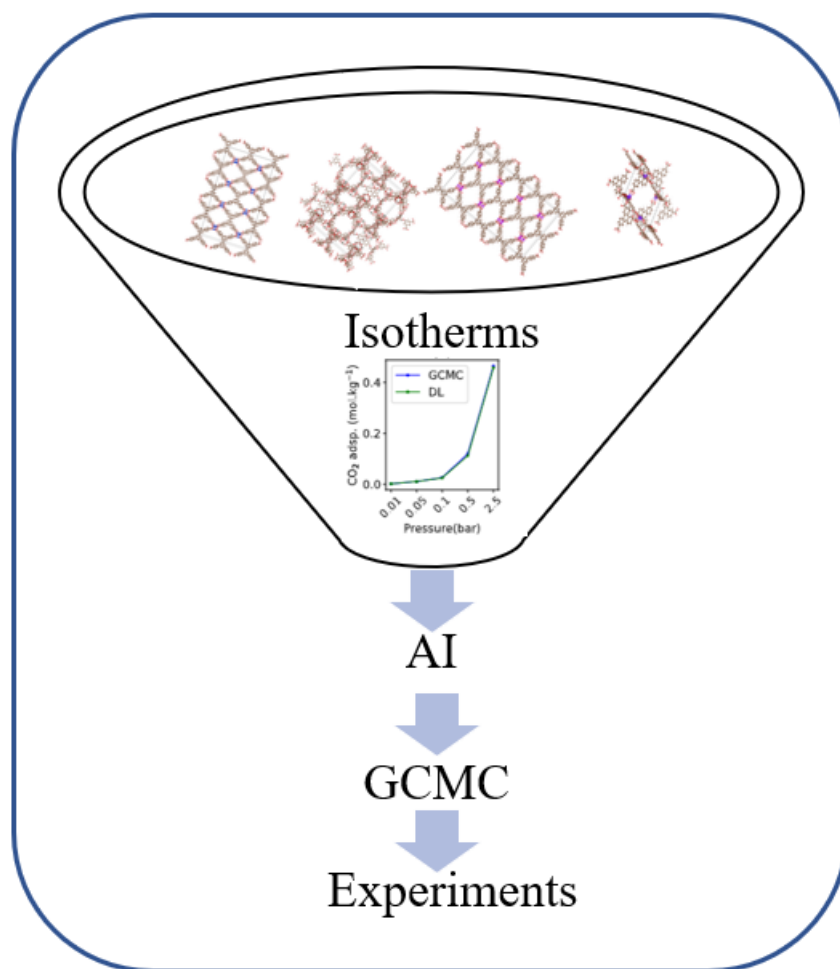


Graphical Abstract

Graph Neural Network Predictions of Metal Organic Framework CO₂ Adsorption Properties

Kamal Choudhary, Taner Yildirim, Daniel W. Siderius, A. Gilad Kusne,
Austin McDannald, Diana L. Ortiz-Montalvo



Highlights

Graph Neural Network Predictions of Metal Organic Framework CO₂ Adsorption Properties

Kamal Choudhary, Taner Yildirim, Daniel W. Siderius, A. Gilad Kusne, Austin McDannald, Diana L. Ortiz-Montalvo

- Accelerated search of high CO₂ adsorption metal organic frameworks
- Orders of magnitude faster predictions with AI techniques
- High-performance MOFs predicted with AI and validated with GCMC simulations

Graph Neural Network Predictions of Metal Organic Framework CO₂ Adsorption Properties

Kamal Choudhary^{a,b}, Taner Yildirim^c, Daniel W. Siderius^c, A. Gilad Kusne^c, Austin McDannald^d, Diana L. Ortiz-Montalvo^e

^a*Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, 20899, MD, USA.,*

^b*Theiss Research, La Jolla, 92037, CA, USA.,*

^c*NIST Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, 20899, MD, USA.,*

^d*Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, 20899, MD, USA.,*

^e*Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, 20899, MD, USA.,*

Abstract

The increasing CO₂ level is a critical concern and suitable materials are needed to capture such gases from the environment. While experimental and conventional computational methods are useful in finding such materials, they are usually slow and there is a need to expedite such processes. We use Atomistic Line Graph Neural Network (ALIGNN) method to predict CO₂ adsorption in metal organic frameworks (MOF), which are known for their high functional tunability. We train ALIGNN models for hypothetical MOF (hMOF) database with 137953 MOFs with grand canonical Monte Carlo (GCMC) based CO₂ adsorption isotherms. We develop high accuracy and fast models for pre-screening applications. We apply the trained model on CoREMOF database and computationally rank them for experimental synthesis. In addition to the CO₂ adsorption isotherm, we also train models for electronic bandgaps, surface area, void fraction, lowest cavity diameter, and pore limiting diameter, and illustrate the strength and limitation of such graph neural network models. For a few candidate MOFs we carry out GCMC calculations to evaluate the deep-learning (DL) predictions.

Keywords: Deep learning, metal organic framework, carbon-capture

1. Introduction

The CO₂ concentration in air is increasing at an alarming rate which could escalate global temperature by 1.5 °C by 2100 [1]. CO₂ capture from pre-combustion plants is an effective way to reduce CO₂ level. Typical materials used in this process are zeolites, calcium chabazite, hydrotalcite, and activated carbon [2], but they usually suffer from low CO₂ working capacities, low tunability, and limited material diversity [3].

Metal organic frameworks (MOFs) are a class of porous materials with inorganic clusters and organic building blocks, which can be repeated to build an extended 3D structure. MOFs provide high thermal stabilities (as high as 500 °C), a wide range of porosities (0.3–0.9), large surface areas (>8000 m²g⁻¹), low densities (0.2 g cm⁻³), a wide range of pore sizes (3–100 Å), active sites that can be used for site-specific adsorption and catalysis, and most importantly a wide range of chemical tunability unlike conventional porous materials [4]. Hence, MOFs are considered as one of the top ten materials for emerging technologies by the International Union of Pure and Applied Chemistry (IUPAC) [5].

In the past, there have been several efforts to use conventional computational methods such as grand canonical Monte Carlo (GCMC) to develop CO₂ isotherm databases for thousands of hypothetical as well as experimentally realizable MOFs [6, 7, 8, 9, 10, 11, 12, 13, 14]. However, such computational methods are still time consuming, especially to screen thousands to millions of MOF candidates. One of the most obvious ways to tackle this issue is using machine learning (ML) methods, where ML methods act as a pre-screening tool for conventional computational methods, which then can be used as a screening tool for experimental designs [15, 16, 17, 18, 19, 20, 21, 20].

The key precursors for applying ML methods for CO₂ absorption in MOF are : 1) large datasets of experimental or computational CO₂ isotherms and other MOF properties, 2) an ML framework that can accurately represent arbitrary MOF crystals and chemical structures. Fortunately, due to materials genome initiative (MGI) [22] and similar initiatives there has been a rapid development for generating large databases of MOFs such as: Computation-Ready, Experimental (CoREMOF), hypothetical (hMOF), and quantum (QMOF) [7, 8, 9, 10, 11, 12, 13, 14]. These databases are based on computational methods such as density functional theory (DFT) or GCMC and the results are compared to experiments wherever applicable [23]. They contain exact coordinates for MOF atomic structures and their properties.

