Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00027d

Rapid prediction of single-site adsorbate probability distributions in metal–organic frameworks using graph neural networks

Jake Burner, Olivier Marchand, Rosa Cicciarella, Marco Gibaldi and Tom K. Woo*

Metal–organic frameworks (MOFs) are porous crystalline materials assembled from inorganic nodes and organic linkers. These materials have garnered significant interest for gas separation and storage applications, particularly because of their porosity and their tunability due to their massive design space. However, navigating such a massive design space poses significant challenges. Atomistic simulation techniques have been applied to accelerate discovery and design of MOFs for various applications. A key property obtained from these simulations is the adsorbate probability distribution (APD). An APD maps the probability of finding an adsorbate molecule in the pore of a MOF at a given temperature and pressure, whose maxima correspond to free energy minima (*i.e.*, binding sites). While APDs and binding sites are not easily accessible experimentally, their generation *via* simulation is tractable. However, high-throughput generation of APDs still requires long simulation times to converge. A machine learning (ML) model to predict APDs would enable the use of this property in data-driven pipelines to identify high performing materials or binding sites. To date, nobody has attempted to apply ML to the prediction of APDs or binding sites of MOFs. In this work, we present DeepAPD – an ML model which predicts APDs at a given temperature and pressure. As an initial proof of concept, the model has been trained on simple spherical adsorbates such as CH_4 and Xe. DeepAPD was found to generate APDs of MOFs at a speedup factor of $>10^5$ in comparison to GCMC. An in-depth discussion of training strategies and dataset size/composition on model performance is presented. It was found that the APDs obtained by ML were sufficiently accurate to get a reliable estimation of binding sites in MOFs, particularly binding sites which have high probability. Finally, the transferability of the ML models was investigated by evaluating the performance of the GNN model on a dataset of experimentally characterized MOFs. We have also implemented the DeepAPD inference code into our binding site identification algorithm to facilitate an end-to-end MOF to binding site prediction. Future work will extend these models to more complex guests such as CO_2, N_2, and H_2O.

Received 20th January 2026
Accepted 8th May 2026

DOI: 10.1039/d6dd00027d

rsc.li/digitaldiscovery

Introduction

Metal–organic frameworks (MOFs) are porous crystalline materials self-assembled from inorganic nodes and organic linkers, which have garnered significant interest for gas separation and storage applications such as CO_2 capture,[1,2] Xe/Kr separations,[3–6] and methane storage.[7,8] Due to their extensive design space, identifying high-performing MOFs for a given separation process is not straightforward. To this end, binding sites of adsorbates in MOFs give not only atomistic insights into the nature of adsorption but also provide a useful synthetic target for high-performing materials. For example, data-driven identification of binding sites in MOFs has resulted in the synthesis of new materials for CO_2 capture from humid flue gas.[9]

A primary obstacle to data mining physisorptive binding sites of MOFs is the lack of experimental data, since such experiments usually rely on complex and challenging techniques. *In situ* neutron powder diffraction (NPD) is a popular method to directly observe the position of adsorbate molecules within MOF pores,[10–12] but access to neutron facilities is limited.[13] While single-crystal XRD (SC-XRD) can be used as an alternative,[14–16] obtaining a single-crystal of a MOF is far from straightforward. Even where SC-XRD is tractable, resolving the location of adsorbates in the pores can be challenging or impossible, particularly for highly delocalized sites where higher resolution synchrotron radiation may be required.[14] For these reasons, direct experimental observation of adsorbates in MOFs is relatively rare (often relying on computational methods) and is usually limited to single-component gas

Department of Chemistry and Biomolecular Sciences, University of Ottawa, 10 Marie Curie Private, Ottawa K1N 6N5, Canada. E-mail: Tom.Woo@uottawa.ca

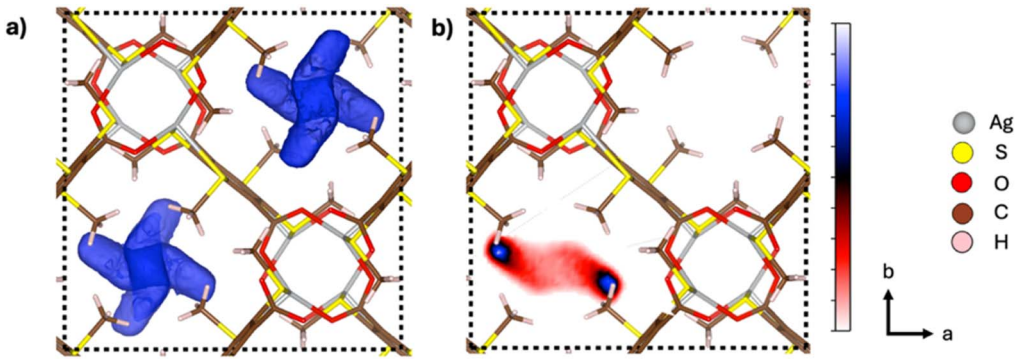

Fig. 1 A GCMC-simulated adsorbate probability distribution (APD) visualized as both (a) a 3D isosurface and (b) a 2D contour plot. The MOF is IJENER_freeONLY from the CoRE MOF database[23] and the adsorbate is methane at 1 bar, 298 K. An isosurface value of 10% of the global maximum was chosen for visualization of the isosurface and the slice through the *c* axis in the contour plot is through the global maximum. The APD was simulated at 1 bar and 298 K. The structures were visualized in VESTA.[24]

streams.[13] Other characterization techniques have been used in the literature to probe adsorbate dynamics in MOFs (*e.g.*, solid-state NMR[17,18]), but still rely to some extent on other characterization techniques and computational methods to elucidate the binding sites.

Atomistic simulations provide an alternative route to binding site identification in MOFs. For example, molecular dynamics (MD) simulations of MOFs have routinely been applied at varying gas loadings to provide insight into binding sites and adsorbate dynamics. So-called grand canonical Monte Carlo (GCMC) simulations have been shown to accurately reproduce experimental adsorption properties of MOFs when the crystal structure is known.[19,20] In such simulations, the system is stochastically perturbed with pre-defined moves (*e.g.*, insertion, deletion, translation, rotation of gas molecules), generating new configurations. Configurations are accepted based on a Boltzmann-weighted probability, biasing acceptance to low-energy configurations. Once equilibrated, statistics are collected from accepted configurations. Over the course of the production phase, adsorbate positions are tracked on a 3D voxel grid, generating an adsorbate probability distribution (APD), where the maxima of the distribution correspond to free energy minima (binding sites) in the MOF pores (Fig. 1). Besides being used to identify binding sites in a material, similarities of APDs between different adsorbates have been used to predict whether a MOF will meet process performance targets for gas separations.[21] This property could also be useful for estimating diffusion rates of MOFs.[22]

However, obtaining well-resolved APDs from GCMC simulations can be compute-intensive depending on the free energy landscape of the material for a given adsorbate, temperature, and pressure. While this is not restrictive for small-scale screenings, it is intractable for larger databases of materials and more complex adsorbates. Therefore, considering the utility of this property, an accurate and rapid method for generating these APDs for MOFs would be highly advantageous. To date, there has been no published attempt to apply machine learning (ML) to predict this property of MOFs and no public database of APDs exists. Such ML models would obviate the need for long GCMC simulations and enable rapid generation of APDs and binding sites for data mining, greatly accelerating MOF design for gas separation or storage applications.

There are a few important conditions that an ML model for generating APDs should satisfy. First, the model should be resolution-independent (*i.e.*, applicable to grids of more than one shape). The reason for this is to maintain a constant spatial resolution between all simulated APDs, which results in vastly different grid shapes across the various cell volumes. The second is the model should ideally not rely on any predefined basis to allow for a wider functional flexibility when learning the APDs. APDs are generically classified as "volumetric data", where data are stored on a grid (voxels), and each voxel corresponds to a value. Notably, at a spatial resolution of 0.15 Å, grid sizes can range anywhere from 10^5 to 10^7 voxels per MOF depending on cell volume, which poses practical challenges related to both storage requirements and efficient ML training. Volumetric data are ubiquitous in computational chemistry, and are particularly common in quantum chemistry (*e.g.*, total electron densities). For this reason, we looked to the quantum chemistry machine learning literature for models which may be adapted to predict APDs. Notably, many traditional approaches to predicting electron densities often rely on kernel regression to predict basis coefficients, where computational complexity grows cubically with number of training examples, or convolutional neural networks (CNNs) which require a fixed grid shape. The equivariant graph neural network (GNN) model "DeepDFT",[25] which has been previously applied to predict ground state electron densities, avoids these challenges and was adapted for use in the present work since it satisfies the criteria outlined above. For further discussion and introduction to GNNs in materials chemistry, the reader is directed to informative reviews on the topic.[26]

As an initial proof of concept, this work demonstrates the application of graph neural networks (GNNs) to the prediction of APDs of simple adsorbates in MOFs. We define simple adsorbates as ones which can be reliably modeled using a single-site force field model. Methane (CH_4) was chosen for its relevance to separation and storage applications, while xenon (Xe) was chosen as a more strongly interacting single-site adsorbate. About 23 thousand MOFs were sampled from the

ARC-MOF database[27] and APDs of Xe and CH_4 were simulated at various pressures. Using these data, a GNN model was trained to predict APDs of each adsorbate at its corresponding state point (T, p). State points for methane were chosen to reflect conditions relevant to methane storage[7,8,19] (65 bar, 1 bar) at 298 K, while 1 bar was chosen for xenon at 298 K, which is close to the conditions relevant to Xe/Kr separations[3–5,28] (20 : 80 Xe : Kr at 1 bar). Therefore, the chosen conditions are relevant to practical separation processes. The models were implemented by adapting the recently published DeepDFT model,[25] which was applied to predict ground-state electron densities of molecules and solid-state materials. The utility of the ML-predicted APDs for binding site identification was assessed by performing an in-depth analysis of the binding sites extracted from ML *vs.* simulated APDs. It was found that the APDs obtained by ML were sufficiently accurate to get a reliable estimation of binding sites in MOFs, particularly binding sites which have high probability or what we will refer to as high occupancy. Finally, the transferability of the ML models was investigated by evaluating the performance of the GNN model on a new dataset of experimentally characterized MOFs from the MOSAEC-DB database.[29] The GNN model has been integrated into our in-house binding site identification algorithm (GALA)[21] to enable user-friendly determination of binding sites without requiring any simulation steps.

Methodology

Adsorbate probability distribution (APD) database

The training and development sets for the GNN models trained in this work were obtained by randomly sampling the ARC-MOF database (version 6.0).[27,30] ARC-MOF was chosen for its wide diversity of organic structural building units (SBUs) and geometries. Overall, APDs were obtained for ~23 K MOFs at each state point for each adsorbate (a total of ~70 K APDs). We hypothesized that the organic SBU chemistry and geometry would have the largest effect on the binding sites of CH_4 and Xe relative to other parameters such as the metal SBU chemistry, and therefore justify the use of random sampling considering the good variety and balance of ARC-MOF in this respect.[27] The set of 23 K MOFs was randomly split 85 : 15 into training and development sets, respectively. The test set was obtained by sampling the MOSAEC-DB database.[29] Only MOFs which were (a) neutral; (b) 3-dimensional; (c) unique; and (d) porous (taken to have a pore limiting diameter >2.4 Å) were sampled from MOSAEC-DB (see the SI for more details).

