PMTransformer: Universal Transfer Learning and Cross-material Few-shot Learning in Porous Materials

Hyunsoo Park[⊥], *Yeonghun Kang*[⊥], *and Jihan Kim**

Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

⊥ These authors contributed equally to this work

ABSTRACT

Porous materials have emerged as a promising solution for a wide range of energy and environmental applications. However, the asymmetric development in the field of MOFs has led to data imbalance when it comes to MOFs versus other porous materials such as COFs, PPNs, and zeolites. To address this issue, we introduce PMTransformer (Porous Material Transformer), a multi-modal pre-trained Transformer model pre-trained on a vast dataset of 1.9 million hypothetical porous materials, including metal-organic frameworks (MOFs), covalent-organic frameworks (COFs), porous polymer networks (PPNs), and zeolites. PMTransformer showcases remarkable transfer learning capabilities, resulting in state-of-the-art performance in predicting various porous material properties. To address the challenge of asymmetric data aggregation, we propose cross-material few-shot learning, which leverages the synergistic effect among different porous material classes to enhance fine-tuning performance with a limited number of examples. As a proof of concept, we demonstrate its effectiveness in predicting bandgap values of COFs using the available MOF data in the training set. Moreover, we established cross-material relationships in porous materials by predicting unseen properties of other classes of porous materials. Our approach presents a new pathway for understanding the underlying relationships between various classes of porous materials, paving the way toward a more comprehensive understanding and design of porous materials.

Introduction

Porous materials possess void spaces that can be exploited for many different applications.[1,2] Depending on the specific nature of the constituent blocks, they can be further categorized into subclasses of materials including metal-organic frameworks (MOFs)[3], covalent organic frameworks (COFs)[4,5], porous polymer networks (PPNs)[5], and zeolites[6]. Since these materials are composed of diverse combinations of molecular building blocks, the nearly infinite chemical design space presents an excellent opportunity to design these materials for a wide range of applications, including gas storage and separation[7], catalysis[8], and drug delivery[9]. And due to the increasing number of experimental and computational structures, recently there have been several works devoted to using a data-science approach to discover and design new porous materials using various different methods.[10,11]

In recent years, machine learning (ML) models have shown promising results in constructing structure-property relationships for porous materials. For instance, Shi et al.[12] have demonstrated the effectiveness of using two-dimensional (2D) energy histogram features, which include structure-gas interaction energies and energy grid gradients at grid points, as descriptors to accurately predict the gas uptake of MOFs. Also, a 3D convolutional neural network (CNN) with 3D voxel, a volume element in 3D space that is analogous to a pixel in 2D space, has been developed as a descriptor for accurate prediction of gas uptake in zeolites.[13] For predicting electronic properties such as band gap, graph neural networks (GNNs) such as Crystal Graph Convolutional Neural Networks (CGCNN)[14] and MatErials Graph Network (MEGNET)[15] have shown high performance. Also, various descriptors have been developed including geometric, chemical, topological features, revised autocorrelations (RAC)[16] and smooth overlap of atomic positions (SOAP)[17]. Recently, MOFTransformer[18], a multi-modal pre-training Transformer, has

been introduced to achieve universal transfer learning in MOFs, showcasing its exceptional ability to transfer learning across various MOF properties.

Despite the potential of machine learning models for predicting material properties in porous materials, their usefulness remains limited by the availability of data. And while MOFs have been extensively explored due to the large number of experimentally reported structures (over 100,000)[19], other porous materials have much smaller number of experimentally reported data. The CoRE COF[20] and Curated COF[21] databases include around 600 experimentally reported COFs, and fewer than 100 PPNs have been synthesized.[22] COFs and PPNs are formed by covalent bonds and strong C-C bonds, respectively, which make them harder to synthesize into crystalline materials due to the lack of reversible reactions.[23] Additionally, zeolites, composed of Si and O atoms have only a bit over two hundred known topologies.[24] This lack of available data for other porous materials poses a significant challenge for developing accurate machine-learning models across all porous materials and perhaps is one of the reasons on why the machine learning works on porous materials thus far has been skewed towards MOFs.

To overcome the challenge of asymmetry data aggregation, it is our opinion that leveraging data from other porous materials represents a promising solution when a specific material class lacks sufficient data (both in terms of number of materials and properties) for model training. For instance, the restricted data availability of only hundreds of COF structures may pose substantial obstacles when it comes to developing machine learning models to predict the properties of COFs. By incorporating data from abundant source materials such as MOFs, the accuracy of the model predictions can be improved through exploiting the potential synergistic effect between the two material classes. To the best of our knowledge, this type of cross-material transfer learning has yet to be explored in any other materials. Indeed, one can envision that such an approach could

enhance the accuracy of machine learning model predictions in overcoming data scarcity challenges through the potential synergistic effect between materials from distinct classes.

In this work, we introduce the Porous Material Transformer (PMTransformer), which is a multi-modal Transformer architecture based on the MOFTransformer and is pre-trained with 1.9 million hypothetical porous materials, including MOFs, COFs, PPNs, and zeolites. The model showcases excellent transfer learning capability across various properties of porous materials, thereby achieving state-of-the-art performance in predicting multiple different properties. To address the challenge of asymmetry data aggregation in porous materials, we propose cross-material few-shot learning to improve predictions of materials lacking available data for their properties by exploiting the uniform characteristics in porous materials. Moreover, we obtain cross-relationships in porous materials by predicting unseen properties of other classes of porous materials. Our approach provides a novel perspective for understanding the underlying and uniform relationships between various classes of porous materials, allowing for the prediction of previously unexplored properties across these material classes.

