doi.org/10.26434/chemrxiv.9894224.v1

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

Ryther Anderson, Achay Biong, Diego Gómez-Gualdrón

Submitted date: 23/09/2019 • Posted date: 27/09/2019

Citation information: Anderson, Ryther; Biong, Achay; Gómez-Gualdrón, Diego (2019): Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. ChemRxiv. Preprint.

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, such as grand canonical Monte Carlo (GCMC), 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 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 neural network only on alchemical species, represented only by their geometry and force field parameters, and used this neural network to predict the loadings of real adsorbates. We focused on predicting room temperature adsorption of small (one- and two-atom) molecules relevant to chemical separations. Namely, argon, krypton, xenon, methane, ethane, and nitrogen. However, we also observed surprisingly promising predictions for more complex molecules, whose properties are outside the range spanned by the alchemical adsorbates. Prediction accuracies suitable for large-scale screening were achieved using simple MOF (e.g. geometric properties and chemical moieties), and adsorbate (e.g. forcefield parameters and geometry) descriptors. Our results illustrate a new philosophy of training that opens the path towards development of machine learning models that can predict the adsorption loading of any new adsorbate at any new operating conditions in any new MOF.

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, such as grand canonical Monte Carlo (GCMC), 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 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 neural network only on alchemical species, represented only by their geometry and force field parameters, and used this neural network to predict the loadings of real adsorbates. We focused on predicting room temperature adsorption of small (one- and two-atom) molecules relevant to chemical separations. Namely, argon, krypton, xenon, methane, ethane, and nitrogen. However, we also observed surprisingly promising predictions for more complex molecules, whose properties are outside the range spanned by the alchemical adsorbates. Prediction accuracies suitable for large-scale screening were achieved using simple MOF (e.g. geometric properties and chemical moieties), and adsorbate (e.g. forcefield parameters and geometry) descriptors. Our results illustrate a new philosophy of training that opens the path towards development of machine learning models that can predict the adsorption loading of any new adsorbate at any new operating conditions in any new MOF.

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 and operating conditions 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 high-throughput screening methods.

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. 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

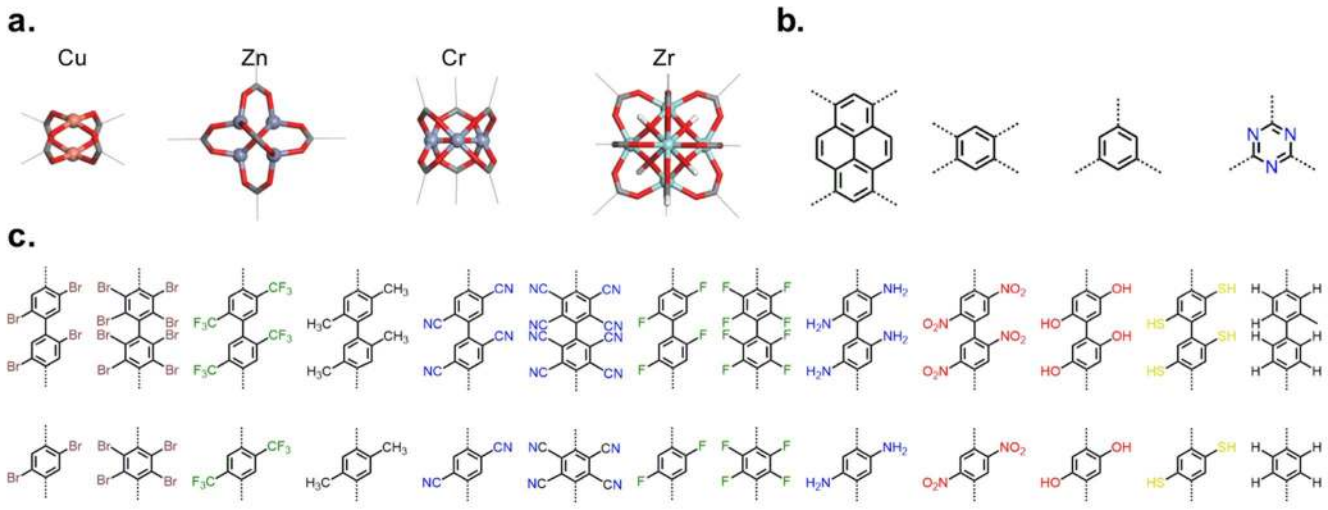

Figure 1. The building blocks used for MOF database construction, dashed lines indicate connections to the rest of the framework; **a)** inorganic (metal-containing) nodes, which include Cu (4-connected), Zn (6-connected), Cr (6-connected), and Zr (8- and 12-connected) oxoclusters, **b**) organic nodes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

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 requiring the fingerprint to include 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 boosted 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, *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*n*H*m*O*x*N*y*S*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 and 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 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 toward universal prediction of adsorption isotherms. Here "universal" should be understood as the ability of such

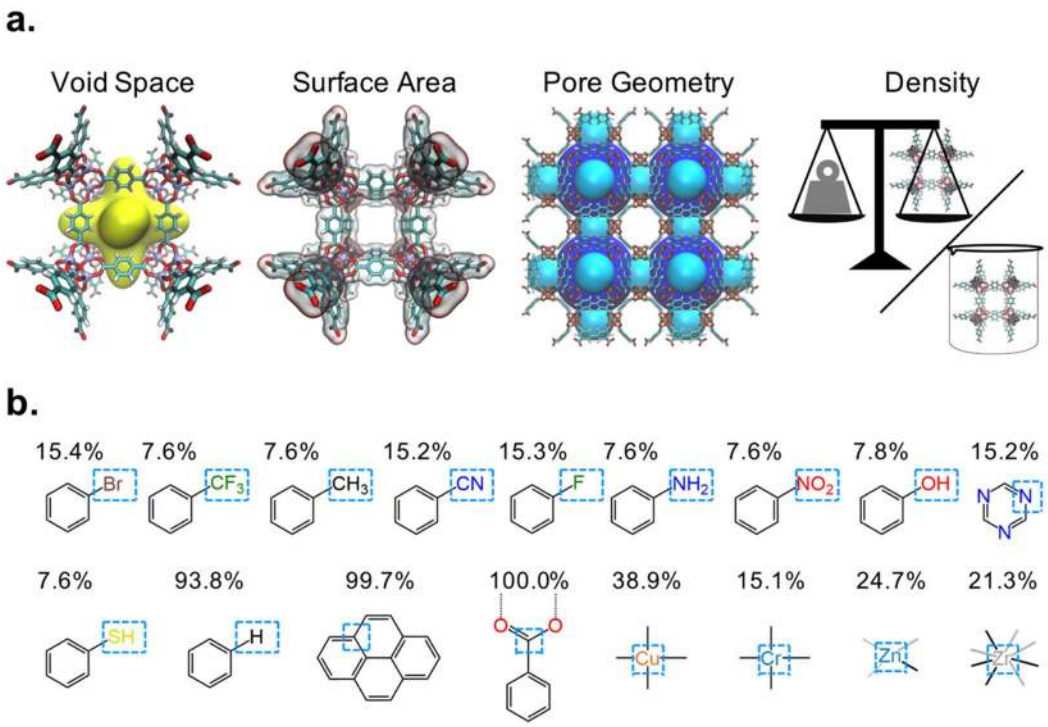

Figure 2. Descriptors constituting the MOF fingerprint. **a)** Five textural properties: void fraction, gravimetric surface area, largest pore diameter (LPD: dark blue sphere), pore limiting diameter (PLD: light blue sphere), and pore size standard deviation (PSSD), and density. **b)** 17 chemical motifs (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.

tools *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) along with different operating conditions. 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 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] 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**. The building blocks were chosen to create a database that provides enough chemical diversity to ensure we explore a variety of interactions with the adsorbates studied in this work, but also with potential adsorbates for future work that could expand on the type of molecules and operating conditions considered for investigation. Atomic charges were assigned to each MOF according to our MBBB approach,[56] which will allow these MOFs to be used in the future for adsorption predictions of more complex adsorbates (for which the consideration of adsorbate-framework electrostatic interactions is required). Each MOF prototype constructed by ToBaCCo was structurally optimized in LAMMPS[57] using the Dreiding[58] force field to describe framework intramolecular forces. 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[59] with a timestep of 10.0 fs with the unit 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. To generate our training data we ran GCMC simulations for 200 one-, two-, and three-site 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