Convergence of the probability distributions was determined by evaluating the similarity of the APDs between replicate unit cells within each simulation cell. The similarity was computed using the Tanimoto coefficient[31] (a generalization of the Jaccard index[32]), which is commonly used to determine the similarity of two equally sized sets. The Tanimoto coefficient is shown in eqn (1) (where $\odot$ denotes the Hadamard product, or element-wise multiplication) for the similarity between two distributions (tensors) **A** and **B**, where $T : (\mathbf{A}, \mathbf{B}) \rightarrow [0, 1], \mathbf{A}, \mathbf{B} \varepsilon \mathbb{R}^{n_x \times n_y \times n_z}$ yields a value of 0 for two completely dissimilar distributions and a value of 1 for two identical distributions. A convergence criterion of $T \geq 0.75$ was used to classify a probability distribution as being sufficiently converged. For simulation cells composed of more than two unit cells, an average coefficient was computed using Tanimoto between all combinations of replicate unit cells in the simulation cell. To reduce stochastic noise of the probability distributions and conserve storage space, each converged probability distribution was "folded" into a single unit cell by averaging the probability density across all replicate unit cells (see the SI for more details). The training, development, and test sets (structures, APDs, and extracted binding sites for each adsorbate and state point) are provided in the SI.

$$T = \frac{\sum A \odot B}{\sum A \odot A + \sum B \odot B - \sum A \odot B} \quad (1)$$

Graph neural networks (GNNs)

The equivariant GNN model "DeepDFT",[25] which has been previously applied to predict ground state electron densities, was used in the present work for reasons outlined in the Introduction. The DeepDFT implementation by Sunshine *et al.*[33] which relies on the ocpmodels Python package (version 0.1.0) was adapted for use in the present paper. This section gives only a brief description of the model and any changes that were implemented, as it is described in detail elsewhere.[25,33] In this work, the modified DeepDFT model will be referred to as the DeepAPD model. The model constructs a graph in which the atoms of the unit cell are the nodes and the edges are determined by a cutoff distance (accounting for periodic boundary conditions), which in this work was optimized to be 6 Å. Associated with each node in the graph is a hidden state which consists of both scalar and vectorial features to ensure E(3) equivariance (*i.e.*, a rotation of a set of nodes will result in an equivalent rotation of their hidden states). The scalar and vectorial hidden states are updated by message passing (aggregating incoming messages from connected edges) using the E(3) equivariant polarizable atom interaction neural network (PaiNN),[34] which characterizes the local environment of each node as an F-dimensional vector, represented as $\varepsilon_i \in \mathbb{R}^{F \times 1}$. The number of rounds of message passing is referred to as the number of interaction layers and is a tunable hyperparameter. The scalar hidden states (s_i) are initialized with trainable embeddings (a_{Z_i}) corresponding to atom type (Z_i), $s_i^0 = a_{Z_i} \in \mathbb{R}^{F \times 1}$, while the vectorial states ($v \rightarrow i$) are initialized as $\vec{v}_i^0 = \vec{0} \in \mathbb{R}^{F \times 3}$.

We wish to tune the trainable parameters of the network to maximize the similarity between the predicted and target probabilities. To do this, DeepAPD makes use of so-called "probe nodes" which do not correspond to any real atoms but are rather a way of inferencing or learning the probability corresponding to a particular point in 3D space. This is done by ensuring that probe nodes (identified by $Z_i = 0$) only receive messages from nodes corresponding to real atoms ($Z_i > 0$). Thus, inserting probe nodes at any points in 3D space with any spatial density of probe nodes will have no impact on the hidden states of the atom nodes. Once trained, the probe nodes are inserted on a regular grid and the resulting probe node

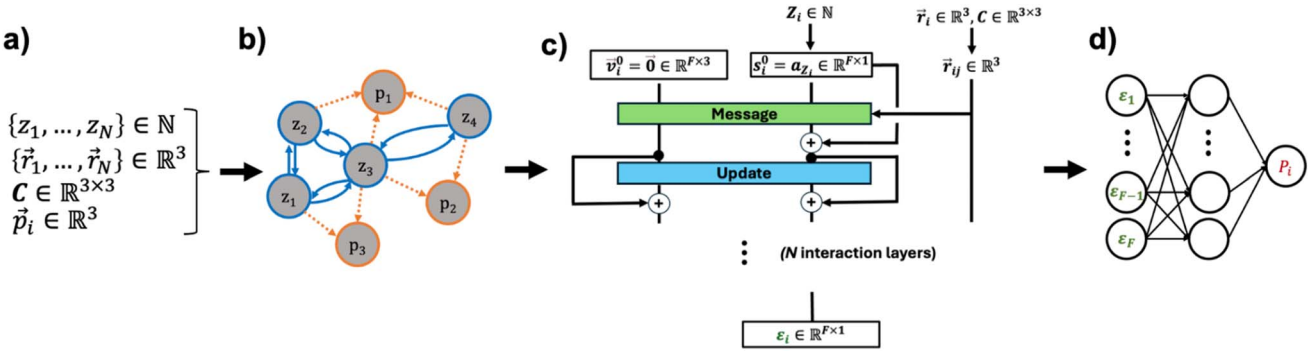

Fig. 2 Schematic showing the DeepAPD architecture. The model takes (a) the atom types/positions, cell vectors, and probe positions as input to (b) construct the structure graph based on a radial cutoff distance, where the probe-atom edges are unidirectional, and atom–atom edges are bidirectional. Messages are constructed and passed/updated N times using the PaiNN interaction network to obtain high-dimensional embeddings ε_i for each probe node in the graph. The embeddings are fed into a (d) final adsorbate-specific readout network to obtain a probability, P_i, at probe position $p \rightarrow i$. The atom nodes ($Z > 0$) and probe nodes ($Z = 0$) use the same interaction network architecture shown in (c), but do not share the same network weights.

embeddings, $\varepsilon_i \in \mathbb{R}^{F\times1}$ if $Z_i = 0$, are fed into an adsorbate-specific 3-layer feedforward neural network (readout network) with F nodes per layer. The readout networks use dropout, an exponential linear unit (ELU) activation function between each layer, and a final softplus activation function to yield a final scalar probability value (P_i) corresponding to probe node i. Each generated APD was normalized to unity. Atom and probe messages are constructed using the same interaction network, but do not share the same network weights. The full model architecture is shown in Fig. 2.

Several changes were made in comparison to previous studies using this ML architecture. First, instead of using mean squared error as the loss function, the negative Tanimoto coefficient (eqn (2)) was minimized, as it was found to slightly improve model performance on the development set, and was more stable during training. Each training batch consisted of only a single MOF, with 6000 targets sampled, where a target corresponds to a single probability value associated with a voxel. Each batch was normalized to unity. Edges were computed considering only atom–atom and atom–probe pairs. Furthermore, during training, probes were not sampled randomly but were rather sampled according to the maxima in each APD since we wish to capture the maxima in the distribution. The maxima were identified according to our in-house binding site localization algorithm, which is described in the Binding site identification section, using an occupancy cutoff of 10%. The probes were then sampled randomly from regions of 20 × 20 × 20 voxels (3.0 Å × 3.0 Å × 3.0 Å) around each maximum for 90% of the probes, and the final 10% of probes being randomly sampled from the grid. Probes were allowed to overlap with framework atoms. Where the models were trained simultaneously on more than one adsorbate (multitask models), the sampled probes were chosen based on the APD of a randomly chosen adsorbate per batch. The loss of multitask models was evaluated by separately evaluating the loss for each adsorbate in each batch, then adding the losses together, and

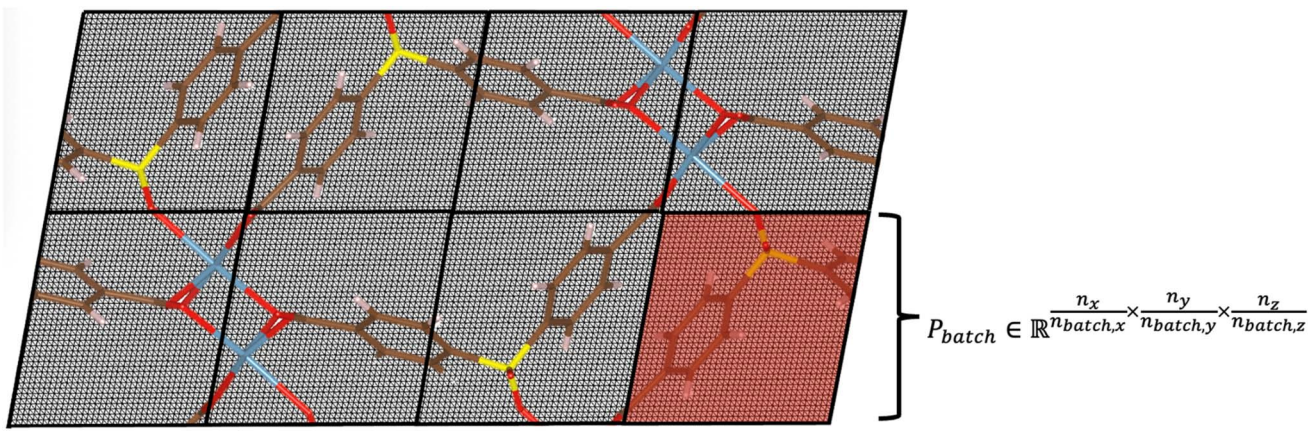

Fig. 3 Regular 3D grid overlayed onto a MOF for interpolation of an APD using the DeepAPD model. The interpolation is done in batches, with a single batch highlighted in red (where $n_{\text{batch},i}$ is the number of points (voxels) in a batch for a given dimension, i). Prior to interpolation of each batch, probe nodes are inserted simultaneously at the centre of each voxel in each batch and a resulting graph is constructed omitting probe–probe edges to reduce memory requirements. A probability value is obtained at each grid point, as shown in Fig. 2. The grid spacing in this example is 0.15 Å (the same grid spacing used for all APDs in this work). The overall APD is updated with the completion of each batch and is normalized to unity after the completion of all batches. The atom colouring is the same as in Fig. 1.

the aggregated loss was backpropagated to update the model parameters.

The trainable parameters of the models were initialized using Glorot initialization[35] with a uniform distribution. The Adam optimizer was used with an initial learning rate of 5×10^{-5}, which was reduced by 3% when the loss on the development set did not improve between two consecutive steps. Instead of early stopping, the model with the best performance on the development set over the course of training was taken as the best model. All loss curves were visually inspected to determine convergence.

During inference, a regularly spaced 3D grid of any arbitrary spatial resolution is superimposed on the unit cell, and prediction is done in batches of probes, with batch size limited only by available GPU/system RAM. The batching procedure is visually depicted in Fig. 3, where probes are inserted at the centre of each voxel. Notably, the memory bottleneck of this procedure involves building edges between each probe and the structure graph within the specified cutoff radius of the model. Since messages are not passed between nodes, batch size has no effect on the predictions. Upon the completion of a batch, the overall probability distribution (initialized as zeros) is updated with the corresponding voxel data of the batch. After all batches are complete, the final APD is then normalized such that the entire distribution sums to unity, and the APD is written to disk in the VASP CHGCAR format.

Computational details

Grand canonical Monte Carlo (GCMC) simulations

All GCMC simulations were performed using version 1.4.0 of fastmc (code available in the SI material). Single-component GCMC simulations were performed for Xe at 1 bar, 298 K and for CH_4 at 298 K and pressures of 1 bar and 65 bar. Supercells of each MOF were constructed such that the minimum simulation cell vector length was 12.5 Å. Steric and dispersion interactions were modeled using the Lennard–Jones (LJ) potential. LJ parameters of Xe and single-site CH_4 were taken from the work of Boato, *et al.*[36] and Martin *et al.*,[37] respectively, while UFF parameters[38] were used for framework atoms. Lorenz-Berthelot mixing rules were used to determine LJ parameters for pairs of atoms of different types. 20 million equilibration steps were used for each MOF. The simulations were run in production until the APDs were deemed converged (see APD database section).

APDs were computed by binning the positions of the LJ site of each adsorbate at each GCMC production step on a real-space 3-dimensional grid. In the GCMC code used, one APD is written per site for each adsorbate molecule. Since this work focuses on single-site adsorbates only, one APD is generated per adsorbate molecule/state point. Furthermore, an equitable binning procedure was used when tracking the positions of adsorbate atoms to accelerate convergence.[21] The resolution of each APD was kept fixed, with each APD having maximum voxel dimensions of 0.15 Å × 0.15 Å × 0.15 Å. At the end of each GCMC simulation, the APD corresponds to the simulation cell, which is necessarily a supercell. Therefore, the simulation cell APD was folded down (spatially averaged) to obtain the APD corresponding to the single unit cell.