Machine learning has been already used in the past to model atomistic properties of MOFs. For instance, Abdi et al.[24] used boosted decision-tree methods to model CO₂ adsorption of MOFs using 1191 data points using literature review. Krishnapriyan et al.[25] used random forest method to model CO₂ and methane uptake including Henry’s coefficients using persistent homology and chemical word embeddings on several datasets. Dureckova et al. [26] used decision-tree methods to CO₂ working capacity and CO₂/H₂ selectivity of gas adsorption in 358400 MOFs. A comprehensive review of all ML literature is beyond the scope of present work and can be found elsewhere[4, 27, 28, 29, 19, 21, 30, 31, 32, 33, 34, 35]. We note that previous ML models are primarily based on cumbersome handcrafted features which can be superseded with deep learning based methods [36]. Also, the application of multi-output ML methods especially for pressure dependent CO₂ adsorption has not been reported earlier to the best of our knowledge. Graph neural network (GNN)/deep learning (DL) methods such as Atomistic Line Graph Neural Network (ALIGNN) [37] can be used to represent arbitrary chemistry and atomic structure with respect to interatomic bonds and angles. GNNs can utilize the full power of deep neural network framework and can be used to model multi-output data as well.

In this work, we apply the ALIGNN model for predicting properties in hMOF and QMOF databases which can be used for accelerated MOF design. We train models for technologically important quantities such as CO₂ adsorption isotherms at different pressures, electronic bandgap, lowest cavity diameter (LCD), pore limiting diameter (PLD), void fraction, volumetric, and gravimetric surface areas. After training the models, we apply them to more CoREMOF database to screen for potential MOF candidates for CO₂ capture. We compare the model predictions with a few experimental measurement results to check accuracy. Additionally, we identify some of the key strength and challenges of the DL framework.

2. Methods

This study uses the QMOF dataset [14] and the hMOF database [8] to train the DL model. The trained DL model is applied to the CoreMOF 2019 database [7] to computationally rank the MOFs based on max CO₂ adsorption.

The QMOF dataset consists of 20425 MOFs that Rosen *et al.* [14] generated using a subset of the CoRE-MOF and Cambridge Structural Database

(CSD) [38] and made them suitable for density functional theory (DFT) calculations. The DFT calculations [39] were carried out at PBE-D3(BJ) level with Vienna *ab initio* Simulation Package (VASP) [40, 41].

The hMOF database consists of 137953 MOF structures which were generated using 102 building blocks from Wilmer *et al.* [9]. Details about the generation of the hMOF database can be found in Wilmer *et al.* [10]. Briefly, the Grand canonical Monte-Carlo (GCMC) calculations were performed using the RASPA code [42]. GCMC requires reliable force-fields for simulation. Interaction energies between non-bonded atoms were computed through the Lennard-Jones (LJ) and Coulomb potentials. LJ parameters for the framework atoms were taken from the Universal Force Field (UFF) [43]. Partial charges and LJ parameters for CO₂ were taken from the TraPPE force field [44]. Each GCMC calculation used a 1000 cycle equilibration period followed by a 1000 cycle production run. A cycle consists of n Monte Carlo steps; where n is equal to the number of molecules (which fluctuates during a GCMC simulation). All simulations included random insertion, deletion, rotation, and translation moves of molecules with equal probabilities.

Unlike the previous database where the MOF structures have been generated using computer algorithms, Computation-Ready, Experimental Metal–Organic Framework (CoREMOF) database contains materials that are derived from synthesized materials following the synthesis and data-refining protocols to provide closer to reality structures. CoREMOF database contains more than 14000 MOF structures [7]. There is no direct overlap between hMOF and CoREMOF database.

Graph neural networks are trained with atomistic line graph neural network (ALIGNN) [37] which is available at <https://github.com/usnistgov/alignn>. ALIGNN has been used to train more than 50 properties of solids and molecules with high accuracy. A MOF structure is represented as a graph using atomic elements as nodes, and atomic bonds as edges. Each node in the atomistic graph is assigned 9 input node features based on its atomic species: electronegativity, group number, covalent radius, valence electrons, first ionization energy, electron affinity, block and atomic volume. The interatomic bond distances are used as edge features with radial basis function upto 8 Å cut-off. We use a periodic 12-nearest-neighbor graph construction. This atomistic graph is then used for constructing corresponding line graph using interatomic bond-distances as nodes and bond-angles as edge features. ALIGNN uses Edge-gated graph convolution for updating nodes as well as edge features. One ALIGNN layer composes an edge-gated graph convolution

on the bond graph with an edge-gated graph convolution on the line graph. The line graph convolution produces bond messages that are propagated to the atomistic graph, which further updates the bond features in combination with atom features. The ALIGNN model is implemented in PyTorch [45] and deep graph library (DGL) [46].

For regression targets we minimize the mean squared error (MSE) loss for 50 epochs using the AdamW optimizer with normalized weight decay of 10^{-5} and a batch size of 64. The learning rate is scheduled according to the one-cycle policy with a maximum learning rate of 0.001. We use 80 initial bond radial basis function (RBF) features, and 40 initial bond angle RBF features. The atom, bond, and bond angle feature embedding layers produce 64-dimensional inputs to the graph convolution layers. The main body of the network consists of 4 ALIGNN layers and 4 graph convolution (GCN) layers, each with hidden dimension 256 nodes. The final atom representations are reduced by atom-wise average pooling and mapped to regression or classification outputs by a single linear layer.

We used National Institute of Standards and Technology (NIST)’s Nisaba cluster to train all ALIGNN models [39]. Each model is trained on a single Tesla V100 SXM2 32 gigabyte Graphics processing unit (GPU), with 8 Intel Xeon E5-2698 v4 CPU cores for concurrently fetching and preprocessing batches of data during training. We use 80 %:10 %: 10 % splits. The 10 % test data is never used during training procedures. Except for the pressure dependent CO₂ adsorption data, which is a multi-output model, all other models are single output regression.

3. Results and discussion

We apply ALIGNN model to predict MOF properties in QMOF and hMOF databases. QMOF is developed using quantum density functional theory method while hMOF is based on classical force field guided GCMC calculations. The QMOF and hMOF materials are composed of some of the combination of the following 9 elements: C, H, O, N, Zn, Cu, Zr, F, Cl, and Br. These datasets are kept in the jarvis-tools package: <https://jarvis-tools.readthedocs.io/en/master/databases.html> [47] in a format to easily apply DL methods. In this section, we will first visualize the data distribution and then discuss the ALIGNN application on these datasets.

