Check for updates

View Article Online
View Journal

MSDE

Molecular Systems Design & Engineering

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: T. Gercina de Vilas, J. F. Fajardo-Rojas, O. Mansurov, R. Devaisher, E. Toberer and D. Gomez-Gualdron, *Mol. Syst. Des. Eng.*, 2026, DOI: 10.1039/D6ME00034G.

This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Design, System, Application

Nanoporous materials, with metal-organic frameworks (MOFs) as their flagship, are promising to harness control of adsorption phenomena, and in turn achieve breakthroughs in a number of engineering applications. With a design space spanning trillions of materials variations, machine learning (ML) adsorption models are crucial to identify optimal designs with suitable adsorption properties for a target application, and to derive data-driven design rules that can further support material development via rational design. But training ML adsorption models comes with its own challenges, some of which can be alleviated with the use of physics-based materials representations. This kind of representation offers advantages such as compatibility with (easier to generate and potentially more informative) artificial training datasets, model applicability beyond a material subclass, and resilience to instances of flawed chemistry-adsorption relationships in the data. However, current physics-based nanomaterial representations tend to require specialized expertise for their creation and/or are prone to raising training scalability issues. The work herein demonstrates two-dimensional interaction parameter histograms (2D-IPHs) as simple, yet informative, material representations that can facilitate the data-efficient, scalable training of ML models that are crucial to screen and obtain data-driven molecular design rules for these materials.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Two-dimensional interaction parameter histograms as a simple and versatile nanoporous material representation for machine learning prediction of adsorption properties

Tatiane Gercina de Vilas,[1,#] Fernando Fajardo-Rojas,[2,4,#] Omar Mansurov,[1,3] Ruby Devaisher,[1] Eric Toberer,[2,4] Diego A. Gómez-Gualdrón[1,4] *

[1] Department of Chemical and Biological Engineering, Colorado School of Mines, 1601 Illinois St., Golden CO 80401, USA
[2]Department of Physics, Colorado School of Mines, 1500 Illinois St., Golden CO 80401, USA
[3]Computer Science Department, Colorado School of Mines, 1500 Illinois St., Golden CO 80401, USA
[4] Materials Science Program, Colorado School of Mines, 1601 Illinois St., Golden CO 80401, USA

#These authors contributed equally
*Corresponding author: dgomezgualdron@mines.edu

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Molecular Systems Design & Engineering Accepted Manuscript

ABSTRACT

Machine-learning (ML) adsorption models are essential to computationally screen nanoporous materials, such as metal-organic frameworks (MOFs). Physics-based MOF representations offer advantages for the training of these ML models such as compatibility with artificial training datasets, model applicability beyond MOFs, and resilience to chemistry-related inaccuracies in the data. However, emerging physics-based MOF representations tend to require specialized expertise for their creation and/or are prone to raising training scalability issues. Here, we demonstrate two-dimensional, interaction-parameter histograms (2D-IPHs) as physics-based MOF representations that are simple, scalable, and informative for adsorption learning. The construction of 2D-IPHs simply needs statistics of the distance of adsorption sites to their closest pore wall atom, along with its interaction parameters. Demonstrating scalability, 2D-IPHs facilitated the use of a multi-million-point, multi-molecule dataset to yield a model that predicts adsorption isotherms for unseen small, non-polar, near-spherical molecules (R^2 = 0.97 - 0.99 for H_2, CH_4, C_2H_8, N_2, Ar, Xe, and Kr). Demonstrating informativeness, 2D-IPHs facilitated training from multi-thousand-point, single-molecule datasets to yield models for: *i)* full adsorption isotherm prediction for small, high-quadrupole and non-spherical molecules (R^2 = 0.98 for CO_2 and C_3H_8), and *ii)* Henry's constant prediction for small, molecules of varied adsorption dependence on dispersion and electrostatic interactions (R^2 = 0.76 - 0.90 for, CO_2, H_2O, and NH_3 and N_2). Moreover, training with 2D-IPHs tended to be robust to training dataset trimming, at least until running into obvious data-scarce scenarios. Even so, in data-scarce scenarios, the use of 2D-IPHs with techniques such as single feature stacking (SFS) and transfer learning (TL) led to significant (even if not total) recovery in model accuracy. Nuances regarding SFS and TL, and the practical screening performance of the models trained herein, are also discussed in this work.

1. INTRODUCTION

Harnessing control of adsorption in nanoporous materials could make a myriad of engineering applications possible.[1,2] Through judicious selection of their constituent building blocks, metal-organic frameworks (MOFs) are promising nanoporous materials to harness this control.[3,4] However, one bottleneck is having to identify, among trillions of possibilities, the precise building block combination that engenders a desired adsorption behavior.[5] Exploring all this combinatorics solely by experiments is intractable, making the prediction of adsorption properties of a prospective MOF by computation necessary.[6] To this end, the virtually *instantaneous* inference of machine learning (ML) models is extremely appealing for high throughput adsorption predictions.[7,8] Thus, there are sustained efforts to develop ML models that can be trained efficiently to gain broad adsorption prediction capabilities in MOFs.[9–15]

Central to the development of ML adsorption models is the development of MOF representations to be used as inputs.[16] Some representations have relied on feature engineering, with the effort in input preparation expected to be compensated with the ability to learn with simpler models and/or smaller datasets.[17–19] One prominent example is atomic property-weighted radial distribution functions.[17] Other representations have sought to minimize input preparation effort, in which case larger datasets and/or more complex models are needed, so that the model "extracts" the features itself.[7,13,20] Examples include, but are not limited to, representing MOFs (or their building blocks) as graphs, directly converting material atoms and bonds into graph nodes and edges, respectively.[21–23] Then, each node and/or edge have been usually described either by a list of "periodic table" properties (e.g., electronegativity) of the atom(s) associated with it or by their chemical identity (e.g. via one-hot encoding[12,21,24,25]). Such representations have been used to predict MOF properties via graph neural networks (GNNs) or variations thereof.[26–29]

Complexity aside, most MOF representations have the model learn a relationship between adsorption and material chemistry,[11–13,16] which is then exploited to perform predictions for new MOFs. However, since ML adsorption models can only be as accurate as the data they are trained on, concerns on the validity of this learned relationship linger due to the common use of *generic* force fields to generate training adsorption data via molecular simulation. Although new force fields can be reparametrized to correct faults in extant training data,[30,31] and ΔML approaches—that focus on learning and correcting model errors—can be used to reduce the demands for new data, the use of chemistry-based representations does not allow to bypass the need for additional model training efforts. Furthermore, representations directly tied to chemical identity may make adsorption property predictions impossible when encountering unseen chemical moieties, fundamentally impairing the utility of the corresponding models for out-of-distribution predictions.

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

A strategy to bypass the above issues is to use physics-based MOF representations, where models learn the relationship between the property of interest and chemistry-agnostic parameters.[10,20,32–35] The "energy histograms" introduced by Snurr and coworkers are notable examples of physics-based representations.[10,20,35] These authors used this representation to teach models to predict adsorption based on the distribution of adsorption site energies. An appealing aspect of energy histograms is their compatibility with simple ML architectures (and benefits derived therein), but drawbacks include their specificity to a particular adsorbate probe (although some histograms can be transferable between molecules if these are significantly similar[20,35]), and the significant simulation expertise needed for their creation.[10,20,32–35] The interaction parameter-embedded cubic grids introduced by Lin and coworkers[33,34] are also notable physics-based representations. These grids have been used as input to models that learn to predict adsorption based on the Lennard Jones (LJ) potential parameters ε and σ and coulomb charge *q* of each grid point.[32,33] The philosophy of this representation arguably follows that of work in other areas that uses ML to solve complicated simulation models by using the simulation model parameters as input.[36] Appealing aspects of these grids are their relative ease of construction and parameter assignment (which does not require simulation expertise), and their non-specificity to a particular adsorbate. Arguably, the greatest drawback of these grids is their matching with complex ML architectures such as 3D-CNNs that may rapidly become prohibitive for training as either the size of the grid or the dataset increases.[8]

The above works demonstrate that physics-based representations can yield models with adsorption property prediction accuracies that are practical for materials screening. However, the noted drawbacks can hinder ML democratization, potentially limiting the widespread use of the above models and approaches for screening of materials for adsorption applications. Accordingly, in this work, we propose the use of 2D-histograms encoding information of MOF LJ and Coulomb interaction potential parameters (2D-IPHs) as a simple, scalable, physics-based MOF representation that can be generated with minimal simulation expertise, and that is highly conducive to adsorption learning by ML models. We envision 2D-IPHs to facilitate model development without the need for significant simulation expertise,[10,20,32–35] extensive computational resources,[8] or specialized ML-oriented feature engineering pipelines. This representation was introduced in earlier work[37] in a less demanding context, where it was used as input for surrogate ML models guiding the "active search" of MOFs. Building upon that foundation, in this work we demonstrate 2D-IPHs to be sufficiently informative in broader and more demanding adsorption prediction tasks. Indeed, to illustrate the suitability, versatility and robustness of this MOF representation, we demonstrate its performance as input to models that predict either full adsorption isotherms or adsorption Henry's constants. Additionally, we illustrate the suitability of the representation for some emerging approaches for data-efficient ML model training such as single-feature stacking (SFS), and inductive transfer learning (TL).

2. COMPUTATIONAL METHODS

2.1. MOF selection. The MOFs selected for this work are a subset of the larger MOFMinE database.[38] This database features 1,036,252 MOFs constructed using ToBaCCo-3.0,[39,40] combining 27 inorganic nodular building blocks, 14 organic nodular building blocks, 19 base edge building blocks (with 13 functionalized variations for each of the latter) into 1,393 topologies (a sample of building blocks is shown in **Fig. 1a**).