Binding site identification

Our in-house tool, the guest atom localization algorithm (GALA) was used to identify the binding sites from the APDs, and is described in detail elsewhere.[21] Briefly, the method first uses a 3D Gaussian filter to reduce statistical noise from the APDs. The implementation makes use of a Gaussian filter in SciPy[39] which convolutes three 1D Gaussian kernels (eqn (2)), where σ is the standard deviation of the Gaussian kernel ($\sigma = 0.10$ was used for all APDs in this work). If more than one maximum was within a 0.45 Å of other maxima, only the maximum with the highest occupancy was retained. Additionally, only maxima greater than 10% of the global maximum in the APD were considered.

$$G(x;\sigma) = -\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (2)$$

Binding site comparison algorithm

A binding site analysis tool was developed to identify and group symmetrically distinct sites between ML and simulated APDs to enable a measure of the quality of binding site predictions from the APDs. The code first parses the ML and simulated GALA output files, extracts the occupancies and energies of each site, and filters out any sites which have an occupancy less than 10% of the global maximum (*i.e.*, relative occupancy of 10%). The relative occupancies are the probability value (height) of each maximum relative to the global maximum in the APD (Fig. S1). Then, the space group of the MOF is identified using the SpacegroupAnalyzer class implemented in the Pymatgen Python package with the tolerance of 0.5, which uses the spglib[40] C library to perform various symmetry-identifying operations. A higher than default tolerance was used to account for slight atomic displacements in experimental crystal structures. Next, pairs of sites from the simulated APD are tested for equivalence under the space group of the crystal using the SpacegroupOperations class from Pymatgen (using a loose tolerance of 0.1 to account for lack of precision in the GCMC simulated APDs), and equivalent sites are grouped together. The same is then repeated for all pairs of sites from the ML APD. In the case of the simulated APDs, even though sites may be symmetrically equivalent, they can have slightly different occupancies, again as a result of lack of precision at the convergence threshold that was set for the APDs (Tanimoto $\geq$ 0.75). Even for the ML APDs, where symmetry equivalent sites should yield identical occupancies, slight distortions in the structure can result in slightly different relative occupancies predicted by the model.

Once the equivalent sets of sites are determined for the ML APD and the simulated APD, the algorithm then identifies the distance between all pairs of sites between ML and simulation. Any pairs which have a distance less than a cutoff (chosen to be 1 Å in this work) are considered to be a match between ML

and simulation and are marked as "matched" sites. If the site found is in a set of equivalent sites, then the ML/simulation pair with the shortest distance is retained and counted as a "match". This site is only counted once to avoid multiple counting of symmetrically identical sites. Any sites, either from ML or simulation, which do not have a match are deemed "missing", and can be missing either from ML (*i.e.*, a site present in only the simulation result) or missing from simulation (*i.e.*, a site present in only the ML result). For matched sites between ML and simulation, their occupancies are compared using mean absolute error (MAE). If there are any missing sites for a particular MOF comparison, the number of unique missing sites is recorded and so is the maximum occupancy within each unmatched equivalent set. This gives insight into how different occupancies are between maxima from both simulated and ML APDs, as well as how many binding sites don't match and whether missing sites tend to be important (high occupancy). A graphical description and example of this methodology is provided in the SI material.

Energy grid adsorbate probability distributions

Energy grids (*i.e.*, guest–host interaction energies on a 3D grid) were computed by making trivial modifications to the energy grid histogram descriptor implemented in version 0.0.9 of the mofdscribe[41] Python library. RASPA2 (ref. 42) was used to compute the guest–host interaction energies required to obtain the energy grids at a grid spacing of 0.15 Å, with modifications to the `MakeASCIGrid` routine in the `grids.c` module. The modifications were made to ensure the grid matched what is generated in our in-house GCMC code, since by default the grid size specified in RASPA2 does not get used as a maximum spacing, resulting in a different number of voxels from the APDs, and preventing direct comparison. The modified source code is provided in the SI material. The same force fields used in the GCMC simulations were applied. In the case of single-site adsorbates at infinite dilution (*i.e.*, $\mu \rightarrow 0$), the guest–host potential energy surface corresponds to the energies of all possible microstates (arising just from translations of the adsorbate). So, the Boltzmann probability distribution may be computed as shown in eqn (3) (where $\beta = (kT)^{-1}$), assuming the conditions specified above. Therefore, any differences between the energy grid APDs and the GCMC APDs is a result of guest–guest interactions/correlations.

$$P_i = \frac{e^{-\beta E_i}}{\sum_{i}^{N_{\text{voxels}}} e^{-\beta E_i}} \tag{3}$$

Data analysis

Uniform manifold approximation and projection (UMAP)[43] as implemented in the Rapids CuML package[44] was used to reduce high-dimensional data to two or three dimensions for visualization purposes. The hyperparameters used can be found in the SI material. The features used in the UMAP dimensionality reduction were not standardized. Geometric and chemical descriptors used in this work were taken from the ARC-MOF[27] and MOSAEC-DB[29] databases.

Results & discussion

Adsorbate probability distribution (APD) database

As a first step of analyzing the diversity of the APDs, the chemical and geometric diversity were assessed. The revised autocorrelation (RAC) descriptors were used to describe MOF chemistry. We hypothesize the ligand chemistry and geometry of the MOFs to be the most influential properties on the adsorption of the gases studied in this work. For this reason, we focus on inspecting the design space covered by the training and test sets considering the geometric descriptors and ligand RAC descriptors, which are publicly available and described in more detail elsewhere.[27,29] Considering the ligand RAC descriptors, the test set of 338 MOFs from MOSAEC-DB span a similar design space as the training set used to train the model, suggesting the test set is not out of distribution from a perspective of ligand chemistry (Fig. S9). This is perhaps unsurprising considering a similar result was found in a previous study when comparing the overall ARC-MOF and MOSAEC-DB databases.[29]

It is clear from the distributions of geometric properties (Fig. S10) that the sampled training sets reflect somewhat closely the distribution of geometric properties of ARC-MOF. Clearly, the procedure of generating the training data biased the selection to MOFs with smaller pores and volume fractions and larger gravimetric densities. This is because APDs of MOFs with smaller pores are likely to converge faster because the binding is likely to be more localized, and MOFs with fewer required production steps were added to the training set with higher priority. To ensure the model performs well on MOFs with larger volume fractions and pore sizes, 338 MOFs from MOSAEC-DB were sampled using farthest point sampling of their geometric and ligand chemistry properties. For this reason, the MOSAEC-DB test set follows a distribution much closer to the overall ARC-MOF database.

Characterizing the diversity of the probability distributions themselves poses a challenge because many probability distribution similarity measures rely on uniform grid shapes. This is not the case for our data since we acquired the APDs at constant spatial resolution and the MOFs vary greatly in cell volume. Therefore, the most straightforward way of comparing the probability distributions is using a global descriptor, such as the Shannon entropy of the distributions. Entropy reflects the randomness or uncertainty associated with a distribution. Therefore, high entropy APDs should be delocalized (*i.e.*, have many maxima) over space, while low entropy APDs should be more localized distributions. To compare entropies between MOFs with different grid shapes, we normalize the entropy of each APD to maximal entropy (*i.e.*, the uniform distribution). This entropy of the flattened APDs relative to a uniform distribution, $H_{\text{rel}}(X) : p(x) \rightarrow [0,1], x \in \mathcal{X}$, is computed according to eqn (4) (see the SI material for more information).

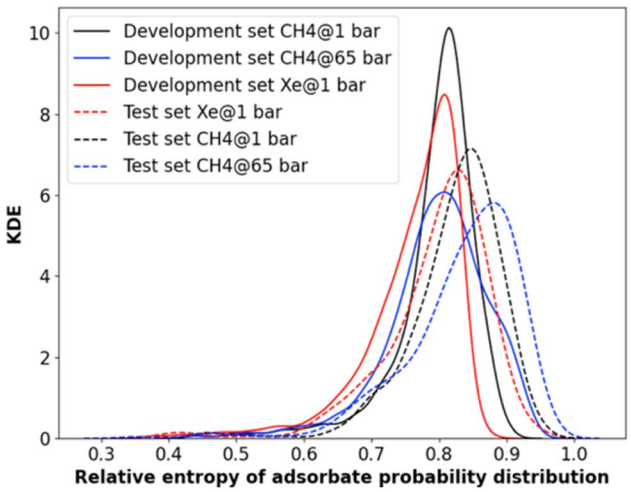

Fig. 4 The distribution of relative entropies of adsorbate probability distributions in the test set and development set for adsorbates under different conditions.

$$H_{\mathrm{rel}}(X) = \frac{H(X)}{H_{\max}(X)} = \frac{\sum_{x\in\mathcal{X}} p(x)\log p(x)}{\log(|\mathcal{X}|)} \quad (4)$$

The relative entropy of the APDs was computed and the kernel density estimation (KDE) of the distribution of entropies was plotted for the test set and training set for the adsorbates under different conditions (Fig. 4). Notably, the entropy of the APDs in the training set are generally smaller than those in the test set, suggesting on average the APDs are more localized for the training set. This is expected, considering the aforementioned distribution of pore sizes in these sets (Fig. S10) and the movement of adsorbates within smaller pores is generally more restricted/localized. The APDs of CH_4 at 65 bar tend to have the widest range of entropies, while the APDs of CH_4 at 1 bar tend to have the smallest range of entropies. In the case of the test set and training set, the APDs of Xe at 1 bar tend to have the smallest average APD, which can be rationalized by the larger size of Xe and stronger interaction energy relative to CH_4, so Xe is expected to be more locally confined within MOF pores. It is perhaps surprising that for some MOFs, the entropy of the CH_4 APDs increases with increasing pressure, suggesting that the APDs become more delocalized. However, the correct interpretation should not be that the distributions are highly diffuse, but rather the distributions have many more highly localized peaks of similar magnitude and therefore results in a more uniform distribution (this is supported by the significant increase in binding sites from low to high pressure, as discussed later).

It is important to note that it is impossible for an APD to have a relative entropy of unity, since this would require non-zero probability in voxels occupied by atoms. To account for this system/geometry-dependent "maximal entropy", one can incorporate the fraction of the unit cell that is free space into the relative entropy metric. Therefore, the relative entropies of the APDs from the training and test sets were plotted as a function of the product of the largest cavity diameter and accessible volume fraction to determine the dependence of the relative entropy on the geometry of the MOF (Fig. 5). It is interesting to note that there is a much larger spread in the data for CH_4 at 1 bar compared to CH_4 at 65 bar. This is because at moderate to large pore sizes at low pressure, if there is a strong enough interaction energy, the binding of the adsorbate can still be localized, with only a few significant maxima. In other words, at low pressure the relative entropy of the distribution has a weak correlation with the geometry of the MOF past a certain pore size. Conversely, at high pressure, the entropy of the APD becomes much more dependent on the geometry of the MOF as is shown by the relatively lower spread in Fig. 5b. At low pressure for both adsorbates (Fig. 5a and c), there are outliers in the overall trend, particularly in the test set where some APDs have high entropy even though the pore size is relatively small (<10 Å), likely due to weak interaction with the framework. A visualization of the change in entropy of the APD going from low pressure to high pressure for two MOFs is given in Fig. S14. In some cases (Fig. S14B), the entropy of the distribution can actually decrease when going to high pressure, but more commonly, the entropy of the APD will increase when going to high pressures, for reasons discussed above (Fig. S14A). As is shown in Fig. S14, the entropies correlate with both the number and spatial extent of the APD. In the case where the relative entropy decreases going from low to high pressure, it is clear

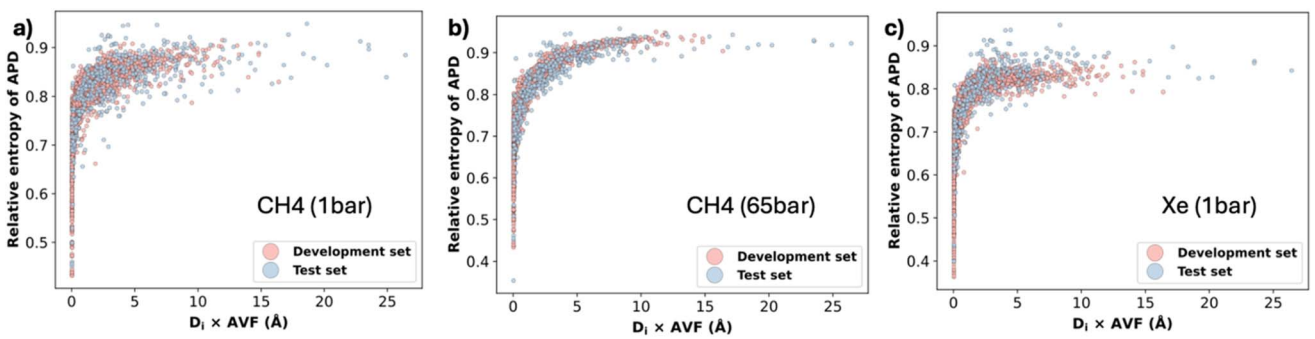

Fig. 5 The relative entropy of APDs of CH_4 at (a) 1 bar and (b) 65 bar, as well as (c) Xe at 1 bar from the training set and test set plotted as a function of the product of largest cavity diameter (D_i) and accessible volume fraction (AVF) of the corresponding MOF structure.

