SUPPLEMENTARY INFORMATION

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

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

Department of Chemistry and Biomolecular Sciences, University of Ottawa, 10 Marie Curie Private, Ottawa K1N 6N5, Canada
*Corresponding author email: Tom.Woo@uottawa.ca

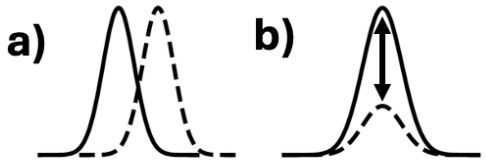

Figure S1. 1-Dimensional depiction of the comparison of two characteristics of maxima from the APDs; a) a difference in position of maxima, and b) a difference in the occupancy of maxima.

Binding site comparison algorithm

An example is shown below to demonstrate how the binding site comparison algorithm works, using the case of GAYXIQ_full from MOSAEC-DB.

1. Identify space group of the MOF to get symmetry opterations
    a. $I4_1/amd$ for this MOF

Origin at $\bar{1}$ at $b\,(c,a)d$, at $0,-\frac{1}{4},\frac{1}{8}$ from $\bar{4}$

Asymmetric unit $\quad 0 \le x \le \frac{1}{2}; \quad -\frac{1}{4} \le y \le \frac{1}{4}; \quad 0 \le z \le \frac{1}{8}$

Symmetry operations

For $(0,0,0)+$ set

(1) 1	(2) $2(0,0,\frac{1}{2})\ \ \frac{1}{4},0,z$	(3) $4^+(0,0,\frac{1}{4})\ \ -\frac{1}{4},\frac{1}{2},z$	(4) $4^-(0,0,\frac{3}{4})\ \ \frac{1}{4},0,z$
(5) $2\ \ \frac{1}{4},y,0$	(6) $2\ \ x,0,\frac{1}{4}$	(7) $2(\frac{1}{2},\frac{1}{2},0)\ \ x,x+\frac{1}{4},\frac{3}{8}$	(8) $2\ \ x,\bar{x}+\frac{1}{4},\frac{1}{8}$
(9) $\bar{1}\ \ 0,0,0$	(10) $a\ \ x,y,\frac{1}{4}$	(11) $\bar{4}^+\ \frac{1}{2},-\frac{1}{4},z;\ \ \frac{1}{2},-\frac{1}{4},\frac{3}{8}$	(12) $\bar{4}^-\ 0,\frac{3}{4},z;\ \ 0,\frac{3}{4},\frac{1}{8}$
(13) $a\ \ x,0,z$	(14) $c\ \ 0,y,z$	(15) $d(\frac{1}{4},-\frac{1}{4},\frac{1}{4})\ \ x+\frac{1}{2},\bar{x},z$	(16) $d(\frac{3}{4},\frac{3}{4},\frac{3}{4})\ \ x,x,z$

For $(\frac{1}{2},\frac{1}{2},\frac{1}{2})+$ set

(1) $t(\frac{1}{2},\frac{1}{2},\frac{1}{2})$	(2) $2\ \ 0,\frac{1}{4},z$	(3) $4^+(0,0,\frac{3}{4})\ \ \frac{1}{4},\frac{1}{2},z$	(4) $4^-(0,0,\frac{1}{4})\ \ \frac{3}{4},0,z$
(5) $2(0,\frac{1}{2},0)\ \ 0,y,\frac{1}{4}$	(6) $2(\frac{1}{2},0,0)\ \ x,\frac{1}{4},0$	(7) $2(\frac{1}{2},\frac{1}{2},0)\ \ x,x-\frac{1}{4},\frac{1}{8}$	(8) $2\ \ x,\bar{x}+\frac{3}{4},\frac{3}{8}$
(9) $\bar{1}\ \ \frac{1}{4},\frac{1}{4},\frac{1}{4}$	(10) $b\ \ x,y,0$	(11) $\bar{4}^+\ \frac{1}{2},\frac{1}{4},z;\ \ \frac{1}{2},\frac{1}{4},\frac{1}{8}$	(12) $\bar{4}^-\ 0,\frac{1}{4},z;\ \ 0,\frac{1}{4},\frac{3}{8}$
(13) $c\ \ x,\frac{1}{4},z$	(14) $b\ \ \frac{1}{4},y,z$	(15) $d(-\frac{1}{4},\frac{1}{4},\frac{3}{4})\ \ x+\frac{1}{2},\bar{x},z$	(16) $d(\frac{1}{4},\frac{1}{4},\frac{1}{4})\ \ x,x,z$

Figure S2. Symmetry operations corresponding to the $I4_1/amd$ space group.

