A Multi-modal Pre-training Transformer for Universal Transfer Learning in Metal-Organic Frameworks

Yeonghun Kang[†, ⊥], *Hyunsoo Park*[†, ⊥], *Berend Smit*[‡], *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

‡ Laboratory of molecular simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Valais, Ecole Polytechnique Fédérale de Lausanne (EPFL), Rue de l'Industrie 17, CH-1951, Sion, Switzerland

⊥ These authors contributed equally to this work

*Correspondence to : jihankim@kaist.ac.kr

ABSTRACT

Metal-organic frameworks (MOFs) are a class of crystalline porous materials that exhibit a vast chemical space due to their tunable molecular building blocks with diverse topologies. Given that an unlimited number of MOFs can, in principle, be synthesized, constructing structure-property relationships through a machine learning approach allows for efficient exploration of this vast chemical space, resulting in identifying optimal candidates with desired properties. In this work, we introduce MOFTransformer, a multi-model Transformer encoder pre-trained with 1 million hypothetical MOFs. This multi-modal model utilizes integrated atom-based graph and energy-grid embeddings to capture both local and global features of MOFs, respectively. By fine-tuning the pre-trained model with small datasets ranging from 5,000 to 20,000 MOFs, our model achieves state-of-the-art results for predicting across various properties including gas adsorption, diffusion, electronic properties, and even text-mined data. Beyond its universal transfer learning capabilities, MOFTransformer generates chemical insights by analyzing feature importance through attention scores within the self-attention layers. As such, this model can serve as a bedrock platform for other MOF researchers that seek to develop new machine learning models for their work.

Introduction

Metal-organic frameworks (MOFs) are a class of crystalline porous materials used for various energy and environmental applications[1-4] due to their excellent properties such as large surface area,[5] high chemical/thermal stability,[6] and tunability.[7] Given that MOFs are composed of thousands of tunable molecular building blocks (i.e., metal nodes and organic linkers), an infinite number of MOFs can, in principle, be synthesized taking into all the different combinations. To efficiently explore this vast MOF search space, it is important to identify the structure-property relationship for a given application. One can then focus on MOFs that contain specific structures that can lead to user-desired properties. To gain information regarding this relationship, high-throughput computational screening approaches has been primarily used by conducting simulations on a large dataset of MOF structures and retroactively identifying the structure/property relationship.[8-11] However, this can be a cumbersome process and more importantly, one would need to conduct independent computational screenings for each of the applications, which requires a vast quantity of computational resources.

An alternative way to discover the structure-property relationship is through a machine-learning (ML) approach, and this methodology has gained a lot of traction lately.[12,13] In particular, geometric descriptors of MOF structures (e.g. void fraction and pore volume) have been used to accurately predict various gas adsorption properties.[14-16] Also, Bucior et al.[17] developed a machine learning model using energy grid histograms as descriptors to predict gas uptake properties. For diffusion properties, Ibrahim et al.[18] developed a machine-learning model to predict N_2/O_2 selectivity and diffusivity using geometric, atom-type, and chemical feature descriptors. For electronic properties, Rosen et al.[19] demonstrated that a graph neural network facilitates capturing the underlying chemical features leading to accurate predictions in the band gap values for the

MOFs. Unfortunately, in all these previous studies, the developed machine-learning model cannot be readily transferred from one application to another. As such, one would need to restart the training process and develop a new machine-learning model from scratch for every different application.

To remedy this issue, one can utilize transfer learning, which incorporates knowledge from one machine learning application to another and, thereby, in principle, saving computational time for subsequent machine learning works. Although transfer learning has been applied in a few cases for MOFs, it is still limited to specific properties (e.g. transfer knowledge from gas uptake to gas diffusivity or between different gas types), limiting their utility.[16,20] To make transferability a feasible solution, a universal transfer learning model that can be applied to all possible properties needs to be constructed. To achieve this, machine-learning models and descriptors should capture two disparate types of features for MOFs: (1) local features (e.g., specific bonds and chemistry makeup of the building blocks) and (2) global features (e.g., geometric and topological descriptors). Although both the local descriptors (e.g. CGCNN,[19,21] chemical descriptors,[18] RACs,[22,23] and building-block embedding.[11,24,25]) and the global features (e.g., geometric features calculated by ZEO++,[26] the histograms of energy-grids.[16,17]) have been developed previously, as far as we know, none of these works have effectively captured both the local and global features to achieve universal transfer learning.

When it comes to multi-modal learning that takes in multiple inputs, the Transformer architecture[27] (initially proposed for sequence data such as language models) has emerged as the dominant modeling network. Given that the Transformer consists of self-attention layers, which enables handling sequences of data in parallel, it facilitates efficient training of neural networks with vast amounts of data. In 2019, Google introduced BERT, a pre-training Transformer encoder in the language model,[28] and demonstrated remarkable performance in transfer learning. By fine-tuning the pre-trained BERT model, it obtained state-of-the-art performance results for many Natural Language Process (NLP) tasks such as question-answering and named entity recognition. Moreover, for computer vision, various vision Transformer architectures have emerged as an alternative solution to convolution neural networks (CNNs).[29] Recently, the pre-trained Transformers' transfer learning strategy has been expanded to multi-modal learning.[30] And finally, the pre-trained multi-modal Transformers achieved state-of-the-art results in vision-language models such as image captioning and vision-question answering.[31-33] Due to its superior performance, the Transformer architectures have recently been adopted to predict various properties of MOFs.[34,35]

In this work, for the first time in MOF research, we introduce the multi-modal Transformer architecture (named “MOFTransformer”), which captures both the local and global features. Our MOFTransformer was pre-trained with 1 million hypothetical MOFs (hMOFs). By fine-tuning the pre-trained MOFTransformer, it showcases excellent prediction capabilities across multiple different properties (e.g., gas uptake, gas diffusivity, electronic properties of MOFs, and text-mined data). Besides its superior performance, this architecture allows chemists to capture insights from attention scores obtained by the attention layers of the MOFTransformer. As such, we believe that this model can serve as a bedrock architecture/model for future machine learning research for the MOF community.

