JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

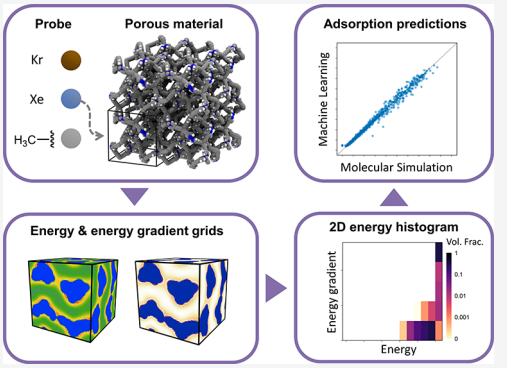
Two-Dimensional Energy Histograms as Features for Machine Learning to Predict Adsorption in Diverse Nanoporous Materials

Kaihang Shi, Zhao Li, Dylan M. Anstine, Dai Tang, Coray M. Colina, David S. Sholl, J. Ilja Siepmann, and Randall Q. Snurr*

Cite This: https://doi.org/10.1021/acs.jctc.2c00798 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A major obstacle for machine learning (ML) in chemical science is the lack of physically informed feature representations that provide both accurate prediction and easy interpretability of the ML model. In this work, we describe adsorption systems using novel two-dimensional energy histogram (2D-EH) features, which are obtained from the probe-adsorbent energies and energy gradients at grid points located throughout the adsorbent. The 2D-EH features encode both energetic and structural information of the material and lead to highly accurate ML models (coefficient of determination $R^2 \sim 0.94–0.99$) for predicting single-component adsorption capacity in metal−organic frameworks (MOFs). We consider the adsorption of spherical molecules (Kr and Xe), linear alkanes with a wide range of aspect ratios (ethane, propane, *n*-butane, and *n*-hexane), and a branched alkane (2,2-dimethylbutane) over a wide range of temperatures and pressures. The interpretable 2D-EH features enable the ML model to learn the basic physics of adsorption in pores from the training data. We show that these MOF-data-trained ML models are transferrable to different families of amorphous nanoporous materials. We also identify several adsorption systems where capillary condensation occurs, and ML predictions are more challenging. Nevertheless, our 2D-EH features still outperform structural features including those derived from persistent homology. The novel 2D-EH features may help accelerate the discovery and design of advanced nanoporous materials using ML for gas storage and separation in the future.



1. INTRODUCTION

The ability of nanoporous materials to provide tailored pore environments for specific molecules has made them strong candidates for applications in energy storage, chemical separations, sensing, and catalysis. Porous carbons such as activated carbon and biochar have been used for decades, for example to remove organic pollutants from water.[1] More recently, there has been extensive research on metal−organic frameworks (MOFs),[2] a class of porous crystalline materials assembled from metal nodes and organic linkers, due to their chemical and structural tunability. Many MOFs have extremely high gravimetric surface area and pore volume, making them attractive for many adsorption application. Porous polymers are purely organic adsorbents that can also be tuned by selection of the constituent monomers to adsorb desired molecules.[3] They combine the adsorption properties of other nanoporous materials with the processability of polymeric materials.[4]

For applications of nanoporous materials such as gas storage or separation, the adsorption uptake of molecules of interest at relevant conditions (e.g., temperature and pressure) is the primary property of interest. Efficient approaches to accurately predict adsorption uptake for a wide range of adsorbate−adsorbent systems across variable thermodynamic conditions can accelerate the identification of promising materials from a large number of real and hypothetical candidates.[5−9] Monte Carlo simulations in the grand canonical ensemble[10] or Gibbs ensemble[11] are the standard method to predict adsorption phase equilibria in porous materials. Monte Carlo simulations with a suitable force field can accurately predict adsorption properties but are time-consuming if thousands of materials need to be screened. To overcome this challenge, many groups are developing machine learning (ML) approaches, where a ML model is trained on adsorption data from either simulations or experiments and then is used to predict the adsorption capacity for new candidates.[12] Such methods may be many orders of magnitude faster than "brute force" molecular simulations (Section 3.6).

A critical factor for a reliable ML model is the engineering of features (often called "descriptors"). A feature set is a machine readable, fix-sized data representation of material characteristics. Typical features to describe adsorbent materials or adsorbate−

Special Issue: Machine Learning for Molecular Simulation

Received: August 2, 2022

Downloaded via UNIV OF CALIFORNIA LOS ANGELES on February 4, 2023 at 07:52:52 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications © XXXX American Chemical Society

Table 1. Adsorption Systems Investigated in This Work[a]

Label[b]	System	T/T_c	P/P_0	Total number of data points[c]	Reference
Kr-1-273	Kr @ 1 bar,273 K	1.3	N/A	2000	36
Kr-10-273	Kr @ 10 bar,273 K	1.3	N/A	2000	36
Xe-1-273	Xe @ 1 bar,273 K	0.94	0.02	2000	36
Xe-10-273	Xe @ 10 bar,273 K	0.94	0.24	2000	36
Eth-4-298	Ethane @ 4 bar,298 K	0.98	0.08	2000	36
Eth-20-298	Ethane @ 20 bar,298 K	0.98	0.41	2000	36
Eth-40-298	Ethane @ 40 bar,298 K	0.98	0.82	2000	36
Pro-1-298	Propane @ 1 bar,298 K	0.81	0.08	2000	36
Pro-5-298	Propane @ 5 bar,298 K	0.81	0.41	2000	36
Pro-10-298	Propane @ 10 bar,298 K	0.81	0.83	2000	36
But-0.24-298	*n*-Butane@ 0.24 bar,298 K	0.70	0.07	2000	This work
But-1.2−298	*n*-Butane@ 1.2 bar,298 K	0.70	0.36	6000	This work
Hex-0.02-298	*n*-Hexane@ 0.02 bar,298 K	0.59	0.1	2000	This work
Hex-10-495	*n*-Hexane@ 10 bar,495.6 K	0.98	0.39	2000	This work
Hex-25-495	*n*-Hexane@ 25 bar,495.6 K	0.98	0.98	2000	This work
DMB-13-477	2,2-dimethylbutane @ 13 bar,477.1 K	0.98	0.5	2000	This work

[a]Critical temperatures, T_c, and saturation pressures, P_0, are available in Table S1. All simulated adsorption data are available in the SI. [b]Label has the general format of "Adsorbate abbreviation-Pressure-Temperature". [c]Total number of data points for training and testing. For each new system (Reference: "This work"), MOFs in the data set were selected randomly from the ToBaCCo database.

adsorbent systems can generally be divided into two categories: (1) structural features that capture geometric information about the pores and (2) chemical or energetic features that reflect interatomic interactions between the adsorbent and some probe adsorbate species. Common structural features include textural properties,[6,13−16] such as the geometric or helium void fraction, adsorbent density, characteristic pore sizes, and surface area. These simple features have been shown to work reasonably well in certain cases, for example, in regression tasks to predict hydrogen storage for pressure swing or pressure−temperature swing use[8,15] and methane storage at high pressure (e.g., 100 bar),[13,17] and in classification tasks to recognize high-performing MOFs for adsorbing CH_4,[13] CO_2,[14] and N_2[14] at low pressures. More advanced structural features have been developed, including Voronoi holograms,[18] a three-dimensional (3D) histogram encoding of probe-accessible fragments of the Voronoi network based on the Voronoi decomposition of a material; a barcode or image representation of material's topology derived from the persistent homology;[16,19,20] and machine generated features using a convolutional neural network from a 3D voxel[21] or crystal graph[22] representation of the material. It was found that ML models using structural features can be effective for materials with simple chemical composition, such as porous carbons[23] and all-silica zeolites.[21] For MOFs, chemical information may be necessary for accurate predictions. Common chemical or energetic features include the number density or percentage of chemical elements or moieties,[24,25] chemical properties of atoms (e.g., metallic or nonmetallic, electronegativity),[24] physical properties related to the functional groups of the MOF linkers,[26] adsorption Henry's coefficient,[27] number density of atom types (e.g., using the atom typing from molecular mechanics force fields[28]), and more recently, the string representation of a material's chemical building blocks.[29,30]

Using both structural and chemical features can enable a more accurate prediction of adsorption,[24−28,31] and incorporating cross information between structural and chemical features can further improve ML model accuracy. Fernandez et al. introduced an atomic property weighted radial distribution function (AP-RDF) as features to predict CH_4, CO_2, and N_2 adsorption.[32] The AP-RDF was designed to encode both geometric and chemical information by introducing the adsorbent atomic properties (e.g., electronegativity, polarizability, and van der Waals volume) to the regular RDF. The AP-RDF descriptor improves the regression quality over conventional textural properties, but the model performance deteriorates at low pressure. Simon et al. proposed the Voronoi energy as a feature, which is the average energy of a probe atom at the accessible Voronoi nodes, the connection of which represents the pore network of the material.[6] The Voronoi energy incorporates both geometrical and energetic information on pores, but it compresses 3D information into a scalar, thus losing spatial details. In addition, the descriptor is biased toward the energy at the pore center, so the full energy landscape in the pore is not captured. Similarly, Fanourgakis et al. proposed an averaged Boltzmann factor for a certain probe atom as a feature;[33,34] instead of inserting a probe into Voronoi nodes, they calculated the averaged Boltzmann factor by randomly placing the probe over the entire space. By changing the size of the probe atom, they prepared a set of features containing nanopore structural information. Another approach is to create a fingerprint based on adsorption data that implicitly encodes both structural and chemical features but allows the ML model to find the higher-order descriptors during the training.[7]

Recently, Bucior et al.[35] proposed using a histogram of the energy felt by a probe species at grid points in a structure as features for ML. In this method, which we will refer to as 1D energy histograms (1D-EH), the energy felt by a probe is first calculated at all points on a regular 3D grid throughout the unit cell. These energies encode both structural and energetic information on the porous material. However, it is not straightforward to use the spatially resolved 3D "energy grids" as features, so Bucior et al. converted them to 1D energy histograms, which loses the spatial information but eliminates the need for cumbersome data augmentation.[21] One attractive property of the histogram representation is that it is invariant to transformations of the unit cell (e.g., translation, rotation, and replication). Li et al. demonstrated the application of these 1D-EH features to the adsorption of short alkanes and Xe/Kr mixtures.[36] A similar approach has been explored by Yu et al. for