Table 1 ML models trained to evaluate the effect of multitask learning, transfer learning, and dataset size on predictive performance. For each model, training set sizes of $N = 100$, 1000, 5000, and 19 335 were investigated

Model name	Training data (adsorbate and state point)	Development data
Xe_1 bar model	Xe (1 bar)	Xe (1 bar)
CH4_65 bar model	CH_4 (65 bar)	CH_4 (65 bar)
CH4_1 bar model	CH_4 (1 bar)	CH_4 (1 bar)
Multitask model	CH_4 (1 bar), CH_4 (65 bar), Xe (1 bar)	CH_4 (1 bar) and Xe (1 bar)[a]

[a] Instead of evaluating an aggregated loss on all adsorbates for the multitask model, separate models were validated during training on a particular adsorbate.

that this is a result of the number of binding sites remaining relatively the same, but with a greater degree of localization.

DeepAPD model training

To explore possible improvements and limitations of the ML models, the effect of various aspects of the training procedure on overall model performance was evaluated. First, the effect of training a model on multiple APDs simultaneously, rather than having a single ML model for each adsorbate/state point (*i.e.*, multitask learning) was evaluated. For this purpose, four different models were trained, which are summarized in Table 1.

It is important to note that some APDs are more expensive to simulate than others depending on the adsorbate, temperature, and pressure. The primary reason for this is related to inefficient sampling of phase space using conventional insertion and deletion moves in simulations of polar adsorbates,[45] or simulations at high pressures[46] or low temperatures.[47] In these limiting cases, equilibration takes longer, and a larger number of production steps are required for convergence of the APDs. Therefore, one may be able to obtain an extensive database of APDs for a simple adsorbate like methane at ambient temperature and pressure, but only tens or hundreds for an adsorbate like water. Along these lines, an important question arises – is there any significant benefit to pre-training a model on one or multiple adsorbates, and using this as a base model for prediction of an entirely different adsorbate for which there are limited data? Transfer learning (TL) was applied to gain more insight into this question.

The effect of transfer learning on both the training efficiency (Fig. 6a) and overall model performance (Fig. 6b) when training a model to predict exclusively CH_4 APDs at 65 bar was investigated. As expected, fine-tuning previously trained models results in an improvement in overall model performance compared to random initialization, particularly when the training set size is small. The largest improvement in overall model performance was observed for $N = 100$, where random initialization and initialization with the Xe_1 bar model yield a loss of −0.78 and −0.81 on the development set, respectively (~4% improvement). It is interesting that the Xe_1 bar model, rather than the CH4_1 bar model, is more effective in transfer learning for the APDs of CH_4 at 65 bar. We speculate that this is because guest–guest interactions play a much more important role in the cases of adsorption of CH_4 at 65 bar and Xe at 1 bar, while guest–host interactions dominate for CH_4 adsorption at 1 bar. For this reason, guest–guest interactions are captured better by the Xe_1 bar model. While guest–host interactions are also important, transferring the guest–host potential energy surface of different adsorbates requires only a modification in the Lennard–Jones potential. This improvement in performance becomes less significant as N increases (*e.g.*, <1%

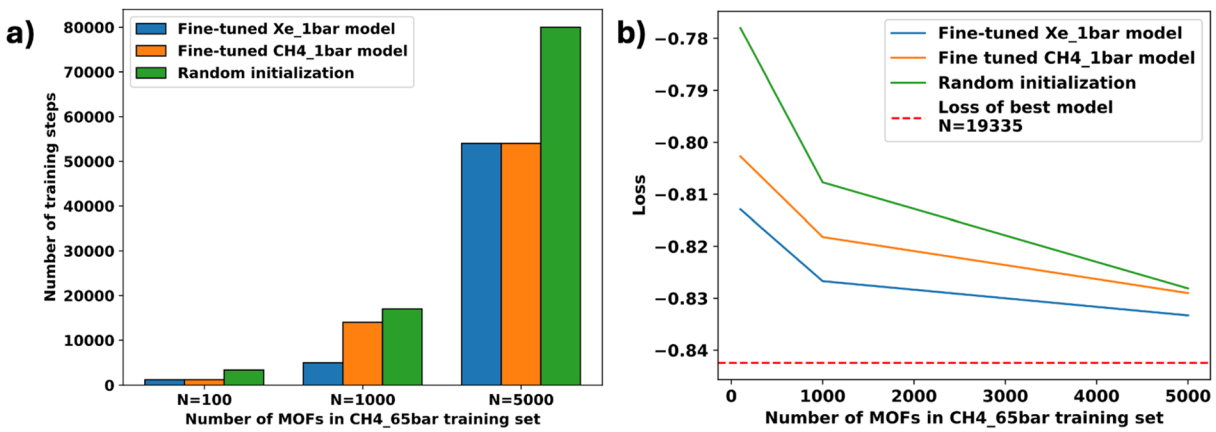

Fig. 6 Effect of transfer learning on (a) the number of training steps required to minimize loss on the development set for varying number of training set sizes ($N = 100$, 1000, 5000), and (b) overall model performance. The results shown are for initializing the CH4_65 bar model with either the pre-trained CH4_1 bar model, Xe_1 bar model, or random initialization. Each pre-trained model had a training set size of 19 335 MOFs. One training step corresponds to a single batch of 6000 targets corresponding to one MOF.

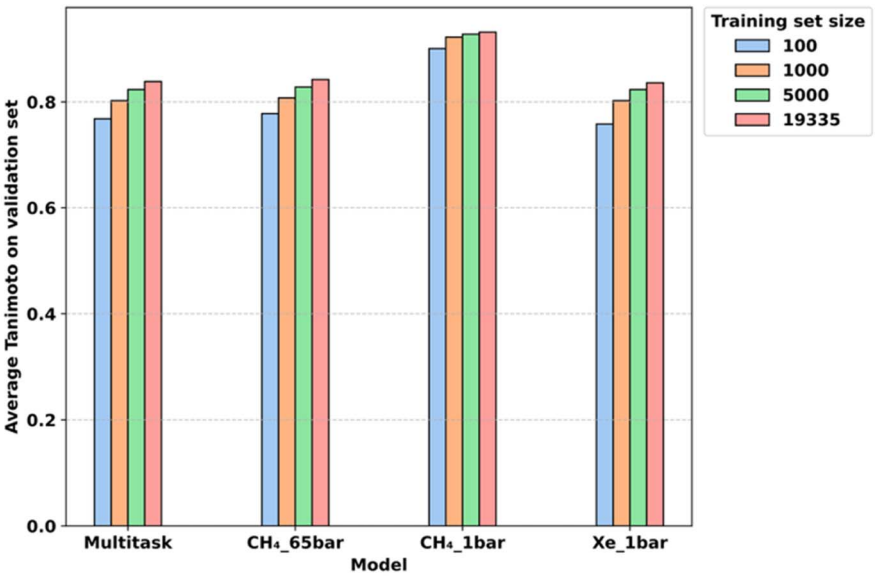

Fig. 7 Effect of dataset size and multitask learning on overall model performance. The multitask model reported here was validated on Xe adsorption data, while the single-site models were validated on their respective dataset.

improvement in model performance for $N = 5000$). Furthermore, it is notable that the TL model fine-tuned from the model trained on the same adsorbate at a different state point (CH4_1 bar model) exhibited a poorer performance boost compared to pre-training on Xe (~2.5% reduction in development loss). While 4% improvement in the development loss seems like a small amount, it is notable that the performance increase in the CH4_65 bar model in going from a dataset size of $N = 100$ to $N = 19\,335$ with random initialization results in a performance increase of less than double (~7.5%). Therefore, even though TL results in only modest improvement in these cases, it is expected that for adsorbates/state points where data are a limiting factor in model performance, the payoff may be significant. In addition to the performance boost, pre-training the model on a large dataset of a different adsorbate and/or state point results in fewer required training steps, requiring as little as 30% of the steps that would be required with random initialization (Fig. 6a).

The general effect of training set size on model performance is summarized in Fig. 7. The multitask model corresponds to the DeepAPD model trained simultaneously on all three APD datasets, as described in the Methodology section. Note that the MOFs in each training set are consistent across different models. Interestingly, and in line with what was seen in the TL analysis, training set sizes of only 100 APDs are sufficient for obtaining >90% maximum performance ($N = 19\,335$) on the development set for the multitask model, which was validated on Xe APDs. The same is observed for the Xe_1 bar model, which is validated on the same dataset. In the case of the CH4_1 bar model, training on the $N = 100$ dataset results in a model which obtains 97% maximum performance, suggesting this is the easiest APD to predict, and will be discussed in more detail later. In terms of the benefit of multitask learning for these APDs, it appears there is little improvement in overall model performance for $N = 19\,335$ when comparing the multitask model (validated on Xe APDs) to the Xe_1 bar model (both achieve a loss of ~−0.84 on the development set). However, since there are no performance losses associated with multitask learning, this model was selected as the final DeepAPD model to avoid having entirely separate GNN models for each adsorbate.

To initially assess the performance of the DeepAPD model, it was evaluated on both the test and development sets, and a Tanimoto coefficient between the resulting APDs was evaluated against the simulated APDs. The average Tanimoto coefficient for each dataset is tabulated in Table 2. As is shown in Fig. 7, the model performs best on the CH_4(1 bar) dataset. The performance on the test set is significantly higher on the development set in the case of the CH_4(1 bar) dataset (~11% higher similarity between ML and simulated APDs), and is modestly higher in the other two datasets (~5% higher similarity between ML and simulated APDs). The Tanimoto is also far less consistent when performing evaluations on the test set, as is evidenced by the significantly higher standard deviations.

To better understand how the Tanimoto similarity corresponds visually to differences in the APDs, the APDs from both simulation and ML of three different MOFs for CH_4 at 65 bar were visualized in Fig. 8. In the case of a YUBRAP_full, which has a high similarity between simulation and ML (Tanimoto = 0.96), the differences between the APDs are very small, and the

Table 2 Average Tanimoto similarity and standard deviation between pairs of ML and simulated APDs for each adsorbate/condition in the test and development (dev) sets

	Tanimoto	
Adsorbate (condition)	dev	Test
CH_4 (1 bar)	0.93 ± 0.03	0.84 ± 0.13
CH_4 (65 bar)	0.83 ± 0.08	0.79 ± 0.12
Xe (1 bar)	0.83 ± 0.06	0.78 ± 0.13

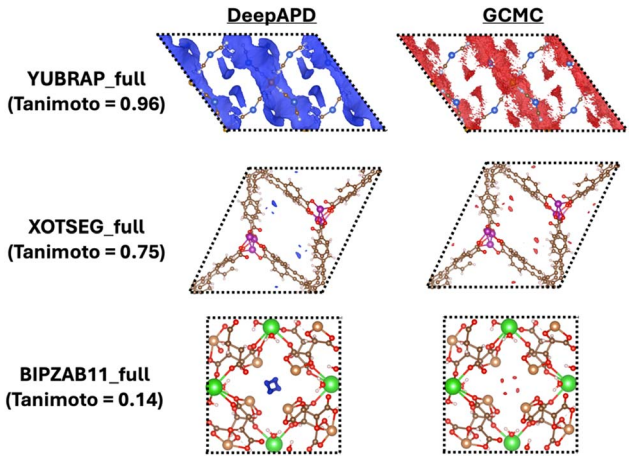

Fig. 8 Comparison of APDs from the MOSAEC-DB test set of CH_4 at 65 bar obtained from DeepAPD (blue) *versus* GCMC (red). An isosurface value of 50% of the global maximum was chosen to emphasize differences in the most probable regions of the APDs.

simulated APD is notably more noisy. In the case of XOTSEG_full, the similarity is slightly worse, and closer to the average of all test sets (Tanimoto = 0.75), and the maxima are more localized. Between ML and simulation, the maxima mostly match in terms of position, but the spatial extent of each maximum at the specified isosurface value – a result which occurs due to differences in relative occupancies. In the last case of BIPZAB11_full, the similarity is very low (Tanimoto = 0.14). In this case, it is clear that instead of having four distinct maxima as is seen in simulation, there is high probability in the areas between them predicted by ML. This visual comparison emphasizes the fact that the Tanimoto is weighted toward maxima.