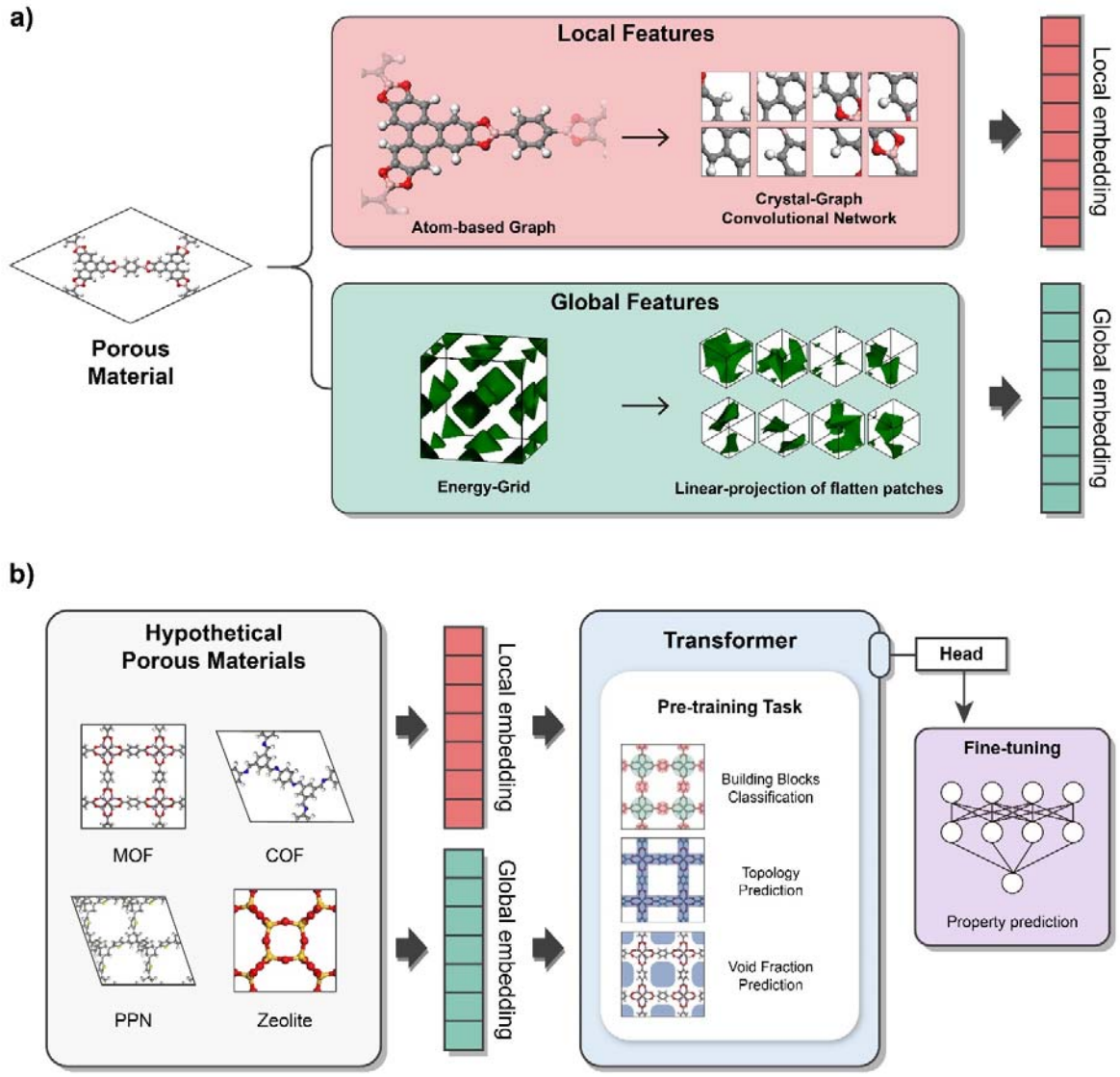


Figure 1. (a) Data representations for porous materials incorporating both local features and global features used with atom-based graphs and energy grids, respectively. (b) Overall schematics of PMTransformer. The model was pre-trained with 1.9 million hypothetical porous materials with three pre-training tasks to capture local and global features in a pre-training stage. In a fine-tuning stage, the PMTransformer is fine-tuned to predict properties of porous materials where its initial weights are initialized with the pre-training weights.

Results

Data representations of MOFs for PMTransformer

Figure 1(a) shows a representative porous materials input data representations for two disparate features (i.e., local features and global features), which serve as inputs of PMTransformer. The local features involve atomistic information related to chemistry of building blocks and specific bonds. The output features of crystal graph convolutional neural networks (CGCNN) were adopted to describe the local features given that they enable capturing atoms' neighbor information such as atom types, distances between neighbor atoms. On the other hand, the global features represent crystalline features including topological and geometric descriptors such as pore volume, surface area, which are captured by the 3D energy grids. The grids are created by calculating interaction energy between a structure and a gas molecules (or gas probe) at each grid point, and can be treated as 3D images, thereby leading to understand the global features. Similar to the Vision Transformer, energy grids are divided by 6 x 6 X 6 patches and flattened by a linear projection. Finally, the local and global embedding are fed into the Transformer encoder of PMTransformer.

Pre-training of PMTransformer

Figure 1(b) illustrates the overall schematic of PMTransformer indicating pre-training and fine-tuning approach to achieve universal transfer learning in porous materials. The pre-training enables our model to learn how to represent the input data in a way that captures its essential features, which can then be used to improve the performance of the model on fine-tuning tasks. The pre-training tasks are designed to enable the model to understand the essential features of porous materials, resulting in superior performance in transfer learning. Previous studies have demonstrated the effectiveness of pre-training tasks designed for MOFs in the MOFTransformer model.[18] The pre-training with topology prediction, void fraction prediction, and metal cluster &

organic linker classification significantly improve transfer learning in MOFs as these tasks facilitate capturing both local and global features of MOFs, which is critical for accurate property prediction.

Building on the pre-training tasks of MOFTransformer, we extended the pre-training tasks to include COFs, PPNs, and zeolites. The pre-training tasks include topology prediction and void fraction prediction for capturing global features of porous materials, and building block classification for capturing local features. Building block classification involves classifying the (1) metal cluster and organic linkers for MOFs, (2) center and linker for COFs and PPNs, and (3) Si and O atoms for zeolites. The accuracies of the pre-training tasks in PMTransformer are comparable to those of MOFTransformer, with topology prediction and building block classification achieving accuracies of 0.98 and 0.99, respectively, and void fraction prediction having a mean absolute error of 0.01.

Construction of Porous Material Database

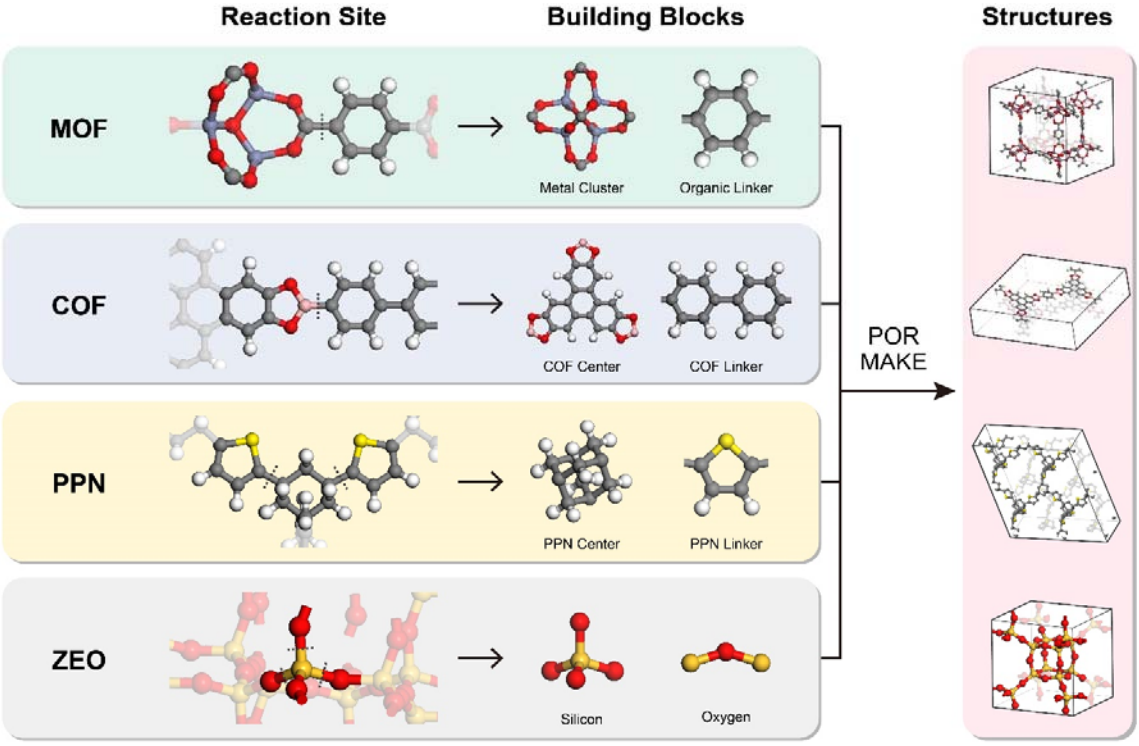
Large and diverse pre-training datasets help the Transformer model learn and comprehend the underlying relationships in pre-training datasets, resulting in improving transfer learning capability in fine-tuning stages. When pre-training the MOFTransformer, one million hypothetical MOFs (hMOFs) were created using the PORMAKE python library[10], with the molecular building blocks and topologies derived from the CoRE MOF, [25] ToBaCCo[26], and RSCR[27] database. In this work, we expanded the pre-training dataset for porous materials to include COFs, PPNs, and zeolites, thereby making it larger and more diverse, as illustrated in Figure 2. Notably, creating pre-training datasets from scratch also facilitates the annotation of topology and building block information for pre-training tasks.

COFs are constructed from organic building blocks with different topologies, linked with covalent bonds. The organic building blocks are relatively rigid backbones that endow the COFs with crystallinity, making them distinct from organic polymers with low crystallinity. The COFs can be synthesized using reactions of boron, triazine, and imine condensation.[2] Various databases of synthesized COFs, including the CURATED COF and CoRE COF databases, as well as hypothetical databases, have been established. For example, Lan et al.[28] developed the Genomic COF database, which contains 471,990 COFs constructed from 130 genetic structure units (GSUs) consisting of 58 centers, 64 linkers, and 8 functional groups with 24 topologies, using reactive sites and quasi-reactive assembly algorithms (QReaxAA). For the pre-training dataset, we constructed a hypothetical COF (hCOF) database using the 130 GSUs and topologies registered in the RCSR database. As shown in Figure 2, we generated 519,606 COFs by PORMAKE, of which only 747 topologies met the constraints with a root mean squared deviation (RMSD) of atomic positions between the building blocks and target node position to measure the strain energy less than 0.3. Notably, the large hCOF dataset, containing numerous COF structures with diverse topologies, enables the PMTransformer to achieve a superior understanding of COFs during pre-training stages.

PPNs constitute a class of porous polymers assembled from tunable building blocks through polymerization reactions, such as homocoupling of tetrahedral monomers. These reactions are typically irreversible, leading to PPNs with exceptional thermal and chemical stability, but amorphous materials. The amorphous nature presents a significant challenge for computational modeling. To address this issue, Martin et al.[29] developed the in-silico PPN database, which utilized a crystalline modeling approach that successfully reproduced the gas adsorption behavior of PPNs. Building on this work, we constructed a diverse hypothetical PPN (hPPN) database for

pre-training by PORMAKE, utilizing the same building blocks from the in-silico PPN database, but with more diverse topologies requiring nodes with four connections for tetrahedral monomers. The building blocks consist of Si, Ge, C, adamantane as centers and 4952 linkers. They result in 277,250 hPPNs including the interpenetrated structures.

Zeolites are a type of crystalline aluminosilicate material composed of silicon, aluminum, and oxygen atoms arranged in tetrahedral structures. Compared to other porous materials like MOFs, COFs, and PPNs, zeolites have a smaller chemical space due to their immutable building blocks. The IZA database[30] currently lists around 250 known zeolite topologies, while the PCOD database[31] was developed using Monte Carlo algorithms and contains many predicted zeolite structures. To prepare for pre-training, we constructed 278 zeolite structures with topologies featuring four connection points using a top-down approach with the RCSR database by PORMAKE. We generated 34,750 zeolites by augmenting these structures through a translational motion in five parts for each cell direction. To supplement the dataset, we randomly selected 65,250 zeolites from the PCOD database, resulting in 100,000 zeolite structures in the pre-training data. The ToposPro software[32] was used to obtain topology information for the structures, but there were still unknown topologies. As such, we labeled these unknown topologies as “unknown topology” during the pre-training stage. All of the generated structures were geometrically optimized using the LAMMPS package[33] with the UFF force field[34].



185

186 **Figure 2.** Construction of diverse and large pre-training dataset for porous materials, including
187 COFs, PPNs, and zeolites, utilized in pre-training the PMTransformer. Hypothetical structures
188 were generated using the PORMAKE Python library, resulting in 1 million hMOFs, 519,606
189 hCOFs, 277,250 hPPNs, and 100,000 zeolites.

190

Fine-tuning results

To evaluate the performance of PMTrasnformer, we compared it with the scratch model (i.e., the default PMTransformer model without any pre-training), the MOFTransformer, which was pre-trained with only MOFs, and several other baseline models, including energy histogram[35], descriptor-based machine learning (ML) model[36], and crystal graph convolutional neural network (CGCNN), using mean absolute errors (MAEs) on different properties of MOFs, COFs, PPNs, and zeolites. The evaluated properties included gas uptake, diffusivity, Henry coefficient, heat of adsorption, stability, and bandgap, as summarized in Table 1, 2.

With regards to the baseline models, the energy histogram model employed the Least Absolute Shrinkage and Selection Operator (LASSO) regression[37], which involved taking an energy histogram that had been converted from energy grids by energy bins. The descriptor-based model utilized 5 geometrical properties (i.e. largest cavity diameter, pore-limiting diameter, gravimetric accessible surface area, volumetric accessible surface area, and volume fraction) as well as 12 chemical properties (i.e. metal type present, number of specified element atoms in unit cell), and 6 additional chemical properties (i.e. total degree of unsaturation, metallic percentage, oxygen to metal ratio, electronegative to total ratio, weighted electronegativity per atom, and nitrogen to oxygen ratio) as inputs. All of these descriptors were used as input to a random forest model. On the other hand, the CGCNN uses atom-based graph representation as inputs, and consists of five convolution layers, one hidden layer after pooling, 64 hidden atom features in convolution layers, and 128 hidden 7 features after pooling.

For the prediction of the MOF properties, the scratch model demonstrated superior performance compared to other baseline models (i.e. energy histogram, descriptor-based ML model, CGCNN) across all properties, as shown in Table 1. It indicates that the data representation of our model

facilitates capturing the underlying feature of MOFs, leading to high performance in predicting various MOF properties. Also, the fine-tuned PMTransformer achieved lower MAE values in all of the MOF properties except for O_2 uptake and N_2 diffusivity compared to the MOFTransformer. This observation indicates that including other porous materials, such as COFs, PPNs, and zeolites, in the pre-training dataset of MOFTransformer leads to higher performance in predicting MOF properties, indicating synergetic effect due to similarity across all porous materials.

For properties of COFs, PPNs, and zeolites, PMTransformer exhibited the lowest MAEs across all properties except for CH_4 uptake at 65 bar in COFs, in which the MOFTransformer had the lowest MAE, as shown in Table 2. Our findings suggest that pre-training with a large set of diverse porous materials, as opposed to pre-training with MOFs alone, plays an important role in improving performance in predicting various properties of porous materials.

Material	Property	Number of Dataset	Energy histogram	Descriptor-based ML	CGCNN	Scratch	MOF Transformer	PM Transformer	Reference
MOF	**H_2 Uptake (100 bar)**	20,000	9.183	9.456	32.864	7.018	6.377	**5.963**	18
MOF	**H_2 diffusivity (dilute)**	20,000	0.644	0.398	0.6600	0.391	0.367	**0.366**	18
MOF	**Band-gap**	20.373	0.913	0.590	0.290	0.271	0.224	**0.216**	38
MOF	**N_2 uptake (1 bar)**	5,286	0.178	0.115	0.108	0.102	0.071	**0.069**	36
MOF	**O_2 uptake (1 bar)**	5,286	0.162	0.076	0.083	0.071	**0.051**	0.053	36
MOF	**N_2 diffusivity (1 bar)**	5,286	7.82e-5	5.22e-5	7.19e-5	5.82e-05	**4.52e-05**	4.53e-05	36
MOF	**O_2 diffusivity (1 bar)**	5,286	7.14e-5	4.59e-5	6.56e-5	5.00e-05	4.04e-05	**3.99e-05**	36
MOF	**CO_2 Henry coefficient**	8,183	0.737	0.468	0.426	0.362	0.295	**0.288**	39
MOF	**Thermal stability**	3,098	68.74	49.27	52.38	52.557	45.875	**45.766**	40

Table 1. Comparison of mean absolute error (MAE) values for various baseline models, scratch, MOFTransformer, and PMTransformer on different properties of MOFs. The bold values indicate the lowest MAE value for each property.

Material	Property	Number of Dataset	Energy histogram	Descriptor-based ML	CGCNN	Scratch	MOF Transformer	PM Transformer	Reference
COF	**CH_4 uptake (65bar)**	39,304	5.588	4.630	15.31	2.883	2.268	**2.126**	[41]
COF	**CH_4 uptake (5.8bar)**	39,304	3.444	1.853	5.620	1.255	**0.999**	1.009	[41]
COF	**CO_2 heat of adsorption**	39,304	2.101	1.341	1.846	1.058	0.874	**0.842**	[42]
COF	**CO_2 log KH**	39,304	0.242	0.169	0.238	0.134	0.108	**0.103**	[42]
PPN	**CH_4 uptake (65bar)**	17, 870	6.260	4.233	9.731	3.748	3.187	**2.995**	[29]
PPN	**CH_4 uptake (1bar)**	17, 870	1.356	0.563	1.525	0.602	0.493	**0.461**	[29]
Zeolite	**CH_4 KH (unitless)**	99,204	8.032	6.268	6.334	4.286	4.103	**3.998**	[43]
Zeolite	**CH_4 Heat of adsorption**	99,204	1.612	1.033	1.603	0.670	0.647	**0.639**	[43]

Table 2. Comparison of mean absolute error (MAE) values for various baseline models, scratch, MOFTransformer, and PMTransformer on different properties of COFs, PPNs, and zeolites. The bold values indicate the lowest MAE value for each property.

Discussion

Cross-material few-shot learning: Prediction of COF Bandgap

Few-shot learning is a promising approach for addressing the challenges posed by limited data availability (typically less than 500) in ML models.[44] In particular, fine-tuning the pre-trained models in vision or language model with only few examples can lead to high performance on unseen tasks. In this work, we applied few-shot learning to the PMTransformer. To address the issue of asymmetry data aggregation in porous materials, we propose a cross-material few-shot learning approach. This approach exploits the synergistic effects from high similarity between different classes of porous materials to improve performance. Specifically, we utilize the relatively abundant number of data for the metal-organic frameworks (MOFs) to train the PMTransformer to predict the properties of other types of porous materials.

Figure 3(a) illustrates the case study application of cross-material few-shot learning to predict the band gap values of the COFs calculated by DFT, where only 400 COF band gap data[45] in the Curated COF dababase are available. The PMTransformer was fine-tuned to predict the COF band gap values by initializing the weights of the model with the weights obtained from the fine-tuned PMTransformer trained on 20,000 MOF bandgaps from the QMOF database. This approach differs from the regular few-shot learning, which involves fine-tuning the PM Transformer with only 400 COF bandgaps. The COF bandgap data was split into 250, 50, and 100 for training, validation, and test. The performance of the few-shot learning and the proposed cross-material few-shot learning methods was compared in terms of mean absolute error (MAE) as the number of training examples ranged from 0 to 250, as shown in Figure 3(b). The results were averaged over five trials on the test set. Notably, the cross-material few-shot learning outperformed the regular few-shot learning method. For instance, when the number of training examples was 250, the cross-material few-shot

learning achieved an r2 score of 0.48, whereas the few-shot learning method achieved an r2 score of only 0.30. These results demonstrate the effectiveness of the proposed cross-material few-shot learning method, which exploits the high similarity among porous materials to achieve a synergetic effect, particularly in cases with limited available data. To further investigate the effect of the number of source materials (MOFs), an ablation study was conducted by varying the number of MOFs used for training from 0 to 20,000 when the number of COF training data was fixed at 250, as shown in Supplementary Figure S6. The results indicate that the performance of the cross-material few-shot learning converged when the number of source material for MOF was at 10,000. Furthermore, we evaluated the cross-material few-shot learning performance of PMTransformer when compared to other ML baseline models such as CGCNN and MEGNET which exhibited high performance in predicting the band gaps in MOFs, as shown in Figure 3(c). The PMTransformer exhibits superior performance compared to other baseline models. This can be attributed to its pre-training, which enabled the PMTransformer to capture general patterns and relationships in porous materials and adapt to new tasks with limited examples. Moreover, it can be observed that the regular few-shot and cross-material few-shot learning in CGCNN and MEGNET do not exhibit a significant improvement in performance compared to the PMTransformer, as demonstrated in Supplementary Figures S7 and S8.

In general, the Transformer architectures[46] are capable of generating attention scores through their attention layers, which reflect the degree of attention the model pays to input features for a given task. These attention scores can be utilized as a tool for feature importance analysis. Figure 3(d) presents the attention scores of representative COF structures, such as COF-1[47], CTF-1[48], and COF-LZU1[49]. In these scores, the larger size of atoms represents higher attention scores, which in turn can be considered as more influential factors in determining band gap. It is important to note

that these structures are composed of benzene rings as linkers and are distinct from their corresponding centers. The analysis of attention scores reveals that the centers have higher attention scores than the linkers, indicating that they play a more prominent role in determining the band gaps. The 2D COFs are known for their ability to extend the π-conjugation system, which leads to greater emphasis being placed on centers that have more than two connection points, as compared to linkers that only have two connection points. Moreover, it is noteworthy that the π-conjugation ability of C-N bonds is a significant aspect to consider. The analysis of attention scores for CTF-1 and COF-LZU-1 indicates that nitrogen atoms within the structures' centers exhibit higher attention scores compared to other atoms. In contrast, the oxygen atoms in the B3O3 rings of COF-1 have relatively lower attention scores among their centers, primarily due to the absence of π-conjugation. This analysis demonstrates the utility of attention scores in providing insights into the underlying factors that determine the band gap of COF structures, thereby facilitating the development of more efficient and accurate models for porous materials.

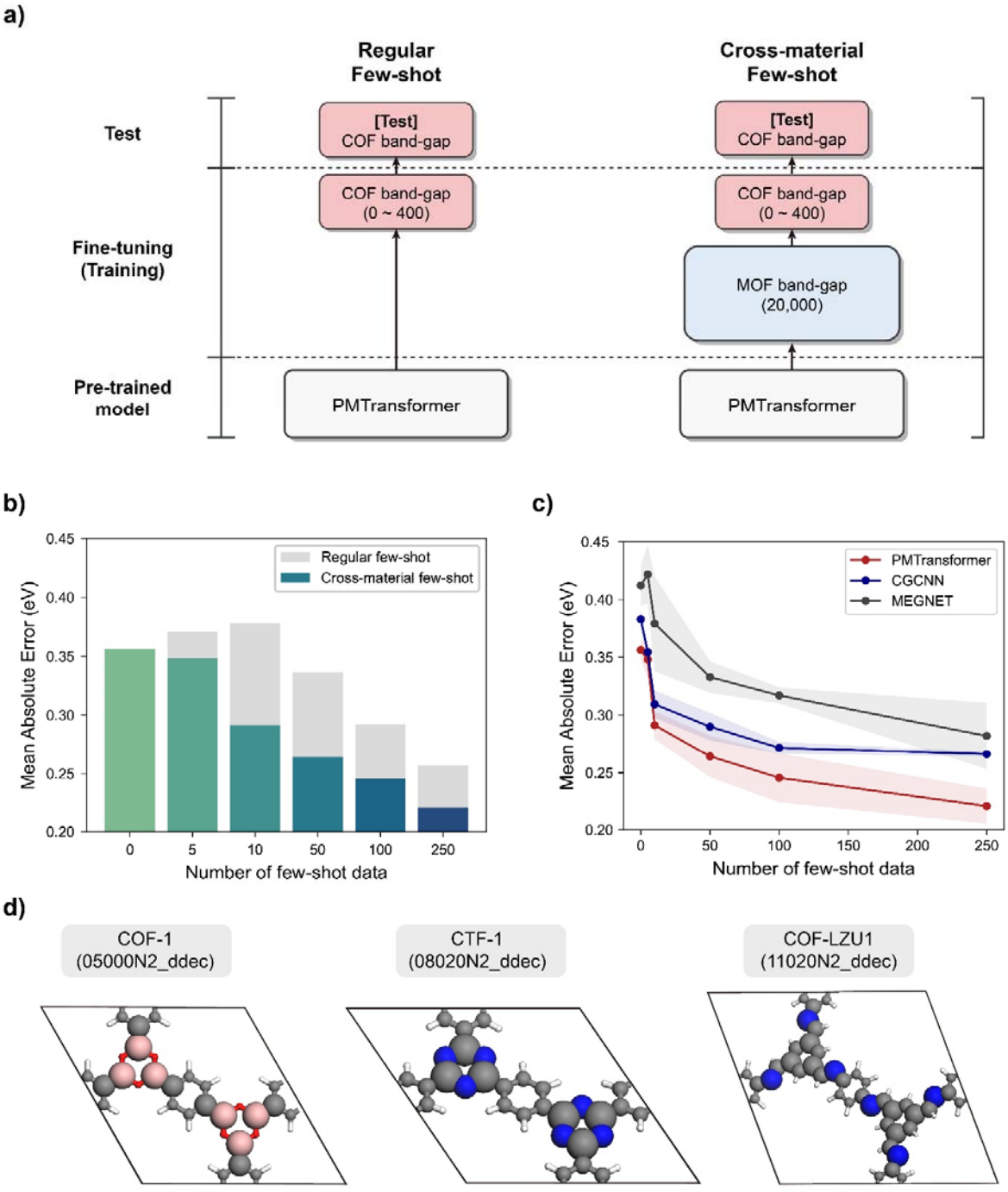


Figure 3. (a) Application of cross-material few-shot learning to predict COF band gaps with limited data. The PMTransformer is fine-tuned using weights from the fine-tuned model on 20,000 MOF band gaps to predict COF band gaps with only 400 examples available. This approach differs from the regular few-shot learning method involving the fine-tuning with only

400 COFs (b) Comparison of MAEs between the regular few-shot and cross-material few-shot results for prediction of band gap of COF as the number of training data (few-shot data) increases from 0 to 250 for PMTransformer. (c) Comparison of MAEs for the cross-material few-shot learning using PMTransformer, CGCNN, and MEGNET as the number of training data (few-shot data) increases from 0 to 250. (d) The schematics for attention scores obtained from the fine-tuned PMTransformer to predict COF band gaps for COF-1, CTF-1, COF-LZU1. The larger atom size represents higher attention scores.

Cross-material relationship in porous materials: H_2 Uptake

Our next case study investigated the cross-material relationship in porous materials and evaluates the ability of PMTransformer to predict unseen properties of other classes of porous materials. The H_2 uptake at 77K and 100 bar was calculated for 5,000 MOFs, COFs, PPNs, and zeolites and randomly split into 4,000 training, 500 validation, and 500 test sets.

In Figure 4(a), a heatmap shows the r2 scores obtained from the PMTransformer fine-tuned with training (or source) materials to predict H_2 uptake and tested on test (or target) materials without further fine-tuning. The diagonal of the heatmap represents the r2 scores when training and test materials are identical. Remarkably, the PMTransformer fine-tuned with MOFs as the training material achieved r2 scores higher than 0.9 when predicting the H_2 uptake of COFs and PPNs in the test set, which is comparable to the r2 scores obtained when training and test materials are the same. These results demonstrate the ability of the PMTransformer model to accurately predict the H_2 uptake of COFs and PPNs when fine-tuned with MOFs as the training material. It is noteworthy that the r2 scores between MOFs, COFs, and PPNs exceed 0.85, indicating their synergetic effect in the cross-material relationship due to their high level of similarity, except when COFs and MOFs are respectively the training and test materials. Conversely, zeolites exhibit low r2 scores, regardless of the source materials, suggesting that zeolites have a lack of synergy with other classes of porous materials.

This observation is supported by the t-SNE plot created by the class tokens from the fine-tuned PMTransformer with MOFs, COFs, PPN, and zeolites in test set, respectively, as illustrated in Figure 4(b). The plot reveals a unique clustering of zeolites, which are positioned solely within the lower H_2 uptake region. It is attributed to the composition of zeolites, which consist primarily of Si and O atoms, resulting in smaller pore sizes and consequently, lower H_2 uptake compared to

other porous materials that are typically composed of molecular building blocks such as metal nodes, organic linkers, and polymer monomers.

Figure 4(c) shows four highlighted structures in the t-SNE plot, where their building blocks and naming are shown in Supplementary Figure S9. **M1** exhibits the highest H_2 uptake value of 55.42 g/L, which is 2D COF with the *hyw* topology, composed of 3,4,9,10-Perylenetetra-carboxylic acid, biphenyl, and 1-cyanopyrene. Interestingly, the high H_2 uptake region in the vicinity of **M1** is mostly populated by COFs, which can be attributed to their void fraction. To investigate this further, the t-SNE plot in Supplementary Figure S10 is colored according to their void fraction values calculated using ZEO++. The COF structures located within the high H_2 uptake region have void fractions ranging between 0.45 and 0.55, which seems to be the optimal range for high H_2 uptake performance, as shown in Supplementary Figure S11. In contrast, **M2** (PPN) and **M3** (MOF) exhibit very low and very high void fraction values, respectively, as depicted in Figure SX, due to their building block. **M2** consists of Ge atoms as centers and short linkers, specifically 1,3-dibromo-1-propanol, while **M3** has a long organic linker, dithieno[3,2-b:2',3'-e]benzene-2,6-dicarboxlyic acid[50]. Additionally, among the zeolites, **M4**, which is sourced from the PCOD database, exhibits the highest H_2 uptake value of 51.48 g/L.

It should be noted that MOFs and PPNs cluster closely together, while COFs are more dispersed, with most located in the highest H_2 uptake region. This behavior can be ascribed to the fact that MOFs and PPNs have common topologies when constructed by PORMAKE, while COFs have a greater diversity of 2D topologies. Indeed, most of entries in the CoRE COF and the CURATED COF database are 2D COF, rather than 3D. Among GSUs from the genomic COF database, the centers are mostly composed of building blocks from 2D COFs than 3D COFs, while building blocks of MOFs and PPNs were derived from 3D structures. This is because the building blocks

of MOFs were obtained from the CoRE COF database, which contains only 3D MOFs in the CSD database. This limitation suggests a need for a more large and diverse pre-training dataset, including 2D MOFs, which would lead to superior transfer learning capability in the fine-tuning stage. Other limitation is the lower accuracy of PMTransformer in predicting zeolite properties. It can be attributed to asymmetry data of zeolites when compared to other porous materials as well as lower diversity of zeolites in porous materials. The pre-training dataset contains only 100,000 zeolites, because their the building blocks (i.e., Si and O) are not tunable, resulting in small chemical space. These limitations must be taken into consideration in future studies.

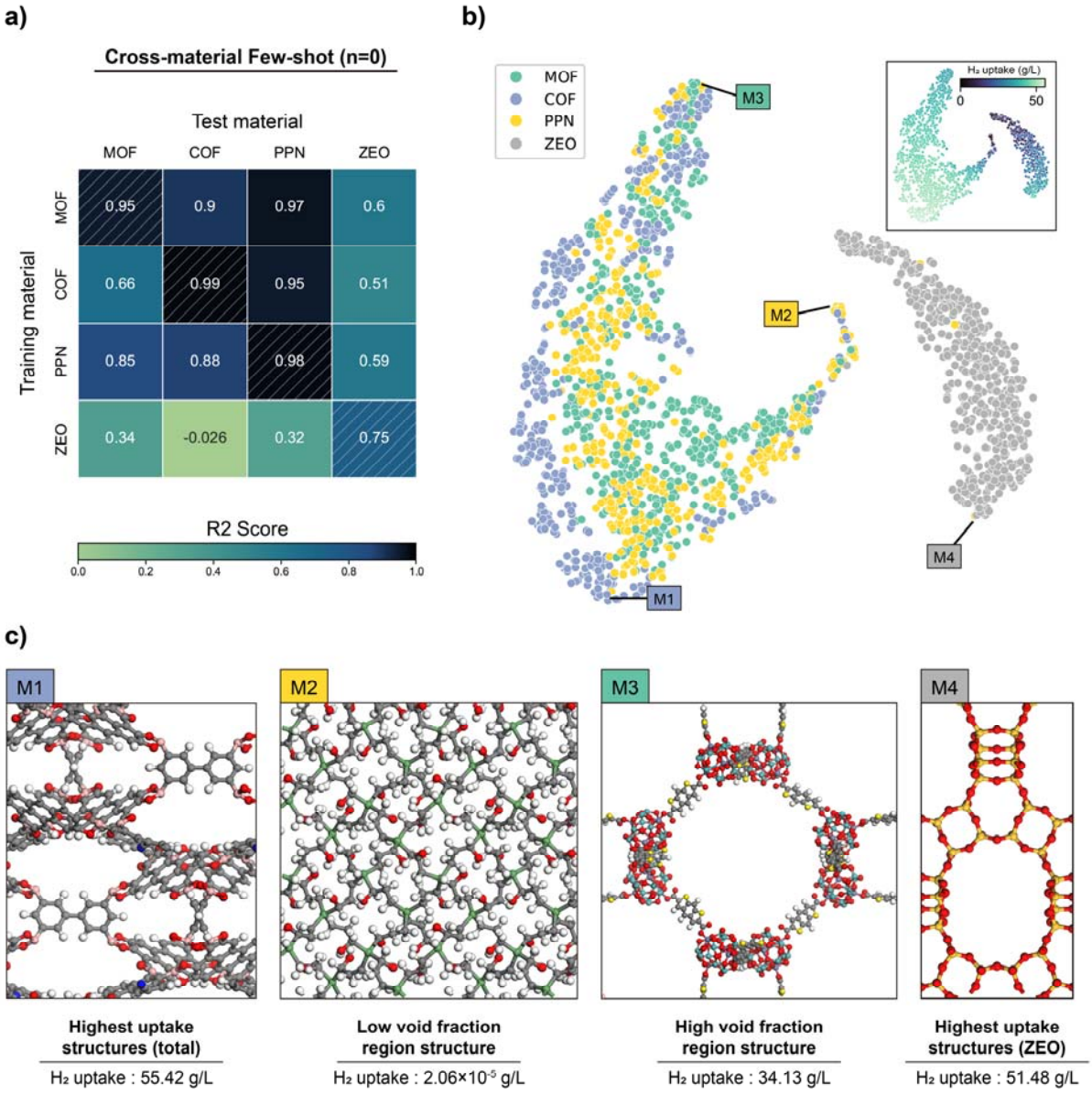


Figure 4. (a) Heatmap of r2 scores obtained from the PMTransformer fine-tuned with training materials to predict H_2 uptake and tested on test materials without further fine-tuning between MOFs, COFs, PPNs and zeolites. (b) A t-SNE plot of class tokens obtained from the fine-tuned PMTransformer with MOFs, COFs, PPN, and zeolites in the test set, with the additional small figure colored by H_2 uptake (c) The t-SNE plot highlights several structures based on their H2 uptake and void fraction characteristics, including **M1** with the highest H_2 uptake, **M2** with low

void fraction, **M3** with high void fraction, and the zeolite structure **M4** with the highest H_2 uptake.

Conclusions

In this work, we present the Porous Material Transformer (PMTransformer) model that combines multi-modal features from MOFs, COFs, PPNs, and zeolites. By pre-training on 1.9 million hypothetical porous materials, our model achieved state-of-the-art performance in predicting various properties of porous materials via fine-tuning. Furthermore, we introduced cross-material few-shot learning to address the challenge of asymmetry data aggregation in porous materials and proposed a method for predicting previously unexplored properties across different material classes (e.g. using MOF data to predict COF properties). Our approach provides an opportunity for understanding the underlying relationships between various classes of porous materials, allowing for the prediction of previously unexplored properties across these material classes, and thus facilitating a more comprehensive understanding and design of porous materials.

Methods

Training details

We adopted a pre-training and fine-tuning approach similar to that used in previous work, MOFTransformer. We note that in few-shot learning, the model is fine-tuned with only a few samples, leveraging the pre-training weights as an initialization. The optimization process in all stages, including pre-training, fine-tuning, and few-shot learning, employed the AdamW[51] optimizer with a learning rate of 10^{-4} and weight decay of 10^{-2}. During the initial phase of the optimization process, the learning rate was gradually increased for the first 5 % of the total epoch and then linearly decayed to zero for the remaining epochs.

During the pre-training stage, the model was trained using a batch size of 1024 for a total of 100 epochs. For fine-tuning and few-shot learning, the model was trained using a smaller batch size of 32 for a total of 20 epochs. The dataset is split randomly into train, validation, and test, with a ratio of 8 : 1 : 1. We adopted the standardization method for scaling the target properties

Computational details for molecular simulation

The H_2 uptake of 5000 MOFs, COFs, PPNs, and zeolites was calculated for cross-material relationships using the RASPA package[52]. The property was used due to its relatively facile calculation with a united atom model. The pseudo-Feynman-Hibbs model was used to describe the H_2 behavior at low temperatures, and Lenard-Jones potentials were fitted to the Feynman-Hibbs potential[53] at T = 77 K. The UFF force field was used for all molecules except H_2, with the Lorentz-Berthelot mixing rule and a cutoff distance of 12.8 Å. To calculate H_2 uptake, GCMC simulations were performed for 10k production cycles at 100 bar and 77 K, with 5k cycles used for initialization.

Conflicts of interest

There are no conflicts to declare.

Author Contributions

H.P and Y.K contributed equally to this work. H.P and Y.K developed PMTransformer and wrote the manuscript with J.K. The manuscript was written through the contributions of all authors. All authors have given approval for the final version of the manuscript.

Data availability

The UFF-optimized CIF files of hypothetical porous materials database used as pre-training dataset are available at https://doi.org/10.6084/m9.figshare.21810147 for MOFs, https://doi.org/10.6084/m9.figshare.22699303 for COFs, PPNs and zeolites. The pre-training PMTransformer model is available at https://doi.org/10.6084/m9.figshare.22698655.v2.

Code availability

The PMTransformer library is based on the MOFTransformer, which is available at at https://github.com/hspark1212/MOFTransformer. From version 2.0.0, the default pre-training model has been changed from MOFTransformer to PMTransformer. For the sake of reproducibility, all results in this manuscript are obtained from a 2.0.0 version of the library, which is available at https://pypi.org/project/moftransformer/2.0.0.

Acknowledgements

H. P., Y. K., and J. K. acknowledge funding from the National Research Foundation of Korea (NRF) under Project Number 2021M3A7C208974513 and 2021R1A2C2003583. This work was supported by the National Supercomputing Center with supercomputing resources including

427 technical support (KSC-2022-CRE-0515). The authors would like to thank Berend Smit and
428 Beatriz Mouriño who provided us with the band gap values calculated by DFT for the CURATED
429 COFs database.

430

References

1 Bennett, T. D., Coudert, F.-X., James, S. L. & Cooper, A. I. The changing state of porous materials. *Nature Materials* **20**, 1179-1187 (2021).

2 Geng, K. *et al.* Covalent organic frameworks: design, synthesis, and functions. *Chemical reviews* **120**, 8814-8933 (2020).

3 Zhou, H.-C., Long, J. R. & Yaghi, O. M. Vol. 112 673-674 (ACS Publications, 2012).

4 Feng, X., Ding, X. & Jiang, D. Covalent organic frameworks. *Chemical Society Reviews* **41**, 6010-6022 (2012).

5 Makal, T. A., Li, J.-R., Lu, W. & Zhou, H.-C. Methane storage in advanced porous materials. *Chemical Society Reviews* **41**, 7761-7779 (2012).

6 Ozin, G. A., Kuperman, A. & Stein, A. Advanced zeolite, materials science. *Angewandte Chemie International Edition in English* **28**, 359-376 (1989).

7 Li, H. *et al.* Recent advances in gas storage and separation using metal–organic frameworks. *Materials Today* **21**, 108-121 (2018).

8 Lee, J. *et al.* Metal–organic framework materials as catalysts. *Chemical Society Reviews* **38**, 1450-1459 (2009).

9 Della Rocca, J., Liu, D. & Lin, W. Nanoscale metal–organic frameworks for biomedical imaging and drug delivery. *Accounts of chemical research* **44**, 957-968 (2011).

10 Lee, S. *et al.* Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces* **13**, 23647-23654 (2021).

11 Yao, Z. *et al.* Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence* **3**, 76-86 (2021).

12 Shi, K. *et al.* Two-Dimensional Energy Histograms as Features for Machine Learning to Predict Adsorption in Diverse Nanoporous Materials. *Journal of Chemical Theory and Computation* (2023).

13 Cho, E. H. & Lin, L.-C. Nanoporous material recognition via 3D convolutional neural networks: Prediction of adsorption properties. *The journal of physical chemistry letters* **12**, 2279-2285 (2021).

14 Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120**, 145301 (2018).

15 Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **31**, 3564-3572 (2019).

16 Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. *The Journal of Physical Chemistry A* **121**, 8939-8954 (2017).

17 Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Physical Review B* **87**, 184115 (2013).

18 Kang, Y., Park, H., Smit, B. & Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence* **5**, 309-318, doi:10.1038/s42256-023-00628-2 (2023).

19 Moghadam, P. Z. *et al.* Development of a Cambridge Structural Database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials* **29**, 2618-2625 (2017).

20 Tong, M., Lan, Y., Yang, Q. & Zhong, C. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chemical Engineering Science* **168**, 456-464 (2017).

21 Ongari, D., Yakutovich, A. V., Talirz, L. & Smit, B. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. *ACS central science* **5**, 1663-1675 (2019).

22 Lu, W. *et al.* Porous polymer networks: synthesis, porosity, and applications in gas storage/separation. *Chemistry of Materials* **22**, 5964-5972 (2010).

23 Haase, F. & Lotsch, B. V. Solving the COF trilemma: towards crystalline, stable and functional covalent organic frameworks. *Chemical Society Reviews* **49**, 8469-8500 (2020).

24 Blatov, V. A., Blatova, O. A., Daeyaert, F. & Deem, M. W. Nanoporous materials with predicted zeolite topologies. *RSC advances* **10**, 17760-17767 (2020).

25 Chung, Y. G. *et al.* Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data* **64**, 5985-5998 (2019).

26 Anderson, R. & Gómez-Gualdrón, D. A. Increasing topological diversity during computational "synthesis" of porous crystals: how and why. *CrystEngComm* **21**, 1653-1665 (2019).

27 O'Keeffe, M., Peskov, M. A., Ramsden, S. J. & Yaghi, O. M. The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. *Accounts of chemical research* **41**, 1782-1789 (2008).

28 Lan, Y. *et al.* Materials genomics methods for high-throughput construction of COFs and targeted synthesis. *Nature Communications* **9**, 5274 (2018).

29 Martin, R. L., Simon, C. M., Smit, B. & Haranczyk, M. In silico design of porous polymer networks: high-throughput screening for methane storage materials. *Journal of the American Chemical Society* **136**, 5006-5022 (2014).

30 Baerlocher, C. Database of zeolite structures. *http://www.iza-structure. org/databases/* (2008).

31 Pophale, R., Cheeseman, P. A. & Deem, M. W. A database of new zeolite-like materials. *Physical Chemistry Chemical Physics* **13**, 12407-12412 (2011).

32 Blatov, V. A., Shevchenko, A. P. & Proserpio, D. M. Applied topological analysis of crystal structures with the program package ToposPro. *Crystal Growth & Design* **14**, 3576-3586 (2014).

33 Thompson, A. P. *et al.* LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **271**, 108171 (2022).

34 Rappé, A. K., Casewit, C. J., Colwell, K., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **114**, 10024-10035 (1992).

35 Bucior, B. J. *et al.* Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Molecular Systems Design & Engineering* **4**, 162-174 (2019).

36 Orhan, I. B., Daglar, H., Keskin, S., Le, T. C. & Babarao, R. Prediction of O2/N2 Selectivity in Metal–Organic Frameworks via High-Throughput Computational Screening and Machine Learning. *ACS Applied Materials & Interfaces* **14**, 736-749 (2021).

37 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267-288 (1996).

38 Rosen, A. S. *et al.* Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **4**, 1578-1597 (2021).

39 Moosavi, S. M. *et al.* Understanding the diversity of the metal-organic framework ecosystem. *Nature communications* **11**, 1-10 (2020).

40 Nandy, A., Duan, C. & Kulik, H. J. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks. *Journal of the American Chemical Society* **143**, 17535-17547 (2021).

41 Mercado, R. *et al.* In silico design of 2D and 3D covalent organic frameworks for methane storage applications. *Chemistry of Materials* **30**, 5069-5086 (2018).

42 Deeg, K. S. *et al.* In silico discovery of covalent organic frameworks for carbon capture. *ACS applied materials & interfaces* **12**, 21559-21568 (2020).

43 Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Science advances* **6**, eaax9324 (2020).

44 Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* **53**, 1-34 (2020).

45 Mouriño, B., Jablonka, K. M., Ortega-Guerrero, A. & Smit, B. In search of Covalent Organic Framework photocatalysts: A DFT-based screening approach. (2023).

46 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

47 Cote, A. P. *et al.* Porous, crystalline, covalent organic frameworks. *science* **310**, 1166-1170 (2005).

48 Kuhn, P., Antonietti, M. & Thomas, A. Porous, covalent triazine-based frameworks prepared by ionothermal synthesis. *Angewandte Chemie International Edition* **47**, 3450-3453 (2008).

49 Ding, S.-Y. *et al.* Construction of covalent organic framework for catalysis: Pd/COF-LZU1 in Suzuki–Miyaura coupling reaction. *Journal of the American Chemical Society* **133**, 19816-19822 (2011).

50 Wang, S., Xiong, S., Wang, Z. & Du, J. Rational Design of Zinc–Organic Coordination Polymers Directed by N-Donor Co-ligands. *Chemistry–A European Journal* **17**, 8630-8642 (2011).

51 Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

52 Dubbeldam, D., Calero, S., Ellis, D. E. & Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation* **42**, 81-101 (2016).

53 Feynman, R. P., Hibbs, A. R. & Styer, D. F. *Quantum mechanics and path integrals*. (Courier Corporation, 2010).