2. Identify binding site coordinates from ML and simulation using GALA. In this example, we see there are *hundreds* of binding sites that are output by GALA, but many are clearly related symmetrically redundant.
    a. Blue = simulation, pink=ML

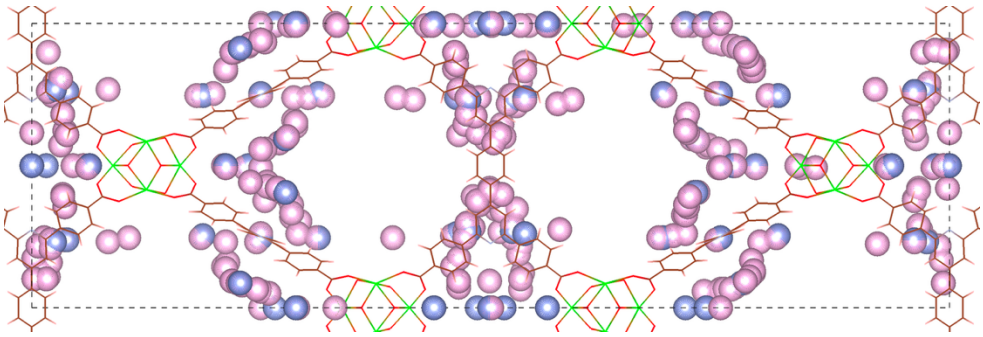

Figure S3. Binding sites identified from ML (pink) and simulation (blue) for GAYXIQ_full.

3.

a. Group simulated binding sites (blue from figure above) based on the symmetry operations of the space group. For this MOF, we get 5 groups of equivalent simulated sites.

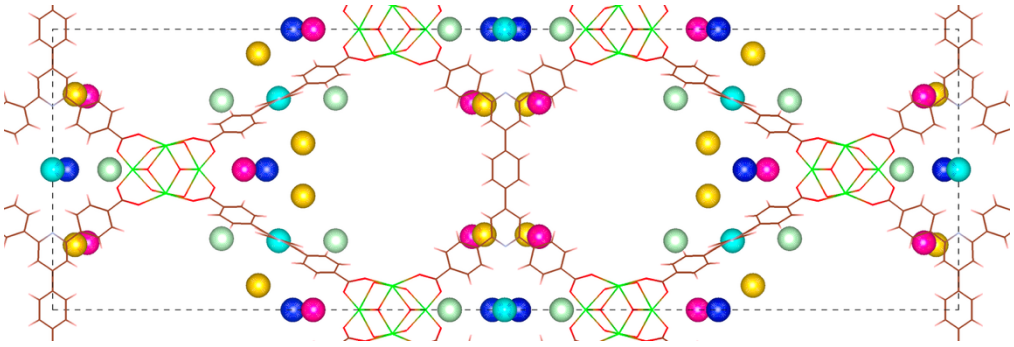

Figure S4. Sets of equivalent binding sites from the simulation APD for GAYXIQ_full identified by colour.

b. Group machine learning binding sites (pink from figure above) based on the symmetry operations of the space group. For this MOF, we get 18 groups of equivalent ML sites.