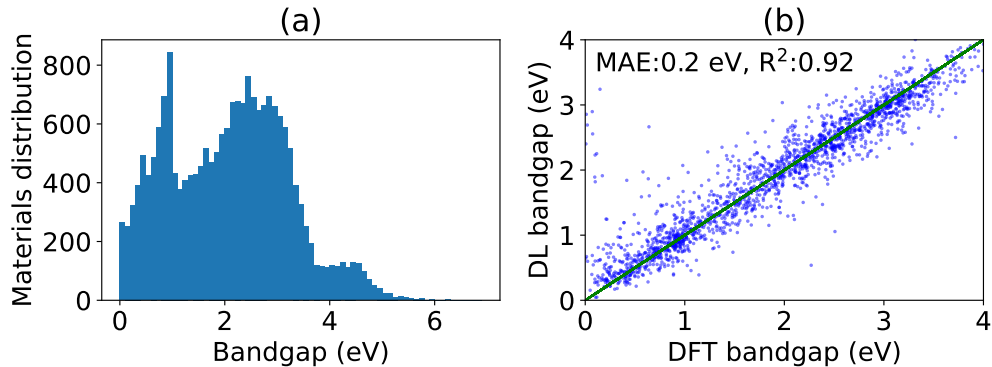


Figure 1: Data visualization of QMOF electronic bandgaps and GNN predictions on the test set. a) Electronic bandgap data distribution in QMOF bandgap distribution database, b) ALIGNN based deep-learning (DL) predictions on 10 % test set.

3.1. Application to QMOF database

Before applying the DL model on QMOF we visualize the properties in QMOF database with about 20425 MOFs. One of the unique features of QMOF is electronic bandgap data which cannot be obtained from classical force field calculations such as in hMOF. Although the focus of the work is on CO₂ adsorption, we note that MOFs can be viewed as photocatalysts and tailoring of bandgaps can provide dual technological benefits of CO₂ reduction as well as solar power generation [48, 49]. We visualize the bandgap data distribution in Fig. 1a. We observe a high peak around 1 eV and 3 eV with data range up to 6 eV.

The developers of QMOF applied several machine-learning models for predicting the bandgaps. They compared well-known machine learning model such as Sine Coulomb matrix, Orbital field matrix [50, 51], Smooth overlap of atomic orbitals (SOAP) [52], and crystal graph convolution neural network (CGCNN) [53] to develop a model for predicting the bandgap of MOFs. They found that the best model was achieved with CGCNN with a mean absolute error (MAE) of 0.274 eV for bandgap prediction. Following a similar training strategy, we trained ALIGNN model on this dataset and find the MAE to be 0.20 eV, which is nearly 27 % improvement in accuracy. A parity plot for the actual DFT data and ALIGNN based predictions on 10 % test set is shown in Fig. 1b. For a perfect agreement, all the data-points would fall upon $y = x$ line, which the data points in Fig. 1b closely resemble. ALIGNN has already

been shown to outperform CGCNN and other models for other solid-state and molecular dataset and the previous results support the findings [37].

3.2. Application to hMOF database

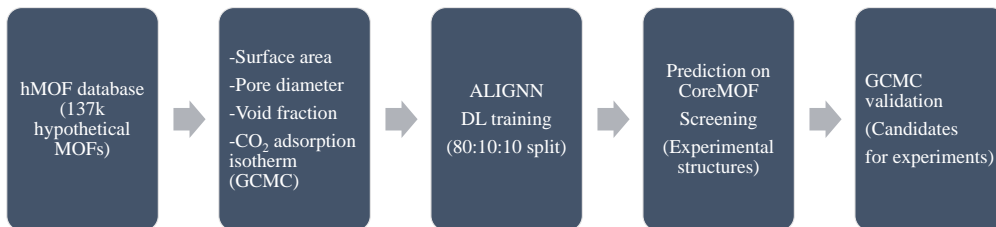


Figure 2: Workflow used in training and applying graph neural network models for several properties in hMOF database.

Next we apply the ALIGNN model for training properties in the hMOF databases. The list of properties include: gravimetric surface area (m^2g^{-1}), volumetric surface area (m^2cm^{-3}), largest cavity diameter (LCD), pore limiting diameter (PLD), void fraction, and pressure dependent CO_2 absorption. All the models except the CO_2 isotherm are single output models. The adsorption isotherm predictions has 5 outputs representing CO_2 absorption at 5 different pressures 0.01, 0.05, 0.1, 0.5, and 2.5 bar at 298 K. In addition to the multi-output model, we also train models for minimum (0.01 bar) and maximum (2.5 bar) pressures (under consideration) to check whether having more data as targets (five versus one) helps improve the model predictions.

In Fig. 2 we show the training process for hMOF which is described below. Before applying any DL, we first visualize the distribution of the above mentioned properties for 137953 MOFs in Fig. 3. As observed in Fig. 3a, MOFs in hMOF have low bulk densities with a peak around 0.7 g

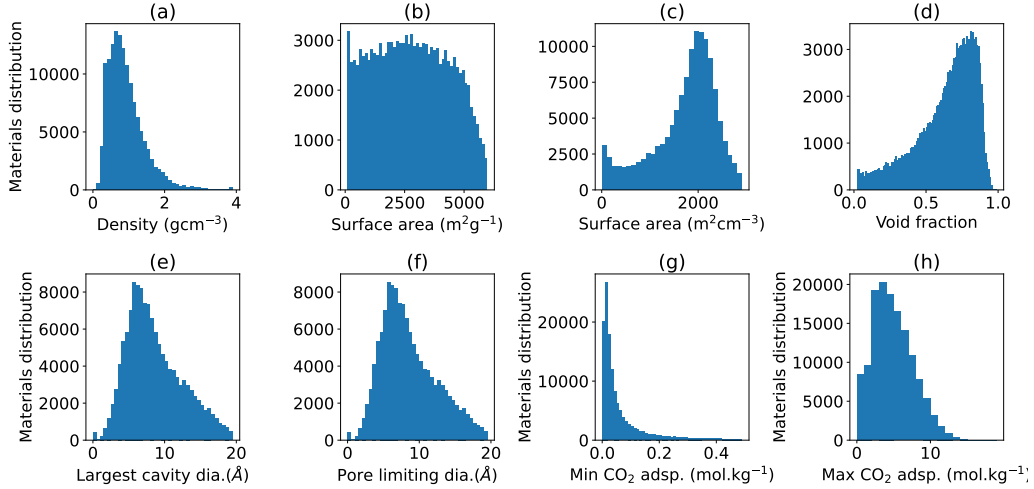


Figure 3: Data distribution visualization of several properties in the hMOF database. a) density, b) gravimetric surface area, c) volumetric surface area, d) void fraction, e) LCD, f) PLD, g) minimum value of CO₂ dsorption among 5 pressure values (i.e. adsorption at 0.0 1bar), and f) maximum value of CO₂ adsorption (i.e. adsorption at 2.5 bar) among 5 pressure values.

cm⁻³, which is much lower than usual solids such as silicon (2.3 g cm⁻³) and aluminum (2.7 g cm⁻³). The gravimetric and volumetric surface areas in Fig. 3b and Fig. 3c varies up to 6000 m² g⁻¹ and 3000 m² cm⁻³, respectively. The gravimetric surface area is almost uniformly distributed across the range while the volumetric surface area peaks around 2000 m² cm⁻³. We note that high volumetric surface area doesn't necessarily imply high gravimetric surface area and vice versa. The void fraction in Fig. 3d shows a peak around 0.8 suggesting high porosity of MOFs. The largest cavity diameters (LCD) and pore limiting diameters (PLD) are other critical parameters for adsorption. Both of them vary up to 20 Å with a peak around 7 Å (as shown in Fig. 3e and Fig. 3f). As mentioned above, hMOF contains CO₂ adsorption at 5 different and monotonically increasing pressure values, so we show the CO₂ adsorption at 0.01 bar and 2.5 bar pressures in Fig. 3g and Fig. 3h, respectively.