Applying DeepAPD to identify binding sites of MOFs

Considering that the Tanimoto is a global similarity measure between two probability distributions, a local description of the model performance would give more insights into the model limitations and guide further improvement, and is more aligned with how the APDs would be used in practice. With the simulated APDs as the ground truth, a straightforward test of the fidelity of the ML-predicted APDs to the simulated APDs is to compare the maxima (binding sites) of the distributions. We focus on two primary characteristics of the maxima, namely the positions and relative occupancies. The relative occupancies are the probability value (height) of each maximum relative to the global maximum in the APD (Fig. S1). Since the development set and test set have different geometric distributions, it will be of value to discuss the performance of DeepAPD on both sets. In addition to this point, it is worth noting that not all datapoints in the development set were used to train the ML model.

Binding sites were extracted from the ML-predicted APDs and simulated APDs using the procedure outlined in the Methodology section. As an initial simple analysis, the most occupied binding sites (global maxima of the APDs) were compared between simulation and ML by evaluating the interaction energy between the adsorbate and framework at the binding site. The parity of the binding energy (*i.e.*, guest–host binding energy) of the most occupied binding sites from the ML APDs *vs.* simulated APDs for the development set is shown in Fig. 9. There is a notable deviation in accuracy when determining the most occupied site of CH_4 at 1 bar ($R^2 = 0.99$, MAPE = 0.91%) in comparison to 65 bar ($R^2 = 0.76$, MAPE = 4.23%). A similar result is observed for the test set, though the performance for predicting CH_4 at 65 bar seems to be slightly better in terms of linear correlation but worse in terms of MAPE (R^2 = 0.90, 4.58%) (Fig. 10). However, the overall trend in predictive performance of APDs remains, and is in agreement with the results presented in Table 2. This difference in accuracy reflects the difficulty of predicting APDs of more strongly interacting adsorbates (Xe) and APDs of adsorbates at high pressures. This result is in line with two previous observations – the fact that training set size had a larger impact on model performance for Xe and CH_4 at 65 bar, and the fact that simulating APDs of more strongly interacting adsorbates and adsorbates at high pressure requires more production steps.

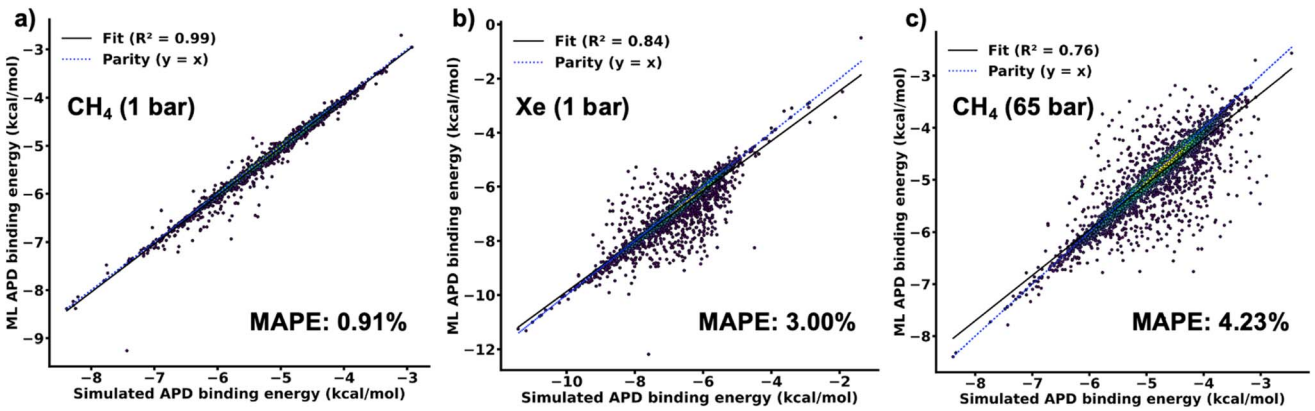

Fig. 9 Parity plots of binding energy for the most occupied binding sites from machine learned *vs.* simulated APDs of (a) CH_4 at 1 bar, (b) Xe at 1 bar, and (c) CH_4 at 65 bar for the development set. MAPE represents mean absolute percentage error. The ML model used for all comparisons was the multitask DeepAPD model. The colour of points represents the KDE density of points.

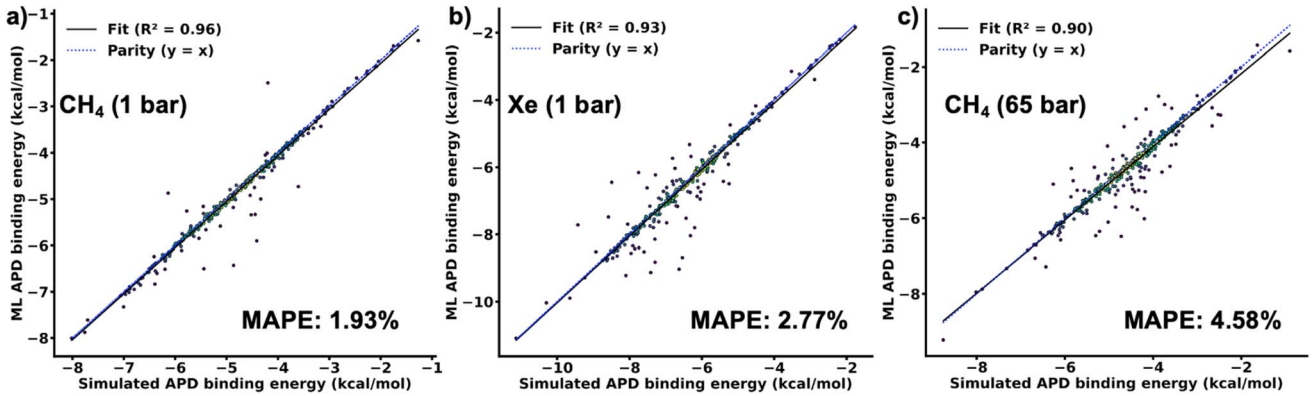

Fig. 10 Parity plots of binding energy for the most occupied binding sites from machine learned *vs.* simulated APDs of (a) CH_4 at 1 bar, (b) Xe at 1 bar, and (c) CH_4 at 65 bar for the test set. MAPE represents mean absolute percentage error. The ML model used for all comparisons was the multitask DeepAPD model. The colour of points represents the KDE density of points.

A possible reason for this deviation in predictive performance of DeepAPD for a simple adsorbate like CH_4 from low to high pressure is likely the increasing complexity of the free energy surface. This is related to the challenges that lead to difficulties in convergence of GCMC simulations.[48] At higher pressures, entropic effects and guest–guest interactions become more important, and capturing these effects poses more challenges than just learning the interaction potential between the adsorbate and the MOF. To support this hypothesis, we compare the binding energy of the most occupied site and the minimum energy binding site (*i.e.*, the site with the strongest guest–host interaction energy) from simulated APDs (Fig. 11). The comparison shows that while the DeepAPD model simply needs to learn a Lennard–Jones potential between the guest and framework atoms (which relies on a simple combination of transferable parameters) to correctly predict the most probable site for CH_4 at 1 bar ($R^2 = 0.99$), the other APDs exhibit some other contributions to the overall free energy of the system (*e.g.*, guest–guest interactions) that results in the most probable binding site not necessarily being the one with the strongest interaction with the framework ($R^2 = 0.74$ and $R^2 = 0.64$ for Xe at 1 bar and CH_4 at 65 bar, respectively). Of course, where the guest–host interaction is the dominant factor in determining the most occupied binding site, a simple energy histogram could be generated, which is trivial to compute in comparison to the APDs. It is important to note when interpreting these results and those in Fig. 9 and 10 that while the most probable site may be one with a low guest–host interaction energy, the minimum energy binding site may still have a high relative occupancy. In the analysis, only the site with 100% occupancy from each MOF was chosen for comparison.

It is important to note that only binding sites with negative binding energies were considered in the analysis presented in Fig. 9 and 10. While all binding sites from simulated APDs have a negative (attractive) guest–host interaction energy, a few sites from the ML APDs have positive (repulsive) binding energies and reveal some outliers and deficiencies in the DeepAPD model. Notably, binding sites from the ML APDs with positive binding energies are rare, as shown in Table S1 (<0.1% of binding sites identified for any adsorbate in the development set are repulsive). However, when these occur, they can be incredibly repulsive (10^3–10^{12} kcal mol^{-1}). Surprisingly, in most cases MOFs which possess outlier ML binding sites do not exhibit any unusual or exotic chemistry. Repulsive binding sites which are predicted with high probability (100% occupancy) are shown in Fig. S11. In the case of CH_4 and Xe at 1 bar, a common chemical motif is responsible for the high energy most probable binding site, shown in Fig. S11(a–c). For CH_4 at 65 bar, binding

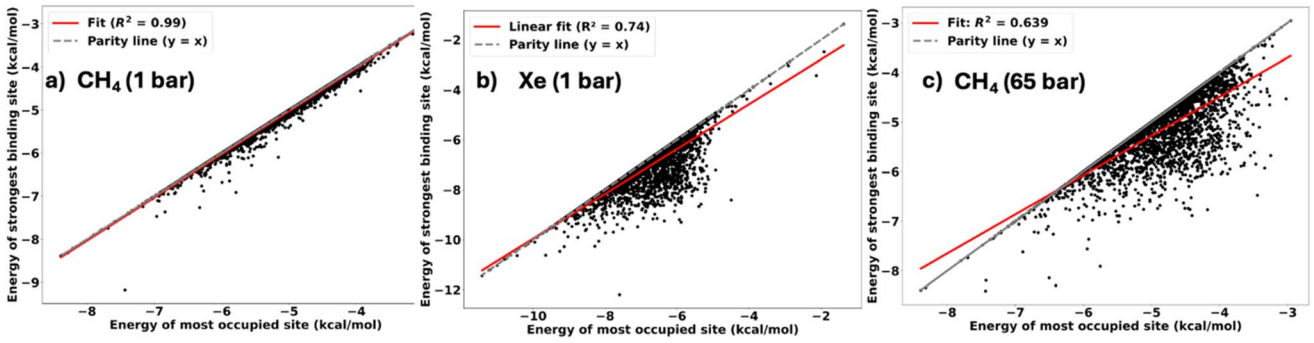

Fig. 11 Parity of the minimum energy binding site *vs.* most occupied binding site for simulated APDs of (a) CH_4 at 1 bar, (b) Xe at 1 bar, and (c) CH_4 at 65 bar. The data shown are from the development set.

Table 3 Number of binding sites that were matched between those extracted from GCMC simulated APDs and estimated APDs for CH_4 at 1 bar, CH_4 at 65 bar, and Xe at 1 bar for the development set. The estimated APDs were obtained *via* either DeepAPD (A) or guest–host interaction energy grids (B)

	(A) Binding sites identified			(B) Binding sites identified		
Adsorbate (conditions)	From ML APD only	From both (matched)	From GCMC APD only	From energy grid APD only	From both (matched)	From GCMC APD only
CH_4 (1 bar)	1247 (4%)	27 152 (85%)	3606 (11%)	507 (2%)	23 718 (76%)	6989 (22%)
CH_4 (65 bar)	9782 (17%)	40 771 (71%)	6660 (12%)	2663 (5%)	21 550 (44%)	25 243 (51%)
Xe (1 bar)	4936 (14%)	26 482 (76%)	3404 (10%)	1416 (5%)	16 389 (53%)	12 773 (42%)

inside N-containing cage-like motifs shown in Fig. S2(e and f) are predicted to be highly responsible for such repulsive binding sites. In some instances, sites that overlap with framework atoms can be predicted to have high probability, as shown in Fig. S2(d). While these cases are exceedingly rare in the development and test sets, physical constraints will be implemented in the DeepAPD model in future. For now, when the model is applied to predict APDs for the purposes of binding site extraction, any binding sites identified with large positive binding energies can be discarded in our binding site extraction tool.

