

Interpretable Machine Learning-Based Predictions of Methane Uptake Isotherms in Metal–Organic Frameworks

Rishi Gurnani, Zhenzi Yu, Chiho Kim, David S. Sholl, and Rampi Ramprasad*



Cite This: *Chem. Mater.* 2021, 33, 3543–3552



Read Online

ACCESS |



Metrics & More

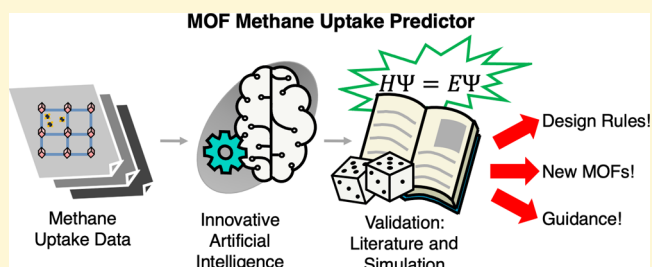


Article Recommendations



Supporting Information

ABSTRACT: Tuning the structure of metal–organic frameworks (MOFs) is a promising pathway toward the development of high-performing materials for methane storage. To aid such discoveries, we introduce techniques for the machine-learned prediction of methane isotherms in MOFs. We demonstrate that our predictors surpass prior benchmarks. We use these models to search for novel (from both a structural and chemical point of view), high-performing MOFs and test them using density functional theory (DFT)-based structural relaxation and molecular simulation of methane adsorption. These simulations reveal that our model generalizes to chemistries not seen during training. One novel candidate, predicted to surpass the 2008 world record for volumetric methane uptake in MOFs, is proposed. Our simulations also reveal that DFT relaxation has a systematic effect on the uptake value. Finally, we interpret the models to discover and present potential MOF–methane uptake structure–property relationships.



1. INTRODUCTION

Natural gas, typically containing 70–90% methane,¹ is a fuel source of interest to the transportation industry. To efficiently store natural gas, a densification process is needed. Porous materials with superior adsorption-based methane storage capability are of interest for this purpose.² Metal–organic frameworks (MOFs) are one such class of porous materials that have attracted attention due to their high surface area,³ stability,⁴ ease of synthesis, chemical tunability,⁵ and efficacy in adsorption and separation of gases.^{6–8}

A MOF consists of a periodic arrangement of constituent building blocks, namely, metal clusters and coordinated organic linkers. Given the different choices one may make for the metal and the organic linkers, the chemical compound space spanned by MOFs is staggeringly large. Searching this space for cases that meet application-specific targets is therefore non-trivial. The scientific community has explored this space using both experimental and physics-based simulation approaches for a variety of end-use applications such as natural gas fuel tanks⁹ and fuel cells.¹⁰ Large data sets from these efforts have been archived by several groups.^{9,11–16}

Surrogate machine learning (ML) models trained on these data sets may be used to accelerate search of the MOF chemical space.¹⁷ These models make predictions orders of magnitude faster than the parent experimental or physics-based simulation approaches used to train the models. Materials design aided by such surrogate models has indeed impacted many domains,^{18,19} including MOF development.^{20–34} A few recent works have led to versatile models for the prediction of gas uptake. For instance, Gharagheizi et al. and Anderson et al. trained models capable of predicting adsorption isotherms of multiple species;^{31,33}

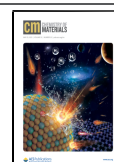
Fanourgakis et al. have demonstrated a capability to predict methane uptake for multiple classes of nanoporous materials,³² while Anderson et al. developed a model for the prediction of hydrogen uptake at any temperature and pressure.²⁵ General challenges remain in any such surrogate model development effort, including modest prediction accuracy,^{26,32,33} inability to generate predictions outside the chemical space of the training set,^{29,31,32} and model interpretability.^{31–33}

In this contribution, we address the above concerns by training and interpreting three ML models which predict methane uptake in MOFs. The development of our methane uptake models involved several critical ingredients captured in Figure 1a: training data, fingerprinting, and deep learning. Our training data was created by curating a large library of 137,953 hypothetical MOFs (hMOFs) developed by Wilmer et al.¹⁶ into a set of chemically valid structures. The Wilmer et al. data set was created using grand canonical Monte Carlo (GCMC) simulations of hMOF structures to determine their gravimetric and volumetric methane isotherms. The hMOF structures themselves were entirely made from the 11 elements shown in Figure 1b. The second vital ingredient of our predictive model building process is a MOF fingerprinting scheme that allowed us to make predictions for MOFs beyond the specific chemistries

Received: December 11, 2020

Revised: April 27, 2021

Published: May 14, 2021



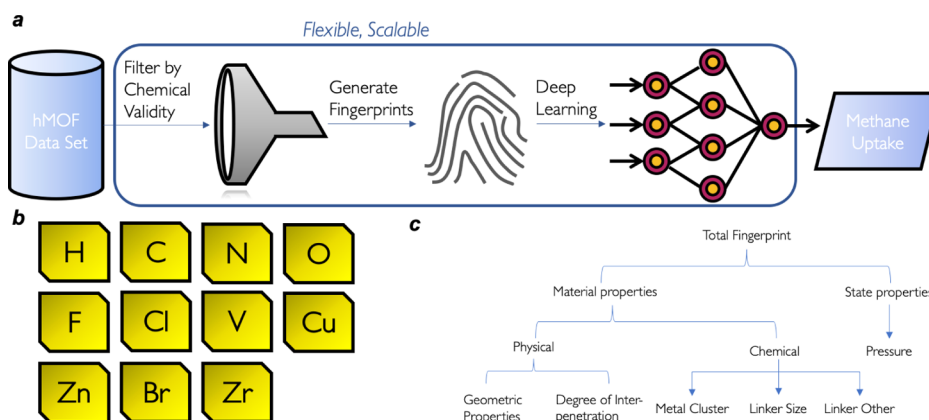


Figure 1. (a) Flowchart describing our computational workflow: start with hMOFs, down-select to remove chemically invalid structures, fingerprint, and map to methane uptake via deep, feed-forward neural networks; (b) symbols of elements which make up hMOF data; (c) hierarchical depiction of our fingerprint which contains geometric information, describes chemistry, captures the degree of interpenetration, and accounts for operating pressure.

(i.e., atom types) contained in the training data set. The fingerprinting process helped convert each MOF to a numerical, machine-readable form. In the present case, the metal and linker parts as well as other higher-order details were encoded in our fingerprints in a comprehensive manner, as schematically shown in Figure 1c. This ensures that, in principle, our approach can be reliably extended to MOFs of arbitrary composition. Finally, the fingerprints were mapped to methane uptake by a feed-forward neural network.

Three such neural networks were trained—one which predicts gravimetric methane uptake at 35 bar and two models which predict the entire volumetric methane uptake isotherm but with varying levels of cost and accuracy. Our predictive models are available for community use and may be accessed from <https://khazana.gatech.edu>. We find that accurate predictions of isotherms in hMOFs can be obtained without knowledge of the unit cell structural details. Moreover, our trained models mined rules for maximizing methane uptake directly from the training data itself. Several of these rules have been reported previously.^{16,28,29,35–37}

We further tested the generalization of one model by predicting the methane uptake of 100,000 novel, hypothetical, MOFs (nMOFs) and compared these predictions against fresh GCMC simulations of methane uptake. The nMOFs were created by substituting the metal clusters of hMOFs with a chemically diverse set of paddle-wheel metal clusters. Given that the original database of hMOFs used as training data in this study (i.e., our “ground truth” data) did not involve optimization of the MOF geometry using density functional theory (DFT)-based simulations (a practice that appears to be common in the MOF community^{16,26,29}), we chose to adopt a similar protocol to create GCMC validation data for our nMOFs. Under this assumption, we found good agreement between our model and GCMC simulation even for “extreme” nMOF chemistries (containing elements in the metal cluster outside of the 11 used to construct hMOFs). This suggests that, in practice, our ML models may generalize reasonably well to arbitrary chemistries.