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

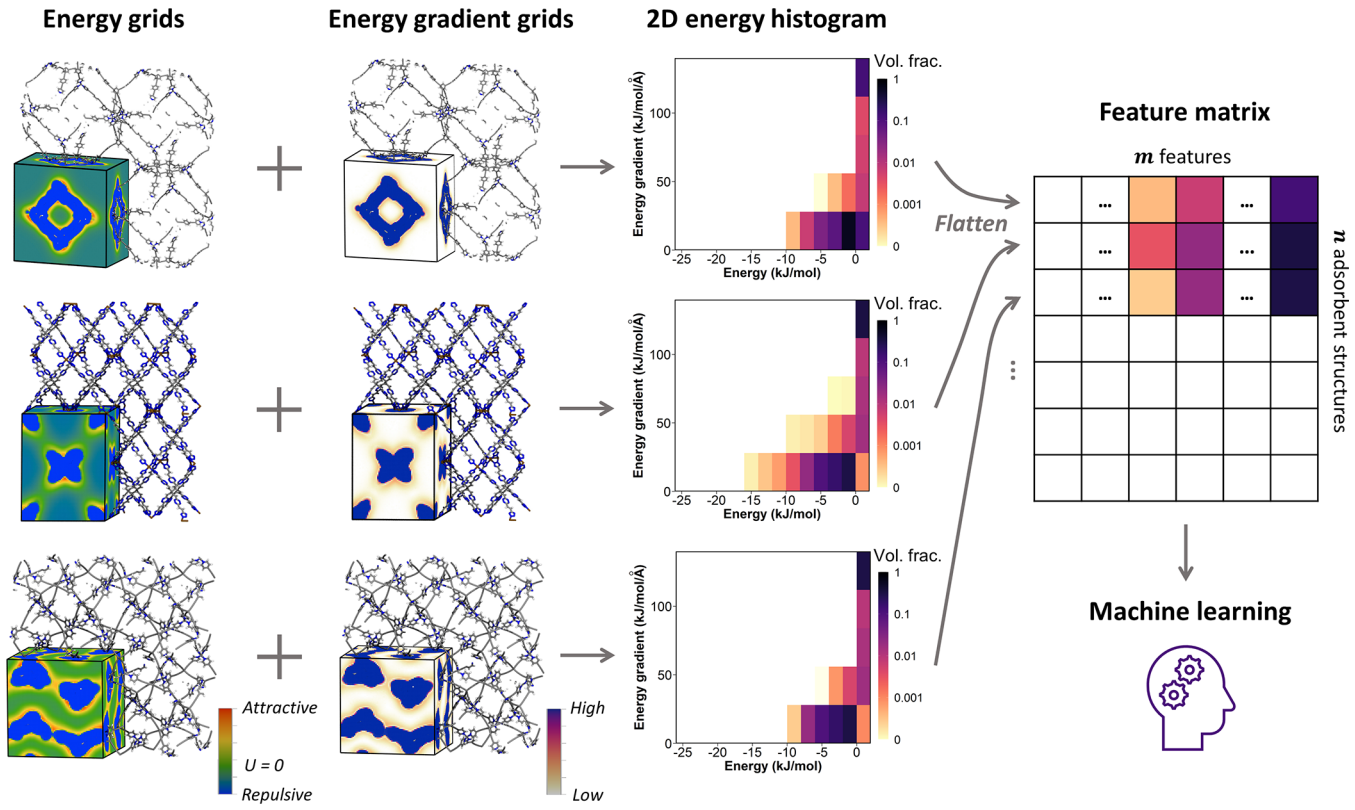


Figure 1. Workflow to construct 2D energy histogram (2D-EH) features. First, the energy and energy gradient felt by a probe particle are evaluated at all grid points in the MOF unit cell. Then, grid points are assigned to the appropriate pixels in a 2D histogram, and the pixel values are normalized with respect to the total number of grid points. To construct a feature matrix for an ML task, the 2D energy histograms are flattened to row vectors and then combined; i.e., each column in the feature matrix corresponds to a pixel in the 2D energy histogram.

predicting adsorption of diverse collections of molecules in MOFs.[37]

In this work, we explore a way to retain some of the spatial information that is lost in 1D-EH features by introducing a second histogram dimension based on the energy gradient, which we term the 2D energy histogram (2D-EH). We show that ML models using these new 2D-EH features can accurately predict single-component adsorption in MOFs for spherical molecules, as well as linear and branched alkanes over a wide range of temperatures and pressures. The capability of our ML model to predict adsorption of alkanes may be useful for high-throughput screening of materials for organic pollutant removal,[38] hydrocarbon storage,[39] and separations.[40−43] By interpreting the 2D-EH features at the atomic level, we show how collapsing the spatially resolved 3D energy and energy gradient grids into a low-dimensional histogram representation still retains some spatial information, which improves the predictions compared to the 1D-EH method. Moreover, we test the transferability of these ML models trained on MOF data to unseen amorphous porous materials (APMs). Finally, we point out a challenge for ML related to capillary condensation in pores.

2. METHODS

2.1. Data Collection. *2.1.1. Metal−Organic Frameworks.* Thousands of MOFs have been synthesized in experiments to date and even more have been generated *in silico*. The structural and chemical diversity of MOFs yields adsorption data that are ideal for training an ML model. In this work, we used MOFs from the ToBaCCo 1.0 database.[44] For each system, we chose a certain number of MOFs randomly from this database for training and testing. See Table 1. There is a total of 13511 MOFs[45] of 41 different topologies in the ToBaCCo 1.0 database. (Originally, 13512 MOFs were reported,[44] but one of them contains no atoms. We excluded this null structure here and in the latest MOFX-DB collection.[45]) The variety of topologies in the ToBaCCo MOFs also leads to a wide distribution in pore size, void fraction, and surface area.[35] All ToBaCCo MOF structures were downloaded from the MOFX-DB Web site (https://mof.tech.northwestern.edu/, accessed Nov. 15, 2020),[45] and we refer to MOFs following the same naming convention as in the MOFX-DB.

2.1.2. Molecular Simulation of Adsorption. We prepared 16 single-component adsorption systems listed in Table 1, with 10 of them taken from Li et al.[36] as benchmark data. These systems represent adsorption of nonpolar spherical, linear, and branched molecules in the ToBaCCo MOFs over a wide range of temperatures (below, near, and above the critical temperature) and pressures (ranging from 2% to 98% of the saturated vapor pressure for subcritical temperatures). The "ground-truth" absolute adsorption uptake in each MOF was obtained from grand canonical Monte Carlo (GCMC) simulations using the RASPA2 package.[46,47] Alkane molecules were represented as united-atom models with potential parameters taken from the TraPPE force field.[48] Nonbonded interactions were modeled by the standard 12-6 Lennard-Jones (LJ) potential. Instead of keeping bond lengths fixed as in the original TraPPE force field, we adopted the default molecular model in the RASPA2 package which treats the chemical bond as a harmonic oscillator. LJ models were used for Kr[49] and Xe[50] atoms. The LJ potential

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 2. Optimized Parameters of 2D-EHs for Different Probes[a]

Probe	Energy [kJ/mol]			Energy gradient [kJ/mol/Å]			Total number of features
	Range	Bin width	Number of bins	Range	Bin width	Number of bins	
Kr	[−30, 0)	2	17	[0, 140)	70	3	51
Xe	[−40, 0)	2	22	[0, 170)	42.5	5	110
CH_3	[−24, 0)	2	14	[0, 112)	28	5	70

[a]A "very attractive" and a "non-negative-energy" bin were added in the energy dimension to account for energy values lying outside the lower and upper bound, respectively; a "very repulsive" bin was also added in the energy gradient dimension to account for energy gradient values higher than the upper bound.

parameters for framework atoms were taken from the Universal Force Field (UFF).[51] All force field parameters and potential forms are reported in Tables S2−S5 in the Supporting Information (SI). Lorentz−Berthelot combining rules[52,53] were applied to estimate the cross-interaction parameters for unlike pairs. All nonbonded interactions were spherically truncated at 12.8 Å for alkane adsorption. The simulation box size was made large enough to ensure that the MOF supercell lattice parameters a, b, and c were at least twice the cutoff radius. The numbers of both initialization and production cycles were set to 3×10^4. All Monte Carlo moves were attempted with equal probability; these moves were translation, rotation, partial and full reinsertion of the component, and swap moves using a continuous fractional component algorithm.[54] For reinsertion and swap moves, the configurational bias algorithm was implemented to enhance the sampling of configurations of chain molecules.[48] A sample simulation input file is provided in Section S2. All adsorption data in this work are reported in volumetric units [cm^3(STP)/cm^3 framework] because the 2D-EH features only encode volumetric information on the structure and no mass or density information is included.

In all GCMC simulations the MOF structure was assumed to be rigid. Although this approach is common in high-throughput calculations of adsorption, systematic studies have suggested that full inclusion of MOF degrees of freedom in adsorption calculations leads to non-negligible effects in a surprising fraction of materials.[55] This observation is a reminder that the predictions of ML models are subject to the same limitations as their training data and that using higher resolution methods that test the impact of these limitations for materials of particular interest is advisible.

2.1.3. Calculation of Textural Properties. All MOF structures were characterized by conventional textural properties. The helium void fraction (VF) was estimated by the Widom insertion method using RASPA2.[49,56] Detailed simulation parameters are listed in Table S6. The total volumetric surface area (VSA, including surface area from both accessible and inaccessible pore network), total gravimetric surface area (GSA), pore limiting diameter (PLD), and largest cavity diameter (LCD) were calculated by Zeo++ v0.3,[57] where framework atom radii were taken from UFF.[51] A nitrogen probe with radius of 1.86 Å[58] was used to assess both the accessibility of the network and the surface area. We used the "-ha" flag to achieve high-accuracy calculations. All textural properties are available in the SI.

2.2. Feature Engineering. The workflow to construct the 2D-EH features is summarized in Figure 1. We first discretized the unit cell of each MOF structure into grid points that are evenly spaced by 0.5 Å along each axial direction; if the edge length cannot be divided by 0.5 evenly, we took the number of grid points along that axial direction to be one (origin) plus the largest integer that does not exceed the quotient. By placing a spherical probe at each grid point i, we calculated the potential energy $\mathcal{V}_i$ and the potential energy gradient as

$$\nabla \mathcal{V}_i = \mathbf{x}\frac{\partial \mathcal{V}_i}{\partial x_i} + \mathbf{y}\frac{\partial \mathcal{V}_i}{\partial y_i} + \mathbf{z}\frac{\partial \mathcal{V}_i}{\partial z_i} = \mathbf{x}\sum_j f_{ij}\frac{x_{ij}}{r_{ij}} + \mathbf{y}\sum_j f_{ij}\frac{y_{ij}}{r_{ij}} + \mathbf{z}\sum_j f_{ij}\frac{z_{ij}}{r_{ij}} \quad (1)$$

where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are unit vectors in Cartesian coordinates. The scalar force between adsorbent atom j and the probe at grid point i is $f_{ij} = -\partial v_{ij}/\partial r_{ij}$, v_{ij} is the pairwise potential energy, and r_{ij} is the scalar distance; variables x_{ij}, y_{ij}, and z_{ij} are the components of the distance vector $\mathbf{r}_{ij}$. The gradient in eq 1 is a vector. We converted it to the (scalar) energy gradient by taking its magnitude (norm):

$$\|\nabla \mathcal{V}_i\| = \sqrt{\left(\sum_j f_{ij}\frac{x_{ij}}{r_{ij}}\right)^2 + \left(\sum_j f_{ij}\frac{y_{ij}}{r_{ij}}\right)^2 + \left(\sum_j f_{ij}\frac{z_{ij}}{r_{ij}}\right)^2} \quad (2)$$

LJ Kr[49] and Xe[50] spheres were used as the probes for the corresponding adsorption systems. For the adsorption of alkanes, a united-atom model of a methyl group was adopted as the probe because it is the most exposed pseudoatom that interacts strongest with the adsorbent.[36,59] LJ parameters of the probes are available in Table S7. The probe-adsorbent energy and energy gradient were evaluated at all grid points. Force fields, the nonbonded interaction scheme, and the simulation box size were consistent with those used in the GCMC simulations of this work and of Li et al.[36] When evaluating the energy gradient in eq 2, the impulsive force was calculated to account for the abrupt change of energy at the cutoff radius, thus retaining the consistency with the energy calculations.[60]

Once ready, we binned the values at the grid points into a 2D histogram in terms of energy and energy gradient (i.e., 2D energy histogram). Each "pixel" in the 2D-EH was normalized by the total number of grid points in the simulation box. The range in each dimension of the 2D histogram was determined based on the statistical significance (see an example in Table S8).[35] For simplicity, the histogram bin width (i.e., the resolution of 2D histogram) was assumed to be dependent only on the probe type, but independent of adsorbate type (i.e., not distinguishing between the different alkanes), adsorption condition, and ML algorithm. We optimized the bin width for three probe types (i.e., Kr, Xe, and methyl group) based on the principle of bias-variance trade-off (see Section S3.3).[61] As shown in Section 3, after determining the resolution of the 2D-EH only once for the methyl probe, the same 2D-EH is robust and transferable to all alkanes in this study at different adsorption conditions. Some grid points have energy and energy gradient values lying outside the bound of the histogram. We

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

included a "very attractive" bin for all energy values smaller than the lower bound in the energy dimension and a "non-negative-energy" bin for all non-negative energy values (including 0). A "very repulsive" bin was also included in the energy gradient dimension to account for all grid points having energy gradient values larger than the corresponding upper bound. With this setup, the summation of all pixel values in the 2D-EH is normalized to unity. Again, we note a prominent advantage of the 2D-EH over the raw grid representation is that the histogram representation is independent of how the unit cell is chosen for the materials.