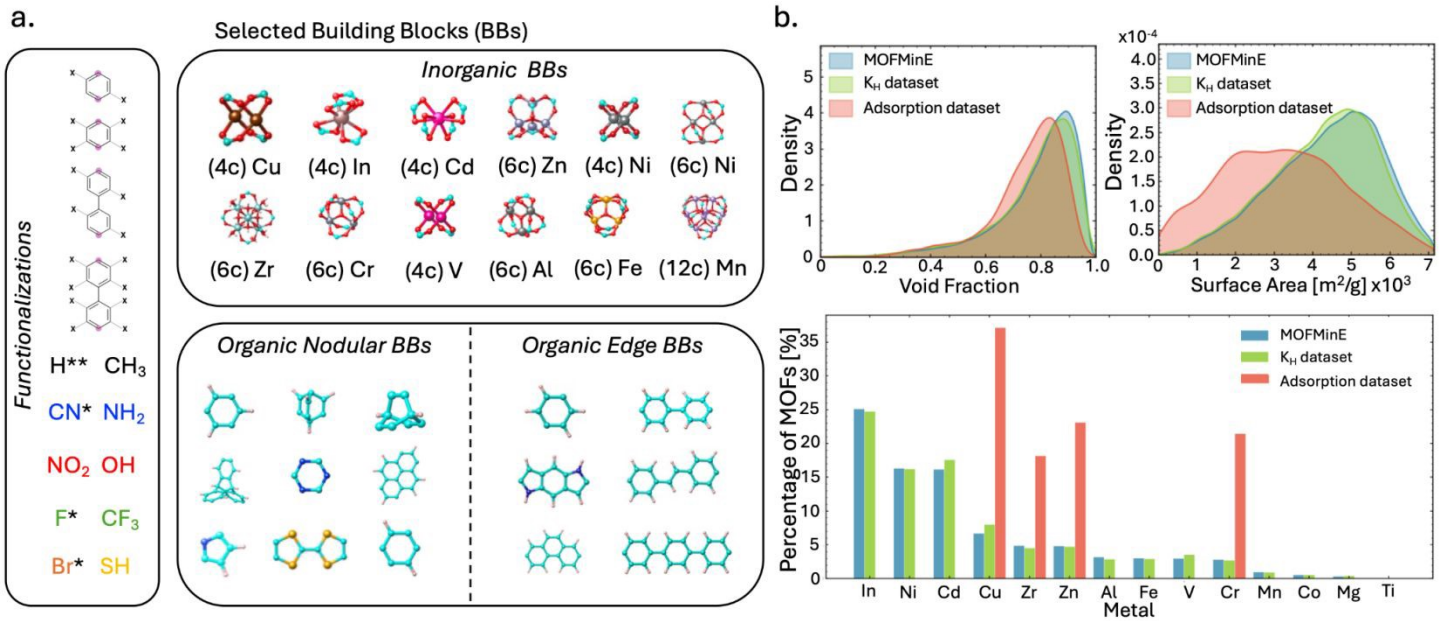

Figure 1. Overview of building blocks and properties in the MOF sets used to train models to predict adsorption loadings and Henry's coefficients (K_H). a) Representative building blocks (BBs) in the MOFMinE database. The top panel shows a sample of inorganic building blocks, while the bottom panel shows a sample of nodular (left) and edge (right) organic building blocks. b) Comparison of the structural and compositional distributions between datasets: the 1,036,252 MOFs in MOFMinE (blue), the 39,950 MOFs used to train models to predict K_H (K_H dataset, green), and the 3,029 MOFs used to predict adsorption loadings (adsorption dataset, red). The top panel reports textural properties, and the bottom panel shows the distribution of metals based on the 14 most-common metals in MOFMinE.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Upon generation, all MOF structures were optimized using an iterative procedure. First, only atomic positions were relaxed, followed by a second relaxation of both atomic positions and unit cell parameters. Optimizations used molecular mechanics with UFF4MOF[41] interaction parameters as implemented in LAMMPS (29 Oct 2020).[42] For Henry's constant calculations, 39,950 MOFs (henceforth referred to as the K_H subset) were randomly drawn from the MOFMinE database. These MOFs reflect the distribution of both textural properties and chemical distribution present in the full MOFMinE dataset. (**Fig. 1b**). For adsorption loading calculations, the selected MOFMinE subset simply corresponds to 3,029 MOFs used in earlier works (henceforth referred to as the adsorption subset).[13,37,43] This MOF subset in many ways reflect the distribution of both textural and chemical distributions of the full MOFMinE dataset, except that it focuses more on MOFs with surface areas in the ~2000 to ~4000 m^2/g range and MOFs with Cu, Zn, Cr and Zr nodes (**Fig. 1b**).

2.2. Adsorption interaction parameters. CH_4, C_2H_6, C_3H_8, N_2 and CO_2 were modeled by TraPPE parameters,[44–47] which are known to reproduce vapor-liquid equilibrium curves for these molecules. Xe and Kr were modeled following a single sphere model with Lennard-Jones parameters reported by Snurr and coworkers,[48] which have been broadly used to study their adsorption-based selectivity.[43,48] Ar was modeled as a single sphere as well, and the parameters used were reported by García-Pérez *et al*.[49] H_2 was modeled following the well-known dispersion attraction-driven sorption study by Darkrim and Levesque.[50] Feynman-Hibbs corrections[51,52] were used for this molecule. H_2O was modeled by TIP4P parameters, which provides good experimental agreement of the O-O interactions between water molecules.[53,54] Adsorption data for alchemical adsorbates was taken from previous work,[13] where they were modeled as one or three-site molecules (reminiscent of three-site models under the TrAPPE force field), with σ and ε LJ parameters, and charge and bond lengths in the 3.0-4.5 Å, 15-250 K, 0.0-0.9 e and 1.0-2.0 Å ranges, respectively. The range for each parameter was selected to be broad, yet physically reasonable, based on the typical values of parameters in real molecules (**Fig. 3d**). The specific combinations of parameters in alchemical species can be found in a previous work.[13] MOF atoms were assigned σ and ε parameters based on UFF4MOF, and charges based on the DFT-calculated charges on their building blocks, according to the MBBB method.[55] Cross-interactions for adsorbate-adsorbate and adsorbate-MOF cases were modeled using Lorentz-Berthelot mixing rules. No interactions between MOF atoms were calculated as these atoms remained fixed during simulations. Cutoffs of 12.8 Å were used for both LJ and Coulomb potentials. Electrostatic interactions beyond the cutoff were modeled using Ewald summation (with a precision of 10^{-6}). No tail corrections were used for LJ potentials.

2.3. Simulations of adsorption loadings. Simulated adsorption loadings were taken from extant datasets or obtained here using grand canonical Monte Carlo (GCMC) with the RASPA.2.0 code.[56] These simulations model the MOF as being in contact with an adsorbate reservoir at fixed chemical potential. At least 5,000 initialization cycles followed by 1,000 data collection cycles were used. Each cycle corresponds to N Monte Carlo trial moves (i.e., translation, insertion/deletion, and rotation), where N for each cycle is set as the highest number between 20 and the number of adsorbates in the simulation cell. Metropolis-Hasting acceptance criteria were used to accept or reject these moves.[57,58] For translation and rotation, these criteria involve the energies of the adsorbed phase configurations. For insertion/deletion, these criteria also involve the chemical potential and, in turn, *fugacity* of the reservoir,[59] which was directly set at desired values generally ranging from 0.01 to 100 bar.

2.4. Henry's constant calculations. Henry's constants (K_H) for H_2O, NH_3, CO_2, and N_2 were calculated at 298 K using the Widom insertion method. Briefly, the molecule of interest was inserted, deleted, and then reinserted 10,000 times at random points within the MOF unit cell, collecting the adsorption energy (ΔU) upon each insertion. Assuming an ideal Rosenbluth weight equal to one, then K_H is calculated as:

$$K_H = \frac{\beta}{\rho_f} < e^{(-\beta \Delta U)} > \qquad (1)$$

where ρ_f is the density of the MOF, β denotes the inverse of RT, and the angular brackets indicate an average quantity. Note that K_H [mol kg^{-1} Pa^{-1}] indicates the inherent affinity of a molecule to a MOF but it is also the slope of the adsorption isotherm at sufficiently small pressures (i.e., dilute adsorption conditions).

2.5. 2D interaction-parameter histograms. As schematized in **Fig. 2a**, to construct 2D-IHPs we first generate an evenly spaced grid of points within the MOF unit cell, with each point intended to represent a potential adsorption site. The spacing between grid points along lines parallel to the unit cell vectors is 1 Å. Thus, the shape and symmetry of the grid is consistent with the shape and symmetry of the MOF unit cell. Note that while the use of cubic grids of fixed extent "sampled" from the MOF regardless of unit cell characteristics is common with sophisticated model architectures such as 3D-CNNs,[34] we prefer our grid approach because it prevents information loss for non-cubic unit cells, among other previously mentioned advantages. For each grid point, we then identify the corresponding closest MOF atom, while considering periodic boundary conditions. As noted in previous work, if done by brute force, this identification can become intractable for large unit cells, so it can be aided by approaches such as KD trees.[37] Each grid point is "embedded" with the distance (d) to the closest MOF atom, as well as the non-bonded interaction parameters (i.e., σ and ε LJ parameters and Coulombic partial charge q).

The above embedding is motivated by the hypothesis that the adsorption energy of each grid point (adsorption site) is greatly influenced by the characteristics of the closest MOF atom. The adsorption energy of each site depends on the particular ε-d, σ-d, and q-d combination associated with it. Thus, analogous to how Snurr and coworkers[10,20] build energy histograms by counting the frequency that certain adsorption energy values occur, we build three 2D-IPHs for each MOF by counting the frequency that certain ε-d, σ-d, and q-d combinations occur. The number of bins for each parameter is fixed to ensure same dimensionality regardless of the MOF size (74 x 20 = 1480 for each 2D-IPH; see example in Fig. S1). To obtain a vectorial representation compatible with multilayer perceptrons (MLPs), each 2D-IPH was "flattened" by

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

concatenating their rows, and the flattened 2D-IPHs were in turn concatenated as well. We found that this simple flattening led to models that were as good as (or better than) those obtained by preceding the flattening with the application of convolutional layers.

The "as is" flattened representation consists of a vector with 4,440 components stemming from the 1,480 bins for each of the three 2D-IPHs, but depending on the size of the dataset, one may decide to reduce the size of this representation. When such reduction was opted in this work, representation components were removed based on their variance across MOFs. The rationale is that components that barely vary across MOFs are less likely to "explain" differences in adsorption behavior and impact model performance. The variance threshold can be modulated depending on how much representation-size reduction is targeted. For instance, we did not implement this reduction when working with the multi-thousand-points dataset in Section 3.3, but we did when working with the multi-million-points dataset in Section 3.1, using a standard deviation threshold of 0.032 that reduced the representation to 44 features. Notably, we found this simple procedure to result in models as good as (or better) than those relying on PCA or autoencoder approaches.

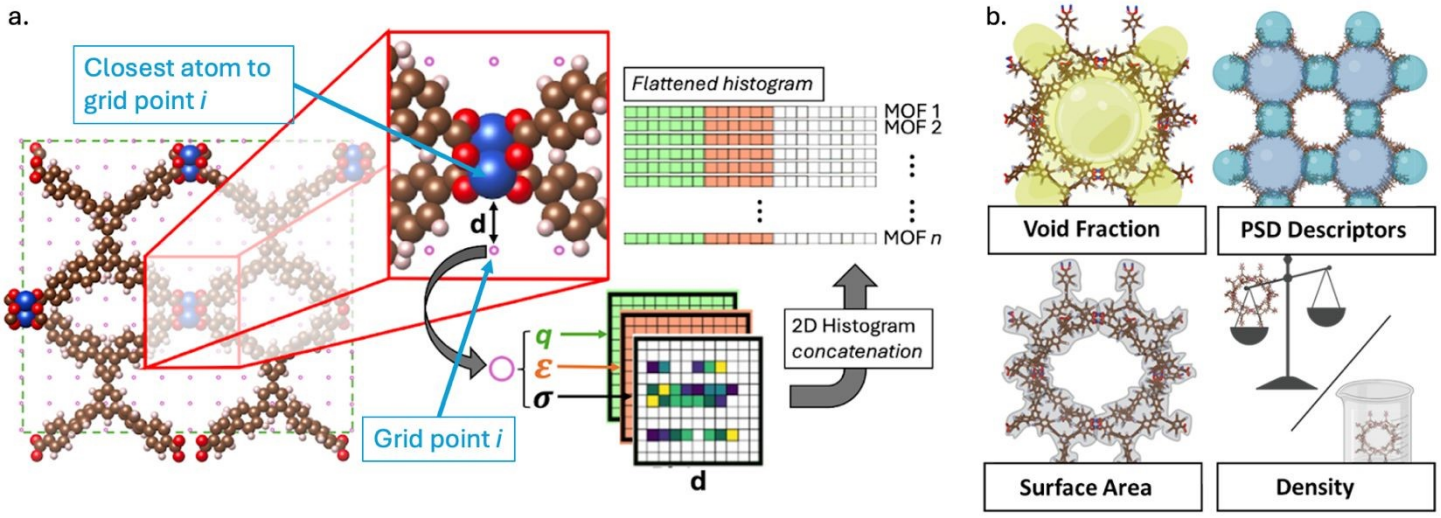

Figure 2. a) Representation of the workflow for constructing the 2D interaction-parameter histograms (2D-IPHs). A uniform grid is generated in the MOF unit cell, and for each grid point the distance to the nearest MOF atom is computed and stored together with the corresponding non-bonded interaction parameters (partial charge (q) for Coulombic interactions and the Lennard-Jones parameters (ε, σ) for dispersion interactions). For each MOF, three 2D-IPHs are built (one for each non-bonded interaction parameter considered) as a function of distance (d). These histograms are then flattened and concatenated to form the final MOF representation used for both K_H and adsorption loading predictions. b) Schematic of the global textural descriptors complementing the 2D-IPH MOF representation for adsorption loading prediction models.

