Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model

Ryther Anderson, Achay Biong, Diego A. Gómez-Gualdrón*

Department of Chemical and Biological Engineering, Colorado School of Mines, Golden CO 80401, USA

ABSTRACT: Tailoring the structure and chemistry of metal-organic frameworks (MOFs) enables the manipulation of their adsorption properties to suit specific energy and environmental applications. As there are millions of possible MOFs (with tens of thousands already synthesized), molecular simulation has frequently been used to rapidly evaluate the adsorption performance of a large set of MOFs. This allows subsequent experiments to focus only on a small subset of the most promising MOFs. In many instances, however, even molecular simulation becomes prohibitively time consuming, underscoring the need for alternative screening methods, such as machine learning, to precede molecular simulation efforts. In this study, as a proof of concept, we trained a neural network—specifically, a multilayer perceptron (MLP)—as the first example of a machine learning model capable of predicting full adsorption isotherms of different molecules not included in the training of the model. To achieve this, we trained our MLP on "alchemical" species— represented only by variables derived from their force field parameters—to predict the loadings of real adsorbates. Alchemical species used for training were small, near-spherical, and nonpolar, enabling the prediction of analogous *real* molecules relevant for chemical separations such as argon, krypton, xenon, methane, ethane, and nitrogen. MOFs were also represented by simple descriptors (e.g. geometric properties and chemical moieties). The trained model was shown to make accurate adsorption predictions for these six adsorbates in both hypothetical and existing MOFs. The MLP presented here is not expected to be applied "as is" to more complex adsorbates with properties not considered during its training. However, our results illustrate a new philosophy of training that can be built upon with the goal of predicting adsorption isotherms in not only a database of MOFs, but also for a database of adsorbates, and over a range of relevant operating conditions.

1. INTRODUCTION

Advanced porous crystals are promising materials in a number of technologies used to mitigate energy- and environment-related problems. For instance, chemical separations requiring large inputs of energy (e.g. cryogenic distillation) could instead be performed using specially tailored porous materials to retain one component selectively (and abundantly),[1–4] ultimately allowing for separation at relatively mild (i.e. non energy-intensive) conditions.[5] Porous crystals include well-known materials such as zeolites,[6] as well as emerging materials such as porous organic cages (POCs),[7] covalent-organic frameworks (COFs)[8] and metal-organic frameworks (MOFs).[9] While crystal tailoring for a specific application is perhaps most readily achieved in MOFs,[10,11] all these materials exhibit an exceptionally large diversity of chemistries and architectures, stemming from the use of different synthetic precursors.[11–14] The number of possible synthetic precursor combinations implies an overwhelming number of possible materials, a number that would be impossible to exhaustively synthesize and experimentally test to find optimal candidates for a specific application.

Consequently, molecular simulation has been frequently used to aid the discovery of porous crystals by performing "computational experiments."[15] For instance, grand canonical Monte Carlo (GCMC) simulations have been used to predict adsorption capabilities in large material databases.[16,17] As the development of more accurate descriptions of relevant intermolecular interactions with new forcefields continues, the matching between GCMC and experiments will continue to improve.[18–20] By using GCMC, one can "narrow down" a large database of materials to a smaller set of potentially high-performing materials on which to devote experimental efforts.[21–24] Through this "hierarchical" approach, GCMC has led to the identification of, for instance, NOTT-101 and SBMOF-1 as high-performing MOFs for CO_2/H_2 and Xe/Kr separation, respectively.[23,25]

However, depending on the size of the database, the number and type of adsorbates involved, the operating conditions, and the number of compositions to be tested, even GCMC simulations can become prohibitively computationally intensive for comprehensive screening. This is a critical drawback if one must solely rely on GCMC for screening, especially considering that recent improvements in algorithms used to "computationally synthesize" porous crystals allow for the creation of databases of unprecedented sizes.[11,26,27] Therefore, building on the hierarchical screening philosophy, a computational "pre-screening" method that allows GCMC to be devoted only to the most promising materials in a database is not only desirable, but potentially necessary to maintain the efficacy of computational high-throughput screening.

Several methods have been considered for pre-screening databases, including estimation of performance metrics using analytical equations with faster-to-calculate descriptors such as Henry's constants[28–30] and surface areas[31,32] as inputs.

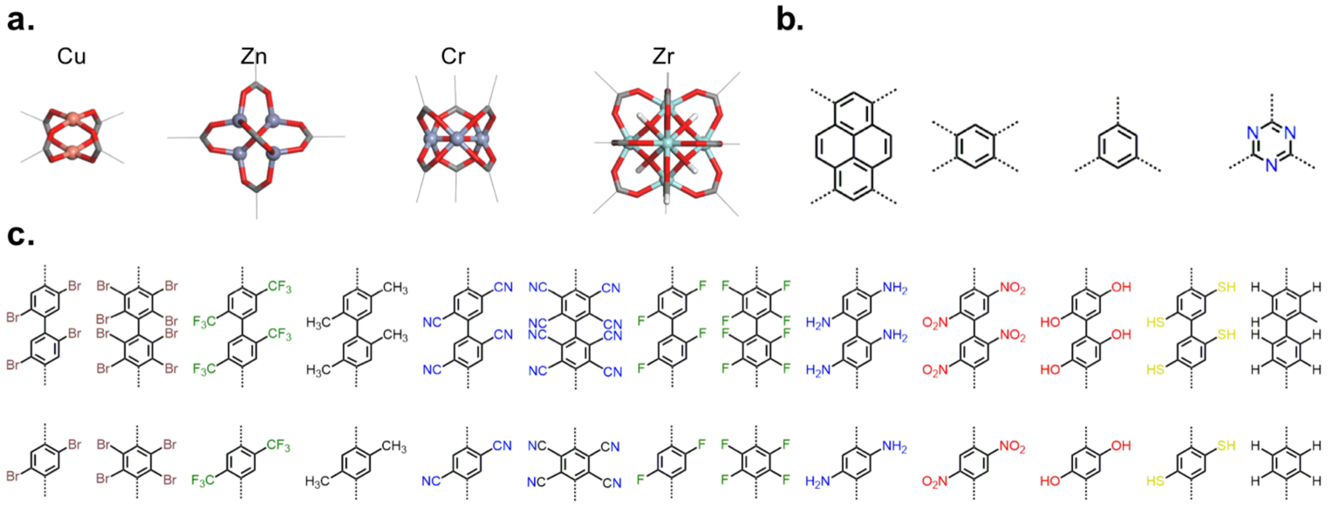


Figure 1. The building blocks used for MOF database construction, dashed lines indicate connections to the rest of the framework; a. inorganic (metal-containing) nodes, which include Cu (4-connected), Zn (6-connected), Cr (6-connected), and Zr (8- and 12-connected) oxoclusters, b. organic nodes (the central part of multitopic organic linkers) c. connecting building blocks (the arms of multitopic organic linkers or the body of ditopic linkers).

However, perhaps the most intriguing prospect is the use of machine learning predictions for the pre-screening stage. Some of the first efforts using machine learning to predict adsorption were presented by Woo and coworkers, who used support vector machines (SVMs) to predict methane adsorption using crystal textural properties such as void fraction, surface area, and pore size as performance descriptors.[33] The array of descriptor values used to represent a material or molecule for training a predictive algorithm is commonly referred to as a “fingerprint”. Woo and coworkers also presented machine learning-based predictions of CO_2 adsorption, made with more complex fingerprints (e.g. atomic property-weighted radial distribution functions) as inputs.[34] In other prominent examples, Smit and coworkers used random forests (RFs) and artificial neural networks (ANNs) to predict Xe/Kr[35] and hydrogen adsorption,[17] respectively, but with a fingerprint that included some simulation-calculated, energy descriptors. Using simple descriptors instead, Fernandez and coworkers used decision trees (DTs) and SVMs to broadly classify materials for CO_2/N_2 separation (e.g. as “potentially good”).[36] Also using new, but still easily-interpretable descriptors, (e.g. metallic percentage, topology, and the chemical identities of building blocks) Srivastava and coworkers predicted methane adsorption using RFs,[37] while Froudakis and coworkers predicted hydrogen and CO_2 adsorption using RFs and SVMs, respectively.[38] Previously, we also predicted loading, selectivity, and working capacity for CO_2 capture from gas mixtures with several different algorithms, finding that the highest accuracy was achieved with gradient boosting machines (GBMs).[39]

The above machine learning efforts have been constrained to generally the same approach: *i)* GCMC is used to simulate the adsorption of a given adsorbate or adsorbate mixture for materials (e.g. MOFs) in a database at a specific operating condition, and then *ii)* an algorithm is trained to predict the simulated adsorption data using material properties—i.e. a material fingerprint—as inputs. It is often noted that the final algorithm could be used to screen new adsorbents, which is an endeavor that may be worthwhile if a new database emerges or the original database grows drastically. However, algorithms trained under this approach can only evaluate new materials for the combination of adsorbates and operating conditions that they were originally trained on. Clearly, this approach severely limits the scope of the predictive algorithms, especially considering that a need to explore the same database for *other* adsorbates (or adsorbate mixtures) and/or *other* operating conditions is more likely to arise than a need to explore *another* database.