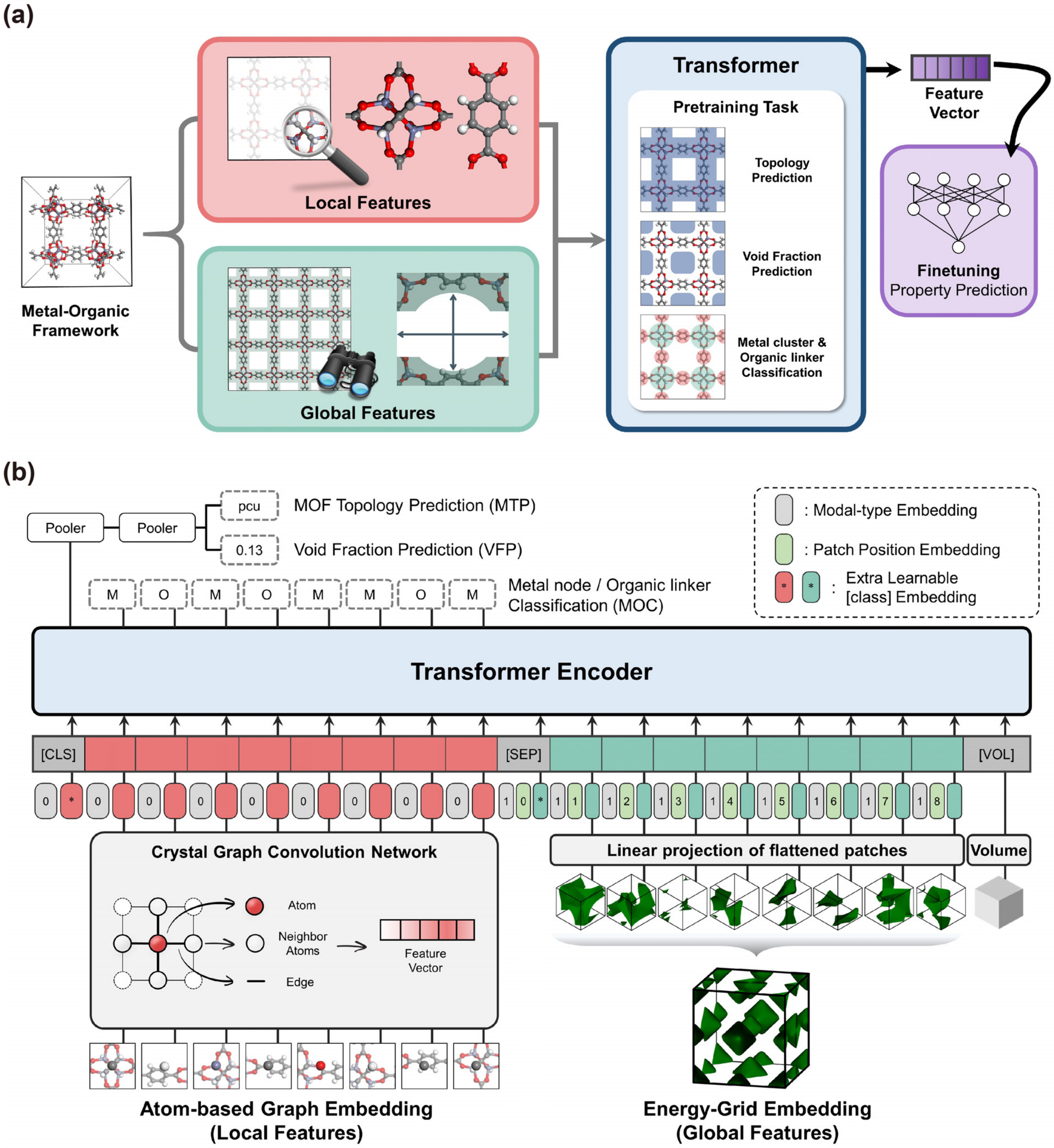
(a)
Metal-Organic Framework
Local Features
Global Features
Transformer
Pretraining Task
Topology Prediction
Void Fraction Prediction
Metal cluster & Organic linker Classification
Feature Vector
Finetuning
Property Prediction
(b)
Pooler
Pooler
pcu
0.13
MOF Topology Prediction (MTP)
Void Fraction Prediction (VFP)
M O M O M M O M
Metal node / Organic linker Classification (MOC)
: Modal-type Embedding
: Patch Position Embedding
: Extra Learnable [class] Embedding
Transformer Encoder
[CLS]
[SEP]
[VOL]
Crystal Graph Convolution Network
Atom
Neighbor Atoms
Edge
Feature Vector
Linear projection of flattened patches
Volume
Atom-based Graph Embedding (Local Features)
Energy-Grid Embedding (Global Features)

Figure 1. (a) Overall schematics of MOFTransformer. The model takes both local and global features as inputs. In a pre-training step, it is trained with three pre-training tasks. In the fine-tuning step, the model is trained to predict desired properties of MOFs using the weights of the pre-trained model as initial weights. (b) The architecture of the MOFTransformer. The input embedding takes atom-based graph embeddings and energy-grid embeddings that serve as local and global features, respectively.

Results

MOFTransformer

The overall schematics of our MOFTransformer is shown in Figure 1(a). To build towards universal transfer learning, both pre-training and fine-tuning strategies are implemented. The objective of pre-training is to allow the MOFTransformer to learn the essential characteristics of a MOF. This pre-trained model serves as a starting point for all subsequent applications. Fine-tuning refers to the process of training the pre-trained models for the specific application at hand (e.g. gas adsorption uptake prediction). Figure 1(b) shows the schematic of the MOFTransformer architecture, which is based on a multi-layer bidirectional Transformer encoder developed by Vaswani et al.[27] The MOFTransformer is a multi-modal Transformer that takes two types of embedding as inputs, each representing the local and global features: (1) atom-based graph embedding (2) energy-grid embedding.

Previously, Xie et al.[21] devised crystal graph convolution neural networks (CGCNN) that transforms atoms (i.e., nodes), bonds (i.e., edges), and their features (i.e., the distance between atoms) into a vector space. Although CGCNN consists of convolutional layers and pooling layers from the original paper, the atom-based graph embedding in the MOFTransformer uses output vectors of the CGCNN without the pooling layers. It allows our model to deal with the atom-wise features without losing information. It should be noted that many atoms in the unit-cell of MOFs have the same embedding from the CGCNN, given that the CGCNN creates the embedding by taking atom types of nodes, distances, and atom types of the neighbor nodes (see Supplementary Figure S1). We grouped the topologically identical atoms and defined these sets as unique atoms (the details of the algorithm are explained in Supplementary Note S1). Removing the information from the overlapping atoms enables efficient training and prevents significant memory issues that frequently appear when training with long sequences of inputs.

When it comes to the energy-grid embedding, the energy grids were calculated using a methane molecule probe that was selected due to its facility in modeling. Universal Force Field,[36] and TraPPE[37] were used to describe adsorbate-adsorbent van der Walls interactions in MOFs and the methane molecule, respectively. The 3D energy grids can be treated as 3D images, which means that the grid points and the energy values of the energy grids serve as pixels and 1-channel colors, respectively. Similar to the Vision Transformer,[29] the MOFTransformer takes 1-dimensional (1D) patches of the flattened 3D energy grids where (H, W, D) are the height, width, and depth of energy grids and (P, P, P) is the patch resolution, and $N = HW\ D/P^3$ is the number of patches. Given that the energy grids were interpolated to 30 × 30 × 30 Å, the height H, weight W, and depth D are 30 Å. The patch size P was set to 5 Å, so the number of patches N is 216.

