1 compares the accuracy of the ML algorithms for predicting hydrogen uptake in MOFs. Coefficient of determination (R^2) and average unsigned error (AUE) were used as performance metrics. SVM variants were trained using min-max feature scaling; unscaled features were used in training the remaining models. The performance of the algorithms as measured by four additional metrics—root-mean-square error (RMSE), explained variance (EV), median absolute error (MAE), and Kendall rank correlation coefficient (τ)—is reported in Tables S6–S9.

Overall, these data indicate that the tree-based ensemble methods are superior to the other methods examined. In particular, the ERT[76,83,104] algorithm exhibited the best performance overall. Boosted decision trees,[76,90–92,102,104] random forest,[76,94,104] and Bagging algorithm variants[76,93,104,120,121] (with tree-based base estimators) are nearly as accurate. The R^2 values for ERT predictions exceed 0.997 for gravimetric capacities, which are equivalent to errors of ~0.14 wt %. Volumetrically, the accuracy of the ERT algorithm is slightly worse than its gravimetric performance: R^2 = 0.967–0.984, equivalent to errors of ~1.1 g-H_2 L^{-1} on average. In general, the worst-performing algorithms were linear regression, ridge regression, and SVM

Table 4. Performance of the extremely randomized trees ML algorithm in predicting UG and UV H_2 capacities of MOFs under PS and TPS conditions

H_2 capacity type	R^2	AUE (capacity units)	RMSE (capacity units)	Kendall τ	MAE (capacity units)
UG at PS (wt %)	0.997	0.14	0.18	0.961	0.10
UV at PS (g-H_2 L^{-1})	0.984	0.97	1.40	0.922	0.69
UG at TPS (wt %)	0.997	0.16	0.23	0.966	0.10
UV at TPS (g-H_2 L^{-1})	0.967	1.32	1.92	0.819	0.91

R^2, AUE, RSME, and MAE represent the coefficient of determination, average unsigned error, root-mean-squared error, and median absolute error, respectively.

with linear kernel. For these algorithms R^2 varies between 0.913 and 0.992 depending on the conditions (i.e., gravimetric/volumetric and PS/TPS). As expected, the linear nature of these algorithms fails to fully capture the nonlinear dependence of output capacities on the multiple input features.

Figure 1 also shows that all the algorithms tested yield more accurate predictions of usable gravimetric (UG) capacities compared with those for usable volumetric (UV) capacities. Likewise, all algorithms more accurately predict usable capacities under PS conditions than under TPS conditions. This reflects the fact that the functional relationships between output capacities (UG/UV) and input features under PS and TPS conditions are likely different, as was observed in previously reported structure(feature)-property(capacity) relationships.[1,6,122] Table 4 summarizes the performance of the ERT algorithm in further detail. A comparison of Tables 1 and 4 indicates that the accuracy of the present ML models surpass previously reported models for H_2 uptake. Furthermore, the present models also appear to be an improvement over earlier models that aim to predict the adsorption capacities of MOFs for any gas species, Table S2. This improved performance can be attributed to the exploration and optimization of multiple ML algorithms, use of an appropriate feature set, and the relatively large size of the present training set.

Figure 2 illustrates the degree of agreement between ERT ML predictions and GCMC calculations of usable H_2 capacities under PS conditions as a function of MOF source database (Figure S4 shows similar data for TPS conditions; see also Table 4). As mentioned above, the present ML models more accurately predict UG capacities than UV capacities. The largest differences between ML and GCMC capacities (Figures 2C, 2F, S4C, and S4F) primarily occur for the real MOF dataset. In principle, these differences may arise either from ML overfitting or from inaccurate GCMC predictions caused by non-ideal/incomplete MOF crystal structure data (i.e., missing atoms, disorder, etc.), as mentioned in previous studies.[1,32,35,123–125] ERT algorithms are fairly robust against overfitting.[83] To examine the possibility for overfitting, test set errors were compared with training set errors, as shown in Figure S5 and Table 4. These data suggest that the outliers are not a consequence of over fitting; hence, inaccuracies in the crystal structure data are proposed as the most likely source of this disagreement.[1,32,35,123–125]

Effect of training set size

Figure 3 illustrates the impact of training set size on the accuracy of the ERT ML models, as quantified using R^2 and AUE (Table S5 summarizes the dataset sizes used in these plots). For training sets containing more than 5,000 MOFs, R^2 and AUE vary slowly and in a monotonic fashion, with AUE decreasing and R^2 increasing. The accuracy of the models is more sensitive to the size of the training set for smaller training sets containing roughly 5,000 or fewer MOFs. Figure S6 highlights the variation in performance for these smaller training sets.

The trends AUE as a function of training set size can be fit to a power law expression of the form $AUE(m) = \alpha m^{\beta} + \gamma$, wherem represents the size of the training set and β is the power law exponent. Fitting this model to the data shown in Figure 3 reveals that the AUE for UG converges faster with training set size ($\beta = -0.37$ and -0.43) than it does for UV ($\beta = -0.16$ and -0.23). A full tabulation of the power law parameters is given in Table S10. Based on these power law expressions, one can determine the necessary size of the training set to achieve a desired level of accuracy. For example, assuming PS operation, to achieve an AUE of approximately 0.25 wt % and 1.5 g-H_2 L^{-1} requires training set sizes (for UG and UV) of less than 300 MOFs randomly selected from the diverse datasets used here.

Univariate feature importance

Figure 4 illustrates the relative importance of the seven crystallographic features in predicting usable hydrogen uptake in MOFs. Feature importance was determined by developing ERT models for each single feature individually. Additional details for these models are provided in the supplemental information. Based on these models, it is evident that pore volume (pv) and void fraction (vf) are the dominant features in predicting H_2 capacity; these two properties appear as the first- or second-most important single features regardless of operating condition or capacity type. The importance of these features can be rationalized by two factors. First, based on the empirical Chahine rule, the pore volume of an MOF correlates with its excess uptake.[113] Second, pore volume and void fraction are related (since $pv = vf\, d^{-1}$)—MOFs with larger pv have larger vf, and vice versa.[1]

Conversely, the largest cavity diameter (lcd) and volumetric surface area (vsa) are the single features whose ML models yield the lowest accuracy. The relative importance of the individual features for predicting UG capacities is: pv > d > vf > gsa > pld > lcd > vsa. This ordering is the same for PS and TPS conditions. In contrast, the importance ordering for UV capacities differs based on the operating condition. Nevertheless, vf and pv remain the two most important single features for both UV conditions, in that order (Figure 4).

Despite their limited input, the single-feature ML models illustrated in Figure 4 achieve high accuracy. For example, any of the

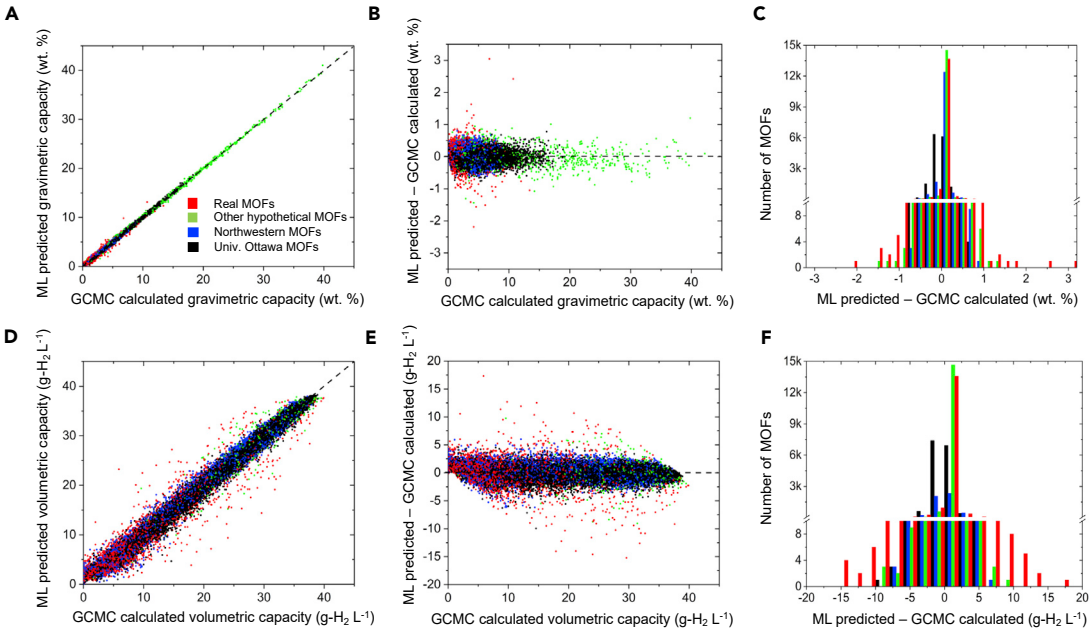


Figure 2. Performance of the ERT algorithm with respect to GCMC calculations for predicting usable H_2 capacities in MOFs

Data were collected at 77 K for a pressure swing (PS) between 100 and 5 bar on a test set of 24,674 MOFs. Different colors represent different categories of MOFs. (A–C) Top and (D–F) bottom panels illustrate performance for usable gravimetric and volumetric capacities, respectively. (A and D) Agreement between ML and GCMC predictions. (B and E) Difference between ML and GCMC as a function of GCMC capacity. (C and F) Distribution of differences in predictions between ML and GCMC.

three independent models for UG-PS based only on pv, d, or vf can predict capacities with $R^2 > 0.95$ and with AUE of less than 0.5 wt %. The accuracy and simplicity of the univariate ML models suggest that they can be used to quickly screen new MOFs for their utility in hydrogen storage. To that end, optimized single-feature ML models for the four categories of usable capacities considered here have been made available for use on the web with an interactive web form or with a python API.[84] Furthermore, the ML models can be downloaded via figshare.[126] These models take as input either pv (for UG predictions) or vf (for UV predictions) of a given MOF. These input data can be quickly calculated from a MOF's crystal structure using modern structure analysis codes.[25,47,127–130] As shown in Figure 4, these models can predict UG with an average error of less than 0.4 wt %, and UV with errors less than 2.2 g-H_2 L^{-1}.

Figure S7 compares the single-feature importance assessments based on ERT ML models (as reported in Figure 4) with three popular methods for determining feature importance: Pearson's correlation coefficient (r),[116–118] Breiman and Friedman's tree-based algorithm as implemented in Scikit-learn,[90,104] and the permutation importance method as implemented in the rfpimp package.[119] It is clear that the feature importance methods do not reproduce *in detail* the rank ordering of feature importance that is suggested by our ERT ML models. Nevertheless, good agreement is evident more broadly. For example, in the case of UG (Figures S7A and S7C), the three feature importance methods suggest that in aggregate pv is the most important feature, while vsa is the least, in agreement with the ERT models (Figures 4A and 4B). Similarly, for UV, the importance methods suggest that vf and lcd are among the most and least important features, respectively. This is the same trend found in the univariate ERT models (Figures 4C and 4D).