While comparison of binding energies is a reasonable initial analysis to determine how well the ML models capture the global maxima of the APDs, there can be many local maxima which are above 10% relative occupancy. For a more specific comparison of the binding sites from the ML *vs.* simulated APDs, the position of the binding sites extracted from each and their corresponding occupancies were directly compared. While this seems like a straightforward comparison, especially for single-site adsorbates, it requires careful consideration of the symmetry of the crystal to ensure that binding sites which are symmetrically equivalent are not multiply counted, which would skew the analysis. The algorithm used to compare binding sites between ML and simulated APDs is described in the Methodology section, and a full description is given in the SI material.

The binding site comparison algorithm yields an abundance of useful information for assessing the fidelity between the ML and simulated APDs. First, it was found that a large proportion of binding sites identified from the ML and simulated APDs match (85%, 76%, and 71% for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively) (Table 3A). This once again follows the same predictive accuracy trend that was observed for identifying the most occupied site (Fig. 9 and 10). Interestingly, the ML model tends to predict sites which are not found from simulation rather than missing sites that are identified from simulation in the cases of CH_4 at 65 bar and Xe at 1 bar. For example, in the case of CH_4 at 65 bar, 17% of all sites are reported by the ML model only. This suggests that the ML model is either overestimating the occupancy of some local maxima (only maxima of relative occupancies of $\geq$10% are considered in this analysis), or the model is introducing false maxima. Still, a reasonable number of binding sites are missed by DeepAPD (between 10–12% depending on the APD), which again could be a result of an underestimation of relative occupancies. For example, if DeepAPD predicts a peak to have an occupancy of 5%, it would be excluded from binding site identification in GALA, so would be counted as a "missed" site. The reason for excluding sites <10% relative occupancy is to focus the analysis on sites which we would consider significant, since the number of total local maxima is often immense.

Before continuing with the analysis of the binding sites obtained from the ML APDs, it is worth discussing when identifying binding sites by a Boltzmann-weighting of the guest–host potential energy surface (PES) may be a reasonable approximation for obtaining binding sites. For example, in the literature, many studies approximate binding sites of MOFs by simply optimizing the position of the adsorbate in an empty MOF.[49–51] In particular, for single-site adsorbates, and at lower pressures, it is possible that such an approximation may be sufficient to obtain a good description of the binding sites in the material, and may even outperform the ML APDs. It is also reasonable to question whether the DeepAPD model has effectively learned the free energy surface instead of simply learning guest–host interaction potentials. To address this, we compare the quality of the binding sites obtained from ML APDs to those obtained from the guest–host PES by using the GCMC binding sites as a reference (Table 3B). Perhaps unsurprisingly, the predictive trend of the energy grid APDs follows that of the ML APDs, with binding sites of CH_4 at 1 bar having the best agreement between the energy grid APDs and simulated APDs (76%) followed by Xe at 1 bar (53%), and finally CH_4 at 65 bar (44%). This is related to the previous discussion on increasing complexity of the free energy surface due to more strongly interacting adsorbates (Xe), and greater contribution of guest–guest interactions to the free energy surface at high pressures (Fig. 11). Most notably, using the ML APDs results in significantly lower overall error in binding site identification in all cases in comparison to using the energy grid APDs. Frequently, sites are found from simulated APDs that are not reported from the energy grid APDs, since guest–guest interactions cause new or shifted maxima in the probability distributions.

The DeepAPD model was further validated on a separate test set composed of structures from the MOSAEC-DB database. In line with initial tests on the correlation of the binding energy of the most occupied site (Fig. 11), the predictive performance of the DeepAPD model is roughly equivalent between the different guests/state points, with ~80% of binding sites matching

Table 4 Number of binding sites that were matched between those extracted from GCMC simulated APDs and ML APDs for CH_4 at 1 bar, CH_4 at 65 bar, and Xe at 1 bar for the test set

	Binding sites identified		
Adsorbate (conditions)	From ML APD only	From both (matched)	From GCMC APD only
CH_4 (1 bar)	39 (2%)	1713 (78%)	452 (20%)
CH_4 (65 bar)	175 (5%)	2787 (82%)	459 (13%)
Xe (1 bar)	102 (5%)	1766 (79%)	362 (16%)

between simulation and ML (Table 4). This improvement in performance in the test set in comparison to the development set is most likely a result of the wider distribution of pore sizes of the test set (Fig. S10). Since the adsorbates are less confined and generally more weakly interacting in the case of large pore MOFs, guest–guest interactions are less dominant, and generally lead to adsorption sites which are more reflective of the guest–host potential energy surface.

Even though many sites are not matched between simulation and ML APDs, it is still unclear whether these sites are important (*i.e.*, of significantly high occupancies). To this end, some statistics were collected about sites which were missing in the development set, and are summarized in Table 5. Notably, even though a large proportion of sites are unmatched for CH_4 at 65 bar and Xe at 1 bar (Table 3A, 29% and 24% of sites, respectively), most of these sites are under 30% occupancy (Table 5, 85% and 81% of sites, respectively). This emphasizes that most sites with an occupancy of >30% are matched between simulated and ML APDs (96% of sites for CH_4 at 65 bar, 95% for Xe at 1 bar and 97% of sites for CH_4 at 1 bar). In terms of correctly predicting the global maximum (*i.e.*, the most occupied binding site from simulation), this was successful for 90%, 76%, and 73% of cases for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively. While most of the time, the correct global maximum is predicted, it is predicted with very high occupancy (>90%) for 98%, 88%, and 87% of MOFs in the development set for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively. For all APDs, the global maximum is reported with an occupancy of >10% in 100% of cases. These results show that high occupancy maxima can be reliably identified with the DeepAPD model, with less confidence in lower occupancy sites. Furthermore, matched binding sites generally exhibit very small errors in position and occupancy, with an MAE of 0.20 Å, 0.24 Å, and 0.27 Å for binding site position and MAE of occupancy of 1.7%, 4.5%, and 4.1% for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively. Notably, the positions of adsorbates are identified on a discrete grid (0.15 Å), so an MAE <0.30 Å corresponds to an average error of <2 voxels.

Expectedly, the ML APDs result in better binding site predictions than energy grid APDs (Table 6). The correct global maximum was identified in only 89%, 68%, and 56% of cases in the development set for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively. In 9–10% of APDs in the case of Xe at 1 bar and CH_4 at 65 bar, the global maximum from simulation is missed entirely (*i.e.*, not found amongst all sites with >10% occupancy). Additionally, while matched sites exhibit very small errors in position, the occupancies exhibit larger errors than observed with the ML APDs, having errors of 3.0%, 9.5%, and 12% for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively.

The results on the MOSAEC-DB test set are slightly worse than those observed on the development set (Table 7), but again are more consistent across all APDs, likely due to a larger pore size distribution. Still, the majority of sites which are not matched between ML and simulation have occupancies below 30% (∼80% of sites for all APDs, which is similar to what was observed in the development set). The global maximum (most occupied site) is also identified correctly with high frequency (89%, 81%, and 78% of cases for CH_4 at 1 bar, Xe at 1 bar, and CH_4 at 65 bar, respectively). In comparison to the development set, identification of the most occupied site is significantly more reliable in the cases of Xe at 1 bar and CH_4 at 65 bar. This is again because of the reasons previously outlined in the analysis

Table 5 Summary of statistics comparing errors in occupancies and identification of maxima between those extracted from APDs obtained with the DeepAPD model and APDs which were obtained from GCMC simulation for various adsorbates/conditions. These statistics correspond to the development set

	Adsorbate/conditions		
	CH_4 (1 bar)	Xe (1 bar)	CH_4 (65 bar)
MAE of occupancies (%)	1.7	4.5	4.1
MAE of distances between matched sites (Å)	0.20	0.24	0.27
Fraction of unmatched sites <30% occupancy (%)	77	81	85
Fraction of MOFs where the most occupied site from simulation matched (%)	90	76	73
Fraction of MOFs where the most occupied site from simulation was found with >90% occupancy (%)	98	88	87
Fraction of MOFs where the most occupied site from simulation was found with any occupancy (%)	100	100	100

Table 6 Summary of statistics comparing errors in occupancies and identification of maxima between those extracted from APDs obtained from guest–host interaction energy grids and APDs which were obtained from GCMC simulation for various adsorbates/conditions. These statistics correspond to the development set

	Adsorbate/conditions		
	CH_4 (1 bar)	Xe (1 bar)	CH_4 (65 bar)
MAE of occupancies (%)	3.0	9.5	12
MAE of distances between matched sites (Å)	0.15	0.21	0.27
Fraction of unmatched sites <30% occupancy (%)	79	66	72
Fraction of MOFs where the most occupied site from simulation matched (%)	86	57	42
Fraction of MOFs where the most occupied site from simulation was found with >90% occupancy (%)	89	68	56
Fraction of MOFs where the most occupied site from simulation was found with any occupancy (%)	100	91	90

Table 7 Summary of statistics comparing errors in occupancies and identification of maxima between those extracted from APDs obtained with the DeepAPD model and APDs which were obtained from GCMC simulation for various adsorbates/conditions. These statistics correspond to the MOSAEC-DB test set

	Adsorbate/conditions		
	CH_4 (1 bar)	Xe (1 bar)	CH_4 (65 bar)
MAE of occupancies (%)	3.3	5.4	5.2
MAE of distances between matched sites (Å)	0.13	0.15	0.19
Fraction of unmatched sites <30% occupancy (%)	77	82	82
Fraction of MOFs where the most occupied site from simulation matched (%)	89	81	78
Fraction of MOFs where the most occupied site from simulation was found with >90% occupancy (%)	99	98	98
Fraction of MOFs where the most occupied site from simulation was found with any occupancy (%)	99	99	99

of Fig. 11. However, it is notable that the MAE of occupancies is slightly higher than what was observed for the development set, which likely contributes to the lower Tanimoto similarity between the ML and simulated APDs (Table 2).

From this analysis, it is evident that prediction of high occupancy maxima of these adsorbates/conditions is reliable using the DeepAPD model. In the case of low occupancy sites (*i.e.*, those with occupancy <30%), a more detailed analysis would be required. However, it is likely that the most occupied binding sites (*i.e.*, those with relative occupancies >30%) are of greater interest in the case of data mining or machine learning applications, where particular adsorbaphores may be targeted. The results also show that while extrapolation of the model to a new dataset shows slightly worse performance in some areas, it is still reliable for the most occupied sites.

While experimental validation is beyond the scope of the present study and direct observation of binding sites in MOFs is incredibly rare, binding sites from GCMC-derived APDs have been shown to be in generally good agreement with binding sites determined from SCXRD experiments.[21] To further support that DeepAPD agrees with experimental observations, the binding sites derived from the DeepAPD adsorbate probability distribution (trained on GCMC data) were compared against experimentally determined binding sites of CH_4 in $Sc_2(BDC)_3$ and Xe in SBMOF-1. The experimental binding sites were obtained from an *in situ* SCXRD experiments under conditions of 230 K and 9 bar (R-factor of 0.0401) and 100 K (unspecified pressure) (R-factor of 0.0703) for the $Sc_2(BDC)_3$ and SBMOF-1 binding sites, respectively.[6,52] The ML binding sites correspond to APD predictions of Xe at 1 bar and CH_4 at 1 bar for SBMOF-1 and $Sc_2(BDC)_3$, respectively. All binding sites were identified using the settings outlined in the binding site identification section in the Methodology. Despite slight differences between the simulated and experimental conditions, the binding sites are in excellent agreement with experiment, where the centre of mass (C atom of CH_4) is shown in Fig. 12a, and xenon is shown in Fig. 12b.

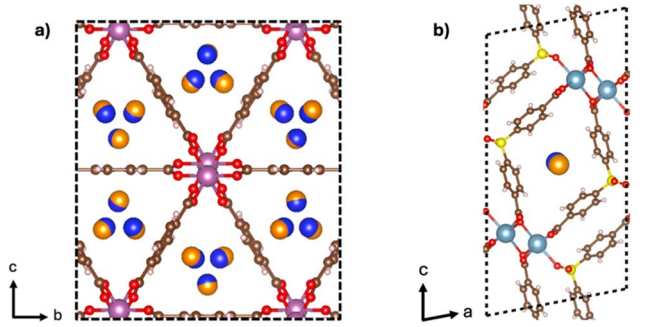

Fig. 12 Binding site comparison of (a) $Sc_2(BDC)_3$@CH_4 and (b) SBMOF-1@Xe between those obtained from experiment (orange) and those extracted from the DeepAPD adsorbate probability distributions (blue). Only the centre of mass (C atom of CH_4) is shown.