2.6. Machine learning model training. All models trained in this work are based on MLP architectures purposely kept at moderate complexity, and developed using Python 3.10.2, TensorFlow 2.15.0, and Scikit-learn 1.5.2, unless otherwise specified (**Supporting Information**). The choice for MLPs was made purposefully to assess the practical informativeness of the proposed 2D-IPHs MOF representation when used with a highly expressive type of model. The models used the flattened histograms as input, in some cases along with select (experimentally obtainable) MOF textural features (**Fig. 2b**). One can think of textural features as globally descriptive, and of the histograms as more locally descriptive (i.e., distributions of local interaction environments) features. All features were normalized or standardized based on training-set statistics.

The models for adsorption loading prediction and for K_H prediction tasks were trained using mean absolute error (MSE) as the loss function. As one of the measures to prevent overfitting, model training considered dropout rates in each trained layer ranging from none to 30%. As needed, early stopping based on validation loss, with patience of 20 epochs and restoration of the best weights, was used as well. Hyperparameter exploration for all models was done via Keras Tuner Bayesian optimization to minimize validation mean absolute error (MAE) over model complexity, dropout, regularization, learning rate, and activation using 40 trial per model. All adsorption prediction results are reported on the same test set of MOFs, corresponding to 1,029 MOFs randomly drawn from the 3,029 MOFs in the adsorption subset. Likewise, K_H prediction results are reported on the same test set of MOFs, corresponding to 7,990 MOFs randomly drawn from the 39,950 MOFs in the K_H subset. To prevent data leakage, these test sets remained unobserved during all corresponding hyperparameter exploration and training processes. Further implementation details, including the data distribution used for each training approach and performance comparison of different representation variants for each prediction task are provided as **Supporting Information**. Note that the physics-grounded nature of the 2D-IPH representation makes it amenable to feature importance analyses, where individual features can be traced back to specific interaction parameter–distance combinations, offering a path to rationalize MOF design (e.g, conceiving chemical moieties that are describe by desirable interaction parameters). However, given the scope of the work herein such interpretability studies have been reserved for subsequent studies.

3. RESULTS AND DISCUSSION.

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

3.1 General adsorption model. The dominant approach to ML adsorption prediction has been to train a model based on an adsorption dataset for a specific adsorbate (usually at a specific condition).[9,17,60,61] Given that datasets are harder to generate for some adsorbates than for others, a more data-efficient approach to adsorption model training could be based on the aggregation of datasets for different adsorbates (even at different conditions) into a larger "master" adsorption dataset. This approach enables training more "general" ML adsorption models that can predict adsorption even for previously unseen adsorbates—based on what the model has learned from other adsorbates. This approach arguably returns to the philosophy of early 20th century analytical adsorption models (e.g., Langmuir,[62,63] BET,[64] and so forth[65]), as it uses adsorbent and adsorbate properties, as well as thermodynamic conditions as input (**Fig. 3a**).

Pursuing this approach, Sholl and coworkers[66] used a large diverse dataset of hydrocarbon-based CHNOPS adsorbates and MOFs for training, to develop a general isotherm equation (via multiple regression genetic programing) to predict single-component adsorption isotherms across this family of adsorbates in MOFs. These authors used counts of chemical moieties as part of the input. In the same general model spirit, Anderson *et. al.*[13] developed a MLP model to predict full isotherms for small, near-spherical, non-polar molecules. Although these authors used physics-based representation for the adsorbate, their MOF representation was still chemistry-based, similar to the atom type-based representation proposed by Fanourgakis *et al.*[67]. Notably, the nature of their adsorbate representation enabled the introduction of "informative" alchemical adsorbates in their dataset to facilitate learning. Intriguingly, Froudakis and coworkers[68] have shown this "alchemical" strategy can be similarly beneficial when applied to the adsorbent, adding to the desirability of a physics-based representation for MOFs.

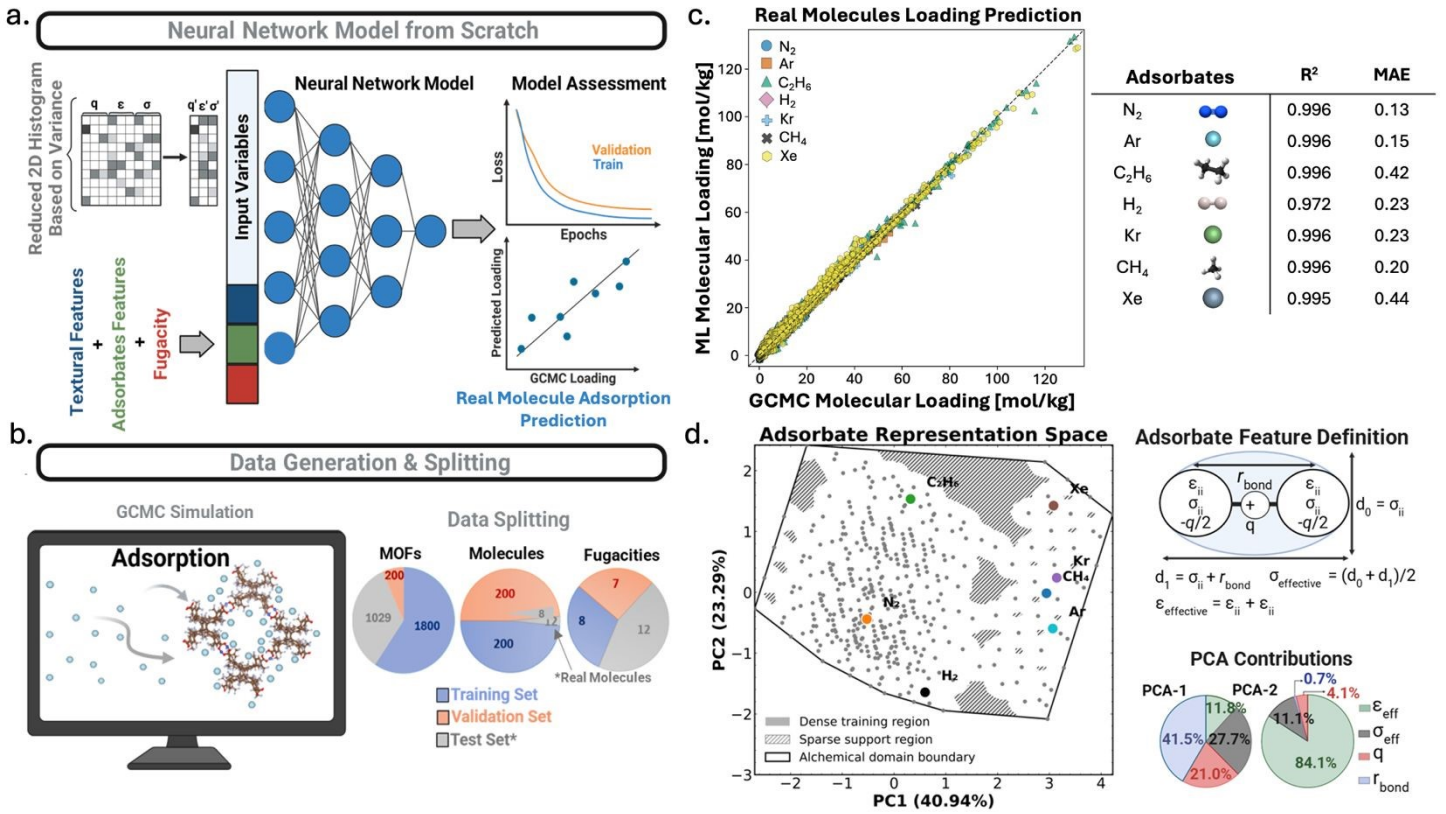

Figure 3. General adsorption model training and predictions. a) Overview of the general adsorption model workflow, indicating model inputs. b) GCMC simulations (left) provide adsorption data across MOFs, adsorbates, and fugacities, which are partitioned (right) into training and validation sets for alchemical adsorbates and a test set for real molecules, the latter which is reserved for prediction evaluation. c) Parity plots comparing general model predictions versus GCMC values for multiple real molecules. R^2 values and mean absolute error (MAE) are presented for each real molecule tested. d) Visualization of the adsorbate representation space obtained by principal component analysis (PCA; left) of the adsorbate descriptors (top right), and the contributions of those descriptors (bottom right) to the PCA components. Gray points in PCA visualization correspond to alchemical adsorbates (training), while colored points indicate real adsorbates (test). The outlined region highlights the adsorbate representation domain spanned during training.

Moreover, such physics-based representation types also open the door to using the models beyond MOFs.[20]

Accordingly, here we leverage the ~5 million-point Anderson *et al*. GCMC dataset[13] to train a general adsorption model that completely relies on physics-based representations, by using our proposed 2D-IPHs to represent the MOFs. We maintained these authors' dataset partition philosophy where no MOF, fugacity, or adsorbate seen in the training and validation datasets is seen in the test set. Importantly, real molecules are only ever seen in the test set (**Fig. 3b**), which enables for assessment of model "zero-shot-like" predictive capability. A variance cutoff equal to 0.032 was used to reduce the histograms, resulting in a 44-component feature vector. Showing that 2D-IPHs lead to scalable model training, a variety

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

of models were readily trained, where the best model had an architecture of a multilayer perceptron with two hidden layers containing 128, and 256 nodes, respectively.

For the targeted molecules CH_4, C_2H_6, N_2, H_2, Ar, Xe and Kr, the new model reached prediction accuracy characterized by R^2 values higher than 0.97 and mean absolute error (MAE) values corresponding to ~6–8% of the mean GCMC adsorption loading for six molecules, and ~22% for H_2 (**Fig. 3c**). These accuracy levels are consistent with state of the art accuracy for adsorption models in MOFs, independently of the MOF representation used.[20,69,70] **Fig. 3c** shows that predictions closely follow the parity line across the explored loading range for all seven molecules, with no obvious loading-dependent deviation from parity. Given the interest in predicting full isotherms, this observation is desirable as it reflects good predictions for low and high fugacity, which correlate with low and high loadings, respectively.