A lingering question (whose answer is largely outside the scope of this study) that still remains is whether accurate optimization of the MOF geometry is necessary prior to the GCMC simulations of gas uptake. To partially address this point, we performed geometry optimizations using DFT for

selected hMOFs and nMOFs. We found that sizable atomic displacements do occur in some cases, which consequently affected the ultimate gas uptake GCMC simulation results. This indicates that future gas adsorption simulations by the community must take factors associated with geometric optimization of MOF structures into account. Despite the caveats above, our ML protocol may still point in the direction of high-performing MOF structures and chemistries in terms of high predicted methane uptake behavior and in terms of design rules derived from the models themselves. Furthermore, as new uptake data (for methane or other gases)—especially that also involve DFT geometry optimizations—become available, the presented models can be re-trained to match the fidelity of the training set data.

2. METHODS

2.1. Data Set. A major pre-requisite for training an ML model is data. In this work, we use a data set created by Wilmer et al. In this data set, the hMOF structures were generated without relaxation by connecting building blocks from existing MOFs. Obtained GCMC results consist of methane uptake at up to 6 pressures (all computed using GCMC simulations) and up to 6 geometric features (see the Supporting Information, Section S1, for a list of geometric features) computed for each of 137,953 hMOFs. We down-selected from this parent set using the following criteria: (1) select only MOFs for which each geometric feature was computed by Wilmer et al. and (2) select only MOFs for which each of its constituent atoms has valid valence (as determined by our Linker SMILES Extractor (LSE), discussed in Section 2.3). These criteria yielded a set of 126,384 MOFs with associated methane loading and complete geometric data. Using this down-selected data set, henceforth referred to as the “curated hMOF” data set, we created three distinct classes of ML models (as discussed in Section 2.4).

2.2. Fingerprinting. The maps learned by ML models require an input space that is machine-readable. Therefore, each [MOF, pressure] pair in the curated hMOF set was converted into a numerical vector, referred to as a “fingerprint”. The components of the fingerprint used in this work are presented in Figure 1c and fall under three categories: state properties, physical properties, and chemical properties. One component of our fingerprint was pressure, a state property. The six computed geometric features present in the curated hMOF data set, \mathbf{fp}_{geo} , and the degree of interpenetration (converted to a one-hot vector, \mathbf{fp}_{ip}) make up the relevant physical properties. We represented all chemical information pertaining to a MOF by $\mathbf{fp}_{\text{chem}}$, which is a concatenation of two numerical vectors: $\mathbf{fp}_{\text{metal}}$ and $\mathbf{fp}_{\text{linker}}$. To encode features into $\mathbf{fp}_{\text{metal}}$, we used the following set of atomic properties to

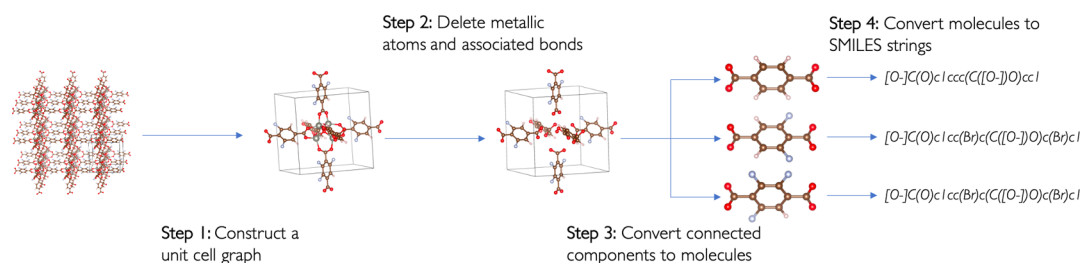


Figure 2. Workflow detailing LSE. First, the unit cell described in the input MOF's structure file is converted into a connected graph. Second, the graph is disconnected by deleting metallic atoms and their bonds. Third, the connected components of the graph are converted into molecules. Finally, each molecule is mapped to its corresponding SMILES string.

represent the metal cluster: ionization potential, Pauling electronegativity, computed atomic radius, coordination number, and electron affinity. Each component of $\mathbf{fp}_{\text{metal}}$ is defined as

$$\mathbf{fp}_{\text{metal},i} = \frac{\sum_{j=1}^{N_E} N_j p_{ij}}{\sum_{j=1}^{N_E} N_j} \quad (1)$$

where p_{ij} is the i^{th} property corresponding to atom type j , $\mathbf{fp}_{\text{metal},i}$ is the i^{th} dimension of $\mathbf{fp}_{\text{metal}}$, N_E is the number of atom types in the cluster, and N_j is the number of atoms of type j present in the cluster. The fingerprinting scheme, used earlier to represent organic polymers with our polymer genome (PG) framework (as described elsewhere³⁸), was used to encode the organic linker features in $\mathbf{fp}_{\text{linker}}$. Since a given MOF can have more than one type of organic linker, an average of the linker fingerprint was used, defined as

$$\mathbf{fp}_{\text{linker}} = \frac{\sum_{i=1}^{N_l} \mathbf{fp}_{\text{linker},i}}{N_l} \quad (2)$$

where $\mathbf{fp}_{\text{linker},i}$ is the PG fingerprint of the i^{th} linker and N_l is the number of connected components extracted from the MOF structure file. Given a valid SMILES string, PG generates a hierarchical fingerprint for the corresponding material spanning the atomic scale to the molecular scale. As a final fingerprint component, we considered the number of atoms present in each of a MOF's constituent linkers, $\mathbf{fp}_{\text{size}}$. The elements of $\mathbf{fp}_{\text{size}}$ are sorted in descending order for each MOF. Our total fingerprint (\mathbf{fp}_{geo} , \mathbf{fp}_{ip} , $\mathbf{fp}_{\text{chem}}$, $\mathbf{fp}_{\text{size}}$, p) contains 440 dimensions.

2.3. Linker SMILES Extraction. Due to the inclusion of $\mathbf{fp}_{\text{linker}}$, our fingerprinting strategy is extendable to those reticular structures containing linkers that have a suitable SMILES string representation. In this study, manual generation of linker SMILES strings was infeasible as the number of unique linkers present in the data is $\sim 10^5$. Thus, we developed the LSE, a scalable program for extracting the SMILES string of each linker present in a given MOF. LSE was implemented using two open source packages, Pymatgen³⁹ and OpenBabel.⁴⁰

As shown in Figure 2 for a representative example, LSE works in four steps. In the first step, a MOF's unit cell is converted into a connected graph (with atoms as vertices and chemical bonds as edges). In the second step, vertices corresponding to metallic atoms are removed from the graph, yielding a disconnected graph with a set of connected components. Each component (containing a large enough number of vertices) is considered as the graph representation of one individual linker in the MOF. In the third step, each connected component is converted into a molecule file. Finally, in the last step, each molecule file is converted to a SMILES string. If any molecule (i.e., linker) is chemically invalid, the MOF is discarded (see the Supporting Information, Section S2, for more details). LSE (as written at the time of publication) is compatible with .cif structure files and strictly requires both at least one metallic atom in the cluster unit and only organic atoms in the linkers.

2.4. Machine Learning. We used the approach illustrated in Figure 1 to train three distinct ML models for prediction of methane loading at 298 K. One model, denoted as the gravimetric uptake Model (GUM), makes predictions for gravimetric methane uptake at 35 bar. A second model predicts volumetric methane uptake at any pressure using all

components of our fingerprint, namely, geometric (\mathbf{fp}_{geo}), chemical ($\mathbf{fp}_{\text{chem}}$), degree of interpenetration (\mathbf{fp}_{ip}), and pressure (p). Similarly, the third model predicts volumetric methane uptake at any pressure but does not require \mathbf{fp}_{geo} . These models are referred to below as the geometric model (GM) and the non-GM (NGM), respectively, because the former requires geometric information, while the latter does not. Each model contains a set of flexible parameters that were trained using neural networks. The three models can therefore be mathematically defined as follows:

$$u_g = \text{GUM}(\mathbf{fp}_{\text{geo}}, \mathbf{fp}_{\text{ip}}, \mathbf{fp}_{\text{chem}}; \theta_1) \quad (3)$$

$$u_v = \text{GM}(\mathbf{fp}_{\text{geo}}, \mathbf{fp}_{\text{ip}}, \mathbf{fp}_{\text{chem}}, p; \theta_2) \quad (4)$$

$$u_v = \text{NGM}(\mathbf{fp}_{\text{ip}}, \mathbf{fp}_{\text{chem}}, p; \theta_3) \quad (5)$$

where u_g is the predicted gravimetric uptake, u_v is the predicted volumetric methane uptake, θ_i are the flexible parameters, and each \mathbf{fp}_x is a distinct component of our MOF fingerprint. To train GUM, we randomly split the hMOF data five times into train, validation, and test sets. 30% of the data were allocated to the test set with the remaining data to be split among the training and validation sets (the fraction of these data to be contained in the validation set, s , is a hyperparameter). These five splits are used to train five models, GUM- i (i is an integer from 1 to 5). We adopted such a training strategy to facilitate comparison of our results with Wu et al.²⁹ To train GMs and NGMs, we adopted the more conventional method of k -fold cross-validation (CV). We chose k to be 10 based on the findings of Breiman and Spector.⁴¹ Before performing 10-fold CV, we performed hyperparameter optimization, the details of which can be found in the Supporting Information, Section S4.

2.5. Model Interpretation. In this work, we use Shapley Additive Explanations (SHAP),⁴² a game-theoretic approach to model interpretation, to compare and interpret the predictions of our models. SHAP treats the features input to an ML model as players and the model itself as a game in which the reward is maximized by maximizing the target property—volumetric methane uptake in our case. The raw outputs of SHAP are importance values to a given model, known as “Shapley values”, of each feature of each data point. Computing the absolute Shapley values, averaged over each data point, yields the mean importance of each feature to the model. SHAP is further useful because it can approximate Shapley values for any model, deep learning or otherwise. Thus, the importance of features of all models which precede or succeed ours can be directly compared with the results presented in this contribution.

2.6. Methane Adsorption Simulations. The methane uptake of novel, hypothetical, MOFs (nMOFs) containing elements not included in the curated hMOF set were predicted using NGM. To test these predictions *in silico*, GCMC simulations of single-component adsorption were performed on the nMOFs using the same method as in the original work of Wilmer et al. on hMOFs. van der Waals interactions for the MOF framework were described by the Universal Force Field,⁴³ while the van der Waals interaction for adsorbates was described by the TrAPPE force field.⁴⁴ Lorentz–Berthelot mixing rules were used to define adsorbate–MOF interactions. Lennard–Jones

interactions were truncated at 12.0 Å with tail corrections. Simulation volumes were expanded to at least 24.0 Å along each dimension, and triclinic periodic boundary conditions were imposed. 10^5 Monte Carlo cycles were used to ensure satisfactory convergence of results. All simulations were carried out using RASPA.⁴⁵

2.7. Density Functional Theory. The effect of relaxation on MOF methane uptake is not well understood. As a step in this direction, we compared the GCMC-simulated methane uptake of both DFT-relaxed and unrelaxed MOFs. Comparisons of DFT-optimized MOF structures with experimentally derived test sets have shown that DFT makes accurate predictions for these materials.⁴⁶ Based on our DFT results, we can glean a better understanding of the effect of relaxation and, together with our ML results, we can give meaningful guidance for experimental syntheses. MOF structure relaxations were calculated using spatially periodic DFT in the Vienna ab initio simulation package,⁴⁷ along with a plane-wave basis set and projected-augmented wave⁴⁸ pseudopotentials. All calculations used the Perdew, Burke, and Ernzerhof⁴⁹ generalized gradient approximation (GGA) exchange–correlation functional with D3 dispersion corrections (PBE-D3).⁵⁰ We simultaneously optimized both the lattice parameters and atomic positions using a plane-wave cutoff energy of 600 eV and Γ -point sampling for Brillouin zone integration. Using a quasi-Newton method, we relaxed geometries until the force on each atom was smaller than 0.05 eV/Å. DFT calculations based on the local density approximation and GGA often fail to describe energetics and geometries around transition metals, which can be corrected by adding a Hubbard-like term.⁵¹ DFT + U was used to describe the strong on-site Coulomb interaction of localized electrons. The strength of the on-site interactions was described by on-site Coulomb parameters U and on-site exchange parameter J . For open-shell 3d or 4d transition metals, we used Hubbard U corrections for the localized d electrons with U values for Cr, Mo, Mn, Ru, and Ni of 3.5, 4.38, 4.0, 4.5, and 6.4 eV, respectively. J values were set to 0 for all elements.^{52,53} Based on previous research, local antiferromagnetic ordering was imposed for each metal dimer and a high-spin electronic state was used for each metal cluster.⁵⁴

3. RESULTS AND DISCUSSION

The results and discussion are organized as follows: first, we benchmark, in order, the performance of GUM, GM, and NGM with respect to prior works. Next, both GM and NGM are interpreted. Finally, we test NGM's predictions via simulation.

3.1. Modeling Gravimetric Methane Uptake. Previous work has been done on the prediction of gravimetric methane uptake in hMOFs. Of these works, we compared GUM against the model proposed by Wu et al.²⁹ as it exhibited the lowest published root-mean-squared error (RMSE). Wu et al. show that including Henry coefficients in the MOF fingerprint greatly improves prediction of methane uptake in hMOFs. However, this improvement comes at a cost. Henry coefficient computations are themselves both manually and computationally expensive and therefore do not significantly reduce the amount of time involved in predicting methane uptake.

The performance metrics, with respect to test (i.e., unseen) data, of our models compared with Wu et al. are shown in Table 1. The performance metrics considered are test set RMSE and test set coefficient of determination (R^2). In addition, the parity plot of the best case model's predictions on its test set of 37,911

Table 1. Performance Metrics of Gravimetric ML Models

model	best		average ^a	
	RMSE (cm ³ /g)	R^2	RMSE (cm ³ /g)	R^2
GUM	9.05	0.992	9.26	0.992
Wu	9.40	0.991	N/A	N/A

^aStatistics shown are computed over five models.

unseen MOFs is shown in Figure 3. Perfect predictions would exhibit an R^2 of 1 and an RMSE of 0.