Binding site similarity

While it was shown that the DeepAPD model has utility for identifying binding sites of MOFs, it is possible there is some utility of the model to generate descriptors of binding environments, either to train ML models, or to use in unsupervised ML techniques to identify common features of sites which adsorb a molecule with some affinity. One way of doing this is using the embeddings of the PaiNN network in the model (ε_i as shown in Fig. 2), which are effectively a high-dimensional representation of the local chemical environment at some arbitrary probe point. These are the embeddings which are used by the DeepAPD model in adsorbate-specific feedforward neural networks to predict the adsorbate probabilities. In other studies, the embeddings of models have been used in a similar way, and is in fact the basis for autoencoders.[53] For example, Fung, *et al.* showed that the embeddings from their convolutional neural network model which mapped the electronic density of states (DOS) of materials to adsorption energies were physically meaningful and proposed using them in inverse design strategies.[54]

DeepAPD embeddings were computed for each of the binding sites and extracted from the Xe (1 bar) APDs in both the development and test sets. First, the 128-dimensional embeddings were reduced to 3 dimensions using UMAP, as described in the methodology section. The embeddings of the binding sites from the development and test sets were then plotted to identify any binding sites in the test set which are significantly different from those in the development set. Since the test set (sampled from MOSAEC-DB) and training set (sampled from ARC-MOF) were previously found to cover a similar design space (Fig. S9 and S10), it is not surprising that the UMAP DeepAPD embeddings of the binding sites from both sets also span the same space (Fig. 13). However, this also provides some evidence that the binding environments in the test set are not out of distribution, rationalizing the similar performance between the development and test sets.

The same dimension-reduced embeddings of the development set were then plotted and coloured according to the xenon–MOF interaction energy at the corresponding site (Fig. 14). While there are no distinct clusters, there is a smooth binding energy gradient across the plotted embeddings. This suggests the model has learned a relationship between local chemical environment and binding energy, and the embeddings are physically meaningful. While the model does not directly predict guest–host interaction energies, these are of course important in the determination of APDs. Assuming Fig. 14 is a faithful representation of the global structure of the data in the original embedding space, the lowest-energy sites are spatially separated from the highest-energy sites, which would be expected. However, after visually inspecting closely grouped structures and their RDFs in Fig. 14, no clear structural motifs or structural patterns were observed. This is most likely because so many structural environments give rise to similar binding energies for a spherical adsorbate such as Xe. However, we hypothesize that non-spherical adsorbates with energetic contributions beyond dispersion (*e.g.*, electrostatics) are more likely to rely on a narrower range of structural motifs for strong binding. This is similar to what was found by Neporozhnii *et al.*, where their ML model trained on the electronic projected DOS of materials showed that many materials shared similar embeddings despite having vastly different structures/compositions.[55]

Computational scaling and speedup

The feasibility of DeepAPD in high-throughput workflows was evaluated by inspecting the computational scaling and comparing wall-times relative to simulation. Notably, the scaling of DeepAPD depends primarily on the size of the MOF (*i.e.*, the number of "chunks") and the density of the MOF (*i.e.*, the size of the graphs and therefore the number of edges required for each "chunk" of the unit cell). For this reason, the scaling of the code with respect to the total number of atoms is not a very informative metric. Fig. S12 shows the scaling of the

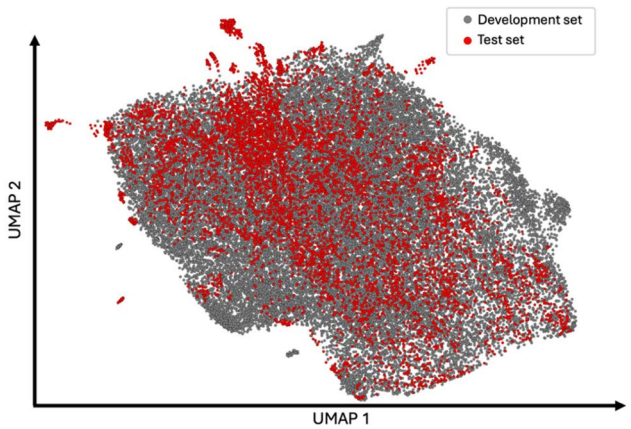

Fig. 13 Universal manifold approximation and project (UMAP) reduced DeepAPD embeddings of binding sites from the test set (red) overlayed on the development set (grey). The binding sites and energies correspond to the Xe (1 bar) sets.

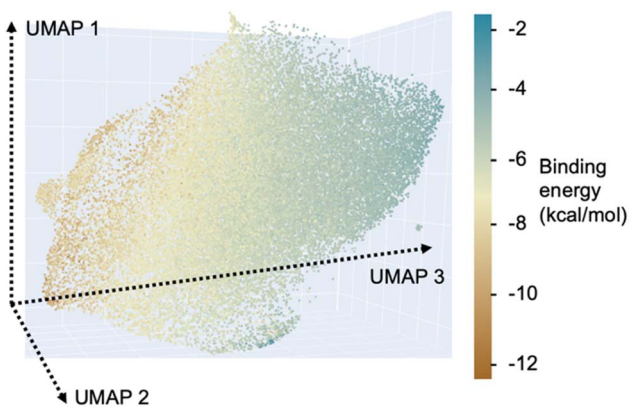

Fig. 14 Universal manifold approximation and project (UMAP) reduced DeepAPD embeddings of binding sites from the development set coloured according to binding energy. The binding sites and energies correspond to the Xe@1 bar sets.

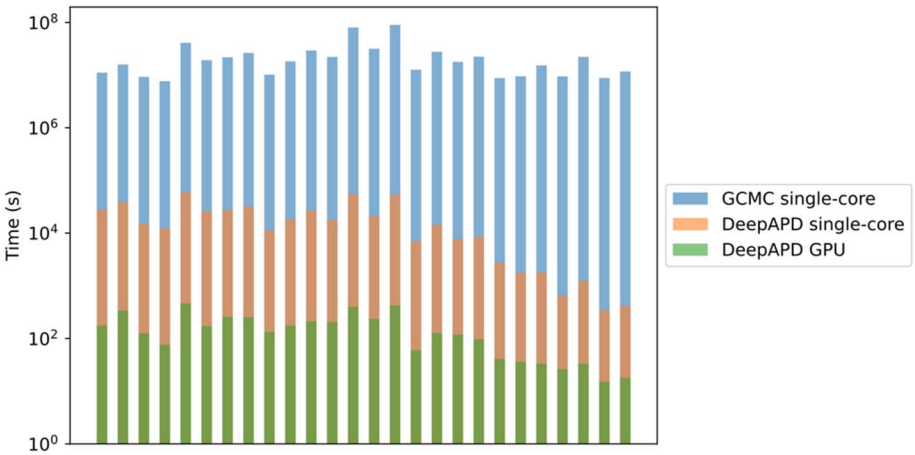

Fig. 15 Relative comparison of wall times to compute and write CH_4@1 bar APDs using GCMC with one core, DeepAPD with one core and DeepAPD with GPU acceleration on a set of 26 MOFs from the MOSAEC-DB test set. Each bar represents one MOF structure (filenames omitted for clarity). The single-core calculations were performed on a single core of an AMD EPYC 9654 CPU, while the GPU accelerated calculations were performed on an NVIDIA H100 GPU and 12 cores of an Intel Xeon Gold 6448Y. The DeepAPD probe batch size used was 1000 in the case of CPU-only execution, while in the GPU-accelerated calculation, a probe batch size of 10 000 was used.

code with respect to number of grid points (*i.e.*, number of total probes required for full evaluation of the APD). The deviation in the scaling is in large part due to the density of the MOFs, as graph construction during each batch starts to become a bottleneck. It is worth noting that the wall-times include the time required to write to disk, which was another notable bottleneck in the prediction of APDs. Nevertheless, for the test set the mean wall-time was only 201 seconds on a single GPU. It is worth noting that probe batch-size could be optimized for each MOF which would lead to significant speedups (to avoid repetitive graph construction), but a constant batch size of 10 000 was chosen. There is also a large spread in computation time because of the geometric diversity of the test set. For context, generating the same data on a single core of an AMD EPYC processor took on average 63 hours for this simplest adsorbate/condition (CH_4 at 1 bar) (Fig. S13).

To obtain a more direct comparison, the GCMC wall times using a single core (AMD EPYC 9654) were compared relative to DeepAPD using a single core (AMD EPYC 9654), and DeepAPD using a GPU (NVIDIA H100 GPU and 12 cores of an Intel Xeon Gold 6448Y CPU) (Fig. 15). Since the test set calculations were run on a highly heterogeneous computing cluster, only a subset of the MOSAEC-DB test set is shown depending on which calculations were run on nodes having an AMD EPYC 9654 CPU. Notably, a batch size of only 1000 probes was used in the CPU-only execution of DeepAPD since it was found that larger batch sizes significantly reduced efficiency when executing the code using only a single core. This is likely because of a shift in bottlenecks (between the DeepAPD model *vs.* graph construction). On average, there is roughly a speedup of $\sim 10^3$ when using DeepAPD on a single core *versus* GCMC on a single core, while the speedup when using GPU acceleration is significantly higher at $\sim 10^5$ relative increase. It is important to emphasize that the average GCMC Tanimoto of the MOFs in the test set is exactly 0.75 ± 0.00, so no computation time was wasted on computing the APDs past the pre-specified convergence criteria. The significant speedup in the single-core execution of DeepAPD shows that even without GPU resources, DeepAPD is a much more rapid way of estimating APDs of adsorbates. When extended to multi-site adsorbates which are more expensive to simulate, the relative speedup will be even more significant since the DeepAPD model execution speed is agnostic to the atom type of the adsorbate, unlike GCMC. For context, the speedup observed for RASPA2 after implementing GPU acceleration was approximately one order of magnitude for GCMC simulations.[56]

Conclusion

The search for high-performing MOFs for gas separation and storage applications has historically relied primarily on global adsorption properties such as uptake, selectivity, and heat of adsorption. However, it has been previously shown that atomistic-level descriptions of adsorption combined with data-driven approaches can be a fruitful approach to identifying synthetic targets. It is also possible that a finer atomistic-level description of adsorption may be required in data-driven workflows connecting the material to its process performance. The development of ML models to predict APDs of materials therefore not only serves a purpose to accelerate binding site identification and data mining but may also serve as a way of generating information-rich material descriptors which could be employed in generative workflows that currently rely on more primitive 3D objects (*e.g.*, energy grids and geometric surfaces). The application of ML to learning APDs of simple single-site adsorbates is a first step in this direction, which will later be expanded to more complex multi-site adsorbates such as CO_2 and even H_2O. When generalizing DeepAPD to multi-site adsorbates, several challenges are expected. Perhaps the most obvious is a more complex potential (*i.e.*, the inclusion of electrostatics), which may be more challenging for the model to

capture. For this reason, newer architecture should be employed, such as those which utilize higher order representations. On a practical note, one should consider whether the model should simultaneously predict all sites for a particular adsorbate, or whether separate models should be used to the predict the APD of each site independently. Finally, automated identification and analysis of resulting binding sites from multi-site APDs will be significantly more complex and require more sophisticated algorithms to assess the performance of DeepAPD for identifying binding sites of multi-site adsorbates in MOFs.

This work has presented an in-depth analysis highlighting the effectiveness of GNNs for predicting APDs, and the utility of the resulting APDs for binding site identification. The APD models were shown to be an accurate and rapid method of identifying binding sites of MOFs, particularly for sites with high probability/occupancy. Importantly, despite the utility of such data, no public database of APDs exists. This work introduces the first such databases for CH_4 at pressures relevant to methane storage (1 bar and 65 bar) and Xe at atmospheric pressure for ~23 K MOFs. We have also implemented the ML models into our automated binding site identification tool to make automated binding site identification straightforward, without requiring a GCMC simulation. With GPU acceleration, DeepAPD can generate APDs up to 10^6 times faster than running a GCMC simulation. We hope that future efforts will prove useful in generative workflows to identify materials with desirable adsorption properties.

Code availability

The DeepAPD code, trained models, and related scripts including the binding site analysis algorithm are available on our group GitHub: https://github.com/uowoolab/DeepAPD. The binding site identification code (GALA) and GCMC code (fastmc) are available on our group GitHub page: https://github.com/uowoolab. The modified version of RASPA used to generate the energy grids is available on GitHub: https://github.com/BurnerJ/RASPA2. All GitHub repositories indicated above are archived in the following Zenodo repository: https://doi.org/10.5281/zenodo.20053445.