Although parity plots give a sense that the ML model performs well, we believe it is always desirable to test models in the context that they are practically used in high-throughput computational screening (HTCS). In this context, one usually wants to use the ML model to identify a small subset of promising MOFs in which more refined (thus more costly) evaluation would be performed. For instance, one may desire to identify the top-100 from ML, hoping that evaluation with, say, GCMC confirms that these MOFs were truly outstanding. Accordingly, one desires strong numerical prediction to be accompanied by strong ranking fidelity, which is the case with the model herein. As shown in **Table 1**, the top-*n* MOFs (from the test set) identified by the general model for H_2 and CH_4 adsorption at a fugacity of 100 bar substantially overlap with the corresponding GCMC-derived top-*n* sets. The overlap was 97+% for the top-100 (about 90th percentile of the test set), and stayed at 90+% even for the top-20 (about 98th percentile of the test set). These adsorption cases were presented here for their relevance to the practical context of screening MOFs for H_2 and CH_4 storage, respectively.

Table 1. Number of MOFs in the ML top-*n* set that are in the actual (i.e., GCMC) top-*n* set, with the top-*n* sets constructed based on MOF H_2 and CH_4 adsorption loadings at a fugacity of 100 bar.

Adsorbate	*n* = 100	*n* = 50	*n* = 20
H_2	97	50	19
CH_4	97	49	19

The real molecule prediction performance facilitated by the alchemical approach can be understood from **Fig. 3d,** which provides a low-dimensional view of the adsorbate representation space through principal component analysis (PCA) of the four descriptors used in the general model (i.e., effective Lennard–Jones diameter (σ_{eff}), effective Lennard–Jones well depth (ε_{eff}), partial charge (q), and bond length r_{bond}). **Fig. 3d** shows that the seven real adsorbates considered in this section project themselves within the representation space spanned by the alchemical species, making the prediction of their adsorption properties closer to an interpolation exercise at which ML is known to excel. With the first two principal components (PC1 and PC2) accounting for 64.23% of the total variance in the adsorbate feature space, and PC1 and PC2 dominated by r_{bond} (41.5%) and ε_{eff} (84.1%), respectively, one can infer that the current real molecule predictions of the general model largely stems from its learning of how adsorbate size (as captured by r_{bond}) and the strength of its predisposition to adsorb (as captured by ε_{eff}) impact adsorption in a given adsorption environment.

3.2 Single-feature stacking. Although the training approach in Section 3.1 is geared towards a "general model," the applicability of the model therein is at this point still restricted to the *types* of adsorbates appearing in the Anderson *et al.* training dataset.[13] Namely, small, near-spherical, non-polar adsorbates. This occurs partly because these are the types of adsorbates that are fully described by the adsorbate features shown in **Fig. 3d**. For instance, the general model underestimates C_3H_8 loadings, presumably because it overestimates the volume of this molecule, which it (incorrectly) assumes to be spherical.[13] On the other hand, the general model underestimates CO_2 loadings, presumably because the adsorption of the alchemical molecules used for training seem not to be as influenced by electrostatic interactions, preventing the model from learning the impact of MOF charges on adsorption. Indeed, while CO_2 is non-polar, it has a significant quadrupole beyond what is seen in the alchemical molecules in the training set, and therefore not learned by the general model.

The above plays into a scenario sometimes encountered in ML, where an extant (primary) model is not sufficiently accurate for a prediction task but could aid to more efficiently train a new (secondary) ML model. This way, in an approach known as single feature stacking (SFS), the data burden for the secondary model may be reduced by using the prediction from the primary model as one of the inputs. Specifically, the (incorrect but informative) prediction of the primary model is used as an input feature for the secondary model, which is then trained to carry the desired task (**Fig. 4a**). Analogous ideas have been explored in other fields such as quantum chemistry, where ML models trained on low-fidelity calculations of a property (e.g., energy) are used as input to train ML models using scarcer high-fidelity data. For instance, Moharreri *et al.*[71] used a sequence of ML models trained to output molecular energy at lower levels of theory as input for a ML model to ultimately predict molecular energy at the B3LYP/aug-cc-pVTZ level.

Accordingly, we decided to test SFS for molecules that fall outside the scope of the general model trained in Section 3.1 (here used as the primary model) such as the abovementioned C_3H_8 and CO_2. The secondary model uses the general model predictions and the MOF histogram features as input. As SFS may be unnecessary in data-abundant scenarios, we examined the efficacy of SFS as a function of data availability. Specifically, by comparing the performance of the SFS model against a model trained from scratch using the same dataset. In this section, full (100%) data availability corresponds to datasets consisting of at least 12,353 GCMC-calculated adsorption loadings for each of CO_2 and C_3H_8, spanning fugacities from 0.01 bar to 100 bar. Data partition closely resembles that used in Section 3.1, with the MOFs following the same splitting as in **Fig. 3b**, and fugacity partition detailed in **Table S2**.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

The performance of the scratch models was relatively stable for C_3H_8 and CO_2 cases as the data availability was reduced from 100% to 30%. A noticeable decline in performance was then apparent when data availability fell to 10% (1,235 data points for training), more pronounced for C_3H_8 case (**Fig. 4b**).

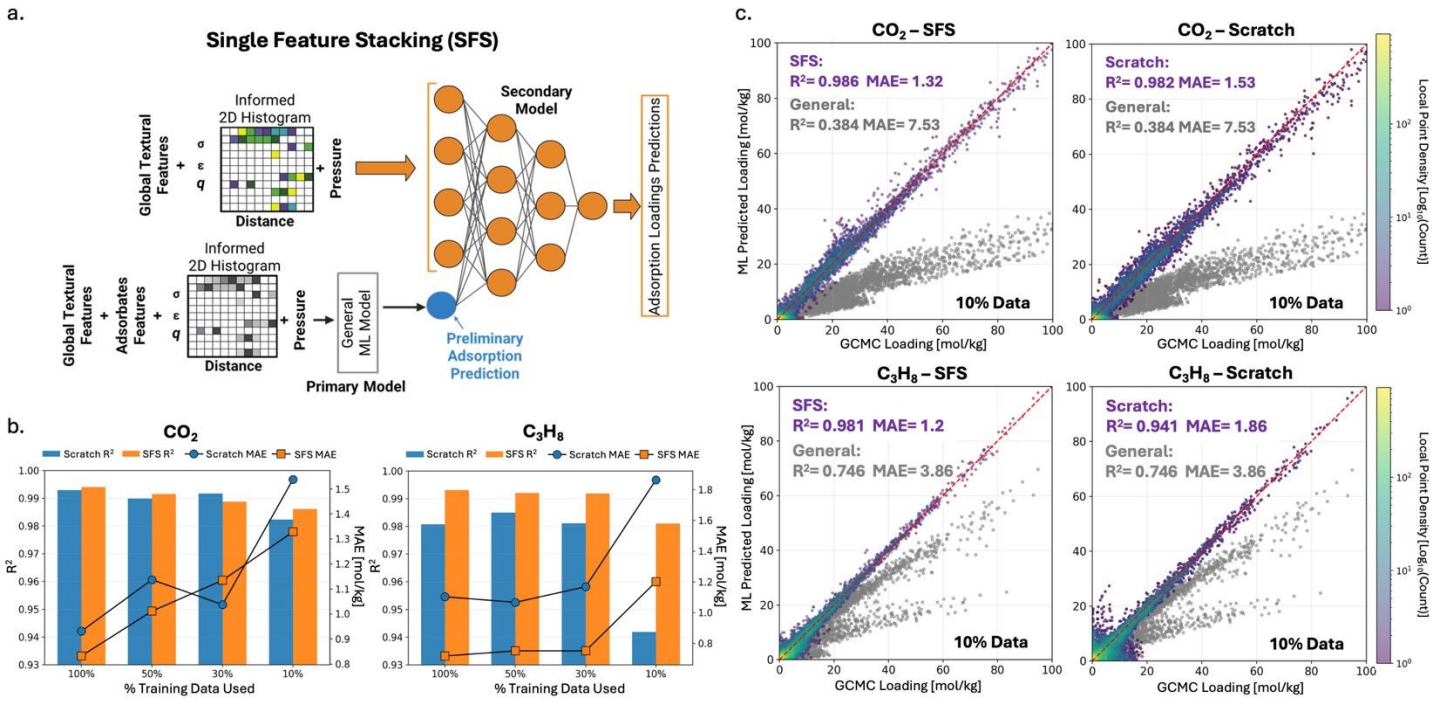

Figure 4: ML prediction for CH_4 and CO_2 adsorption loading using single-feature stacking (SFS). a) Schematic of the SFS approach. Predictions from the general (primary) model are used as an additional input feature to train a specialized (secondary) model for the target adsorbate, which still uses 2D-IPHs as part of the inputs. b) R^2 values (bars) and mean absolute error (MAE; lines) for model predictions as a function of the fraction of single-adsorbate training data used to train scratch (orange) and SFS (blue) models for CO_2 and C_3H_8 (100% = 12,533 datapoints). c) ML prediction versus GCMC adsorption loadings for CO_2 (top) and C_3H_8 (bottom). The "as is" general (primary) model (gray points) is used as reference in each plot. The results for the scratch model predictions (right plots) and the SFS model predictions (left plots), when trained on 10% of the available data (i.e., 1,253 data points) are shown colored by point density. R^2 values and MAE are reported as insets.

Table 2. Number of MOFs in the ML top-*n* set that are in the actual (i.e., GCMC) top-*n* set, with the top-*n* sets constructed based on MOF CO_2 and C_3H_8 adsorption loading at a fugacity of 1 bar. Results presented as a function of data (%) availability for training.

CO_2				
Model	**Data (%)**	**n = 100**	**n = 50**	**n = 20**
SFS	100	54	23	4
Scratch		50	17	5
SFS	50	52	24	4
Scratch		40	18	5
SFS	30	35	12	2
Scratch		36	15	4
SFS	10	35	18	3
Scratch		29	12	2
C_3H_8				
Model	**Data (%)**	**n = 100**	**n = 50**	**n = 20**
SFS	100	80	36	13
Scratch		58	21	6
SFS	50	78	37	10
Scratch		69	26	8
SFS	30	86	39	9
Scratch		68	27	8
SFS	10	71	30	8
Scratch		31	7	1

Still, even at this data availability level, the scratch models for both CO_2 and C_3H_8 maintained R^2 values above 0.94, establishing that the 2D-IPHs preserve sufficient MOF information useful for adsorption loading predictions, even in data-limited scenarios. Nevertheless, the SFS models exhibited greater robustness to data scarcity, and tended to outperform the corresponding scratch models, although more clearly so with respect to MAE (especially in the C_3H_8 case), and at 10% data availability (i.e., 1,235 data points). For the sake of visualization, parity plots for the predictions for C_3H_8 and CO_2 scratch and SFS models at 10% data availability are shown in **Fig. 4c**. The improvement of both scratch and SFS model over the general model (gray points) is evident. However, the differences between the SFS and scratch models are more visually subtle despite a 14-35 % improvement in MAE by the SFS over the scratch models (for CO_2 and C_3H_8, respectively). Thus, as done with the general model in Section 3.1, we decided to compare the SFS and scratch models in a practical screening scenario.