Optimized parameters of the 2D-EHs are summarized in Table 2 for different probes. It was found that the bins in the energy dimension are finer than those in the energy gradient dimension because the energy appears in the Boltzmann factor weighing the different microstates and appearing in the MC acceptance rules. Eventually, the 2D histograms were flattened into row vectors and stacked together to form a $n \times m$ feature matrix, as illustrated in Figure 1, where n is the number of adsorbents in the data set and m is the total number of pixels (features) of the 2D-EH (see Table 2 for the value of m).

2.3. Machine Learning and Data Analysis. The correlation function, $y = F(\mathbf{X})$, between the 2D-EH feature vector, $\mathbf{X}$, and the corresponding adsorption capacity of the material, y, was estimated using supervised ML algorithms. The form of the correlation function is unknown *a priori*. As a starting point, we tested least absolute shrinkage and selection operator (LASSO) regression.[62] The functional form of LASSO is a multiple linear regression with an L1 regularization term included in the loss function to reduce model overfitting. Most importantly, LASSO automatically selects important features by zeroing the weights of the least important features. In cases where the correlation function is nonlinear, we tried a Random Forest (RF) regression algorithm.[63] The RF regression falls into the category of ensemble learning. A RF model consists of many independent regression trees (for a regression task), and the results are averaged from all trees, thus reducing the likelihood of overfitting the training data set. Gradient boosting[64] is another class of ensemble learning methods. Like RF, gradient boosting also combines several weak learners into a stronger learner. However, it differs from RF in the way that gradient boosting trains tree predictors sequentially and each iteration focuses on the residuals of the data.[65] In particular, we employed extreme gradient boosting (XGB),[66] which is an efficient implementation of the gradient boosting algorithm. Finally, we tried a simple class of feed-forward artificial neural networks, i.e., multilayer perceptron (MLP).[65] Unlike RF and XGB, an MLP with many hidden layers is able to learn features at various levels of abstraction. We designed an MLP architecture having four hidden layers with a dropout layer added after the first hidden layer to prevent overfitting. ML model training and testing were performed with multiple Python and R packages. Details on the ML methods and hyperparameter tunning are available in Section S4 in the SI. For all systems we used a 50/50 data split for training and testing purposes except for the case of *n*-butane adsorption at 1.2 bar, 298 K, where more data points are available, and an 80/20 data splitting was used. Code is hosted at https://github.com/snurr-group/2D-energy-histogram. Trained ML models and other necessary supporting data are stored at 10.5281/zenodo.5481697.

To measure the predictive accuracy, the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root-mean-square error (RMSE) were calculated,

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \overline{y}_i)^2} \tag{3}$$

$$\mathrm{MAE} = \frac{\sum_i^n |y_i - \hat{y}_i|}{n} \tag{4}$$

$$\mathrm{MAPE} = \frac{1}{n} \sum_i^n \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\% \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}} \tag{6}$$

where n is the total number of data points and $\overline{y}_i$ is the mean of all adsorption amounts from GCMC (y_i). In practice, for a porous material with small pores (LCD < 3 Å), GCMC simulation predicted $y_i = 0$ while ML models predicted a small but nonzero value. This led to a problematic implementation of the conventional MAPE definition where y_i is in the denominator. Here we adopted a slightly different MAPE definition, where the predicted amount, $\hat{y}_i$, is present in the denominator.[67] This definition of MAPE tells us the expected percentage error given a predicted value. This implementation of MAPE helps avoid the singularity problem (i.e., division by zero) in our case.

3. RESULTS AND DISCUSSION

3.1. Benchmark of 2D Energy Histogram Features. We first compared the performance of 2D-EH features against the 1D-EH features on the data set reported by Li et al.[36] (see Table 1). Figure 2 shows parity plots comparing the GCMC data for Kr adsorption at 1 bar, 273 K, with ML predicted values using 2D-EH features. Even with most of the points concentrated at low loading and a few points spreading out in the middle and high loading range (>50 cm^3_{STP}/cm^3), both LASSO and RF models can establish a good correlation using the 2D-EH features, with LASSO performing slightly better ($R^2 = 0.99$, compared to $R^2 = 0.97$ for RF). We note that analogous ML models from Li et al. using the 1D-EH features produced a lower level of correlation on the same data set (with $R^2 \sim 0.83-0.85$).[36] Although our ML models achieve an overall good fitting quality, LASSO sacrifices predictive capability at low loadings (<3 cm^3_{STP}/cm^3) by systematically overestimating the adsorption capacity (inset of Figure 2a). In contrast, the RF model improves the prediction quality in the low-uptake regime (inset of Figure 2b) but at the cost of marginally less predictive ability in the high loading range (>75 cm^3_{STP}/cm^3), thus leading to a slightly higher RMSE value (3.4 cm^3_{STP}/cm^3 compared to 2.1 cm^3_{STP}/cm^3 for LASSO). The difference in fitting quality at low loading between LASSO and RF models also explains a slightly higher MAPE value in the LASSO case (i.e., MAPE = 8.4% for LASSO versus MAPE = 4.3% for RF), where systematic deviations at low loading have larger impact on the MAPE metric. In practice, the RF model appears suitable to cases where uptake is present in the denominator of the objective metric, such as in the prediction of selectivity, while the LASSO regression model is preferred when predicting adsorption capacity larger than ~3 cm^3_{STP}/cm^3.

Figure 3 summarizes the evaluation metrics of ML predictions using 2D-EH features and those using 1D-EH features[36] for Kr, Xe, and alkanes up to C_3. Comparing different ML models, LASSO regression using 2D-EH features works better for

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

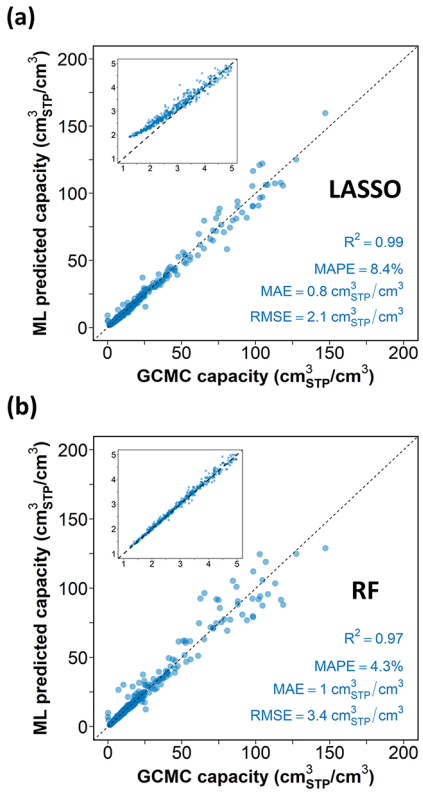


Figure 2. Parity plots comparing GCMC results for Kr adsorption at 1 bar, 273 K, against ML prediction (on 1000 testing data) using 2D-EH features and different ML models: (a) LASSO regression and (b) RF. Inset is an enlarged view of the corresponding parity plot at low capacity (<5 cm^3_{STP}/cm^3). Both LASSO and RF give good predictions, with LASSO giving a better overall fitting quality and RF being particularly good at low loadings. See Figure S4 for parity plots on the training data.

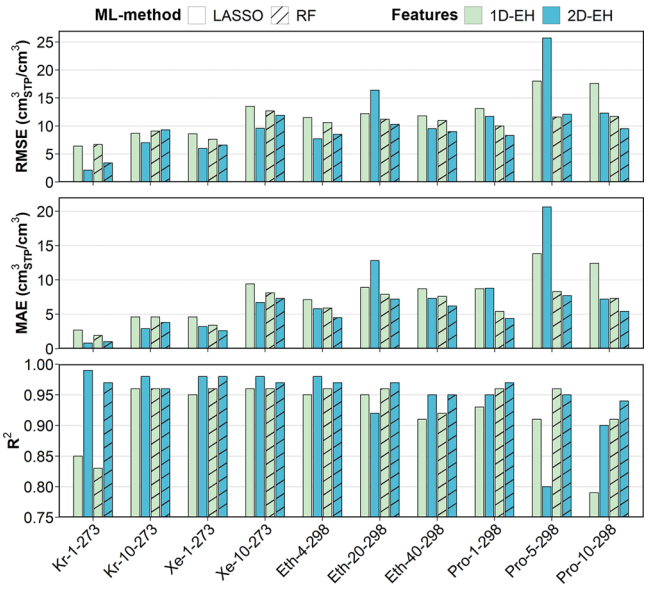


Figure 3. Benchmark comparison of ML prediction on the testing data using 2D-EH features (this work) and 1D-EH features.[36] In general, 2D-EH features lead to better predictions than the 1D-EH features for both spherical and chain molecules. Corresponding parity plots for each system are available in Figures S4–S13. See Table 1 for the meaning of the system labels.

spherical molecules with an average of 7% and 23% reduction in MAE and RMSE compared to those of the RF models, indicating an approximate linear correlation between the input features and adsorption capacity. This linear behavior has been observed before[35,36] and is expected in view of the nature of energy-histogram-based features. We discuss this further in Section 3.3. Unlike spherical molecules, short-chain alkanes expand over several grid points at a time, and nonlinear correlation functions (such as the RF model) yield improved performance, while LASSO regression can produce systematic errors; for example, see the LASSO prediction using 2D-EH features in the "Pro-5-298" case in Figure 3 and the corresponding parity plots in Figure S12. Comparing different features, when LASSO models were implemented for adsorption of spherical molecules, replacing the 1D-EH features with our 2D-EH features reduces the MAE by more than 40% on average; in the case of Kr adsorption at 1 bar and 273 K, this number is about 70%. When RF models were applied for short-chain alkanes, 2D-EH features reduce the MAE by 15% on average compared to 1D-EH features. We note that Li et al. used 1 Å for the grid size. We found that a finer grid size of 0.5 Å leads to a 4.6% decrease in MAE compared to a coarser grid size of 1 Å for the "Eth-40-298" case using the RF model (Figure S20). We thus used a grid size of 0.5 Å to prioritize the accuracy, with the intention to test the applicability of our 2D-EH features to different systems.

We also trained baseline RF models using conventional textural properties as features (VF, VSA, GSA, LCD, PLD), and corresponding ML performance metrics on the testing data are summarized in Table S11. For all systems considered in Figure 3, while RF models using 2D-EH features always maintain high predictive accuracy with $R^2 > 0.94$, baseline models show an increase in performance from low pressure to high pressure with R^2 ranging from 0.81 to 0.97. The improvement of 2D-EH features over textural features is especially significant for adsorption of spherical molecules and for alkanes at low relative pressure ($P/P_0 \sim 0.1$), with an average of 63% reduction in MAE. For adsorption of alkanes at medium to high relative pressure, the performances of baseline models and RF models using 2D-EH features are comparable. The good predictive accuracy of baseline models at these conditions can be understood by incrementally stronger correlation between the adsorption capacity and structural properties (e.g., VF, VSA, GSA) as the loading increases.

The 2D-EH features are closely related to the Henry's constant of corresponding probes. We calculated the Henry's constant for Kr and Xe at 273 K using the Widom insertion method. For alkane systems, the cost of calculating the energy histogram for a single spherical probe species is significantly lower than that for calculating the Henry's constant for a complex molecule such as *n*-hexane. Therefore, we calculated the Henry's constant of a single methyl probe at 298 K for a consistent comparison with 2D-EH features. Testing metrics of baseline RF models using both textural properties and Henry's constant as features are summarized in Table S12. As expected, adding the Henry's constant to the baseline feature set improves the ML predictions for spherical molecules and for alkanes at low relative pressure, where RF models using only textural features fall short. Nevertheless, the 2D-EH features still lead to better ML predictions under these conditions, and an average of 33% reduction in MAE can be achieved.