Conflicts of interest

There are no conflicts to declare.

Data availability

The GCMC-simulated APDs of the training, development, and MOSAEC-DB test set and the binding sites of the development set and set for CH_4 at 1 bar, CH_4 at 65 bar, and Xe at 1 bar can be found on Zenodo at https://doi.org/10.5281/zenodo.16800893, https://doi.org/10.5281/zenodo.16801034, and https://doi.org/10.5281/zenodo.16801181, respectively. The ML-predicted APDs and corresponding binding sites of the development set and test set can be found on Zenodo (https://doi.org/10.5281/zenodo.16808817).

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6dd00027d.

Acknowledgements

Financial support from the Natural Sciences and Engineering Research Council of Canada (DISCOVERY Grant), the University of Ottawa, MITACS (Accelerate), and TotalEnergies is greatly appreciated. We particularly acknowledge D. Cooper and the TotalEnergies High Performance Computing team for providing resources and technical assistance that were instrumental to the completion of this work. The computing resources provided by the University of Ottawa and the Digital Research Alliance of Canada are also appreciated. JB acknowledges Dr Cecile Pereira, Peigen Xie, and Dr Saman Alavi for helpful discussions.

References

1 J.-B. Lin, *et al.*, A scalable metal–organic framework as a durable physisorbent for carbon dioxide capture, *Science*, 2021, **374**, 1464–1469.
2 C. Charalambous, *et al.*, A holistic platform for accelerating sorbent-based carbon capture, *Nature*, 2024, **632**, 89–94.
3 J. Francis Kurisingal, D. Won Kim and C. S. Hong, Effective approaches to boost Xe/Kr separation in Metal–Organic Frameworks: A review, *Coord. Chem. Rev.*, 2024, **507**, 215731.
4 M. H. Mohamed, *et al.*, Trailblazing Kr/Xe Separation: The Birth of the First Kr-Selective Material, *ACS Appl. Mater. Interfaces*, 2024, **16**, 29364–29373.
5 K.-A. Zhou, *et al.*, Significant Enhancement of Xe Adsorption and Xe/Kr Separation of Metal–Organic Framework Induced by Pore Channel Shaping and Pore Polarity Engineering, *ACS Mater. Lett.*, 2024, **6**, 2809–2815.
6 D. Banerjee, *et al.*, Metal–organic framework with optimally selective xenon adsorption and separation, *Nat. Commun.*, 2016, **7**, ncomms11831.
7 Y. Peng, *et al.*, Methane Storage in Metal–Organic Frameworks: Current Records, Surprise Findings, and Challenges, *J. Am. Chem. Soc.*, 2013, **135**, 11887–11894.
8 W. Xie, *et al.*, Methane Storage and Purification of Natural Gas in Metal–Organic Frameworks, *ChemSusChem*, 2025, **18**, e202401382.
9 P. G. Boyd, *et al.*, Data-driven design of metal–organic frameworks for wet flue gas CO2 capture, *Nature*, 2019, **576**, 253–256.
10 A. J. Campanella, B. A. Trump, A. J. Gosselin, E. D. Bloch and C. M. Brown, Neutron diffraction structural study of CO_2 binding in mixed-metal CPM-200 metal–organic frameworks, *Chem. Commun.*, 2020, **56**, 2574–2577.
11 D. Duong, *et al.*, Observation of binding of carbon dioxide to nitro-decorated metal–organic frameworks, *Chem. Sci.*, 2020, **11**, 5339–5346.
12 M. Asgari, *et al.*, Understanding How Ligand Functionalization Influences CO_2 and N2 Adsorption in a Sodalite Metal–Organic Framework, *Chem. Mater.*, 2020, **32**, 1526–1536.

13 Y. Chen, W. Lu, M. Schröder and S. Yang, Analysis and Refinement of Host–Guest Interactions in Metal–Organic Frameworks, *Acc. Chem. Res.*, 2023, **56**, 2569–2581.
14 H. Yang, *et al.*, Visualizing Structural Transformation and Guest Binding in a Flexible Metal–Organic Framework under High Pressure and Room Temperature, *ACS Cent. Sci.*, 2018, **4**, 1194–1200.
15 I. Gonzalez, *et al.*, Structural characterization of framework–gas interactions in the metal–organic framework Co_2 (dobdc) by *in situ* single-crystal X-ray diffraction, *Chem. Sci.*, 2017, **8**, 4387–4398.
16 R. M. Main, *et al.*, In Situ Single-crystal X-ray Diffraction Studies of Physisorption and Chemisorption of SO_2 within a Metal–Organic Framework and Its Competitive Adsorption with Water, *J. Am. Chem. Soc.*, 2024, **146**, 3270–3278.
17 S. Chen, B. E. G. Lucier, P. D. Boyle and Y. Huang, Understanding The Fascinating Origins of CO_2 Adsorption and Dynamics in MOFs, *Chem. Mater.*, 2016, **28**, 5829–5846.
18 V. Martins, *et al.*, Cold, Hot, Dry, and Wet: Locations and Dynamics of CO2 and H2O Co-Adsorbed in an Ultramicroporous MOF, *Inorg. Chem.*, 2023, **62**, 11152–11167.
19 R. B. Getman, Y.-S. Bae, C. E. Wilmer and R. Q. Snurr, Review and Analysis of Molecular Simulations of Methane, Hydrogen, and Acetylene Storage in Metal–Organic Frameworks, *Chem. Rev.*, 2012, **112**, 703–723.
20 J. G. McDaniel, S. Li, E. Tylianakis, R. Q. Snurr and J. R. Schmidt, Evaluation of Force Field Performance for High-Throughput Screening of Gas Uptake in Metal–Organic Frameworks, *J. Phys. Chem. C*, 2015, **119**, 3143–3152.
21 T. D. Burns, *Pores to Process: The In Silico Study of Metal–Organic Frameworks from Crystal Structure to Industrial Pressure Swing Adsorption for Postcombustion Carbon Capture and Storage*, University of Ottawa, Ottawa, Canada, 2022.
22 A. D. E. F. Gonçalves *et al.*, Fundamental of CO_2 Adsorption and Diffusion in Sub-nanoporous Materials: Application to CALF-20, *arXiv*, 2025, Preprint, arXiv:2507.07791, DOI: 10.48550/arXiv.2507.07791.
23 Y. G. Chung, *et al.*, Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals, *Chem. Mater.*, 2014, **26**, 6185–6192.
24 K. Momma and F. Izumi, VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data, *J. Appl. Crystallogr.*, 2011, **44**, 1272–1276.
25 P. B. Jørgensen and A. Bhowmik, Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids, *Npj Comput. Mater.*, 2022, **8**, 1–10.
26 P. Reiser, *et al.*, Graph neural networks for materials science and chemistry, *Commun. Mater.*, 2022, **3**, 93.
27 J. Burner, *et al.*, ARC–MOF: A Diverse Database of Metal–Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning, *Chem. Mater.*, 2023, **35**, 900–916.
28 M. T. Kapelewski, J. Oktawiec, T. Runčevski, M. I. Gonzalez and J. R. Long, Separation of Xenon and Krypton in the Metal–Organic Frameworks M2(m-dobdc) (M=Co, Ni), *Isr. J. Chem.*, 2018, **58**, 1138–1143.
29 M. Gibaldi, *et al.*, MOSAEC-DB: a comprehensive database of experimental metal–organic frameworks with verified chemical accuracy suitable for molecular simulations, *Chem. Sci.*, 2025, **16**, 4085–4100.
30 J. Burner *et al.*, *ab initio REPEAT Charge MOF Database (ARC-MOF)*, DOI: 10.5281/zenodo.13891643, 2024.
31 T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*, 1958.
32 P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et du Jura, *Bull. Société Vaudoise Sci. Nat.*, 1901, **37**, 547.
33 E. M. Sunshine, M. Shuaibi, Z. W. Ulissi and J. R. Kitchin, Chemical Properties from Graph Neural Network-Predicted Electron Densities, *J. Phys. Chem. C*, 2023, **127**, 23459–23466.
34 K. Schütt, O. Unke and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, pp. 9377–9388.
35 X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
36 G. Boato and G. Casanova, A self-consistent set of molecular parameters for neon, argon, krypton and xenon, *Physica*, 1961, **27**, 571–589.
37 M. G. Martin and J. I. Siepmann, Transferable Potentials for Phase Equilibria. 1. United-Atom Description of *n*-Alkanes, *J. Phys. Chem. B*, 1998, **102**, 2569–2577.
38 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. I. Goddard and W. M. Skiff, UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
39 P. Virtanen, *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Methods*, 2020, **17**, 261–272.
40 A. Togo, K. Shinohara and I. Tanaka, Spglib: a software library for crystal symmetry search, *Sci. Technol. Adv. Mater. Methods*, 2024, **4**, 2384822.
41 K. M. Jablonka, A. S. Rosen, A. S. Krishnapriyan and B. Smit, An Ecosystem for Digital Reticular Chemistry, *ACS Cent. Sci.*, 2023, **9**, 563–581.
42 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials, *Mol. Simul.*, 2016, **42**, 81–101.
43 L. McInnes, J. Healy and J. Melville, Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv*, 2020, Preprint, arXiv:1802.03426, DOI: 10.48550/arXiv.1802.03426.
44 C. J. Nolet *et al.*, Bringing UMAP Closer to the Speed of Light with GPU Acceleration, *arXiv*, 2021, Preprint, arXiv:2008.00325, DOI: 10.48550/arXiv.2008.00325.

45 A. Torres-Knoop, *et al.*, Optimization of Particle Transfers in the Gibbs Ensemble for Systems with Strong and Directional Interactions Using CBMC, CFCMC, and CB/CFCMC, *J. Phys. Chem. C*, 2016, **120**, 9148–9159.

46 A. Torres-Knoop, A. Poursaeidesfahani, T. J. H. Vlugt and D. Dubbeldam, Behavior of the Enthalpy of Adsorption in Nanoporous Materials Close to Saturation Conditions, *J. Chem. Theory Comput.*, 2017, **13**, 3326–3339.

47 A. Rahbari, *et al.*, Recent advances in the continuous fractional component Monte Carlo methodology, *Mol. Simul.*, 2021, **47**, 804–823.

48 B. Mazur, L. Firlej and B. Kuchta, Efficient Modeling of Water Adsorption in MOFs Using Interpolated Transition Matrix Monte Carlo, *ACS Appl. Mater. Interfaces*, 2024, **16**, 25559–25567.

49 A. S. Rosen, J. M. Notestein and R. Q. Snurr, Identifying promising metal–organic frameworks for heterogeneous catalysis *via* high-throughput periodic density functional theory, *J. Comput. Chem.*, 2019, **40**, 1305–1318.

50 A. Sriram, *et al.*, The Open DAC 2023 Dataset and Challenges for Sorbent Discovery in Direct Air Capture, *ACS Cent. Sci.*, 2024, **10**, 923–941.

51 Y. Li, X. Jin, E. Moubarak and B. A. Smit, Refined Set of Universal Force Field Parameters for Some Metal Nodes in Metal–Organic Frameworks, *J. Chem. Theory Comput.*, 2024, **20**, 10540–10552.

52 S. R. Miller, *et al.*, Single Crystal X-ray Diffraction Studies of Carbon Dioxide and Fuel-Related Gases Adsorbed on the Small Pore Scandium Terephthalate Metal Organic Framework, $Sc_2(O_2CC_6H_4CO_2)_3$, *Langmuir*, 2009, **25**, 3618–3626.

53 Z. Yao, *et al.*, Inverse design of nanoporous crystalline reticular materials with deep generative models, *Nat. Mach. Intell.*, 2021, **3**, 76–86.

54 V. Fung, G. Hu, P. Ganesh and B. G. Sumpter, Machine learned features from density of states for accurate adsorption energy prediction, *Nat. Commun.*, 2021, **12**, 88.

55 I. Neporozhnii *et al.*, Navigating Materials Space with ML-Generated Electronic Fingerprints, *Chemrxiv*, 2024, DOI: 10.26434/chemrxiv-2023-j1szt-v2.

56 Z. Li, *et al.*, Efficient Implementation of Monte Carlo Algorithms on Graphical Processing Units for Simulation of Adsorption in Porous Materials, *J. Chem. Theory Comput.*, 2024, **20**, 10649–10666.