For the practical screening performance, we quantified the ability of the models to recover top-performing MOFs for CO_2 and C_3H_8 adsorption at 1 bar. This low-pressure condition is more challenging than the high-pressure condition used in **Table 1**, but may be relevant to bridge CO_2 capture and

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

compression steps in some configurations for direct-air capture (DAC),[72] and to design C_3H_8 separation from light hydrocarbon streams. **Table 2** shows that across all data availability levels, SFS models consistently identify a larger fraction of the true high-performing MOFs than the corresponding scratch models. This behavior is better observed for C_3H_8 where, for instance, even at 100% data availability, the SFS model captures 38% more top-100 MOF than the corresponding scratch model—the SFS model recovers 80% of the top-100 MOFs). Across all data availability levels for this molecule, the SFS model recovers ~52%, ~122% and ~214% more MOFs from the top-100, top-50, and top-20, respectively. These results demonstrate that 2D-IPHs-based SFS improves not only pointwise predictive accuracy but also ranking robustness, a key criterion in hierarchical HTCS applications where identifying a small subset of top candidates is often more important than minimizing global error metrics. For CO_2, the advantage is more modest since SFS recovers ~14% and ~25% more top-performing MOFs at top-100 and top-50, respectively, with no consistent edge at top-20. This contrast suggests that CO_2 adsorption presents a more tractable learning problem than C_3H_8 adsorption, at least at the conditions and data levels examined here.

3.3. Henry's constant prediction. K_H has been used for screening materials based on affinity to molecules of interest.[73] Usually as an early filter in hierarchical HTCS, leveraging their low calculation cost compared to adsorption simulations.[74–78] However, while K_H calculations are less computationally intensive, obtaining K_H is not a trivial exercise once the number of MOFs start to surpass a few hundred thousands. An important consideration given the overwhelming size of the MOF design space. Notably, in ML endeavors, K_H has been most commonly used as part of ML inputs for adsorption loading prediction,[79,80] with studies demonstrating that its use can substantially improve prediction accuracy.[80] However, fewer ML efforts have focused on the prediction of K_H itself.[33,81,82] The challenge with ML K_H prediction is that K_H is a dilute regime property, and at this regime adsorption is controlled by finer structural and chemical

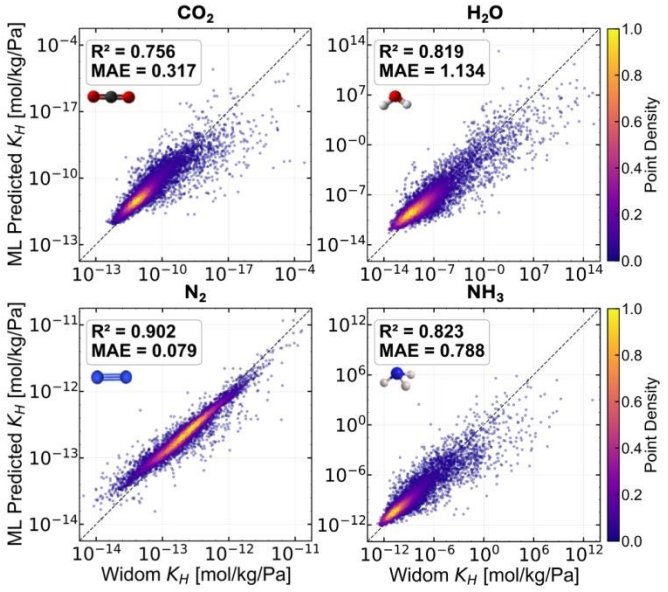

Fig. 5. ML predicted versus simulated Henry's constant (K_H) values for CO_2, H_2O, N_2, and NH_3. Color indicates the point density within each dataset. Models were trained to predict $\log_{10}(K_H)$, thus R^2 values and mean absolute error (MAE) of $\log_{10}(K_H)$ are reported as insets in each plot.

details of the MOF pores rather than by more "summarizing" features such as pore volume or surface area. Based on all the above we considered K_H prediction a good test for the 2D-IPHs MOF representation. Given the much more tractable dataset, for the 2D-IHPs of the MOFs in the K_H dataset, we used the complete (flattened) 2D-IHPs as input to the model. Our best model architecture, shared by the scratch and TL models, consisted of a MLP with three hidden layers with 128 nodes each. Further details on the MLP training can be found as Supplementary Information.

The ML model predictions on the 7,990 MOFs in the test set for this task are summarized in **Fig. 5** for N_2, CO_2, H_2O and NH_3. These molecules are expected to pick up on different aspects of the 2D-IPHs as they provide different levels of K_H dependence on dispersion and electrostatic interactions. For instance, due to their different quadrupole/dipole moments, dependence on electrostatics is low for N_2, middle for CO_2, and high for H_2O and NH_3. As K_H spans values across a wide range of orders of magnitude, we focused on the prediction of $\log(K_H)$ as usually done in the literature.[33,81,83] The mean absolute error (MAE) for $\log(K_H)$ for these molecules ranged from 0.07 to 1.11, the ratio between the mean absolute deviation to the mean (MAD) and MAE ranged between 2.4 and 3.8, whereas R^2 ranged from 0.76 to 0.90.

The best K_H predictions were achieved for the N_2 model (R^2 = 0.90, MAE = 0.08 , MAD/MAE = 3.8). This observation is consistent with persistent findings that adsorption behavior can be more easily learned and predicted for molecules in which this phenomenon is primarily dominated by dispersion interactions.[81,83] The models developed for NH_3, H_2O, and CO_2 exhibit comparable predictive performance based on the abovementioned metrics, which is consistent with the added role that electrostatic interactions play in the adsorption of these molecules. As hinted above, NH_3 and H_2O possess dipole moments that N_2 does not, and while CO_2 also lacks a dipole moment, its quadrupole moment is almost double that of N_2.

To place the predictive performance of our K_H models in context, we compare against representative studies in the literature. Such inspection shows that K_H prediction appears to be an inherently more demanding regression task than high-pressure adsorption loading prediction. Similar to us, Lin and coworkers[33] also worked with single-molecule datasets based on calculated (log of) K_H's in a distribution of 15,415 materials (MOFs and zeolites). These authors used (ε, σ, q)-embedding 3DCNN material representations as model input, yielding a root-mean squared error (RMSE) of 0.88 and 0.28 for CO_2 and CH_4, respectively, these results are consistent with our observations that predictions involving electrostatics are more difficult. While recognizing that we used a larger dataset, we note that the RMSE for our K_H prediction models in the CO_2 and N_2 cases are lower at 0.53 and 0.12. In this case, we compare N_2 with CH_4 due to their similar characteristics (small, non-polar, near-spherical). In a different approach to predict (log of) K_H in 45 hydrocarbon-based CHNOPS adsorbates,

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Sholl and coworkers[81,83] used multi-molecule datasets and combined four atomic-property radial distribution functions (AP-RDF) with energy-histograms (or variations thereof) generated with a CH_4 probe to represent the MOF. These authors leverage the adsorbate aggregation approach (see Section 3.1), worked with different molecules to us, and we cannot decouple the importance of the (chemistry-based) AP-RDFs from that of the energy histograms in their models. However, their range of R^2 values (0.67 to 0.97) is comparable (if with a higher upper bound) to ours (again, 0.76 to 0.90). Similar to our observations, their prediction accuracy depended on the characteristics of the molecules. Based on the above observations, the 2D-IPH-leveraging models herein appear competitive with current approaches in the field for K_H prediction.

We believe the above highlights the informativeness of the 2D-IPHs despite their simplicity. But as done in preceding sections, we now contextualize the performance of the models trained herein in practical terms. We find that if these models were to be used in a hierarchical screening to identify presumed top MOFs based on their affinity for these molecules, one would find that the top-800 MOFs according to the ML model would contain between 569 and 709 MOFs (71% to 88%) of the true top-800 (**Table 3**). This top MOF identification performance follows the decreasing order N_2, H_2O, NH_3, CO_2 once again being higher for the molecule whose adsorption is largely controlled by dispersion interactions. The top-800 roughly corresponds to the 90th percentile for the 7,990-MOF test set for K_H, which makes this identification endeavor analogous to finding the top-100 for the 1,029-MOF test for adsorption. When going for a more stringent effort, such as identifying the 97th percentile (top-200) MOFs, the identification success rate remains above 50% even for the worst case (CO_2). Overall, the numbers in **Table 3** indicate top-*n* recovery performances comparable or better than those reported in **Table 2** for low-pressure CO_2 adsorption. Accordingly, the 2D-IPHs facilitate models promising for hierarchical MOF screening based on adsorbate affinity for the MOF.

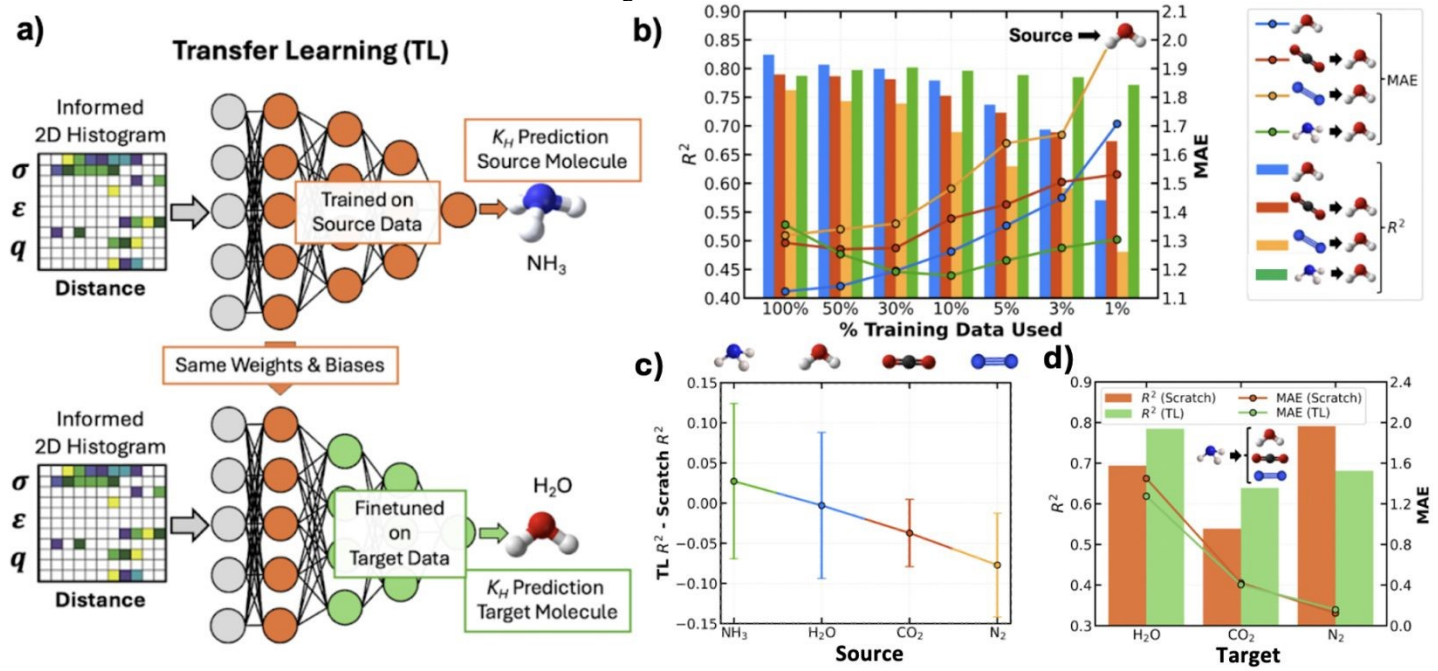

Figure 6. ML prediction of the (log of) Henry's coefficient ($\log_{10}(K_H)$) for multiple molecules using transfer learning (TL). a) Schematic of the inductive TL strategy using the prediction of a source task (e.g., NH_3; top) to predict a target task (e.g., H_2O; bottom). b) R^2 values (bars) and mean absolute error (MEA; lines) performance for TL models using CO_2, N_2, and NH_3 as source molecules to predict H_2O, relative to scratch models trained directly on H_2O data under varying levels of data availability. Schematic representation of the source-target pairs is included. c) Relative change in R^2 between TL and scratch models for each source-target molecule pair as a function of source molecules. Error bars represent the standard deviation of the metric. d) Comparison of predictive performance for H_2O, CO_2, and N_2 using NH_3 as source molecule, showing R^2 values (bars) and mean squared error (MSE; lines) for TL (green) and corresponding scratch (orange) models. Panels c and d are results for 3% of the original training dataset, hence corresponding to 958 MOFs.