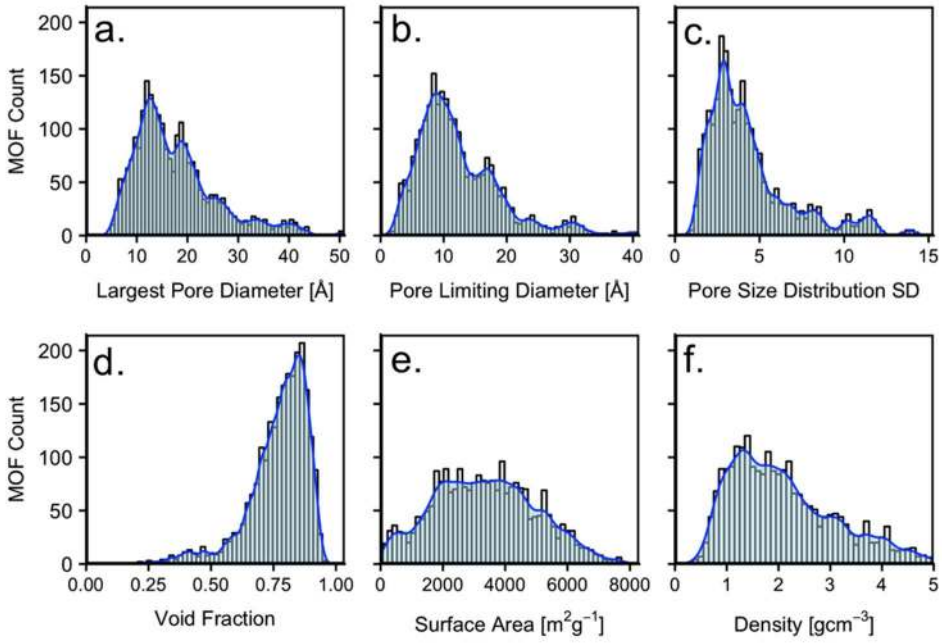

Figure 3. Histograms of the textural properties used as descriptors in MOF fingerprint.

alchemical species). To generate our validation data, we ran GCMC simulations for 200 one-, two-, and three-site 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.

LJ parameters for helium correspond to those used by Smit and coworkers,[68] LJ parameters for argon were taken from Perez and coworkers,[69] and LJ parameters for krypton, and xenon correspond to those used by Sikora and coworkers.[70] The parameters for these adsorbates is summarized in **Table S5**. Methane, ethane, propane, butane, isobutane, benzene, and nitrogen, were modeled according to the TraPPE force-field developed by Siepmann and coworkers.[71,72] Accordingly, methane, ethane, propane, 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 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.[71,72] LJ parameters and charges for hydrogen were taken from the Darkim-Levesque model,[73,74] 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.

2.3 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, here we focused on both one-site and two/three-site 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). 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 parameters that control dispersion and electrostatic interactions. Ultimately, we used the effective LJ parameters ($\epsilon_{\text{effective}}$ and $\sigma_{\text{effective}}$) for each adsorbate along with the maximum charge magnitude (zero for one atom adsorbates) and the bond length (also 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/three site adsorbates, $\epsilon_{\text{effective}}$ was the sum of the ϵ_{ii} of the different sites, and $\sigma_{\text{effective}}$ was:

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

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 the test set, $\epsilon_{\text{effective}}$ 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 (essentially an extension of equation 1 to adsorbates with more than one bond). Similarly, bond length was taken to be the longest distance between atom coordinates in the lowest energy geometry (as calculated according to the relevant force field, see below). Four other adsorbate fingerprints were considered, these are discussed in detail in **Section S2**.

2.4. MOF fingerprinting. We tested seven different MOF fingerprints (see results in **Table S6**), 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), framework density (ρ_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 chemical diversity of the studied MOFs is apparent from the frequency that each moiety appears in the database (the percentages shown in **Fig. 2**) and the structural diversity of the studied MOF is apparent from the histograms of the textural properties observed in our database (shown 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,[60] while GSA was calculated by rolling a nitrogen-sized spherical probe along the framework surface.[61] Both of these calculations were done in RASPA-2.0.[62] LPD and PLD were calculated using zeo++.[63] 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. The number density of the various chemical moieties was calculated using an in-house pattern recognition code.

2.5. Adsorption Simulations. RASPA-2.0[62] was used to perform all GCMC simulations, in which 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). 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. Adsorbate-adsorbate interactions were modeled using Lennard-Jones (LJ) potentials to describe dispersion interactions and Coulomb's law to describe charge-charge interactions. To be consistent with previous work considering adsorption of charged two-atom species[61,64–66] we do not consider adsorbate-framework electrostatic interactions. Available Dreiding forcefield[58] parameters were assigned to framework atoms. Otherwise UFF[67] parameters were used. Lorentz-Berthelot mixing rules were used to calculate parameters for interactions between atoms not explicitly parametrized. Adsorbate force-field parameters were assigned as discussed on Section 2.2.

3. NEURAL NETWORK TRAINING

3.1. Model training. Here we trained a multilayer perceptron (MLP, see **Fig. S1**) to predict the adsorption data obtained from GCMC simulations. All MLPs were trained using Keras[75] through the SciKit-learn[76] Python module. First, before training the final MLP model, we investigated different network hyperparameter configurations in order to determine which configuration(s) 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 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 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 parameters on a grid, keeping topology constant. The parameters considered and their final values (**Table S7**) are presented in **Section S2**.

The final model resulting from the above tuning procedure was then tested on the real adsorbates in the test set (see below). Additional data, demonstrating the reproducibility of our model, is presented in **Figure S2**. 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. 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 (i.e. if validation error did not improve for 20 epochs in a row, training was terminated and the lowest error model from the previous epochs was taken to be the model error).

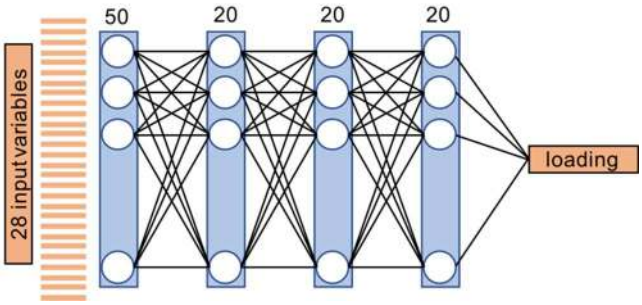

Figure 4. The configuration of our final model. The number of nodes in each hidden layer are shown above the corresponding layer.

