Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Representation Learning for Long-Chain Hydrocarbon Adsorption in Zeolites

Yachan Liu
Department of Chemical Engineering, University of Massachusetts Amherst
Ping Yang
Department of Chemical Engineering, University of Massachusetts Amherst
Gustavo Perez
College of Information and Computer Sciences, University of Massachusetts Amherst
Aaron Sun
College of Information and Computer Sciences, University of Massachusetts Amherst
Wei Fan
Department of Chemical Engineering, University of Massachusetts Amherst
Subhransu Maji
College of Information and Computer Sciences, University of Massachusetts Amherst
Peng Bai
pengbai@umass.edu

Department of Chemical Engineering, University of Massachusetts Amherst

Article

Keywords:

Posted Date: January 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-6689839/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at Journal of Materials Chemistry A on October 6th, 2025. See the published version at https://doi.org/10.1039/D5TA04285B.

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Representation learning for long-chain hydrocarbon adsorption in zeolites†

Yachan Liu,[a] Ping Yang,[a] Gustavo Perez,[b] Aaron Sun,[b] Wei Fan,[a] Subhransu Maji,*[b] and Peng Bai*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Zeolites are a class of crystalline nanoporous materials known for their ability to discriminate molecules based on size and shape. Such molecular shape selectivity arises from the precise 3-dimensional arrangement of zeolite framework atoms and the resulting non-covalent interactions. In this work, Henry's constants (k_H) for n-octadecane adsorption in all-silica zeolites were used as the target property in a systematic effort to optimize and evaluate various machine-learning representations widely used for materials modeling, including convolutional neural networks (ConvNets) with 3D volumetric grids (ZeoNet, as proposed by *J. Mater. Chem. A* **2023**, 11, 17 570), ConvNets with 2D multi-view images, Vision Transformers with 3D volumetric grids, PointNet and EdgeConv with point clouds of atomic coordinates and solvent-accessible surface, and graph-based neural networks (CGCNN, MEGNet, M3GNet, and MACE). ZeoNet was found to vastly outperform other representations, achieving a correlation coefficient of $r^2 = 0.973$ and a mean-squared error (MSE) of 4.4 in $\ln k_H$. In comparison, the best-performing graph model for this task, M3GNet, obtained $r^2 = 0.888$ and MSE $= 18.5$, reflecting the difficulty of graph models in capturing subtle structural variations and long-ranged spatial correlations. ZeoNet also exhibited excellent transferability to other hydrocarbon molecules, including mono- and di-branched C18 isomers and linear C24 and C30 alkanes. Fine-tuning of pre-trained ZeoNet using training sets of 2100 samples can achieve the same level of performance as ZeoNet trained from scratch using over 10 000 samples. Finally, multi-task learning, which trains a single model with a shared representation and multiple prediction heads, was shown to improve the accuracy in predicting adsorption selectivity without compromising the prediction of Henry's constant for individual molecules.

1 Introduction

Zeolites are crystalline aluminosilicates that are the most widely used industrial catalysts by value.[1] Zeolites contain cages and channels of molecular dimensions that give rise to their ability to preferentially adsorb and react upon molecules of different sizes and shapes. Due to this molecular shape selectivity, zeolite catalysts are often likened to enzymes,[2] and an important direction of research in zeolite science is to develop the ability to identify the most suitable framework structure for a target application. While this is traditionally accomplished through physics-based modeling using first-principles calculations or statistical mechanical simulations, machine learning (ML) has been increasingly applied to expedite the prediction of structure-performance relationships.

An expanding array of ML representations has been investigated for encoding material structures. Among them, the most popular is perhaps graph-based representations, of which nodes and edges map intuitively to atoms and chemical bonds. The crystal graph convolutional neural networks (CGCNN)[3] and SchNet[4] are two of the earliest graph neural networks (GNNs) developed for periodic materials. Here, the initial node and edge feature vectors include atomic properties and distances between atoms, respectively. CGCNN was trained on eight different properties of crystals, including formation energy, absolute energy, band gap, Fermi energy, bulk moduli, shear moduli, Poisson's ratio, and metal/semiconductor classifications, calculated using density functional theory (DFT) for a large, diverse set of inorganic crystals by the Materials Project.[5] Later, GNN models improved upon the earlier studies by incorporating additional geometric attributes beyond atomic distances[6] as well as adding global attributes (e.g., MEGNet[7]). Apart from node updates, angle,[8]

[a] *Department of Chemical Engineering, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States of America. Fax: +1-413-545-1647; Tel: +1-413-545-6189; E-mail: pengbai@umass.edu*

[b] *College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States of America. E-mail: smaji@cs.umass.edu*

† Electronic Supplementary Information (ESI) available: Details of model architecture; Additional tables of numerical data for model performance; Additional Figures of training convergence. See DOI: 10.1039/cXsm00000x/

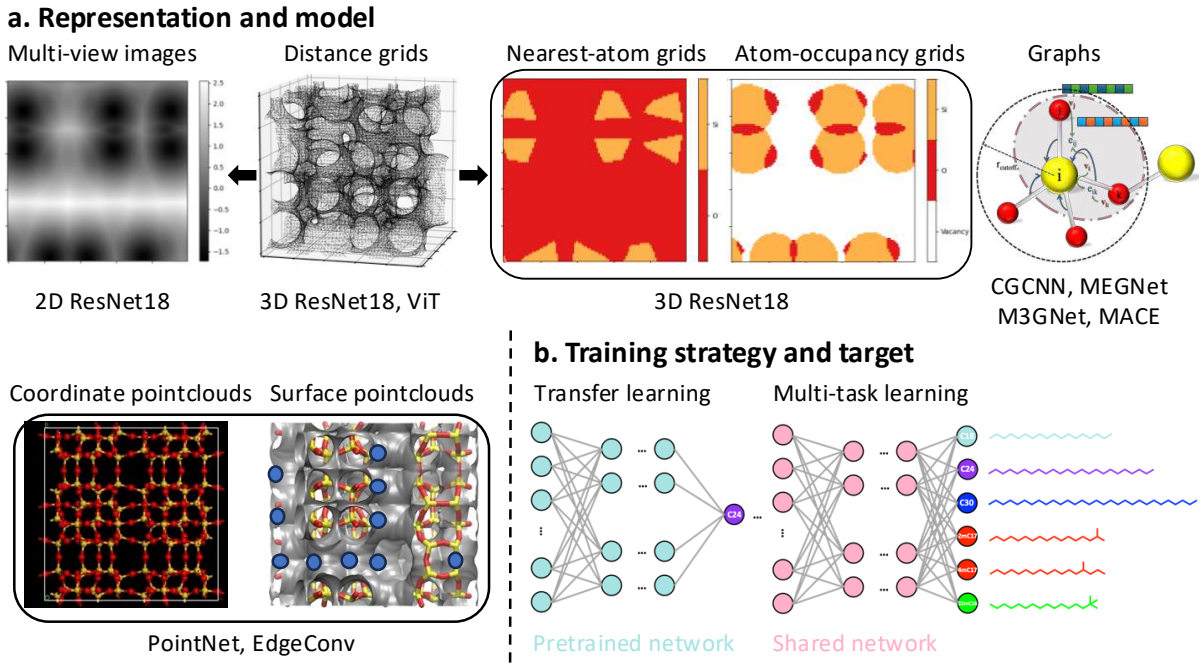


Fig. 1 (a) Representations and models and (b) learning strategies examined in this work for predicting adsorption in zeolites. For clarity of visualization, nearest-atom and atom-occupancy grids are shown in 2D.

edge,[9] and global embeddings[7] can also be updated. In addition, some adapted more sophisticated readout functions beyond average pooling (e.g., global attention mechanism[7,10,11]). GNNs also form the basis of many popular ML potentials such as M3GNet[10] and MACE.[12] However, most of these models encode local features, an issue that has just started to be addressed.[13–15] As will be discussed, significantly larger neighborhood sizes and deeper message passing than typically used, at the cost of substantially higher computational expenses, may be needed for graph-based models to describe the sum of non-covalent interactions that define the subtle shape selectivity in zeolites.

Motivated by the successes of convolutional neural networks (ConvNets) in perceiving shapes in image recognition, Lin et al. adapted a LeNet/AlexNet-based ConvNet architecture to operate on 3D volumetric data to predict the adsorption of methane or CO_2.[16,17] Here, the volumetric representation was constructed as a binary occupancy grid, where each voxel was marked as either zero or one depending on whether it is occupied by framework atoms. Building on ConvNets but targeting larger, flexible guest molecules, we proposed ZEONET, based on a finer-grained 3D volumetric representation of distance grids, in which each grid location was assigned a floating number measuring its distance to the nearest solvent accessible surface.[18] Several modern ConvNet architectures were evaluated, including AlexNet, VGG, ResNet, and DenseNet, of which the ResNet18 architecture was found to offer the best trade-off between expressiveness and efficiency for the task of predicting Henry's constants (k_H, mol kg^{-1} MPa^{-1}) for the adsorption of n-octadecane, a long-chain hydrocarbon with significant conformational flexibility. The best-performing ZEONET achieved $r^2 = 0.977$ and a mean-squared error (MSE) of 3.8 in $\ln k_H$, which corresponds to an error of 9.3 kJ mol^{-1} in adsorption free energy. Instead of being represented as 3D volumetric data, 3D objects can also be sliced or projected into multiple 2D views and processed by image-based ConvNets. The multi-view representation can potentially benefit from highly optimized ConvNet architectures that have been pre-trained on extremely large image datasets. This approach was used, for example, for object detection in X-ray and magnetic resonance images.[19]

The Transformer architecture was initially designed for machine translation tasks and operates on sequences of word tokens.[20] Vision Transformers (ViTs) extend the architectural design to computer vision tasks[21] and rely on the self-attention mechanism to capture both local and global dependencies without convolutions. Transformers scale better than ConvNets and can potentially outperform ConvNets with larger models and datasets. Kang et al. introduced MOFTransformer to predict gas adsorption, diffusion, and electronic properties in metal-organic frameworks (MOFs).[22] MOFTransformer combines atom-based graph embeddings and energy-grid embeddings to represent local and global features. The graph embeddings use the output vectors of CGCNN, while the energy-grid embeddings take 3D energy grids describing the interactions between MOFs and a methane probe molecule. Like ViTs, MOFTransformer flattens the 3D energy grids into a sequence of patches supplemented with a positional embedding and a modal-type embedding (with zero and one for graph and energy-grid embeddings, respectively) that are summed element-wise to form the input tokens.

Finally, atomistic systems can be directly modeled as point clouds. DeFever et al. performed molecular simulations for different phases of Lennard-Jones, water, and mesophase systems and employed PointNet[23] to directly process atomic positions as point clouds. They found that PointNet was able to classify the crystal

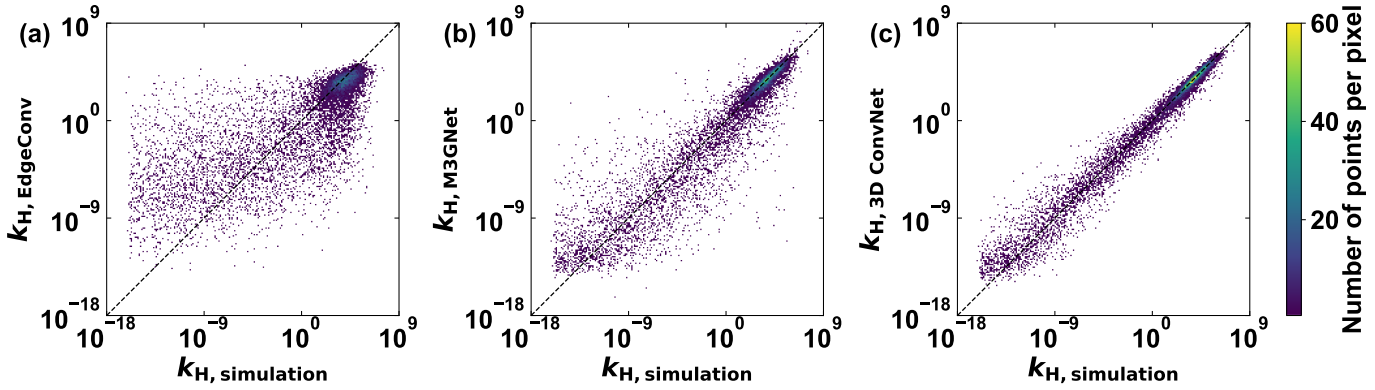


Fig. 2 Scatter plot of the testing performance, comparing Henry's constants for n-octadecane adsorption predicted by (a) EdgeConv, (b) M3GNet, and (c) 3D ConvNet and target values from Monte Carlo simulations.

phases with an overall accuracy of 99.2%.[24] Alternatively, point clouds can also be constructed by sampling points from, say, the van der Waals surface of a molecule. Using this approach, Wang et al. represented each point with its coordinates and electrostatic potential and used PointNet to classify toxicity.[25]

In this work, a comprehensive evaluation of the various representations was conducted on the same discriminative task of predicting the adsorption of n-octadecane in all-silica zeolites (Fig. 1). Results will be presented for (1) ZEONET,[18] in particular 3D ResNet18 using alternative volumetric representations that may be easier to extend to other elements (e.g., aluminum and phosphorus), including nearest-atom grids, atom-occupancy grids, and multi-channel grids combining distance, nearest-atom, and atom-occupancy information; (2) pre-trained 2D ResNet18 on multi-view distance images; (3) ViT[21] on 3D distance grids; (4) PointNet[23] and EdgeConv[26] on point clouds constructed from both atomic positions and points sampled from molecular surfaces; and (5) graph models including CGCNN,[3] MEGNet,[7] M3GNet,[10] and MACE.[12] We also note that an MLP model and an XGBoost model were presented in our previous study to establish a performance baseline.[18] The two models achieved an r^2 of 0.783 and MSE of 35.7 (MLP) and an r^2 of 0.841 and MSE of 26.2 (XGBoost) using five hand-crafted geometric descriptors as input, including pore-limiting diameter, largest-cavity diameter, surface area, pore volume, and density. As shown in Fig. 2 and Table 1, ZEONET still vastly outperformed the other representations. To test the transferability and robustness of the ZEONET representation, we compared transfer learning and multi-tasking for correctly predicting Henry's constant and selectivity for six C18-C30 isomers, which are essential performance characteristics for the hydro-isodewaxing application used for the production of synthetic lubricants.[27]

2 Methodologies

2.1 Adsorption dataset

As in our previous study,[18] the dataset of long-chain hydrocarbon adsorption was used for all ML tasks. This dataset was produced from a computational screening study that used Monte Carlo (MC) simulations to predict the adsorption of six long-chain hydrocarbon molecules that differ in molecular weight and architecture: linear C18, C24, and C30 (n-octadecane, n-tetracosane, and n-triacontane) and mono- and di-branched C18 isomers (2- / 4-methylheptadecanes, 2C17 and 4C17; 2,2-dimethylhexadecane, 22C16).[27] The intermolecular potentials used were developed for a diverse range of molecules and zeolite structures and their accuracy has been validated extensively against experiments.[28] In total, the dataset includes Henry's constant, k_H, in 402 experimentally synthesized structures cataloged by the Structure Commission of the International Zeolite Association (IZASC)[29] and 331 172 computationally predicted structures from the Predicted Crystallography Open Database (PCOD).[30]

In hydro-isodewaxing, zeolites are used to catalyze the selective isomerization of long-chain alkanes in the range of 16 - 40 carbons into slightly branched isomers of similar molecular weight, thereby reducing the pour point and improving the viscosity index of the lubricant product. For this catalytic transformation, the match between the catalyst pore environment and alkane size and shape is critical, as partly captured by the adsorption Henry's constants; ideal zeolites favor linear alkanes while minimizing over-cracking of branched products. However, Henry's constants characterize adsorption behavior at the zero-loading limit. As demonstrated by the comparison to the simulations of high-pressure, liquid-phase multi-component mixtures, extrapolating properties at infinite dilution to reaction operating conditions is far from being perfect.[27] Extending the results reported in this work to cover the full pressure range is therefore an important next step.

For the comparison of model representations, the ML task remains the same as the original ZEONET;[18] i.e., predicting $\ln k_H$ (or roughly, adsorption free energy). Therefore, zeolites with $k_H = 0$ were removed, leaving 100 520 structures (269 IZA zeolites and 100 251 PCOD zeolites). For multi-task learning, the dataset included all zeolites for which $k_H \neq 0$ for at least one isomer, resulting in 101 651 structures. The datasets in both cases underwent the same training/validation/testing split of roughly 7:2:1, but for multi-task learning, structures with $k_H = 0$ entries were removed from the testing set to ensure a balanced assessment; see Table 2 for a detailed breakdown. The validation set was used to evaluate model performance for hyperparameter tuning or for early stopping during model training. The testing set was not used during

Table 1 Model performance for predicting C18 adsorption in zeolites, comparing the accuracy and efficiency of the best performing models in each architectural family including (1) 3D ConvNets, (2) multi-view 2D ConvNets, (3) Vision Transformers, (4) point clouds, and (5) GNNs.

Model	r^2	MSE	Time [second per sample]
(1) 3D ResNet w. distance grids	0.973	4.4	0.13
(2) Multi-view 2D ResNet[a]	0.912	14.7	
(3) Vision Transformer[a,b]	0.869	21.8	
(4) EdgeConv w. atomic coordinates	0.610	65.1	0.11
(5) M3GNet	0.888	18.5	0.044

[a] Performance on the validation set.
[b] Data were obtained by training from scratch. Using 3D ResNet18 features pretrained for the same task gives $r^2 = 0.962$ and MSE $= 6.4$.

any training stages and served as an unbiased assessment of final model performance. The testing set was large enough that its (random) composition will not influence any of our conclusions (see Table 2 of ref.[18]). For transfer learning, we constructed a small random subset of 2100 training and 600 validation samples from the corresponding full set used in multi-task learning. All zeolites within this subset have non-zero k_H values for all six isomers. The model performance was assessed using the same testing set as in multi-task learning.

2.2 Training

For all modeling work, PyTorch v1.11.0 was used with an Nvidia RTX 2080TI or A100 GPU as the accelerator. All models were trained to predict $\ln k_H$ using MSE as the loss function except for multi-task learning (see later).

2.3 3D ConvNets

3D ResNet18 models were trained using the same model architecture, training protocol, and hyperparameters as previously optimized for ZEONET,[18] with data augmentation techniques such as random translations and rotations. In particular, the input grid has a linear dimension (L) of 45 Å and a grid resolution (Δd) of 0.45 Å. The Adam optimizer, a learning rate of 0.001, and a batch size of 16 were used. The training and inference times for 3D ResNet18 were about 0.2 s and 0.13 s per sample, both dominated by grid file IO, while the MC simulations took hours to complete.[18] In principle, the number of epochs is selected such that roughly equivalent amounts of samples are processed in the training run. As before, training on the full dataset used 30 epochs, while training on the smaller subsets (for transfer learning) would take around 1000 epochs. However, we observed that transfer training usually converged within five epochs. We therefore limit the total number of epochs to 100 as an upper bound for training on the smaller subsets.

For transfer learning, a model pretrained for a different task was ported over but with the last fully connected (FC) layer reinitialized randomly. All layers were then fine-tuned, with learning rates for the last FC layer and for all previous layers set to 0.001 and 0.0001, respectively.

Two new 3D volumetric representations were examined in this work. In the **nearest-atom grid**, each voxel was assigned the atomic number of the framework atom that the voxel is closest to, while in the **atom-occupancy grid**, each voxel was assigned the atomic number of the framework atom occupying the voxel or zero if it is not occupied (see Fig. 1). The atom-occupancy grid can thus be viewed as a more detailed version of the binary occupancy grid used by Lin et al.[16,17] Specifically, to construct these two volumetric grids for a zeolite, two auxiliary structures were prepared, one containing only silicon atoms and the other containing only oxygen atoms. Zeo++, version 0.3,[31] was used to calculate the distance grids for the two auxiliary structures, d_O and d_{Si}, with a grid resolution of 0.15 Å. The distance values were calculated between a grid point and the center of framework atoms. In Zeo++ calculations, Si and O atoms have radii of $r_{Si} = 2.1$ Å and $r_O = 1.52$ Å. For the nearest-atom grid, a grid location was assigned the atomic number of oxygen if $d_O < d_{Si}$, and vice versa. For the atom-occupancy grid, each grid location was assigned zero if $d_{Si} > r_{Si}$ and $d_O > r_O$, the atomic number of oxygen if $r_O - d_O > r_{Si} - d_{Si}$, or the atomic number of silicon otherwise.

2.4 2D ConvNets

For 2D ConvNets, a ResNet18 model pre-trained on ImageNet1K[32] was fine-tuned with a batch size of 16, the Adam optimizer, a learning rate of 0.0001, and 30 epochs. Only translation augmentations were used. Three representations were examined:

2D distance images Three 2D slices were extracted from the 3D distance grid, each taken at a random index along one of the three axes, and stacked together to give a 3-channel image of size $(3, 100, 100)$.

3D distance images The 3D distance grid was processed as an image by treating the z-axis as the channel dimension, giving an input tensor of size $(100, 100, 100)$.

2D multi-view images Here, a view is a 3-channel image, where each channel is a 2D slice, giving a tensor of size $(\text{num of views}, 3, 100, 100)$. Three approaches for constructing multi-view representations were considered. In the first approach, each view was formed in the same manner as described above for a 2D distance image, while in the other two approaches, a view was formed by randomly selecting an axis and taking either three random slices perpendicular to that axis or three equidistant slices at a random index i, $i-3$, and $i+3$ (i.e., spaced 1.35 Å apart). The multiple views were processed through the

Table 2 Dataset split for transfer learning and multi-task learning

Full set for multi-task learning and for pre-training during transfer learning			
	Training	Validation	Testing
Total	71166	20375	10110
$k_{\text{H, C18}} \neq 0$	70369	20142	7254
$k_{\text{H, C24}} \neq 0$	62506	17964	7254
$k_{\text{H, C30}} \neq 0$	51599	14710	7254
$k_{\text{H, 2C17}} \neq 0$	68635	19671	7254
$k_{\text{H, 4C17}} \neq 0$	66295	19029	7254
$k_{\text{H, 22C16}} \neq 0$	63114	18109	7254
$k_{\text{H, all}} \neq 0$	51032	14563	7254
Subset for transfer learning			
	Training	Validation	Testing
$k_{\text{H, all}} \neq 0$	2100	600	7254

same ResNet18 and aggregated at a view pooling layer by taking the maximum feature value at each spatial location across all views, before passing through the last fully connected layer in ResNet18.

For 3D distance images and 2D multi-view images, a linear adapter layer was employed at the beginning of the ConvNet, which was observed to improve performance by facilitating the adaptation of new input data, possibly out of domain, to pre-existing weights in the first layer of the ConvNets.[33]

2.5 Vision Transformers

A standard ViT architecture[21] was used, in which the 3D distance grid was split into patches that were flattened into a sequence of tokens and optionally, supplemented with positional embeddings. The Adam optimizer was used with a fixed learning rate of 0.005 or a learning rate of 0.005 - 0.0001 with a weight decay of 0.1. The batch size was set to 64 (without positional embedding) or 32 (with positional embedding). We explored several architectural variations by changing the patch size (20, 25, and 50), the number of ViT blocks (2, 6, and 18), the number of heads in each block (4, 8, 12, 16, and 18), and the hidden dimension used in the linear mapper (64, 128, 256, 512, 768, and 1440). Finally, we tested a ResNet3D encoder with unfrozen pre-trained 3D ResNet features.

Positional embeddings of size $(l+1,d)$ employ d-dimensional vectors to encode spatial information and provide location-specific context to l input tokens. Common positional embeddings were experimented with, including 1D learnable embeddings, 3D learnable embeddings, and 1D/3D sinusoidal embeddings. Similar to the 2D positional embedding of the ViT paper,[21] the 3D positional embedding trains three sets of embeddings, one for each Cartesian axis and with size $d/3$. Based on the coordinates of a patch, the three embeddings are then concatenated to get the final positional embedding for that patch. The learnable embeddings are optimized during training and initialized either randomly or using sinusoidal functions, which provide a smooth and continuous representation of both absolute and relative positional information:

$$P(i,2j) = \sin\left(\frac{i}{10000^{2j/d}}\right) \text{ and } P(i,2j+1) = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

where $i \in \{1,...,l\}$ is the position index and j is the dimension index.

2.6 Point clouds

As mentioned earlier, two types of point clouds were constructed. For the first type, a supercell of a desired size was generated for each zeolite, and a specified number of atoms was taken from the supercell as the point clouds input. For the second type, surface meshes were first generated from the 3D distance grid using the marching cubes algorithm[34] for the isosurface with $d = 0$. Points were then uniformly sampled from the surface of the mesh with a probability proportional to the mesh area. During training, random shuffling is applied as a data augmentation technique.

PointNet[23] and **EdgeConv**[26] were trained to model the point-cloud representations using the Adam optimizer with a learning rate of 0.001. The number of points is constrained by the available GPU memory, model architecture, and batch size. The maximum number of points tested was 4096 for PointNet and 2048 for EdgeConv. EdgeConv processes n points as input, computes a set of edge features between each point and its k neighbors at an EdgeConv layer, and aggregates the features within each set to determine the EdgeConv responses for each point. The ratio of k to n was set as $20:1024$ to yield a consistent "neighborhood size" when the total number of points n is varied. Periodic boundary conditions were used to construct the input volume from the unit cell, but were not considered for the edge feature calculations. The training time for PointNet and EdgeConv with 2048 points was ~ 0.4 s/sample for surface sampling point clouds and ~ 0.1 s/sample for atom coordinate point clouds.

2.7 Graphs

The molecular graphs were generated from the Crystallographic Information File (CIF) for each zeolite using pymatgen.[35] Pair-

wise distances were calculated between atoms and used to determine the creation of edges based on a user-selected cutoff value. Different atom and edge features were used per the prescription of each graph model, as described below, while a detailed description of the model architectures is given in the SI. All models were trained with the Adam optimizer with a learning rate of 0.001, a batch size of 16, and a total of 50 epochs (or 30 epochs for CGCNN). Representative training times on NVIDIA A100 GPU are reported below.

Crystal Graph Convolutional Neural Networks (CGCNN) [3] Atom features for CGCNN were encoded as a 92-dimensional vector using one-hot encoding and included the group number, period number, and block type of the atom in the periodic table, its electronegativity, covalent radius, number of valence electrons, first ionization energy, electron affinity, and atomic volume. Edges were created between an atom and its N_{neighbor} nearest neighbors within a given radius, r_{cut}. Edge features were interatomic distances, d, converted to a vector, the element of which was calculated as $\exp\left[-(d-\mu)^2/(2\sigma_\mu^2)\right]$. The set of Gaussian functions used was centered at $0 \leq \mu \leq r_{\text{cut}}$ with a spacing of $\Delta\mu$ and a variance of $\sigma_\mu^2 = 0.08$ Å^2. In addition, the number of graph convolutional layers, N_{conv}, was varied. The training time for CGCNN with $N_{\text{neighbor}} = 100$, $r_{\text{cut}} = 8$ Å, $\Delta\mu = 0.2$ Å, and $N_{\text{conv}} = 3$ was ~ 0.13 s/sample.

MatErials Graph Network (MEGNet) [7] The input graph for MEGNet was generated from pymatgen structures using the Deep Graph Library.[36] Atom features were represented using one-hot encoding of the atomic number. Bond features were interatomic distance for atoms within a cutoff distance, r_{cut}, which is converted to a vector using a set of N_b Gaussian functions evenly distributed between 0 and $r_{\text{cut}} + 1.0$ with $\sigma_\mu^2 = 1$ Å^2. MEGNet also includes a state attribute vector, initialized as two zeroes, for global message passing. In addition, the number of MEGNet blocks, N_{block}, was varied. The training time for MEGNet with $r_{\text{cut}} = 8$ Å, $N_b = 100$, and $N_{\text{block}} = 3$ was ~ 0.032 s/sample.

Materials Graph Networks with 3-body Interactions (M3GNet) [10] Building on MEGNet, M3GNet includes atomic coordinates and a 3×3 matrix of lattice parameters as the input to allow for its use as machine-learning force fields. M3GNet also includes explicitly 3-body interactions. The cutoff distance, r_{cut}, and the 3-body cutoff distance, $r_{3-\text{body}}$, were studied in this work. The training time for M3GNet with $r_{\text{cut}} = 10$ Å and $r_{3-\text{body}} = 4$ Å was ~ 0.044 s/sample.

Multi-Atomic Cluster Expansion (MACE) [12] Combining equivariant message passing with a higher order local descriptor, the Atomic Cluster Expansion (ACE),[37] MACE has shown high performance and low computational cost for atomic energies and forces on a series of benchmark datasets such as rMD17, 3BPA, and QM9.[12,38] The inputs of MACE model were graph representations of atomic structures produced by ASE.[39] Each graph encoded atom features using one-hot encoding of atomic number, pairwise interatomic distances using Bessel functions, and angles using spherical harmonics. Only the cutoff distance, r_{cut}, was studied in this work. The training time for MACE with $r_{\text{cut}} = 8$ Å was ~ 0.067 s/sample.

3 Results and Discussion

3.1 Representations

3D multi-channel volumetric grids 3D distance grids were introduced in our previous study[18] for use with ConvNets to predict the adsorption of n-octadecane, a long-chain hydrocarbon molecule that exhibits significant conformational flexibility. As a volumetric representation, 3D distance grids allow for an accurate description of shape-selective adsorption in all-silica zeolites. However, the chemical composition of the zeolite framework is only implicitly captured in distance grids through the different radii of framework atoms. Al-exchanged zeolites, for example, will differ only slightly from their all-silica counterparts due to the similar radii of aluminum vs. silicon atoms, both of which are shadowed by outlying framework oxygen atoms. To augment distance grids, we investigate two other volumetric representations, nearest-atom grids and atom-occupancy grids, that encode framework elemental compositions (Fig. 1a).

As compared in Table 3, all three volumetric representations show very similar performance, with r^2 and MSE values ranging from 0.972 and 4.6 for distance grids to 0.966 and 5.7 for nearest-atom grids. It was previously estimated that the size of the testing set (containing $\geq$10,000 samples) allows for r^2 and MSE to be determined to be within 0.002 and 0.3, respectively,[18] which means that the performance for distance grids and atom-occupancy grids is indistinguishable statistically, while the performance is likely slightly worse for nearest-atom grids. In Table 3, multi-channel volumetric grids were also tested, including two-channel (distance + nearest-atom grids and distance + atom-occupancy grids) and three-channel versions. All three cases show a slightly better performance, with r^2 and MSE reaching 0.976 and 3.8, respectively, for the three-channel volumetric representation.

It is worth noting that, unlike point clouds and graph models discussed later, 3D ConvNets presented here are not invariant with respect to translations and rotations, although data augmentation was used to force the trained models to be insensitive to these transformations. To test model sensitivity with respect to translated or rotated inputs, ten random zeolites were selected from the testing set, and for each, ten structure variants were generated by performing random translations and rotations on the original structure. The testing set was used here to ensure that the model has not seen any structural variants of the selected zeolites during training. As shown in Fig. S27, 3D ConvNet achieved an r^2 of 0.987 and MSE of 0.6, with a mean variance of 0.25; by comparison, M3GNet obtained an r^2 of 0.897 and MSE of 4.7, demonstrating that for the purpose of materials screening, the lack of perfect translational and rotational invariance in ZEONET is more than compensated by its higher accuracy for the target application.

2D view-based representations Fig. 3 summarizes the performance of 2D view-based representations. As the baseline, 2D distance images achieved $r^2 = 0.758$ and MSE $= 40.3$, which are worse than the performance of a multi-layer perceptron trained on simple geometric descriptors such as pore sizes and surface ar-

Table 3 Validation accuracy for 3D ResNet18 trained on distance grids, nearest-atom grids, and atom-occupancy grids.

Single-channel grids			
	Distance grids	Nearest-atom grids	Atom-occupancy grids
r^2	0.972	0.966	0.971
MSE	4.6	5.7	4.8
Multi-channel grids			
	Distance grids Nearest-atom grids	Distance grids Atom-occupancy grids	Distance grids Nearest-atom grids Atom-occupancy grids
r^2	0.976	0.975	0.976
MSE	3.9	4.2	3.8

eas (with $r^2 = 0.783$ and MSE $= 35.7$; see Table 1 of Liu *et al.* [18]). As described in Methodologies, these 2D distance images comprise three random slices taken from the three axes and thus contain a rather limited perspective of the zeolite structure. The use of a linear adapter (i.e., the three-axes series with a single view) slightly improves the accuracy. As the number of views increases, the performance of all three view-based representations steadily increases, reaching a plateau above six or seven views. The approach of constructing views by taking three random slices along a single axis as its channels is significantly worse at all view counts: r^2 is lower by 0.04-0.13 and MSE is larger by 6-20. By contrast, taking closely spaced, equidistant slices along the same axis produced the best multi-view model at eight views, showing an r^2 value of 0.912 and MSE of 14.7. This representation outperforms the other two multi-view constructions, presumably because its channel dimension contains voxel values at the same spatial coordinates (in contrast to three separate axes) and with consistent spacing (in contrast to random slices along the same axis). However, its performance decreases substantially at lower view counts, likely due to the narrower spatial regions from which slices are taken. Finally, 3D distance images, which treat the z-axis as the channel dimension, perform slightly worse: when the linear adapter is not used, the r^2 is smaller by 0.02, while the MSE is higher by 15.

Vision Transformers A ViT architecture consisting of two blocks and four attention heads per block was used first to optimize the schemes for data augmentation, patch size, hidden dimension, and positional encoding. As shown in Table 4, translation augmentation led to a surprising decrease in the r^2 and an increase in MSE by 0.17 and 28, respectively, compared to the case when no augmentation was used. Similarly, positional encoding appeared to offer little beneficial effect. The learned embedding and the sinusoidal embedding both caused a change of ~ 0.1 in r^2 and 2-6 in MSE relative to when no positional encoding was used. Given these observations, no input augmentation or positional encoding was employed when studying the effect of patch size and hidden dimension. The optimization of these two hyperparameters showed that a patch size of 20-25 and a hidden dimension of 128-256 yielded nearly optimal results. A bigger patch size of 50 decreased the r^2 by 0.2 and increased the MSE by 35, while a larger hidden dimension of 512 decreased the r^2 by 0.12 and increased the MSE by 19. Next, the choices of learning rate, weight decay, and dropout schemes were slightly varied. As shown in Table 5, a typical dropout rate of 0.1 reduced the r^2 and MSE by about 0.01 and 2.0, respectively, while a learning rate of 0.0001 with a weight decay of 0.1 improved the r^2 and MSE by about 0.04 and 6.9, respectively. Therefore, in the last set of experiments, the smaller learning rate and weight decay were used to optimize the number of ViT blocks and attention heads. The number of hidden dimensions was varied roughly in proportion to the changes in the number of ViT blocks.[21] Table 5 shows that larger ViT models exhibited significantly higher accuracies, with the r_2 increasing by 0.14 and the MSE decreasing by 23. However, the training of even larger ViTs seemed unstable, unable to reach performance plateau in our tests. Overall, the highest performance of ViTs trained from scratch was achieved with a patch size of 20, 18 ViT blocks, 18 attention heads per block, and a hidden dimension of 1440, with $r^2 = 0.869$ and MSE $= 21.8$, which are substantially worse than for 3D ConvNets.

Finally, we tested ViTs with their encoder initialized to pre-trained 3D ResNet18 features, including one with $\Delta d = 0.15$ Å (input dimension, $L = 15$ Å) and another with a more optimal $\Delta d = 0.6$ Å (input dimension, $L = 60$ Å). The former achieved $r^2 = 0.907$ and MSE $= 15.4$, which are better than the corresponding 3D ResNet18 by 0.03 and 4.7, respectively. The latter reached $r^2 = 0.962$ and MSE $= 6.4$, nearly identical to the corresponding 3D ResNet18. We attribute these performance gains mostly to the pre-trained 3D ResNet18 features rather than the ViT architecture. To realize the power of Transformers, a much larger model and/or dataset might be needed.

Point clouds Comparing the two point-cloud representations, points sampled from the $d = 0$ isosurface of the 3D distance grid yielded rather poor performance, as shown in Table 6. The r^2 ranges from 0.128 to 0.395, while the MSE lies between 100.8 and 145.3. With $n = 1024$ points, increasing the 3D grid resolution from 0.45 to 0.3 Å, with a corresponding reduction of input volume from $L = 45$ to 30 Å, improves the performance of both PointNet and EdgeNet by ~ 0.18 in r^2. This suggests that 1024 points are insufficient to capture the shape of zeolite pores over a volume of $45 \times 45 \times 45$ Å^3, which was found to be the minimal size for predicting C18 adsorption with 3D ConvNets.[18] While adding more points seems the obvious solution to this problem,

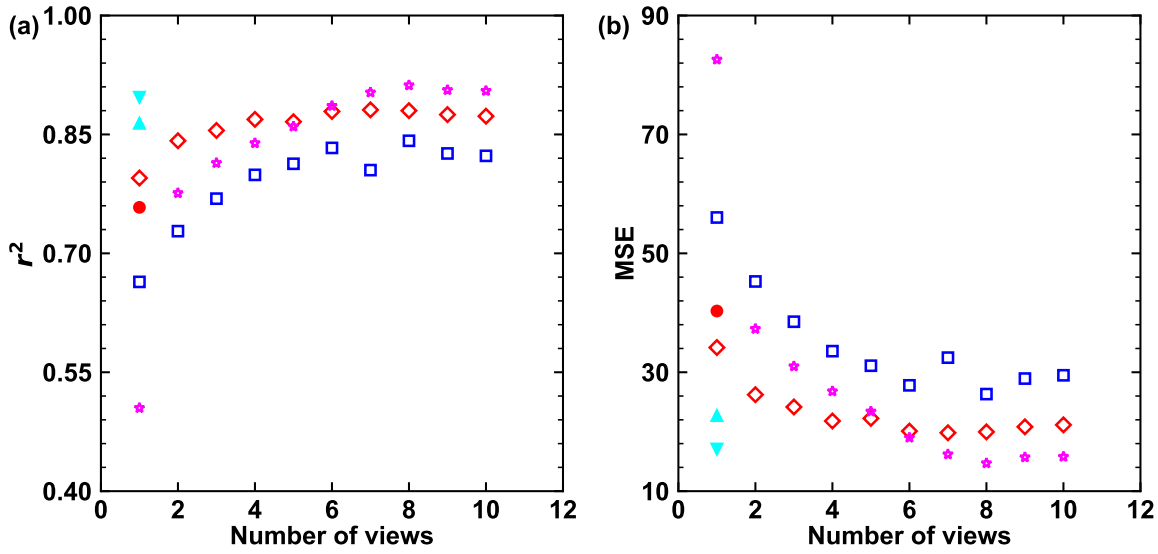


Fig. 3 Validation performance of 2D ResNet18 as a function of number of views, showing (a) r^2 and (b) MSE for 2D multi-view images with views constructed from three axes (red diamonds), with views constructed from a single axis at three random indices (blue squares), and with views constructed from a single axis at closely spaced, equidistant indices (magenta stars). For comparison, 2D distance images and 3D distance images with and without a linear adapter are shown as red circles, cyan up-triangles, and cyan down-triangles, respectively. Here, a view or 2D image has three channels, while in the 3D distance image, the z-axis is viewed as the channel dimension, giving 100 channels; see Methodologies for details.

sampling points from the isosurface substantially increases the overhead of training. In addition, a higher point count potentially requires many more training cycles (Fig. S6 and S7). This difficulty in training is reflected in the performance regression of PointNet trained on 16 384 points with the same number of epochs. Given the trends observed so far, we do not anticipate substantially higher levels of performance with a practical number of points.

The other point-cloud representation processes atomic coordinates directly. Fig. 4 examines the performance of PointNet and EdgeConv as a function of number of points, n. When points are randomly selected within the unit cell, the performance of PointNet gradually increases from $r^2 = 0.391$ and MSE $= 101.6$ at $n = 128$ to $r^2 \sim 0.44$ and MSE ~ 95 at $n =$ 512-1024. Similarly, EdgeConv also improves moderately with more points, reaching a maximum of $r^2 = 0.34$ and MSE $= 110$ at around $n =$ 512-1024. The performance plateau is likely determined by the distribution of the number of atoms within the unit cell among all zeolites, which has a mean of $\langle n \rangle = 199$ and a standard deviation of $\sigma_n = 99$. Zeolite SFV has the most atoms per unit cell ($n = 2352$). Adding more points above this value (by duplication) will not yield any benefit due to the PointNet architecture and its use of max pooling.[23]

Next, the effect of input volume was investigated. In this series of experiments, the unit cell was replicated along all three axes to create a supercell, from which a cubic region of increasing dimension was taken out and points were randomly selected within the cube. In this setup, both PointNet and EdgeConv show a much more pronounced improvement with increasing values of n. At $L = 30$ Å, the performance of PointNet increases from $r^2 = 0.218$ and MSE $= 130.4$ ($n = 256$) to 0.489 and 86.9 ($n = 2048$), while the performance of EdgeConv increases from $r^2 = 0.175$ and MSE $= 137.5$ ($n = 256$) to 0.610 and 65.1 ($n = 2048$). Interestingly, with the same number of points, for example at $n = 1024$, the performance of both models decreases with L, possibly arising from the fact that the point clouds were unable to include all atoms within the larger cubes. In fact, using the median density of $\langle \rho_{\mathrm{Si/O}} \rangle = 51$ per nm^3 (Table 3 of Liu *et al.*[18]), the number of atoms would be expected to increase from $n = 1377$ to 11 016 when L increases from 30 to 60 Å. Compared to PointNet, EdgeConv appears to handle sparser point clouds better, thus benefiting more significantly from a larger input volume, which is likely due to the explicit use of neighbor information in computing point features.

Overall, the highest performance achieved with point-cloud representations, $r^2 = 0.610$ and MSE $= 65.1$, is substantially worse than models trained on simple geometric descriptors ($r^2 = 0.783$ and MSE $= 35.7$; see Table 1 of Liu *et al.*[18]). It is worth noting that both PointNet and EdgeConv have been demonstrated to learn orientation-independent features of an unordered set of points and both models achieved good performance in 3D computer vision tasks including object classification and segmentation on popular datasets such as ShapeNet.[40] The poor performance observed here reflects the difficulty of point clouds, without an excessive number of points, in capturing subtle pore-shape variations that control the adsorption characteristics of different zeolites.

Graph models As one of the earliest graph models for periodic materials, CGCNN was evaluated to establish a performance baseline for other graph models. However, as shown in Table 7, CGCNN with default settings ($r_{\mathrm{cut}} = 6$ Å, $\Delta\mu = 0.2$ Å, $N_{\mathrm{conv}} = 3$, and $N_{\mathrm{neighbor}} = 12$) exhibited extremely poor performance. A pa-

Table 4 Validation accuracy for ViT trained with varying augmentation, patch size, hidden dimension, and positional encoding. These parameters were studied one at a time while keeping the others at near optimal values: no data augmentation, a patch size of 20, a hidden dimension of 256, and no positional encoding.

	r^2	MSE
Augmentation		
None	0.756	40.5
Translation	0.585	68.9
Patch size		
20	0.756	40.5
25	0.747	42.0
50	0.538	76.6
Δd (Å) of input to 3D ResNet18		
0.15	0.907	15.4
0.6	0.962	6.4

	r^2	MSE
Hidden dimension		
64	0.705	48.9
128	0.757	40.3
256	0.756	40.5
512	0.640	59.7
Positional encoding		
None	0.753	49.2
1D learned	0.740	43.1
3D sinusoidal	0.760	47.6

Table 5 Validation accuracy for ViT, showing the effect of ViT blocks, attention heads, learning rate scheduler, and dropout.

Num of blocks	Hidden dimension	Num of heads	Base learning rate	Weight decay	Dropout	r^2	MSE
2	256	8	0.005	0	0	0.732	44.8
2	256	16	0.005	0	0	0.763	39.5
2	256	16	0.005	0	0	0.732	44.8
2	256	16	0.005	0.01	0	0.785	35.9
2	256	16	0.001	0.03	0	0.802	32.9
2	256	16	0.0001	0.1	0	0.805	32.6
2	256	16	0.005	0.01	0.1	0.773	37.9
2	256	16	0.001	0.03	0.1	0.790	35.1
2	256	16	0.0001	0.1	0.1	0.799	33.6
6	768	12	0.0001	0.1	0	0.855	24.2
6	768	12	0.0001	0.1	0.1	0.842	26.3
18	1440	18	0.0001	0.1	0	0.869	21.8
18	1440	18	0.0001	0.1	0.1	0.862	22.9

Table 6 Validation performance of PointNet and EdgeConv on surface sampling point clouds. The points were sampled from the $d = 0$ isosurface of the $100 \times 100 \times 100$ 3D distance grids with a grid resolution of 0.45 Å.

Num of points	PointNet		EdgeConv	
	r^2	MSE	r^2	MSE
1024[a]	0.310	115.1	0.395	100.8
1024	0.128	145.3	0.224	129.4
2048	0.212	131.4		
4096	0.187	135.5		

[a] Grid resolution of 0.3 Å.

rameter search was therefore conducted to optimize model hyperparameters. As the edge cutoff was increased to $r_{\text{cut}} \geq 8$ Å, the r^2 increased from 0.104 to 0.251, while the MSE decreased from 149 to 125. The non-smooth variation of these values with r_{cut} reflects the training difficulty due to the overall poor model performance. Keeping r_{cut} at 8 Å while increasing the number of Gaussian functions or the number of convolutional layers does not yield significant benefits, but a wide spacing beyond $\Delta\mu \geq 0.4$ Å is detrimental, and the minor improvement from increasing N_{conv} comes with the cost of substantially longer training times and GPU memory consumption.

The sizable effect of r_{cut} motivated us to investigate ways to incorporate a larger number of neighbor atoms when aggregating node features. While the original paper of CGCNN stated that only two atoms sharing a Voronoi face are considered to be connected,[3] we discovered that the latest CGCNN code base simply constructs edges between an atom and its N_{neighbor} nearest neighbors. Therefore, the neighborhood size is controlled by both r_{cut} and N_{neighbor}. Table 7 shows that the performance of CGCNN improves almost linearly with N_{neighbor}, reaching $r^2 = 0.781$ and MSE $= 36.5$ at $N_{\text{neighbor}} = 100$, which roughly extends to a radius of 7.8 Å (estimated based on a median density of $\langle \rho_{\text{Si/O}} \rangle = 51$ per nm^3). Given the relatively slow speed of CGCNN, further increases of r_{cut} and N_{neighbor} were not pursued.

Similar trends were observed for MEGNet, M3GNet, and MACE (Fig. 5 and numerical data in Tables S5-7). The default settings of these models ($r_{\text{cut}} = 45$ Å) performed rather poorly, giving r^2 values ranging from 0.241 (MEGNet) to 0.544 (MACE) and the MSE ranging from 127 (MEGNet) to 76 (MACE). The performance im-

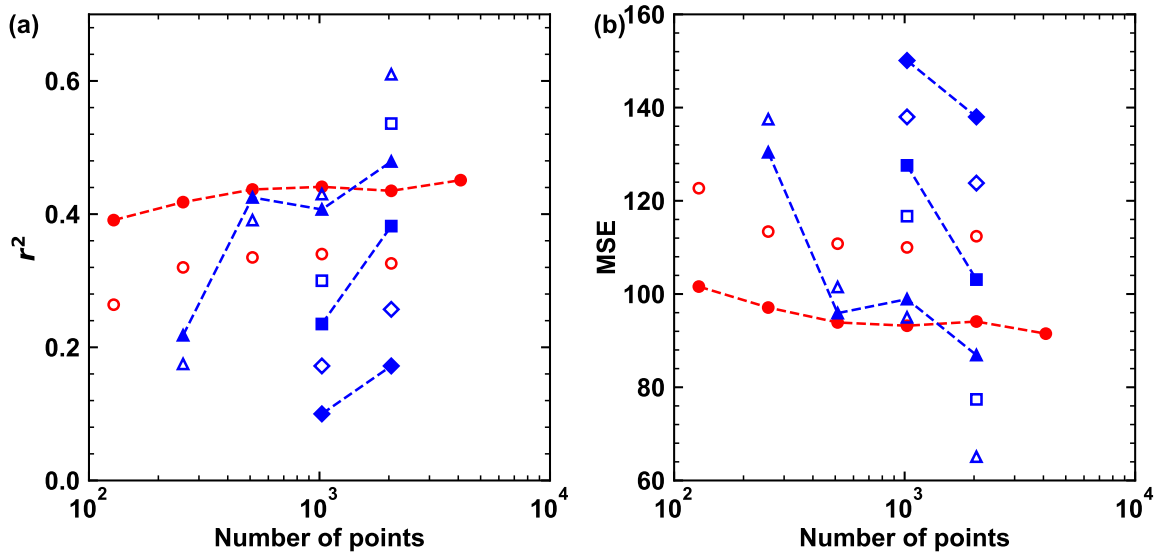


Fig. 4 Validation performance: (a) r^2 and (b) MSE, as a function of number of points for PointNet (filled symbols) and EdgeConv (empty symbols) trained on point clouds of atomic coordinates within the unit cell (red circles), 30 Å cubes (blue up-triangles), 45 Å cubes (blue squares), and 60 Å cubes (blue diamonds).

Table 7 Validation accuracy of CGCNN with different model settings. The default setting and the best-performing setting are indicated in bold.

N_{neighbor}	r_{cut}	$\Delta\mu$	N_{conv}	r^2	MSE
12	4	0.2	3	0.150	142
12	**6**	**0.2**	**3**	0.104	149
12	8	0.2	3	0.251	125
12	8	0.05	3	0.239	127
12	8	0.1	3	0.213	131
12	8	0.4	3	0.063	156
12	8	0.8	3	0.055	157
12	8	1.0	3	0.092	151
12	8	0.2	4	0.281	120
12	8	0.2	5	0.283	120
12	8	0.2	6	0.288	119
20	8	0.2	3	0.330	112
40	8	0.2	3	0.408	98.7
60	8	0.2	3	0.569	71.9
100	8	0.2	3	**0.781**	**36.5**

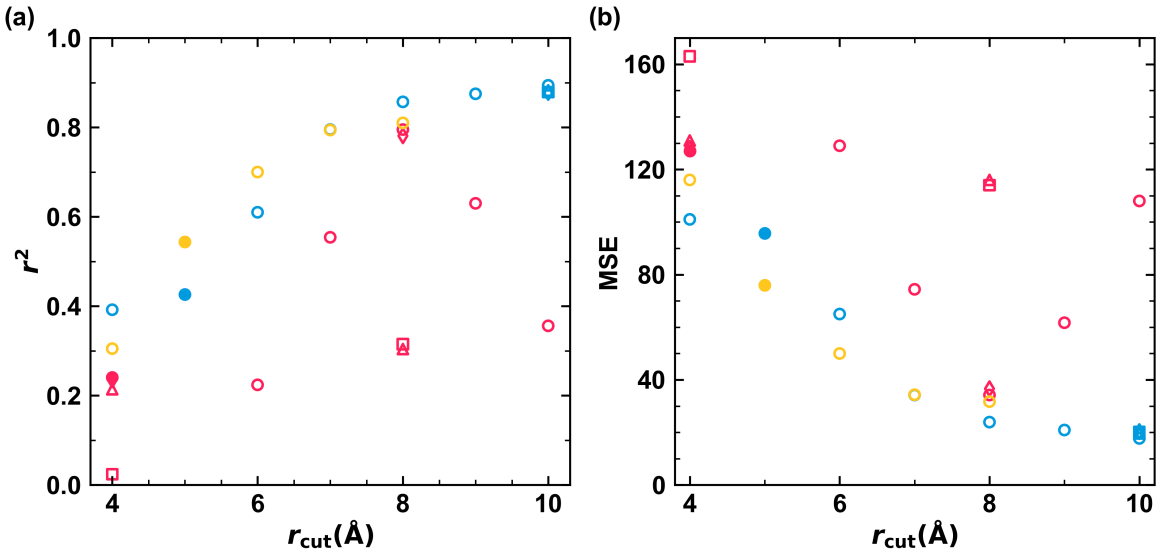


Fig. 5 Validation performance of graph models as a function of r_{cut}, showing (a) r^2 and (b) MSE for MEGNet (red), M3GNet (blue), and MACE (yellow). The default settings for each model are indicated by filled circles. For MEGNet, circles represent results obtained with $N_b = 100$ and $N_{\text{block}} = 3$, while diamonds, squares, and triangles correspond to $N_b = 40,\ 20$, and 10 ($r_{\text{cut}} = 4$ Å), or $N_{\text{block}} = 4,\ 5$, and 6 ($r_{\text{cut}} = 8$ Å), respectively. For M3GNet, circles represent results obtained with $r_{3-\text{body}} = 4$ Å, while diamonds, squares, and triangles correspond to $r_{3-\text{body}} = 3,\ 5$, and 6 Å, respectively.

proves substantially upon increasing r_{cut} to 8-10 Å. The best performance was obtained for M3GNet with $r_{\text{cut}} = 10$ Å, reaching $r^2 = 0.894$ and MSE= 17.7. For MEGNet, reducing the number of basis functions slightly decreased the performance, while surprisingly, adding MEGNet blocks ($N_{\text{block}} \geq 5$) resulted in a significant decline in its performance, likely caused by a more difficult training process (see Fig. S23). For M3GNet, varying $r_{3-\text{body}}$ between 3 and 6 Å did not lead to discernible differences, indicating the short-range nature of 3-body effects. Compared to CGCNN and MEGNet, the substantial GPU memory requirements of M3GNet and MACE pose a significant challenge for fine-tuning model hyperparameters, especially for further increasing r_{cut}. Overall, the results here underscore the importance of cutoff radii for graph models to capture long-range interactions. They also highlight the difficulty of graph models in learning accurate shape selectivity in zeolites dominated by such long-range, non-covalent interactions.