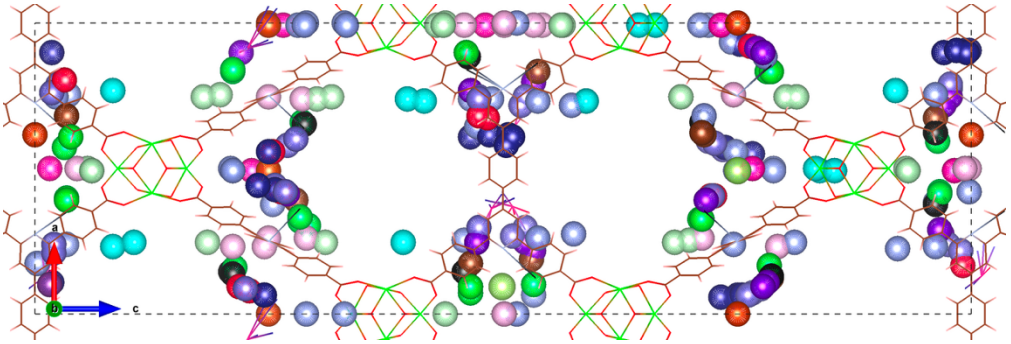

Figure S5. Sets of equivalent binding sites from the ML APD for GAYXIQ_full identified by colour.

4.

a. Identify matching pairs of sites. Within each equivalent grouping (one equivalent set for ML and one equivalent set for simulation), the pair of matching sites with the smallest distance is retained as a "representative" pair for the group. In this case, all simulated sites have been matched (see figure in step 3a), without multiple-counting due to symmetry. These matches all have a distance of 0 (perfect match), with the exception of the cyan site which is off by ~0.15 A. Labels correspond to (ML equivalent set ID, simulated equivalent set ID) and occupancy errors ($\boldsymbol{\varepsilon_{occ}}$). The groups are ordered by occupancy (e.g., in this case, the ML algorithm got the right ordering of sites, with the exception of the yellow site, which had an occupancy of ~10% from simulation.

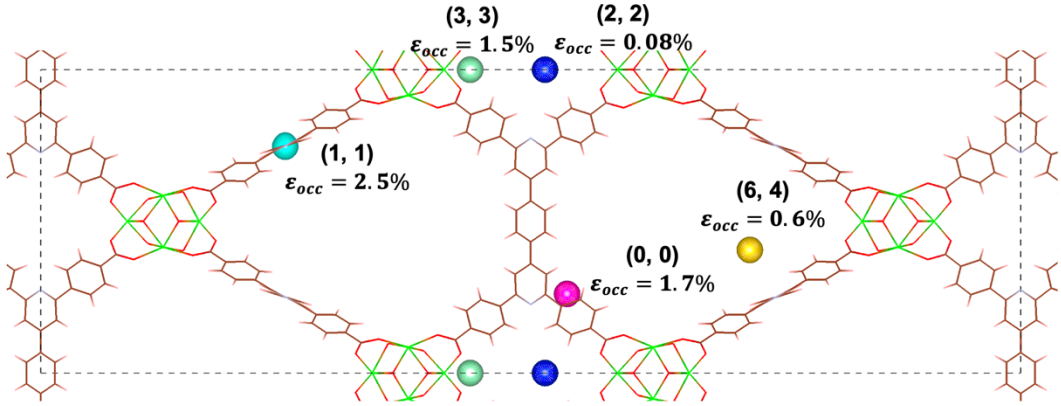

Figure S6. Matched binding sites between ML and simulation, where only a single binding site is represented from each equivalent set of binding sites. The ordered sites by occupancy are identified by indices in parentheses, while the mean absolute error in the occupancy between simulation and ML is shown.

b. Identify missing sites (any which were not matched). In this case, only ML sites were not matched. 6 of the ML sites were matched to at least one simulation site in the 5 groups. This leaves 12 ML groups of sites unmatched. These are shown below, with the occupancy of each site labelled. In this case, the occupancies of the missing sites are generally very low (majority are <13%).

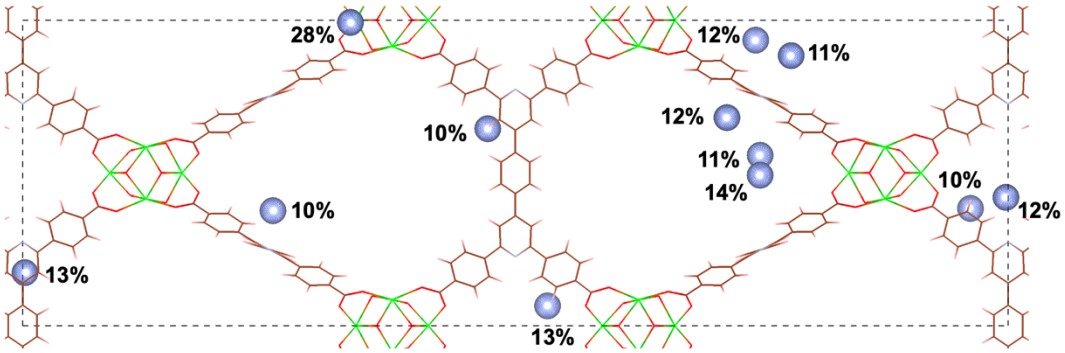

Figure S7. Missing sites from machine learning (blue), i.e., sites which were not matched to simulated sites. All simulated sites were matched in this instance. The occupancy of each missing site is labelled.