The MOFTransformer model is derived from the BERT-based model[28] (L=12, H=768, A=12), where L is the number of blocks, H is the hidden size, and A is the number of self-attention heads. Similar to BERT's class and separate tokens, the class token [CLS] and the separate token [SEP], which are learnable embedding layers, are located at the first position and between the two types of embedding, respectively (see Figure 1(b)). The [CLS] token is a head token of the Transformer blocks and predicts desired properties by adding a single pooling layer for the pre-training and fine-tuning tasks. Apart from these, a volume token [VOL], which is the normalized cell volume, is added at the final position of the input embedding because the interpolation of the energy grids leads to a loss of information regarding the volume of the original energy grids. Finally, position embedding and modal-type embedding, which are also learnable embedding layers, are added to the input embedding by the element-wise summation. The position embedding is a vector that encodes the position of the sequence, and the modal-type embedding encodes the two types of embedding to 0 and 1.

Understanding MOF descriptors

It is important to recognize how MOF descriptors (i.e., local features and global features) influence the properties of MOFs. As shown in Figure 2, H_2 uptake, H_2 diffusivity, and band gap were selected as case-study applications for MOFs that represent adsorption, diffusion, and electronic properties, respectively. Figure 2(a-c) shows the structure-property maps obtained from the molecular simulations for each of these applications. For H_2 uptake and diffusivity, the data was taken from our fine-tuning dataset (20,000 structures). The band gap values are obtained from the QMOF database (version 13) with the PBE functional that includes a total of 20,373 structures. From Figure 2(a-b), it can be seen that the H_2 uptake and diffusivity increase with accessible volume fraction and are strongly dependent on the MOF topology due to the correlation between topology and void fraction. Meanwhile, the band gap exhibits no correlation with accessible volume fraction and topology, which is reasonable given that electronic properties are more dependent on local chemical features as opposed to global geometric features.

On top of this, Figure 2(d-f) shows the correlation between the MOF properties and the types of metal atoms. It can be seen that the dependence on metal atoms is lowest for H_2 uptake while highest for the band gap energy. And similar trends can be found for the organic linkers (see Supplementary Figure S2). Along with the aforementioned geometric analysis, Figure 2(d-f) confirms that adsorption and diffusion properties rely more on global features, while electronic properties rely more on local features. Apart from these, some properties like O_2 diffusivity (which is more dependent on electronic effects than H_2 diffusivity) and CO_2 Henry coefficient have more complex correlations between features and properties (see Supplementary Figure S3). As such, this illustrates the importance of integrating both local and global features within the Transformer to enable universal transferability across different applications.

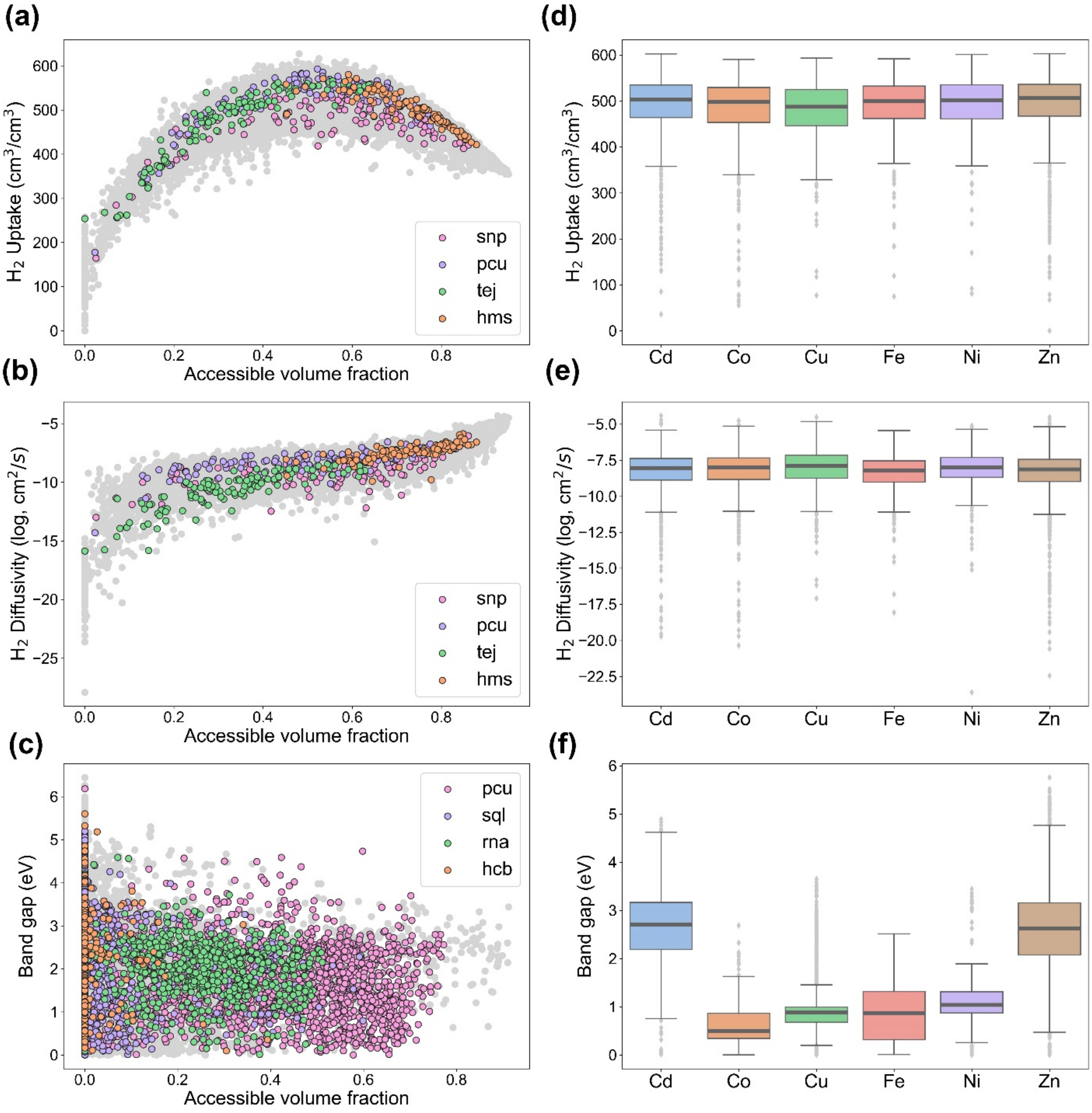

Figure 2. (a-c) Scattered plots showing the relationship between accessible volume fraction and various properties (i.e., gas uptake, diffusivity, and bandgap). Gray dots represent the MOFs from each database, and colored dots represent MOFs with the top four topologies obtained from MOFid.[38] (d-f) The box plot of properties (adsorption, diffusion, and band gap) for each metal type. The dark line in the center of the box represents the median.