3.2 Transfer learning

The benchmarks presented in the preceding section and in our prior work[18] demonstrate the superior performance of 3D ConvNets with volumetric grids in predicting the adsorption strength of a long-chain linear alkane. In particular, 3D ResNet18 with single-channel distance grids offers the best balance of model accuracy and computational efficiency. Our previous study further indicates that a minimum of 10 000 training samples is needed to reach peak accuracy.[18] In this section, the robustness and transferability of ZEONET were evaluated for five other tasks, including the adsorption of mono- and di-branched C18 isomers and linear C24 and C30.

As shown in Table 8, ZEONET trained from scratch on the full dataset, the size of which is given in Table 2, showed excellent performance for all four C18 isomers, with $r^2 > 0.93$ and MSE < 4.5. As the carbon number increases, the performance of ZEONET declines, with C30 exhibiting a precipitous drop in the r^2 to 0.802 and an increase in the MSE to over 38. It is likely that the distance grids optimized for C18 (i.e., $L = 45$ Å and $\Delta d = 0.45$ Å) are inadequate for the much longer C30 molecule. As a rough estimate, considering a C-C bond length of 1.54 Å and a C-C-C angle of 114°, a fully stretched C30 would extend over 38.5 Å.

Table 8 also includes the performance metrics of ZEONET trained from scratch using only 2100 training samples. Unsurprisingly, the r^2 decreases by 0.04-0.07 and the MSE increases by 2-4 for C18 isomers and C24, while the performance degrades even more for C30, by 0.13 in the r^2 and 25 in the MSE. In comparison, transfer learning yielded much more promising results. For transfer learning, 3D ResNet18 was initialized with a pretrained model and fine-tuned using a smaller learning rate for all but the output layer (see Methodologies). As shown in Table 9, models pretrained for C18 adsorption improved the r^2 and MSE values for the other five molecules by 0.03-0.09 and 2-17, respectively, with the largest gains observed for C30. These performance metrics are comparable to or even slightly better than training from scratch using the full training set. Finally, the additional columns of Table 9 show that models pretrained on any of the six tasks (i.e., C18, C24, C30, 2C17, 4C17, or 22C16) provide a similar level of benefits for transfer learning to the other tasks, averaging improvements of 0.05-0.06 and 2.5-6.2 in the r^2 and MSE, respectively. The smallest transferability is again found with models pretrained on C30 adsorption, but no adsorption task stands out as the best base model for the purpose of transfer learning.

Table 8 Testing-set performance of 3D ResNet18 with single-channel distance grids trained from scratch using the full dataset or using the transfer-learning subset for the adsorption of six alkane molecules. Note that the training/validation/test split is different from our previous work,[18] as described in Methodologies.

	Full set		Subset	
	r^2	MSE	r^2	MSE
C18	0.942	1.8	0.892	3.4
C24	0.914	6.1	0.858	10.2
C30	0.802	38.1	0.672	63.2
2C17	0.932	2.5	0.882	4.3
4C17	0.935	2.8	0.859	6.2
22C16	0.947	4.3	0.906	7.7

Table 9 Transfer learning performance evaluated on the testing set, comparing models pre-trained for one adsorption task using the full dataset (designated as Pt. C18, C24, C30, 2C17, 4C17, and 22C16) and fine-tuned for the other five tasks using a training set of 2100 samples.

	Pt. C18		Pt. C24		Pt. C30		Pt. 2C17		Pt. 4C17		Pt. 22C16	
	r^2	MSE	r^2	MSE	r^2	MSE	r^2	MSE	r^2	MSE	r^2	MSE
C18	0.942	1.8	0.945	1.7	0.927	2.3	0.947	1.6	0.947	1.7	0.933	2.1
C24	0.933	4.8	0.914	6.1	0.916	6.0	0.924	5.4	0.93	5.0	0.917	5.9
C30	0.761	46.1	0.766	45.0	0.802	38.1	0.737	50.7	0.747	48.8	0.731	51.8
2C17	0.939	2.2	0.938	2.3	0.931	2.5	0.932	2.5	0.945	2.0	0.933	2.4
4C17	0.943	2.5	0.925	3.3	0.917	3.7	0.944	2.5	0.935	2.8	0.923	3.4
22C16	0.940	4.9	0.940	4.9	0.940	4.9	0.948	4.3	0.945	4.5	0.947	4.3

3.3 Multi-task learning

In the previous section, separate ConvNet models were trained for the different adsorption tasks, despite the possibility that their convolutional layers may have learned similar features. Here, a multi-task learning strategy is examined, where a single model with multiple regression heads is trained to predict all six adsorption labels simultaneously, promoting shared representations that may improve generalization to novel tasks. Multi-task learning also allows for the possibility of including in the loss functions the selectivity between molecules, defined as $S_{ij} = \ln(k_{\mathrm{H},i}/k_{\mathrm{H},j})$. By maintaining the relative differences between molecules, the joint model may be better suited for optimizing separation processes.

In particular, the following loss functions were tested that give different weights to Henry's constants ($\ln k_{\mathrm{H},i}$) vs. selectivity:

1. Average MSE $= \sum_{i=1}^{6} \mathrm{MSE}_i/6 = \sum_{i=1}^{6} \left(\ln \hat{k}_{\mathrm{H},i} - \ln k_{\mathrm{H},i}\right)^2/6$
2. Average selectivity $S = \sum_{i=2}^{6} \left|\hat{S}_{1i} - S_{1i}\right|/5 = \sum_{i=2}^{6} \left|\ln\left(\hat{k}_{\mathrm{H},1}/\hat{k}_{\mathrm{H},i}\right) - \ln\left(k_{\mathrm{H},1}/\ln k_{\mathrm{H},i}\right)\right|/5$
3. $\mathrm{MSE}_{\mathrm{C18}} + S$
4. $\mathrm{MSE}_{\mathrm{C18}} + 10S$

where $i = 1,2,\ldots,6$ corresponds to C18, C24, C30, 2C17, 4C17, and 22C16, respectively.

The resulting models were then evaluated for their predictions of the six Henry's constants (Tables 10), in addition to their ability to capture selectivity trends (Table 11). Options 1, 3, and 4 for the loss function all yielded models that showed similar performance as models trained separately for each molecule from scratch (Table 8). The biggest drop in performance was found for C30 with Option 3, with the r^2 decreasing by 0.044 and the MSE increasing by 8.6, but most other molecules display r^2 and MSE values that are within 0.02 and 3 of the individually trained models. By contrast, Option 2 failed terribly in predicting Henry's constants, due to the complete absence of $\ln k_{\mathrm{H},i}$ in its loss function design. Finally, as shown in Table 11, multi-task models 1, 3, and 4 exhibited improved accuracy in predicting the selectivity, reducing $|\Delta S|$ by 5 to 53% relative to individually trained models (Options 5 or 6), with the largest gains observed for the mono-branched C18 isomers. Interestingly, even Option 1 showed an overall improvement of 18%, despite not including selectivities in its loss function. This observation reinforces the premise that a shared representation can promote model generalization across different tasks.

4 Conclusions

A collection of machine-learning representations widely used for materials modeling was optimized and evaluated for predicting the logarithm of Henry's constants, $\ln k_{\mathrm{H}}$, for n-octadecane adsorption in all-silica zeolites, a task that stresses the ability of the machine-learning models to capture subtle structural variations and long-ranged spatial correlations. ZEONET, which comprises 3D ConvNets and volumetric grids, was extended to work on nearest-atom grids and atom-occupancy grids. Atom-occupancy grids reached similar performance to distance grids used in the original ZEONET,[18] while nearest-atom grids were slightly worse. Combining either grid with the distance grid further improved the performance, achieving $r^2 = 0.976$ and MSE $= 3.8$ (Table 3). These alternative volumetric representations allow for the extension of ZEONET to encode systems with more diverse elemental compositions. Work is underway to examine their performance

Table 10 Testing-set performance of multi-task ResNet18 models trained with different loss functions, showing r^2 and MSE values of the model prediction for individual molecules.

Sorbate	(1) MSE		(2) Selectivity		(3) $\mathrm{MSE}_{\mathrm{C18}} + S$		(4) $\mathrm{MSE}_{\mathrm{C18}} + 10S$	
	r^2	MSE	r^2	MSE	r^2	MSE	r^2	MSE
C18	0.944	1.7	−1.920	90.7	0.955	1.4	0.940	1.9
C24	0.899	7.2	−0.307	93.4	0.933	4.8	0.923	5.5
C30	0.817	35.3	0.367	122.1	0.758	46.7	0.787	41.1
2C17	0.940	2.2	−1.520	91.6	0.948	1.9	0.930	2.6
4C17	0.934	2.9	−1.129	93.8	0.941	2.6	0.934	2.9
22C16	0.939	5.0	−0.185	96.6	0.937	5.1	0.936	5.3
Composite r^2	0.885		−0.241		0.868		0.875	
Average MSE		9.1		98.0		10.4		9.9

Table 11 Mean-absolute error of selectivity on the testing set, calculated using multi-task ResNet18 models trained with different loss functions (Options 1-4), using six ResNet18 models trained separately for each molecule from scratch (Table 8; Option 5), and using six fine-tuned ResNet18 models (Pt. C18 of Table 9; Option 6).