Recently, Sholl and coworkers[40] underscored the low diversity of adsorbates so far considered in computational screening by noting that most adsorption studies on material databases focused on CO_2, CH_4 and H_2. This focus is mainly driven by interest in energy storage and carbon capture. However, the potential of advanced porous crystals extends to applications involving a much larger diversity of adsorbates. For instance, current commercial applications of MOFs involve unusual adsorbates such as 1-methylcyclopropene and boron trifluoride.[41] Other potential applications in refrigeration,[42] medicine,[43] protection against chemical warfare agents,[44] and a myriad of chemical separations,[45–47] involve many other adsorbates (e.g. CH_3OH, O_2, H_2O, H_2S). Separations relevant to the oil and gas industry can involve complex mixtures of $C_nH_mO_xN_yS_z$ adsorbates.[48] Recognizing the need for faster ways to predict adsorption for a diversity of adsorbates, Sholl and coworkers[40] tried predicting isotherms for 24 adsorbates using the Langmuir model with simulation-calculated Henry’s constants and saturation loadings. Two caveats to this approach are its lack of extensibility to non-Langmuir-shaped isotherms, and the need to calculate new Henry’s constants and saturation loadings for new temperatures. However, these caveats could be potentially overcome using machine learning.

In recent work,[49] we found that a single multi-layer perceptron (MLP), a class of ANN, was able to predict full hydrogen isotherms and isobars, which requires predicting adsorption at temperatures and pressures not included in the training data. That is, the algorithm is required to learn the behavior of loading with changes in temperature and pressure, for a diverse range of materials (and thus isotherm/isobar shapes). In the cited work, we used inherent material properties (similar to those discussed previously), temperature (T) and pressure (P), and the relevant force field parameter describing

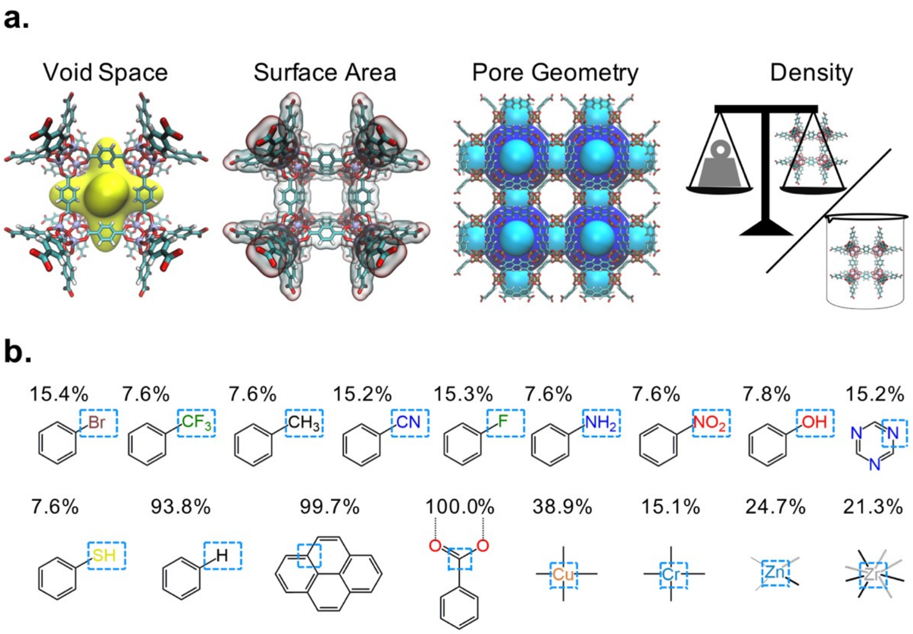


Figure 2. Descriptors constituting the MOF fingerprint. a. Six textural properties: void fraction, gravimetric surface area, largest pore diameter (LPD: dark blue sphere), pore limiting diameter (PLD: light blue sphere), pore size standard deviation (PSSD), and density. b. 17 chemical motifs (boxed), for which their respective number density in each MOF was calculated. The percentage of MOFs in the database that contain each motif is listed at the top.

the "chemistry" of adsorbate/adsorbent interactions as our descriptors. The success of including operational (T and P) and adsorbate-dependent descriptors (force field parameters) motivated us to investigate the suitability of machine learning as a tool that could make universal adsorption predictions possible. A prerequisite for such "universal" tool is the ability *to predict adsorption for molecules for which it was not originally trained*.

Given that there is ongoing debate on the scope of machine learning and the best strategies to train machine learning models even when focused on a specific adsorbate or mixture, a first step toward the development of a universal model is to study whether the same machine learning model that is used to predict the adsorption of a given adsorbate can actually be used to predict the adsorption of a different adsorbate. Accordingly, the work herein focuses on demonstrating such capability, considering the substantial increase in the complexity of the data that arises when including different adsorbates (even simple ones, as those considered here) along with different operating conditions. Additionally, an underlying theme in our work is to make the machine learning algorithm as accurate as possible while keeping model inputs brief, easily interpretable, and *obtainable with minimal computational effort*. To generate training data, we focused on the adsorption of 200 alchemical species (i.e. adsorbates that do not necessarily have force-field parameters that replicate the structure, chemistry or physics of any *real* atom or molecule) at room temperature in a relatively small, topologically and chemically diverse database of 2,400 MOFs created using our Topologically-Based Crystal Constructor (ToBaCCo) code.[10,11] We tested the model on real adsorbates (Ar, Kr, Xe, methane, ethane, and N_2) partly chosen due to their relevance to several gas storage [31,43,50] and chemical separation[1,51–55] applications. We limited the number of MOFs in our database to keep the number of simulations needed to generate the requisite data reasonable.

2. DATA GENERATION

2.1 Database construction. The ToBaCCo-3.0 code[10,11] was used to "computationally synthesize" 2,400 MOFs of 50 topologies using the building blocks illustrated in **Fig. 1**. This selection includes commonly observed oxometallic nodes and a diversity of non-metal chemical moieties found in existing MOFs. The selection of building blocks was made aiming to maximize the diversity of topologies and framework-adsorbate interactions in a relatively small database, which in turn helped us maintain the number of simulations needed to generate training data reasonable. Each MOF prototype constructed by ToBaCCo was structurally optimized in LAMMPS (version 31 Mar 2017) [56] using the Dreiding[57] force field. In Dreiding, functional forms and force constants for bond and angle terms are independent of atom types, but equilibrium bond lengths and angles are unavailable for some metals. In such cases, we used the crystallographic bond lengths/angles from experimental CIFs as the equilibrium bond lengths/angles. For the optimization, we used an iterative approach where in each iteration the atom coordinates were first optimized using the fast inertial relaxation engine (FIRE) algorithm developed by Bitzek et al[58] with a timestep of 10.0 fs with the MOF cell parameters fixed. Then, the atom positions and unit cell parameters were optimized together using a conjugate gradient algorithm. For each iteration, the first and second step optimizations were stopped when the change in energy between consecutive geometries divided by the energy of the last geometry was less than 1.0×10^{-6} and no atom experienced a force larger than 1.0×10^{-6} kcal/mol Å^{-1}. The iterations were stopped when the energy change from the previous iteration to the current iteration was less than 1.0×10^{-6} kcal/mol.

2.2. Training, validation, and test set data generation. Before any GCMC simulations were run, we randomly split our 2,400 MOFs into 1,800 training MOFs, 200 validation MOFs, and 400 test MOFs.

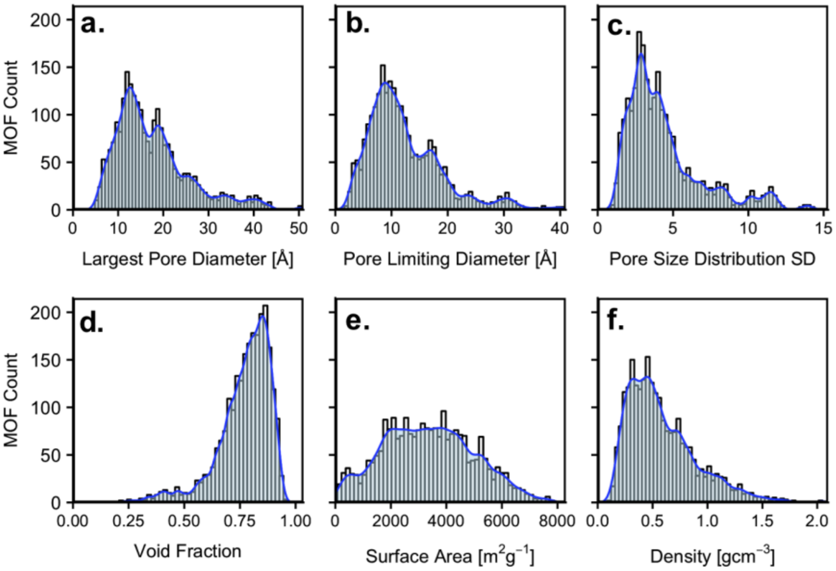


Figure 3. Histograms of the textural properties used as descriptors in the MOF fingerprint.