As implemented, all ML models using 2D-EH features predict adsorption capacity in volumetric units, but the conversion to

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

gravimetric capacity is straightforward. We illustrate this conversion with Kr adsorption data in Figure S21. The converted gravimetric capacities have slightly worse R^2 compared to the original volumetric data, but the MAPE remains the same after conversion (see eq 5).

3.2. Adsorption of Alkanes in MOFs. Having demonstrated the advantage of 2D-EH features over 1D-EH and typical baseline features in simple systems, we implemented 2D-EH features to predict the adsorption of larger linear and branched alkanes. We employed four ML regression algorithms: LASSO, RF, XGB, and MLP, and the comparison is summarized in Figure 4. The LASSO regression leads to a lower predictive

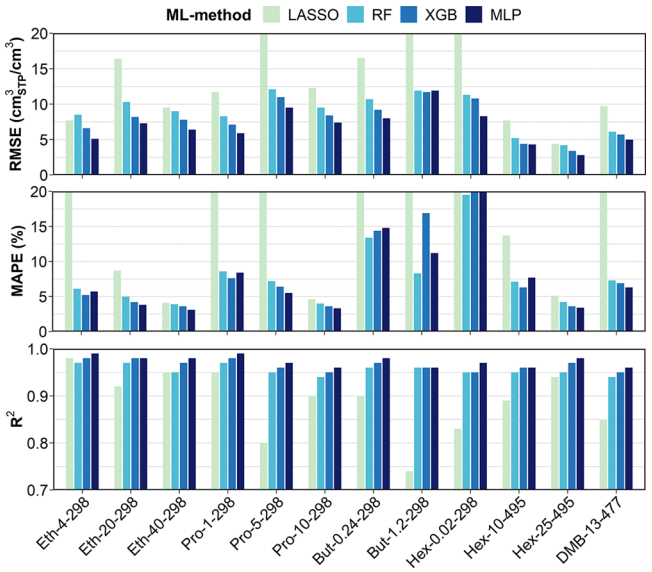


Figure 4. Bar plots for ML performance metrics on the testing data using 2D-EH features. Both MAPE and RMSE are cut off at a value of 20 in the plots; some systems have MAPE and RMSE that exceed the cutoff value (see full data in Table S13). The results show that nonlinear ML models, such as RF, XGB, and MLP, are suitable for predicting adsorption behavior of chain molecules using only the methyl group as a probe. Corresponding parity plots for each system are available in Figures S8–S19. See Table 1 for the meaning of the system labels.

accuracy than the nonlinear models for these alkane systems, consistent with the findings reported in Section 3.1; in some cases, LASSO results in a 3-fold increase in MAPE compared to the RF predictions. By applying nonlinear ML models, i.e., RF, XGB, and MLP, the adsorption predictions improve, with R^2 ranging from 0.94 to 0.99. Among the three nonlinear models, MLP provides the best prediction. High predictive accuracy on the testing data (R^2: 0.96–0.99) indicates that MLP with 2D-EH features can accurately predict adsorption of both linear and branched alkanes over a wide range of relative pressures and temperatures. Figure 5 shows parity plots comparing GCMC data and MLP predictions for three representative systems, i.e., propane, *n*-hexane, and 2,2-dimethylbutane adsorption at different relative pressures. Similar performance in both training and testing data sets suggests that the current MLP architecture is robust. Even using RF with default hyperparameters in R, the adsorption capacity can be predicted reasonably well for all alkanes examined here, with R^2 ranging from 0.94 to 0.97. XGB regression shows results of intermediate quality between RF and MLP (R^2 of 0.95–0.98).

It should be noted that, given a MOF structure, the same 2D-EH features based on the energy and energy gradient grids of a simple methyl probe were used for all alkanes, including both linear hexane and its branched isomer 2,2-dimethylbutane. This convenience in preparation of features can be attributed to the similar guest–host interactions for all alkane systems, where no site-specific interaction (e.g., Coulombic interaction) was considered in the simulation. Using these features, the ML model implicitly learns how multisite molecules adsorb in these complex pore environments—also learning the effects of adsorbate–adsorbate interactions.

In comparison to baseline RF models using only textural properties as features (Table S11), for alkanes larger than C_3, 2D-EH features outperform textural features at low relative pressure ($P/P_0 \sim 0.1$), with an average of 50% reduction in MAE. This number decreases to 26% at a relative pressure of 0.5. At high relative pressure near saturation ("Hex-25-495" case), the performance of both types of features is similar. We also trained baseline MLP models using textural features (Table S11). We found that, unlike the case of 2D-EH features where MLP models improve the predictive accuracy, MLP models with textural features lead to similar (and in some cases almost identical) performance compared to RF models. Similar to observations in Section 3.1, adding the Henry's constant of a single methyl probe at 298 K to the textural feature set improves the ML predictions at low relative pressure (Table S12). This new feature set combining both Henry's constant and textural properties, however, still underperforms the 2D-EH features, leading to a MAE that is on average 39% higher at low relative pressure for alkanes larger than C_3.

Although disguised by high R^2 values of ML predictions based on 2D-EH features, the MAPEs are unusually high for some systems, e.g., But-0.24-298, But-1.2-298, and Hex-0.02-298. The parity plots of these systems exhibit systematic outliers in the middle and high adsorption capacity ranges (e.g., Figures S14–S16). The systematic deviation in ML prediction is associated with the presence of capillary condensation and hysteresis near the investigated adsorption conditions. Capillary condensation and hysteresis can be observed in mesopores at (deep) subcritical temperatures as a first-order phase transition proceeding from a low-density adsorbed phase to a high-density adsorbed phase. For adsorption at higher temperatures close to the critical temperature, such as for *n*-hexane adsorption at 495.6 K ($T/T_c = 0.98$), capillary condensation does not occur, and the MAPEs and RMSEs are reduced. Further analysis of systems displaying capillary condensation is provided in Section 3.5.

3.3. Interpretation of 2D-EH Features and ML Models. One- and two-dimensional energy histograms lack the spatial information provided in the 3D energy grid, and yet these lower dimension histograms can serve as successful descriptors for predicting adsorption. Figure 6a helps explain why this is so. It can be seen that the 2D-EH representation essentially classifies the grid points in the 3D structure in terms of well-defined regions (or adsorption sites) roughly based on the distance to the framework walls, and the volume fraction of a region corresponds to the value of a 2D-EH feature (i.e., a pixel in the 2D energy histogram). For example, the 2D-EH feature X_{70} corresponds to the volume fraction of the space that overlaps with the framework, and feature X_{14} corresponds to the volume fraction of the "open space" where the probe cannot feel the framework. The 1D-EH features work in a similar manner, but they cannot distinguish grid points close to the framework from those far away from the framework when the grid points have the

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

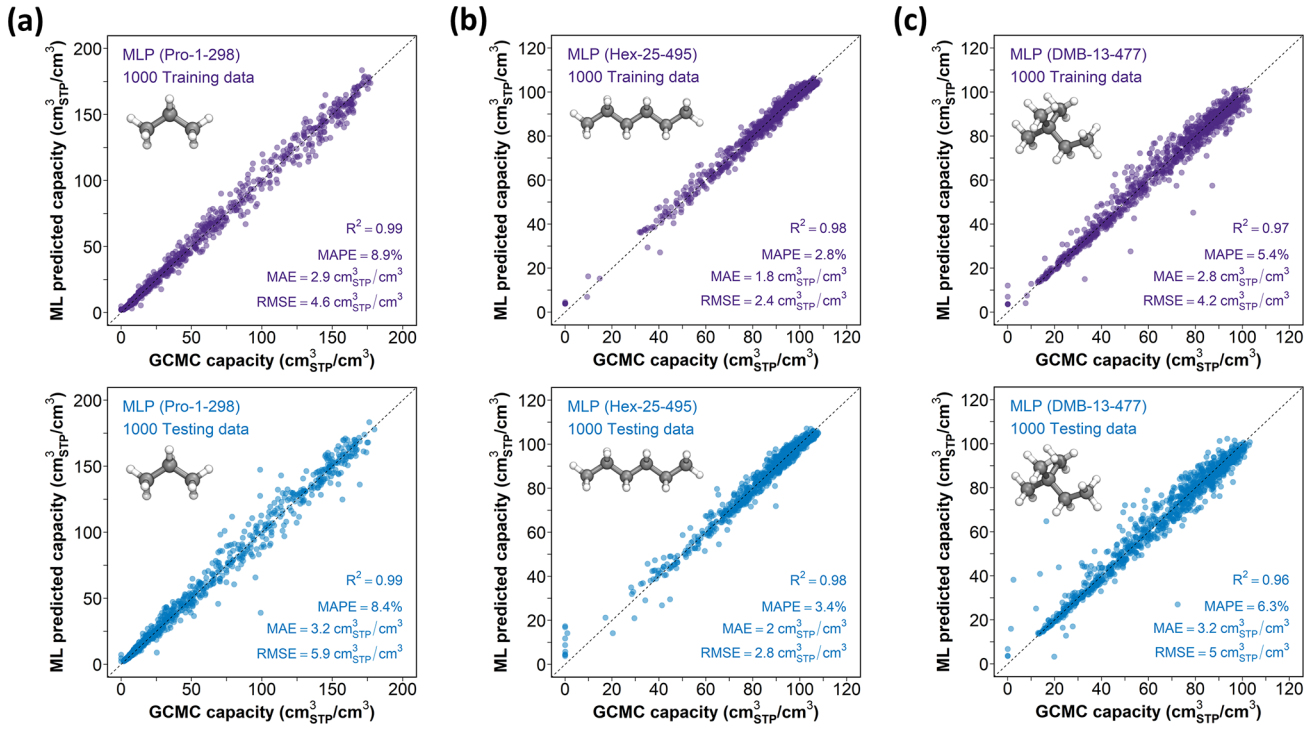


Figure 5. Example parity plots (training on the top and testing on the bottom) comparing GCMC results and MLP predictions using 2D-EH features for adsorption of different alkanes: (a) propane at 1 bar, 298 K (P/P_0 = 0.08); (b) *n*-hexane at 25 bar, 495.6 K (P/P_0 = 0.98); and (c) 2,2-dimethylbutane at 13 bar, 477.1 K (P/P_0 = 0.5), where P_0 is the saturation pressure of the adsorbate at the operational temperature reported in Table S1.

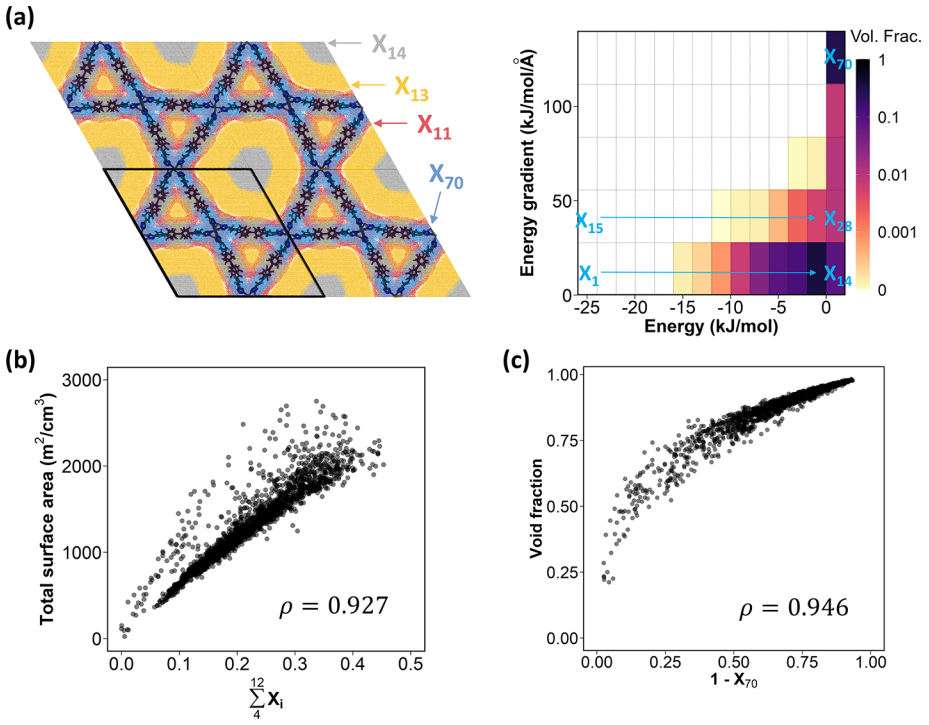


Figure 6. Interpretation of 2D-EH features. (a) Left: 3D supercell of tobmof-1269 with grid points superimposed and colored based on different 2D-EH pixels; the black parallelepiped is the unit cell of the super structure. Right: corresponding 2D energy histogram evaluated by a methyl probe. We assign a notation X_i (i = 1, 2, ..., 70) to each 2D-EH feature variable (i.e., pixel). (b) Total volumetric surface area versus the summation of 2D-EH feature values that represent the fraction of grid points in the unit cell that have energy ranging from −20 to −2 kJ/mol and energy gradient ranging from 0 to 28 kJ/mol/Å. (c) Helium void fraction versus $1 - X_{70}$, where feature X_{70} represents the fraction of grid points in the unit cell that have energy and energy gradient larger than 0 and 112 kJ/mol/Å, respectively. In both (b) and (c), the data points correspond to the 2000 MOF structures examined for adsorption of *n*-hexane at 10 bar, 495.6 K. The Pearson correlation coefficient, ρ, is close to 1 in both cases, suggesting a strong linear correlation between the conventional textural properties and the linear combination of 2D-EH features.