Sorbate	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
C24	1.23	1.09	1.01	0.99	1.47	1.46
C30	3.50	3.12	3.64	3.33	3.84	4.30
2C17	0.63	0.55	0.58	0.58	1.24	1.02
4C17	0.81	0.72	0.75	0.72	1.14	0.98
22C16	1.25	1.01	1.14	1.06	1.32	1.25
Average	1.48	1.30	1.42	1.34	1.80	1.80

on encoding MOF materials. It is also worth further investigation to extract from the learned 3D ConvNets what structural features drive the adsorption behavior of different sorbate molecules. The reader is referred to initial attempts at feature attribution in our previous work.[18]

Using pre-trained ImageNet1K, several strategies for constructing multi-view representations from slices of the distance grid were examined. The performance of these image-based ConvNets increases steadily with the number of views, reaching a maximum with six or more views. The best performance was found for distance images where views are constructed from equidistant slices, with $r^2 = 0.912$ and MSE = 14.7. Vision Transformer, ViT, was found to be sensitive to hyperparameters and challenging to optimize, which achieved $r^2 = 0.869$ and MSE = 21.8 when training from scratch. Supplying pretrained 3D ResNet18 features resulted in performance on par with 3D ResNet18 itself. Without a much larger model or dataset, ViT does not seem to offer compelling advantages over the more robust 3D ConvNets.

Both point-cloud representations were found unsuitable for this task. The best model, which is EdgeConv operating on the coordinates of 2048 atoms, achieved only an r^2 of 0.610 and MSE of 65.1. Using points sampled from the molecular surface resulted in even poorer performance and was hampered by significant sampling cost. Finally, popular graph-based neural networks, including state-of-the-art models used for machine-learning intermolecular potentials (MLIPs), were not able to achieve the same level of accuracy as 3D ConvNets. In particular, the default settings of CGCNN, MEGNet, M3GNet, and MACE all perform terribly, with $r^2 < 0.55$ and MSE > 75. Their performance increases substantially with the size of the local environment used for calculating node features, as controlled by r_{cut} (and additionally N_{neighbor} in the case of CGCNN). The highest performance, achieved by M3GNet with $r^2 = 0.888$ and MSE = 18.5, was still significantly lower than 3D ConvNets, highlighting the difficulty of graph-based models in describing the detailed 3-dimensional arrangement of atoms and the resulting non-covalent interactions. It remains to be seen how well MC simulations using MLIPs can predict such adsorption properties.

The transfer learning results demonstrated the ability of ZEONET to generalize to other hydrocarbon adsorption tasks, including mono- and di-branched C18 isomers and linear C24 and C30 alkanes. Using only 2100 training samples, ZEONET can achieve comparable performance as training from scratch using over 10 000 samples. Among the different adsorption tasks, C30 presents the greatest difficulty, with the highest r^2 achieved being only 0.802, regardless of whether models were trained from scratch or via fine-tuning a pretrained model. Finally, multi-task learning, which trains a single representation with multiple prediction heads, was shown to improve the accuracy in predicting adsorption selectivity without compromising the prediction of Henry's constants for individual molecules. While this study focuses on the adsorption of flexible, long-chain molecules in zeolites, the observations may offer general insights into the behavior of various material representations for systems dominated by individually weak, collectively important non-covalent interactions that give rise to the rich variety of shape selectivity and confinement effects.

Acknowledgements

This project was supported by the Defense Advanced Research Projects Agency (DARPA) under Grant No. D24AP00322-00. Funding for the GPU cluster used in this work was provided by the National Science Foundation under Grant No. CNS-1919334. The research also used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC Award No. ER-CAP0028612 and of the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) through Allocation No. CTS190069.

Conflicts of interest

There are no conflicts to declare.

Data availability statement

The data supporting the findings of this work are available in the supplementary information (SI). The code and datasets used in this work can be found at https://r.bai.group/ZeoNet. Supplementary information: details of model architecture; additional tables of numerical data for model performance; additional figures of training convergence. See DOI: https://doi.org/10.1039/d5ta04285b.

Notes and references

1 T. Maesen, *Studies in Surface Science and Catalysis*, 2007, **168**, 1–12.

2 P. Del Campo, C. Martínez and A. Corma, *Chemical Society Reviews*, 2021, **50**, 8511–8595.

3 T. Xie and J. C. Grossman, *Physical Review Letters*, 2018, **120**, 145301.

4 K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, Advances in Neural Information Processing Systems, 2017, pp. 992–1002.

5 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.

6 J. Gasteiger, J. Groß and S. Günnemann, International Conference on Learning Representations (ICLR), 2020.

7 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chemistry of Materials*, 2019, **31**, 3564–3572.

8 K. Choudhary and B. DeCost, *npj Computational Materials*, 2021, **7**, 185.

9 K. Yan, Y. Liu, Y. Lin and S. Ji, Advances in Neural Information Processing Systems, 2022, pp. 15066–15080.

10 C. Chen and S. P. Ong, *Nature Computational Science*, 2022, **2**, 718–728.

11 S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu and J. Hu, *Physical Chemistry Chemical Physics*, 2020, **22**, 18141–18148.

12 I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner and G. Csanyi, Advances in Neural Information Processing Systems, 2022.

13 P. Loche, K. K. Huguenin-Dumittan, M. Honarmand, Q. Xu, E. Rumiantsev, W. B. How, M. F. Langer and M. Ceriotti, *Fast and flexible range-separated models for atomistic machine learning*, 2024, arXiv preprint arXiv:2412.03281.

14 S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko and K.-R. Müller, *Science Advances*, 2023, **9**, eadf0873.

15 L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car *et al.*, *The Journal of Chemical Physics*, 2022, **156**, 124107.

16 E. H. Cho and L.-C. Lin, *The Journal of Physical Chemistry Letters*, 2021, **12**, 2279–2285.

17 T.-H. Hung, Z.-X. Xu, D.-Y. Kang and L.-C. Lin, *The Journal of Physical Chemistry C*, 2022, **126**, 2813–2822.

18 Y. Liu, G. Perez, Z. Cheng, A. Sun, S. C. Hoover, W. Fan, S. Maji and P. Bai, *Journal of Materials Chemistry A*, 2023, **11**, 17570–17580.

19 H. Kang, L. Xia, F. Yan, Z. Wan, F. Shi, H. Yuan, H. Jiang, D. Wu, H. Sui, C. Zhang *et al.*, *IEEE transactions on medical imaging*, 2020, **39**, 2606–2614.

20 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Advances in Neural Information Processing Systems*, 2017, **30**, 5998–6008.

21 A. Dosovitskiy, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2020, arXiv preprint arXiv:2010.11929.

22 Y. Kang, H. Park, B. Smit and J. Kim, *Nature Machine Intelligence*, 2023, **5**, 309–318.

23 C. R. Qi, H. Su, K. Mo and L. J. Guibas, *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*, 2017, arXiv:1612.00593.

24 R. S. DeFever, C. Targonski, S. W. Hall, M. C. Smith and S. Sarupria, *Chemical Science*, 2019, **10**, 7503–7515.

25 L. Wang, L. Zhao, X. Liu, J. Fu and A. Zhang, *Environmental Science & Technology*, 2021, **55**, 9958–9967.

26 Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein and J. M. Solomon, *Dynamic Graph CNN for Learning on Point Clouds*, 2019, arXiv:1801.07829.

27 P. Bai, M. Y. Jeon, L. Ren, C. Knight, M. W. Deem, M. Tsapatsis and J. I. Siepmann, *Nature Communications*, 2015, **6**, 1–9.

28 P. Bai, M. Tsapatsis and J. I. Siepmann, *The Journal of Physical Chemistry C*, 2013, **117**, 24375–24387.

29 C. Baerlocher and L. McCusker, *Database of Zeolite Structures*, 2022, `http://www.iza-structure.org/databases/`.

30 R. Pophale, P. A. Cheeseman and M. W. Deem, *Physical Chemistry Chemical Physics*, 2011, **13**, 12407–12412.

31 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous and Mesoporous Materials*, 2012, **149**, 134–141.

32 PyTorch, *Pre-trained ResNet18 on ImageNet1K*, 2022, `https://pytorch.org/vision/stable/models.html`.

33 G. Pérez and S. Maji, 2022 26th International Conference on Pattern Recognition (ICPR), 2022, pp. 3048–3055.

34 W. E. Lorensen and H. E. Cline, *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

35 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. Chevriver, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314–319.

36 M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li and Z. Zhang, *Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks*, 2019, arXiv preprint arXiv:1909.01315.

37 R. Drautz, *Physical Review B*, 2019, **99**, 014104.

38 D. P. Kovács, I. Batatia, E. S. Arany and G. Csányi, *The Journal of Chemical Physics*, 2023, **159**, 044118.

39 A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.

40 A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi and F. Yu, *Shapenet: An information-rich 3d model repository*, 2015, arXiv preprint arXiv:1512.03012.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MLZeolitesModelComparisonrevSI.pdf