Table 3. Number of MOFs in the ML top-*n* set that are in the actual (i.e., simulation) top-*n* set, with the top-*n* sets constructed based on MOF Henry's constant.

Adsorbate	n = 800	n = 400	n = 200
N_2	709	335	158
H_2O	649	291	120
NH_3	607	265	116
CO_2	569	245	104

3.4 Transfer learning. An emerging approach to facilitate data-efficient training of ML models to predict adsorption is transfer learning (TL).[84–86] Previous works indicate that the choice of MOF descriptors can either facilitate or hinder TL,[85,87] thus we considered a TL exercise as another good test for the 2D-IPHs MOF representation. Thus, in this section, we explore TL, each time from one of the models trained in Section 3.3 to predict $\log(K_H)$ for one of N_2, CO_2, H_2O and NH_3 to new models to predict $\log(K_H)$ for the remaining three molecules. We performed inductive TL by having the source and target models

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

share the same configuration of nodes and layers, as well as having the target model retain the same weights and biases as the source model in some layers (**Fig. 6a**). This is an exercise colloquially referred to as "freezing layers." We found that freezing just one layer yielded significantly better target models than when freezing two or three layers (see **Section S5**). This observation illustrates that while freezing layers is what enables knowledge transfer, the target models still need to be given sufficient "breathing room" (i.e., retrainable layers) to readapt to the new (target) task.

As brought up for SFS, data-abundant scenarios can make TL unnecessary, so we examined TL at different data availability levels ranging from 100% to 1% with respect to the data available in Section 3.3. For all data availability levels, a baseline for comparison was established by training a corresponding model from scratch. For all molecules, the behavior of the scratch models was qualitatively similar as data availability decreased. Namely, accuracy declines slowly initially, and then rapidly once below a certain critical data availability level. For instance, for H_2O, this critical level is ~10% of the original data (~3,196 MOFs), below which R^2 for scratch models started to significantly drop below ~0.8 (**Fig. 6b**). For all molecules, with data availability above the critical level, TL only yielded models that at best (with a suitable source task) had comparable accuracy to the corresponding scratch model. This observation is consistent with the increasingly held view that when data is abundant, sophisticated ML approaches are not necessary to achieve high predictive performance.[88–90]

With a suitable source task, however, TL can make the target models quite resilient to declines in data availability, resulting in significant impact on model accuracy as data becomes scarce. For instance, with K_H prediction for NH_3 as the source task, the target models for K_H prediction for H_2O show an R^2 declining only from ~0.79 to ~0.77 as data availability decreased from 100% to 1% (~318 MOFs). Thus, at the latter data availability level TL clearly outperforms the scratch model, whose R^2 was only ~0.57 (i.e., 26 % less than with TL) (**Fig. 6b**). Reflecting the importance of the source task, less resilience for the target models for K_H prediction for H_2O was achieved with K_H prediction for CO_2 and N_2 as the source tasks. To the point that TL from K_H prediction for N_2 to prediction for H_2O never yielded a target model outperforming the scratch (**Fig. 6b**).

Table 4. Number of MOFs in the ML top-*n* set that are in the actual (i.e., simulation) top-*n* set, with the top-*n* sets constructed based on MOF Henry's constant for N_2, CO_2, H_2O and NH_3. Results for the scratch and TL models at 3% data availability (958 MOFs) and NH_3 Henry's constant prediction as the source task (except when NH_3 was the target task, in which case H_2O Henry's constant prediction was used instead).

Model	Data (%)	n = 800	n = 400	n = 200
$NH_3 \rightarrow H_2O$	3	624	274	110
Scratch		576	252	102
$H_2O \rightarrow NH_3$	3	592	246	105
Scratch		530	191	78
$NH_3 \rightarrow CO_2$	3	496	207	92
Scratch		500	192	75
$NH_3 \rightarrow N_2$	3	586	247	111
Scratch		626	278	125

Examining what constitutes a suitable source task, we find two factors to be at play: *i)* the inherent informativeness of the source task (i.e., how much information the source task forces the model to extract from the histograms), and *ii)* the similarity of the target and source tasks (i.e., to what extent they demand the same kind of information from the histograms). Regarding "*i*", comparing the average R^2 improvement for target models with respect to the corresponding scratch models reveals the prediction of K_H for NH_3 to be the most inherently informative task (see **Fig. 6c** for the 3% data availability case). To be sure, the average R^2 improvement can vary significantly, and all source tasks can have target tasks for which they yield a positive average R^2 improvement. However, only the prediction of K_H for NH_3 leads to clearly positive average improvement in R^2. Further cementing the importance of inherent informativeness, note that it is dramatically more effective to do TL from NH_3 to N_2 than from N_2 to NH_3 (**Table S12 – S13**), even though task similarity is identical in both cases. Interestingly, this is the case even though the source models to predict K_H for NH_3 generally have lower R^2 than the source models to predict K_H for N_2. We believe this is consistent with NH_3 adsorption being more complex than N_2 adsorption, and thus forcing the source model to extract more information out of the histograms.

Regarding "*ii*", notice that while prediction of K_H for NH_3 is the most informative task, it is more successful the more similar to it the target task is. For instance, at 3% data availability, the most successful target models are for the prediction of K_H for H_2O and CO_2 (**Fig. 6d**). The target model for the prediction of K_H for H_2O (CO_2) attained an R^2 of 0.79 (0.64), corresponding to a 13 % (16 %) improvement over the corresponding scratch model. This observation can be understood based on NH_3, H_2O and CO_2 adsorption all being greatly driven by electrostatic interactions. Note that while the target model for the prediction of K_H for N_2 may have somewhat higher R^2 (0.68) than for the CO_2 case, this is simply due to the prediction of K_H for N_2 being a simpler task, with the N_2 scratch model still outperforming the TL model in this case.

We end with a practical test for the TL models. Specifically, assessing the ability of these models to identify the top-*n* performing MOFs across different tiers (*n* = 800, 400, and 200) as done in Section 3.3. Following up from **Fig. 6b-d,** we focus on TL models at 3% data availability, using NH_3 K_H prediction as the source task (except when NH_3 K_H prediction was the target task, in which case H_2O K_H prediction was the source task) (**Table 4**). Trends on top-*n* identification performance agree with those for R^2 and MAE, as performance was better the more similar the source and target tasks were.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Accordingly, the best TL performances were for the $NH_3 \rightarrow H_2O$ and $H_2O \rightarrow NH_3$ TL scenarios (7 to 26% more top-*n* MOFs detected than the scratch model for n=800, 400, 200). In contrast, in the $NH_3 \rightarrow N_2$ case, the TL model never outperformed the scratch model. As for absolute identification efficacy, the use of TL across cases listed in **Table 4** resulted in success rates for top-800 identification ranging from 62% to 73% across the studied adsorbates. For the more stringent top-200 identification, this success rate ranged from 46% to 56%. As with R^2 and MAE, we find the best TL scenarios to significantly slow down the decline in model performance. For instance, for the $NH_3 \rightarrow H_2O$ and $H_2O \rightarrow NH_3$ TL models, the top-800 identification success rate was kept in the 74-78% range, whereas for the models at full data availability (Section 3.3) was a somewhat higher 76-81%. For the more stringent top-200 identification case, the analogous success numbers are 53-55% for TL vs. 58-60% for models at full data availability.

CONCLUSIONS

Here we demonstrated that two-dimensional interaction-parameter histograms (2D-IPHs) provide a simple, inexpensive, physics-based MOF representation for adsorption learning. These 2D-IPHs capture the distributions of adsorption sites, when the latter are characterized by their electrostatic and dispersion interaction potential features. Physics-based MOF representations offer a pathway toward adsorption models that are transferable across materials, can exploit alchemical datasets, and are more resilient to inaccuracies in molecular simulation models used to generate data. Across a variety of prediction tasks, 2D-IPHs consistently enabled models to accurately learn adsorption properties, and gain hierarchical screening-relevant, ranking fidelity. For instance, 2D-IPHs facilitated a model for prediction of full adsorption isotherms for multiple unseen real molecules within the small, non-polar, near-spherical class. Then, for adsorbates that fall outside the scope of the above model, 2D-IPHs supported efficient specialization of a new model through single-feature stacking. Additionally, 2D-IPHs facilitated prediction of Henry's constants across molecules spanning dispersion- and electrostatics-dominated adsorption, while also showing the facilitation of inductive transfer learning when training data were severely limited.

Taken together, these results indicate that 2D-IPHs present highly appealing features such as: *i)* scalability as indicated by its application to large MOF unit cells and multimillion-point datasets; *ii)* applicability across adsorption regimes ranging from the dilute regime to the pore saturation-regime as indicated by prediction for both Henry's constants and full adsorption isotherms; and *iii)* effectiveness with data-efficient training strategies (here assessed in single feature stacking and transfer learning scenarios), and *iv)* simplicity, which may increase 2D-IPHs appeal to the broader adsorption community, including experimentalists. Accordingly, we anticipate that this representation will be valuable to develop computational material discovery pipelines for adsorption-based application, not only for MOFs, but also for other nanoporous materials. Furthermore, we envision that improvements to the 2D-IPHs can be done while retaining the same construction philosophy, such as the parameter-gradient statistics analogous to the energy-gradients statistics that have been used to improve the informativeness of energy histogram representations.

SUPPORTING INFORMATION

The Supporting Information is available free of charge.

Additional details on computational infrastructure and data availability, machine learning architectures and training protocols; data processing and hyperparameter exploration; model selection and reproducibility procedures; fugacity grids used for GCMC simulations; supplementary learning curves; optimal network configurations; and complementary analyses for adsorption loading, K_H prediction, single-feature stacking, and transfer-learning studies. (PDF)

ACKNOWLEDGEMENTS

This work was funded through NSF grants CBET-2450909 and OAC-2118201 (HDR: Institute for Data-Driven Dynamics Design). R.D. contributions to this project were funded through NSF REU Grant DMR 1950924. Calculations were made possible thanks to the supercomputing cluster Mio at the Colorado School of Mines.

REFERENCES

(1) Thommes, M.; Schlumberger, C. Characterization of Nanoporous Materials. *Annual Review of Chemical and Biomolecular Engineering* **2021**, *12*, 137–162. https://doi.org/10.1146/annurev-chembioeng-061720-081242.

(2) Ren, E.; Guilbaud, P.; Coudert, F.-X. High-Throughput Computational Screening of Nanoporous Materials in Targeted Applications. *Digital Discovery* **2022**, *1* (4), 355–374. https://doi.org/10.1039/D2DD00018K.

(3) Petit, C. Present and Future of MOF Research in the Field of Adsorption and Molecular Separation. *Current Opinion in Chemical Engineering* **2018**, *20*, 132–142. https://doi.org/10.1016/j.coche.2018.04.004.

(4) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341* (6149), 1230444. https://doi.org/10.1126/science.1230444.

(5) Lee, S.; Kim, B.; Cho, H.; Lee, H.; Lee, S. Y.; Cho, E. S.; Kim, J. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **2021**, *13* (20), 23647–23654. https://doi.org/10.1021/acsami.1c02471.

(6) Ongari, D.; Talirz, L.; Smit, B. Too Many Materials and Too Many Applications: An Experimental Problem Waiting for a Computational Solution. *ACS Cent. Sci.* **2020**, *6* (11), 1890–1900. https://doi.org/10.1021/acscentsci.0c00988.

(7) Park, J.; Kim, H.; Kang, Y.; Lim, Y.; Kim, J. From Data to Discovery: Recent Trends of Machine Learning in Metal–Organic Frameworks. *JACS Au* **2024**, *4* (10), 3727–3743. https://doi.org/10.1021/jacsau.4c00618.

(8) Gómez-Gualdrón, D. A.; Vilas, T. G. de; Ardila, K.; Fajardo-Rojas, F.; Pak, A. J. Machine Learning to Design Metal–Organic Frameworks: Progress and Challenges from a Data Efficiency Perspective. *Mater. Horiz.* **2025**. https://doi.org/10.1039/D5MH01467K.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

(9) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Klontzas, E.; Froudakis, G. A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials. *J. Phys. Chem. A* **2019**, *123* (28), 6080–6087. https://doi.org/10.1021/acs.jpca.9b03290.

(10) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal–Organic Frameworks. *Mol. Syst. Des. Eng.* **2019**, *4* (1), 162–174. https://doi.org/10.1039/C8ME00050F.

(11) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145* (5), 2958–2967. https://doi.org/10.1021/jacs.2c11420.

(12) Cui, J.; Wu, F.; Zhang, W.; Yang, L.; Hu, J.; Fang, Y.; Ye, P.; Zhang, Q.; Suo, X.; Mo, Y.; Cui, X.; Chen, H.; Xing, H. Direct Prediction of Gas Adsorption via Spatial Atom Interaction Learning. *Nat Commun* **2023**, *14* (1), 7043. https://doi.org/10.1038/s41467-023-42863-6.

(13) Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16* (2), 1271–1283. https://doi.org/10.1021/acs.jctc.9b00940.

(14) 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* (15), 7681–7689. https://doi.org/10.1021/jp4006422.

(15) Mohamed, S. A.; Jiang, J. Computational Design of Metal–Organic Frameworks with Triangular Adsorbaphores for Highly Selective Adsorption of m-Xylene. *Nanoscale* **2025**, *17* (37), 21546–21553. https://doi.org/10.1039/D5NR02310F.

(16) Sung, I.-T.; Cheng, Y.-H.; Hsieh, C.-M.; Lin, L.-C. Machine Learning for Gas Adsorption in Metal–Organic Frameworks: A Review on Predictive Descriptors. *Ind. Eng. Chem. Res.* **2025**, *64* (4), 1859–1875. https://doi.org/10.1021/acs.iecr.4c03500.

(17) 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* (27), 14095–14105. https://doi.org/10.1021/jp404287t.

(18) 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* (12), 4459–4475. https://doi.org/10.1021/acs.chemmater.5b01475.

(19) Krishnapriyan, A. S.; Haranczyk, M.; Morozov, D. Topological Descriptors Help Predict Guest Adsorption in Nanoporous Materials. *J. Phys. Chem. C* **2020**, *124* (17), 9360–9368. https://doi.org/10.1021/acs.jpcc.0c01167.

(20) Shi, K.; Li, Z.; Anstine, D. M.; Tang, D.; Colina, C. M.; Sholl, D. S.; Siepmann, J. I.; Snurr, R. Q. Two-Dimensional Energy Histograms as Features for Machine Learning to Predict Adsorption in Diverse Nanoporous Materials. *J. Chem. Theory Comput.* **2023**, *19* (14), 4568–4583. https://doi.org/10.1021/acs.jctc.2c00798.

(21) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301.

(22) Wang, R.; Zhong, Y.; Bi, L.; Yang, M.; Xu, D. Accelerating Discovery of Metal–Organic Frameworks for Methane Adsorption with Hierarchical Screening and Deep Learning. *ACS Appl. Mater. Interfaces* **2020**, *12* (47), 52797–52807. https://doi.org/10.1021/acsami.0c16516.

(23) Choudhary, K.; Yildirim, T.; Siderius, D. W.; Kusne, A. G.; McDannald, A.; Ortiz-Montalvo, D. L. Graph Neural Network Predictions of Metal Organic Framework CO2 Adsorption Properties. *Computational Materials Science* **2022**, *210*, 111388. https://doi.org/10.1016/j.commatsci.2022.111388.

(24) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294.

(25) Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for Improved Materials Property Predictions. *npj Comput Mater* **2021**, *7* (1), 185. https://doi.org/10.1038/s41524-021-00650-1.

(26) Raza, A.; Waqar, F.; Sturluson, A.; Simon, C.; Fern, X. Towards Explainable Message Passing Networks for Predicting Carbon Dioxide Adsorption in Metal-Organic Frameworks. arXiv December 2, 2020. https://doi.org/10.48550/arXiv.2012.03723.

(27) Rosen, A. S.; Fung, V.; Huck, P.; O'Donnell, C. T.; Horton, M. K.; Truhlar, D. G.; Persson, K. A.; Notestein, J. M.; Snurr, R. Q. High-Throughput Predictions of Metal–Organic Framework Electronic Properties: Theoretical Challenges, Graph Neural Networks, and Data Exploration. *npj Comput Mater* **2022**, *8* (1), 112. https://doi.org/10.1038/s41524-022-00796-6.

(28) Zhang, X.; Shang, W. Prediction of Rn, N2, and O2 Adsorption and Separation in MOFs and MOF@AC Composites via Machine Learning and Graph Neural Networks. *Ind. Eng. Chem. Res.* **2025**, *64* (50), 24227–24239. https://doi.org/10.1021/acs.iecr.5c04453.

(29) Teng, Y.; Tan, H.; Huang, W.; Shan, G. Atomic-Level Interpretable Multimodal Graph Neural Network for Predicting Carbon Dioxide Adsorption in Metal-Organic Frameworks. *Commun Phys* **2025**, *8* (1), 491. https://doi.org/10.1038/s42005-025-02399-1.

(30) Haldoupis, E.; Borycz, J.; Shi, H.; Vogiatzis, K. D.; Bai, P.; Queen, W. L.; Gagliardi, L.; Siepmann, J. I. Ab Initio Derived Force Fields for Predicting CO2 Adsorption and Accessibility of Metal Sites in the Metal–Organic Frameworks M-MOF-74 (M = Mn, Co, Ni, Cu). *J. Phys. Chem. C* **2015**, *119* (28), 16058–16071. https://doi.org/10.1021/acs.jpcc.5b03700.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

(31) 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* (4), 1477–1488. https://doi.org/10.1021/ct500094w.

(32) Sarikas, A. P.; Gkagkas, K.; Froudakis, G. E. Gas Adsorption Meets Deep Learning: Voxelizing the Potential Energy Surface of Metal-Organic Frameworks. *Sci Rep* **2024**, *14* (1), 2242. https://doi.org/10.1038/s41598-023-50309-8.

(33) Hung, T.-H.; Xu, Z.-X.; Kang, D.-Y.; Lin, L.-C. Chemistry-Encoded Convolutional Neural Networks for Predicting Gaseous Adsorption in Porous Materials. *J. Phys. Chem. C* **2022**, *126* (5), 2813–2822. https://doi.org/10.1021/acs.jpcc.1c09649.

(34) Cho, E. H.; Lin, L.-C. Nanoporous Material Recognition via 3D Convolutional Neural Networks: Prediction of Adsorption Properties. *J. Phys. Chem. Lett.* **2021**, *12* (9), 2279–2285. https://doi.org/10.1021/acs.jpclett.1c00293.

(35) Li, Z.; Bucior, B. J.; Chen, H.; Haranczyk, M.; Siepmann, J. I.; Snurr, R. Q. Machine Learning Using Host/Guest Energy Histograms to Predict Adsorption in Metal–Organic Frameworks: Application to Short Alkanes and Xe/Kr Mixtures. *The Journal of Chemical Physics* **2021**, *155* (1), 014701. https://doi.org/10.1063/5.0050823.

(36) Subraveti, S. G.; Li, Z.; Prasad, V.; Rajendran, A. Physics-Based Neural Networks for Simulation and Synthesis of Cyclic Adsorption Processes. *Ind. Eng. Chem. Res.* **2022**, *61* (11), 4095–4113. https://doi.org/10.1021/acs.iecr.1c04731.

(37) Liu, T.-W.; Fajardo-Rojas, F.; Addish, S.; Martinez, E.; Gomez-Gualdron, D. A. MOFs to Enhance Green NH3 Synthesis in Plasma Reactors: Hierarchical Computational Screening Enhanced by Iterative Machine Learning. *ACS Appl. Mater. Interfaces* **2024**, *16* (49), 68506–68519. https://doi.org/10.1021/acsami.4c11396.

(38) Niyongabo Rubungo, A.; Fajardo-Rojas, F.; Gómez-Gualdrón, D. A.; Dieng, A. B. Highly Accurate and Fast Prediction of MOF Free Energy via Machine Learning. *J. Am. Chem. Soc.* **2025**, *147* (52), 48035–48045. https://doi.org/10.1021/jacs.5c13960.

(39) 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. *Crystal Growth & Design* **2017**, *17* (11), 5801–5810. https://doi.org/10.1021/acs.cgd.7b00848.

(40) Anderson, R.; Gómez-Gualdrón, D. A. Increasing Topological Diversity during Computational "Synthesis" of Porous Crystals: How and Why. *CrystEngComm* **2019**, *21* (10), 1653–1665. https://doi.org/10.1039/C8CE01637B.

(41) Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T. Extension of the Universal Force Field to Metal–Organic Frameworks. *J. Chem. Theory Comput.* **2014**, *10*. https://doi.org/10.1021/ct400952t.

(42) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19. https://doi.org/10.1006/jcph.1995.1039.

(43) Anderson, R.; Gómez-Gualdrón, D. A. Deep Learning Combined with IAST to Screen Thermodynamically Feasible MOFs for Adsorption-Based Separation of Multiple Binary Mixtures. *J Chem Phys* **2021**, *154* (23), 234102. https://doi.org/10.1063/5.0048736.

(44) Eggimann, B. L.; Sun, Y.; DeJaco, R. F.; Singh, R.; Ahsan, M.; Josephson, T. R.; Siepmann, J. I. Assessing the Quality of Molecular Simulations for Vapor–Liquid Equilibria: An Analysis of the TraPPE Database. *J. Chem. Eng. Data* **2020**, *65* (3), 1330–1344. https://doi.org/10.1021/acs.jced.9b00756.

(45) Potoff, J. J.; Siepmann, J. I. Vapor–Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen. *AIChE Journal* **2001**, *47* (7), 1676–1682. https://doi.org/10.1002/aic.690470719.

(46) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102* (14), 2569–2577. https://doi.org/10.1021/jp972543+.

(47) Shah, M. S.; Siepmann, J. I.; Tsapatsis, M. Transferable Potentials for Phase Equilibria. Improved United-Atom Description of Ethane and Ethylene. *AIChE Journal* **2017**, *63* (11), 5098–5110. https://doi.org/10.1002/aic.15816.