We note that some of these quantities (*e.g.* PLD, LCD, and void fraction) can be easily calculated for the whole dataset of potential MOFs using conventional computational methods such as Zeo++ [54]. Using DL based methods to predict these properties will elucidate the strengths and

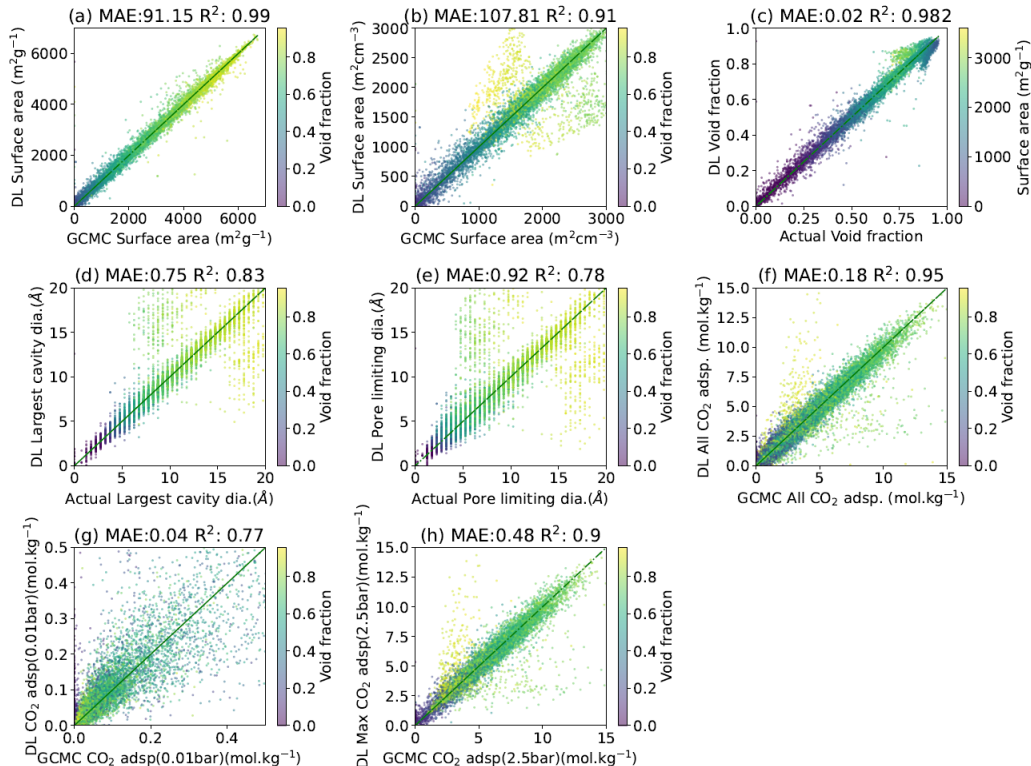


Figure 4: ALIGNN predictions and GCMC actual value comparisons of several properties on 10 % held test set (13796 MOFs) in the hMOF dataset. a) gravimetric surface area, b) volumetric surface area, c) void fraction, d) largest cavity diameter, e) pore limiting diameter, f) adsorption at five different pressures, g) adsorption at 0.01 bar, and h) adsorption at 2.5 bar.

weaknesses of the DL models. However, predicting adsorption isotherms can be prohibitively expensive for the whole dataset by traditional means (*e.g.* GCMC). Therefore the DL methods can be useful as a surrogate model, with the usefulness characterized by the appropriate performance metrics.

We trained the models for the above mentioned quantities for 50 epochs because further iterations do not improve the validation accuracy as shown in Fig. S1 for 100 epochs. Additionally, we have tried different cutoff and max-neighbors parameters such as: 1) cutoff 8 Å, 12 k-nearest neighbors (kNN), 2) cutoff 20 Å, 12 kNN, 3) cutoff 8 Å and kNN 24 with no significant change in MAE. The computational cost analysis for these trainings are also provided in the supplementary information (Table S1).

The parity plots for predictions on the test dataset by the ALIGNN models for different properties is shown in Fig 4. For an ideal model the DL and actual target values should line on $y = x$ straight line. Corresponding performance metric for the trained models are charted in Table. 1. In addition to the performance metric such as the mean absolute error (MAE), we also provide the mean absolute deviation (MAD) which serves as the baseline for the model predictions. Usually, a good machine learning model has MAE five times lower than its MAD. The MAD:MAE ratio also serves as a good comparison metric across different properties. From Table 1, we observe that most of our models have MAD:MAE greater than 5, with highest value for gravimetric surface area and the minimum value for the pore limiting diameter. Same behavior can be observed from the parity plots.

Comparing Fig. 4a and Fig. 4b, we observe that gravimetric surface area is easier to predict than the volumetric surface area. Currently we don't have a good explanation for this behavior. The void fraction is easier to predict as well as shown in Fig. 4c. The LCD and PLD parity plots in Fig, 4d and Fig. 4e suggest they are much difficult to predict. This shows one of the limitations of the ALIGNN model where it cannot capture long range order based properties. In Fig. 4f, we show the DL predicted data against the actual GCMC data for all the pressures in one plot. Similar plots for the minimum and maximum pressures are shown in Fig. 4 g and 4h. Interestingly, the full-isotherm data based model has lower overall MAE as well as lower MAD:MAE in comparison with the maximum pressure adsorption model. This result suggests that training multi-output model could be advantageous for adsorption data especially at higher pressure values. The colorbar in the plots represent void fraction with the exception of Fig. 4c (surface area is represented). We observe that most of the deviations from $y = x$ line in the test set occur on high void fraction MOFs (such as upper yellow part of Fig. 4b, 4d, 4e and 4h). This can be due to the k-nearest neighbor and relatively shorter cut-off (8 Å) strategies for graph construction in ALIGNN models. Larger k-nearest neighbors and cut-off impose higher computational expense, a challenge for ALIGNN.

In Table 1, the first 5 properties are geometric while the rest are adsorption properties. It is interesting to note that while it is difficult for ALIGNN to predict the geometric properties to high accuracy, ALIGNN is able to predict the adsorption properties accurately. The geometric properties are features of the long range crystallographic order, while the graph used in the ALIGNN model considers neighbors only up to 8 Å away. It is therefore reasonable

Table 1: ALIGNN performance on hMOF properties on 10 % held test dataset in terms of mean absolute error (MAE), mean absolute deviation (MAD), coefficient of determination (R^2), and root mean-square error (RMSE).