Pre-training Results

The pre-training tasks play an essential role in determining the effectiveness of the transfer learning performance. Three pre-training tasks were designed to capture the essential features of the MOFs: (1) MOF topology prediction (MTP), (2) void fraction prediction (VFP), and (3) metal cluster/organic linker classification (MOC). For the MTP task, the model was trained to predict the 1,079 topologies of MOFs by adding a classification head, which consists of a single dense layer to the [CLS] token. The list of topologies is summarized in Supplementary Table S1. For the VFP task, the model is trained to predict accessible void fraction calculated by ZEO++[26] by adding single dense layers to the [CLS] token. Finally, the MOC task was performed as it would enable the model to learn the features separately stemming from each metal node and organic linker. The binary classification (determining a given MOF atom as belonging to the metal or the organic linker) is conducted for the atom-wise features of atom-based embedding. The accuracies of MTP and MOC were 0.97, 0.98 and the MAE of VFP was 0.01.

Next, we visualized the embedding vector of the pre-training model in a two-dimensional space using t-SNE, and PCA methods, as shown in Figure 3. Figure 3(a) shows a result of a t-SNE plot for the embedding vector of class tokens with the top 10 frequently appearing topologies in the dataset. Figure 3(a) shows that MOFs with different topologies are clustered together and segregated from other MOFs, indicating that proper learning has occurred. And the same pattern of results was seen for all topologies (see Supplementary Figure S4). Furthermore, it is interesting to note that the PCA plots exhibit the distribution of the embedding vector that gradually increases according to the void fraction, as shown in Figure 3(b). This indicates that the embedding vectors are clustered with similar values of void fraction. These results demonstrate that the pre-training model is successfully trained to capture the critical features of the MOFs.

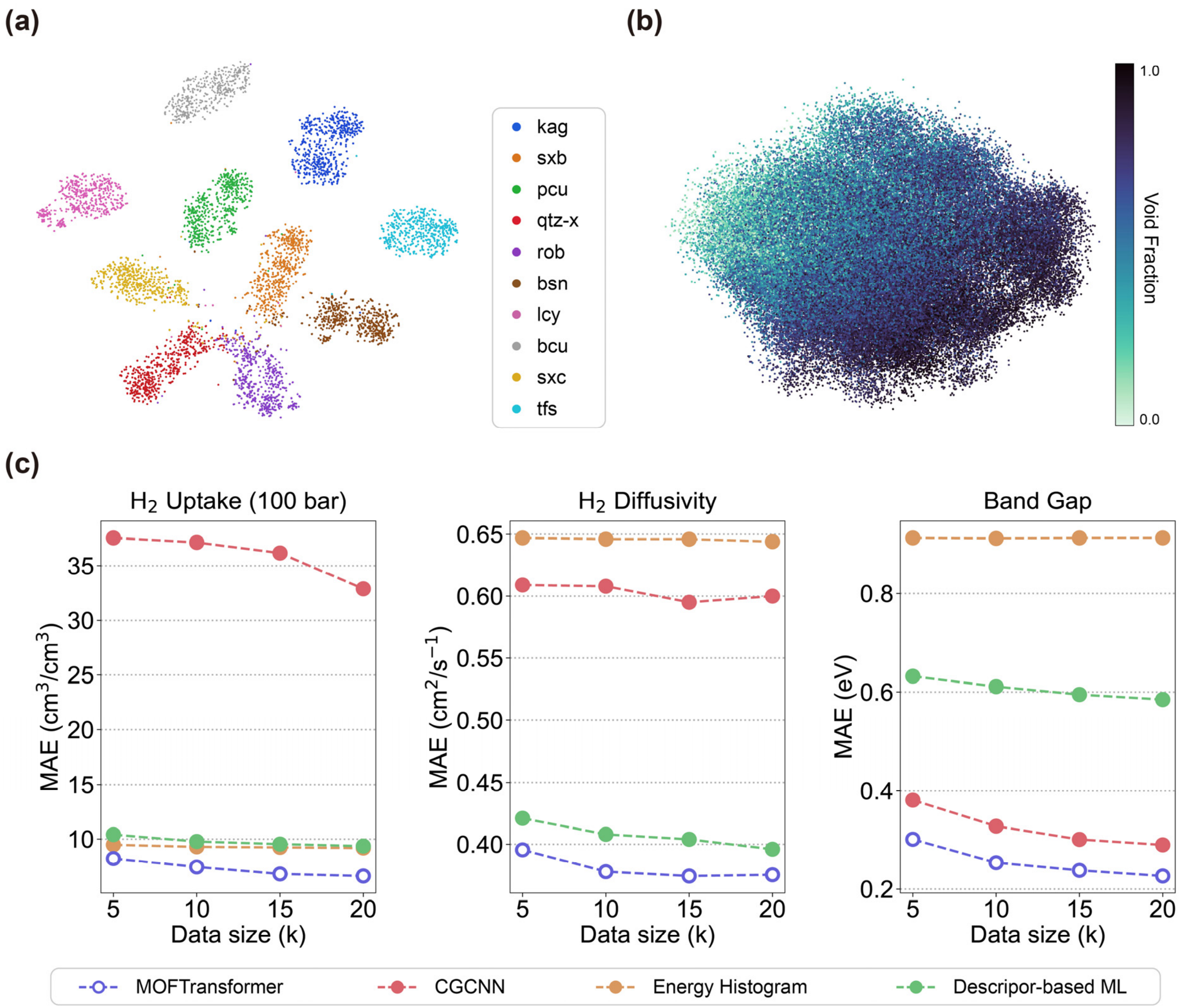

Figure 3. (a) For the top 10 frequently appearing topologies, the t-SNE plot embeds the class tokens of the pre-training model. (b) The class tokens of the pre-trained model are embedded by the PCA method, and a void fraction determines their colors. (c) Plots of MAE results of the fine-tuning model and three baseline models with datasets of H2 uptake, H2 diffusivity, and band gap according to dataset size from 5,000 to 20,000. The baseline models are machine learning models that were respectively used to predict gas uptake, diffusivity, and band gap values.