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

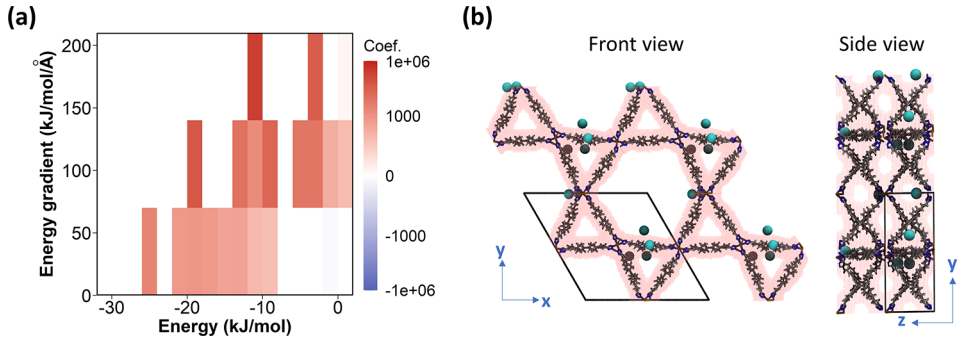


Figure 7. Interpretation of the LASSO model for Kr adsorption at 1 bar, 273 K. (a) 2D heatmap of LASSO coefficients. Pixels in this 2D heatmap have one-to-one correspondence to pixels in the 2D energy histogram; see an example histogram in Figure S24 for a Kr probe. (b) Front and side views of the GCMC snapshot in tobmof-1269, plotted with VMD software.[70] Kr molecules are shown as cyan spheres. Grid points that favor adsorption (i.e., grid points contributing to the 2D-EH features that have positive LASSO coefficients) are overlaid on the structure as red shading.

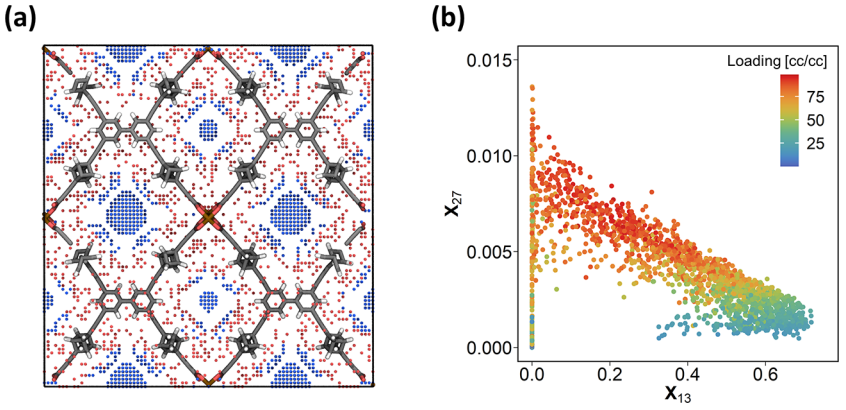


Figure 8. Illustration of the importance of the energy gradient in prediction of alkane adsorption. (a) 3D structure of tobmof-4936 with grid points associated with 2D-EH features X_{13} (blue) and X_{27} (red); see Figure 6a for their location in the 2D-EH. Both features fall into the same energy range but have different energy gradient ranges. (b) 2D-EH feature X_{27} versus X_{13} for 2000 MOFs with points colored according to *n*-hexane adsorption capacity at 10 bar, 495.6 K. Feature variable X_{27} is more efficient in splitting the data than X_{13} in tree-based ML algorithms, such as XGB and RF.

same energy. Compared to using the 3D grid representation of the material directly (see ref 21 for example), the 2D-EH facilitates the ML task by grouping voxels into a smaller number of features in a physical way. Based on this reasonable space decomposition, it is worth considering the relationship between 2D-EH features and conventional textural properties. Figure 6b shows that there is roughly a linear correlation between the total surface area and the summation of some 2D-EH features that corresponds to the volume fraction of pore space near the solid framework. The latter resembles the "binding fraction" proposed by Bobbitt and coauthors, which was designed to quickly screen a large number of MOFs for hydrogen storage.[68] Figure 6c shows that the quantity $1 - X_{70}$ correlates with the helium void fraction. The apparent linear correlation in both cases is confirmed by a high Pearson correlation coefficient. We note that the linear combination of 2D-EH features that can be correlated with example textural properties is not unique due to the correlation between 2D-EH features (see Figure S23). The ability to encode structural information enables the 2D-EH features to accurately predict adsorption at both low pressure (where surface interactions dominate) and high pressure (where the pore volume of material dominates). Results presented in Figure 6 are based on energy and energy gradient grids evaluated by a methyl group, but the demonstrated characteristics are probe-independent (see Figure S24 for the Kr probe).

Interpreting 2D-EH features in the way illustrated in Figure 6a also provides insight into the predictive capability of the LASSO regression model for spherical molecules. Since a spherical molecule represented by a single interaction site always falls into a region that is represented by a 2D-EH feature, the coefficient of LASSO regression directly implies whether that particular region in the porous structure is favorable for adsorption. Figure 7a shows a heatmap of LASSO coefficients learned from Kr adsorption data at 1 bar, 273 K. LASSO coefficients that are positive indicate 2D-EH features that positively contribute to the adsorption capacity. To visualize this idea at the atomic level, we collected grid points that were classified to the 2D-EH features that have a positive LASSO coefficient and mapped them back to the 3D structure. Figure 7b shows these grid points superimposed on the MOF structure for the example of tobmof-1269. We also overlap a GCMC snapshot of Kr molecules onto the structure. We can see that the "favorable" regions learned by the LASSO model are consistent with the information conveyed by the GCMC snapshot, which suggests that our ML model learns some basic physics of adsorption from the data pattern. Nevertheless, we should be cautious about overinterpreting the LASSO coefficients. Our caution is based on two points:

(a) A larger LASSO coefficient may not necessarily indicate that the corresponding feature is more important than the others in the prediction. Normalizing (in the range of [0, 1], as in our case) or standardizing (with zero mean and unit variance) the 2D-EH features will not affect the model's overall prediction but will alter the coefficients significantly (Figure S25).

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

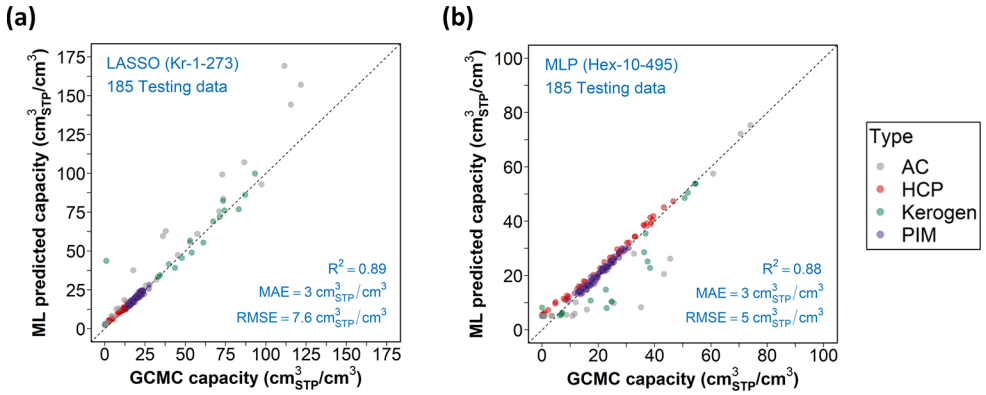


Figure 9. Parity plots comparing GCMC results and ML predictions for adsorption in APMs. (a) Kr adsorption at 1 bar, 273 K. (b) *n*-Hexane adsorption at 10 bar, 495.6 K. ML models used here are the same as those originally introduced in Sections 3.1 and 3.2 and were trained with MOF data only.

(b) The 2D-EH features are not mutually independent. Pearson correlation coefficients shown in Figure S23 confirm that some features are linearly correlated or anticorrelated. This suggests that, for highly collinear variables, the LASSO regression may randomly select one of them (typically the first one enters the model) and shrink the coefficients of the rest. In this sense, it is hard to say one feature is more important than other correlated features. This problem may be solved by preprocessing 2D-EH features with principle component analysis[69] to form a compact and independent feature set.

In contrast to spherical molecules, articulated molecules, such as the alkanes examined here, do not reside in only one represented region at a time but expand over multiple voxels in 3D space. To illustrate why the inclusion of the energy gradient helps the ML prediction, such as in tree-based algorithms (e.g., XGB and RF), we picked 2D-EH feature X_{27} from Figure 6a, which is identified as among the top in feature importance in predicting *n*-hexane adsorption at 10 bar, 495.6 K (Figure S26). For comparison, we also picked another feature X_{13}. Both features fall into the same energy range from −2 to 0 kJ/mol in the 2D energy histogram, but feature X_{27} corresponds to grid points having higher energy gradient (28 to 56 kJ/mol/Å) and feature X_{13} corresponds to grid points having lower energy gradient (0 to 28 kJ/mol/Å); see histogram in Figure 6a. In the 3D structure, grid points belonging to X_{27} are closer to the framework than those belonging to X_{13}, as shown in Figure 8a. Intuitively, the grid points close to the framework should contribute differently to adsorption than those having the same energy but farther away from the framework. To confirm this, Figure 8b shows the values of X_{27} versus X_{13} for 2000 MOF structures in the "Hex-10-495" data set (Table 1). Mimicking the way that tree-based methods split the data, if we draw a vertical line (i.e., splitting data based on X_{13}), we find mixed high-loading and low-loading data on both sides of the line (poor classification). While if we draw a horizontal line (i.e., splitting the data based on the feature X_{27}), we find a better classification of the data. This example shows that grid points close to the framework may not have the same influence on the ML prediction as those far away from the framework but possessing the same energies, and our 2D-EH features are able to distinguish these grid points.

3.4. Transferability to Amorphous Porous Materials. We also tested the generalizability of our ML models using 2D-EH features to amorphous porous materials (APMs). Compared to the ToBaCCo MOFs used for ML training, APMs have much smaller pore sizes (e.g., LCD < 20 Å, Figure S28) and more complex pore morphologies due to their amorphous nature. Therefore, APMs are a good set of materials for testing the transferability of ML models to regions that are not emphasized by the training data.