Property	Unit	MAD	MAE	MAD:MAE	R^2	RMSE
Grav. surface area	$\text{m}^2 \text{ g}^{-1}$	1430.82	91.15	15.70	0.99	180.89
Vol. surface area	$\text{m}^2 \text{ cm}^{-3}$	561.44	107.81	5.21	0.91	229.24
Void fraction	No unit	0.16	0.017	9.41	0.98	0.03
LCD	\AA	3.44	0.75	4.56	0.83	1.83
PLD	\AA	3.55	0.92	3.86	0.78	2.12
All adsp	mol kg^{-1}	1.70	0.18	9.44	0.95	0.49
Adsp at 0.01bar	mol kg^{-1}	0.12	0.04	3.00	0.77	0.11
Adsp at 2.5bar	mol kg^{-1}	2.16	0.48	4.50	0.90	0.97

that ALIGNN, designed to capture short-range atomistic interactions, has difficulty predicting the long-range geometric features. Despite this, ALIGNN is able to predict the adsorption properties well, which leads us to conclude that the adsorption properties are largely determined by the short-range interactions. Conventional works with MOFs have focused on optimizing the long-range geometric features in attempts to optimize the adsorption properties [55, 56, 57]. This work shows that the short-range atomistic interactions are the primary factors in the adsorption properties. Note that previous works on predicting CO_2 adsorption with descriptor based ML models only used a certain portion of hMOF dataset or were trained on entirely different datasets, hence, it is not fair to compare those performances with the current work. However, we note that CO_2 adsorptions models had R^2 ranging from 0.45 to 0.94 [24, 25, 26].

In addition to the parity plots, we also analyze the best and worst isotherm predictions using the multi-output isotherm predictor model. We sort the test data based predictions with respect to MAE and show top 5 results in Fig. 5a to Fig. 5e and worst values from Fig. 5f to Fig. 5j. Clearly, the best predictions are on par with the target data. For the worst predictions, the multi-output predictions are higher in some cases and lower in others compared to actual GCMC data implying non-uniform trends.

To investigate the reason why the DL model works well for a certain class of materials, we carried out several correlation analyses between the errors of multi-output isotherm predictor model and materials properties. We calculate the mean of absolute errors between actual and DL predictor

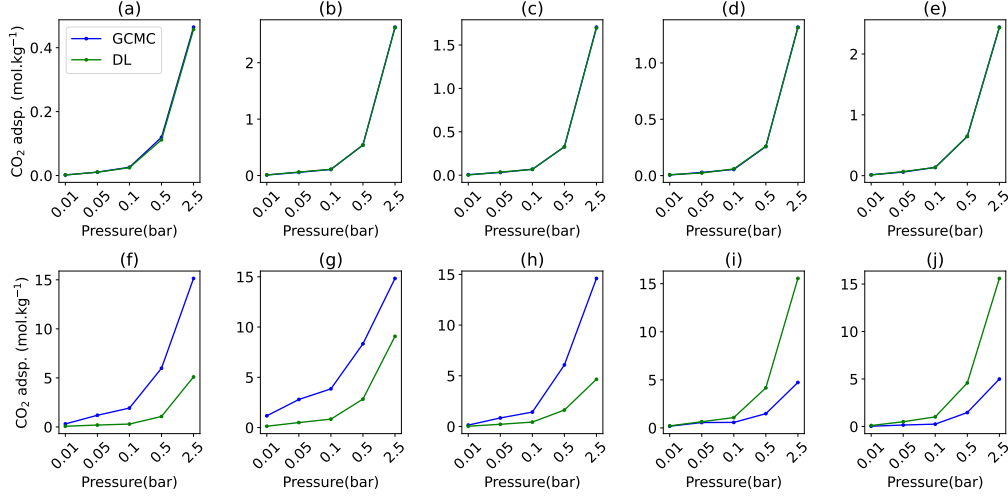


Figure 5: Comparison of actual GCMC and ALIGNN predictions of CO₂ adsorption at different pressures values and at 298 K. The model was tested on the 10 % held set of 13796 materials, out of which 5 best and 5 worst predictions are shown. a-e) 5 best predictions, f-j) 5 worst predictions.

values. From Table 2, we find the minimum and maximum CO₂ intake and density of the MOFs relate the most to the errors in the DL predictions. Note that there are relatively few examples of low-density materials, hence DL is less likely to generalize well for these set of materials.

Table 2: Pearson and Spearman correlation between mean of mean absolute errors of the CO₂ absorptions on test set and physical properties.

Property	Unit	Pearson	Spearman
Density	g m^{-3}	0.04	0.22
Grav. surface area	$\text{m}^2 \text{g}^{-1}$	-0.07	-0.10
Vol. surface area	$\text{m}^2 \text{cm}^{-3}$	-0.04	-0.10
Void fraction	No unit	-0.02	-0.2
N atoms	No unit	0.1	0.2
PLD	\AA	-0.07	-0.23
LCD	\AA	-0.05	-0.21
Min adsp	mol kg^{-1}	0.33	0.56
Max adsp	mol kg^{-1}	0.27	0.40

Next, we evaluate the accuracy of the DL model based predictions by

comparing them with a few experimental and GCMC calculation results. Over the years we have experimentally characterized a large number of MOFs using same measurement protocol and very accurate high-pressure sievert apparatus[58]. We compare the experimental CO₂ adsorption at 2.5 bar for some of these previously well characterized MOFs [59, 58, 60, 61] with DL predictions in Fig. 6. We observe excellent comparison between DL and experimental results. We find that the DL and experimental data are in reasonable agreement with mean absolute difference for CO₂ adsorption and surface area as 1.54 mol.kg⁻¹ and 399.75 m².g⁻¹ respectively.

Moreover, we apply the DL model for predicting maximum CO₂ adsorption trained on hMOF database to screen high CO₂ adsorbing MOFs in the CoREMOF database. Note that there is almost no overlap between the MOFs in these two databases. Only 16 MOFs were confirmed to be overlapping between these two databases as mentioned in the supplementary article information of ref.[7]. The hMOF database is based on hypothetical structures while CoreMOF uses experimental structures which have been processed following several protocols. DL prediction for a MOF takes less than a few seconds. We rank the MOFs and provide the DL based predictions for MOFs in the supplementary information. Some of the high rank materials predicted by our DL model are: ATEYUV (10.7 mol.kg⁻¹), WEHJUQ (10.2 mol.kg⁻¹), ATEYOP (9.94 mol.kg⁻¹), HOJLID (9.86 mol.kg⁻¹), and MINCUJ (9.81 mol.kg⁻¹), shown in Fig. 7 and Fig. 8. The crystal structure visualization for these MOFs as well as other screened materials are available at <https://mof.tech.northwestern.edu>

In order to check the DL predictions, we carried out accurate GCMC simulations and compare the results with DL predictions. In Fig. 8a, the comparison of GCMC vs DL predictions for CO₂ on top 10 screened candidates shows the DL method can indeed predict high CO₂ adsorption MOFs. The GCMC data are obtained with the similar computation set-up as of hMOF database but with larger cutoff (up to 18 Å) and atomic charges of MOFs. The partial atomic charges of the framework are obtained by message passing neural networks model developed in Ref.[62] and included in the CO₂ adsorption simulations. Although the GCMC values do not perfectly align with the DL data, the trends are noteworthy. We note that nine out of ten predicted structures give a CO₂ adsorption above 10 mol/kg at 2.5 bar, which is quite high compared to all other experimentally well known MOFs. Fig. 8a clearly suggests that DL can predict high performing MOFs which are also validated with the GCMC calculations. We find that the mean ab-