Fine-tuning Results

Figure 3(c) shows the fine-tuning results for predicting H_2 uptake (100 bar), H_2 diffusivity, and band gap, which were obtained from GCMC, MD, and DFT simulations, respectively. While 1 million hMOFs were used for the pre-training step, a relatively smaller number of MOFs (i.e., 5,000 to 20,000) were used for training during the fine-tuning stages. The performance of the fine-tuning is compared with the three baseline models (i.e., the energy histogram,[17] descriptor-based ML model,[18] and CGCNN[19,21]) as these have shown high performance in predicting gas uptake, diffusivity, and band gap, respectively. And from these comparisons, it can be seen that the MOFTransformer outperforms all of these other models, demonstrating both its superior performance as well as transferable capabilities. It is worth noting that the MatErials Graph Network (MEGNET)[39] outperforms the CGCNN in predicting the band gaps of MOFs[40]. The MEGNET utilizes global state attributes such as system temperature as well as atomic and bond attributes as inputs. However, graph network models like CGCNN and MEGNET may have difficulty in effectively predicting properties that rely on global features such as gas uptake and diffusivity for MOFs. This is due to the larger crystal system of MOFs, which is characterized by a larger number of atoms and defined by metal clusters and organic linkers as nodes and edges, respectively. As a result, the MOFTransformer exhibits strong performance in universal transfer learning for MOFs compared to graph network models. The ablation studies of the fine-tuning to demonstrate the effect of the data size on the pre-training tasks are explained in the Supplementary Note S2.

To demonstrate further transferability across different applications, the MOFTransformer was fine-tuned for various properties summarized in Table 1. Table 1 shows a performance comparison between our fine-tuned model and the machine-learning models used in other works. And it can be seen that the MOFTransformer model has either similar or higher performance (i.e., higher R^2 score or lower MAE) across all properties. In particular, it is worth noting the robustness of our model across different gas types, even though the probe molecule used to generate energy grids was CH_4. The reason is that overall shape of energy grids is relatively insensitive to the type of probe molecule which has little effect on our model to learn global features from energy-grid embeddings. In addition, the MOFTransformer can accurately predict properties at ambient condition, given that N_2, O_2 uptake and diffusivity were calculated at 1 bar and 298 K. Moreover, our model extends well to showcase lower MAE than the machine learning model using revised autocorrelations (RAC)[41] with geometric features as descriptors to predict solvent removal stability and thermal stability collected by text-mining. It is worth highlighting that our model showcases high performance in predicting the experimental data like text-minded data as well as the calculated properties. This result suggests that one can easily obtain high-performance structure-properties relationships by using our pre-trained model and fine-tuning it without needing to develop a new model from scratch.

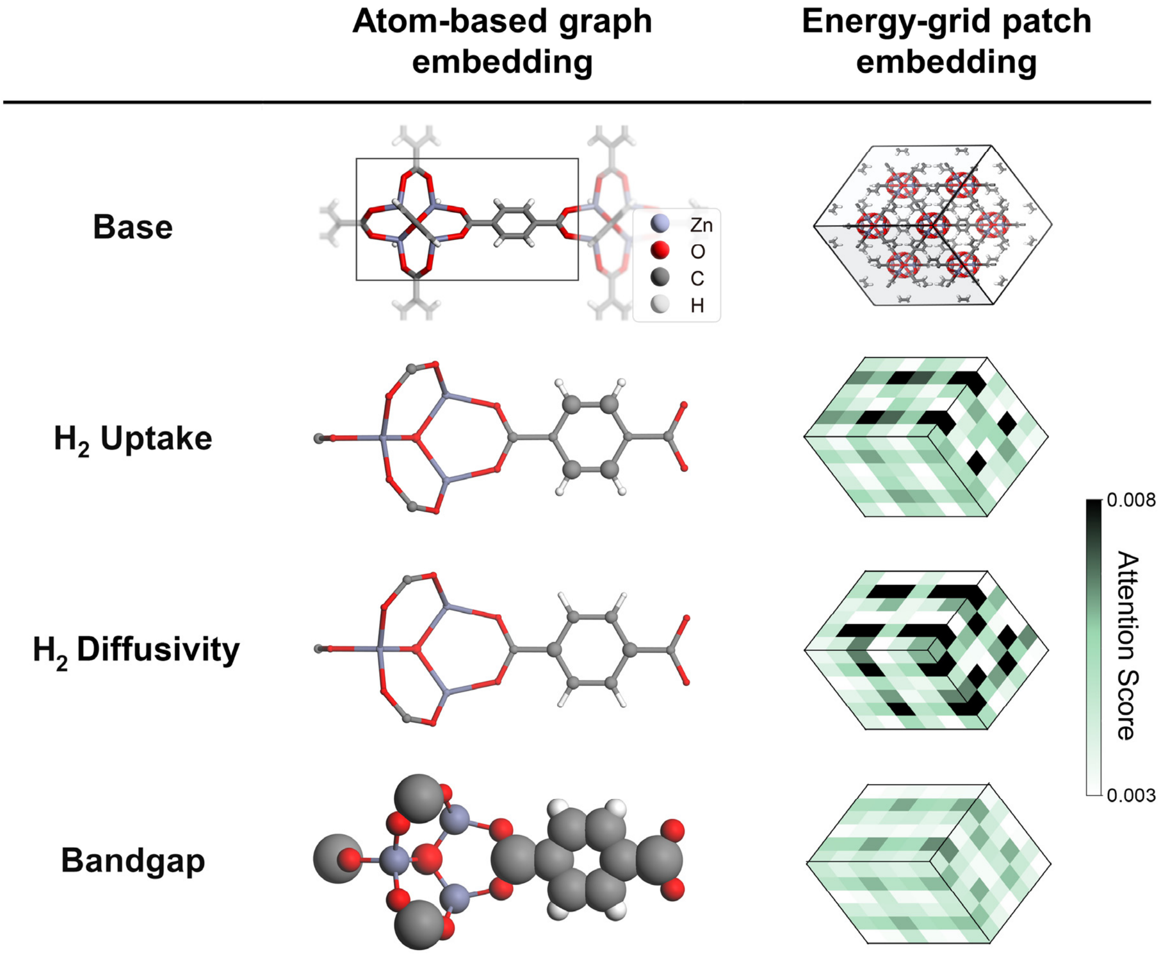

Figure 4. The schematics for attention score of atom-based embedding and energy-grid embedding in IRMOF-1. (left) Repeating building blocks models in IRMOF-1 with atomic size proportional to attention score. (right) Energy-grids that represent attention scores by color. The original form of the IRMOF-1 is shown in the "base."

Discussion

Apart from the universal transfer learning, feature importance and its interpretation can lead to a better understanding of the relationship between the MOF structures and their properties. Given that attention scores measure how much the model should pay attention to inputs when predicting desired properties, attention layers of the Transformer were assigned high attention scores to input features according to their importance. From the fine-tuning models that predict H_2 uptake, H_2 diffusivity, and band gap, feature importance analysis was implemented for IRMOF-1, which is one of the representative isoreticular MOFs. Figure 4 shows both the repeating building blocks models, which represent the metal cluster and organic linker, of IRMOF-1 (representing local features) and the 6×6×6 patches of energy-grids (representing global features). The sizes of atoms in the repeating building block models are scaled according to the attention scores obtained by the atom-based embeddings. And the colors of the patches are proportional to the attention scores obtained from the energy-grid embeddings. As can be seen from Figure 4, the atom-based embeddings are assigned with low attention scores (e.g. visualized by small atom sizes) when predicting H_2 uptake and diffusivity. On the other hand, the energy-grid embeddings are assigned with high attention scores, which is in accordance with the fact that H_2 uptake and diffusivity rely more on the global features. Meanwhile, for the band gap prediction, there is a reversal in trend as the atom-based graph embeddings have higher attention scores compared to energy-grid embeddings as the band gap is more dependent on the local features. The additional feature importance analysis for other properties (e.g. O_2 diffusivity and CO_2 Henry coefficient) were also conducted (see Supplementary Figure S8). Note that the feature importance analysis via attention scores is in line with previous findings and a chemist's intuition.

Beyond the case study of IRMOF-1, we implemented an in-depth analysis of feature importance for the atom-based graph and the energy-grid embeddings for band gap and H_2 uptake, respectively. Given that the band gap is defined by the difference between the conduction-band minimum (CBM) and the valance-band maximum (VBM), one might think that the atoms that exhibit strong peaks at the CBM and VBM play a critical role in determining its value. Interestingly, we identified that the atoms with peaks at the CBM and VBM strongly correlate with the atoms having high attention scores. Figure 5(a) shows the repeating building blocks models of IRMOF-1, 2, 3, and Ni-IRMOF-1 and their density of state (DOS) plots. IRMOF-2 and IRMOF-3 are variants of the IRMOF-1 structure with the BDC linker functionalized by $-Br$ and $-NH_2$. For IRMOF-2 and IMROF-3, the atoms that are part of the organic linkers (i.e., C, H, N, Br) have higher attention scores than those from the metal clusters (i.e., Zn, O). Consistent with these results, the atoms of the organic linker have peaks at the CBM and VBM compared to those of the metal clusters. Meanwhile, for the Ni-IRMOF-1 (which has Ni instead of Zn compared to the IRMOF-1), the atoms that belong to the metal cluster have higher attention scores and stronger peaks at the CBM and VBM compared to the organic linkers. These tendencies are consistent with other examples that were randomly selected in the QMOF database (see Supplementary Figure S9). Apart from these, we confirmed that the feature importance analysis could capture the underestimation of the band gap calculated by the PBE functional (see Supplementary Note S3). Hence, these results demonstrate that the fine-tuned model successfully learns the chemical features that are the more important when it comes to the band gap predictions.

When it comes to the energy-grid embeddings, one could argue that the patches located near the metal atoms have an important role on determining the gas uptake [42] Indeed, from the fine-tuned model to predict H_2 uptake, the 8 highest attention scores from the 6x6x6 energy-grid patches of IRMOF-1 are located near the metal atoms as shown in Figure 5(b). The metal atoms can make stronger bonding with adsorbates than other atoms such C, H, O, resulting in lower energy values for energy-grid patches near the metal atoms. Based on these observations, one can infer that the energy values of energy-grid patches can have an impact on determining attention score. Therefore, we plotted the relationship between the energy values of energy-grid points and the attention scores for each patch to further illustrate this relationship. The minimum energy values are normalized by their corresponding structure (or unit cell), which is represented on the y-axis of Figure 5(c). Figure 5(c) suggests that the energy-grid points with high attention scores tend to have relatively low energy values, as seen in the patches near the metal atoms. It is essential to highlight the fact that the scatter points within the high attention region (attention score > 0.008) exhibit a lower difference of energy than 20 kJ/mol.

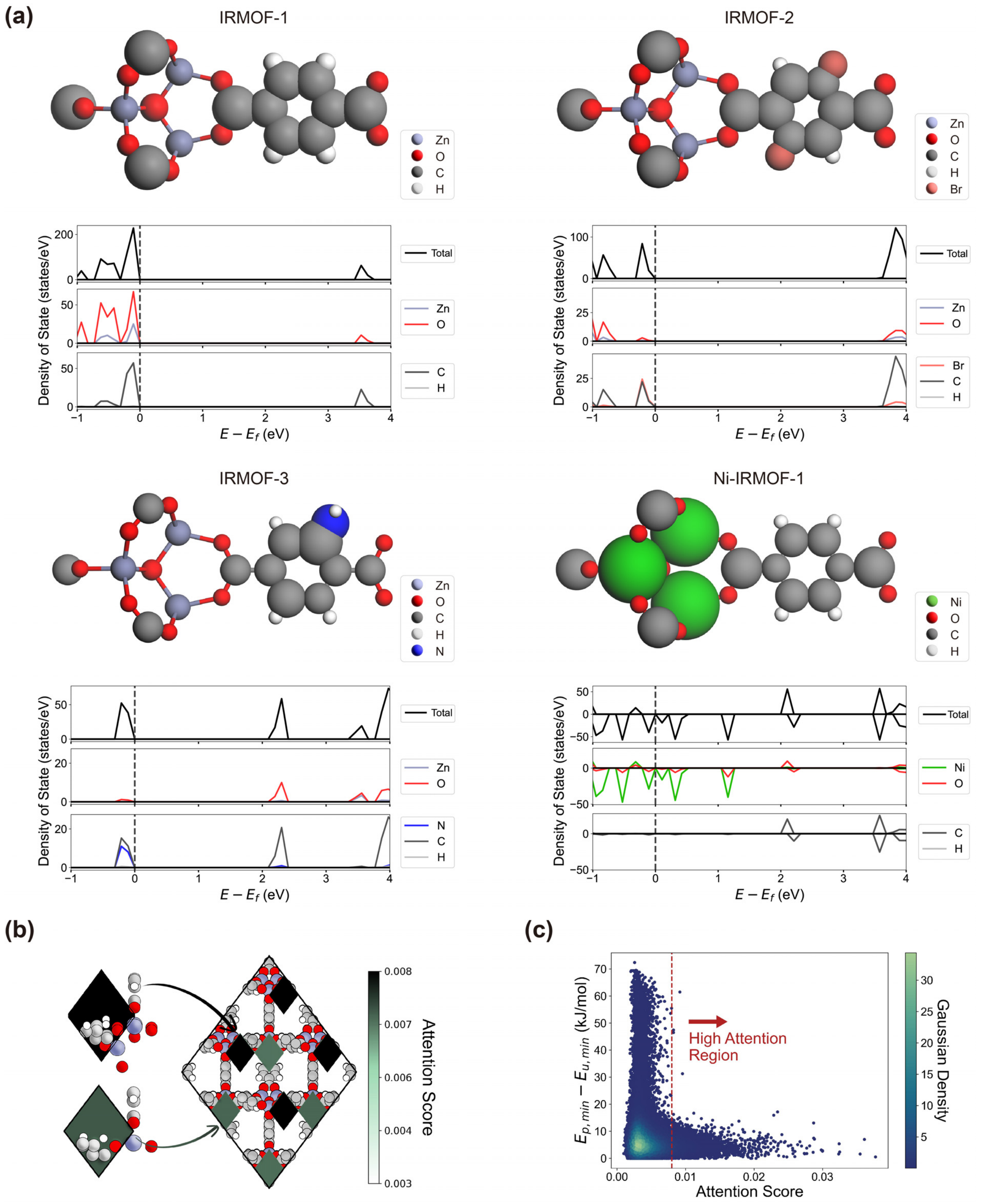

Figure 5. (a) Schematics of attention score for atom-based embedding, and density of state (DOS) plots for IRMOF-1, 2, 3, and Ni-IRMOF-1. The atomic sizes of repeating building blocks model

are proportional to the attention score. E means the energy, and E_f indicates the Fermi level. Positive and negative values of DOS indicate spin-up and spin-down channels, respectively. (b) Schematic of high attention score patches of energy-grid embedding for IRMOF-1. (c) Scattered plot for the difference of minimum energy between patch and unit cell according to energy-grid embedding. $E_{p,min}$ refers the minimum energy of the patch, and $E_{u,min}$ refers to the minimum energy of the unit cell. The red line (x = 0.008) distinguishes between high and low attention regions.

Conclusions

For the first time, we introduced a multi-modal pre-trained Transformer to capture both local and global features of MOFs. The model facilitates capturing the chemistry of metal nodes and organic linkers from the CGCNN and the information on geometric and topological features such as pore volume and topology from the energy grids. By fine-tuning the MOFTransformer model, our model outperforms all of the other state-of-the-art machine learning model across various different properties, showing its universal transferability as well as superior performance. Moreover, the model can provide insights by analyzing the feature importance from attention scores obtained from attention layers of the fine-tuned model. We believe that this model can be used as a bedrock model for other MOF researchers who wish to start their machine learning work and, as such, can help accelerate materials discovery and research within the field of porous materials.

Methods

Construction of hMOFs

The hMOFs used to train our MOFTransformer were constructed using PORMAKE,[11] a Python library that can generate MOFs by combining building blocks with different topologies. These building blocks and the topologies were obtained from ToBaCCo,[43] CoREMOF (with all of the solvents removed),[44] and RCSR database.[45] Altogether, 1 million and 20,000 hMOFs were generated for the pre-training, and fine-tuning dataset, respectively, and the details of building hMOFs are explained in Supplementary Note S4. All of the generated structures were geometrically optimized using the LAMMPS[46] package with the UFF force field.[36]

Computational details for molecular simulation

For the fine-tuning dataset, H_2 uptake and diffusivity (or diffusion coefficient) were selected to represent adsorption and diffusive properties. H_2 was selected to enable facile calculation while being different from the guest molecule (i.e., methane) used for the energy grid construction. The calculations were conducted using the RASPA package.[47] For the H_2 molecule, a united atom model was adopted. Also, the pseudo-Feynmna-Hibbs model was used to express the H_2 behavior at low temperature, which leads to fitting the Lenard-Jones (LJ) potentials to Feynman-Hibbs potential at T = 77 K.[48,49] Except for the H_2 molecules, the UFF force field was used with the Lorentz-Berthelot mixing rule and a cutoff distance of 12.8 Å.

For H_2 uptake calculation, the GCMC calculation was performed at 100 bar and 77 K for 10k production cycles with 5k cycles used for the initialization. Diffusivity (or diffusion coefficient) was calculated at infinite dilution at 77 K using the MD simulation. Given that the intermolecular interactions of the H_2 atoms are ignored for the infinite dilution simulation, it may sometimes lead to the initial configurations of the H_2 atoms captured within the small pores of MOFs. The initial configurations were obtained from the MC simulation without infinite dilution for 5k cycles to prevent this from happening. Then, the MD simulations were conducted by NVT ensemble with 1 fs time step.[18,50] The simulations were run for 3 million cycles, with 1k cycles used for the initialization and 10k cycles for equilibration. The guest molecules' mean-squared displacement (MSD) was computed every 10k cycles, and the diffusion coefficient was obtained using the slope of the MSD through Einstein's relation.[51]

Pre-training and Fine-tuning

In the pre-training step, AdamW[52] optimizer with a learning rate of 10^{-4} and weight decay of 10^{-2} was used in all three tasks. The model was trained with a batch size of 1,024 during 100 epochs. The pre-training dataset was randomly split into training, validation, test sets with the number of 800,000, 126,611, 100,000, respectively. The learning rate was warmed up during the first 5 % of the total epoch and then was linearly decayed to zero for the remaining epochs.

For fine-tuning, the MOFTransformer is trained to predict the desired properties with the model initialized by the converged weights from the pre-trained model. By adding a single dense layer to the class token, all model weights are fine-tuned to predict desired properties of MOFs. Given that the relatively small datasets are used during the fine-tuning step, the model was trained with a batch size of 32 during 20 epochs whose optimizer and learning rates are the same as those of the pre-training step. The fine-tuning dataset was randomly split into training, validation, test sets in a ratio of 0.8:0.1:0.1. For scaling the target properties, the standardization method was adopted. Training details of the three baseline models for comparison of the fine-tuning models are explained in Supplementary Note S5.

Conflicts of interest

There are no conflicts to declare.

Author Contributions

Y.K and H.P contributed equally to this work. Y.K and H.P developed MOFTransformer 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

Data used in this work are available via Figshare (10.6084/m9.figshare.21155506). It provides the pre-trained model and the atom-based graph embeddings and the energy-grid embeddings used as inputs of the MOFTransformer for CoREMOF, QMOF database .as well as fine-tuning data. In addition, The UFF-optimized CIF files of hypothetical MOFs used in this work are available via Figshare (10.6084/m9.figshare.21810147)

Code availability

The MOFTransformer library is available at https://github.com/hspark1212/MOFTransformer. Documents for the library is available at https://hspark1212.github.io/MOFTransformer which provides up-to-date documentation for pre-training, fine-tuning, and feature importance analysis with the MOFTransformer. For the sake of reproducibility, all results in this manuscript are obtained from a 1.0.1 version of MOFTransformer library, which is available at https://pypi.org/project/moftransformer/1.0.1.

Acknowledgements

H. P., Y. K., and J. K. acknowledge funding from National Research Foundation of Korea (NRF) under Project Number 2021M3A7C208974513. This work was supported by the National Supercomputing Center with supercomputing resources including technical support (KSC-2021-CRE-0460). BS is supported by the PrISMa Project, which is funded through the ACT programme (Accelerating CCS Technologies, Horizon2020 Project No 294766). Financial contributions made from: BEIS together with extra funding from NERC and EPSRC, UK; RCN, Norway; SFOE, Switzerland and US-DOE, USA, are gratefully acknowledged. Additional financial support from TOTAL and Equinor, is also gratefully acknowledged.

References

1 Freund, R. *et al.* The current status of MOF and COF applications. *Angewandte Chemie International Edition* **60**, 23975-24001 (2021).

2 Kumar, S. *et al.* Green synthesis of metal–organic frameworks: A state-of-the-art review of potential environmental and medical applications. *Coordination Chemistry Reviews* **420**, 213407 (2020).

3 Qian, Q. *et al.* MOF-based membranes for gas separations. *Chemical reviews* **120**, 8161-8266 (2020).

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

5 Deng, H. *et al.* Large-pore apertures in a series of metal-organic frameworks. *science* **336**, 1018-1023 (2012).

6 Ding, M., Cai, X. & Jiang, H.-L. Improving MOF stability: approaches and applications. *Chemical Science* **10**, 10209-10230 (2019).

7 Wang, C., Liu, D. & Lin, W. Metal–organic frameworks as a tunable platform for designing functional molecular materials. *Journal of the American Chemical Society* **135**, 13222-13234 (2013).

8 Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chemical Society Reviews* **43**, 5735-5749 (2014).

9 Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature* **576**, 253-256 (2019).

10 Daglar, H. & Keskin, S. Recent advances, opportunities, and challenges in high-throughput computational screening of MOFs for gas separations. *Coordination Chemistry Reviews* **422**, 213470 (2020).

11 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).

12 Altintas, C., Altundal, O. F., Keskin, S. & Yildirim, R. Machine learning meets with metal organic frameworks for gas storage and separation. *Journal of Chemical Information and Modeling* **61**, 2131-2146 (2021).

13 Chong, S., Lee, S., Kim, B. & Kim, J. Applications of machine learning in metal-organic frameworks. *Coordination Chemistry Reviews* **423**, 213487 (2020).

14 Ahmed, A. & Siegel, D. J. Predicting hydrogen storage in MOFs via machine learning. *Patterns* **2**, 100291 (2021).

15 Simon, C. M. *et al.* The materials genome in action: identifying the performance limits for methane storage. *Energy & Environmental Science* **8**, 1190-1199 (2015).

16 Lim, Y. & Kim, J. Application of transfer learning to predict diffusion properties in metal–organic frameworks. *Molecular Systems Design & Engineering* (2022).

17 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).

18 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).

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

20 Ma, R., Colon, Y. J. & Luo, T. Transfer learning study of gas adsorption in metal–organic frameworks. *ACS applied materials & interfaces* **12**, 34041-34048 (2020).

21 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).

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

23 Nandy, A. *et al.* MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Scientific Data* **9**, 1-11 (2022).

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

25 Lim, Y., Park, J., Lee, S. & Kim, J. Finely tuned inverse design of metal–organic frameworks with user-desired Xe/Kr selectivity. *Journal of Materials Chemistry A* **9**, 21175-21183 (2021).

26 Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **149**, 134-141 (2012).

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

28 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

29 Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

30 Hu, R. & Singh, A. in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1439-1449.

31 Zhou, L. *et al.* in *Proceedings of the AAAI Conference on Artificial Intelligence.* 13041-13049.

32 Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J. & Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

33 Kim, W., Son, B. & Kim, I. in *International Conference on Machine Learning.* 5583-5594 (PMLR).

34 Cao, Z., Magar, R., Wang, Y. & Farimani, A. B. MOFormer: Self-Supervised Transformer model for Metal-Organic Framework Property Prediction. *arXiv preprint arXiv:2210.14188* (2022).

35 Chen, P., Jiao, R., Liu, J., Liu, Y. & Lu, Y. Interpretable Graph Transformer Network for Predicting Adsorption Isotherms of Metal–Organic Frameworks. *Journal of Chemical Information and Modeling* **62**, 5446-5456 (2022).

36 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).

37 Martin, M. G. & Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *The Journal of Physical Chemistry B* **102**, 2569-2577 (1998).

38 Bucior, B. J. *et al.* Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design* **19**, 6682-6697 (2019).

39 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).

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 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).

42 Koizumi, K., Nobusada, K. & Boero, M. Hydrogen storage mechanism and diffusion in metal–organic frameworks. *Physical Chemistry Chemical Physics* **21**, 7756-7764 (2019).

43 Colón, Y. J., Gomez-Gualdron, D. A. & Snurr, R. Q. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Crystal Growth & Design* **17**, 5801-5810 (2017).

44 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).

45 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).

46 Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **117**, 1-19 (1995).

47 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).

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

49 Fischer, M., Hoffmann, F. & Fröba, M. Preferred hydrogen adsorption sites in various MOFs—a comparative computational study. *ChemPhysChem* **10**, 2647-2657 (2009).

50 Daglar, H., Erucar, I. & Keskin, S. Exploring the performance limits of MOF/polymer MMMs for O2/N2 separation using computational screening. *Journal of Membrane Science* **618**, 118555 (2021).

51 Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der physik* **369**, 253-287 (1921).

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

Table 1. A table of fine-tuning results with the publicly accessible databases of MOFs that include the properties calculated by GCMC, MD, and even text-mining data. The results of the machine learning models used in the paper on the databases are summarized to compare the performance.

Property	MOFTransformer	Original paper	Number of data	Remarks	Ref
N_2 uptake	R2 : 0.78	R2 : 0.71	5,286	CoREMOF	18
O_2 uptake	R2 : 0.83	R2 : 0.74	5,286	CoREMOF	18
N_2 diffusivity	R2 : 0.77	R2 : 0.76	5,286	CoREMOF	18
O_2 diffusivity	R2 : 0.78	R2 : 0.74	5,286	CoREMOF	18
CO_2 henry coefficient	MAE : 0.30	MAE : 0.42	8,183	CoREMOF	22
Solvent removal stability classification	ACC : 0.76	ACC : 0.76	2,148	Text-mining data	40
Thermal stability regression	R2 : 0.44 (MAE : 45°C)	R2 : 0.46 (MAE : 44°C)	3,098	Text-mining data	40