Multivariate feature importance

Figure 5 illustrates how the accuracy of the ML models varies with the number and combination of features. Assuming 7 features, $2^7 - 1 = 127$ possible combinations exist. For a given number of features, Figure 5 plots the combination of features resulting in the highest accuracy model. (The supplementary file [Table S11] summarizes the performance for all 508 possible feature combinations and capacity/operating condition types.) As expected, Figure 5 shows that ML accuracy generally increases as the number of input features increases. As previously discussed, when limited to a single feature, vf yields the best accuracy for predicting UV, while pv is the best choice for UG. When the feature set is extended to 2 features, the combination of d and pv is the optimal choice among the $\binom{7}{2} = 21$ possible pairs regardless of the capacity (UG versus UV) or operating condition (PS versus TPS). For larger numbers of features, the optimal feature combination depends upon the operating condition and the capacity type. Based on the AUE, whose value tends to plateau as more features are added, highly accurate ML models can be generated using only 5 input features (Table 5). These data lend further support to the notion that the accuracy of a given ML model depends on both the number and identity of the input features. As a slightly more accurate alternative to the univariate web models described above, a subset of the

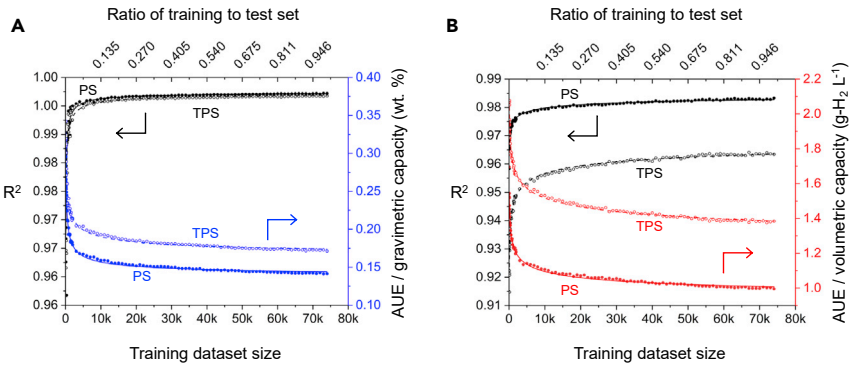


Figure 3. ML performance versus training set size

Performance of ERT ML models for predicting usable (A) gravimetric and (B) volumetric H_2 capacity as a function of training set size and the ratio of training to test set size. One hundred 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 AUE (right axis, blue and red for UG and UV, respectively). Lines represent a power law fit to the data.

present multivariate ML models that use 4, 5, and 7 input features are also available on the web using an interactive web form and via a python API.[84] The ML models can also be downloaded via figshare.[126]

H_2 uptake in unseen MOFs

Figure 6 illustrates the H_2 storage capacities of 820,039 MOFs as predicted by the 7-feature ERT ML models developed here. (This dataset is publicly accessible via HyMARC data hub.[70]) These MOFs are referred to as "unseen", in that they have not been included in the training or test sets used to develop the models. Figures 6A and 6B show UV capacities as functions of UG capacities under PS and TPS conditions, respectively. Both plots exhibit a rapid increase in UV at low values of UG, and reach a maximum in UV at UG values of approximately 9 wt %. Beyond the maximum, UV decreases relatively slowly with increasing UG. These trends are consistent with our earlier findings derived from GCMC calculations on smaller datasets.[1,2,6]

In the case of PS operation, the maximum UV across the MOFs in the dataset is 37.4 g-H_2 L^{-1}; for TPS operation the maximum UV is 48.5 g-H_2 L^{-1}. In the case of UG, the maximum value predicted is 39 wt % for PS operation and 42 wt % for TPS. These values can be placed in context by comparing against the Department of Environment hydrogen storage targets, which stipulate system-level hydrogen densities of 5.5 wt % and 40 g-H_2 L^{-1} by 2025 and 6.5 wt %/50 g-H_2 L^{-1} longer-term ("Ultimate target").[6] Given that the tank and balance-of-plant for the storage system have non-zero mass and volume, the MOFs examined here cannot meet the Ultimate target for UV, regardless of operating condition.[12] More optimism exists, however, for meeting the gravimetric targets given the high UG exhibited by these systems on a MOF-only basis. Of course, an additional challenge is to identify MOFs that excel both gravimetrically and volumetrically.[1,2,6,31,131]

It is also helpful to compare the performance predictions in Figures 6A and 6B with that of state-of-the-art materials. In the case of PS operation, our previous study demonstrated that PCN-610 (NU-100) exhibits a hydrogen capacity of 10.1 wt % and 35.5 g-H_2 L^{-1},[1] which, to our knowledge, is the best combination of gravimetric and volumetric capacities

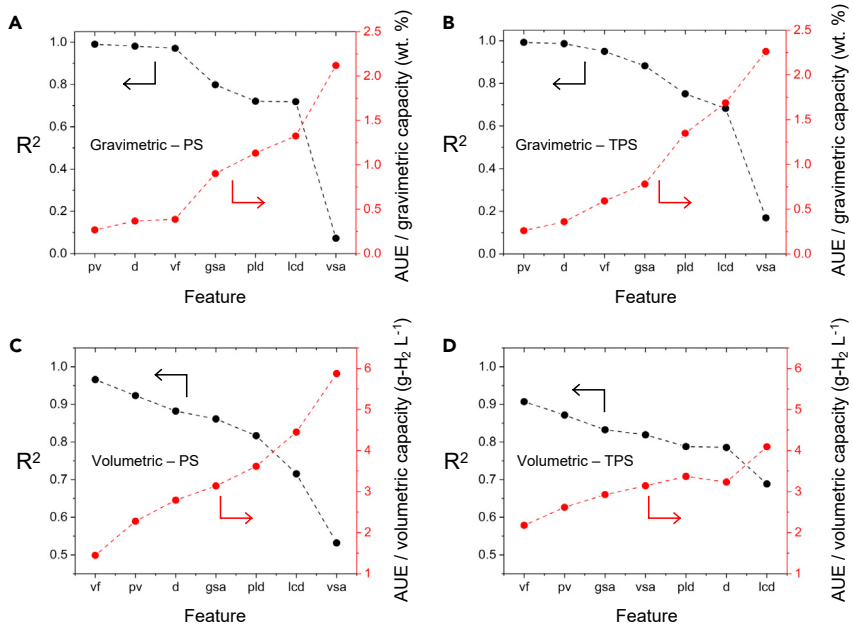


Figure 4. Univariate feature importance in predicting usable H_2 capacities in MOFs

Feature importance was determined by developing distinct ERT models for each individual feature. The accuracy of the resulting models was assessed using R^2 (left axis; black dataset) and AUE (right axis; red dataset). Models were trained on a dataset of 74,201 MOFs and tested on a set of 24,674 MOFs. pv, pore volume; d, density; vf, void fraction; gsa, gravimetric surface area; pld, pore limiting diameter; lcd, largest cavity diameter; vsa, volumetric surface area.

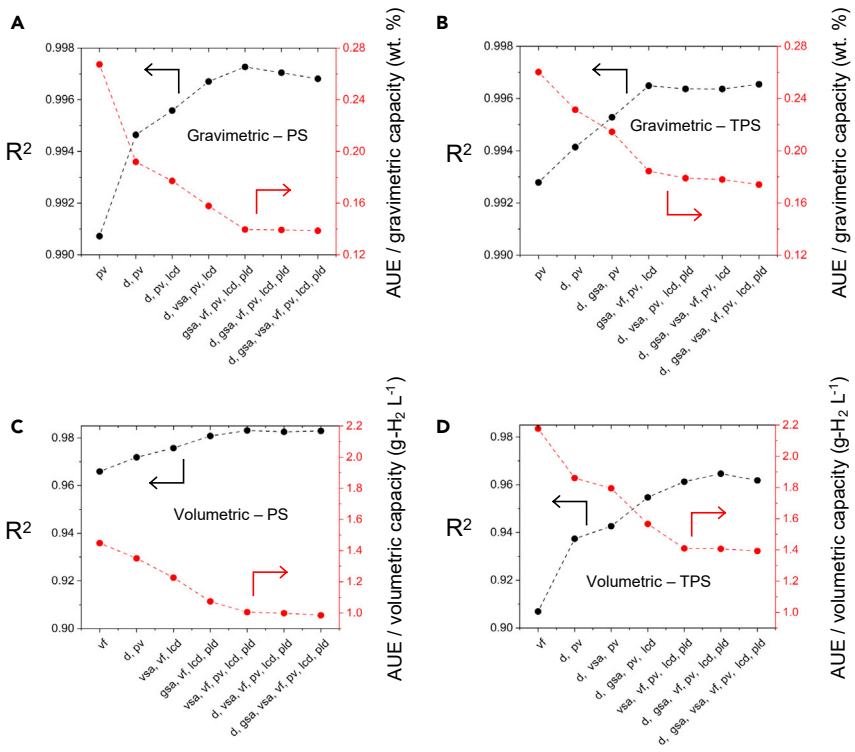


Figure 5. Multivariate feature importance in predicting usable H_2 capacities in MOFs

The accuracy of ERT ML models, as determined by R^2 and AUE, was determined as a function of the number and combination of input features. Each data point represents the most accurate feature combination for a given number of features. ERT models were trained on a dataset of 74,201 MOFs. R^2 and AUE were calculated using a test of 24,674 MOFs. Feature abbreviations are defined in Figure 4.

reported for any MOF under these conditions. The data in Figure 6A reveal that 16,345 MOFs can, in principle, exceed this capacity on both a UG and UV basis. In the case of TPS operation (Figure 6B), MOF-5 remains the benchmark, which a measured capacity of 7.8 wt % and 51.9 g-H_2 L^{-1}.[2] Figure 6D shows that only 21 MOFs out-perform MOF-5 under these conditions.

Regarding the accuracy of the present ML predictions, Table 4 shows that the AUE of these models are on the order of 0.15 wt % and 1.3 g-H_2 L^{-1}. Although these errors are small, a more rigorous validation of the ML can be achieved with GCMC calculations. Thus, GCMC calculations were performed on a subset of MOFs that ML predicted to exhibit high UV and UG capacities. These MOFs fall within the rectangular regions shown in Figures 6A and 6B, and exhibit capacities that meet or exceed 36 g-H_2 L^{-1} and 7.5 wt % for PS conditions and 48 g-H_2 L^{-1} and 7.5 wt % under TPS conditions. In total, 21,700 compounds were re-examined with GCMC based on their ML-predicted PS capacities, and another 7,901 were re-examined for TPS.

Figure 6C compares ML and GCMC predictions for usable capacities for 21,700 high-capacity MOFs under PS conditions. The strong overlap in the two datasets further highlights the accuracy of the ML models. A total of 8,187 MOFs were predicted by GCMC to out-perform PCN-610/NU-100 under these conditions. A summary of the 10 highest-capacity MOFs, sorted based on their GCMC capacities, is provided in Table 6 (a more extensive listing is provided in Table S12). The highest-capacity MOFs are all hypothetical compounds: five originate from the ToBaCCo database,[59] two are from the University of Ottawa database,[37] and the remainder are from the Northwestern[36] database. These MOFs all exhibit high surface areas (average = 5,746, range = 4,346–7,835 m^2 g^{-1}) and large void fractions of 0.89, on average. The range of these property values are consistent with those reported in an earlier study,[1,132,133] and suggest that maximizing the surface area is an important design guideline for PS operation. The highest-capacity MOF, mof_7642,[59] is predicted to exhibit capacities of 11.1 wt % and 40.5 g-H_2 L^{-1}, surpassing that of PCN-610/NU-100, the record-holder under PS conditions. The crystal structure of mof_7642 is shown in Figure 7A.