To generate our training data, we ran GCMC simulations for 200 one-, and two-atom alchemical adsorbates at fugacities of 1, 5, 10, 50, 75, and 100 bar in the 1,800 training MOFs (fugacity was used as opposed to pressure, so we did not have to calculate critical constants for each alchemical species). To generate our validation data, we ran GCMC simulations for 200 one-, and two-atom alchemical adsorbates (all were entirely different than any adsorbate used for computing the training data) at fugacities of 2.5, 30, 60, 80, and 90 bar in the 200 validation MOFs. The LJ parameters, charges and bond-lengths used to generate all of the alchemical species considered are given in **Tables S1-S4**. To generate our test data, we ran GCMC simulations for 12 *real* adsorbates (argon, krypton, methane, xenon, nitrogen, ethane, helium, hydrogen, propane, butane, isobutane, and benzene) at fugacities of 1, 2.5, 5, 10, 25, 50, 60, 75, 80, and 100 bar in the 400 test MOFs. Note that the real adsorbates considered here were distinct from the alchemical species used to generate the training/validation data, i.e. no real adsorbate had the same LJ, charge, or bond-length parameters as the alchemical ones. GCMC simulations for six real adsorbates (argon, krypton, methane, xenon, nitrogen, ethane) were run in 1,528 MOFs from the CoRE MOF database[59,60] at fugacities of 1, 5, 10, 50, 75, and 100 bar for additional testing data including experimental MOFs.

LJ parameters for helium correspond to those used by Smit and coworkers,[61] LJ parameters for argon were taken from Perez and coworkers,[62] and LJ parameters for krypton, and xenon correspond to those used by Sikora and coworkers.[63] The parameters for these adsorbates are summarized in **Table S5**. Methane, ethane, propane, n-butane, isobutane, benzene, and nitrogen, were modeled according to the TraPPE force-field developed by Siepmann and coworkers.[64,65] Accordingly, methane, ethane, propane, n-butane/isobutane, and benzene are modeled as a one, two, three, four, and six uncharged sites, respectively, while nitrogen is modeled with three charged sites in order to reproduce electric quadrupoles. For nitrogen, the two atoms are each assigned LJ parameters and charges, while a "dummy" site at the center of mass is only assigned a charge.[64,65] LJ parameters and charges for hydrogen were taken from the Darkim-Levesque model,[66,67] which is also a three-site model, however, the two atoms are each assigned only a charge while a dummy site at the center of mass is assigned LJ parameters and a charge. **Fig. S1** shows the generic one- and two-atom force field models used for GCMC simulations.

2.3. MOF fingerprinting. We tested seven different MOF fingerprints (see results in **Table S7**), and found that a fingerprint combining six MOF textural properties—helium void fraction (V_F), gravimetric surface area (GSA), largest pore diameter (LPD), pore limiting diameter (PLD), inverse framework density ($1/\rho_F$), and the pore size standard deviation (PSSD)—together with the number density of 17 distinct MOF chemical moieties resulted in sufficiently accurate predictions. The descriptors for the fingerprint are illustrated in **Fig. 2**. While we did identify another feature set—which we nominally refer to as the *bag-of-atoms*—that provided slightly more accurate predictions, we determined that the slight increase in model accuracy was not worth the significant increase in model complexity required to use this descriptor set (further details are provided in **Section S2**).

The range of chemistries covered by our constructed MOFs is shown in **Fig. 2**, with the associated percentages indicating the frequency with which it appears in the database. The range of textural properties covered by the constructed MOFs is shown by the histograms in **Fig. 3**. Our optimal fingerprint can be considered simple because it is limited to 23 easily-calculated descriptors instead of the hundreds needed when using atomic-property weighted radial distribution functions[34] or other high dimensional descriptors (e.g. bag-of-atoms). V_F was calculated using the Widom insertion method with helium as the probe molecule,[68] while GSA was calculated by rolling a nitrogen-sized spherical probe along the framework surface.[69] Both of these calculations were implemented in RASPA-2.0.[70] LPD and PLD were calculated using zeo++ (version 0.2.2).[71] PSSD, as a measure of pore polydispersity, was calculated by taking the weighted standard deviation of the pore size distribution (also calculated using RASPA-2.0), where each pore diameter was weighted by the distribution value. For each MOF, the number density of a given chemical moiety was calculated by counting the number of times that moiety appeared in the unit cell and dividing by volume of the latter.

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

$$\sigma_{\text{effective}} = (2\sigma_{ii} + r_{\text{bond}})/2 \qquad (1)$$

which is the average of σ_{ii} and the end-to-end length of the molecule if we consider the diameter of each atom to be σ_{ii}. For the more complex adsorbates considered in **Section 4.6**, $\epsilon_{\text{effective}}$ was taken to be the sum of all ϵ_{ii} values, and $\sigma_{\text{effective}}$ was taken to be the average of the shortest dimension and longest dimension (an extension of Eq. 1 to adsorbates with more than one bond). Similarly, the bond length, r_{bond}, was taken to be the longest distance between atom coordinates in the lowest energy geometry (as calculated according to the relevant force field, see above). **Fig. S2** shows an example of how the fingerprint described above is calculated for both mono- and diatomic adsorbates. Four other adsorbate fingerprints were considered, these are discussed in detail in **Section S2**.

2.5. Adsorption Simulations. RASPA-2.0[70] was used to perform all GCMC simulations. In grand canonical simulations chemical potential, volume, and temperature are kept constant. Chemical potentials were calculated directly from fugacity. Simulations consisted of 2,000 initialization cycles (no data recording) and 2,000 production cycles (data recording) for all one and two atom adsorbates. Each cycle consists of *N* Monte Carlo moves (translation, rotation, or insertion/deletion), where *N* is the highest value between 20 and the number of adsorbates in the simulation cell. For propane, n-butane, isobutane, and benzene, simulations consisted of 15,000 initialization cycles followed by 5,000 production cycles, and configurational bias was used during insertion moves. Adsorbate-adsorbate interactions were modeled using Lennard-Jones (LJ) potentials to describe dispersion interactions and Coulomb's law to describe charge-charge interactions. Available Dreiding forcefield[57] parameters were assigned to framework atoms, otherwise UFF[72] parameters were used. Framework-adsorbate electrostatic interactions were ignored as the partial charges of the adsorbates were small. The negligible effect of this choice on loadings was shown earlier for N_2 adsorption at 77 K,[69] but was tested and verified again here on our test set MOFs (**Fig. S3**). This test was possible by assigning atomic charges to each MOF according to our MBBB approach.[73] Lorentz-Berthelot mixing rules were used to calculate parameters for interactions between atoms not explicitly parametrized. Adsorbate forcefield parameters were assigned as discussed in **Section 2.2**.

3. NEURAL NETWORK TRAINING

3.1. Model training. Here we trained a multilayer perceptron (MLP, see **Fig. S4**) to predict the adsorption data obtained from GCMC simulations. All MLPs were trained using Keras[74] through the SciKit-learn[75] Python package (all version numbers for Python packages used during training are given in **Table S6**). First, before training the final MLP, we investigated different network hyperparameter configurations to determine which hyperparameter values were likely to give an accurate final model. During this procedure (called tuning) we assessed model performance using both mean absolute error (MAE) and mean absolute percentage error (MAPE) on the validation set. These errors were selected as they both have useful and relevant physical interpretations (and are what we are most concerned with minimizing when predicting loading).

Tuning was performed using a two-step procedure. First, we exhaustively investigated diverse neural network topologies from one to eight hidden layers with between 10 and 50 nodes (in increments of 10 nodes) in each layer, keeping all other hyperparameters fixed, to find a class of network topologies which generally gave the most accurate results. We found that one, two, and three hidden layer networks, while making reasonably accurate and reproducible predictions, had higher error than deeper MLPs. On the other hand, we found that many deep networks (more than five layers) had low minimum error on the validation set, but were sensitive, i.e. we observed large oscillations in the validation set error across epochs and with slight changes in network topology for these networks. Therefore, we selected a four-hidden-layer topology for our final model, as it was both highly accurate, and robust in its predictions. Second, and after settling upon this network topology, we varied other important net hyperparameters on a grid, keeping topology constant. The hyperparameters considered and their final values (**Table S8**) are presented in **Section S2**.

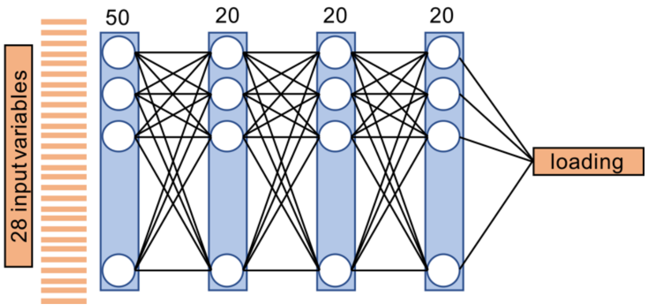


Figure 4. The configuration of our final model. The number of nodes in each hidden layer are shown above the corresponding layer.

The final model resulting from the above tuning procedure was then tested on the real adsorbates in the test set MOFs, and in the CoRE MOFs (see below). Additional data, demonstrating the reproducibility of our model, is presented in **Fig. S5**. The architecture of this final model is shown in **Fig. 4**. We reiterate that there were *no shared MOFs or adsorbates* between the training, validation, and test sets, during any model training.

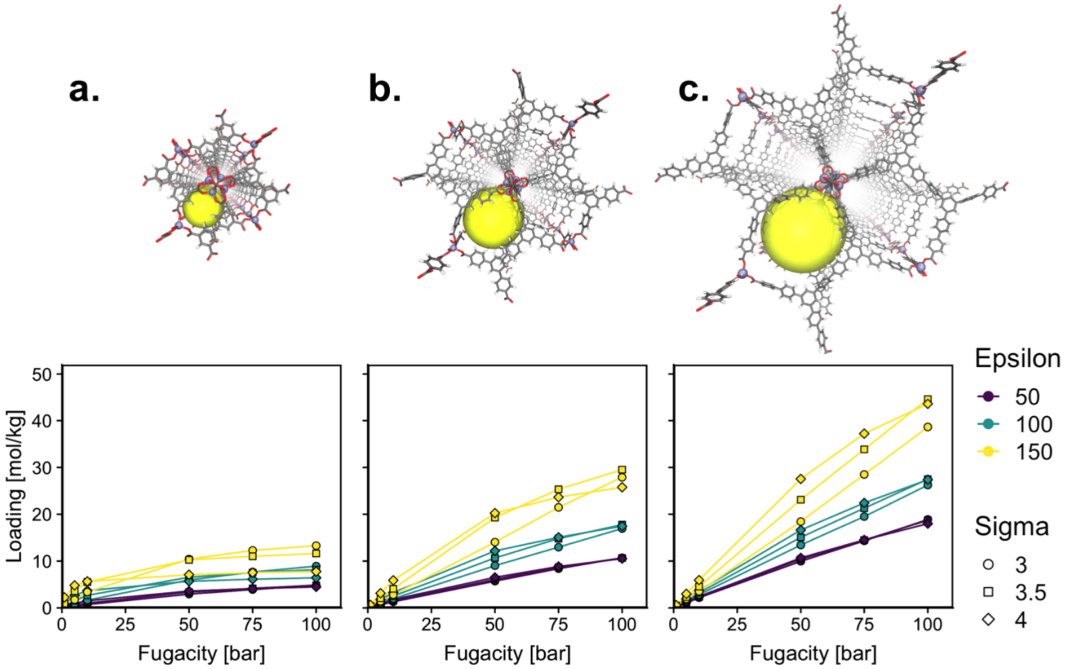


Figure 5. Isotherms for alchemical adsorbates considered in a. a representative small pore MOF (LPD=6.5 Å), b. a representative intermediate pore MOF (LPD=13.7 Å), and c. a representative large pore MOF (LPD=19.5 Å). All the MOFs shown are of the **mcn** topology. The large pore in each case is illustrated by a yellow sphere.

In addition, there were no shared fugacities between the training and validation set. We considered both shared and unshared fugacities between the training and test set. Every network considered was trained for a maximum of 500 epochs. Early termination with a patience of 20 epochs was employed to prevent over-fitting. That is, if validation loss (MAE for the final MLP) did not improve for 20 epochs in a row, training was terminated and the lowest loss from the previous epochs was taken to be the model error. A measure of importance for each of the 28 descriptors used as input for the neural network is given in **Table S9**.

4. RESULTS AND DISCUSSION

4.1. Model predictive ability for simple adsorbates. Before discussing the overall predictive performance of our final MLP, we discuss briefly, from an intuitive perspective, some of the things that the model must learn. **Fig. 5** shows adsorption isotherms for a subset of the single-atom alchemical adsorbates (from the training set) in three MOFs. These MOFs are representative of structures with small (LPD ~ 7 Å), intermediate (LPD ~ 14 Å, or near the first peak in the LPD histogram shown in **Fig. 3**) and large (LPD ~ 20 Å, or near the second peak in the LPD histogram shown in **Fig. 3**) pores, respectively. The adsorbates with the largest ϵ_{ii} and σ_{ii} have the highest loading at low fugacities in all three MOFs. Thus, the model thus must learn that intrinsic adsorbate-adsorbent interactions play a dominant role in controlling adsorbate loadings at low fugacity. However, the larger ϵ_{ii} and σ_{ii} are, the more easily saturation is reached as fugacity increases, giving rise to pore size limitations. For instance **Fig. 5** shows that, at high fugacity, and in small pore MOFs, a smaller molecule (σ_{ii}= 3) with a weaker interaction (ϵ_{ii}= 100), can have a higher loading than a larger molecule (σ_{ii}= 4) with a stronger interaction (ϵ_{ii}= 150). However, this is no longer the case in the large pore MOF. The model, consequently, must learn that pore size limitations play a dominant role for larger adsorbates in smaller pores. Of course, this is just a brief glimpse into the intricacies of the interplay between MOF and adsorbate properties and fugacities (operating conditions) that determine adsorbate loading, and which the model must learn.

Next, we consider the predictive performance of our final model on a set of real adsorbates by comparing our model predictions to GCMC calculated loading of argon, methane, krypton, xenon, ethane, and nitrogen at 10 fugacities from 1 to 100 bar in the 400 test set MOFs. We computed several measures of model performance for each adsorbate, which are presented in **Table 1**. Specifically, we consider mean absolute percentage error (MAPE), mean absolute error (MAE), and Pearson correlation (*R*). Perfect predictions would have zero MAPE and MAE and an *R* value of unity.

Table 1. Model performance metrics of our final model for loading predictions made on the test set MOFs.

Adsorbate	MAPE [%]	MAE [mol//kg]	R
Argon	4.4	0.17	0.999
Methane	4.6	0.24	0.999
Krypton	4.7	0.30	0.999
Xenon	4.5	0.42	0.999
Ethane	4.3	0.42	0.999
Nitrogen	4.0	0.14	0.999

Values of *R* close to one indicate a very strong linear correlation between GCMC loadings and those predicted by the MLP. As a point of comparison, our final model predicted the validation set loadings (200 alchemical adsorbates in 200 MOFs at 6 fugacities) with a MAPE of 3.0 % and a MAE of 0.18 mol/kg. We note that MAPE is biased towards adsorbates with higher loadings, since a larger absolute error may still be a relatively low absolute percentage error. For example, while nitrogen and argon predictions are visibly accurate (and have

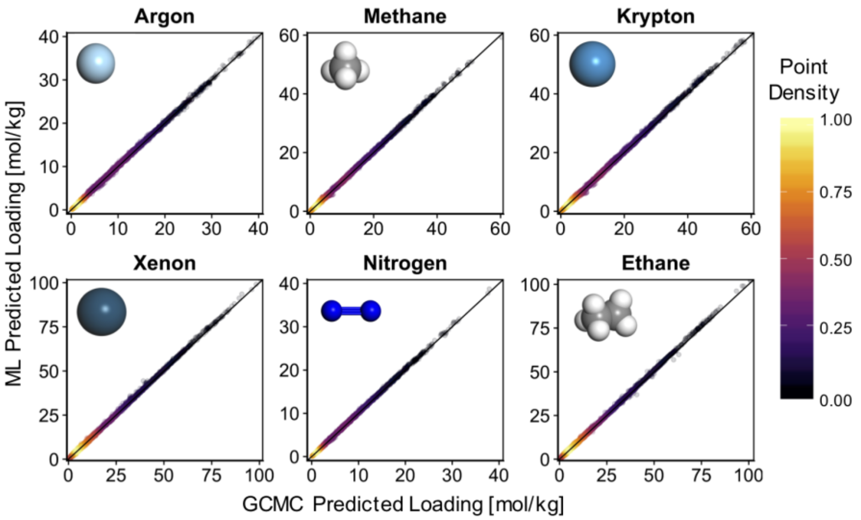


Figure 6. Parity plots comparing the predictions of the final MLP model for the six indicated adsorbates versus GCMC-calculated values in the 400 test set MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings).

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

Parity plots showing the MLP predicted loadings (at all 10 fugacities considered in the test set) versus the corresponding GCMC simulated loadings for each adsorbate provide a more complete picture of the predictive capabilities of the final MLP (**Fig. 6**). Perfect predictions would result in all the points in these plots falling on the diagonal line. It is clear from **Table 1** and **Fig. 6** that the final MLP predicts loadings of the six adsorbates considered here exceedingly well, as expected, given that the same model predicted the loadings of 200 similar alchemical adsorbates with similar accuracy. Parity plots, however, do not give a complete picture of the performance of a model trained to predict adsorption loading. Not only should the model predict individual loading points correctly, but it should also predict related points in the correct order and all with a similar level of accuracy. That is, the model should be able to reproduce full isotherms, which means it is accurate at each point, and also predicts the shape of the isotherm.

4.2. Predictive ability for full isotherms. Now we proceed to illustrate the ability of our model to predict full adsorption isotherms. This is a necessary if one aims to, for instance, couple machine learning predictions for pure components with IAST theory (*when applicable*) to rapidly obtain mixture adsorption data to screen MOFs for chemical separation applications. **Fig. 7** compares isotherms predicted by the MLP (continuous line) and those obtained from GCMC simulations (points) for methane and ethane (plots for the other adsorbates are given in **Fig. S6**). To get a more accurate picture of the GCMC-simulated isotherms we ran simulations in the test set MOFs at fugacities not included in the training set (empty points in **Fig. 7**).

As it is unfeasible to present the isotherm comparison for all the test cases studied here (six adsorbates in 400 MOFs), we chose to present five isotherms per adsorbate. However, to provide a fair picture of prediction accuracy, we aimed to present a range of “best to worst” cases. To do so, we first ranked all the isotherms predicted by our MLP according to their isotherm mean percentage error (IMPE). For each MLP-calculated isotherm point for which we also had a GCMC-simulated value (10 fugacities for each MOF), we estimated the absolute percentage error (APE). The mean of this set of APE values was taken to be the IMPE. The IMPE values were then used to classify the MLP-predicted isotherms into five quantiles—the 0.00 (Q1), 0.25 (Q2), 0.50 (Q3), 0.75 (Q4), and 1.00 (Q5) quantiles. Thus, Q1 isotherms are representative of the *best* predicted isotherms according to our IMPE metric, Q5 isotherms are representative of the *worst*, and Q3 isotherms are representative of an “average” (or median) prediction. **Fig. 7** and **Fig. S6** present one isotherm from each quantile (the one nearest to the IMPE quantile value).

Q1 isotherms are quantitively correct for all adsorbates, Q5 isotherms tend to be qualitatively correct but can deviate more significantly from GCMC-simulated values in some pressure ranges. However, as machine learning predictions are intended for use in high throughput screening of MOFs (or other porous crystals) some less than stellar predictions are acceptable as long as the vast majority of predictions are good. This is the case even for isotherms in the Q3 and Q4 quantiles. For instance, Q3 isotherms (the “average” prediction accuracy) have IMPEs ranging from 3.23 % (for nitrogen) to 3.93% (for krypton). As a point of comparison, Dokur and Keskin[76] showed that that a difference of well over 10% can be observed in GCMC predicted loadings. of methane and nitrogen (albeit in CO_2/N_2 and CO_2/CH_4 mixtures) when switching between using UFF and Dreiding LJ parameters for MOF atoms, and that these errors likely do not affect high throughput screening results significantly. Accordingly, the accuracy reached by the trained MLP is certainly suitable to accelerate materials discovery by utilizing it as part of a hierarchical screening strategy.

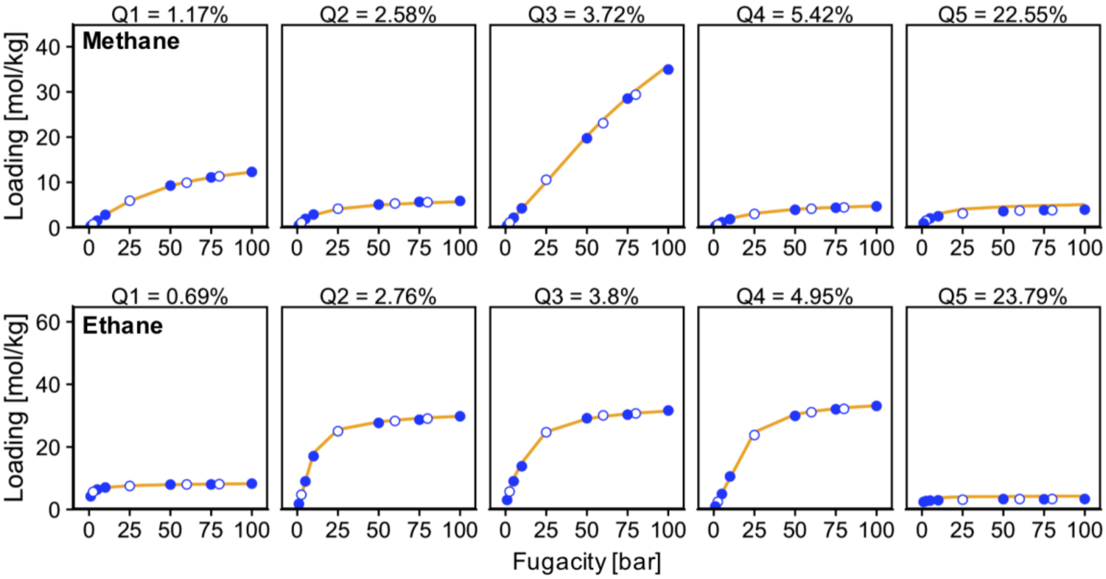


Figure 7. Isotherms for methane (top) and ethane (bottom), for the Q1 (0.00), Q2 (0.25), Q3 (0.50), Q4 (0.75), and Q5 (1.00) quantiles of isotherm mean percentage error (IMPE, with corresponding values shown). Points are GCMC simulated values (filled points correspond to fugacities included in training, empty points were not included in training), and orange lines show MLP predictions.

4.3. Transferability of model for predictions on MOFs with inaccessible pores. Some MOFs have pores that are inaccessible to certain adsorbates due to large energy barriers that separate these pores from the rest of the pore volume. Since the GCMC algorithm can insert adsorbates into such inaccessible pores, in some cases explicit "pore blocking" must be done to obtain accurate adsorption predictions. However, we did not use such blocking during the generation of training data as this would have required preparing different blocking schemes for each MOF depending on the force-field parameters for each adsorbate considered (note that which pores are inaccessible depends on adsorbate size and/or interaction energy) before running GCMC simulations and calculating textural properties. We reasoned that this procedure was unnecessarily onerous because—as proven with our use of alchemical adsorbates—the training simulation data does not have to be "realistic" to be meaningful to teach the machine learning model how adsorption depends on descriptor values. Therefore, we posited that a model trained without pore blocking considerations would still be able to correctly predict adsorption in cases where pore blocking was necessary simply by using the new set of textural properties that results from pore blocking as input for the model. As discussed in **Section S3** and illustrated in the parity plots in **Fig. S7**, the above hypothesis was shown to be true, further validating our training strategy.

4.4. Transferability of model for predictions on experimental MOFs. Although we aimed to make our database as chemical and topologically diverse as possible, one may speculate as to whether this diversity is sufficient to train a model that is applicable to the space of experimentally known MOFs. One way to assess this is to test our model on a MOF database that is not subject to the biases present in our database. Thus, we tested our model predictions on 1,528 out of 12,479 structures in the CoRE MOF database, which is a collection of experimental MOF structures curated by Chung and coworkers.[59,60] The details of our CoRE MOF selection procedure are described in **Section S4,** but the majority of discarded CoRE MOFs simply had elements not encountered in the database we used for training the model. **Fig. 8**, which shows predictions on the CoRE MOFs, is analogous to **Fig. 6** in that it compares our model predictions for argon, methane, krypton, xenon, nitrogen and ethane at six fugacities with GCMC simulations. Although the MAE, MAPE, and *R* values calculated from the model predictions (**Table S10)** on the CoRE MOFs show a decrease in overall predictive ability, the parity plots in **Fig. 8** and histograms of MAPE values (**Fig. S8**) show similar error distributions for predictions on the CoRe MOFs and the test set for our constructed database, albeit with a longer tail towards higher MAPEs for predictions on the former.

Consistent with the errors for prediction of individual adsorption loadings, full isotherm predictions by the model on CoRE MOFs remained accurate, albeit with higher errors. For instance, for the Q3 IMPE values in the CoRE MOFs vary between 6.7 % (ethane) and 11.8 % (nitrogen). Analogous plots to **Fig. 7** for the CoRe MOFs are shown in **Fig. S9**. At this point it is important to note that a significant portion of the CoRE MOFs used in our testing have chemical motifs and building blocks not seen in our database, which is likely the source of the heightened model error. Our model naturally accounts for this to a certain extent but cannot be expected to predict the influence of all the different building blocks and chemistries extant in the CoRE MOFs. For instance, while the copper motif in our database always corresponds to Cu in a paddlewheel, we can consider any Cu atom in the CoRE MOFs to be part of the copper motif, e.g. Cu as part of an "infinite" node (as in MOF-74), or as a single-atom node, and not necessarily bound to carboxylates. Such Cu motifs will influence the adsorption differently than Cu in a paddlewheel, but likely in a manner that is still correlated to the Cu number density. More specifically, for any motif that consists of a unique element (i.e. -Br, -SH, and the metals), we count any instance of that element when computing number densities. But for motifs that correspond to a particular *type* of an element (i.e. aromatic carbon, carbon in the -CN functional group, etc.) we ensure that the element is actually of that type when computing number densities. Therefore, any types of certain elements present in the CoRE MOFs that are not included in our fingerprint (e.g. triple-bonded carbon) are not explicitly considered by our model, but will be accounted for to a certain extent by the MOF textural properties.

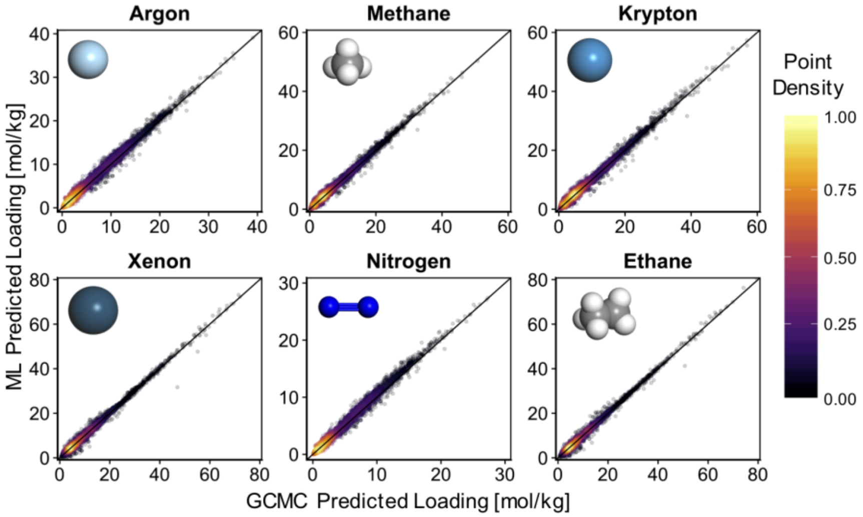


Figure 8. Parity plots comparing the predictions by the final MLP model for the six indicated adsorbates versus GCMC-calculated values in 1,528 CoRE MOFs. Points color indicate the point density in the plot (the highest density is observed at low loadings).

4.5. Material ranking accuracy based on performance metrics. For machine learning to be effectively used in hierarchical screening, it needs to correctly rank MOFs (or whichever type of porous crystal is being studied) according to some performance metric. This way, it guarantees that the most promising MOFs are studied with more accurate methods in subsequent screening stages. Thus, here, we assess the ability of our model to identify top-performing materials. One important consideration at this point is that material performance in chemical separations and gas storage often depends on adsorption properties at more than one pressure/fugacity. For instance, the working capacity, which is the difference between adsorption loadings at a high and a low pressure is a common performance metric. Thus, to evaluate the ability of our model to rank MOFs, we decided to include the ranking based on working capacities, in particular for the 100 bar ↔ 5 bar fugacity swing, we calculated for the six real adsorbates on the 400 MOF test set, and the CoRE MOF subset introduced in Section 4.4.

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

The values of *S* and the number of correctly identified top-100 MOFs are given in **Table 2** for both loadings (at 100 bar) and working capacities (for the 100 bar ↔ 5 bar fugacity swing). As an overall indication of the similarity between MLP rankings and GCMC rankings notice that the values of *S* are very close to one in all cases. As for the ability of the MLP model to capture MOFs in the top-100 for each adsorbate case, when our model was applied to the test MOFs, it was able to correctly identify at least 98 (97) of the top 100 MOFs based on the considered loadings (working capacities). When applied to the CoRE MOFs, our model correctly picked at least 91 (93) of the top-100 MOFs. Evidently, the MLP does a slightly better job ranking MOFs in the test set than in the CoRE MOFs. This is not surprising considered *(i)* the less accurate MLP predictions in the CoRE MOFs, *(ii)* the one order of magnitude higher number of CoRE MOFs considered here relative to the number of test set MOFs, which makes the selection of the top 100 inherently more difficult. As a complement to **Table 2**, **Figure S10** shows our model performance for selecting the top *N* MOFs (*N* ranges from 1 to 100) from both the test set and the CoRE MOF subset. From **Figure S10** we can observe that our model makes accurate selections (above 90% correct placement) even when considering *N* much less than 100. Based on the above results, we can conclude our MLP model is able to pick out the vast majority of high-performing MOFs from diverse databases as a first screening step.

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

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

$\sigma_{\text{effective}}$, *ii)* more sites, and *iii)* significantly different shapes than any alchemical species included in training. Thus, to make adsorption predictions for these adsorbates, the MLP model must extrapolate.

Table 2. Comparison of MOF rankings between MLP predictions and GCMC simulations as indicated by the Spearman rank correlation coefficients (*S*) and the number of MOFs in the top-100 from GCMC simulation that were also found in the top-100 from MLP predictions.

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

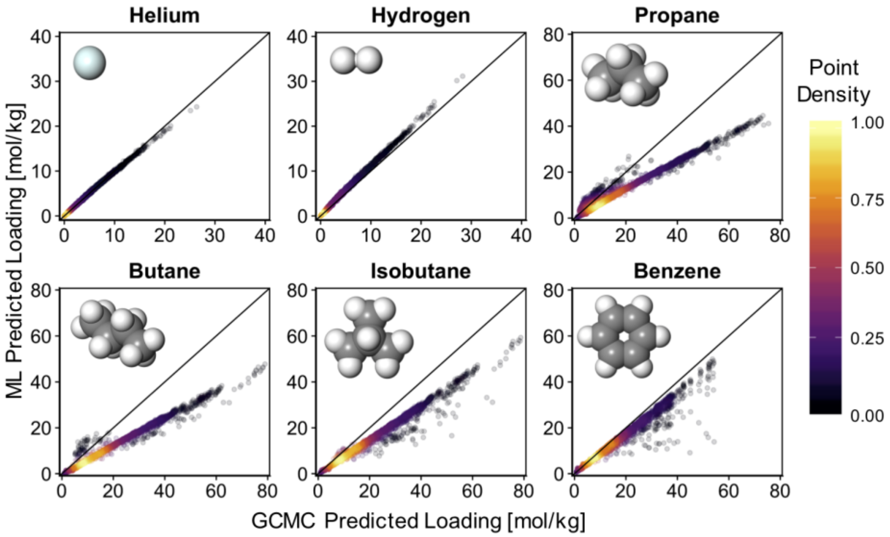


Figure 9. Parity plots comparing the predictions of the final MLP model for the six indicated (extrapolated) adsorbates versus GCMC-calculated values in the 400 test set MOFs. These adsorbates possess properties outside the ranges covered by the alchemical adsorbates used during model training. Points color indicates the point density in the plot (the highest density is observed at low loadings).

The parity plots in **Fig. 9** provide a visual comparison between MLP predictions and GCMC data for the six "extrapolated" adsorbates (*R*, MAE and MAPE values are given in **Table S11)**. Despite the extrapolation, predictions for helium and hydrogen remain quite good, making the present model applicable to, for instance, screen MOFs for CH_4/H_2 and CH_4/He separations. In contrast, adsorption predictions for propane, n-butane, isobutane, and benzene generally fail, but do so in an interesting fashion. Namely, MLP predictions and GCMC values are linearly correlated, but with systematic underprediction. Additionally, the linear correlations present a notable number of outliers, most of which to correspond to low loadings. Intriguingly, these "low-loading outliers" tend to center around the parity line. Further analysis of the predictions for the *hydrocarbons* (**Fig. S11**) revealed that MLP (under)predictions that linearly correlate with GCMC data correspond to cases where the loading is more than ~50% of the saturation value. On the other hand, MLP predictions clustering around the parity line correspond to cases where the loading is less than ~25% of the saturation value. Since a material approaches saturation as fugacity increases, the MLP model is expected to predict low-fugacity isotherm points better than high-fugacity ones. This is confirmed by inspecting select isotherms (**Fig. S12-S13)**, where the MLP model seems to capture the isotherm shape reasonably well but simply fails to correctly capture the saturation loading.