APMs considered were hyper-cross-linked polymers (HCPs), polymers of intrinsic microporosity (PIMs), activated carbons (AC), and kerogens. We collected united-atom models for 120 microporous polymer conformations (composed of unique types of 9 HCPs and 15 PIMs) from previous work.[71−73] Thyagarajan and Sholl[74] built a database of rigid APMs, which includes a collection of 68 and 16 published structures of AC[75−81] and kerogen,[82] respectively. We excluded 19 large AC structures from the collection that would make GCMC simulations prohibitively slow (Figure S27). Calculation details of the 2D-EH features for these structures are consistent with those described in Section 2.2. GCMC simulations of adsorption followed the same protocols as in Section 2.1.2. Nonbonded LJ parameters for AC and kerogen were taken from UFF,[51,74] and nonbonded LJ parameters for polymers were taken from the united-atom TraPPE force field.[59] During the simulation, we assumed that the structures of the amorphous materials were rigid. We note that this assumption was used for the purpose of evaluating ML predictions on APMs, and in practice, adsorption-induced restructuring can significantly affect predictions in organic microporous glasses.[83,84] We selected two representative adsorbate systems for testing: (1) Kr at 1 bar, 273 K, and (2) *n*-hexane at 10 bar, 495.6 K. Both systems are at a high relative temperature, T/T_c, close or above 1. Thus, we can test the transferability of well-behaved ML models without complications from the potential problem of capillary condensation. All adsorption data in APMs are available in the SI.

In the case of Kr adsorption at 1 bar, 273 K, we employed the LASSO model for MOFs that is originally introduced in Section 3.1 to predict adsorption in unseen APMs. Figure 9a shows the parity plot comparing GCMC data and LASSO predictions. Good evaluation metrics (MAE = 3 cm^3_{STP}/cm^3) for the 185 testing data points highlight the robustness and transferability of the LASSO model. Similarly, in Figure 9b, the MLP model that is introduced in Section 3.2 shows good predictions in general for all types of APMs for *n*-hexane adsorption at 10 bar, 495.6 K (MAE = 3 cm^3_{STP}/cm^3 comparable to that tested on MOF structures, i.e., $MAE_{MOF} \sim 3$ cm^3_{STP}/cm^3, see Figure S17). Our

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

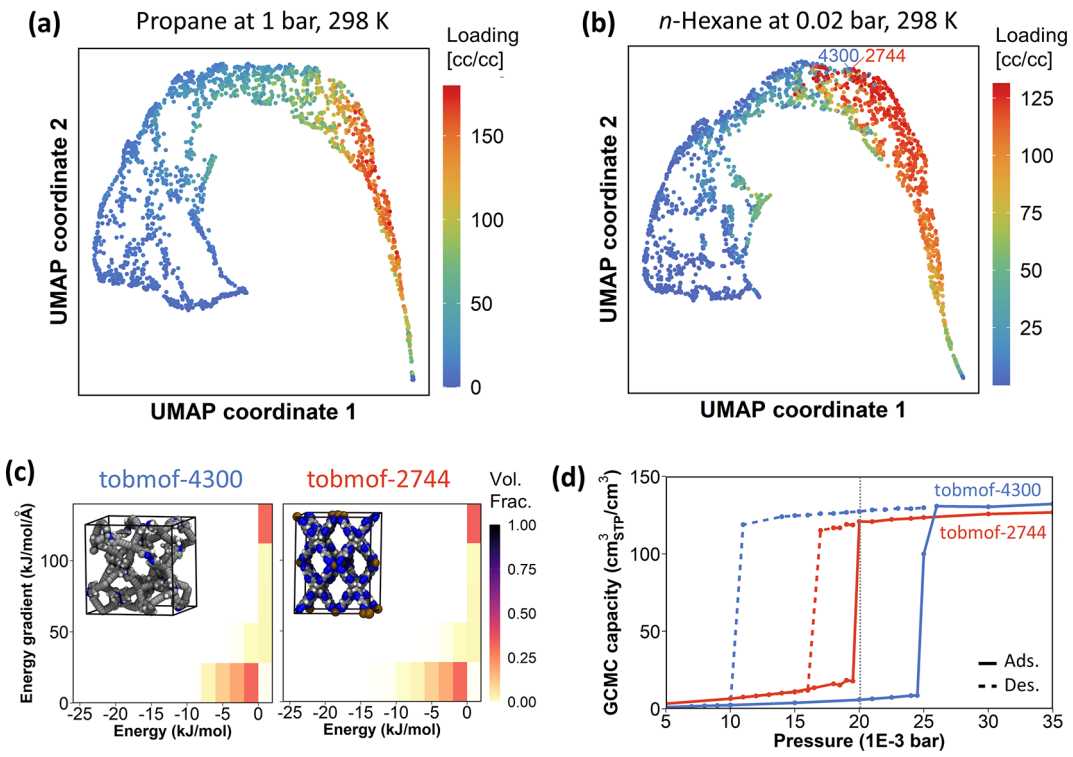


Figure 10. UMAP analysis of the effectiveness of 2D-EH features and illustration of capillary condensation in MOFs. (a) MOF structures projected in a 2D UMAP space, with points colored according to the propane adsorption capacity at 1 bar, 298 K, and (b) *n*-hexane adsorption capacity at 0.02 bar, 298 K. (c) 2D energy histograms for two selected MOFs marked in (b); insets show the unit cells. (d) Adsorption (solid line) and desorption (dashed line) isotherms of *n*-hexane at 298 K for selected MOFs, where capillary condensation and hysteresis occur near the studied pressure, 0.02 bar.

findings suggest that, despite being trained with ToBaCCo MOFs that have generally large pores (Figure S28), ML models using 2D-EH features show excellent transferability to different classes of porous materials with smaller pores and more complex pore morphologies.

3.5. Challenges in Predicting Adsorption with Capillary Condensation. As pointed out in Section 3.2, the occurrence of capillary condensation leads to systematic deviations in our ML predictions. The poor ML performance at deep subcritical temperatures is not limited to alkane adsorption but also occurs for spherical molecules, such as methane (modeled as a single LJ sphere) adsorbed at 112 K ($T/T_c \approx 0.59$ and $P/P_0 = 0.48$, Figure S29). For many applications of adsorption in gas storage and separation that occur at or above the critical temperature of the gases of interest, one does not run into this problem. For example, hydrogen storage is usually at or above 77 K (T_c = 33.2 K), methane storage is at 298 K (T_c = 190.6 K), and CO_2 capture is usually above 298 K (T_c = 304 K). In this section, we try to answer the following questions: (1) Why do 2D-EH features perform poorly at deep subcritical temperatures? (2) How well do 2D-EH features handle capillary condensation compared to structural features, such as those derived from persistent homology?

To answer the first question, we applied a dimensionality reduction method, specifically Uniform Manifold Approximation and Projection (UMAP),[85] to learn the data topology in a high-dimensional manifold and then project the high-dimensional data onto a low-dimensional space. The UMAP method is an unsupervised learning technique. Thus, it takes only the 2D-EH features of each structure as the input but not the adsorption data. Details on the UMAP calculations are provided in Section S8.1. We favor the UMAP method over the t-distributed Stochastic Neighbor Embedding (t-SNE) method[86] mainly because the UMAP uses the cross-entropy cost function, which is arguably better than the Kullback−Leibler divergence of t-SNE in preserving the global structure of the data; this allows us to identify similar and different MOF structures based on the 2D-EH features. MOF structures with similar 2D-EH features are clustered together in the UMAP space, while structures with distinct 2D-EH features are separated from one another.

Figure 10a shows the UMAP embedding space of MOF structures considered for propane adsorption at 1 bar, 298 K, along with the adsorption data, which is shown by the colors of the points. We can see that low-loading regions (blue) are well separated from the high-loading region (red), and there are smooth transitions between regions. This well behaved UMAP space indicates that the 2D-EH features are sufficient to distinguish between similar and different structures in terms of adsorption capacity. This observance supports the good ML regression of the data in Figure 5a. In contrast, for *n*-hexane adsorption at 0.02 bar, 298 K, where systematic deviations are present in the ML prediction, MOF structures with drastically different loadings are clustered together in some parts of the UMAP space, as shown in Figure 10b. This suggests a nonuniqueness (one-to-many) problem in the regression. We selected tobmof-4300 and tobmof-2744 to illustrate this problem. Figure 10c shows that these two structures share very similar (almost identical) 2D energy histograms. In the adsorption isotherms shown in Figure 10d, tobmof-2744 shows a sharp increase in the isotherm before the studied pressure 0.02 bar. This sharp increase corresponds to capillary condensation in pores with 19.9 Å diameter (LCD) where the adsorbed phase goes through a first-order phase transition from a low-density state to a high-density state. In comparison, the abrupt filling in

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

tobmof-4300 (LCD = 15 Å) occurs at a higher pressure and the uptake at 0.02 bar is low. We note that the relation between the condensation pressure and pore diameter does not follow the Kelvin equation.[87,88] This might be attributed to the dependence of capillary condensation on topology and pore connectivity in MOFs. Given similar 2D-EH features for tobmof-2744 and tobmof-4300, our ML model is unable to distinguish them. To compensate the increase in cost function due to different loadings in these two "similar" structures, the trained ML model predicts an average of the two. This also explains the underestimation in ML prediction at the high loading range and overestimation at the low/middle range (Figures S16 and S30). A further analysis confirms that the ML prediction is difficult for MOFs that specifically have pores about 20 Å in diameter (Figure S30); this pore size roughly corresponds to the occurrence of capillary condensation of *n*-hexane at 0.02 bar, 298 K. Moreover, it is possible that a small fraction of GCMC training data is unreliable. Preliminary results indicate that the accurate determination of adsorption amount near condensation pressure is challenging in MOFs using standard GCMC simulations because the system can get stuck in a metastable state.

We also compared the performance of the 2D-EH features with other structural features when capillary condensation is present. Persistent homology is an advanced algebraic method for discerning topological features of data.[16,89] We derived persistent homology from alpha shape filtration,[90] during which the appearance and disappearance of loops and voids inside the alpha shape were recorded. To vectorize the extracted topological information for ML tasks, the persistence image method was adopted.[91] We prepared two types of persistent images for each MOF structure. The first type records the loops in the structure, which reflects the channel/window size and surface texture (e.g., benzene ring). We shall refer to the corresponding persistent image as 1D persistent homology (1D-PH) features. The other type encodes the size of cavities (pores) in the structure along with the size of the narrowest window that is directly connected to the pore. We shall refer to the second type as 2D persistent homology (2D-PH) features. Figure S31 shows an example of the 2D-PH for IRMOF-1. A detailed explanation about the persistent homology and its calculation is available in Section S8.2. For reference, we also included conventional textural properties (TX) in the feature set, i.e., VF, VSA, GSA, PLD, and LCD.

Table 3 summarizes the evaluation metrics of RF predictions for *n*-hexane adsorption at 0.02 bar, 298 K, using different types of features. Overall, the 2D-EH features outperform all structural features and their combination when capillary condensation is present, with an average of 53% and 44% reduction in MAE and RMSE, respectively. Moreover, our 2D-EH features are more efficient than persistent homology features, considering the much smaller size of the 2D-EH features (70) compared to 1D-PH (839) and 2D-PH features (624). It is interesting to note that the RF model based on 1D-PH features has similar performance to those based on 2D-PH features, TX features, and combinations thereof. This indicates that, at low pressure (P/P_0 = 0.1), low-dimensional topological information on channel/window size plays an important role in predicting the adsorption of *n*-hexane. It should be noted that 1D-PH and 2D-PH features share some overlapping information, such as "bottleneck" window size. Another observation is that TX features have similar (and even slightly better) performance than the persistent homology features. Compared to structural features that are derived from purely geometric information,[16] the helium void fraction (VF) calculated in this work also encodes some energetic information (see Section 2.1.3), which is useful for predicting low-pressure loadings. Most importantly, the RF regression model using the full feature set continues to exhibit systematic errors in the prediction (Figure S32). This does not indicate that structural information is not important for predicting the onset of capillary condensation, but rather that simply combining individual feature sets does not provide critical cross-information between the structure and energy. For a system with more training data as in the case of *n*-butane adsorption at 298 K, 1.2 bar, we made similar observations (data available in Table S14). These findings indicate that further research on capillary condensation is needed to guide the design of effective ML features for adsorption prediction.