(48) 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* (7), 2217–2223. https://doi.org/10.1039/C2SC01097F.

(49) 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* (27), 9976–9979. https://doi.org/10.1021/jp803753h.

(50) Darkrim, F.; Levesque, D. Monte Carlo Simulations of Hydrogen Adsorption in Single-Walled Carbon Nanotubes. *J. Chem. Phys.* **1998**, *109* (12), 4981–4984. https://doi.org/10.1063/1.477109.

(51) Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q. Evaluating Topologically Diverse Metal–Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci.* **2016**, *9* (10), 3279–3289. https://doi.org/10.1039/C6EE02104B.

(52) Getman, R. B.; Bae, Y.-S.; Wilmer, C. E.; Snurr, R. Q. Review and Analysis of Molecular Simulations of Methane, Hydrogen, and Acetylene Storage in Metal–Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 703–723. https://doi.org/10.1021/cr200217c.

(53) Moghadam, P. Z.; Fairen-Jimenez, D.; Snurr, R. Q. Efficient Identification of Hydrophobic MOFs: Application in the Capture of Toxic Industrial Chemicals. *J. Mater. Chem. A* **2015**, *4* (2), 529–536. https://doi.org/10.1039/C5TA06472D.

(54) Ghosh, P.; Kim, K. C.; Snurr, R. Q. Modeling Water and Ammonia Adsorption in Hydrophobic Metal–Organic Frameworks: Single Components and Mixtures.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

J. Phys. Chem. C **2014**, *118* (2), 1102–1110. https://doi.org/10.1021/jp410758t.

(55) 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* (1), 365–376. https://doi.org/10.1021/acs.jctc.7b00841.

(56) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: Molecular Simulation Software for Adsorption and Diffusion in Flexible Nanoporous Materials. *Molecular Simulation* **2016**, *42* (2), 81–101. https://doi.org/10.1080/08927022.2015.1010082.

(57) Gowers, R. J.; Farmahini, A. H.; Friedrich, D.; Sarkisov, L. Automated Analysis and Benchmarking of GCMC Simulation Programs in Application to Gas Adsorption. *Molecular Simulation* **2018**, *44* (4), 309–321. https://doi.org/10.1080/08927022.2017.1375492.

(58) Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57* (1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

(59) Panagiotopoulos, A. Z. Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble. *Molecular Physics* **2002**, *100* (1), 237–246. https://doi.org/10.1080/00268970110097866.

(60) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials. *J. Phys. Chem. C* **2020**, *124* (13), 7117–7126. https://doi.org/10.1021/acs.jpcc.9b10766.

(61) 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* (17), 3056–3060. https://doi.org/10.1021/jz501331m.

(62) Langmuir, I. THE ADSORPTION OF GASES ON PLANE SURFACES OF GLASS, MICA AND PLATINUM. *J. Am. Chem. Soc.* **1918**, *40* (9), 1361–1403. https://doi.org/10.1021/ja02242a004.

(63) Swenson, H.; Stadie, N. P. Langmuir's Theory of Adsorption: A Centennial Review. *Langmuir* **2019**, *35* (16), 5409–5426. https://doi.org/10.1021/acs.langmuir.9b00154.

(64) Brunauer, S.; Emmett, P. H.; Teller, E. Adsorption of Gases in Multimolecular Layers. *J. Am. Chem. Soc.* **1938**, *60* (2), 309–319. https://doi.org/10.1021/ja01269a023.

(65) Dubinin, M. M.; Radushkevich, L. V. Equation of the Characteristic Curve of Activated Charcoal. *Proceedings of the Academy of Sciences of the USSR, Physical Chemistry Section* **1947**, *55*, 331–333.

(66) Gharagheizi, F.; Tang, D.; Sholl, D. S. Selecting Adsorbents to Separate Diverse Near-Azeotropic Chemicals. *J. Phys. Chem. C* **2020**, *124* (6), 3664–3670. https://doi.org/10.1021/acs.jpcc.9b10955.

(67) Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. E. A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142* (8), 3814–3822. https://doi.org/10.1021/jacs.9b11084.

(68) Fanourgakis, G. S.; Gkagkas, K.; Froudakis, G. Introducing Artificial MOFs for Improved Machine Learning Predictions: Identification of Top-Performing Materials for Methane Storage. *The Journal of Chemical Physics* **2022**, *156* (5), 054103. https://doi.org/10.1063/5.0075994.

(69) Dong, M.; Wu, X.; Cai, W. Prediction of Methane Adsorption Isotherms in Metal–Organic Frameworks by Neural Networks: Two-Dimensional Energy Gradient Feature and Masked Learning Mechanism. *J. Phys. Chem. B* **2025**, *129* (43), 11333–11349. https://doi.org/10.1021/acs.jpcb.5c05384.

(70) Deng, Z.; Sarkisov, L. Engineering Machine Learning Features to Predict Adsorption of Carbon Dioxide and Nitrogen in Metal–Organic Frameworks. *J. Phys. Chem. C* **2024**, *128* (24), 10202–10215. https://doi.org/10.1021/acs.jpcc.4c01692.

(71) Moharreri, E.; Pardakhti, M.; Srivastava, R.; Suib, S. L. Energy–Geometry Dependency of Molecular Structures: A Multistep Machine Learning Approach. *ACS Comb. Sci.* **2019**, *21* (9), 614–621. https://doi.org/10.1021/acscombsci.9b00028.

(72) Wiegner, J. F.; Grimm, A.; Weimann, L.; Gazzani, M. Optimal Design and Operation of Solid Sorbent Direct Air Capture Processes at Varying Ambient Conditions. *Ind. Eng. Chem. Res.* **2022**, *61* (34), 12649–12667. https://doi.org/10.1021/acs.iecr.2c00681.

(73) Gonzalez, J.; Mukherjee, K.; Colón, Y. J. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data* **2023**, *68* (1), 291–302. https://doi.org/10.1021/acs.jced.2c00443.

(74) 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* (1), 5912. https://doi.org/10.1038/ncomms6912.

(75) Chung, Y. G.; Bai, P.; Haranczyk, M.; Leperi, K. T.; Li, P.; Zhang, H.; Wang, T. C.; Duerinck, T.; You, F.; Hupp, J. T.; Farha, O. K.; Siepmann, J. I.; Snurr, R. Q. Computational Screening of Nanoporous Materials for Hexane and Heptane Isomer Separation. *Chem. Mater.* **2017**, *29* (15), 6315–6328. https://doi.org/10.1021/acs.chemmater.7b01565.

(76) 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* (14), 4571–4579. https://doi.org/10.1021/acs.chemmater.8b00843.

(77) Yıldız, T.; Erucar, I. Revealing the Performance of Bio-MOFs for Adsorption-Based Uremic Toxin Separation Using Molecular Simulations. *Chemical Engineering Journal* **2022**, *431*, 134263. https://doi.org/10.1016/j.cej.2021.134263.

(78) Li, B.; Gong, S.; Cao, P.; Gao, W.; Zheng, W.; Sun, W.; Zhang, X.; Wu, X. Screening of Biocompatible MOFs

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

for the Clearance of Indoxyl Sulfate Using GCMC Simulations. *Ind. Eng. Chem. Res.* **2022**, *61* (19), 6618–6627. https://doi.org/10.1021/acs.iecr.2c00283.
(79) Zhang, X.; Zheng, Q.-R.; He, H.-Z. Machine-Learning-Based Prediction of Hydrogen Adsorption Capacity at Varied Temperatures and Pressures for MOFs Adsorbents. *Journal of the Taiwan Institute of Chemical Engineers* **2022**, *138*, 104479. https://doi.org/10.1016/j.jtice.2022.104479.
(80) Orhan, I. B.; Le, T. C.; Babarao, R.; Thornton, A. W. Accelerating the Prediction of CO2 Capture at Low Partial Pressures in Metal-Organic Frameworks Using New Machine Learning Descriptors. *Commun Chem* **2023**, *6* (1), 214. https://doi.org/10.1038/s42004-023-01009-x.
(81) Yu, X.; Choi, S.; Tang, D.; Medford, A. J.; Sholl, D. S. Efficient Models for Predicting Temperature-Dependent Henry's Constants and Adsorption Selectivities for Diverse Collections of Molecules in Metal–Organic Frameworks. *J. Phys. Chem. C* **2021**, *125* (32), 18046–18057. https://doi.org/10.1021/acs.jpcc.1c05266.
(82) Choi, S.; Sholl, D. S.; Medford, A. J. D–MOPH–25: Diverse MOF–Molecule Pairs for Henry's Constants Prediction*. *Mach. Learn.: Sci. Technol.* **2025**, *6* (3), 035058. https://doi.org/10.1088/2632-2153/ae0241.
(83) Choi, S.; Sholl, D. S.; Medford, A. J. Gaussian Approximation of Dispersion Potentials for Efficient Featurization and Machine-Learning Predictions of Metal–Organic Frameworks. *J. Chem. Phys.* **2022**, *156* (21), 214108. https://doi.org/10.1063/5.0091405.
(84) Kang, Y.; Park, H.; Smit, B.; Kim, J. A Multi-Modal Pre-Training Transformer for Universal Transfer Learning in Metal–Organic Frameworks. *Nat Mach Intell* **2023**, *5* (3), 309–318. https://doi.org/10.1038/s42256-023-00628-2.
(85) Ma, R.; Colón, Y. J.; Luo, T. Transfer Learning Study of Gas Adsorption in Metal–Organic Frameworks. *ACS Appl. Mater. Interfaces* **2020**, *12* (30), 34041–34048. https://doi.org/10.1021/acsami.0c06858.
(86) Cai, Z.; Li, W.; Chung, Y. G.; Li, S.; Liang, T.; Wu, T. Transfer Learning-Assisted Computational Screening of Metal-Organic Frameworks and Covalent-Organic Frameworks for the Separation of Xe/Kr Noble Gas. *Separation and Purification Technology* **2024**, *348*, 127752. https://doi.org/10.1016/j.seppur.2024.127752.
(87) Zhang, W.; Fang, Y.; Ma, Z. The Effect of Task Similarity on Deep Transfer Learning. In *Neural Information Processing*; Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.-S. M., Eds.; Springer International Publishing: Cham, 2017; pp 256–265. https://doi.org/10.1007/978-3-319-70096-0_27.
(88) Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in the Deep Learning Era; IEEE, 2017; pp 843–852.
(89) Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* **2009**, *24* (2), 8–12. https://doi.org/10.1109/MIS.2009.36.
(90) Sutton, R. *The Bitter Lesson*. https://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Molecular Systems Design & Engineering Accepted Manuscript

View Article Online
DOI: 10.1039/D6ME00034G

Open Access Article. Published on 01 June 2026. Downloaded on 6/3/2026 5:19:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Data availability

The data supporting this article have been included as part of the supplementary information (SI):

Supplementary Information: Additional details on machine learning architectures and training protocols; data processing and hyperparameter exploration; model selection and reproducibility procedures; fugacity grids used for GCMC simulations; supplementary learning curves and parity plots; optimal network configurations; and complementary analyses for adsorption loading, K_H prediction, single-feature stacking, and transfer-learning studies. (PDF)

Additional codes and files to reproduce results can be found at https://osf.io/vu3hn/overview?view_only=cc825a2815bd43529214c8c32be82519

Molecular Systems Design & Engineering Accepted Manuscript