What, then, did we learn from testing the model on "extrapolated" adsorbates? Our MLP predicts the adsorption of hydrogen and helium well, likely because these molecules are small, non-polar, and near-spherical (resembling the alchemical adsorbates used for model training). The fact that the MLP was able to extrapolate well for these two adsorbates suggests that the model learned some fundamentals about the physics of adsorption. However, for the hydrocarbons, the trends observed in the MLP predictions are likely related to the non-spherical shapes of these adsorbates. Considering that the MLP was trained to predict saturation loadings (which depends on how molecules pack within MOF pores) of molecules that are near-spherical, and that the highly non-spherical hydrocarbons *will pack differently* than these near-spherical adsorbates, the overserved systematic error is not surprising. In fact, the packing of the hydrocarbons is likely to be more efficient than our model predicts, given that ellipsoids (a shape more representative of the geometries of the considered hydrocarbons) can pack more densely than spheres of the same volume.[77,78] In addition, the volume of a sphere with a diameter of $\sigma_{\text{effective}}$ is greater than the actual volume occupied by the hydrocarbons in the GCMC simulation (when calculating this actual volume as that occupied by overlapped spheres of diameter σ_{ii}). This is likely why the MLP systematically underestimates loadings for fugacities where the MOF is at or near saturation, but not so for fugacities where the MOF is far from saturation (where the molecules are isolated, and framework-adsorbate interactions are more important). The systematic loading underestimation near saturation follows the order n-butane > propane > isobutane > benzene, which is consistent with the order of asymmetry of these molecules. Based on this, one would expect that future work on developing an MLP applicable to more complex molecules than those considered in the present work should include measures of adsorbate shape/asymmetry as part of the adsorbate fingerprint (perhaps along with descriptors of MOF pore shape). The introduction these additional features could "correct the slope" for the linear correlations in **Fig. 9**.

5. CONCLUSIONS

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

ASSOCIATED CONTENT

Supporting Information.

Adsorbate force-field parameters; additional information about multi-layer perceptrons and their training; additional figures relating to the final model predictive ability.

Four comma-separated values (CSV) files containing the training, validation, test set data, and CoRe MOF adsorption data, respectively.

Saved final MLP model (HDF5 format).

AUTHOR INFORMATION

Corresponding Author

* dgomezgualdron@mines.edu

Funding Sources

NSF CAREER

ACKNOWLEDGMENTS

D.A.G.-G. acknowledges funding from NSF CAREER (CBET 1846707). Simulations were made possible by the Mio supercomputer cluster at Colorado School of Mines

REFERENCES

(1) Anderson, R.; Schweitzer, B.; Wu, T.; Carreon, M. A.; Gómez-Gualdrón, D. A. Molecular Simulation Insights on Xe/Kr Separation in a Set of Nanoporous Crystalline Membranes. *ACS Appl. Mater. Interfaces* **2018**, *10*, 582–592.

(2) Kulkarni, A. R.; Sholl, D. S. Screening of Copper Open Metal Site MOFs for Olefin/Paraffin Separations Using DFT-Derived Force Fields. *J. Phys. Chem. C* **2016**, *120*, 23044–23054.

(3) Vermoortele, F.; Maes, M.; Moghadam, P. Z.; Lennox, M. J.; Ragon, F.; Boulhout, M.; Biswas, S.; Laurier, K. G. M.; Beurroies, I.; Denoyel, R.; et al. P-Xylene-Selective Metal–Organic Frameworks: A Case of Topology-Directed Selectivity. *J. Am. Chem. Soc.* **2011**, *133*, 18526–18529.

(4) Demir, H.; Stoneburner, S. J.; Jeong, W.; Ray, D.; Zhang, X.; Farha, O. K.; Cramer, C. J.; Siepmann, J. I.; Gagliardi, L. Metal–Organic Frameworks with Metal–Catecholates for O2/N2 Separation. *J. Phys. Chem. C* **2019**, *123*, 12935–12946.

(5) Sholl, D. S.; Lively, R. P. Seven Chemical Separations to Change the World. *Nature* **2016**, *532*, 435–437.

(6) Baerlocher, C.; McCusker, L. B. Database of Zeolite Structures.

(7) Tozawa, T.; Jones, J. T. A.; Swamy, S. I.; Jiang, S.; Adams, D. J.; Shakespeare, S.; Clowes, R.; Bradshaw,

D.; Hasell, T.; Chong, S. Y.; et al. Porous Organic Cages. *Nat. Mater.* **2009**, *8*, 973.

(8) Côté, A. P.; Benin, A. I.; Ockwig, N. W.; O'Keeffe, M.; Matzger, A. J.; Yaghi, O. M. Porous, Crystalline, Covalent Organic Frameworks. *Science.* **2005**, *310*, 1166 LP – 1170.

(9) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science.* **2013**, *341*, 1230444.

(10) 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.

(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) Bureekaew, S.; Schmid, R. Hypothetical 3D-Periodic Covalent Organic Frameworks: Exploring the Possibilities by a First Principles Derived Force Field. *CrystEngComm* **2013**, *15*, 1551–1562.

(13) Turcani, L.; Greenaway, R. L.; Jelfs, K. E. Machine Learning for Organic Cage Property Prediction. *Chem. Mater.* **2019**, *31*, 714–727.

(14) Earl, D. J.; Deem, M. W. Toward a Database of Hypothetical Zeolite Structures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449–5454.

(15) Boyd, P. G.; Lee, Y.; Smit, B. Computational Development of the Nanoporous Materials Genome. *Nat. Rev. Mater.* **2017**, *2*, 17037.

(16) 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. The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* **2015**, *8*, 1190–1199.

(17) 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. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29*, 2844–2854.

(18) Franz, D.; Forrest, K. A.; Pham, T.; Space, B. Accurate H2 Sorption Modeling in the Rht-MOF NOTT-112 Using Explicit Polarization. *Cryst. Growth Des.* **2016**, *16*, 6024–6032.

(19) Campbell, C.; Gomes, J. R. B.; Fischer, M.; Jorge, M. New Model for Predicting Adsorption of Polar Molecules in Metal–Organic Frameworks with Unsaturated Metal Sites. *J. Phys. Chem. Lett.* **2018**, *9*, 3544–3553.

(20) Lin, L.-C.; Lee, K.; Gagliardi, L.; Neaton, J. B.; Smit, B. Force-Field Development from Electronic Structure Calculations with Periodic Boundary Conditions: Applications to Gaseous Adsorption and Transport in Metal–Organic Frameworks. *J. Chem. Theory Comput.* **2014**, *10*, 1477–1488.

(21) Moghadam, P. Z.; Islamoglu, T.; Goswami, S.; Exley, J.; Fantham, M.; Kaminski, C. F.; Snurr, R. Q.; Farha, O. K.; Fairen-Jimenez, D. Computer-Aided Discovery of a Metal–Organic Framework with Superior Oxygen Uptake. *Nat. Commun.* **2018**, *9*, 1378.

(22) Gomez-Gualdron, D. A.; Gutov, O. V; Krungleviciute, V.; Borah, B.; Mondloch, J. E.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Snurr, R. Q. Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chem. Mater.* **2014**, *26*, 5632–5639.

(23) 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. In Silico Discovery of Metal-Organic Frameworks for Precombustion CO2 Using a Genetic Algorithm. *Sci. Adv.* **2016**, *2*, e1600909.

(24) 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.* **2011**, *4*, 83.

(25) Banerjee, D.; Simon, C. M.; Plonka, A. M.; Motkuri, R. K.; Liu, J.; Chen, X.; Smit, B.; Parise, J. B.; Haranczyk, M.; Thallapally, P. K. Metal–Organic Framework with Optimally Selective Xenon Adsorption and Separation. *Nat. Commun.* **2016**, *7*, ncomms11831.

(26) Keupp, J.; Schmid, R. TopoFF: MOF Structure Prediction Using Specifically Optimized Blueprints. *Faraday Discuss.* **2018**, *211*, 79–101.

(27) 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.

(28) Li, S.; Chung, Y. G.; Snurr, R. Q. High-Throughput Screening of Metal–Organic Frameworks for CO2 Capture in the Presence of Water. *Langmuir* **2016**, *32*, 10368–10376.

(29) Chung, Y. G.; Bai, P.; Haranczyk, M.; Leperi, K. T.; Li, P.; Zhang, H.; Wang, T. C.; Duerinck, T.; You, F.; Hupp, J. T.; et al. Computational Screening of Nanoporous Materials for Hexane and Heptane Isomer Separation. *Chem. Mater.* **2017**, *29*, 6315–6328.

(30) Bai, P.; Jeon, M. Y.; Ren, L.; Knight, C.; Deem, M. W.; Tsapatsis, M.; Siepmann, J. I. Discovery of Optimal Zeolites for Challenging Separations and Chemical Transformations Using Predictive Materials Modeling. *Nat. Commun.* **2015**, *6*, 5912.

(31) Goldsmith, J.; Wong-Foy, A. G.; Cafarella, M. J.; Siegel, D. J. Theoretical Limits of Hydrogen Storage in Metal–Organic Frameworks: Opportunities and Trade-Offs. *Chem. Mater.* **2013**, *25*, 3373–3382.

(32) Gómez-Gualdrón, D. A.; Wilmer, C. E.; Farha, O. K.; Hupp, J. T.; Snurr, R. Q. Exploring the Limits of Methane Storage and Delivery in Nanoporous Materials. *J. Phys. Chem. C* **2014**, *118*, 6941–6951.

(33) 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.

(34) 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.

(35) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What Are the Best Materials To Separate a Xenon/Krypton Mixture? *Chem. Mater.* **2015**, *27*, 4459–4475.

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

(37) 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.

(38) Borboudakis, G.; Stergiannakos, T.; Frysali, M.; Klontzas, E.; Tsamardinos, I.; Froudakis, G. E. Chemically Intuited, Large-Scale Screening of MOFs by Machine Learning Techniques. *npj Comput. Mater.* **2017**, *3*, 40.

(39) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30*, 6325–6337.

(40) Tang, D.; Wu, Y.; Verploegh, R. J.; Sholl, D. S. Efficiently Exploring Adsorption Space to Identify Privileged Adsorbents for Chemical Separations of a Diverse Set of Molecules. *ChemSusChem* **2018**, *11*, 1567–1575.

(41) Scott, A. Round two for MOF commercialization https://cen.acs.org/articles/95/i24/Round-two-MOF-commercialization.html (accessed May 31, 2019).

(42) Rieth, A. J.; Wright, A. M.; Rao, S.; Kim, H.; LaPotin, A. D.; Wang, E. N.; Dincă, M. Tunable Metal–Organic Frameworks Enable High-Efficiency Cascaded Adsorption Heat Pumps. *J. Am. Chem. Soc.* **2018**, *140*, 17591–17596.

(43) DeCoste, J. B.; Weston, M. H.; Fuller, P. E.; Tovar, T. M.; Peterson, G. W.; LeVan, M. D.; Farha, O. K. Metal–Organic Frameworks for Oxygen Storage. *Angew. Chemie Int. Ed.* **2014**, *53*, 14092–14095.

(44) Matito-Martos, I.; Moghadam, P. Z.; Li, A.; Colombo, V.; Navarro, J. A. R.; Calero, S.; Fairen-Jimenez, D. Discovery of an Optimal Porous Crystalline Material for the Capture of Chemical Warfare Agents. *Chem. Mater.* **2018**, *30*, 4571–4579.

(45) Mon, M.; Bruno, R.; Ferrando-Soria, J.; Armentano, D.; Pardo, E. Metal–Organic Framework Technologies for Water Remediation: Towards a Sustainable Ecosystem. *J. Mater. Chem. A* **2018**, *6*, 4912–4947.

(46) Barnett, B. R.; Gonzalez, M. I.; Long, J. R. Recent Progress Towards Light Hydrocarbon Separations Using Metal–Organic Frameworks. *Trends Chem.* **2019**, *1*, 159–171.

(47) Evans, A.; Luebke, R.; Petit, C. The Use of Metal–Organic Frameworks for CO Purification. *J. Mater. Chem. A* **2018**, *6*, 10570–10594.

(48) Marshall, A. G.; Rodgers, R. P. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.* **2004**, *37*, 53–59.

(49) 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.

(50) Peng, Y.; Krungleviciute, V.; Eryazici, I.; Hupp, J. T.; Farha, O. K.; Yildirim, T. Methane Storage in Metal–Organic Frameworks: Current Records, Surprise Findings, and Challenges. *J. Am. Chem. Soc.* **2013**, *135*, 11887–11894.

(51) Ockwig, N. W.; Nenoff, T. M. Membranes for Hydrogen Separation. *Chem. Rev.* **2007**, *107*, 4078–4110.

(52) Ockwig, N. W.; Nenoff, T. M. Membranes for Hydrogen Separation. *Chem. Rev.* **2010**, *110*, 2573–2574.

(53) Grande, C. A.; Cavenati, S.; Da Silva, F. A.; Rodrigues, A. E. Carbon Molecular Sieves for Hydrocarbon Separations by Adsorption. *Ind. Eng. Chem. Res.* **2005**, *44*, 7218–7227.

(54) Mckee, D. W. United States Patent Office, 2009.

(55) Cheng, H. C.; Hill, F. B. Separation of Helium-Methane Mixtures by Pressure Swing Adsorption. *AIChE J.* **1985**, *31*, 95–102.

(56) Plimpton, S. Fast Parallel Algorithms for Short-Range

Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(57) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(58) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbsch, P. Structural Relaxation Made Simple. *Phys. Rev. Lett.* **2006**, *97*, 170201.

(59) 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.

(60) 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. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**.

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

(62) García-Pérez, E.; Parra, J. B.; Ania, C. O.; Dubbeldam, D.; Vlugt, T. J. H.; Castillo, J. M.; Merkling, P. J.; Calero, S. Unraveling the Argon Adsorption Processes in MFI-Type Zeolite. *J. Phys. Chem. C* **2008**, *112*, 9976–9979.

(63) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. Thermodynamic Analysis of Xe/Kr Selectivity in over 137 000 Hypothetical Metal–Organic Frameworks. *Chem. Sci.* **2012**, *3*, 2217–2223.

(64) Potoff, J. J.; Siepmann, J. I. Vapor–Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE J.* **2001**, *47*, 1676–1682.

(65) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569–2577.

(66) Levesque, D.; Gicquel, A.; Darkrim, F. L.; Kayiran, S. B. Monte Carlo Simulations of Hydrogen Storage in Carbon Nanotubes. *J. Phys. Condens. Matter* **2002**, *14*, 9285–9293.

(67) Darkrim, F.; Levesque, D. Monte Carlo Simulations of Hydrogen Adsorption in Single-Walled Carbon Nanotubes. *J. Chem. Phys.* **1998**, *109*, 4981–4984.

(68) Widom, B. Some Topics in the Theory of Fluids. *J. Chem. Phys.* **1963**, *39*, 2808–2812.

(69) Bae, Y.-S.; Yazaydın, A. Ö.; Snurr, R. Q. Evaluation of the BET Method for Determining Surface Areas of MOFs and Zeolites That Contain Ultra-Micropores. *Langmuir* **2010**, *26*, 5475–5483.

(70) 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.

(71) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149*, 134–141.

(72) Rappé, 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.

(73) Argueta, E.; Shaji, J.; Gopalan, A.; Liao, P.; Snurr, R. Q.; Gómez-Gualdrón, D. A. Molecular Building Block-Based Electronic Charges for High-Throughput Screening of Metal–Organic Frameworks for Adsorption Applications. *J. Chem. Theory Comput.* **2018**, *14*, 365–376.

(74) François Chollet. Keras. *GitHub Repos.* **2015**.

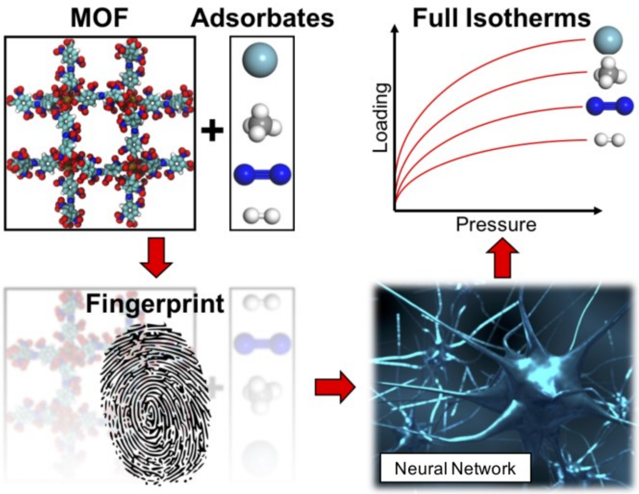
(75) Klikauer, T. Scikit-Learn: Machine Learning in Python. *TripleC* **2016**, *14*, 260–264.

(76) Dokur, D.; Keskin, S. Effects of Force Field Selection on the Computational Ranking of MOFs for CO2 Separations. *Ind. Eng. Chem. Res.* **2018**, *57*, 2298–2309.

(77) Man, W.; Donev, A.; Stillinger, F. H.; Sullivan, M. T.; Russel, W. B.; Heeger, D.; Inati, S.; Torquato, S.; Chaikin, P. M. Experiments on Random Packings of Ellipsoids. *Phys. Rev. Lett.* **2005**, *94*, 1–4.

(78) Donev, A.; Stillinger, F. H.; Chaikin, P. M.; Torquato, S. Unusually Dense Crystal Packings of Ellipsoids. *Phys. Rev. Lett.* **2004**, *92*, 1–4.

TOC


MOF
Adsorbates
+
Full Isotherms
Loading
Pressure
Fingerprint
Neural Network