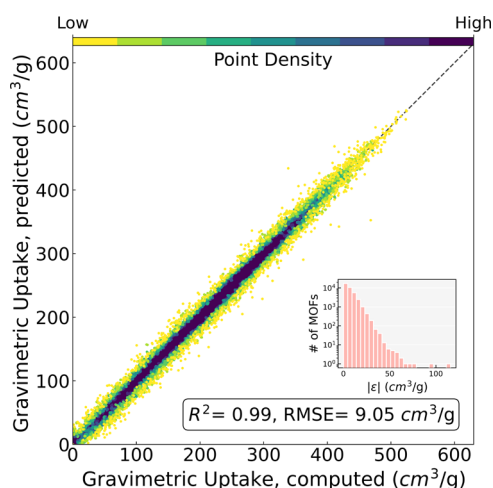


Figure 3. Computed gravimetric methane uptake vs predicted gravimetric methane uptake for 37,911 unseen MOFs predicted by the best case model. Perfect predictions fall on top of the black, dashed parity line. The density of points is shown in color to convey the ratio of predictions which fall near to and far from the parity line. Inset: this information is also presented as a histogram with absolute error ($|\Delta|$) on the x-axis and the corresponding number of MOFs (in the log scale) on the y-axis. R^2 and RMSE values for the unseen MOFs are shown in the bottom right.

Looking at Table 1, both our average and best models outperform Wu et al. We emphasize that this improvement comes with the additional benefit of circumventing the Henry constant computations that were used by Wu et al. We attribute the elevated performance and efficiency of our model to two major factors. The first is our inclusion of rich linker-related features which are easily computed from the SMILES strings of linkers. This underscores the importance of LSE, which offers rapid access to SMILES strings of MOF linkers at a scale not feasible manually. The second factor is that LSE was used to discard chemically invalid hMOFs (see the Supporting Information, Section S2). The invalidity of these MOFs indicates that their corresponding loading data contain irrelevant information and spurious correlations which, if learned by the model, undermine downstream predictive tasks.

3.2. Modeling Volumetric Methane Uptake Isotherms.

We now turn to a different learning problem—predicting volumetric methane isotherms. To address this problem, we re-used our computational pipeline to build two models (GM and NGM). Below, we compare GM and NGM against the models proposed by Fanourgakis et al.³² as their work exhibited the lowest published RMSE to date. Notably, as in our GM, Fanourgakis et al. also use the geometric information of each hMOF as part of their fingerprint.

3.2.1. Geometric Models. We performed 10-fold CV to train 10 GMs, which we refer to as GM- j , where j indicates the fold number. The test set predictions made by each GM- j are plotted together in Figure 4a. The performance metrics of the best GM (GM-4), the average of all GMs, and the average performance of the models produced by Fanourgakis et al. are all shown in Table 2. The performance metrics considered are test set RMSE, test set R^2 , and test set average maximum 10 percent error (AM10PE). Perfect predictions would exhibit an AM10PE of 0.

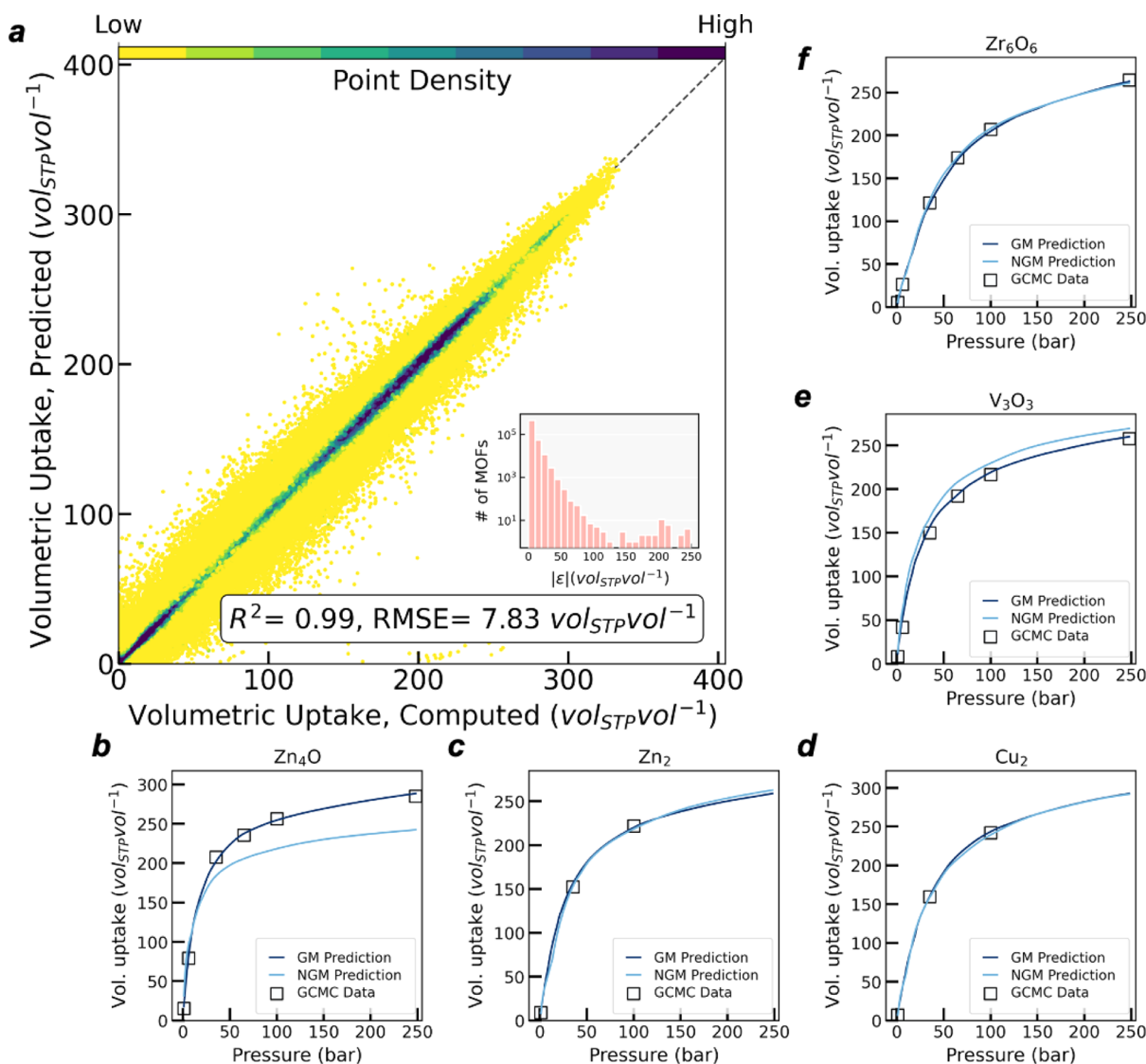


Figure 4. (a) Computed volumetric methane uptake vs predicted volumetric methane uptake for each of 533,430 measurements in the curated hMOF data set. Perfect predictions fall on top of the black, dashed parity line. The density of points is shown in color to convey the ratio of predictions which fall near to and far from the parity line. Inset: this information is also presented as a histogram with absolute error ($|\epsilon|$) on the x-axis and the corresponding number of MOFs (in the log scale) on the y-axis. R^2 and RMSE are shown in the bottom right. (b–f) NGM-5 and GM-5 predicted isotherms (light-blue and dark-blue curves, respectively) vs GCMC data (white squares) for five randomly selected MOFs from the test set shared by both models. The metal cluster contained in each MOF is noted in the title of its corresponding plot. Each structure is noted in the [Supporting Information](#), Section S5.

Table 2. Performance Metrics of Volumetric ML Models

model	weighted average ^a			best ^b		
	RMSE ^c	R^2	AM10PE	RMSE ^c	R^2	AM10PE
Fanourgakis	8.90	0.908	N/A	N/A	N/A	N/A
GM	7.83	0.992	18.74	7.01	0.994	16.70
NGM	18.64	0.955	46.26	17.22	0.962	42.75

^aStatistics shown are computed over 10 models. ^b“Best” models are GM-4 and NGM-5. ^cGiven in units of $\text{vol}_{\text{STP}}\text{vol}^{-1}$.

As illustrated in [Table 2](#) and [Figure 4a](#), with the exception of a few cases, GMs are able to predict volumetric methane uptakes with high accuracy. GMs surpass Fanourgakis et al. in RMSE by 22% in the best case. We emphasize that this improvement in RMSE came despite our inclusion of high-pressure volumetric methane uptake data points not modeled by Fanourgakis et al. Inclusion of these data points skews unbounded metrics, such as RMSE, that grow as the scale of the target variable (volumetric

methane uptake in this case) grows. Furthermore, GM (and NGM) is able to predict methane uptake at any pressure, while the models proposed by Fanourgakis et al. are restricted to predicting uptake at four discrete pressures.

3.2.2. Non-Geometric Models. 10-fold CV was also performed to train 10 NGMs which we refer to as NGM-*j*. The test sets used to train NGM-*j* matched those used to train GM-*j* (e.g., GM-1 and NGM-1 are both trained on one subset,

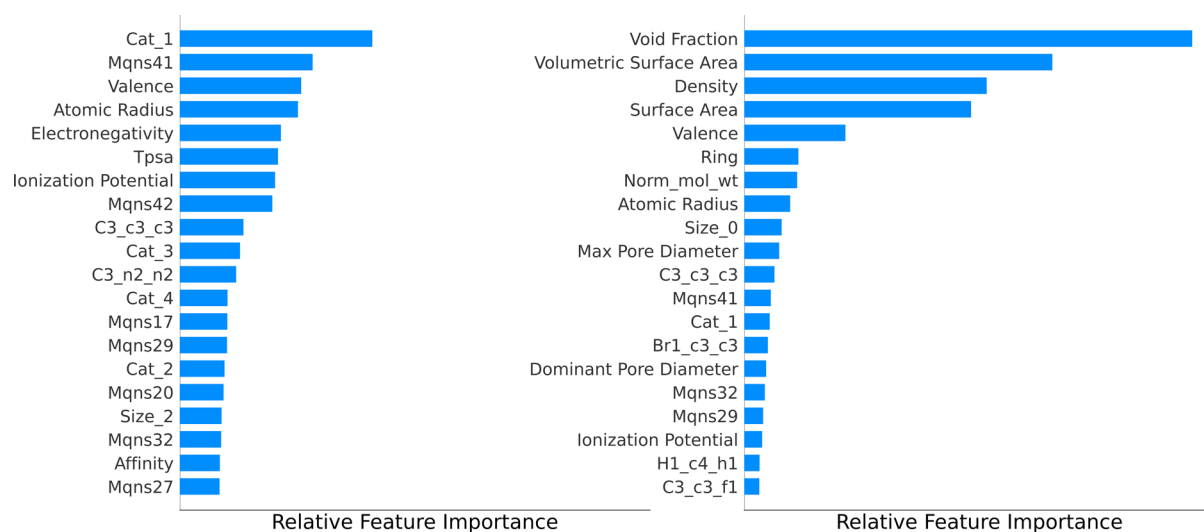


Figure 5. Feature importance, as ranked by SHAP, for (a) NGM-4 and (b) GM-5. Each panel plots the 20 most important features of its respective model. The feature's name is shown on the y-axis, and its corresponding importance is shown on the x-axis. See the [Supporting Information](#), Section S6, for a description of the features.

GM-2 and NGM-2 are both trained on another subset, and so forth).

The average performance metrics of the NGMs are shown in [Table 2](#), along with the performance metrics of the most accurate NGM, NGM-5. Interestingly, the average NGM outperforms Fanourgakis et al. in terms of R^2 even though the NGM does not require the geometric information required by both Fanourgakis et al. and GMs. Although NGMs are an improvement over Fanourgakis et al., they are less accurate than GMs. This decrease in performance is expected, considering the fact that NGMs do not use the rich structural information contained in fp_{geo} . This reduces accuracy in some cases. For example, NGMs cannot distinguish between structural polymorphs, while GMs can (so long as each polymorph exhibits a distinct fp_{geo}). However, this advantage of GMs comes at a cost. First, GMs require the computation of fp_{geo} . Second, even before these computations, a steeper cost arises—the construction of the MOF's unit cell (which, if relaxed, may take a few days using DFT or several hours using molecular dynamics). NGMs circumvent these constraints while producing predictions with a reasonable level of accuracy.

As an additional comparison between GMs and NGMs, five MOFs—each containing a different metal cluster—were randomly selected from the fifth test set (i.e., the test set used to train both GM-5 and NGM-5). For each MOF (see the [Supporting Information](#), Section S5, for the corresponding structures), the isotherms predicted by NGM-5 and GM-5 are plotted in [Figure 4b–f](#) against its volumetric methane data from the GCMC simulations of Wilmer et al. These cases corroborate the fact that while GMs provide superior predictive accuracy, NGMs perform competitively at a fraction of the cost.

3.3. Model Interpretations. We employed neural networks due to a favorable balance between scaling to the large number of structures in the curated hMOF data set and predictive accuracy. However, the hundreds of thousands of trainable parameters present in our neural networks make such models difficult to interpret directly. Model interpretability techniques attempt to address this issue by determining what feature(s) a given ML model weighs most heavily when making predictions. We employed SHAP in this work to explore this issue.

Out of more than 400 features, [Figure 5](#) ranks the 20 most important features (excluding pressure) to NGM-5 and GM-4—our best NGM and GM, respectively—as deduced by SHAP. Some observations stand out when comparing the two rankings. First, void fraction, volumetric surface area, and density are the most important features to GM-4. It is not surprising that these features are important; similar conclusions have emerged from multiple high-throughput studies of adsorption of MOFs and other materials.^{28,29,55} Second, both maximum and dominant pore diameters are highly important features to GM-4. Similarly, previous studies report optimum pore diameters of 4 and 8 Å (the diameter required to fit one and two methane molecules, respectively) for methane uptake.^{16,35,36} Third, the feature importance distribution of NGM-5 is much broader than that of GM-4. This is consistent with the previous two observations. NGM-5 does not have the benefit of powerful geometric features and is thus forced to squeeze out information from a larger number of comparatively weak features to make an accurate prediction. Fourth, Cat-1, a binary feature related to whether or not a MOF is interpenetrated, is the 13th most important feature in GM-4 but is the most important feature in NGM-5. This suggests that much of the information present in the geometric features is also contained in Cat-1. Intuitively, this makes sense; an interpenetrated MOF tends to, for example, exhibit a higher density than its non-interpenetrated counterpart. Fifth, C3_c3_c3, which indicates the occurrence frequency of three concatenated 3-fold carbon atoms in the linker, is a consistently important feature. Indeed, previous work has shown carbon atomic number density to be an important factor for methane uptake in MOFs.²⁹ Sixth, linker sizes (i.e., Size_i, the i^{th} element of fp_{size} in [Figure 5](#)) are consistently ranked as important features to both GM and NGM. Similarly, previous work found that most MOFs with short linkers avoid network interpenetration,³⁷ a property which generally increases a MOF's uptake.

Given SHAP's ability to identify several established and intuitive structure–property relationships (SPRs) from our ML models, it would not be surprising if a few or many of the remaining important features shown in [Figure 5](#) constitute yet-to-be discovered SPRs. For example, the presence of valence as a consistently important feature indicates that metal clusters of a

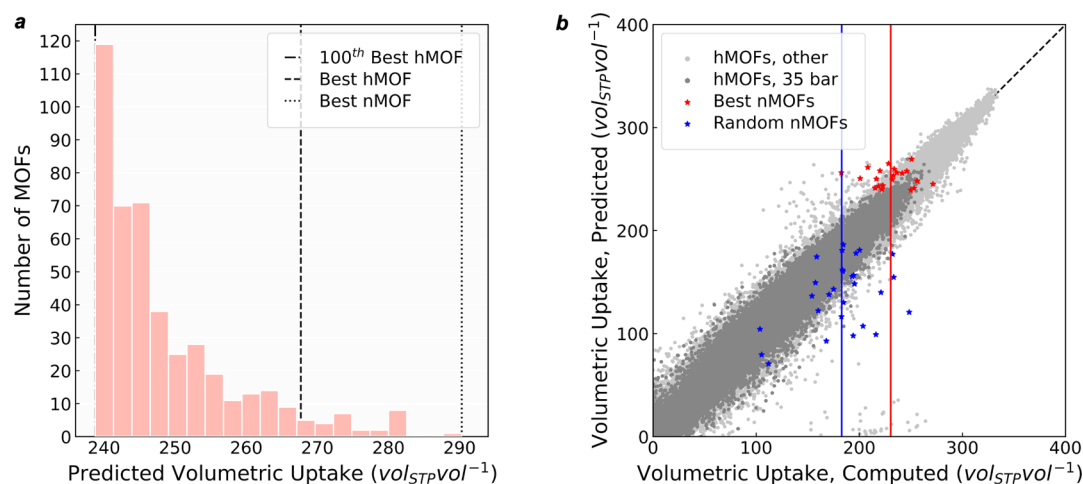


Figure 6. (a) 429 nMOFs predicted to be high-performing. GCMC-simulated volumetric methane uptake at 35 bar of both the 100th best and best paddle-wheel hMOFs shown as vertical lines along with predicted volumetric methane uptake at 35 bar of the best nMOF. (b) NGM predictions of the best nMOFs (red stars), random nMOFs (blue stars), and GM predictions of hMOFs (uptake at 35 bar shown in dark gray; uptake at other pressures shown in light gray). Average GCMC-simulated methane uptake of the best and random nMOFs shown by red and blue lines, respectively.

specific coordination number may tend to maximize methane uptake. In addition, both Br1_c3_c3 and C3_c3_f1 (aX_bY_cZ: a, b, and c are contiguous heavy atoms, X is the number of atoms bonded to a, Y is the number of atoms bonded to b, and Z is the number of atoms bonded to c) are important features to GM. This suggests that incorporation of electronegative functionalities (e.g., Br–R and F–R) in linkers can significantly impact methane uptake. We present these possible relationships with hopes that they may be investigated and, if corroborated, used to design high-performing materials.

3.4. Using NGM to Design Paddle-Wheel MOFs. The primary advantage of NGMs over GMs is the ability to make accurate predictions without costly computations. This provides a pathway to use NGMs for the ultrafast screening of MOF candidates. We illustrate this capability by maximizing the volumetric methane uptake at 35 bar in paddle-wheel MOFs (i.e., MOFs for which the metal cluster is composed of two metal atoms and coordinated to four oxygen atoms plus two nitrogen atoms).

Using paddle-wheel hMOFs and a set of paddle-wheel metals⁵⁶ (a complete list in the [Supporting Information](#), Section S7), we generated fingerprints for 100,000 novel candidate MOFs, which we call nMOFs. We then used NGM-5 to predict volumetric methane uptake at 35 bar for each nMOF. It is worth emphasizing that with the NGM approach, a full structure prediction for each nMOF is not required, so these predictions could be made very rapidly. We define a “high-performing” MOF as one whose volumetric methane uptake at 35 bar is larger than 238 vol_{STP} vol⁻¹. As shown in [Figure 6a](#), 429 nMOFs (0.429% of the total) were predicted to be high-performing.

We conducted further studies on 21 of the highest-performing (with respect to NGM-5-predicted uptake) nMOFs and 30 randomly chosen nMOFs, making sure to include several nMOFs with synthesis-friendly metal clusters, namely, Mg or Al. For each of these nMOFs, we performed GCMC simulations of methane uptake at 35 bar on an unrelaxed atomic configuration (see the [Supporting Information](#), Section S7, for details on generation of unrelaxed structures). This is appropriate because our present model was trained from the data of Wilmer et al. which include only MOF structures that were not subjected to

DFT geometry optimizations. As shown in [Figure 6b](#), we found good agreement between NGM and GCMC results.

A few observations from [Figure 6b](#) deserve mention. First, it is clear that those nMOFs which NGM predict to be high-performing are, in general, clustered tightly at appreciably high uptake values compared to random nMOFs which exhibit a wider distribution centered around a comparatively low uptake value. Second, the best nMOFs tend to have large uptake values compared to the hMOFs. Third, not only do the nMOFs compare favorably with hMOFs but also they compare favorably with synthesized MOFs. In fact, 13 nMOFs display a GCMC-simulated uptake that would surpass the 2008 world-record synthetic MOF^{16,57} (to our knowledge, this record has not since been surpassed). We also find that prediction errors in nMOFs are comparable to those of the hMOFs with the exception of a few aberrant cases (see the [Supporting Information](#), Section S8). For these cases, when predictions were made using the GM, the error in the predictions relative to the GCMC simulation results decreased significantly (see the [Supporting Information](#), Section S9). In other words, for these aberrant cases, the geometric information is particularly important for accurate methane uptake prediction.

We noted that the predictions in [Figure 6](#) were made on unrelaxed MOF structures to be consistent with the underlying Wilmer et al. data used to train all our models. To understand the impact of geometry optimization on methane uptake in general, we performed DFT relaxation on the unrelaxed atomic configuration of each nMOF and hMOF and simulated methane uptake on these new structures. These GCMC calculations showed poor agreement with the simulations on unrelaxed structures (see the [Supporting Information](#), Tables S1 and S2) and therefore also showed poor agreement with NGM. Of the 13 unrelaxed nMOFs with GCMC-simulated methane uptake surpassing the world record, one structure retains world-record methane loading after DFT relaxation. This nMOF is made of a Mg metal cluster and does not contain interpenetrating frameworks. The SMILES strings of its two unique linkers are C1=CC=C(C2=C1C3=C(C=C2)C=CC(=C3)C([O])=O)C([O])=O and N#CC#CC#N. An image of the DFT-relaxed unit cell is provided in the [Supporting Information](#), Section S10, and

the associated .cif file is included as [Supporting Information Structure File 1](#).

The disagreement in simulated uptake between the unrelaxed and DFT-relaxed nMOFs is an important point. This difference is likely due to the fact that DFT relaxation results in large atomic displacements relative to the initial, unrelaxed configuration. However, this result is not unique to the nMOFs. The hMOFs used as templates for the nMOFs themselves also experience large displacements during DFT relaxation (see the [Supporting Information](#), Section S11). This mandates that future efforts, aimed at gas uptake data generation via GCMC-based simulations, utilize relaxed equilibrium structures for MOFs. While this may lead to a significant escalation of the computational cost expended during data generation, it must be noted that this added effort will be worth the investment as downstream ML models will be more robust. Moreover, the training data generation constitutes a one-time cost, while the downstream ML model prediction cost will remain as low as that of the present models.

We end this section by emphasizing that the time required to predict the target value of all 100,000 nMOFs using NGM-5 was 3.4 s on one 32 GB Tesla V100-PCIE GPU—many orders of magnitude faster than the time required to perform 100,000 GCMC methane uptake simulations.

4. OUTLOOK

In this work, we applied a computational pipeline involving sophisticated linker SMILES extraction, MOF fingerprinting, and deep learning protocols to develop ML models of methane uptake (both single-pressure and room-temperature isotherms) by MOFs. Design rules, some corroborated by decades of research and others novel, were deduced using ML models. These models surpass previous efforts to predict methane uptake in hMOFs with respect to accuracy. One model, NGM, does not require geometric information of the MOF unit cell and therefore accelerates design. Furthermore, these models are shown to generalize to MOF chemistries beyond those used to train the model. This widens the scope of MOFs which can be reliably studied with ML. The speed with which the ML models can predict methane uptake allows identification of attractive candidates (e.g., those with high methane uptake) by rapidly screening large lists of candidate MOFs. This was demonstrated here, and the identified candidates were validated through explicit GCMC simulations.

It should be noted though that our models can predict methane uptake only to the same level of accuracy as that of the training data, which in our case involved GCMC methane uptake simulations of hMOF structures whose geometry was not previously optimized. One finding of this work is that optimization of the MOF geometry prior to GCMC simulations may be required to achieve high enough fidelity of both the GCMC simulations and, consequently, the downstream ML models. When such higher-fidelity data become available, the computational pipeline described in this contribution can be applied to re-train, perhaps through transfer learning, more robust ML models.

This contribution has implications for future studies of reticular structures. To start with, our fingerprinting strategy is directly applicable to covalent organic frameworks (COFs). Further, LSE can be directly used by others to obtain the SMILES strings of MOF linkers, which—as have been demonstrated by us—improve accuracy of ML models when meaningfully parsed. Adaptation of LSE to work with COFs

requires further study, while the approach is not extendable to, for example, zeolites.

We have focused here on predicting methane adsorption isotherms in MOFs, and our work aimed at identifying new materials considered only the adsorption loading at a single pressure (35 bar). In detailed development of materials for end-use applications, it is important to understand not simply the adsorption loading but the swing capacity that is achievable, the potential impact of adsorbing impurities, the material's stability, etc. Making strong connections between material properties and optimization of process design and operation for gas storage and separation remains challenging.⁵⁸ The ML-enabled approaches we have illustrated here can contribute to the development of materials in these settings by providing efficient ways to select promising materials from among enormous numbers of options. The chosen materials identified in this work serve as starting points for closer examination with more detailed models and experiments.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.chemmater.0c04729>.

List of geometric figures, structure rejection by LSE, CV technique, ML parameters and their optimization, structures used to compare GM and NGM, feature descriptions, generation of nMOFs, comparing NGM predictions of nMOFs and hMOFs, correcting nMOF predictions by NGM with GM, highest-performing nMOF, DFT relaxation of hMOFs and nMOFs, and nMOFs which did not reach DFT equilibrium ([PDF](#))

■ AUTHOR INFORMATION

Corresponding Author

Rampi Ramprasad — School of Materials Science and Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States; orcid.org/0000-0003-4630-1565; Email: rampi.ramprasad@mse.gatech.edu

Authors

Rishi Gurnani — School of Materials Science and Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States; orcid.org/0000-0002-2744-2234

Zhenzi Yu — School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States

Chiho Kim — School of Materials Science and Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States; orcid.org/0000-0002-1814-4980

David S. Sholl — School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, 30332 Atlanta, Georgia, United States; orcid.org/0000-0002-2771-9168

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.chemmater.0c04729>

Notes

The authors declare no competing financial interest.

NGM and LSE are available for public use. LSE is available at <https://github.com/Ramprasad-Group/LSE>, while NGM can be found at <https://khazana.gatech.edu>.

ACKNOWLEDGMENTS

R.G. acknowledges Dr. Rohit Batra for insightful discussions. The hMOF data was obtained by Dr. Rohit Batra from Dr. Randall Snurr. This work was supported as part of the Center for Understanding and Control of Acid Gas-Induced Evolution of Materials for Energy (UNCAGE-ME), an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Basic Energy Sciences, under award no. DE-SC0012577.

REFERENCES

- (1) Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M. In silico Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *J. Am. Chem. Soc.* **2014**, *136*, 5006–5022.
- (2) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating metal-organic frameworks for natural gas storage. *Chem. Sci.* **2014**, *5*, 32–51.
- (3) Farha, O. K.; Özgür Yazaydın, A.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat. Chem.* **2010**, *2*, 944–948.
- (4) Slater, A. G.; Cooper, A. I. Function-led design of new porous materials. *Science* **2015**, *348*, aaa8075.
- (5) Islamoglu, T.; Goswami, S.; Li, Z.; Howarth, A. J.; Farha, O. K.; Hupp, J. T. Postsynthetic Tuning of Metal-Organic Frameworks for Targeted Applications. *Acc. Chem. Res.* **2017**, *50*, 805–813.
- (6) Herm, Z. R.; Wiers, B. M.; Mason, J. A.; van Baten, J. M.; Hudson, M. R.; Zajdel, P.; Brown, C. M.; Masciocchi, N.; Krishna, R.; Long, J. R. Separation of Hexane Isomers in a. *Science* **2013**, *340*, 960–964.
- (7) Bloch, E. D.; Queen, W. L.; Krishna, R.; Zadrozny, J. M.; Brown, C. M.; Long, J. R. Hydrocarbon Separations in a Metal-Organic Framework with Open Iron(II) Coordination Sites. *Science* **2012**, *136*, 5271.
- (8) McDonald, T. M.; et al. Cooperative insertion of CO₂ in diamine-appended metal-organic frameworks. *Nature* **2015**, *519*, 1038.
- (9) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chem. Mater.* **2014**, *26*, 6185–6192.
- (10) Wen, X.; Zhang, Q.; Guan, J. Applications of metal-organic framework-derived materials in fuel cells and metal-air batteries. *Coord. Chem. Rev.* **2020**, *409*, 213214.
- (11) Anderson, R.; Gómez-Gualdrón, D. A. Increasing topological diversity during computational “synthesis” of porous crystals: how and why. *CrystEngComm* **2019**, *21*, 1653–1665.
- (12) Keupp, J.; Schmid, R. MOF structure prediction using specifically optimized blueprints. *Faraday Discuss.* **2018**, *211*, 79–101.
- (13) Boyd, P. G.; Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **2016**, *18*, 3777–3792.
- (14) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985–5998.
- (15) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal-Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17*, 5801–5810.
- (16) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **2012**, *4*, 83–89.
- (17) Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2020**, 1–24.
- (18) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilia, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785–796.
- (19) Gómez-Bombarelli, R.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120–1127.
- (20) Aghaji, M. Z.; Fernandez, M.; Boyd, P. G.; Daff, T. D.; Woo, T. K. 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**, *2016*, 4505–4511.
- (21) Tsamardinos, I.; Fanourgakis, G. S.; Greasidou, E.; Klontzas, E.; Gkagkas, K.; Froudakis, G. E. An Automated Machine Learning architecture for the accelerated prediction of Metal-Organic Frameworks performance in energy and environmental applications. *Microporous Mesoporous Mater.* **2020**, *300*, 110160.
- (22) Fernandez, M.; Trefiak, N. R.; Woo, T. K. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. *J. Phys. Chem. C* **2013**, *117*, 14095–14105.
- (23) Gopalan, A.; Bucior, B. J.; Bobbitt, N. S.; Snurr, R. Q. Prediction of hydrogen adsorption in nanoporous materials from the energy distribution of adsorption sites*. *Mol. Phys.* **2019**, *117*, 3683–3694.
- (24) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. *Mol. Syst. Des. Eng.* **2019**, *4*, 162–174.
- (25) Anderson, G.; Schweitzer, B.; Anderson, R.; Gómez-Gualdrón, D. A. Attainable Volumetric Targets for Adsorption-Based Hydrogen Storage in Porous Crystals: Molecular Simulation and Machine Learning. *J. Phys. Chem. C* **2019**, *123*, 120–130.
- (26) Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb. Sci.* **2017**, *19*, 640–645.
- (27) Fanourgakis, G. S.; Gkagkas, K.; Tylanakis, E.; Klontzas, E.; Froudakis, G. A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials. *J. Phys. Chem. A* **2019**, *123*, 6080–6087.
- (28) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. Large-scale quantitative structure-property relationship (QSPR) analysis of methane storage in metal-organic frameworks. *J. Phys. Chem. C* **2013**, *117*, 7681–7689.
- (29) Wu, X.; Xiang, S.; Su, J.; Cai, W. Understanding Quantitative Relationship between Methane Storage Capacities and Characteristic Properties of Metal-Organic Frameworks Based on Machine Learning. *J. Phys. Chem. C* **2019**, *123*, 8550–8559.
- (30) Zhang, X.; Cui, J.; Zhang, K.; Wu, J.; Lee, Y. Machine Learning Prediction on Properties of Nanoporous Materials Utilizing Pore Geometry Barcodes. *J. Chem. Inf. Model.* **2019**, *59*, 4636–4644.
- (31) Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16*, 1271–1283.
- (32) Fanourgakis, G. S.; Gkagkas, K.; Tylanakis, E.; Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 3814–3822.
- (33) Gharagheizi, F.; Tang, D.; Sholl, D. S. Selecting Adsorbents to Separate Diverse Near-Azeotropic Chemicals. *J. Phys. Chem. C* **2020**, *124*, 3664–3670.
- (34) Batra, R.; Chen, C.; Evans, T. G.; Walton, K. S.; Ramprasad, R. Prediction of water stability in metal-organic frameworks using machine learning. *Nat. Mach. Intell.* **2020**, *2*, 704.
- (35) Düren, T.; Sarkisov, L.; Yaghi, O. M.; Snurr, R. Q. Design of new materials for methane storage. *Langmuir* **2004**, *20*, 2683–2689.

- (36) Wu, H.; Simmons, J. M.; Liu, Y.; Brown, C. M.; Wang, X.-S.; Ma, S.; Peterson, V. K.; Southon, P. D.; Kepert, C. J.; Zhou, H.-C.; Yildirim, T.; Zhou, W. Metal-Organic Frameworks with Exceptionally High Methane Uptake: Where and How is Methane Stored? *Chem.—Eur. J.* **2010**, *16*, 5205–5214.
- (37) Alezi, D.; Belmabkhout, Y.; Suyetin, M.; Bhatt, P. M.; Weseliński, L. J.; Solovyeva, V.; Adil, K.; Spanopoulos, I.; Trikalitis, P. N.; Emwas, A.-H.; Eddaoudi, M. MOF Crystal Chemistry Paving the Way to Gas Storage Needs: Aluminum-Based *sof*-MOF for CH₄, O₂, and CO₂ Storage. *J. Am. Chem. Soc.* **2015**, *137*, 13308–13318.
- (38) Tran, H. D.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, 171104.
- (39) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.
- (40) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel. *J. Cheminf.* **2011**, *3*, 1–14.
- (41) Breiman, L.; Spector, P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int. Stat. Rev.* **1992**, *60*, 291.
- (42) Lundberg, S. M.; Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, 4766–4775.
- (43) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (44) Wick, C. D.; Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 4. United-Atom Description of Linear and Branched Alkenes and Alkylbenzenes. *J. Phys. Chem. B* **2000**, *104*, 8008–8016.
- (45) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **2016**, *42*, 81–101.
- (46) Nazarian, D.; Ganesh, P.; Sholl, D. S. Benchmarking density functional theory predictions of framework structures and properties in a chemically diverse test set of metal-organic frameworks. *J. Mater. Chem. A* **2015**, *3*, 22432–22440.
- (47) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169–11186.
- (48) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953–17979.
- (49) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (50) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
- (51) Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N. Density Functional Theory in Transition-Metal Chemistry: A Self-Consistent Hubbard *U* Approach. *Phys. Rev. Lett.* **2006**, *97*, 103001.
- (52) Wang, L.; Maxisch, T.; Ceder, G. Oxidation energies of transition metal oxides within the GGA + *U* framework. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 195107.
- (53) Mann, G. W.; Lee, K.; Cococcioni, M.; Smit, B.; Neaton, J. B. First-principles Hubbard *U* approach for small molecule binding in metal-organic frameworks. *J. Chem. Phys.* **2016**, *144*, 174104.
- (54) You, W.; Liu, Y.; Howe, J. D.; Tang, D.; Sholl, D. S. Tuning Binding Tendencies of Small Molecules in Metal–Organic Frameworks with Open Metal Sites by Metal Substitution and Linker Functionalization. *J. Phys. Chem. C* **2018**, *122*, 27486–27494.
- (55) Fang, H.; Kulkarni, A.; Kamakoti, P.; Awati, R.; Ravikovitch, P. I.; Sholl, D. S. Identification of High-CO₂-Capacity Cationic Zeolites by Accurate Computational Screening. *Chem. Mater.* **2016**, *28*, 3887–3896.
- (56) Tranchemontagne, D. J.; Mendoza-Cortés, J. L.; O'Keeffe, M.; Yaghi, O. M. Secondary building units, nets and bonding in the chemistry of metal–organic frameworks. *Chem. Soc. Rev.* **2009**, *38*, 1257–1283.
- (57) Ma, S.; Sun, D.; Simmons, J. M.; Collier, C. D.; Yuan, D.; Zhou, H.-C. Metal-Organic Framework from an Anthracene Derivative Containing Nanoscopic Cages Exhibiting High Methane Uptake. *J. Am. Chem. Soc.* **2008**, *130*, 1012–1016.
- (58) Park, J.; Landa, H. O. R.; Kawajiri, Y.; Realff, M. J.; Lively, R. P.; Sholl, D. S. How Well Do Approximate Models of Adsorption-Based CO₂ Capture Processes Predict Results of Detailed Process Models? *Ind. Eng. Chem. Res.* **2020**, *59*, 7097–7108.