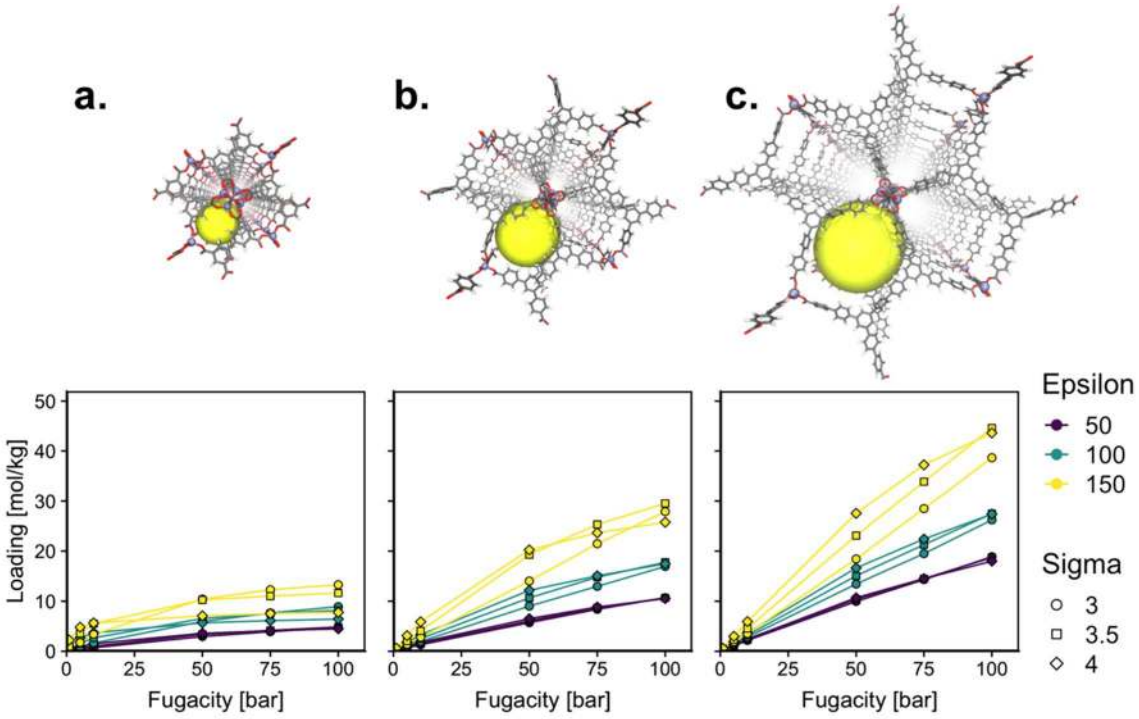

Figure 5. Isotherms for alchemical adsorbates considered in **a.** a representative small pore MOF (LPD=6.5 Å), **b.** a representative intermediate pore MOF (LPD=13.7 Å), and **c.** a representative large pore MOF (LPD=19.5 Å). All the MOFs shown are of the **mcn** topology. The large pore is illustrated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

4. RESULTS AND DISCUSSION

4.1. Model predictive ability for simple adsorbates. Before discussing the overall prediction performance of our final MLP, we discuss briefly, from an intuitive perspective, what the model must learn. **Fig. 5** shows adsorption isotherms for a subset of the training set alchemical adsorbates. 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. From the isotherms in **Fig. 5** one can see that the model must learn that intrinsic adsorbate-adsorbent interactions play a dominant role in controlling adsorbate loadings at low fugacity, but that surface area and void fraction start to play a dominant role at moderate and high fugacities, respectively. For instance, consider how the loading of the alchemical adsorbate with the largest ϵ_{ii} and $\sigma_{ii}$is restricted in the small pore MOF, resulting in lower loadings than adsorbates with significantly smaller ϵ_{ii} values. In the large pore MOF, this effect is reversed, with this adsorbate achieving nearly the highest loading out of all the adsorbates.

Next, we consider our final model predictive performance 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 (methane, ethane, and nitrogen are modeled with one, two, and three sites, respectively) 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), Pearson correlation (R), and Spearman correlation (S). Perfect predictions would have zero MAPE and MAE. Values of R close to one indicate a very strong linear correlation between GCMC loadings and those predicted by the MLP. Values of S close to one indicate that the MLP predictions increase nearly monotonically with the GCMC simulated values. As a point of comparison, our final model predicted the validation set loadings (200 alchemical adsorbates in 200 MOFs at 5 fugacities) with a MAPE of 3.2 % and a MAE of 0.19 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 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.

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

Adsorbate	MAPE [%]	MAE [mol//kg]	R	S
Argon	4.8	0.17	0.999	0.999
Methane	5.1	0.28	0.999	0.999
Krypton	5.4	0.39	0.999	0.999
Xenon	5.1	0.44	0.999	0.999
Ethane	4.2	0.37	0.999	0.999
Nitrogen	4.8	0.16	0.999	0.999

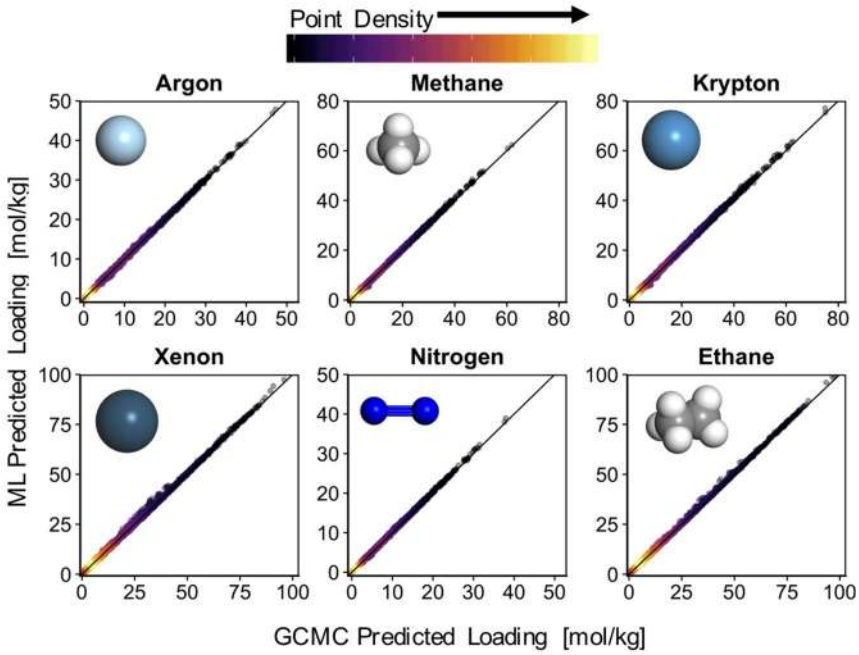

Figure 6. Parity plots comparing the predictions of the final MLP model for the six indicated adsorbates versus GCMC-calculated values. The listed real adsorbates possess properties within the ranges covered by the alchemical adsorbates used during model training. Points color indicate the point density in the plot (the highest density is observed at low loadings).

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 model performs extremely well on the six adsorbates considered here 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 ability 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 adsorbate cases are given in **Fig. S3**) in test set MOFs. The isotherms predicted by the MLP were constructed from loadings predicted at 100 different fugacities between 1 and 100 bar (notably, it took only minutes to obtain these 100-point isotherms for 400 MOFs). 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 (size 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 median 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 median 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 the *best* predicted isotherms according to our IMPE metric, Q5 isotherms are predicted the *worst*, and Q3 isotherms are average predictions. **Fig. 7** and **Fig. S3** present one isotherm from each quantile (the one nearest to the IMPE quantile value).

Q1 isotherms are essentially 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 acceptable. 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 1.74 % (for ethane) to 3.18% (for krypton). As a point of comparison, Dokur and Keskin [77] 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 effect 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.

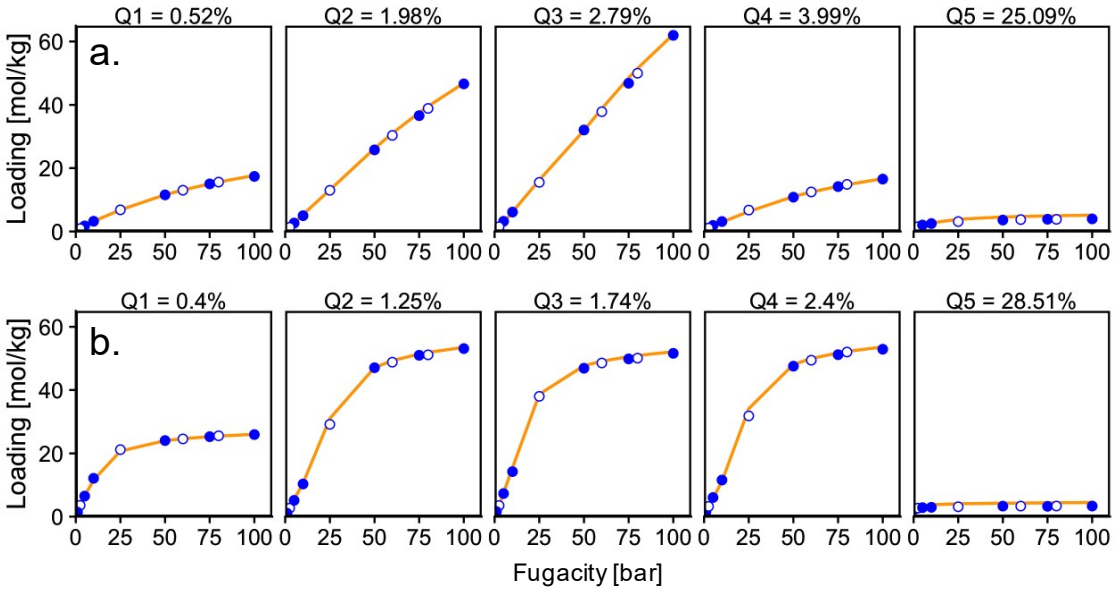

Figure 7. Isotherms for **a.** methane and **b.** ethane, for the Q1 (0.00), Q2 (0.25), Q3 (0.50), Q4 (0.75), and Q5 (1.00) quantiles of isotherm median percentage error (IMPE, with corresponding values shown). Points are GCMC simulated values (filled correspond to fugacities included in training, empty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

4.3. Material ranking accuracy based on performance metrics. For machine learning to effectively be 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, as the final endeavor of this work, we assessed the ability of our model to identify top-performing materials.

Table 2. The number of MOFs in the top 100 from GCMC data that were encountered in the top 100 predicted using data predicted by machine learning.

Adsorbate	# Correctly Placed	
	Loading@100 bar [mol/kg]	Working Capacity [mol/kg]
Argon	99	100
Methane	99	98
Krypton	98	97
Xenon	100	99
Ethane	98	100
Nitrogen	99	99

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. Specifically, we focused on the ability of our model to identify MOFs in the "top 100" of the 400 MOF test set according to their loadings at 100 bar and their working capacity for a 100 bar ↔ 5 bar fugacity swing.

Table 2 summarizes the ability of the MLP to rank MOFs according to their loading and working capacity. For loadings at 100 bar, the top 100 MOFs predicted using machine learning contains between 98 (Kr case) and 100 (Xe case) of the MOF found in the top 100 constructed using GCMC data. In the case of working capacities case, the top 100 constructed from machine learning predictions contains between 97 (Kr case) and 100 (Ar and ethane cases) of the MOFs found in the top 100 constructed with GCMC data. Given that ranking materials in this manner is one of the most important uses of a machine learning algorithm used to predict adsorption loading, our model could be useful even for relatively complex adsorbates (see below and **Table S10**) as a first step for screening MOF databases for adsorption performance of diverse chemicals.

4.4. Testing the limits of the current MLP model (extrapolation). It is well-known that machine learning models are meant to work within the confines of property values delimited by the training data. To assess to what extent the trained MLP "breaks" for adsorbates whose properties are outside the ranges considered in the training, we decided to compare MLP predictions to GCMC data for adsorbates with properties outside those considered during training (including more complex adsorbates). Note that this endeavor was partly done as a preliminary test to inform future studies on more complex adsorption scenarios. He, (ϵ_{ii} and σ_{ii} smaller than for any alchemical species) H_2 (a three-site model, that contrary to any three-site alchemical adsorbate only has LJ parameters *at* the central site), 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) were chosen for this test.

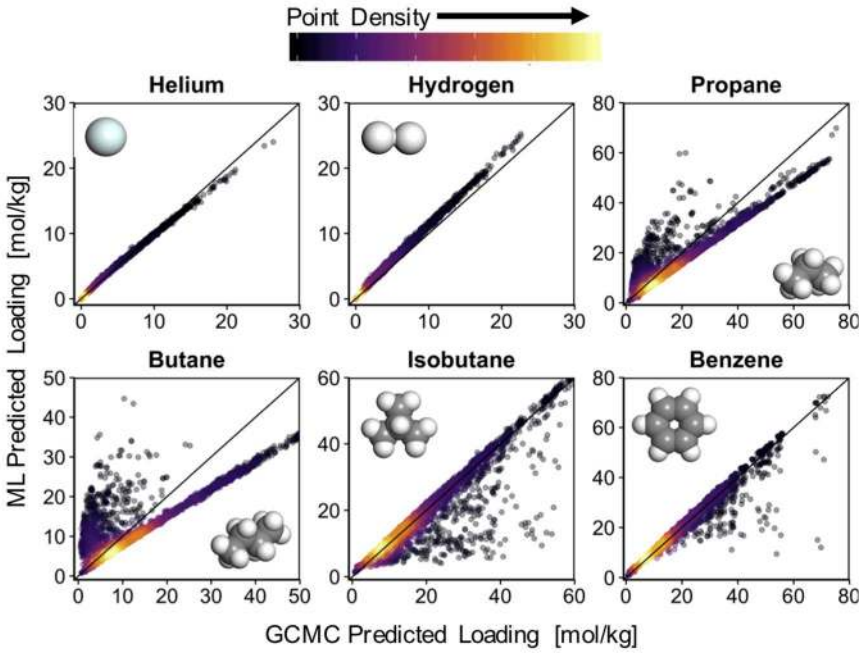

Figure 8. Parity plots comparing the predictions of the final MLP model for the six indicated adsorbates versus GCMC-calculated values. The listed real adsorbates possess properties outside the ranges covered by the alchemical adsorbates used during model training. Points color indicate the point density in the plot (the highest density is observed at low loadings).

The parity plots in **Fig. 8** provide a visual comparison between MLP predictions to GCMC data for the above species, with the accuracy metrics provided in **Table S8** (isotherm comparisons analogous to **Fig. 7** are presented in **Fig. S4)**. The MLP predictions for all these species presented a high correlation with the corresponding GCMC data (the lowest R was 0.877 for n-butane). The best predictions (MAE < 0.6 kg/mol) were for helium and hydrogen loading, for which the MLP slightly, but systematically, under- and overestimated, respectively, the GCMC-calculated loading. The next best MLP predictions (MAE < 2.1 kg/mol) were for isobutane and benzene, for which the bulk of the points fall near the parity line, but for which the presence of outliers is apparent. The least accurate MLP predictions (MAE < 4.1 kg/mol) were for propane and n-butane), for which the bulk of the points follow a linear, but systematically underestimated relationship with GCMC data, and also present outliers. Note, however, that the number of outliers for the last four adsorbates is not that high when considering the point density (indicated by point color in **Fig. 8**).

The majority of outliers for adsorption predictions for propane and n-butane occur at the lowest fugacities (1.0 and 2.5 bar) in MOFs with higher than average void fraction (> 0.77) and LPD (> 17.7). Specifically, the average V_F (LPD) of MOFs in the top-50 outlying points, assessed using MAPE for propane and butane loading predictions are 0.87 (19.4) and 0.90 (23.7), respectively. In contrast, most of the outliers for adsorption predictions for isobutane and benzene are at intermediate fugacities (between 2.5 and 10 bar) in MOFs with relatively small void fraction (LPD). The averages of this property in outliers for isobutane and benzene are 0.68 (16.2) and 0.64 (16.6), respectively.

In light of the statistics above, one can speculate on the reasons for prediction inaccuracies in these "extrapolated" adsorbates. The systematic underpredictions for the bulk of the propane and n-butane may be related to their flexible character compared to the rigid alchemical adsorbates used in the training data, as well as a higher aspect ratio. This is consistent with the absence of systematic overestimation for isobutane and benzene adsorption predictions, which present a low aspect ratio and little flexibility.

The outliers for isobutane and benzene occur because the MLP model has trouble capturing the isotherm shape for these molecules around the fugacities near the saturation fugacity (which is rather low for these large molecules). Thus, the inaccuracy could be due to the model being blind to packing effects in small pores. The outliers for propane and n-butane occur because the MLP model overestimates adsorption of these molecules in large pore MOFs (where the lower adsorption enthalpies have a harder time compensating the loss of entropy during adsorption). Thus, the inaccuracy could be due to the model being blind to entropic losses that are important in flexible molecules. We reiterate, however, that since extrapolation behavior is difficult to predict, the above analysis is solely based on chemical intuition, and, while plausible, cannot be taken as absolutely certain. This analysis can, on the other hand, guide the selection of descriptors in future, more generalizable deep learning models.

Despite the presence of outliers, one would expect that the ability of the MLP to rank MOFs according to their loading and/or working capacity for these "extrapolated" adsorbates is still quite good due to highly linear relationships between MLP predictions and GCMC data. This is confirmed according to the data in **Table S9** (which is analogous to **Table 2)**, most notably for loadings at 100 bar. Indeed, for this property, at least 98 out of the 100 top MOFs are correctly identified by the MLP model. For working capacity on the other hand, discounting benzene, at least 89 out of the top 100 MOFs are identified. The least

accurate ranking ability is for benzene, but even in that case the MLP model captures about half (46) of the MOFs in the top-100. The anomalous MLP performance in ranking benzene is due to inaccuracy in the isotherm region around the saturation fugacity (see above and **Fig. S4f**). Nonetheless, the MLP accuracy to both predict adsorption of He, H_2, propane, n-butane, isobutane and benzene in MOFs and rank the MOFs based on adsorption of these species is higher than we were expecting considering that their properties are (in some cases far) outside the ranges covered by the training data.

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 different adsorbates at different pressures. Key to accomplishing these prediction capabilities was the inclusion of thermodynamic conditions (here fugacity), adsorbate force field parameters as model inputs, but above all the inclusion alchemical adsorbates as part of the training data. 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, similar to the C_3 and above alkanes, 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.

Three comma-separated values (CSV) files containing the training, validation, and test set 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* **10,** 582–592 (2018).

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* **120,** 23044–23054 (2016).

3. Vermoortele, F. *et al.* p-Xylene-Selective Metal–Organic Frameworks: A Case of Topology-Directed Selectivity. *J. Am. Chem. Soc.* **133,** 18526–18529 (2011).

4. Demir, H. *et al.* Metal–Organic Frameworks with Metal–Catecholates for O2/N2 Separation. *J. Phys. Chem. C* **123,** 12935–12946 (2019).

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

6. Baerlocher, C. & McCusker, L. B. Database of Zeolite Structures.

7. Tozawa, T. *et al.* Porous organic cages. *Nat. Mater.* **8,** 973 (2009).

8. Côté, A. P. *et al.* Porous, Crystalline, Covalent Organic Frameworks. *Science.* **310,** 1166 LP – 1170 (2005).

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

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.* **17,** 5801–5810 (2017).

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

12. Bureekaew, S. & Schmid, R. Hypothetical 3D-periodic covalent organic frameworks: exploring the possibilities by a first principles derived force field. *CrystEngComm* **15,** 1551–1562 (2013).

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

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

15. Boyd, P. G., Lee, Y. & Smit, B. Computational development of the nanoporous materials genome. *Nat. Rev. Mater.* **2,** 17037 (2017).

16. Simon, C. M. *et al.* The materials genome in action: identifying the performance limits for methane storage. *Energy Environ. Sci.* **8,** 1190–1199 (2015).

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

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.* **16,** 6024–6032 (2016).

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.* **9,** 3544–3553 (2018).

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.* **10,** 1477–1488 (2014).

21. Moghadam, P. Z. *et al.* Computer-aided discovery of a metal–organic framework with superior oxygen uptake. *Nat. Commun.* **9,** 1378 (2018).

22. Gomez-Gualdron, D. A. *et al.* Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chem. Mater.* **26,** 5632–5639 (2014).

23. Chung, Y. G. *et al.* In silico discovery of metal-organic frameworks for precombustion CO2 using a genetic algorithm. *Sci. Adv.* **2,** e1600909 (2016).

24. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4,** 83 (2011).

25. Banerjee, D. *et al.* Metal–organic framework with optimally selective xenon adsorption and separation. *Nat. Commun.* **7,** ncomms11831 (2016).

26. Keupp, J. & Schmid, R. TopoFF: MOF structure prediction using specifically optimized blueprints. *Faraday Discuss.* **211,** 79–101 (2018).

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

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* **32,** 10368–10376 (2016).

29. Chung, Y. G. *et al.* Computational Screening of Nanoporous Materials for Hexane and Heptane Isomer Separation. *Chem. Mater.* **29,** 6315–6328 (2017).

30. Bai, P. *et al.* Discovery of optimal zeolites for challenging separations and chemical transformations using predictive materials modeling. *Nat. Commun.* **6,** 5912 (2015).

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.* **25,** 3373–3382 (2013).

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* **118,** 6941–6951 (2014).

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* **117,** 7681–7689 (2013).

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* **117,** 14095–14105 (2013).

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.* **27,** 4459–4475 (2015).

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.* **5,** 3056–3060 (2014).

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.* **19,** 640–645 (2017).

38. Borboudakis, G. *et al.* Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Comput. Mater.* **3,** 40 (2017).

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.* **30,** 6325–6337 (2018).

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* **11,** 1567–1575 (2018).

41. Scott, A. Round two for MOF commercialization. *Chemical Engineering and News* (2019). Available at: https://cen.acs.org/articles/95/i24/Round-two-MOF-commercialization.html. (Accessed: 31st May 2019)

42. Rieth, A. J. *et al.* Tunable Metal–Organic Frameworks Enable High-Efficiency Cascaded Adsorption Heat Pumps. *J. Am. Chem. Soc.* **140,** 17591–17596 (2018).

43. DeCoste, J. B. *et al.* Metal–Organic Frameworks for Oxygen Storage. *Angew. Chemie Int. Ed.* **53,** 14092–14095 (2014).

44. Matito-Martos, I. *et al.* Discovery of an Optimal Porous Crystalline Material for the Capture of Chemical Warfare Agents. *Chem. Mater.* **30,** 4571–4579 (2018).

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* **6,** 4912–4947 (2018).

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

47. Evans, A., Luebke, R. & Petit, C. The use of metal–organic frameworks for CO purification. *J. Mater. Chem. A* **6,** 10570–10594 (2018).

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

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* **123,** 120–130 (2019).

50. Peng, Y. *et al.* Methane Storage in Metal–Organic Frameworks: Current Records, Surprise Findings, and Challenges. *J. Am. Chem. Soc.* **135,** 11887–11894 (2013).

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

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

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.* **44,** 7218–7227 (2005).

54. Mckee, D. W. United States Patent Office. *Journal of the American Society for Naval Engineers* **39,** 620–622 (2009).

55. Cheng, H. C. & Hill, F. B. Separation of helium-methane mixtures by pressure swing adsorption. *AIChE J.* **31,** 95–102 (1985).

56. Argueta, E. *et al.* Molecular Building Block-Based Electronic Charges for High-Throughput Screening of Metal–Organic Frameworks for Adsorption Applications. *J. Chem. Theory Comput.* **14,** 365–376 (2018).

57. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **117,** 1–19 (1995).

58. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94,** 8897–8909 (1990).

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

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

61. 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* **26,** 5475–5483 (2010).

62. Dubbeldam, D., Calero, S., Ellis, D. E. & Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **42,** 81–101 (2016).

63. 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.* **149,** 134–141 (2012).

64. Gómez-Gualdrón, D. A. *et al.* Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **9,** 3279–3289 (2016).

65. Walton, K. S. & Snurr, R. Q. Applicability of the BET Method for Determining Surface Areas of Microporous Metal−Organic Frameworks. *J. Am.*

Chem. Soc. **129,** 8552–8556 (2007).

66. Yu, J., Ma, Y. & Balbuena, P. B. Evaluation of the Impact of H2O, O2, and SO2 on Postcombustion CO2 Capture in Metal–Organic Frameworks. *Langmuir* **28,** 8064–8071 (2012).

67. 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.* **114,** 10024–10035 (1992).

68. Ongari, D. *et al.* Accurate Characterization of the Pore Volume in Microporous Crystalline Materials. *Langmuir* **33,** 14529–14538 (2017).

69. García-Pérez, E. *et al.* Unraveling the Argon Adsorption Processes in MFI-Type Zeolite. *J. Phys. Chem. C* **112,** 9976–9979 (2008).

70. 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.* **3,** 2217–2223 (2012).

71. Potoff, J. J. & Siepmann, J. I. Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE J.* **47,** 1676–1682 (2001).

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

73. Levesque, D., Gicquel, A., Darkrim, F. L. & Kayiran, S. B. Monte Carlo simulations of hydrogen storage in carbon nanotubes. *J. Phys. Condens. Matter* **14,** 9285–9293 (2002).

74. Darkrim, F. & Levesque, D. Monte Carlo simulations of hydrogen adsorption in single-walled carbon nanotubes. *J. Chem. Phys.* **109,** 4981–4984 (1998).

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

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

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

SUPPLEMENTARY INFORMATION

Adsorption Isotherm Predictions for Multiple Molecules in MOFs using the same Deep Learning Model

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

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

* dgomezgualdron@mines.edu

Table of contents

Section S1. Adsorbate Force Field Parameters

■ One-site Adsorbates

Table S1. Parameters for single-site alchemical adsorbates *used for training*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]
50	3.0
100	3.5
150	4.0
200	4.5
250	

Table S2. Parameters for single-site alchemical adsorbates *used for validation*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]
60	3.10
75	3.25
125	3.75
175	4.25
225	

Table S3. Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) *used for training*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]	**Charge** [e]	**Bond Length** [Å]
15	3.00	0.0	1.0
30	3.70	0.2	1.3
45	4.50	0.5	1.6
60		0.9	
95			

Table S4. Parameters for two- and three-site alchemical adsorbates (two-site adsorbates correspond to a maximum magnitude charge of zero) *used for validation*; all possible combinations were considered.

ϵ_{ii} [K]	σ_{ii} [Å]	**Charge** [e]	**Bond Length** [Å]
25	3.25	0.0	1.1
35	3.50	0.3	1.3
50	4.00	0.6	1.5
70		0.9	
80			

Table S5. Lennard-Jones parameters for real single site real adsorbates.

Adsorbate	Epsilon [K]	Sigma [Å]
Helium	10.9	2.64
Argon	124.1	3.42
Methane	148.0	3.73
Krypton	166.4	3.64
Xenon	221.0	4.10

Section S2. Details of Model Training

■ Overview of the Multilayer Perceptron Algorithm

A multilayer perceptron (MLP) is a type of feedforward artificial neural network, which consists of at least one hidden layer (for a least three layers including the input and output layer). The goal of the MLP, as any other machine learning algorithm, is to predict a response Y given a set X of variables. For instance, when the set X_i for observation i is fed into the input layer (one node for each variable in the set X_i), the network (once trained) will provide the corresponding response Y_i as output. The output is typically a real number (when doing regression) or an integer (when doing classification).

Each hidden layer consists on N nodes, which are assigned an activation function (a non-linear function) and a weight. Each node in a hidden layer receives a weighted summation of the outputs of the preceding layer, and then applies the activation function to it. The output of each node is applied different weights into the summations that are received by the nodes in the next layer. The process is done until the output layer is reached, which consists of one node with a linear (for regression) or step (for classification) activation function.

The training of an MLP consists of finding the weights assigned to the output of each node, which result in matching as closely as possible an array of known responses i given the corresponding variable sets X_i (these known responses and variables sets are referred to as the training set). Thus, during training, an MLP learns (unveils) the relationship between variable sets and their corresponding response from a (hopefully) broad and diverse dataset. The trained MLP can then be applied to new variable sets X_j for which the response is unknown.

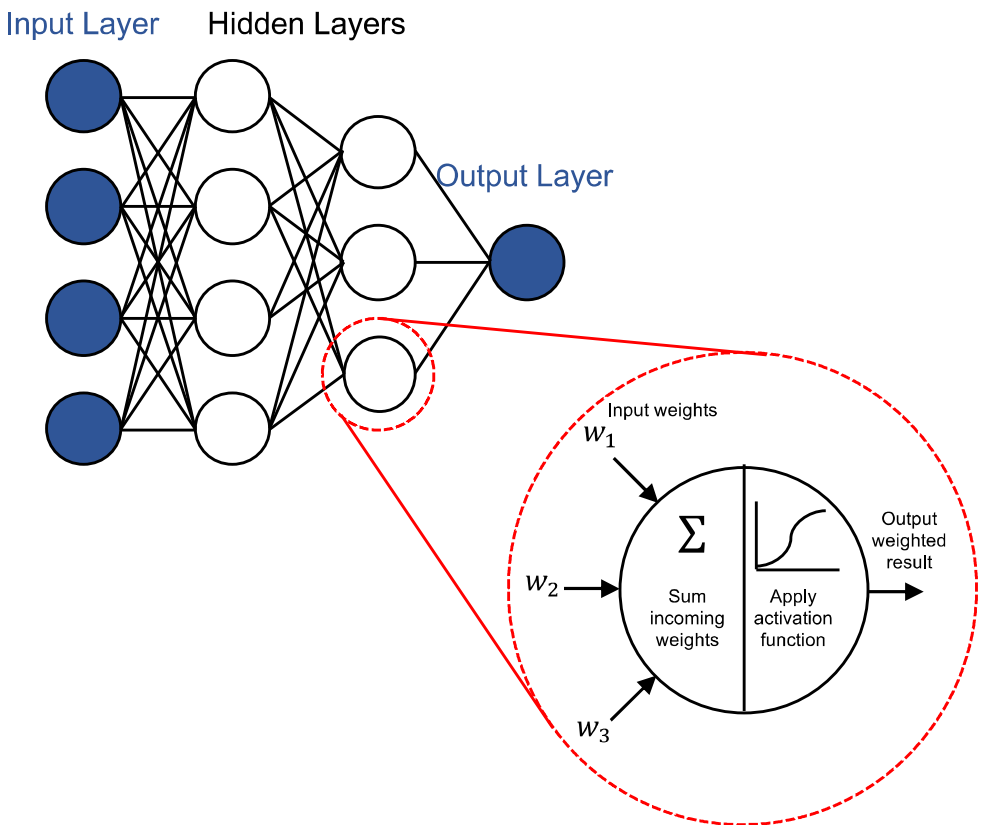

Figure S1. Schematic of a simple MLP, with two hidden layers (white nodes). The magnification shows an individual node, in which the input weights are combined (in this case as a sum), fed into the activation function (a sigmoid is shown), then output with the node weight.

The optimal weights are found using a supervised learning algorithm known as *backpropagation*.[1] The weights of each node are initialized, usually randomly or according to some probability distribution, and the prediction tested for a single input or a set of inputs. The error of this prediction is then calculated (i.e. the loss function) and used to calculate the partial derivate of model error with respect to each weight in the network.[1] The calculation of these derivatives is backpropagation. The calculated derivatives (i.e. the gradient of error) can then be used in an optimization function which minimizes the loss function.

There are many so-called "hyperparameters" which influence the final weights, and therefore the output of an MLP model. Here, we will focus our discussion on hyperparameters tuned during the training of our models. These are *i)* the architecture or *topology* of the net, which is the number of hidden layers and the number of nodes in each layer, *ii)* the activation function applied to the incoming weights for each node (e.g. sigmoid, ReLU, or linear), *iii)* how the initial weights are assigned, i.e. the *initialization* procedure, *iv)* the *optimizer* used to minimize the loss function, *v)* the number of times (*epochs*) all of the training data is passed through the backpropagation algorithm, and *vi) batch size* which is the number of observations passed through the net before updating the weights (between one and the total number of observations). **Table S6** shows the final hyperparameter set for each of our MLPs (the four base models and the top model).

■ Additional Training Results

Before training any neural networks, we tested 28 different descriptor sets using an XGBoost model.[2] These included four adsorbate descriptors together with nine MOF descriptors.To reduce training time, these errors were calculated on models trained using only 10% (randomly selected) of the total training data and validated on 10% of the validation data. Adsorbate descriptors considered were *(i)* $\epsilon_{e\text{ffective}}$, $\sigma_{e\text{ffective}}$, and the maximum charge magnitude, *(ii)* $\epsilon_{e\text{ffective}}$, $\sigma_{e\text{ffective}}$, and the maximum charge magnitude multiplied by the bond length, *(iii)* $\epsilon_{e\text{ffective}}$, $\sigma_{e\text{ffective}}$, maximum charge, and bond length, *(iv)* $\epsilon_{e\text{ffective}}$, $\sigma_{e\text{ffective}}$, maximum charge, bond length, and the aspect ratio (shortest dimension divided by the longest dimension). We found XGBoost models trained on *(i)* and *(ii)* (together with MOF descriptors) were generally less accurate that *(iii)* and *(iv)*. We, therefore, chose adsorbate descriptor set *(iii)* as it yeilded increased accuracy over *(i)* and *(ii)* and *(iv)* was not more accurate than *(iii)* despite the extra variable.

The error of the XGBoost models trained on the seven MOF descriptors considered, together with adsorbate descriptor set *(iii)* are shown in **Table S6**. The chemical motifs descriptor set includes the chemical motifs described in the main text (**Figure 2**). Two atomic property-weight radial distribution functions (AP-RDFs) were used in the AP-RDF descriptor, with framework epsilon and sigma (LJ parameters) as the two atomic properties (see ref [3] for details). The bag-of-atoms descriptor was calculated by dividing each MOF unit cell into $6 \times 6 \times 6$ cuboids and calculating the sum of the framework atom epsilon and sigma in each cuboid (each cuboid had two variables, the sum of epsilons and the sum of sigma), then normalizing by the total number of framework atoms to make the descriptor intensive. While the bag-of-atoms descriptor, together with textural properties proved the most accurate model, we found that the massively increased complexity of the data (438 total MOF descriptors verses 23 for the next most accurate), and, thus, the models that would be required to learn that data, was not worth the marginal decrease in error.

Table S6. XGBoost model error when trained on the indicated MOF descriptor set, together with adsorbate descriptor set *(iii)*, see above. The selected set is in bold.

MOF Descriptor Set	**MAPE** [%]	**MAE** [mol/kg]
Textural	31.7	2.91
AP-RDFs	31.8	2.97
Chemical Motifs	32.7	3.15
Bag-of-atoms	36.5	3.93
AP-RDFs + Textural	30.7	2.96
Chemical Motifs + Textural	**30.6**	**2.92**
Bag-of-atoms + Textural	28.5	2.85

After selecting the descriptors to use we proceeded to tune our final model. First we tuned network topology on a grid, where between one and eight hidden layers were considered, where each hidden layer could have between 10 and 50 nodes (in increments of 10), the only restriction being that each layer could not have more nodes than the previous layer (with the exception of the first layer). After settling upon the final topology, we tuned other important hyperparameters on a grid. Namely, the activation function (we considered sigmoid, and ReLU), the optimizer used in the backpropagation algorithm (we considered, stochastic gradient descent, Adam,[4] and NAdam[4]), the method used to initialize the network weights (we considered Glorot normal,[5] Glorot uniform,[5]

LeCun normal,[6] and LeCun uniform[6]), the learning rate used in the optimizer (we considered values of 0.1, 0.05, 0.01, 0.005, 0.0001, and 0.00005), the batch size (we considered values of 10 to 100 in increments of 10), and the loss function (we considered mean absolute error, mean squared error, mean absolute percentage error, and mean squared logarithmic error). The fina values for these hyperparemters are presented in **Table S7**. Note that we used early termination with a patience of 20 for every network in this procedure; a maximum of 500 epochs was allowed, and no networks hit that maximum.

Table S7. The optimal hyperparameter values for our final MLP found using a grid search. Topology is denoted by N_1-...-N_M, where N_i is the number of nodes in each hidden layer.

Hyperparameter	**Values**
Topology	50-20-20-20
Activation	sigmoid
Optimizer	Adam[4]
Initializer	Glorot normal[5]
Learning rate	0.0001
Batch Size	50
Loss	MAE

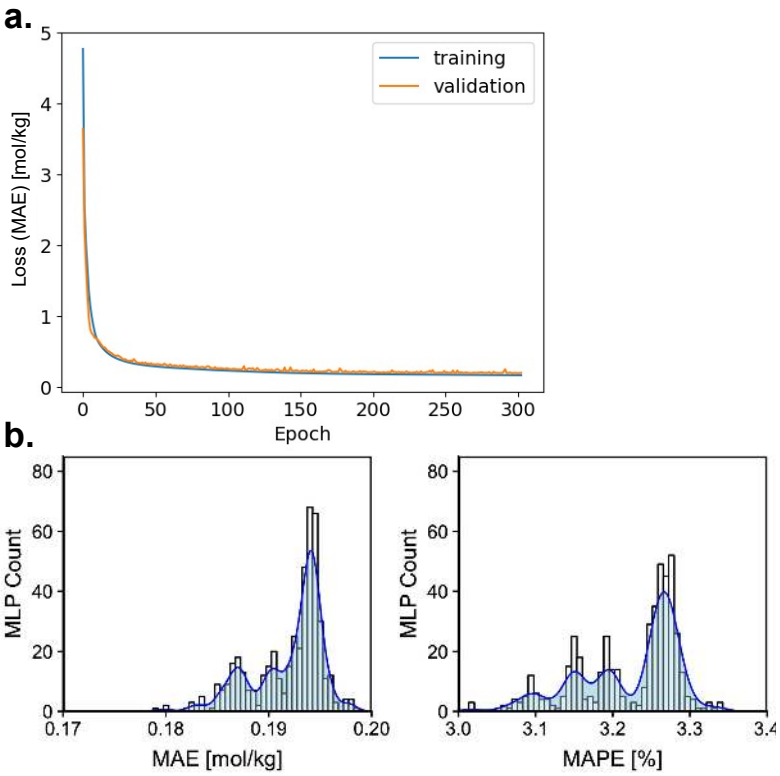

Figure S2. a. Shows the learning curve of our final model, demonstrating that our training set is representative of the validation set (and vice-versa) and that the model was not overfit by fitting over too many epochs at the learning rate used. b. shows the error variability in 500 replications of our final model trained with different random seeds, MAE varies between 0.18 and 0.20 mol/kg and MAPE varies between 3.0 and 3.35 %.

Section S3. Additional Model Prediction Data

Table S8. Model performance metrics of our final model for loading predictions made on adsorbates with properties outside of range spanned by the training set (extrapolative predictions).

Adsorbate	**MAPE** [%]	**MAE** [mol//kg]	**R**	**S**
Helium	20.3	0.16	0.998	0.999
Hydrogen	19.4	0.55	0.998	0.999
Propane	34.3	4.42	0.943	0.899
Butane	70.0	4.54	0.877	0.771
Isobutane	14.2	2.09	0.933	0.934
Benzene	9.4	1.37	0.971	0.990

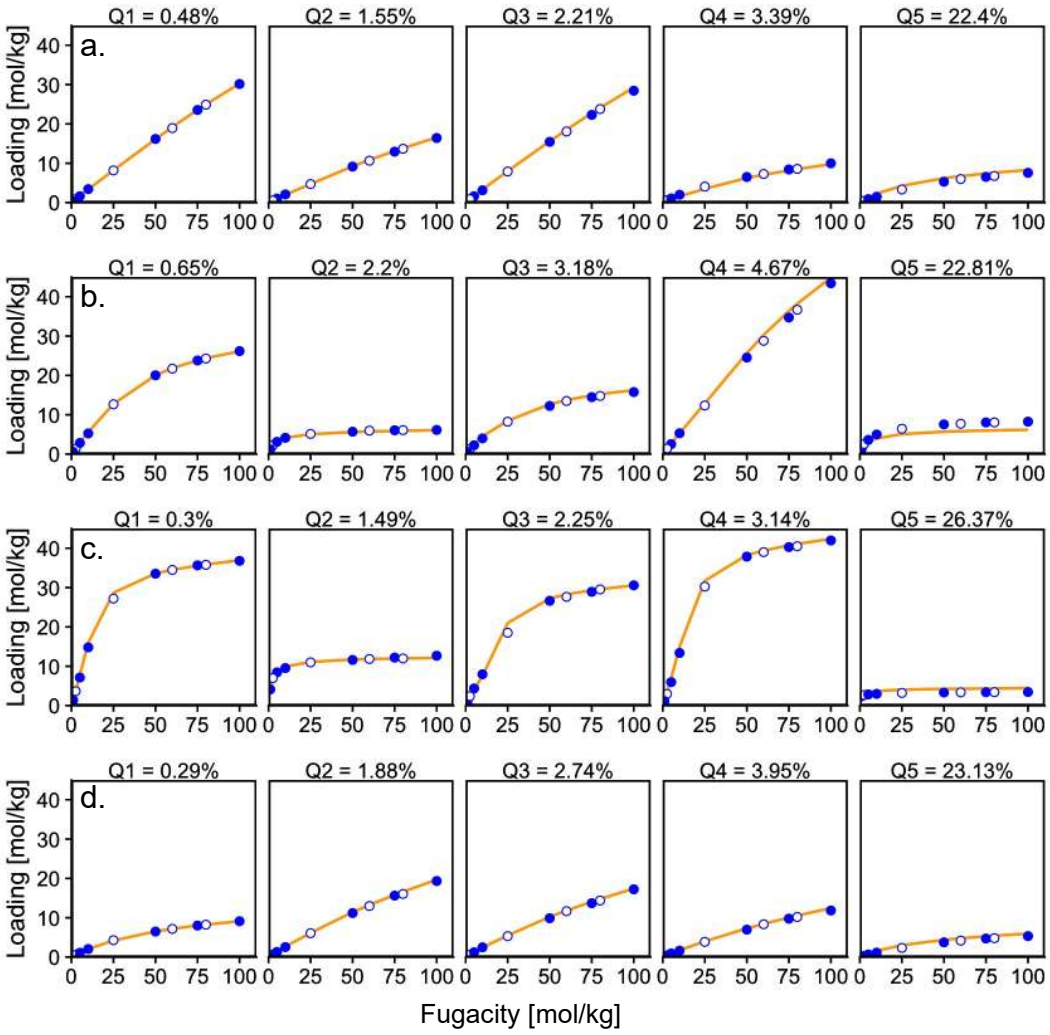

Figure S3. Plots analogous to **Fig. 7** in the main text for predictions on the remaining simple adsorbates, with adsorbates: a. argon, b. krypton, c. xenon, and d. nitrogen. Fugacities corresponding to filled, blue points were included in training, empty points correspond to pressures not included in training, orange lines correspond to model predictions.

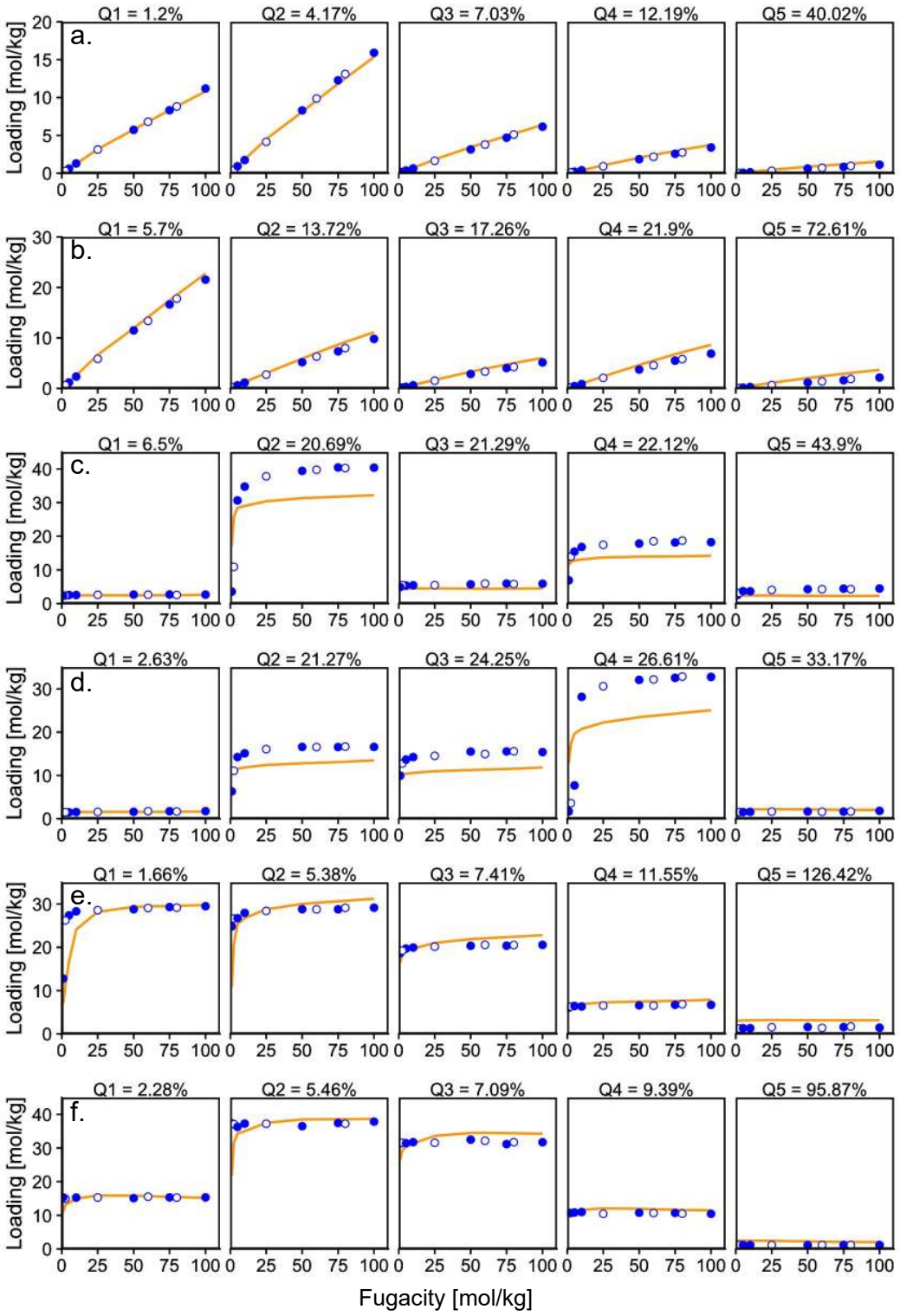

Figure S4. Plots analogous to **Fig. 7** in the main text for extrapolated predictions, with adsorbates: a. helium, b. hydrogen, c. propane, d. butane, e. isobutane, f. benzene. Fugacities corresponding to filled, blue points were included in training, empty points correspond to pressures not included in training, orange lines correspond to model predictions.

Table S9. The number of MOFs in the top 100 from GCMC data that were encountered in the top 100 predicted using data predicted by machine learning for the six "extrapolated" adsorbates.

Adsorbate	# Correctly Placed	
	Loading@100 bar [mol/kg]	Working Capacity [mol/kg]
Helium	99	99
Hydrogen	98	98
Propane	99	94
Butane	98	95
Isobutane	99	89
Benzene	99	46

References

1. Rojas, R. *Neural Networks: A Systematic Introduction*. (Springer-Verlag, 1996).
2. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016**,** 785–794 (Association for Computing Machinery, 2016).
3. 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* **117,** 14095–14105 (2013).
4. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* 1–15 (2014).
5. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (eds. Teh, Y. W. & Titterington, M.) **9,** 249–256 (PMLR, 2010).
6. LeCun, Y., Bottou, L., Orr, G. B. & Muller, K.-R. Efficient BackProp. in *Neural Networks: tricks of the trade* (Springer, 1998).