5. Output a summary of this comparison and save the results to a Python Pickle file.

Matches (ML index, Sim index): [(0, 0), (1, 1), (2, 2), (3, 3), (6, 4), (8, 4)]

Unmatched ML sites: [53, 82, 96, 99, 115, 121, 141, 154, 186, 224, 233, 281]

Unmatched ML occupancies: [27.71, 14.48, 12.83, 12.78, 12.45, 12.32, 11.84, 11.64, 11.06, 10.51, 10.4, 10.03]

Unmatched Sim sites: [ ]

Unmatched Sim occupancies: [ ]

MAE of occupancies (%): 1.1458333333333321

MAE of distances: 0.10383788336737991

Delta E (most occupied binding energies (kcal/mol): 0.03056100000000006

Figure S8. A sample output from the binding site analyzer script.

Data Diversity

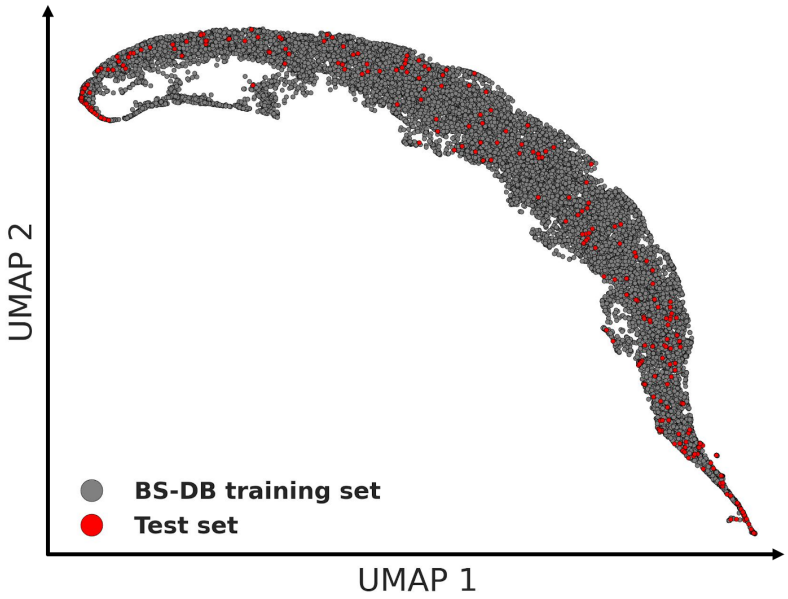

Figure S9. Plot of revised autocorrelation (RAC) descriptors of the organic ligand chemistry of MOFs in the ML training set (sampled from ARC-MOF) and test set (sampled from MOSAEC-DB) projected onto two dimensions using uniform manifold approximation and projection (UMAP).