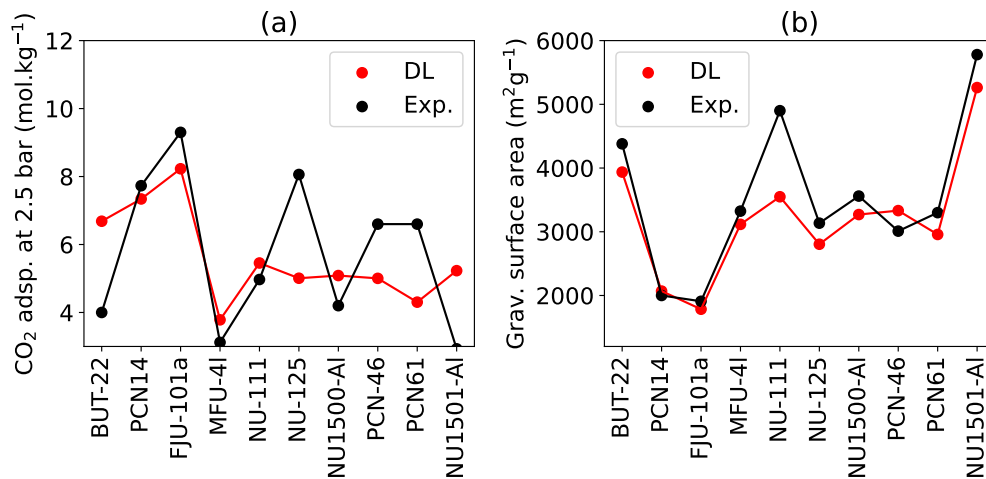


Figure 6: Comparison of experimental measurements and ALIGNN (DL) predictions for a few previously characterized MOFs from literature. a) CO₂ adsorption , b) gravimetric surface area.

solute difference between the GCMC and DL values for CO₂ adsorption at 2.5 bar for these candidates is 4.2 mol.kg⁻¹ which is almost 9 times higher than that of MAE reported in Table 1. Next we compare the DL and GCMC bases surface area predictions for these top 10 screened MOFs. We notice that the trends in the predictions (as shown in Fig. 8b) are much better than that of CO₂ adsorption. We find the mean absolute difference between these values as 697.2 m².g⁻¹ which is 7 times higher than MAE of the DL model obtained for gravimetric surface area in Table. 1.

We studied the effect of Cutoff (both van der Waals and Coulomb interaction) and the results are shown in Fig. S2 and Fig. S3. We noticed that for Coulomb interaction 8 Å is good enough for convergence. For vdW interaction, we observe a 5 % decrease in adsorption at high pressures. However, we note that taking a larger vdW cutoff does not affect the general trend in the adsorption data which is clear from Fig. S3. The agreement between DL and GCMC calculations with 8 Å cutoff are much better compared to 16 Å cutoff. This is probably due to the fact that we used 8 Å cutoff in DL model. Since cutoff does not change the general trend in the isotherm, the overall effect is just to shift the adsorption isotherm and therefore it won't affect the predictive feature of our DL model.

Experimental synthesis and characterizations of these is beyond the scope

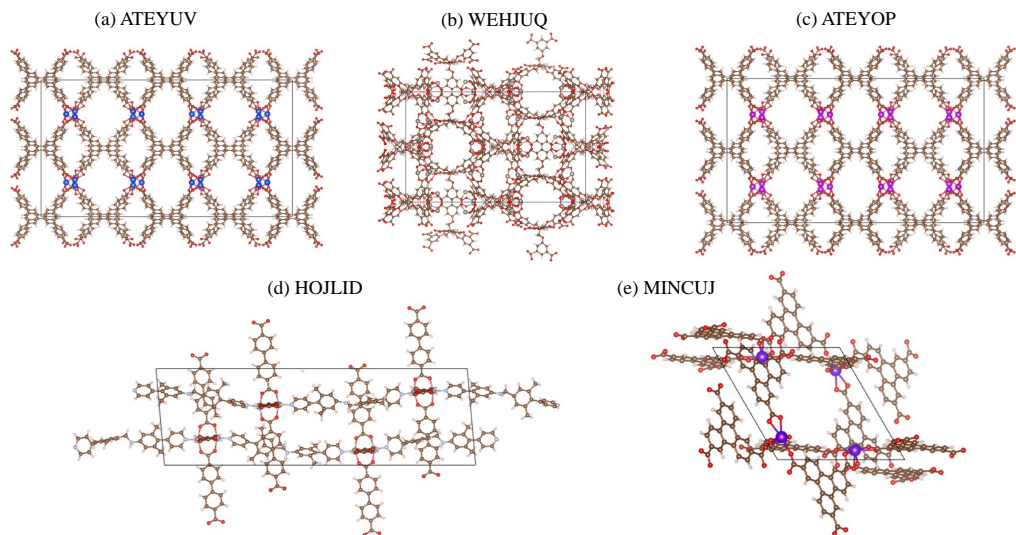


Figure 7: ALIGNN predicted some examples of the high- CO_2 adsorbing MOFs from CoREMOF DB. Complete list is provided in the supplementary information. a) ATEYUV, b) WEHJUQ, c) ATEYOP, d) HOJLID, and e) MINCUJ. The crystal structure visualization for these MOFs as well as other screened materials are available at <https://mof.tech.northwestern.edu>.

of the present work and will be carried out in future. The complete list of MOFs during the above screening procedure is provided in the supplementary information. The supplementary information include the name identifier of the MOF (id), DL based gravimetric surface area (dl_sag), gravimetric surface area reported in CoREMOF (actual_sag), density, DL based CO_2 isotherm predictions at the five pressure bar points, and DL based maximum CO_2 predictions at 2.5 bar. We note that some of the MOFs in the CoREMOF database have multiple disordered atoms, which are not physical. In the supplementary information, we designate such structures with yes or no (Y/N) and we recommend MOFs with no multiple atoms only. The trained ALIGNN (DL) models for the properties trained above are also distributed in the ALIGNN GitHub repo <https://github.com/usnistgov/alignn/blob/main/alignn/pretrained.py> as pre-trained models for making quick predictions on arbitrary MOFs.

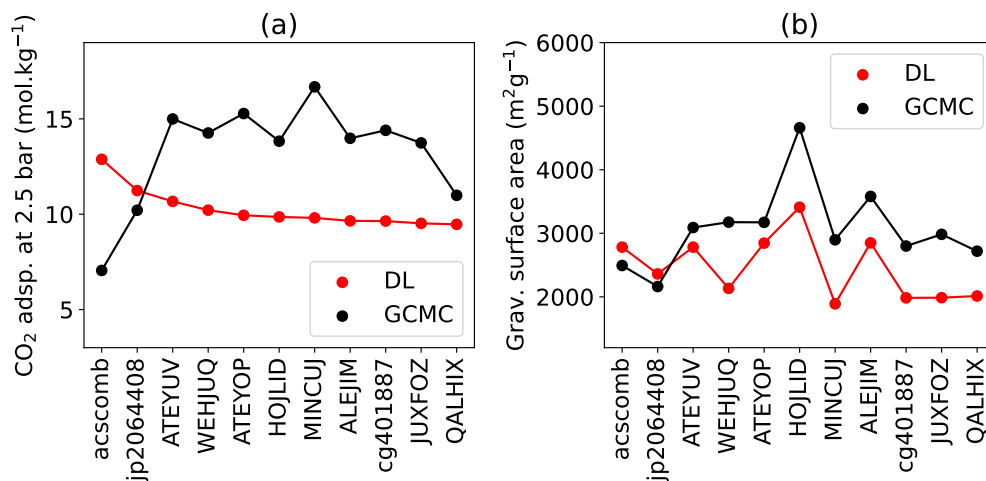


Figure 8: Comparison of GCMC and ALIGNN (DL) predictions for top 10 screened candidates from CoREMOF DB. a) CO₂ adsorption , b) gravimetric surface area.