Table 3. RF Predictions for Adsorption for *n*-Hexane at 298 K, 0.02 bar, Using Different Sets of Features[a]

Feature set	R^2	*MAE*	*MAPE* [%]	*RMSE*
2D-EH	0.95	5.9	19.5	11.3
1D-PH	0.80	14.3	45.5	21.5
2D-PH	0.81	13.4	45.5	21.1
TX	0.83	11.5	42.4	20.1
1D-PH+ 2D-PH	0.83	12.9	44.0	20.1
1D-PH+ 2D-PH+ TX	0.85	11.6	42.2	18.7
2D-EH+ 1D-PH+ 2D-PH+ TX	0.95	5.7	18.6	11.1

[a]Metrics presented are for 1000 testing data points. MAE and RMSE are in units of cm^3_{STP}/cm^3.

3.6. Methods Benchmarking. The ML workflow for materials screening and discovery is significantly more efficient than brute-force GCMC simulations. The following computational time was estimated using a combination of Intel Xeon E5-2650 v3 @ 2.30 GHz and E5-2680 v4 @ 2.40 GHz. To simulate adsorption of simple molecules such as Xe, each GCMC simulation took on average 2 CPU hours to finish. For more complex molecules such as 2,2-dimethylbutane, a single GCMC simulation took on average 25 CPU hours. One can imagine that a screening task using brute force GCMC simulations can become prohibitive as the cost per simulation increases with increasing molecular complexity and when the number of materials increases. For the examples in this work, molecular simulations for 2,000 MOF structures required a total of 4,000 CPU hours for Xe and 50,000 CPU hours for 2,2-dimethylbutane. In contrast, the ML workflow only required about 58 CPU hours to compute the energy and energy gradient grids for 2000 randomly chosen ToBaCCo MOF structures, another 2−12 CPU hours to bin these grids into 2D energy histograms depending on the histogram resolution, and less than 10 min to fit and run the regression analyses (RF model, for example). Once the ML model is validated, the computational cost of prediction for additional structures is trivial (less than a second per material). Thus, to perform the task of screening 2,000 MOF structures, the ML approach (once trained) requires only ∼70 CPU hours versus 4,000 to 50,000 CPU hours for GCMC, a two to 3 orders of magnitude speedup. It is noteworthy that because a single probe sphere (CH_3 group) was used for all alkanes in this work, the computational cost of our ML approach can be further reduced by reusing the grid information from the other alkanes.

Although GCMC simulations in this work were performed in serial, the computational advantage of the ML workflow still

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

holds in comparison to a parallel GCMC implementation based on graphics processing unit (GPU). The publicly available, GPU-based code, GOMC,[92] is reported to accelerate a GCMC simulation by a factor of 2−13, which is much less than the several orders of magnitude speedup achieved by the ML workflow. In addition, the efficiency of the parallel simulation heavily relies on the implementation of the simulation algorithm and the structure of the code. In some cases, the GOMC code was found to be even slower than the serial RASPA code used in this work, such as for CO_2 adsorption in IRMOF-1 with only LJ interactions.[92]

It is also useful to consider the computational cost of our ML approach relative to other similar approaches. Achieving a higher predictive accuracy with our 2D-EH features compared to 1D-EH features[35] required only a 11% increase in wall time to calculate the additional energy gradient (Figure S22). Our current code handles the energy and energy gradient calculations separately, so additional efficiencies reducing the extra time associated with the 2D-EH features would be achieved by grouping common calculation tasks for obtaining the energy and energy gradient.

4. CONCLUSIONS

In this work, we have proposed and tested 2D energy histograms (2D-EH) as features for ML regression models to predict adsorption in nanoporous materials. The 2D-EH is constructed by first placing a probe particle (a spherical atom or a united-atom methyl group for alkane molecules) at evenly spaced points throughout the adsorbent to calculate the energy and energy gradient at these grid points. These grid results are binned together into a 2D histogram, which is further flattened to be used as a material input representation for ML models. We found that compared to the previous 1D-EH version where only the energy is considered, including energy gradients into the histogram features retains important information on the 3D energy landscape of the adsorbent−adsorbate system and thus improves the ML predictions of adsorption capacity.

We have demonstrated the effectiveness of 2D-EH features by training multiple ML models to predict single-component adsorption in MOFs. Here we considered spherical species (Kr and Xe), linear alkanes with a wide range of aspect ratios (ethane, propane, *n*-butane, and *n*-hexane), and a branched alkane (2,2-dimethylbutane) over a wide range of temperatures and pressures. We found using 2D-EH features with the LASSO model significantly improves the predictive accuracy for spherical molecules compared to that using 1D-EH features. For adsorption of alkanes, nonlinear ML models employed in this work (RF, XGB, and MLP) all show highly accurate predictions with $R^2 \sim 0.94-0.99$. The advantage of 2D-EH features over baseline features that consist of textural properties (VF, VSA, GSA, LCD, PLD) and Henry's constant has also been demonstrated, with significant improvement found at low pressure range.

Physically, each 2D-EH feature (or histogram pixel) represents the volume fraction of a well-defined region in the adsorbent, and the definition of regions is roughly based on the distance to the adsorbent surface. This sensible way to decompose the space allows a ML model to learn some basic adsorption physics from the training data. In addition, we have shown that it is possible to correlate a linear combination of 2D-EH features to the textural properties of the material, such as void fraction and surface area, again showing that 2D-EH features encode both structural and energetic information on the adsorption system. Our ML models have also been shown to be generalizable and transferrable. ML models that were trained with only MOF data show excellent predictive capabilities for adsorption in unseen amorphous porous materials, including hyper-cross-linked polymers, polymers of intrinsic microporosity, activated carbons, and kerogens. Future work will focus on extending 2D-EH features to polar molecules, such as water and CO_2, where Coulombic interactions and molecular orientations play an essential role in the adsorption.

Finally, we identified a challenge for future ML studies in adsorption systems, namely, making predictions for systems displaying capillary condensation and hysteresis. The insufficiency of our 2D-EH features at deep subcritical temperature has been elucidated in the low-dimensional UMAP space. Nevertheless, our 2D-EH features still perform better than structural features such as those derived from persistent homology. Further investigations on capillary condensation in MOFs are needed to solve this challenge.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00798.

Details about feature calculation, ML methods and hyperparameters tuning, additional parity plots comparing ML prediction and GCMC results, persistent homology calculation, and other supporting figures and tables (PDF)

GCMC data for adsorption in MOFs and in APMs (XLSX)

Textural properties of APMs (XLSX)

Textural properties of ToBaCCo 1.0 MOFs (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Randall Q. Snurr − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0003-2925-9246; Email: snurr@northwestern.edu

Authors

Kaihang Shi − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0002-0297-1746

Zhao Li − *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0001-5035-4614

Dylan M. Anstine − *Department of Materials Science and Engineering, University of Florida, Gainesville, Florida 32611, United States; George and Josephine Butler Polymer Research Laboratory, University of Florida, Gainesville, Florida 32611, United States*

Dai Tang − *School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Coray M. Colina − *Department of Materials Science and Engineering, University of Florida, Gainesville, Florida 32611, United States; George and Josephine Butler Polymer Research Laboratory and Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0003-2367-1352

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

David S. Sholl − *School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; Transformational Decarbonization Initiative, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0002-2771-9168

J. Ilja Siepmann − *Department of Chemistry and Chemical Theory Center, University of Minnesota, Minneapolis, Minnesota 55455, United States; Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455, United States;* orcid.org/0000-0003-2534-4507

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00798

Funding

This research was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences, under Award DE-FG02-17ER16362. Simulations in this work were made possible by the high-performance computing facility Quest at Northwestern University, the high-performance computing services core facility (RRID:SCR_022168) provided by North Carolina State University, the high-performance computing system HiPerGator 2.0 at the University of Florida, and the Hive cluster at the Georgia Institute of Technology which is supported by the National Science Foundation under Grant Number 1828187. Simulations in this research were supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, U.S.A. This research also used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC Award BES-ERCAP0020094.

Notes

The authors declare no competing financial interest.

Code and example simulation files are available at https://github.com/snurr-group/2D-energy-histogram. Other associated data including trained ML models and CIF files for APMs are available at 10.5281/zenodo.5481697.

REFERENCES

(1) Gupta, V. K.; Saleh, T. A. Sorption of Pollutants by Porous Carbon, Carbon Nanotubes and Fullerene- An Overview. *Environ. Sci. Pollut. Res.* **2013**, *20* (5), 2828−2843.

(2) Zhou, H.-C.; Long, J. R.; Yaghi, O. M. Introduction to Metal-Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 673−674.

(3) Sanders, D. F.; Smith, Z. P.; Guo, R.; Robeson, L. M.; McGrath, J. E.; Paul, D. R.; Freeman, B. D. Energy-Efficient Polymeric Gas Separation Membranes for a Sustainable Future: A Review. *Polymer (Guildf).* **2013**, *54* (18), 4729−4761.

(4) Wang, S.; Pomerantz, N. L.; Dai, Z.; Xie, W.; Anderson, E. E.; Miller, T.; Khan, S. A.; Parsons, G. N. Polymer of Intrinsic Microporosity (PIM) Based Fibrous Mat: Combining Particle Filtration and Rapid Catalytic Hydrolysis of Chemical Warfare Agent Simulants into a Highly Sorptive, Breathable, and Mechanically Robust Fiber Matrix. *Mater. Today Adv.* **2020**, *8*, 100085.

(5) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal-Organic Frameworks. *Nat. Chem.* **2012**, *4* (2), 83−89.

(6) 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.

(7) Sun, Y.; DeJaco, R. F.; Li, Z.; Tang, D.; Glante, S.; Sholl, D. S.; Colina, C. M.; Snurr, R. Q.; Thommes, M.; Hartmann, M.; Siepmann, J. I. Fingerprinting Diverse Nanoporous Materials for Optimal Hydrogen Storage Conditions Using Meta-Learning. *Sci. Adv.* **2021**, *7* (30), No. eabg3983.

(8) Ahmed, A.; Siegel, D. J. Predicting Hydrogen Storage in MOFs via Machine Learning. *Patterns* **2021**, *2* (7), 100291.

(9) Rahimi, M.; Moosavi, S. M.; Smit, B.; Hatton, T. A. Toward Smart Carbon Capture with Machine Learning. *Cell Reports Phys. Sci.* **2021**, *2* (4), 100396.

(10) Norman, G. E.; Filinov, V. S. Investigations of Phase Transitions by a Monte-Carlo Method. *High Temp.* **1969**, *7* (2), 216−222.

(11) Panagiotopoulos, A. Z.; Quirke, N.; Stapleton, M.; Tildesley, D. J. Phase Equilibria by Simulation in the Gibbs Ensemble. *Mol. Phys.* **1988**, *63* (4), 527−545.

(12) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120* (16), 8066−8129.

(13) 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.

(14) Fernandez, M.; Barnard, A. S. Geometrical Properties Can Predict CO_2 and N_2 Adsorption Performance of Metal-Organic Frameworks (MOFs) at Low Pressure. *ACS Comb. Sci.* **2016**, *18* (5), 243−252.

(15) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29* (7), 2844−2854.

(16) 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.

(17) 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.

(18) Martin, R. L.; Smit, B.; Haranczyk, M. Addressing Challenges of Identifying Geometrically Diverse Sets of Crystalline Porous Materials. *J. Chem. Inf. Model.* **2012**, *52* (2), 308−318.

(19) Zhang, X.; Cui, J.; Zhang, K.; Wu, J.; Lee, Y. Machine Learning Prediction on Properties of Nanoporous Materials Utilizing Pore Geometry Barcodes. *J. Chem. Inf. Model.* **2019**, *59* (11), 4636−4644.