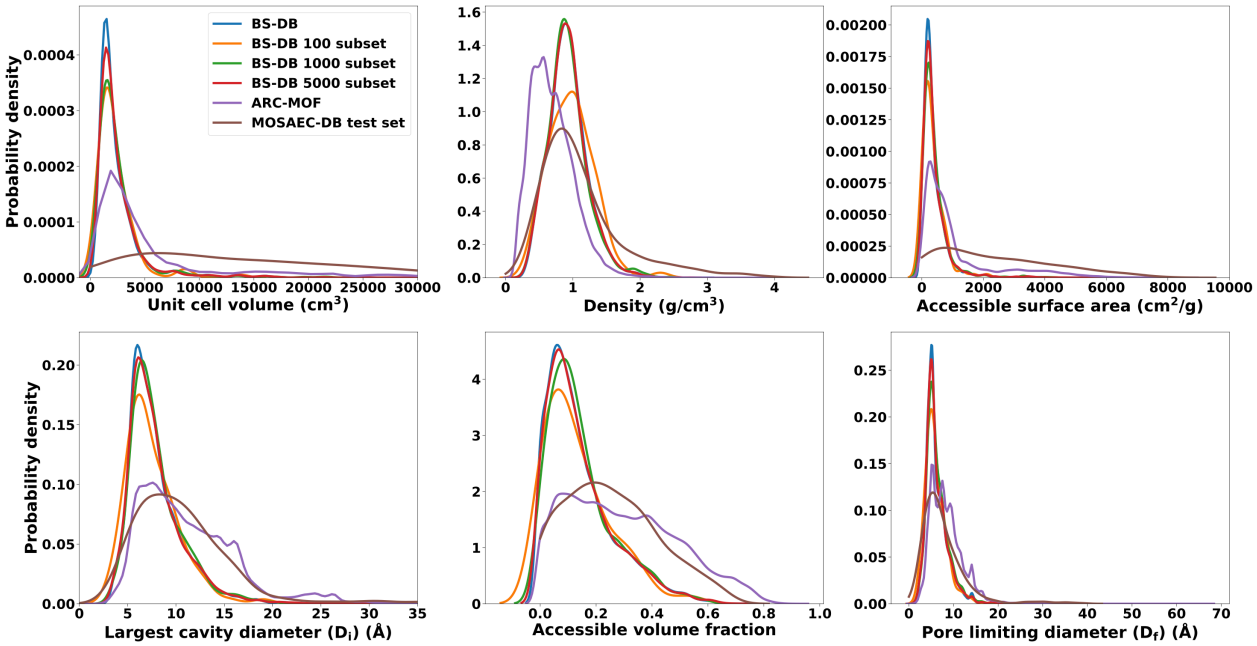

Figure S10. Kernel density estimation (KDE) plots showing the distribution of geometric properties of the ARC-MOF database compared to the test set (sampled from MOSAEC-DB) and various training sets used in this work.

Outliers from DeepAPD

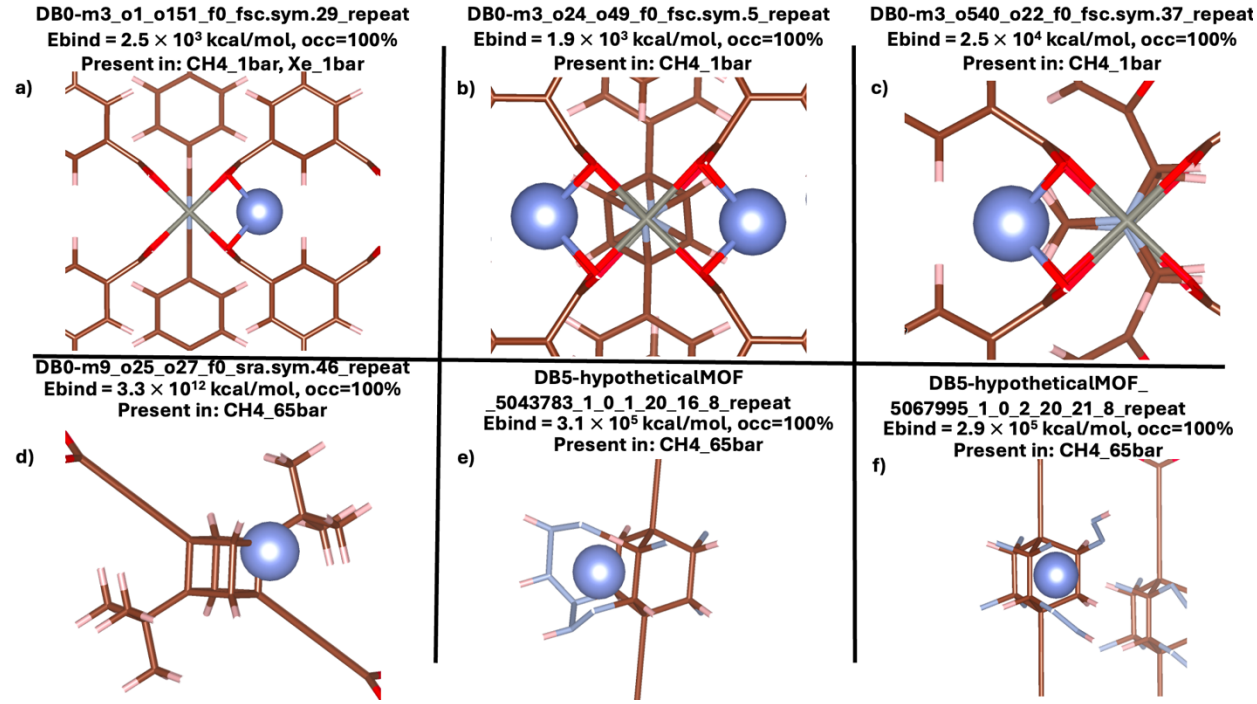

Figure S11. Examples of high probability binding sites from APDs predicted by DeepAPD possessing high binding energy.

Table S1. Statistics on the binding sites in the development set obtained from the ML APDs which possess a positive (repulsive) binding energy.

	Adsorbate/Conditions		
	CH_4 (1 bar)	**Xe (1 bar)**	**CH_4 (65 bar)**
Fraction of MOFs where at least one binding site has a positive energy (%)	0.52	0.08	0.25
Fraction of binding sites with a positive binding energy (%)	0.06	0.01	0.02
Fraction of MOFs where the most occupied site has a positive binding energy	0.17	0.03	0.08

DeepAPD Timings

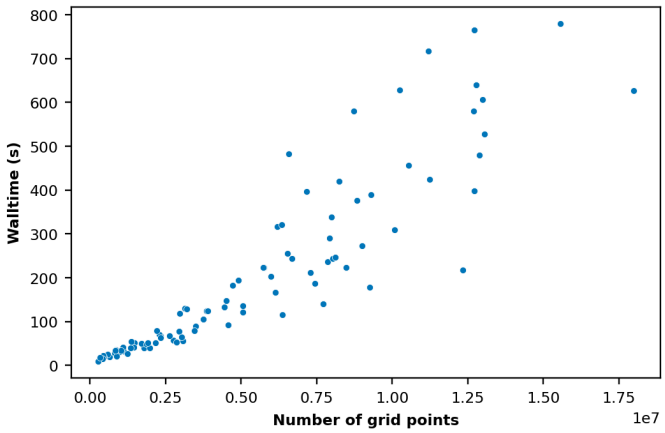

Figure S12. Wall-time of generating CH_4@1bar APDs with 0.15 Å as a function of the number of grid points for the MOSAEC-DB test set using a single NVIDIA H100 GPU and 12 cores of an Intel Xeon Gold 6448Y CPU. A probe batch size of 10,000 was used. The average wall-time was 201 seconds and the maximum wall-time was 779 seconds (~13 minutes).

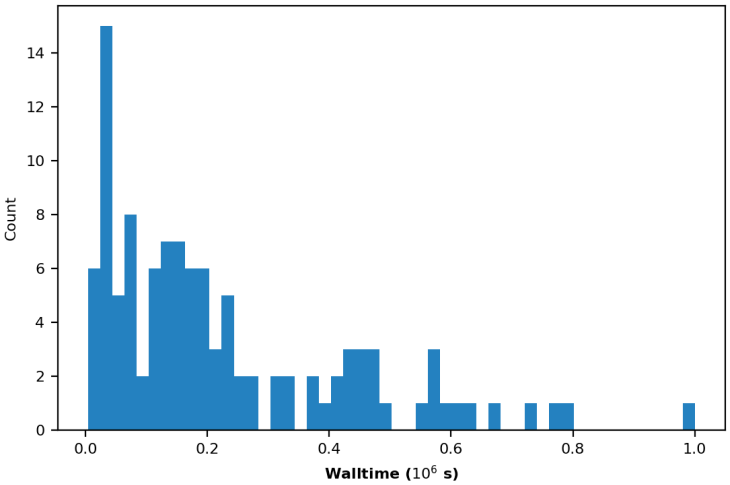

Figure S13. Distribution of wall-times to generate CH_4@1bar APDs of the MOSAEC-DB test set with GCMC on a single core of an AMD EPYC 9654 CPU (base clock speed of 2.4 GHz). The mean wall-time was 229,413 seconds (63 hours). The maximum wall-time was ~278 hours.

Relative entropies

The relative entropy was derived as a metric of defining the uniformity of the APDs with respect to a fully uniform distribution:

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

The maximum entropy distribution (i.e., a completely uniform distribution) was derived as follows (where for a uniform distribution, each point has equal probability, i.e., $p(X = X_n) = \frac{1}{N}$ for a normalized distribution):

$$H_{max}(X) = -\sum_{n=1}^{N} p(X = X_n) \log p(X = X_n) = -\sum_{n=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) = -N \times \frac{1}{N} \log N = -\log\ N$$

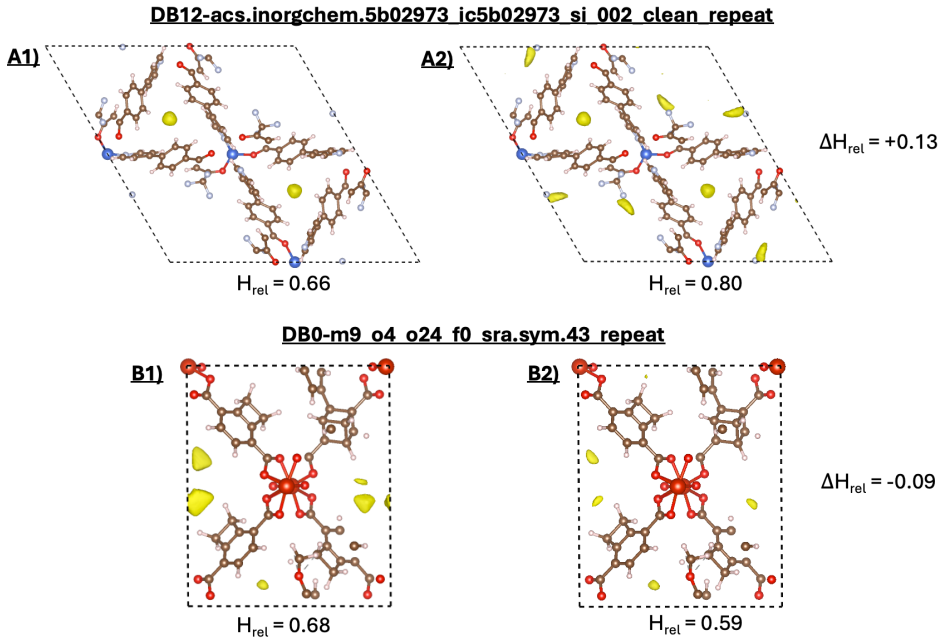

Figure S14. The relative entropy of two MOFs A) DB12-acs.inorgchem.5b02973_ic5b02973_si_002_clean_repeat and B) DB0-m9_o4_o24_f0_sra.sym.43_repeat, with isosurfaces visualized at an isosurface value of 10% of the global maximum, highlighting large positive changes and large negative changes between the CH_4 APDs at A1,B1) 1 bar and A2,B2) 65 bar.

Machine learning details

UMAP parameters

For each UMAP visualization (referenced by figures), the hyperparameters in Table S2 were used.

Table S2. UMAP parameters used for dimensionality reduction of binding site high-dimensional neural network embeddings from DeepAPD (the 'Euclidean' metric was used in each case).

Figure	min_dist	n_neighbours
S9 (2 components)	0.1	60
14 (2 components)	0.1	250
15 (3 components)	0.1	250

For tuning of the DeepAPD models, the ranges of hyperparameters in Table S3 were optimized using a grid search. In most cases, there was very little impact on final model performance amongst the searched hyperparameter space (a change in loss of 1-2% on the development set).

Table S3. Hyperparameters of DeepDFT model and loss functions considered in grid search for optimally performing model.

Parameter	Range searched	Optimal
Distance cutoff (graph generation)	3 Å – 8 Å	6 Å
Initial learning rate	1×10^{-4} – 1×10^{-6}	5×10^{-5}
Loss function	Mean squared error, Kullback-Liebler divergence, Jensen-Shannon divergence, Tanimoto, and combinations thereof	Tanimoto
Number of message passing rounds (layers)	2, 3, 4	3
Hidden channels	128, 256	128
Number of radial basis functions	32, 64	32

The hyperparameters of the readout feedforward neural network were also optimized according to a hyperparameter search corresponding to parameters in Table S4.