4. Conclusion

In this article, we have applied atomistic line graph neural network (ALIGNN) to predict several properties such as volumetric and gravimetric surface area, void fraction, maximum and minimum CO₂ adsorption, adsorption isotherms at several pressures and electronic bandgaps in hMOF and QMOF datasets. High accuracy models are obtained for surface area and CO₂ adsorption that can be used for pre-screening applications. We have applied the trained model to pre-screening high-performance MOFs in the CoREMOF database and predicted several promising MOF for CO₂ capture. Moreover, we carry out GCMC calculations for a few screened candidates and compare them with DL predictions showing promising trends. This study demonstrates some of the key strength for using GNN over usual conventional methods. We also discuss some of the weaknesses of the model that motivate further development for the GNN model. Our work also establishes that GNN models can not only be successfully applied for molecules and solids but MOF structures as well.

5. Acknowledgements

K.C. thanks National Institute of Standards and Technology for computational and funding support.

6. Conflict of interest

The authors declare no conflict of interest.

7. Supporting Information

Effect of increasing number of epochs during model training, effect of different cutoffs on absorption behavior and computational expense are provided in the supplementary information. A list of DL based screened MOF candidates is provided in the supporting information.

8. Data availability

The MOF dataset trained in this work can be obtained from JARVIS-Tools database page (<https://jarvis-tools.readthedocs.io/en/master/databases.html>).

References

- [1] B. Kirtman, S. B. Power, A. J. Adedoyin, G. J. Boer, R. Bojariu, I. Camilloni, F. Doblas-Reyes, A. M. Fiore, M. Kimoto, G. Meehl, et al., Near-term climate change: projections and predictability, Cambridge University Press, 2013.
- [2] A. Lee, G. Xiao, P. Xiao, K. Joshi, R. Singh, P. A. Webley, High temperature adsorption materials and their performance for pre-combustion capture of carbon dioxide, *Energy Procedia* 4 (2011) 1199–1206.
- [3] K. Stocker, M. Ellersdorfer, M. Lehner, J. G. Raith, Characterization and utilization of natural zeolites in technical applications, *BHM Berg- und Hüttenmännische Monatshefte* 162 (4) (2017) 142–147.
- [4] C. Altintas, O. F. Altundal, S. Keskin, R. Yildirim, Machine learning meets with metal organic frameworks for gas storage and separation, *Journal of Chemical Information and Modeling* 61 (5) (2021) 2131–2146.
- [5] F. Gomollón-Bel, Ten chemical innovations that will change our world: Iupac identifies emerging technologies in chemistry with potential to make our planet more sustainable, *Chemistry International* 41 (2) (2019) 12–17.

- [6] C. Altintas, G. Avci, H. Daglar, A. Nemati Vesali Azar, S. Velioglu, I. Erucar, S. Keskin, Database for co₂ separation performances of mofs based on computational materials screening, *ACS applied materials & interfaces* 10 (20) (2018) 17257–17268.
- [7] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, et al., Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019, *Journal of Chemical & Engineering Data* 64 (12) (2019) 5985–5998.
- [8] N. S. Bobbitt, J. Chen, R. Q. Snurr, High-throughput screening of metal–organic frameworks for hydrogen storage at cryogenic temperature, *The Journal of Physical Chemistry C* 120 (48) (2016) 27328–27341.
- [9] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, *Nature chemistry* 4 (2) (2012) 83–89.
- [10] C. E. Wilmer, O. K. Farha, Y.-S. Bae, J. T. Hupp, R. Q. Snurr, Structure–property relationships of porous materials for carbon dioxide separation and capture, *Energy & Environmental Science* 5 (12) (2012) 9849–9856.
- [11] B. J. Sikora, C. E. Wilmer, M. L. Greenfield, R. Q. Snurr, Thermodynamic analysis of xe/kr selectivity in over 137000 hypothetical metal–organic frameworks, *Chemical Science* 3 (7) (2012) 2217–2223.
- [12] Y. J. Colón, D. A. Gomez-Gualdrón, R. Q. Snurr, Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications, *Crystal Growth & Design* 17 (11) (2017) 5801–5810.
- [13] D. A. Gómez-Gualdrón, Y. J. Colón, X. Zhang, T. C. Wang, Y.-S. Chen, J. T. Hupp, T. Yildirim, O. K. Farha, J. Zhang, R. Q. Snurr, Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage, *Energy & Environmental Science* 9 (10) (2016) 3279–3289.
- [14] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, Machine learning the quantum-chemical

properties of metal–organic frameworks for accelerated materials discovery, *Matter* 4 (5) (2021) 1578–1597.

- [15] R. K. Vasudevan, K. Choudhary, A. Mehta, R. Smith, G. Kusne, F. Tavazza, L. Vlcek, M. Ziatdinov, S. V. Kalinin, J. Hattrick-Simpers, Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics, *MRS communications* 9 (3) (2019) 821–838.
- [16] G. R. Schleder, A. C. Padilha, C. M. Acosta, M. Costa, A. Fazzio, From dft to machine learning: recent approaches to materials science—a review, *Journal of Physics: Materials* 2 (3) (2019) 032001.
- [17] J. Schmidt, M. R. Marques, S. Botti, M. A. Marques, Recent advances and applications of machine learning in solid-state materials science, *npj Computational Materials* 5 (1) (2019) 1–36.
- [18] H. Ohno, Y. Mukae, Machine learning approach for prediction and search: application to methane storage in a metal–organic framework, *The Journal of Physical Chemistry C* 120 (42) (2016) 23963–23968.
- [19] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, T. K. Woo, Rapid and accurate machine learning recognition of high performing metal organic frameworks for co₂ capture, *The journal of physical chemistry letters* 5 (17) (2014) 3056–3060.
- [20] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, Big-data science in porous materials: materials genomics and machine learning, *Chemical reviews* 120 (16) (2020) 8066–8129.
- [21] S. Chong, S. Lee, B. Kim, J. Kim, Applications of machine learning in metal-organic frameworks, *Coordination Chemistry Reviews* 423 (2020) 213487.
- [22] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, et al., New frontiers for the materials genome initiative, *npj Computational Materials* 5 (1) (2019) 1–23.