A search in the CCDC[134] was performed to identify MOFs that have been synthesized that are similar to the high-capacity compounds identified here. The existence of similar MOFs may suggest synthetic procedures that could be adapted to the present systems. The top 5 MOFs under PS conditions contain relatively long tritopic linkers. In the case of mof_7642, this search identified the interpenetrated MOF RANCEQ[135] as having a similar index of 0.82. Interpenetration is fairly common in MOFs (such as mof_7642) with longer linkers, and is generally undesirable for achieving high uptake. Nevertheless, several examples of successful synthesis of MOFs with long, multi-topic linkers that do not undergo interpenetration, have been reported. These include MOF-180 and MOF-200,[136] the PCN-6X series,[137] and NOTT-112.[138] The next four PS candidates in Table 6 exhibit pillared Zn paddlewheel clusters with long ditopic linkers. Karagiaridi et al.[139] demonstrated the feasibility of synthesizing pillared paddlewheel MOFs with long linkers; the SALEM-X series are examples.[139] Finally, str_m3_o5_o20_f0_nbo.sym.1.out is based on a Zn paddlewheel cluster and a ditopic linker. HOFSUS (CSD Refcode) is an example of such a MOF.[140]

Figure 6D provides a similar comparison between ML predictions and GCMC calculations for MOFs expected to exhibit high capacities under TPS conditions. Under these conditions, only 95 MOFs were predicted by GCMC to out-perform MOF-5. A summary of the 10 highest-capacity MOFs, sorted by their GCMC capacities, is provided in Table 6 (see Table S13 for a more extensive tabulation). As found for PS operation, all of the top performing candidates are hypothetical compounds. One difference with the PS case is that all of these MOFs originate from the University of Ottawa database.[37] Furthermore, none of the highest-capacity MOFs identified for PS operation appear as top candidates for TPS. Comparing the highest-capacity MOFs for

Table 5. Optimal combinations of features for predicting UG and UV H_2 storage capacities at PS and TPS conditions

Condition	Feature combination	No. of features	R^2	AUE	RMSE	Kendall τ
UG at PS	gsa, vf, pv, lcd, pld	5	0.997	0.14 wt %	0.19 wt %	0.959
UG at TPS	d, vsa, pv, lcd, pld	5	0.996	0.18 wt %	0.25 wt %	0.959
UV at PS	vsa, vf, pv, lcd, pld	5	0.983	1.01 g-H_2 L^{-1}	1.45 g-H_2 L^{-1}	0.920
UV at TPS	vsa, vf, pv, lcd, pld	5	0.961	1.41 g-H_2 L^{-1}	2.10 g-H_2 L^{-1}	0.814

both operating conditions, it can be seen that the high-capacity TPS MOFs systematically exhibit lower surface areas (average = 4,073 m^2 g^{-1}), smaller void fractions (average = 0.83), and higher densities. Hence, the categories of MOFs that maximize uptake under PS and TPS conditions exhibit distinct properties. These differences suggest that maximizing the surface area—which, as discussed above, is desirable for maximizing PS capacity—is not advantageous for TPS operation. This behavior can be explained by trends in *total* capacities,[6] which the TPS capacities reported here approximate. More specifically, it is known that total volumetric capacities are maximized for intermediate values of the surface area; for larger surface areas the volumetric capacity decreases.

Returning to the list of promising MOFs for TPS operation, Table 6 reports that the highest-capacity MOF, str_m1_o1_o11_f0_pcu.sym.102.out, has a GCMC-predicted capacity of 10.4 wt % and 53.1 g-H_2 L^{-1}. This capacity surpasses that of MOF-5, which, to our knowledge, holds the capacity record under these conditions. The crystal structure of this MOF is shown in Figure 7B.

The top 10 MOFs under TPS conditions contain the same Zn metal cluster and terephthalic acid linkers, where the linkers have been modified with varying functional groups. The slight differences in the capacities of these MOFs can be traced to differences in the functional groups. A similarity search based on str_m1_o1_o11_f0_pcu.sym.117.out identified 40 similar MOFs. Approximately 30 of these (for example, HIFTOG, MIBQAR, UNIGEE, VUSJUP, and ZELROZ) contain Zn metal clusters and linkers based on variants of terephthalic acid.

Figures S8 and S9 and Table S14 quantify the differences between ML and GCMC predictions on the subset of high-capacity MOFs shown in Figures 6C and 6D. For PS operation, the AUE of ML relative to GCMC is 0.24 wt % and 0.66 g-H_2 L^{-1}, while for TPS the AUE is 0.24 wt % and 1.28 g-H_2 L^{-1}. Both sets of errors are comparable with the errors reported in Table 4 for the original test set of MOFs. Figures S8C and S8F and S9(c,f) plot the frequency distribution of the differences between GCMC and ML. These distribution plots suggest that the largest differences occur for predictions involving real MOFs and for hypothetical MOFs extracted from databases other than those from Northwestern,[36] University of Ottawa,[37] and BJT.[81] (These MOFs are referred to as "other hypothetical MOFs" in Figure 6). These MOFs, along with the real compounds, exhibit higher structural diversity than those contained in the other databases. For example, the diversity of the topologies used in the ToBaCCo[59] and Zr-MOFs[45] databases and in the linkers used in MTV-MOF[82] database are larger than what is found in the databases from Northwestern,[36] University of Ottawa,[37] and BJT.[81]

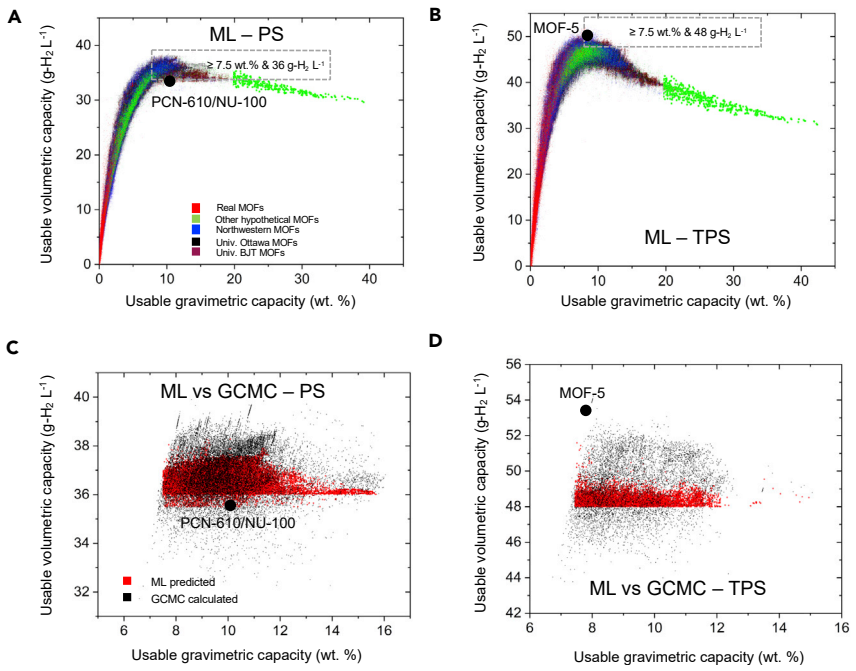


Figure 6. ML predictions of H_2 capacities for 820,093 unseen MOFs

Predicted capacities for (A) PS and (B) temperature + PS operation. Colors indicate the originating database for a given MOF. (C and D) Validation of ML-predicted capacities for the highest-capacity MOFs identified by ML; shown in the rectangular regions in (C and D) using GCMC simulations. For comparison, the capacities of PCN-610/NU-100 (PS: 10.1 wt %, 35.5 g-H_2 L^{-1}) and MOF-5 (TPS: 7.8 wt %, 51.9 g -H_2 L^{-1}) are shown.[1]

Table 6. Highest-capacity MOFs, as identified by ML and verified with GCMC, under pressure swing and temperature + pressure swing conditions

Name	Source	Density (g cm^{-3})	Grav. surface area (m^2 g^{-1})	Vol. surface area (m^2 cm^{-3})	Void fraction	Pore volume (cm^3 g^{-1})	Largest cavity diameter (Å)	Pore limiting diameter (Å)	Usable grav. capacity (wt %)		Usable vol. capacity (g-H_2 L^{-1})	
									GCMC	ML	GCMC	ML
Pressure swing												
mof_7642	ToBaCCo	0.30	5,561	1,695	0.89	2.93	12.8	11.8	11.1	10.3	40.5	37.4
mof_7690	ToBaCCo	0.30	5,715	1,706	0.89	2.98	12.8	12.0	11.3	10.4	40.3	37.3
mof_7594	ToBaCCo	0.40	5,070	2,031	0.86	2.15	11.2	9.7	8.6	7.9	39.9	37.0
mof_7210	ToBaCCo	0.29	5,936	1,730	0.89	3.04	13.4	11.7	11.4	10.5	39.8	37.1
mof_7738	ToBaCCo	0.25	6,054	1,502	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	5,926	1,820	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	5,073	1,583	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	5,818	1,787	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	5,860	1,800	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	4,772	1,882	0.87	2.22	14.1	9.6	8.7	8.1	39.7	37.2
Temperature + pressure swing												
str_m1_o1_o11_f0_pcu.sym.102.out	UO	0.45	4,352	1,974	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	4,162	1,977	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	4,263	2,006	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	4,326	2,005	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	3,703	2,138	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	4,359	1,978	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	3,417	2,035	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	4,330	2,007	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	4,436	1,980	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	3,507	2,043	0.83	1.42	12.4	9.8	8.2	7.7	52.5	48.1

Here, NW and UO refer to the Northwestern[36] and University of Ottawa databases.[37] Grav., gravimetric; Vol., volumetric.

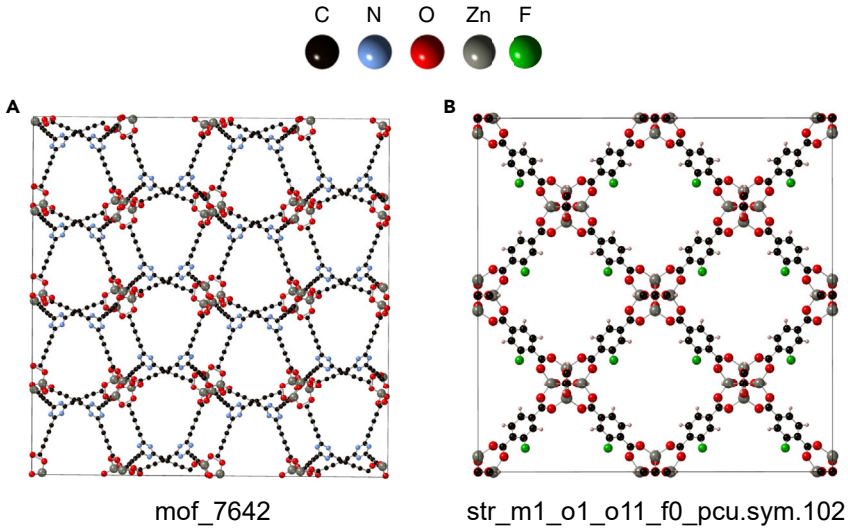


Figure 7. Crystal structures of high-capacity MOFs

Highest-capacity MOFs under (A) PS and (B) temperature + PS conditions. These MOFs originate from the ToBaCCo[59] and University of Ottawa[37] databases, respectively.

DISCUSSION

Limitations of this study

As described previously, some of the high-capacity MOFs identified here may prove difficult to synthesize. Although this limitation applies primarily to the hypothetical MOFs, in some cases real MOFs are also known to undergo framework collapse during activation, which would reduce capacity.[1,2] Nevertheless, future improvements to synthesis techniques may overcome these limitations—what is difficult to make today may be possible in the future. Secondly, our models do not distinguish between realistic MOFs having non-defective crystal structures and those for which the structures are defective/unrealistic. Unrealistic structures can result from incomplete or imperfect virtual solvent removal and the presence of partial occupancies or symmetry disorder in the crystal structure.[31] Consequently, a defective/unrealistic MOF could be erroneously predicted to be a promising candidate. Follow-up calculations using GCMC and visual inspection of the crystal structure are recommended for all promising candidates identified by ML. Finally, the ML models developed here are non-interpretable, "black-box" models. Although these models are demonstrated to be highly accurate, additional effort is required to assess the relative importance of their input data. (The approach demonstrated here for evaluating feature importance involved the development of multiple models with varying numbers and combinations of features.) Alternatively, interpretable white-box ML models could be developed to provide more insight into feature importance. However, our experience suggests that white-box models generate less accurate predictions.

Concluding remarks

The H_2 storage capacities of nearly a million MOFs have been predicted via ML. The predictions span a diverse collection of MOFs sourced from 19 databases and reveal performance under two operating conditions: PS and temperature + PS. More than a dozen ML algorithms were benchmarked, with the ERT method found to be the most accurate. The resulting ML models are accessible on the web at the HyMARC data hub.[84] These models allow for accurate, rapid screening of the hydrogen storage properties of new MOFs using minimal structural data as input; only a single feature is needed for the simplest models.

The accuracy of the ML models was characterized as a function of training set size and the number/combination of input features. Regarding the dependence on the training set, the accuracy of the models can be well described using a simple power law function of the training set size. The dependence on the number and combination of input features was determined by evaluating 508 independent ML models generated from all possible combinations of the seven features. The most important features for predicting H_2 uptake are pore volume (for gravimetric capacity) and void fraction (for volumetric capacity).

Using these models, 8,282 MOFs are identified that have the potential to exceed the capacities of state-of-the-art materials under usable conditions. The identified MOFs are predominantly hypothetical compounds, which (for PS operation) exhibit low densities (<0.31 g cm^{-3}) in combination with high surface areas (>5,300 m^2 g^{-1}), void fractions (~0.90), and pore volumes (>3.3 cm^3 g^{-1}). These MOFs are suggested as targets for experimental synthesis.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Prof. Donald Siegel, djsiege@umich.edu.

Materials availability

This study did not generate new reagents.

Data and code availability

- Original data have been deposited to HyMARC Data Hub: https://datahub.hymarc.org/dataset/computational-prediction-of-hydrogen-storage-capacities-in-mofs
- Interactive ML models: https://sorbent-ml.hymarc.org/
- Python API: https://sorbent-ml.hymarc.org/
- Downloadable ML models and instructions: https://doi.org/10.6084/m9.figshare.14173520.v1

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100291.

ACKNOWLEDGMENTS

Financial support was provided by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, grant no. DE-EE0007046. Computing resources were provided by the NSF via grant 1531752 MRI: Acquisition of Conflux, A Novel Platform for Data-Driven Computational Physics (Tech. Monitor: Ed Walker). The authors acknowledge Jesse Adams, Dr. Zeric Hulvey, Ms. Courtney Pailing, Mr. Nick Wunder, Ms. Nalinrat Guba,

and Dr. Caleb Phillips for facilitating web hosting of the ML models and the development of an application programmers interface. A.A. acknowledges Profs. Randall Snurr and Tom Woo for providing access to their MOF databases; Dr. Maciej Haranczyk for use of the Zeo++ code and the mail-order MOF database; and Prof. Adam J. Matzger, Dr. Antek G. Wong-Foy, Dr. Saona Seth, Dr. Yiyang Liu, Dr. Suresh Kuthuru, and M. Veensra for helpful discussions.

AUTHOR CONTRIBUTIONS

A.A. conducted the computational components of the project. Both authors contributed to the drafting of the paper and to the project idea.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 24, 2021
Revised: May 10, 2021
Accepted: May 26, 2021
Published: June 24, 2021

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., 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.
3. Wong-Foy, A.G., Matzger, A.J., and Yaghi, O.M. (2006). Exceptional H_2 saturation uptake in microporous metal-organic frameworks. J. Am. Chem. Soc. *128*, 3494–3495.
4. Satyapal, S., Petrovic, J., Read, C., Thomas, G., and Ordaz, G. (2007). The U.S. Department of Energy's National Hydrogen Storage Project: progress towards meeting hydrogen-powered vehicle requirements. Catal. Today *120*, 246–256.
5. Greene, D.L., and Duleep, G. (2013). Worldwide Status of Hydrogen Fuel Cell Vehicle Technology and Prospects for Commercialization. U.S. Department of Energy. https://www.hydrogen.energy.gov/pdfs/progress13/xi_1_greene_2013.pdf.
6. Allendorf, M.D., Hulvey, Z., Gennett, T., Ahmed, A., Autrey, T., Camp, J., Seon Cho, E., Furukawa, H., Haranczyk, M., Head-Gordon, M., et al. (2018). An assessment of strategies for the development of solid-state adsorbents for vehicular hydrogen storage. Energy Environ. Sci. *11*, 2784–2812.
7. Yang, J., Sudik, A., Wolverton, C., and Siegel, D.J. (2010). High capacity hydrogen storage materials: attributes for automotive applications and techniques for materials discovery. Chem. Soc. Rev. *39*, 656–675.
8. Long, J.R. (2015). Hydrogen Storage in Metal-Organic Frameworks. . U.S. Department of Energy, Hydrogen and Fuel Cells Program 2015 Annual Merit Review Proceedings: Project ST103. https://www.hydrogen.energy.gov/pdfs/review15/st103_long_2015_o.pdf.
9. U.S. Department of Energy. (n.d.) DOE Technical Targets for Onboard Hydrogen Storage for Light-Duty Vehicles, https://energy.gov/eere/fuelcells/doe-technical-targets-onboard-hydrogen-storage-light-duty-vehicles.
10. Astiaso Garcia, D., Barbanera, F., Cumo, F., Di Matteo, U., and Nastasi, B. (2016). Expert opinion analysis on renewable hydrogen storage systems potential in Europe. Energies *9*, 963.
11. Riis, T., Sandrock, G., Ulleberg, Ø., and Vie, P.J.S. (2006). Hydrogen storage R&D: priorities and gaps. In Hydrogen Production and Storage: R&D Priorities and Gaps (International Energy Agency)), pp. 19–33.
12. Purewal, J., Veenstra, M., Tamburello, D., Ahmed, A., Matzger, A.J., Wong-Foy, A.G., Seth, S., Liu, Y., and Siegel, D.J. (2019). Estimation of system-level hydrogen storage for metal-organic frameworks with high volumetric storage density. Int. J. Hydrogen Energy *44*, 15135–15145.
13. Manoharan, Y., Hosseini, S.E., Butler, B., Alzhahrani, H., Senior, B.T.F., Ashuri, T., and Krohn, J. (2019). Hydrogen fuel cell vehicles; current status and future prospect. Appl. Sci. *9*, 2296.
14. Makridis, S.S. (2016). Hydrogen storage and compression. In Methane and Hydrogen for Energy Storage, R. Carriveau and D.S.-K. Ting, eds. (The Institution of Engineering and Technology)), pp. 1–28.
15. Veenstra, M., Purewal, J., Xu, C., Yang, J., Blaser, R., Sudik, A., Siegel, D., Ming, Y., Liu, D., Hang, C., et al. (2015). Ford/BASF-SE/UM Activities in Support of the Hydrogen Storage Engineering Center of Excellence. U.S. Department of Energy, Office of Scientific and Technical Information, 10.2172/1296578.
16. Öhrström, L. (2015). Let's talk about MOFs—topology and terminology of metal-organic frameworks and why we need them. Crystals *5*, 154–162.
17. Fischer, R.A., and Schwedler, I. (2014). Terminologie von Metall-organischen Gerüstverbindungen und Koordinationspolymeren (IUPAC-Empfehlungen 2013). Angew. Chem. Int. Ed. *126*, 7209–7214.
18. Batten, S.R., Champness, N.R., Chen, X.-M., Garcia-Martinez, J., Kitagawa, S., Öhrström, L., O'Keeffe, M., Paik Suh, M., and Reedijk, J. (2013). Terminology of metal-organic frameworks and coordination polymers (IUPAC Recommendations 2013). Pure Appl. Chem. *85*, 1715–1724.
19. Thommes, M., Kaneko, K., Neimark, A.V., Olivier, J.P., Rodriguez-Reinoso, F., Rouquerol, J., and Sing, K.S.W. (2015). Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (IUPAC Technical Report). Pure Appl. Chem. *87*, 1051–1069.
20. Batten, S.R., Champness, N.R., Chen, X.-M., Garcia-Martinez, J., Kitagawa, S., Öhrström, L., O'Keeffe, M., Suh, M.P., and Reedijk, J. (2012). Coordination polymers, metal-organic frameworks and the need for terminology guidelines. CrystEngComm *14*, 3001.
21. O'Keeffe, M. (2014). Nets, tiles, and metal-organic frameworks. APL Mater. *2*, 124106.
22. Tranchemontagne, D.J., Mendoza-Cortés, J.L., O'Keeffe, M., and Yaghi, O.M. (2009). Secondary building units, nets and bonding in the chemistry of metal-organic frameworks. Chem. Soc. Rev. *38*, 1257.
23. Reymond, J.-L. (2015). The chemical space project. Acc. Chem. Res. *48*, 722–730.
24. Kontijevskis, A. (2017). Mapping of drug-like chemical universe with reduced complexity molecular frameworks. J. Chem. Inf. Model. *57*, 680–699.
25. Martin, R.L., Smit, B., and Haranczyk, M. (2012). Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. J. Chem. Inf. Model. *52*, 308–318.
26. Sun, D., Sun, F., Deng, X., and Li, Z. (2015). Mixed-metal strategy on metal-organic frameworks (MOFs) for functionalities expansion: Co substitution induces aerobic oxidation of cyclohexene over inactive Ni-MOF-74. Inorg. Chem. *54*, 8639–8643.
27. Deng, H., Doonan, C.J., Furukawa, H., Ferreira, R.B., Towne, J., Knobler, C.B., Wang, B., and Yaghi, O.M. (2010). Multiple functional groups of varying ratios in metal-organic frameworks. Science *327*, 846–850.
28. Park, J., Kim, H., Han, S.S., and Jung, Y. (2012). Tuning metal-organic frameworks with open-metal sites and its origin for enhancing CO_2 affinity by metal substitution. J. Phys. Chem. Lett. *3*, 826–829.
29. Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G.P., Wood, P.A., Ward, S.C., and Fairen-Jimenez, D. (2017). Development of a Cambridge structural database subset: a collection of metal-organic frameworks for past, present, and future. Chem. Mater. *29*, 2618–2625.
30. Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge Structural Database. Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater. *72*, 171–179.

31. Goldsmith, J., Wong-Foy, A.G., Cafarella, M.J., and Siegel, D.J. (2013). Theoretical limits of hydrogen storage in metal–organic frameworks: opportunities and trade-offs. Chem. Mater. *25*, 3373–3382.

32. Altintas, C., Avci, G., Daglar, H., Nemati Vesali Azar, A., Erucar, I., Velioglu, S., and Keskin, S. (2019). An extensive comparative analysis of two MOF databases: high-throughput screening of computation-ready MOFs for CH_4 and H_2 adsorption. J. Mater. Chem. A. *7*, 9593–9608.

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

34. Chung, Y.G., Haldoupis, E., Bucior, B.J., Haranczyk, M., Lee, S., Zhang, H., Vogiatzis, K.D., Milisavljevic, M., Ling, S., Camp, J.S., et al. (2019). Advances, updates, and analytics for the computation-ready, experimental metal-organic framework database: CoRE MOF 2019. J. Chem. Eng. Data *64*, 5985–5998.

35. Chen, T., and Manz, T.A. (2020). Identifying misbonded atoms in the 2019 CoRE Metal-Organic Framework Database. RSC Adv. *10*, 26944–26951.

36. Wilmer, C.E., Leaf, M., Lee, C.Y., Farha, O.K., Hauser, B.G., Hupp, J.T., and Snurr, R.Q. (2011). Large-scale screening of hypothetical metal-organic frameworks. Nat. Chem. *4*, 83–89.

37. 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. Eur. J. Inorg. Chem. *2016*, 4505–4511.

38. Martin, R.L., Lin, L.C., Jariwala, K., Smit, B., and Haranczyk, M. (2013). Mail-order metal-organic frameworks (MOFs): designing isoreticular MOF-5 analogues comprising commercially available organic molecules. J. Phys. Chem. C *117*, 12159–12167.

39. Bao, Y., Martin, R.L., Haranczyk, M., and Deem, M.W. (2015). In silico prediction of MOFs with high deliverable capacity or internal surface area. Phys. Chem. Chem. Phys. *17*, 11962–11973.

40. Witman, M., Ling, S., Anderson, S., Tong, L., Stylianou, K.C., Slater, B., Smit, B., and Haranczyk, M. (2016). In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. Chem. Sci. *7*, 6263–6272.

41. Chung, Y.G., Gómez-gualdrón, D.A., Li, P., Leperi, K.T., Deria, P., Zhang, H., Vermeulen, N.A., Stoddart, J.F., You, F., Hupp, J.T., et al. (2016). In Silico Discovery of Metal-Organic Frameworks for Precombustion CO2 Capture Using a Genetic Algorithm. Sci. Adv. *2*, e1600909.

42. 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.

43. 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.

44. Anderson, R., and Gómez-Gualdrón, D.A. (2019). Increasing topological diversity during computational "synthesis" of porous crystals: how and why. CrystEngComm *21*, 1653–1665.

45. Gomez-Gualdron, D.A., Gutov, O.V., Krungleviciute, V., Borah, B., Mondloch, J.E., Hupp, J.T., Yildirim, T., Farha, O.K., and Snurr, R.Q. (2014). Computational design of metal-organic frameworks based on stable zirconium building units for storage and delivery of methane. Chem. Mater. *26*, 5632–5639.

46. Bao, Y., Martin, R.L., Simon, C.M., Haranczyk, M., Smit, B., and Deem, M.W. (2015). In silico discovery of high deliverable capacity metal-organic frameworks. J. Phys. Chem. C *119*, 186–195.

47. Willems, T.F., Rycroft, C.H., Kazi, M., Meza, J.C., and Haranczyk, M. (2012). Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous Mesoporous Mater. *149*, 134–141.

48. Addicoat, M.A., Coupry, D.E., and Heine, T. (2014). AuToGraFS: automatic topological generator for framework structures. J. Phys. Chem. A *118*, 9607–9614.

49. Boyd, P.G., and Woo, T.K. (2016). A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. CrystEngComm *18*, 3777–3792.

50. Gómez-Gualdrón, D.A., Colón, Y.J., Zhang, X., Wang, T.C., Chen, Y.-S., Hupp, J.T., Yildirim, T., Farha, O.K., Zhang, J., and Snurr, R.Q. (2016). Evaluating topologically diverse metal-organic frameworks for cryo-adsorbed hydrogen storage. Energy Environ. Sci. *9*, 3279–3289.

51. Yao, Z., Sánchez-Lengeling, B., Bobbitt, N.S., Bucior, B.J., Kumar, S.G.H., Collins, S.P., Burns, T., Woo, T.K., Farha, O., Snurr, R.Q., et al. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. Nat. Mach. Intell. *3*, 76–86.

52. Sadus, R.J. (1999). Molecular Simulation of Fluids: Theory, Algorithms, and Object-Orientation (Elsevier).

53. Allen, M.P., and Tildesley, D.J. (1989). Computer Simulation of Liquids (Oxford University Press).

54. Frenkel, D., and Smit, B. (2001). Understanding Molecular Simulation: From Algorithms to Applications, 2nd ed. (Academic Press, Inc.).

55. Hill, T.L. (1986). An Introduction to Statistical Thermodynamics (Dover Publications).

56. Dubbeldam, D., Torres-Knoop, A., and Walton, K.S. (2013). Molecular simulation on the inner workings of Monte Carlo codes. Mol. Simul. *39*, 14–15.

57. 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.

58. Martin, R.L., Simon, C.M., Smit, B., and Haranczyk, M. (2014). *In silico* design of porous polymer networks: high-throughput screening for methane storage materials. J. Am. Chem. Soc. *136*, 5006–5022.

59. Colón, Y.J., Gómez-Gualdrón, D.A., and Snurr, R.Q. (2017). Topologically guided, automated construction of metal-organic frameworks and their evaluation for energy-related applications. Cryst. Growth Des. *17*, 5801–5810.

60. Boyd, P.G., Moosavi, S.M., Witman, M., and Smit, B. (2017). Force-field prediction of materials properties in metal-organic frameworks. J. Phys. Chem. Lett. *8*, 357–363.

61. Thornton, A.W., Simon, C.M., Kim, J., Kwon, O., Deeg, K.S., Konstas, K., Pas, S.J., Hill, M.R., Winkler, D.A., Haranczyk, M., et al. (2017). Materials genome in action: identifying the performance limits of physical hydrogen storage. Chem. Mater. *29*, 2844–2854.

62. Bobbitt, N.S., and Snurr, R.Q. (2019). Molecular Simulation Molecular Modelling and Machine Learning for High-Throughput Screening of Metal-Organic Frameworks for Hydrogen Storage Molecular Modelling and Machine Learning for High-Throughput Screening of Metal-Organic Frameworks for Hydrogen Storage. Mol. Simul. *45*, 1069–1081.

63. Borboudakis, G., Stergiannakos, T., Frysali, M., Klontzas, E., Tsamardinos, I., and Froudakis, G.E. (2017). Chemically intuited, large-scale screening of MOFs by machine learning techniques. NPJ Comput. Mater. *3*. https://doi.org/10.1038/s41524-017-0045-8.

64. Broom, D.P., Webb, C.J., Hurst, K.E., Parilla, P.A., Gennett, T., Brown, C.M., Zacharia, R., Tylianakis, E., Klontzas, E., Froudakis, G.E., et al. (2016). Outlook and challenges for hydrogen storage in nanoporous materials. Appl. Phys. A *122*, 151.

65. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.

66. Wahiduzzaman, M., Walther, C.F.J., and Heine, T. (2014). Hydrogen adsorption in metal-organic frameworks: the role of nuclear quantum effects. J. Chem. Phys. *141*, 064708.

67. Durette, D., Bénard, P., Zacharia, R., and Chahine, R. (2016). Investigation of the hydrogen adsorbed density inside the pores of MOF-5 from path integral grand canonical Monte Carlo at supercritical and subcritical temperature. Sci. Bull. *61*, 594–600.

68. Fischer, M., Hoffmann, F., and Fröba, M. (2009). Preferred hydrogen adsorption sites in various MOFs—a comparative computational study. ChemPhysChem *10*, 2647–2657.

69. Furukawa, H., Miller, M.A., and Yaghi, O.M. (2007). Independent verification of the saturation hydrogen uptake in MOF-177 and establishment of a benchmark for hydrogen adsorption in metal-organic frameworks. J. Mater. Chem. *17*, 3197.

70. Ahmed, A., and Siegel, D.J. (2019). HyMARC Datahub. https://datahub.hymarc.org/dataset/computational-prediction-of-hydrogen-storage-capacities-in-mofs.

71. 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.

72. 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.

73. Fernandez, M., Trefiak, N.R., and Woo, T.K. (2013). Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. J. Phys. Chem. C *117*, 14095–14105.

74. 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.

75. Nanoporous Materials Genome Center. http://www.chem.umn.edu/nmgc/.

76. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning (Springer).

77. Dorugade, A.V., and Kashid, D.N. (2010). Alternative Method for Choosing Ridge Parameter for Regression. Appl. Math. Sci. *4*, 447–456.

78. Van Wieringen, W.N. (2020). Lecture Notes on Ridge Regression. arXiv, 1509.09169 https://arxiv.org/abs/1509.09169.

79. Smola, A.J., Smola, A.J., and Schölkopf, B. (2004). A tutorial on support vector regression. Stat. Comput. *14*, 199–222.

80. Bucior, B.J., Bobbitt, N.S., Islamoglu, T., Goswami, S., Gopalan, A., Yildirim, T., Farha, O.K., Bagheri, N., and Snurr, R.Q. (2019). Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. Mol. Syst. Des. Eng. *4*, 162–174. https://doi.org/10.1039/c8me00050f.

81. Lan, Y., Yan, T., Tong, M., and Zhong, C. (2019). Large-scale computational assembly of ionic liquid/MOF composites: synergistic effect in the wire-tube conformation for efficient CO_2/CH_4 separation. J. Mater. Chem. A *7*, 12556–12564.

82. Li, S., Chung, Y.G., Simon, C.M., and Snurr, R.Q. (2017). High-throughput computational screening of multivariate metal-organic frameworks (MTV-MOFs) for CO_2 capture. J. Phys. Chem. Lett. *8*, 19.

83. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. *63*, 3–42.

84. 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/.

85. Boyd, P.G., Chidambaram, A., García-Díez, E., Ireland, C.P., Daff, T.D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S.M., Maroto-Valer, M.M., et al. (2019). Data-driven design of metal-organic frameworks for wet flue gas CO_2 capture. Nature *576*, 253–256.

86. Boyd, P.G., Chidambaram, A., García-Díez, E., Ireland, C.P., Daff, T.D., Bounds, R., Gładysiak, A., Schouwink, P., Moosavi, S.M., Maroto-Valer, M.M., et al. (2019). Data-driven design of metal-organic frameworks for wet flue gas CO_2 capture, Materials Cloud Archive 2018.0016/v3 (2019). Nature *576*, 253–256. https://doi.org/10.24435/materialscloud:2018.0016/v3.

87. Snurr, R.Q. (2016). Reduced-mHOF-database. https://github.com/snurr-group/Reduced-hMOF-database.

88. 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.

89. Wolpert, D.H., and Macready, W.G. (1997). No free lunch theorems for optimization. IEEE Trans. Evol. Comput. *1*. https://doi.org/10.1109/4235.585893.

90. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (2017). Classification and Regression Trees (Routledge).

91. Freund, Y., and Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. *55*, 119–139.

92. Drucker, H. (1997). Improving regressors using boosting techniques. In ICML '97 Proc. Fourteenth Int. Conf. Mach. Learn, pp. 107–115. https://dl.acm.org/doi/10.5555/645526.657132.

93. Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

94. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

95. Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232.

96. Chang, C.-C., and Lin, C.-J. (2001). LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology *2*, 1–27.

97. Platt, J.C., and Platt, J.C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers (MIT Press), pp. 61–74.

98. Buhmann, M.D. (2002). Radial Basis Functions: Theory and Implementations (Cambridge University Press).

99. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. J. Mach. Learn. Res. *9*, 1871–1874.

100. Rifkin, R.M., and Lippert, R.A. (2007). Notes on Regularized Least Squares. MIT http://128.30.100.62:8080/media/fb/ps/MIT-CSAIL-TR-2007-025.pdf.

101. Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. *46*, 175–185.

102. Freund, Y., and Schapire, R.E. (1999). A short introduction to boosting. J. Jpn. Soc. Artif. Intell. *14*, 771–780.

103. Fernández-Delgado, M., Sirsat, M.S., Cernadas, E., Alawadi, S., Barro, S., and Febrero-Bande, M. (2019). An extensive experimental survey of regression methods. Neural Networks *111*, 11–34.

104. 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.

105. Richard, M.-A., Bénard, P., and Chahine, R. (2009). Gas adsorption process in activated carbon over a wide temperature range above the critical point. Part 1: modified Dubinin-Astakhov model. Adsorption *15*, 43–51.

106. Gomez-Gualdron, D.A., Wang, T.C., García-Holley, P., Sawelewa, R.M., Argueta, E., Snurr, R.Q., Hupp, J.T., Yildirim, T., and Farha, O.K. (2017). Understanding volumetric and gravimetric hydrogen adsorption trade-off in metal-organic frameworks. ACS Appl. Mater. Interfaces *9*, 33419–33428.

107. Düren, T., Bae, Y.-S., and Snurr, R.Q. (2009). Using molecular simulation to characterise metal-organic frameworks for adsorption applications. Chem. Soc. Rev. *38*, 1237.

108. Allendorf, M.D., Bauer, C.A., Bhakta, R.K., and Houk, R.J.T. (2009). Luminescent metal-organic frameworks. Chem. Soc. Rev. *38*, 1330.

109. Gómez-Gualdrón, D.A., Moghadam, P.Z., Hupp, J.T., Farha, O.K., and Snurr, R.Q. (2016). Application of consistency criteria to calculate BET areas of micro- and mesoporous metal-organic frameworks. J. Am. Chem. Soc. *138*, 215–224.

110. Himanen, L., Geurts, A., Foster, A.S., and Rinke, P. (2019). Data-driven materials science: status, challenges, and perspectives. Adv. Sci. 1900808.

111. Wei, J., Chu, X., Sun, X., Xu, K., Deng, H., Chen, J., Wei, Z., and Lei, M. (2019). Machine learning in materials science. InfoMat *1*, 338–358.

112. Fanourgakis, G.S., Gkagkas, K., Tylianakis, E., Klontzas, E., and Froudakis, G. (2019). A robust machine learning algorithm for the prediction of methane adsorption in nanoporous materials. J. Phys. Chem. A. acs.jpca.9b03290. https://doi.org/10.1021/acs.jpca.9b03290.

113. Panella, B., Hirscher, M., and Roth, S. (2005). Hydrogen adsorption in different carbon nanostructures. Carbon N. Y. *43*, 2209–2214.

114. Balderas-Xicohténcatl, R., Schlichtenmayer, M., and Hirscher, M. (2017). Volumetric Hydrogen Storage Capacity in Metal–Organic Frameworks. Energy Technol. *6*, 578–582.

115. Moosavi, S.M., Chidambaram, A., Talirz, L., Haranczyk, M., Stylianou, K.C., and Smit, B. (2019). Capturing chemical intuition in synthesis of metal-organic frameworks. Nat. Commun. *10*, 539.

116. Zwillinger, D., and Kokoska, S. (2000). Standard Probability and Statistics Tables and Formulae (CRC Press).

117. Oliphant, T.E. (2007). Python for scientific computing. Comput. Sci. Eng. *9*, 10–20.

118. Millman, K.J., and Aivazis, M. (2011). Python for scientists and engineers. Comput. Sci. Eng. *13*, 9–12.

119. Parrt, T., and Turgutlu, K.. Rfpimp 1.3.4. https://github.com/parrt/random-forest-importances.

120. Frank, E., Hall, M.A., and Witten, I.H. (2016). The WEKA Workbench. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

121. Kuhn, M., and Johnson, K. (2013). Applied predictive modeling (Springer).

122. Witman, M., Ling, S., Grant, D.M., Walker, G.S., Agarwal, S., Stavila, V., and Allendorf, M.D. (2020). Extracting an empirical intermetallic hydride design principle from limited data via interpretable machine learning. J. Phys. Chem. Lett. *11*, 40–47.

123. Sturluson, A., Huynh, M.T., Kaija, A.R., Laird, C., Yoon, S., Hou, F., Feng, Z., Wilmer, C.E., Colón, Y.J., Chung, Y.G., et al. (2019). The role of molecular modelling and simulation in the discovery and deployment of metal-organic frameworks for gas storage and separation. Mol. Simul. *45*, 1082–1121.

124. Barthel, S., Alexandrov, E.V., Proserpio, D.M., and Smit, B. (2018). Distinguishing Metal-Organic Frameworks. Cryst. Growth Des. *18*, 1738–1747.

125. Chen, T., and Manz, T.A. (2019). A collection of forcefield precursors for metal-organic frameworks. RSC Adv. *9*, 36492–36507.

126. Ahmed, A., and Siegel, D.J. (2021). Machine learning models for predicting hydrogen storage in metal-organic frameworks. Figshare. https://doi.org/10.6084/m9.figshare.14173520.v1.

127. Pinheiro, M., Martin, R.L., Rycroft, C.H., Jones, A., Iglesia, E., and Haranczyk, M. (2013). Characterization and comparison of pore landscapes in crystalline porous materials. J. Mol. Graph. Model. *44*, 208–219.

128. Pinheiro, M., Martin, R.L., Rycroft, C.H., and Haranczyk, M. (2013). High accuracy geometric analysis of crystalline porous materials. CrystEngComm *15*, 7531–7538.

129. Ongari, D., Boyd, P.G., Barthel, S., Witman, M., Haranczyk, M., and Smit, B. (2017). Accurate Characterization of the Pore Volume in Microporous Crystalline Materials. Langmuir *33*, 14529–14538.

130. Sarkisov, L., Bueno-Perez, R., Sutharson, M., and Fairen-jimenez, D. (2020). Material Informatics with PoreBlazer v4.0 and CSD MOF Database. Chem. Mater. *32*, 9849–9867.

131. Chen, Z., Li, P., Anderson, R., Wang, X., Zhang, X., Robison, L., Redfern, L.R., Moribe, S., Islamoglu, T., Gómez-Gualdrón, D.A., et al. (2020). Balancing volumetric and gravimetric uptake in highly porous materials for clean energy. Science *368*, 297–303.

132. Camp, J.S., Stavila, V., Allendorf, M.D., Prendergast, D., and Haranczyk, M. (2018). Critical Factors in Computational Characterization of Hydrogen Storage in Metal-Organic Frameworks Critical Factors in Computational Characterization of Hydrogen Storage in Metal-Organic Frameworks. J. Phys. Chem. C *122*, 18957–18967.

133. Churchard, A.J., Banach, E., Borgschulte, A., Caputo, R., Chen, J.C., Clary, D., Fijalkowski, K.J., Geerlings, H., Genova, R.V., Grochala, W., et al. (2011). A multifaceted approach to hydrogen storage. Phys. Chem. Chem. Phys. *13*, 16955–16972.

134. MacRae, C.F., Sovago, I., Cottrell, S.J., Galek, P.T.A., McCabe, P., Pidcock, E., Platings, M., Shields, G.P., Stevens, J.S., Towler, M., et al. (2020). Mercury 4.0: from visualization to analysis, design and prediction. J. Appl. Crystallogr. *53*, 226–235.

135. Manos, M.J., Markoulides, M.S., Malliakas, C.D., Papaefstathiou, G.S., Chronakis, N., Kanatzidis, M.G., Trikalitis, P.N., and Tasiopoulos, A.J. (2011). A highly porous interpenetrated metal-organic framework from the use of a novel nanosized organic linker. Inorg. Chem. *50*, 11297–11299.

136. Furukawa, H., Ko, N., Go, Y.B., Aratani, N., Choi, S.B., Choi, E., Yazaydin, A.Ö., Snurr, R.Q., O'Keeffe, M., Kim, J., et al. (2010). Ultrahigh porosity in metal-organic frameworks. Science *329*, 424–428.

137. Yuan, D., Zhao, D., Sun, D., and Zhou, H.-C. (2010). An isoreticular series of metal-organic frameworks with dendritic hexacarboxylate ligands and exceptionally high gas-uptake capacity. Angew. Chem. Int. Ed. *49*, 5357–5361.

138. Yan, Y., Telepeni, I., Yang, S., Lin, X., Kockelmann, W., Dailly, A., Blake, A.J., Lewis, W., Walker, G.S., Allan, D.R., et al. (2010). Metal-organic polyhedral frameworks: high H_2 adsorption capacities and neutron powder diffraction studies. J. Am. Chem. Soc. *132*, 4092–4094.

139. Karagiaridi, O., Bury, W., Tylianakis, E., Sarjeant, A.A., Hupp, J.T., and Farha, O.K. (2013). Opening metal-organic frameworks. Vol. 2: inserting longer pillars into pillared-paddlewheel structures through solvent-assisted linker exchange. Chem. Mater. *25*, 3499–3503.

140. Zheng, X., Huang, Y., Duan, J., Wang, C., Wen, L., Zhao, J., and Li, D. (2014). A microporous Zn(II)-MOF with open metal sites: structure and selective adsorption properties. Dalt. Trans. *43*, 8311–8317.

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] ρ_{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	ρ_{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, ρ_{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	ρ_{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	ρ_{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	ρ_{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	ρ_{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	ρ_{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	ρ_{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	ρ_{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:

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

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

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

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

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

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

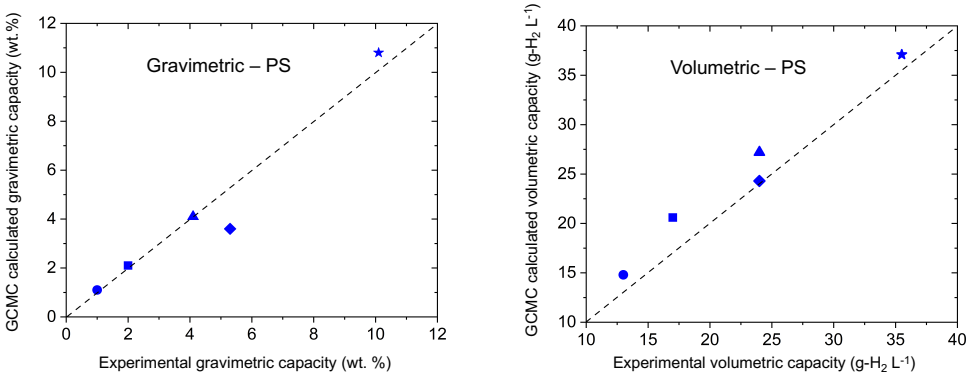


Figure S1. Comparison between experiments and GCMC calculations of H_2 capacities for a benchmark set of open-metal-site MOFs for pressure swing operation: HKUST-1 (■), NOTT-112 (◆), Cu-MOF-74 (●), NU-125 (▲), NU-100/PCN-610 (★).[1,24]

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	2.63×10^{-3}	2.0	2.1	17	20.6
FOPFAS	NOTT-112	9.24×10^{-4}	5.3	3.6	24	24.3
LENKIA	Cu-MOF-74	4.91×10^{-3}	1.0	1.1	13	14.8
REWNEO	NU-125	1.09×10^{-3}	4.1	4.1	24	27.2
HABQUY/GAGZEV	NU-100/ PCN-610	4.47×10^{-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

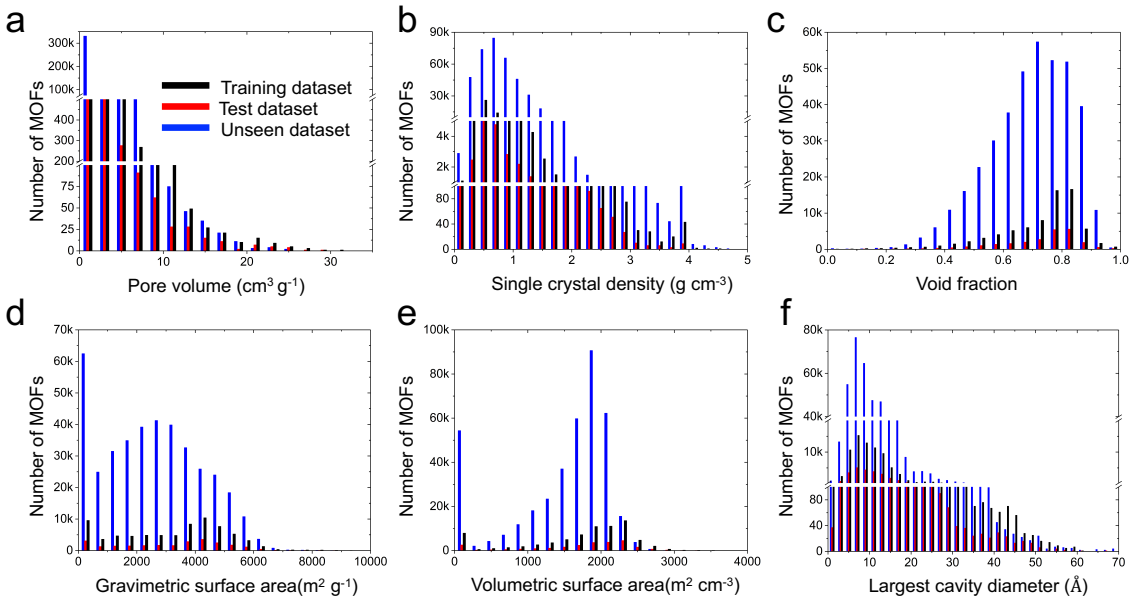


Figure S2. Distribution of 6 crystallographic features in 3 different datasets used in this study. (a) pore volume, (b) single crystal density, (c) void fraction, (d) gravimetric surface area, (e) volumetric surface area, and (f) largest cavity diameter.

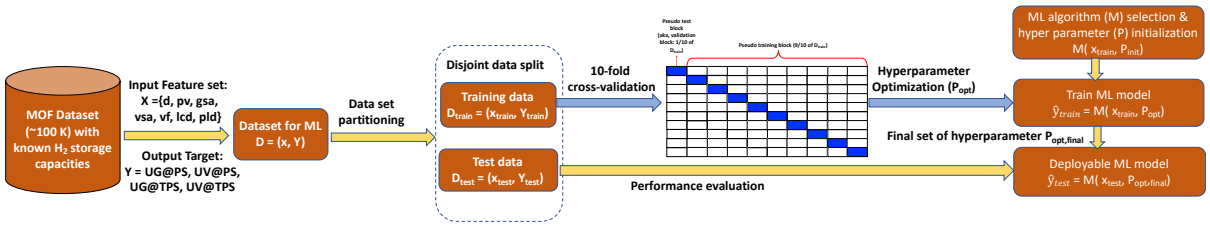


Figure S3. Machine learning work-flow.

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 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

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 τ	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 τ	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 τ	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

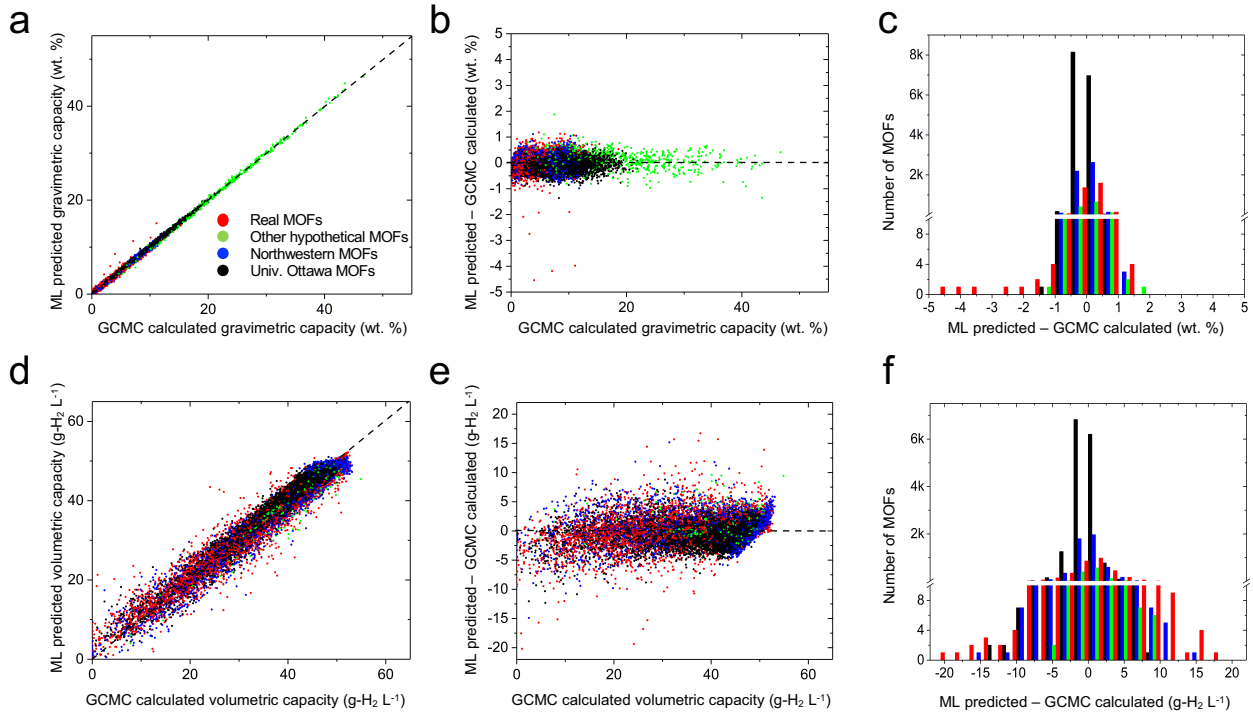


Figure S4. Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable H_2 capacities in MOFs. Data is collected under TPS conditions on a test set of 24,674 MOFs. Different colors represent different categories of MOFs. Top **(a-c)** and bottom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

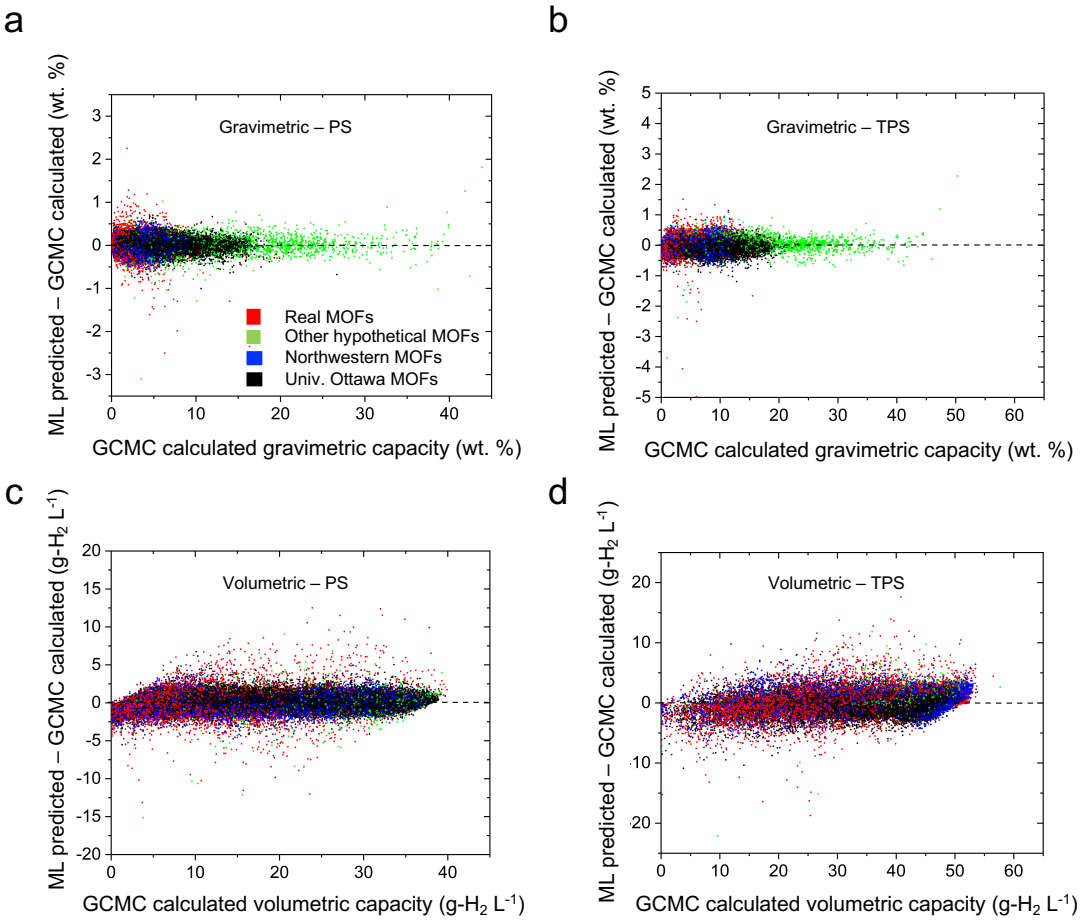


Figure S5. Difference between ML and GCMC as a function of GCMC capacity for the training set of 74,201 MOFs. Performance of the Extremely Randomized Trees ML algorithm with respect to GCMC calculations for predicting usable H_2 capacities in MOFs. Data is collected under PS **(a, c)** and TPS **(b,d).** Different colors represent different categories of MOFs. Top **(a, b)** and bottom **(c,d)** panels illustrate performance for usable gravimetric and volumetric capacities, respectively.

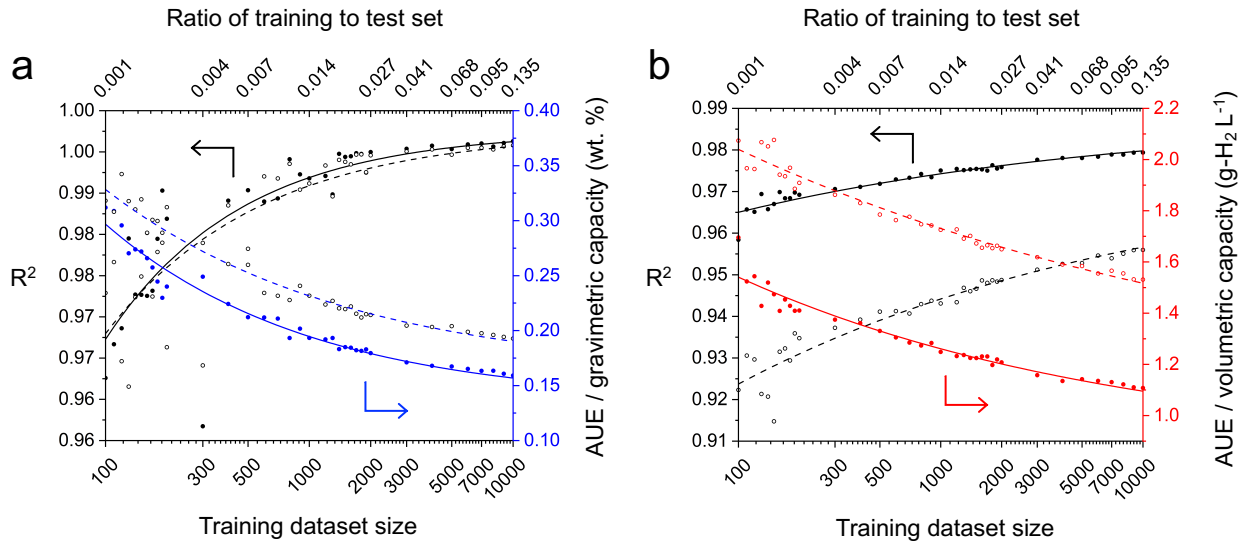


Figure S6. Performance of Extremely Randomized Trees ML models as a function of training set size and the ratio of training to test set size. (a) Usable gravimetric and (b) volumetric H_2 capacity. 100 different training sets ranging in size between 100 and 74,021 MOFs were examined. A common set of 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 ε represents the metric of accuracy (here average unsigned error or AUE). α, β, and γ are the power-law coefficient, exponent, and constant, respectively.

Condition	β (scaling factor)	α (coefficient)	γ (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

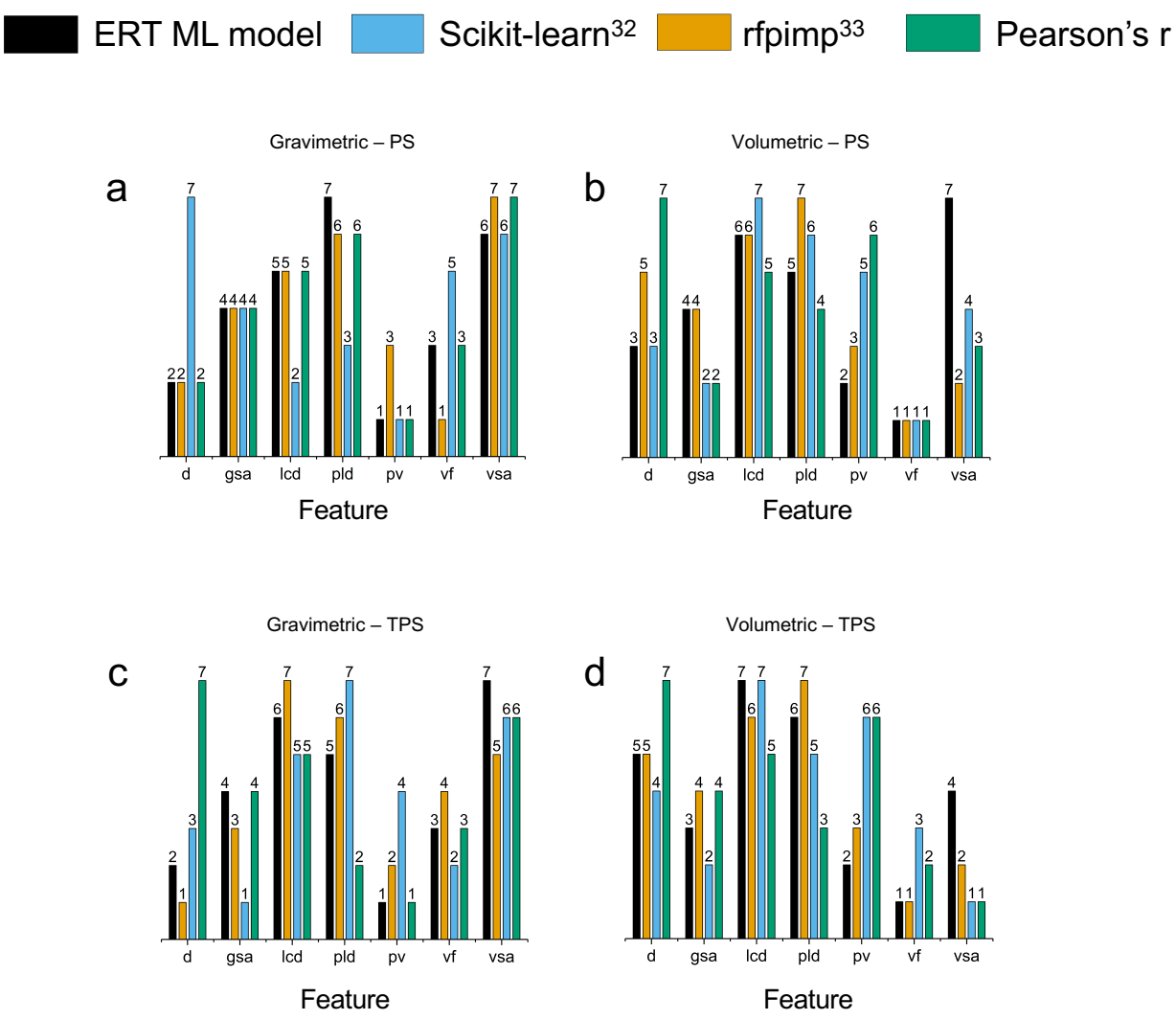


Figure S7. Relative importance of seven features in predicting H_2 storage in MOFs. [32,33] Features are ranked 1(most important) through 7 (least important). Four different methods were used: Pearson's correlation coefficient (r), Breiman and Fried-man's tree-based algorithm as implemented in Scikit-learn,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

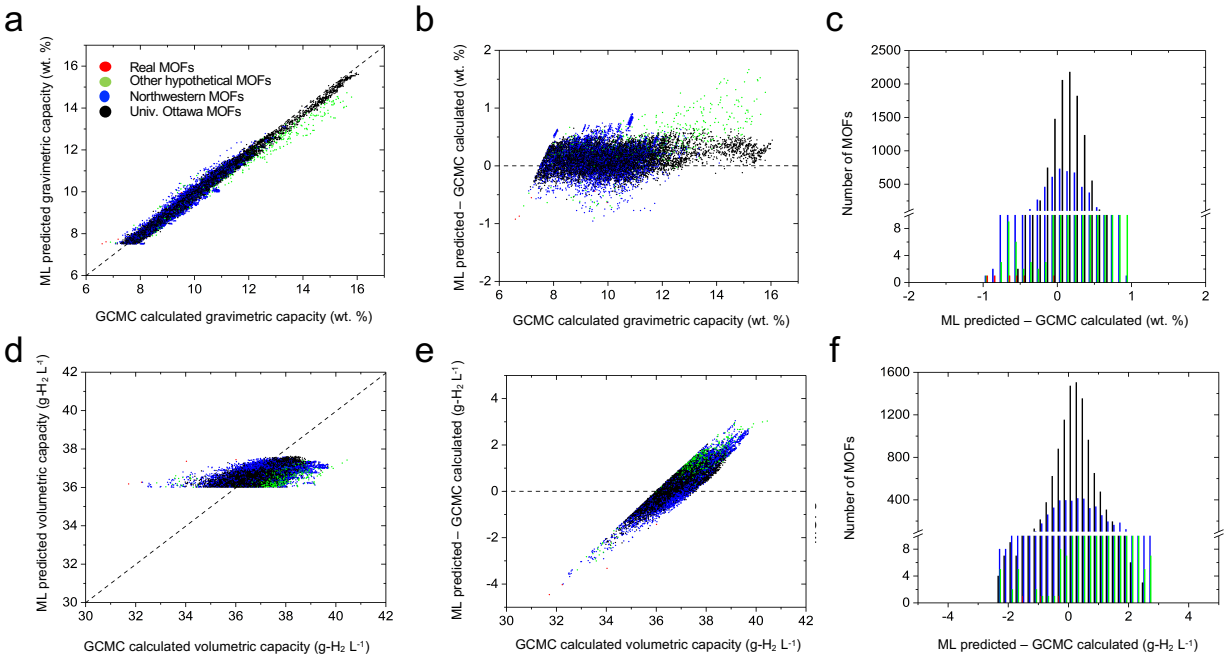


Figure S8. Comparison of GCMC calculations with ML predictions for the 21,700 highest-capacity MOFs predicted by ML for PS conditions. Top (**a**-**c**) and bottom (**d-f)** panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (**a**, **d**) show the 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**	

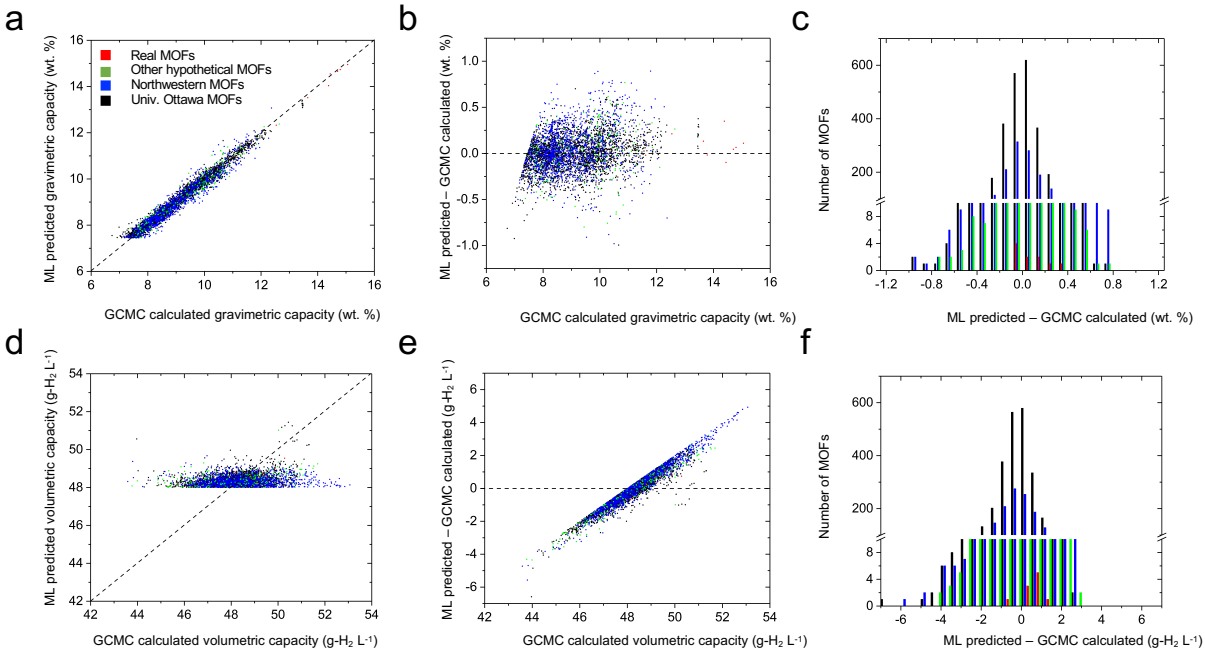


Figure S9. Comparison of GCMC calculations with ML predictions for the 7,901 highest-capacity MOFs predicted by ML for TPS conditions. Top (**a**-**c**) and bottom (**d-f)** panels illustrate the performance for gravimetric and volumetric capacities, respectively. Left panels (**a**, **d**) show the 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 ($g\text{-}H_2\ L^{-1}$)	UG (wt. %)	UV ($g\text{-}H_2\ L^{-1}$)
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.