Table S4. Hyperparameters of the readout feedforward neural networks considered in the grid search for optimally performing DeepDFT model.

Parameter	Range searched	Optimal
# layers	2, 3, 4	3
Final activation function	Softmax, none	Softmax
Dropout probability	0, 0.2, 0.4, 0.5	0.2
# nodes/layer	128, 256	128

Test set composition

Of the 10,877 MOFs in MOSAEC-DB which were a) neutral; b) 3-dimensional; c) unique; and d) porous (taken to have a pore limiting diameter > 2.4 Å), 500 MOFs were sampled using farthest point sampling based on the geometric and SBU ligand chemistry revised autocorrelation (RAC) descriptors, taken from the original database publication. Finally, to limit computational expense, MOFs for which the APDs were obtained for all three adsorbate/state conditions within 10 days on a single CPU per state point were kept, composing the MOSAEC-DB test set (338 MOFs). Considering the exceptionally large unit cells of some of these MOFs, and the requirement for a minimal supercell to evaluate convergence (see below), the simulation cells were often very large, increasing computational resources beyond what was used per MOF for the training/development sets.

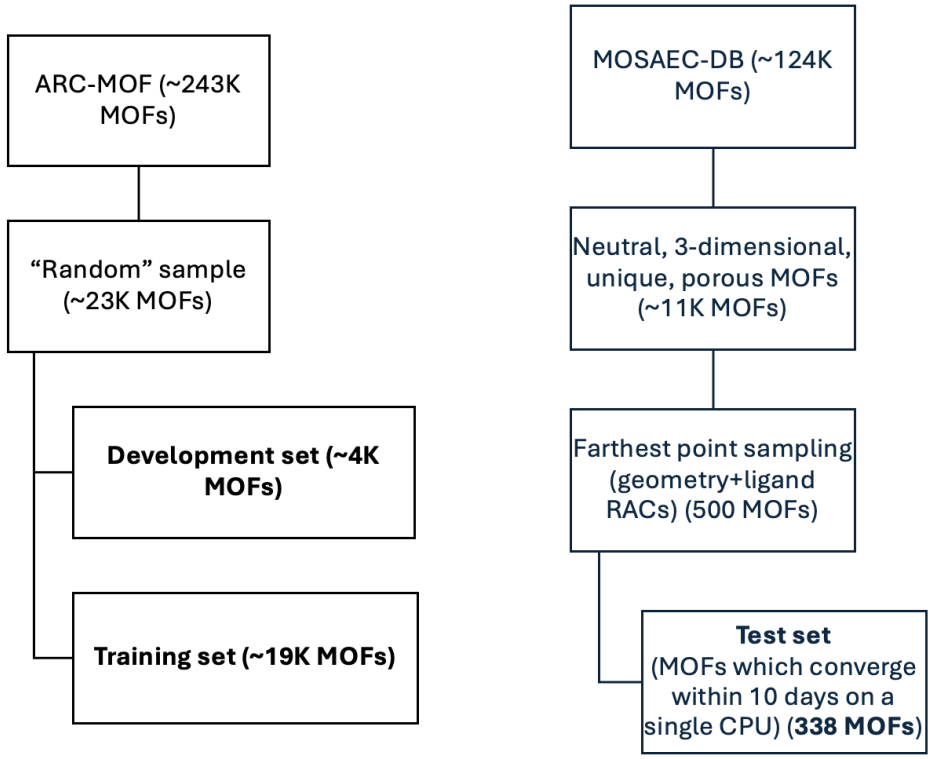

Figure S15. Schematic showing database curation workflow.

Tanimoto Convergence

Initially, 20 million production steps were used for each GCMC simulation. For any simulations which did not converge with this many production steps, an additional simulation was run without equilibration with the same number of production steps, starting from the last accepted configuration. The two resulting distributions were then added together, and convergence was tested again. This process was repeated until a probability distribution with a Tanimoto coefficient of at least 0.75 was obtained.