- [23] W. Sun, L.-C. Lin, X. Peng, B. Smit, Computational screening of porous metal-organic frameworks and zeolites for the removal of so₂ and no_x from flue gases, *AIChE Journal* 60 (6) (2014) 2314–2323.
- [24] J. Abdi, F. Hadavimoghaddam, M. Hadipoor, A. Hemmati-Sarapardeh, Modeling of co₂ adsorption capacity by porous metal organic frameworks using advanced decision tree-based models, *Scientific reports* 11 (1) (2021) 1–14.
- [25] A. S. Krishnapriyan, J. Montoya, M. Haranczyk, J. Hummelshøj, D. Morozov, Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks, *Scientific reports* 11 (1) (2021) 1–11.
- [26] H. Dureckova, M. Krykunov, M. Z. Aghaji, T. K. Woo, Robust machine learning models for predicting high co₂ working capacity and co₂/h₂ selectivity of gas adsorption in metal organic frameworks for precombustion carbon capture, *The Journal of Physical Chemistry C* 123 (7) (2019) 4133–4139.
- [27] R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. A. Gómez-Gualdrón, Role of pore chemistry and topology in the co₂ capture capabilities of mofs: from molecular simulation to machine learning, *Chemistry of Materials* 30 (18) (2018) 6325–6337.
- [28] A. Guda, S. Guda, A. Martini, A. Bugaev, M. Soldatov, A. Soldatov, C. Lamberti, Machine learning approaches to xanes spectra for quantitative 3d structural determination: The case of co₂ adsorption on cpo-27-ni mof, *Radiation Physics and Chemistry* 175 (2020) 108430.
- [29] J. Burner, L. Schwiedrzik, M. Krykunov, J. Luo, P. G. Boyd, T. K. Woo, High-performing deep learning regression models for predicting low-pressure co₂ adsorption properties of metal-organic frameworks, *The Journal of Physical Chemistry C* 124 (51) (2020) 27996–28005.
- [30] Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, et al., Inverse design of nanoporous crystalline reticular materials with deep generative models, *Nature Machine Intelligence* 3 (1) (2021) 76–86.

- [31] D. Tang, F. Gharagheizi, D. S. Sholl, Adsorption-based separation of near-azeotropic mixtures—a challenging example for high-throughput development of adsorbents, *The Journal of Physical Chemistry B* 125 (3) (2021) 926–936.
- [32] J. M. Findley, D. S. Sholl, Computational screening of mofs and zeolites for direct air capture of carbon dioxide under humid conditions, *The Journal of Physical Chemistry C* 125 (44) (2021) 24630–24639.
- [33] X. Zhang, T. Zhou, K. Sundmacher, Integrated metal–organic framework and pressure/vacuum swing adsorption process design: Descriptor optimization, *AIChE Journal* (2022) e17524.
- [34] T.-H. Hung, Z.-X. Xu, D.-Y. Kang, L.-C. Lin, Chemistry-encoded convolutional neural networks for predicting gaseous adsorption in porous materials, *The Journal of Physical Chemistry C* (2022).
- [35] G. S. Fanourgakis, K. Gkagkas, E. Tylianakis, G. E. Froudakis, A universal machine learning algorithm for large-scale screening of materials, *Journal of the American Chemical Society* 142 (8) (2020) 3814–3822.
- [36] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular graph convolutions: moving beyond fingerprints, *Journal of computer-aided molecular design* 30 (8) (2016) 595–608.
- [37] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, *npj Computational Materials* 7 (1) (2021) 1–8.
- [38] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, Development of a cambridge structural database subset: a collection of metal–organic frameworks for past, present, and future, *Chemistry of Materials* 29 (7) (2017) 2618–2625.
- [39] NIST disclaimer, Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. such identification is not intended to imply recommendation or endorsement by nist, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. (2021).

- [40] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Physical review B* 54 (16) (1996) 11169.
- [41] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Physical review b* 59 (3) (1999) 1758.
- [42] D. Dubbeldam, S. Calero, D. E. Ellis, R. Q. Snurr, Raspa: molecular simulation software for adsorption and diffusion in flexible nanoporous materials, *Molecular Simulation* 42 (2) (2016) 81–101.
- [43] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, W. M. Skiff, Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations, *Journal of the American chemical society* 114 (25) (1992) 10024–10035.
- [44] J. J. Potoff, J. I. Siepmann, Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen, *AIChE journal* 47 (7) (2001) 1676–1682.
- [45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, *Advances in neural information processing systems* 32 (2019) 8026–8037.
- [46] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, et al., Deep graph library: Towards efficient and scalable deep learning on graphs., *arXiv e-prints* (2019).
- [47] K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. H. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, et al., The joint automated repository for various integrated simulations (jarvis) for data-driven materials design, *npj Computational Materials* 6 (1) (2020) 1–13.
- [48] D. Y. Lee, D. V. Shinde, S. J. Yoon, K. N. Cho, W. Lee, N. K. Shrestha, S.-H. Han, Cu-based metal–organic frameworks for photovoltaic application, *The Journal of Physical Chemistry C* 118 (30) (2014) 16328–16334.

- [49] X. Guo, L. Liu, Y. Xiao, Y. Qi, C. Duan, F. Zhang, Band gap engineering of metal-organic frameworks for solar fuel productions, *Coordination Chemistry Reviews* 435 (2021) 213785.
- [50] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Physical review letters* 108 (5) (2012) 058301.
- [51] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, *Physical Review B* 87 (18) (2013) 184115.
- [52] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space, *Physical Chemistry Chemical Physics* 18 (20) (2016) 13754–13769.
- [53] T. Xie, J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Physical review letters* 120 (14) (2018) 145301.
- [54] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials, *Microporous and Mesoporous Materials* 149 (1) (2012) 134–141.
- [55] A. Asghar, N. Iqbal, T. Noor, Ultrasonication treatment enhances mof surface area and gas uptake capacity, *Polyhedron* 181 (2020) 114463.
- [56] R. L. Martin, M. Haranczyk, Exploring frontiers of high surface area metal-organic frameworks, *Chemical Science* 4 (4) (2013) 1781–1785.
- [57] H.-M. Wen, B. Li, L. Li, R.-B. Lin, W. Zhou, G. Qian, B. Chen, A metal-organic framework with optimized porosity and functional sites for high gravimetric and volumetric methane storage working capacities, *Advanced Materials* 30 (16) (2018) 1704792.
- [58] Y. Peng, V. Krungleviciute, I. Eryazici, J. T. Hupp, O. K. Farha, T. Yildirim, Methane storage in metal-organic frameworks: Current records, surprising findings, and challenges, *Journal of American Chemical Society* 135 (32) (2013) 11887–11894.

- [59] J. M. Simmons, H. Wu, Z. Wei, T. Yildirim, Carbon capture in metal-organic frameworks; a comparative study, *Energy & Environmental Science* 4 (6) (2011) 2177–2185.
- [60] B. Wang, X. Zhang, H. Huang, Z. Zhang, T. Yildirim, W. Zhou, S. Xiang, B. Chen, A microporous aluminum-based metal-organic framework for high methane, hydrogen, and carbon dioxide storage, *Nano Research* 14 (2) (2021) 507–511.
- [61] B. J. Bucior, N. S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O. K. Farha, N. Bagheri, R. Q. Snurr, Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks, *Molecular Systems Design & Engineering* 4 (1) (2019) 162–174.
- [62] A. Raza, A. Sturluson, C. M. Simon, X. Fern, Message passing neural networks for partial charge assignment to metal-organic frameworks, *Journal of Physical Chemistry C* 124 (35) (2020) 19070–19082.