(20) Krishnapriyan, A. S.; Montoya, J.; Haranczyk, M.; Hummelshøj, J.; Morozov, D. Machine Learning with Persistent Homology and Chemical Word Embeddings Improves Prediction Accuracy and Interpretability in Metal-Organic Frameworks. *Sci. Rep.* **2021**, *11* (1), 8888.

(21) 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.

(22) 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.

(23) Zhang, Z.; Schott, J. A.; Liu, M.; Chen, H.; Lu, X.; Sumpter, B. G.; Fu, J.; Dai, S. Prediction of Carbon Dioxide Adsorption via Deep Learning. *Angew. Chem.* **2019**, *131* (1), 265−269.

(24) Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb. Sci.* **2017**, *19* (10), 640−645.

(25) 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.

(26) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO 2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30* (18), 6325−6337.
(27) Wu, X.; Xiang, S.; Su, J.; Cai, W. Understanding Quantitative Relationship between Methane Storage Capacities and Characteristic Properties of Metal-Organic Frameworks Based on Machine Learning. *J. Phys. Chem. C* **2019**, *123* (14), 8550−8559.
(28) 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.
(29) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal-Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19* (11), 6682−6697.
(30) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3* (1), 76−86.
(31) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Framework Ecosystem. *Nat. Commun.* **2020**, *11* (1), 4068.
(32) 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.
(33) 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.
(34) 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.
(35) 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.
(36) 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. *J. Chem. Phys.* **2021**, *155* (1), 014701.
(37) 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.
(38) Sigmund, G.; Gharasoo, M.; Hüffer, T.; Hofmann, T. Deep Learning Neural Network Approach for Predicting the Sorption of Ionizable and Polar Organic Pollutants to a Wide Range of Carbonaceous Materials. *Environ. Sci. Technol.* **2020**, *54* (7), 4583−4591.
(39) Yang, C.; Kaipa, U.; Mather, Q. Z.; Wang, X.; Nesterov, V.; Venero, A. F.; Omary, M. A. Fluorous Metal-Organic Frameworks with Superior Adsorption and Hydrophobic Properties toward Oil Spill Cleanup and Hydrocarbon Storage. *J. Am. Chem. Soc.* **2011**, *133* (45), 18094−18097.
(40) 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.
(41) Sholl, D. S.; Lively, R. P. Seven Chemical Separations to Change the World. *Nature* **2016**, *532* (7600), 435−437.
(42) 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.
(43) Gharagheizi, F.; Tang, D.; Sholl, D. S. Selecting Adsorbents to Separate Diverse Near-Azeotropic Chemicals. *J. Phys. Chem. C* **2020**, *124* (6), 3664−3670.
(44) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal-Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17* (11), 5801−5810.
(45) Bobbitt, N. S.; Shi, K.; Bucior, B. J.; Chen, H.; Tracy-Amoroso, N.; Li, Z.; Sun, Y.; Merlin, J. H.; Siepmann, J. I.; Siderius, D. W.; Snurr, R. Q.MOFX-DB: An Accessible Online Database of Computational Adsorption Data for Nanoporous Materials. *J. Chem. Eng. Data***2023**, in press. DOI: 10.1021/acs.jced.2c00583
(46) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: Molecular Simulation Software for Adsorption and Diffusion in Flexible Nanoporous Materials. *Mol. Simul.* **2016**, *42* (2), 81−101.
(47) *RASPA2*. https://github.com/iRASPA/RASPA2 (accessed May 16, 2022).
(48) Martin, M. G.; Siepmann, J. I. Novel Configurational-Bias Monte Carlo Method for Branched Molecules. Transferable Potentials for Phase Equilibria. 2. United-Atom Description of Branched Alkanes. *J. Phys. Chem. B* **1999**, *103* (21), 4508−4517.
(49) Talu, O.; Myers, A. L. Reference Potentials for Adsorption of Helium, Argon, Methane, and Krypton in High-Silica Zeolites. *Colloids Surfaces A Physicochem. Eng. Asp.* **2001**, *187*−*188*, 83−93.
(50) Hirschfelder, J. O.; Curtiss, C. F.; Bird, R. B.*Molecular Theory of Gases and Liquids*; Wiley: New York, 1964; DOI: 10.1063/1.3061949.
(51) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024−10035.
(52) Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der Kinetischen Theorie der Gase. *Ann. Phys.* **1881**, *248* (1), 127−136.
(53) Berthelot, D. Sur Le Mélange Des Gaz. *Compt. Rendus* **1898**, *126*, 1703−1706.
(54) Shi, W.; Maginn, E. J. Continuous Fractional Component Monte Carlo: An Adaptive Biasing Method for Open System Atomistic Simulations. *J. Chem. Theory Comput.* **2007**, *3* (4), 1451−1463.
(55) Yu, Z.; Anstine, D. M.; Boulfelfel, S. E.; Gu, C.; Colina, C. M.; Sholl, D. S. Incorporating Flexibility Effects into Metal-Organic Framework Adsorption Simulations Using Different Models. *ACS Appl. Mater. Interfaces* **2021**, *13* (51), 61305−61315.
(56) Widom, B. Some Topics in the Theory of Fluids. *J. Chem. Phys.* **1963**, *39* (11), 2808−2812.
(57) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134−141.
(58) Bae, Y.-S.; Yazaydın, A. Ö.; Snurr, R. Q. Evaluation of the BET Method for Determining Surface Areas of MOFs and Zeolites That Contain Ultra-Micropores. *Langmuir* **2010**, *26* (8), 5475−5483.
(59) 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.
(60) Shi, K.; Shen, Y.; Santiso, E. E.; Gubbins, K. E. Microscopic Pressure Tensor in Cylindrical Geometry: Pressure of Water in a Carbon Nanotube. *J. Chem. Theory Comput.* **2020**, *16* (9), 5548−5561.
(61) James, G.; Witten, D.; Hastie, T.; Tibshirani, R.*An Introduction to Statistical Learning*; Springer Texts in Statistics; Springer New York: New York, NY, 2013; Vol. *103*; DOI: 10.1007/978-1-4614-7138-7.
(62) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58* (1), 267−288.
(63) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(64) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189−1232, DOI: 10.1214/aos/1013203451.

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(65) Géron, A.*Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed.; O'Reilly Media, Inc.: 2019.
(66) Chen, T.; Guestrin, C.XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, U.S.A., 2016; Vol. *13-17-Augu*, pp 785−794, DOI: 10.1145/2939672.2939785.
(67) Geeen, K.; Tashman, L. Percentage Error: What Denominator? *Foresight Int. J. Appl. Forecast.* **2009**, No. 12, 36−40.
(68) Bobbitt, N. S.; Chen, J.; Snurr, R. Q. High-Throughput Screening of Metal-Organic Frameworks for Hydrogen Storage at Cryogenic Temperature. *J. Phys. Chem. C* **2016**, *120* (48), 27328−27341.
(69) Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* **1933**, *24* (6), 417−441.
(70) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33−38.
(71) Kupgan, G.; Liyana-Arachchi, T. P.; Colina, C. M. Pore Size Tuning of Poly(Styrene-Co-Vinylbenzyl Chloride-Co-Divinylbenzene) Hypercrosslinked Polymers: Insights from Molecular Simulations. *Polymer (Guildf).* **2016**, *99*, 173−184.
(72) Hart, K. E.; Colina, C. M. Estimating Gas Permeability and Permselectivity of Microporous Polymers. *J. Membr. Sci.* **2014**, *468*, 259−268.
(73) Rukmani, S. J.; Liyana-Arachchi, T. P.; Hart, K. E.; Colina, C. M. Ionic-Functionalized Polymers of Intrinsic Microporosity for Gas Separation Applications. *Langmuir* **2018**, *34* (13), 3949−3960.
(74) Thyagarajan, R.; Sholl, D. S. A Database of Porous Rigid Amorphous Materials. *Chem. Mater.* **2020**, *32* (18), 8020−8033.
(75) Farmahini, A. H.; Opletal, G.; Bhatia, S. K. Structural Modelling of Silicon Carbide-Derived Nanoporous Carbon by Hybrid Reverse Monte Carlo Simulation. *J. Phys. Chem. C* **2013**, *117* (27), 14081−14094.
(76) Nguyen, T. X.; Cohaut, N.; Bae, J.-S.; Bhatia, S. K. New Method for Atomistic Modeling of the Microstructure of Activated Carbons Using Hybrid Reverse Monte Carlo Simulation. *Langmuir* **2008**, *24* (15), 7912−7922.
(77) Farmahini, A. H.; Bhatia, S. K. Effect of Structural Anisotropy and Pore-Network Accessibility on Fluid Transport in Nanoporous Ti3SiC2 Carbide-Derived Carbon. *Carbon N. Y.* **2016**, *103*, 16−27.
(78) de Tomas, C.; Suarez-Martinez, I.; Marks, N. A. Graphitization of Amorphous Carbons: A Comparative Study of Interatomic Potentials. *Carbon N. Y.* **2016**, *109*, 681−693.
(79) de Tomas, C.; Suarez-Martinez, I.; Vallejos-Burgos, F.; López, M. J.; Kaneko, K.; Marks, N. A. Structural Prediction of Graphitization and Porosity in Carbide-Derived Carbons. *Carbon N. Y.* **2017**, *119*, 1−9.
(80) de Tomas, C.; Suarez-Martinez, I.; Marks, N. A. Carbide-Derived Carbons for Dense and Tunable 3D Graphene Networks. *Appl. Phys. Lett.* **2018**, *112* (25), 251907.
(81) Powles, R. C.; Marks, N. A.; Lau, D. W. M. Self-Assembly of Sp2-Bonded Carbon Nanostructures from Amorphous Precursors. *Phys. Rev. B* **2009**, *79* (7), 075430.
(82) Bousige, C.; Ghimbeu, C. M.; Vix-Guterl, C.; Pomerantz, A. E.; Suleimenova, A.; Vaughan, G.; Garbarino, G.; Feygenson, M.; Wildgruber, C.; Ulm, F.-J.; Pellenq, R. J.-M.; Coasne, B. Realistic Molecular Model of Kerogen's Nanostructure. *Nat. Mater.* **2016**, *15* (5), 576−582.
(83) Kupgan, G.; Demidov, A. G.; Colina, C. M. Plasticization Behavior in Polymers of Intrinsic Microporosity (PIM-1): A Simulation Study from Combined Monte Carlo and Molecular Dynamics. *J. Membr. Sci.* **2018**, *565*, 95−103.
(84) Anstine, D. M.; Tang, D.; Sholl, D. S.; Colina, C. M. Adsorption Space for Microporous Polymers with Diverse Adsorbate Species. *npj Comput. Mater.* **2021**, *7* (1), 53.
(85) McInnes, L.; Healy, J.; Melville, J.UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2018, DOI: 10.48550/arXiv.1802.03426.
(86) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2625.
(87) Thomson, W. LX. On the Equilibrium of Vapour at a Curved Surface of Liquid. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1871**, *42* (282), 448−452.
(88) Shi, K.; Santiso, E. E.; Gubbins, K. E. Current Advances in Characterization of Nano-Porous Materials: Pore Size Distribution and Surface Area. *Porous Materials: Theory and Its Application for Environmental Remediation* **2021**, 315−340.
(89) Edelsbrunner, H.; Harer, J. Persistent Homology-a Survey. *Surveys on Discrete and Computational Geometry* **2008**, *453*, 257−282.
(90) Edelsbrunner, H.*Alpha Shapes-a Survey*; 2009.
(91) Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushtanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L. Persistence Images: A Stable Vector Representation of Persistent Homology. *J. Mach. Learn. Res.* **2017**, *18*, 1−35.
(92) Nejahi, Y.; Soroush Barhaghi, M.; Mick, J.; Jackman, B.; Rushaidat, K.; Li, Y.; Schwiebert, L.; Potoff, J. GOMC: GPU Optimized Monte Carlo for the Simulation of Phase Equilibria and Physical Properties of Complex Fluids. *SoftwareX* **2019**, *9*, 20−27.

https://doi.org/10.1021/acs.jctc.2c00798
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX