Metal-organic frameworks meet Uni-MOF: a transformer-based gas adsorption detector

Jingqi Wang[1,2,+], Jiapeng Liu[3,4,+], Hongshuai Wang[2,5], Guolin Ke[2], Linfeng Zhang[2,4], Jianzhong Wu[6,*], Zhifeng Gao[2,*], and Diannan Lu[1,*]

[1]Department of Chemical Engineering, Tsinghua University, Beijing, 100084, China
[2]DP Technology, Beijing, 100080, China
[3]School of Advanced Energy, Sun Yat-Sen University, Shenzhen, 518107, China
[4]AI for Science Institute, Beijing, 100190, China
[5]Jiangsu Key Laboratory for Carbon-Based Functional & Materials Devices, Institute of Functional & Nano Soft Materials (FUNSOM), Soochow University, Suzhou, 215123, China
[6]Department of Chemical and Environmental Engineering, University of California, Riverside, CA, 92521, United States
[*]corresponding author: ludiannan@mail.tsinghua.edu.cn; gaozf@dp.tech; jwu@engr.ucr.edu
[+]these authors contributed equally to this work

ABSTRACT

Gas separation is crucial for industrial production and environmental protection, with metal-organic frameworks(MOFs) offering a promising solution due to their tunable structural properties and chemical compositions. Traditional simulation approaches, such as molecular dynamics, are complex and computationally demanding. Although feature engineering-based machine learning methods perform better, they are susceptible to overfitting because of limited labeled data. Furthermore, these methods are typically designed for single tasks, such as predicting gas adsorption capacity under specific conditions, which restricts the utilization of comprehensive datasets including all adsorption capacities. To address these challenges, we propose Uni-MOF, an innovative framework for large-scale, three-dimensional MOF representation learning, designed for universal multi-gas prediction. Specifically, Uni-MOF serves as a versatile "gas adsorption detector" for MOF materials, employing pure three-dimensional representations learned from over 631,000 collected MOF and COF structures. Our experimental results show that Uni-MOF can automatically extract structural representations and predict adsorption capacities under various operating conditions using a single model. For simulated data, Uni-MOF exhibits remarkably high predictive accuracy across all datasets. Impressively, the values predicted by Uni-MOF correspond with the outcomes of adsorption experiments. Furthermore, Uni-MOF demonstrates considerable potential for broad applicability in predicting a wide array of other properties.

Introduction

Gas separation[1–3] is a significant industrial challenge that requires immediate attention, given its critical role in various applications. For examples, separating CH_4 from CO_2 is essential for obtaining high-quality natural gas and effectively achieving the carbon capture, utilization and storage for environmental reasons[4,5]. Gas separation also has implications for other fields, such as the production of high-purity oxygen[6,7] and nitrogen[8,9] for industrial purposes and the purification of noble gases for medical diagnosis and lasers[10,11]. Given its importance, research in this area is crucial for advancing technology and meeting industry demands.

Metal-organic frameworks (MOFs) have emerged as a kind of promising material in the field of gas separation due to their unique properties[12–14]. MOFs are composed of metal ions and organic ligands, which provide them with highly ordered pore structures and adjustable aperture sizes. These properties make them ideal for various gas separation applications[15–17]. The ability to control the pore size and chemical composition of MOFs allows for selective adsorption and separation of different gases. MOFs with different pore sizes exhibit varied capacities for gas adsorption[18], and tuning the chemical composition[19] can affect the preference for adsorbate gases. The ability to selectively adsorb and separate different gases makes MOFs a promising material for various industrial and environmental applications[20,21].

While the potential of MOFs for gas adsorption is promising, accurately predicting their adsorption capacity remains a challenge. Molecular dynamics (MD)[22,23], Monte Carlo (MC)[24] and other simulation/calculation methods[25,26] have been applied to provide reference values, but these approaches are computationally expensive and complicated for implementation, limiting their application to large-scale, multi-gas and high-throughput calculations. Moreover, the vast range of operating conditions for gas adsorption further complicates the predictions.

Machine learning techniques have demonstrated significant potential in accurately predicting properties of crystalline materials[27–29], reducing the cost of traditional trial-and-error experiments, and eliminating the need for expensive simulations. However, these methods often rely on *ad hoc* feature engineering based on expert domain knowledge, leading to overfitting and biased performance when using a limited amount of labeled data. With the emergence of deep learning, graph neural networks (GNNs)[30,31] and Transformers[32–34] have proven successfully in predicting MOF properties. These models directly incorporate structural information such as chemical bonds, atoms, and spatial coordinates as inputs, and automatically learn structural features through data-driven approaches. Thanks to the powerful representation capacity of these deep-learning frameworks, learned features can effectively eliminate biases introduced by feature engineering mentioned earlier.

Despite their high performance and powerful predictive capabilities, existing models for predicting adsorption properties are typically designed for single tasks, specifically predicting the adsorption uptake of a particular gas under certain conditions. However, the available datasets for these single task predictions are often limited, thereby hindering the models' generalizability and full utilization of their capabilities. On the other hand, the combination of labeled data from various adsorbate gases across different temperature and pressure environments can create a substantial dataset suitable for training across the entire working conditions. The increased data size may also enhance the models' ability to generalize and improve their practical industrial use. Therefore, a unified adsorption framework is necessary for advancing these models. Additionally, integrating representation learning (or pre-training) for large-scale unlabeled MOF structures may further improve the model's performance as well as representation ability. The pre-training trick has been widely implemented in combination with large-scale models to discover new drugs[35], where pure three-dimensional molecular structures were used to pre-train these models. Experimental results have also demonstrated that pre-trained models outperform previous methods, particularly in property prediction[33], suggesting remarkable improvement through pre-training.

Inspired by this, we propose the Uni-MOF framework as a universal solution for predicting gas adsorption of MOFs under different conditions using structural representation learning. To our best knowledge, Uni-MOF is the first framework of its kind and acts as a comprehensive "gas adsorption detector" for MOF materials. Our framework is easy to use and allows for module selection. Additionally, it effectively addresses the issue of overfitting by integrating various cross-system absorption labeled data with representation learning from massive amounts of unlabeled structural data. This compensates for the lack of high-quality and insufficient data, ultimately leading to higher accuracy in gas adsorption predictions. Our study utilized a self-supervised learning approach on a database containing over 631,000 MOF and COF structures. The results were remarkable, demonstrating a high prediction accuracy. Fine-tuning experiments revealed that the Uni-MOF framework is robust in databases with ample data. When applied to databases with sufficient sampling of working conditions, our Uni-MOF framework is able to screened high-performance adsorbents under high pressure accurately by feeding only the labeled data obtained at low pressure by home-brew simulations. We must stress that Uni-MOF provides a convenient approach for high pressure adsorption capacities, which are generally more computationally demanding for traditional simulations. The results are consistent with experimental screening outcomes. Furthermore, the performance of Uni-MOF on cross-system datasets exceeded that on single-system tasks. By leveraging support from other gas adsorption data, Uni-MOF accurately predicted the adsorption properties of unknown gases. Extensive pre-training on three-dimensional structures enabled Uni-MOF to effectively learn MOF structures. T-distributed stochastic neighbor embedding(t-SNE) analysis confirmed that the fine-tuning stage can further learn structural features and effectively identify structures with distinct adsorption behaviors. This indicates a strong correlation between learned representations and the target values of gas adsorption.

Overview

The Uni-MOF framework comprises pre-training on three-dimensional nanoporous crystals and fine-tuning for multi-task prediction in downstream applications. Figure 1 provides a schematic representation of Uni-MOF framework. The pre-training of three-dimensional crystal materials significantly enhances the prediction performance of downstream tasks, particularly for large-scale unlabeled data. To address the issue of inadequate supervised training datasets, we collected an extensive dataset of MOF structures and generated over 300,000 MOFs using ToBaCCo.3.0[36,37]. Similar to the masking tagging task in BERT[38] and Uni-Mol[35], Uni-MOF employs a prediction task for masked atoms, thereby promoting the pre-trained models to acquire an in-depth understanding of the materials' spatial structures. To enhance the robustness of pre-training and generalize the learned representation, we introduced noises to the original coordinates of MOFs, as depicted in Figure 1 **left**. In the pre-training stage, we devised two tasks: 1) reconstructing the pristine three-dimensional positions from the noisy data, and 2) predicting the masked atoms. These tasks can augment the model's robustness and improve downstream prediction performance.

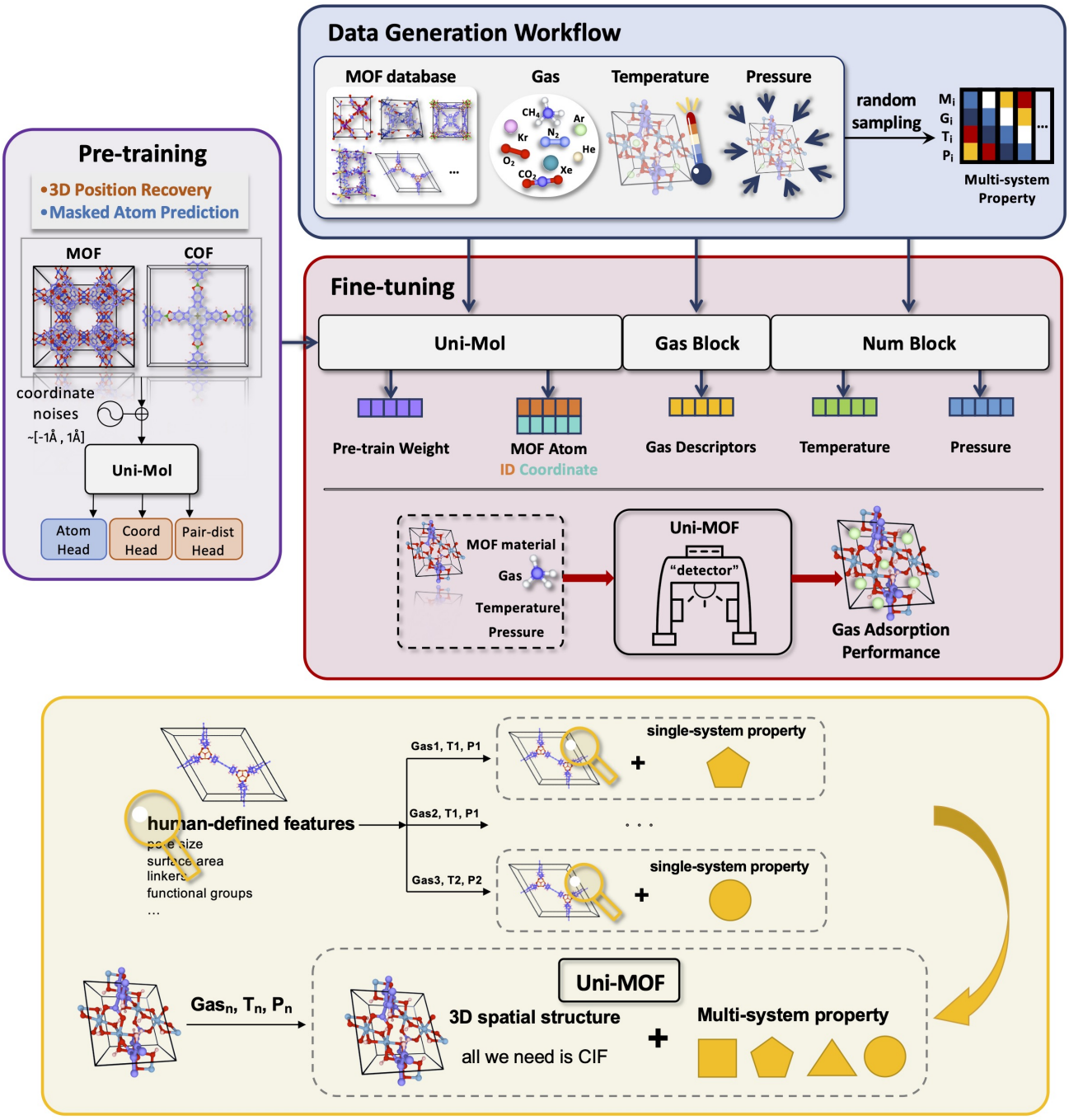

Figure 1. Schematic overview of Uni-MOF framework. Left: pre-training workflow. In the pre-training stage, in addition to predicting the types of masked atoms, a three-dimensional position denoising task was used to learn the three-dimensional spatial representation. Uniform noise of [-1 Å, 1 Å] is added to the 15% of atomic coordinates randomly, and then the spatial position encoding is calculated based on the corrupted coordinates. **Upper right:** data generation workflow. Cross-system performance datasets can be collected or generated by random sampling of different operating conditions. **Center right:** workflow of Uni-MOF fine-tuning. A unified gas adsorption prediction model Uni-MOF is built by the embedding of pre-trained weight, MOF material, gas, temperature and pressure. **Bottom right:** overall workflow of Uni-MOF. For the universal Uni-MOF framework, no additional analytical calculations for materials are required, and the properties under varied working conditions can be predicted based solely on the crystallographic information file (CIF) of MOF materials.

In addition to diverse spatial configurations, a comprehensive set of material property data points is also crucial for model training. To enrich the dataset, we established a custom data generation process (as illustrated in Figure 1 **upper right**). For

example, we utilized the CoRE MOF database that comprises successfully synthesized MOFs, gases that are significant in the separation field, as well as the common temperature and pressure operating range under the corresponding system. By randomly sampling from these various materials, gases, temperatures, and pressure pools, a significant volume of adsorption uptake data can be generated for Uni-MOF fine-tuning. This data generation process improves the efficiency of data generation and can form a widely sampled dataset, Table 1 lists all the databases applied in this study. The simulation-derived database with diversity is beneficial for model fine-tuning and optimization, ultimately accomplishing the objective of virtual screening for material performance.

The fine-tuning of Uni-MOF depicted in Figure 1 **center right** is based on the extraction of representations acquired through pre-training, as well as the generation and collection of extensive datasets using our home-brew workflows. During the fine-tuning process, we trained the model using over 631,000 MOFs and COFs with labels across various adsorption conditions, enabling accurate prediction of adsorption capacities. With the diverse database of cross-system targeted data, the fine-tuned Uni-MOF can predict the multi-system adsorption property of MOFs under arbitrary states, including different gases, temperatures, and pressures. As a result, Uni-MOF is a unified and readily available framework for predicting adsorption performance of MOF adsorbents.

Above all, Uni-MOF obviates the requirement for additional labor in identifying human-defined structural features. Instead, the crystallographic information file (CIF) of MOFs, along with pertinent gas, temperature, and pressure parameters, suffices. The self-supervised learning strategy and abundant databases ensure that Uni-MOF can foretell nanoporous material properties in a wide range of operating parameters, thereby rendering it a proficient "gas adsorption detector" for MOF materials.

Uni-MOF: A Universal "Gas Adsorption Detector" for MOFs

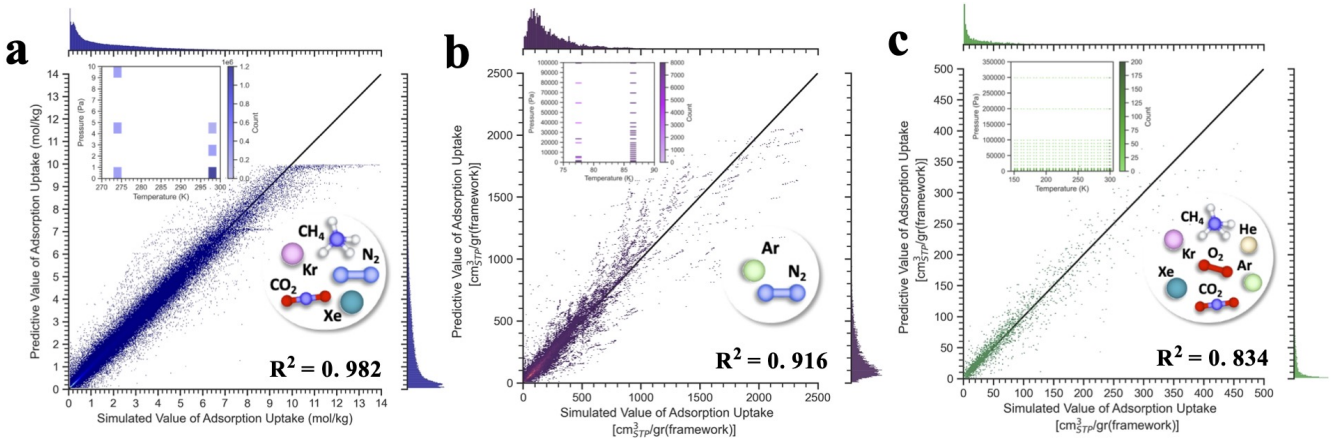

Figure 2. The correlation between predicted and simulated value of gas adsorption amount for **a**, Database of hMOF_MOFX_DB (mol/kg), **b**, Database of CoRE_MOFX_DB (cm^3_{STP}/g) and **c**, Database of CoRE_MAP_DB (cm^3_{STP}/g). Sub-figure is the distribution of temperature (K) and pressure (Pa) for each database.

In order to evaluate the predictive capability of Uni-MOF as a comprehensive framework for adsorption performance prediction, two mixed-state databases for gas adsorption, namely hMOF_MOFX-DB and CoRE_MOFX-DB, were compiled with adsorbate gases consisting of [CO_2, N_2, CH_4, Kr, Xe] and [N_2, Ar], respectively. In addition, the CoRE_MAP_DB database was generated via our home-brew Monte Carlo simulation workflow for adsorption uptake of seven gases (CO_2, CH_4, Ar, Kr, Xe, O_2, He).

Since the data sources of the three databases are different, we conducted model training for each database separately in order to ensure the consistency of data sets, details of these three databases, including temperature and pressure ranges, are listed in Table S1. To prevent data bias and ensure that the test set remained unseen by the model, we divided the data set into 8:1:1 according to MOF structures instead of randomly splitting. The model with the highest validation set R^2 (coefficient of determination) was saved as the optimal model, and the final R^2 was reported using the weights of this model to reasonably avoid over-fitting.

The collected datasets, hMOF_MOFX_DB and CoRE_MOFX_DB, exhibit relatively concentrated temperature and pressure distributions, as depicted in Figure 2 **a,b**. Notably, both databases offer adequate data, with over 2,000,000 and 400,000 data points, respectively. The prediction results demonstrate that Uni-MOF is remarkably robust when applied to databases that possess sufficient data with relatively concentrated operating states, such as hMOF_MOFX_DB and CoRE_MOFX_DB, with R^2 values of 0.98 and 0.92, respectively.

In contrast, our CoRE_MAP_DB database, which we have learned from Table S1, contains slightly more than 10,000 data points. It encompasses an extensive sampling of the adsorption of seven adsorbed gases in over 12,000 MOFs, covering a temperature range of 150-300K and a pressure range of 1Pa-3bar, as depicted in the sub-figure of Figure 2 **c**. For such a widely distributed database, Uni-MOF can still achieve excellent prediction accuracy with an R^2 value of 0.83, demonstrating its good generalizability.

Experimental adsorption uptake prediction

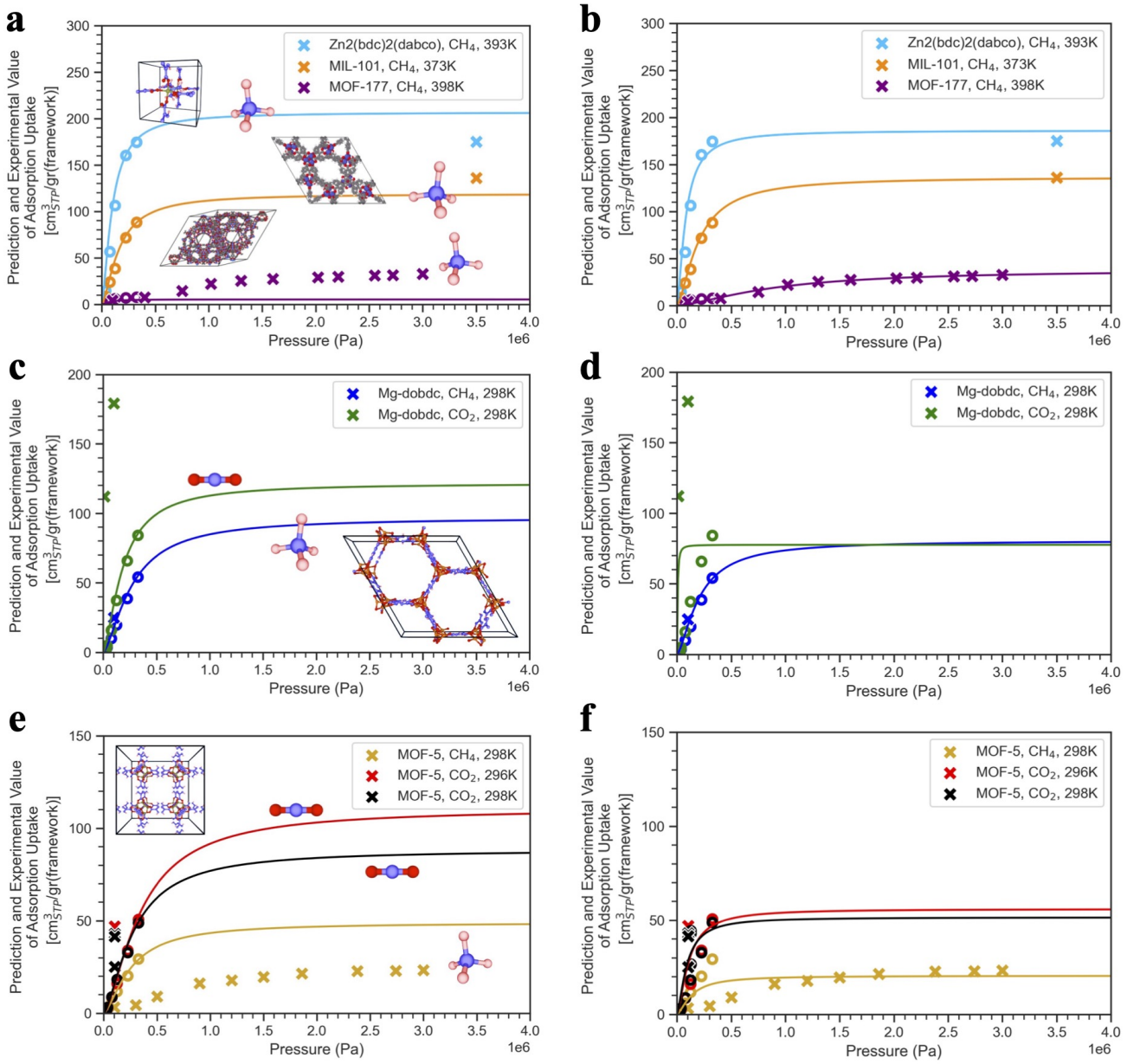

Figure 3. Langmiur adsorption isotherms based on low-pressure predictions and high-pressure experimental values. **a** Uni-MOF predicted and **b** experimentally corrected adsorption isotherms of methane adsorption in Zn2(bdc)2(dabco), MIL-101 and MOF-177 at 393K, 373K and 398K, respectively; **c** Uni-MOF predicted and **d** experimentally corrected adsorption isotherms of methane and carbon-dioxide adsorption in Mg-dobdc at 298K; **e** Uni-MOF predicted and **f** experimentally corrected adsorption isotherms of methane and carbon-dioxide adsorption in MOF-5 at 298K and 296K. The hollow dot represents the predicted value of Uni-MOF, and the cross dot represents the experimental data referenced from previous literature. Different colors represent different adsorption operating conditions.

Despite the excellent prediction performance of Uni-MOF on simulated results, we are wondering how the framework would behave in comparison to the real experimentally collected results. In this study, we have chosen commonly used laboratory materials, such as MOF-5 and MOF-177,[39–42], and compared the predicted gas adsorption capacity with existing experimental data. Considering that CoRE_MAP_DB database contains a diverse range of operating conditions and experimentally validated MOF structures, we chose the weights trained using this database to predict the adsorption performance of experimental materials. Figure 3 displays the predicted results under different conditions, and Table S3 provides detailed information.

Figure 3 **a** presents the Langmuir adsorption isotherm[43] obtained by fitting the predicted methane adsorption capacity under low pressure (less than 5 bar). It shows that the high pressure adsorption capacity displayed by the Langmuir adsorption isotherm is consistent with the experimental values. Specifically, the high pressure adsorption capacity of (Zn2(bdc)2(dabco), methane, 393K) > (MIL-101, methane, 373K) > (MOF-177, methane, 398K). This suggests that Uni-MOF framework is capable of accurately screening high performance adsorbents under high pressure based solely on prediction adsorption capacity under low pressure. Furthermore, the experimental values are introduced to correct the adsorption isotherms. In Figure 3 **b**, one can observe that the corrected adsorption isotherms have a strong correlation with experimental adsorption capacity to some extent. The results exhibit that Uni-MOF not only has the ability to screen the adsorption performance of the same gas in different materials but also can accurately screen the adsorption performance of different gases in the same material (Figure 3 **c, d**) or at different temperatures (Figure 3 **e, f**).

In the foreseeable future, the intersection of Artificial Intelligence(AI) and materials science will necessitate the resolution of practical and scientific issues. Nonetheless, the attainment of process implementation by AI in the realm of machine learning techniques that entail copious amounts of data remains a formidable challenge, given the dearth of experimental data and the diverse array of synthetic technology and characterization conditions implicated. Our research has made a significant stride in materials science by incorporating operating conditions into the Uni-MOF framework to ensure data adequacy and enable screening functions that are consistent with experimental findings.

Cross-system forecasting

In order to showcase the predictive capabilities of Uni-MOF with regards to cross-system properties, five materials were randomly selected from each of the six systems (carbon-dioxide at 298K, methane at 298K, krypton at 273K, xenon at 273K, nitrogen at 77K and argon at 87K) contained in databases hMOF_MOFX_DB and CoRE_MOFX_DB, which have been thoroughly sampled in terms of temperature and pressure. The predicted and simulated values of gas adsorption uptake at varying pressures were then compared, with the results presented in Figure 4 **a**-**f**. It is evident that, due to the fact that the adsorption isotherms were obtained purely through simulated values, the predicted values of adsorption uptake generated by Uni-MOF for the hMOF_MOFX_DB and CoRE_MOFX_DB databases align closely with the simulated values across all cases. This finding is further supported by the high prediction accuracy demonstrated in Figure 2 **a** and **b**.

Given the ability of Uni-MOF to predict properties across systems, we were intrigued by its potential to forecast the adsorption capacity of unknown gases. The CoRE_MAP_DB database contains a diverse array of adsorption data points for various gases, such as methane, carbon dioxide, argon, krypton, xenon, oxygen, and nitrogen. To evaluate the predictive capability of Uni-MOF, we divided the CoRE_MAP_DB data points by adsorbate gas and predicted the adsorption capacity for each gas separately. The resulting predictions are depicted in Figure 4 **g** and summarized in Table S6. Remarkably, Uni-MOF demonstrated robustness in predicting the adsorption capacity of unknown gases, achieving a high prediction accuracy (R^2) of 0.85 for krypton and a prediction accuracy above 0.35 for all unknown gases.

Unlike the prediction of adsorption between different materials, the prediction of adsorption behavior for unknown gases is a formidable challenge. Variations in molecular size, surface adsorption energy, and inter-molecular forces among different gas types have a significant impact on the adsorption mechanism and behavior, as illustrated in Figure 4 **h**. Despite this complexity, Uni-MOF exhibits exceptional generalizability, as evidenced by its ability to accurately predict the adsorption properties of unknown gases with only the support of adsorption data from other gases.

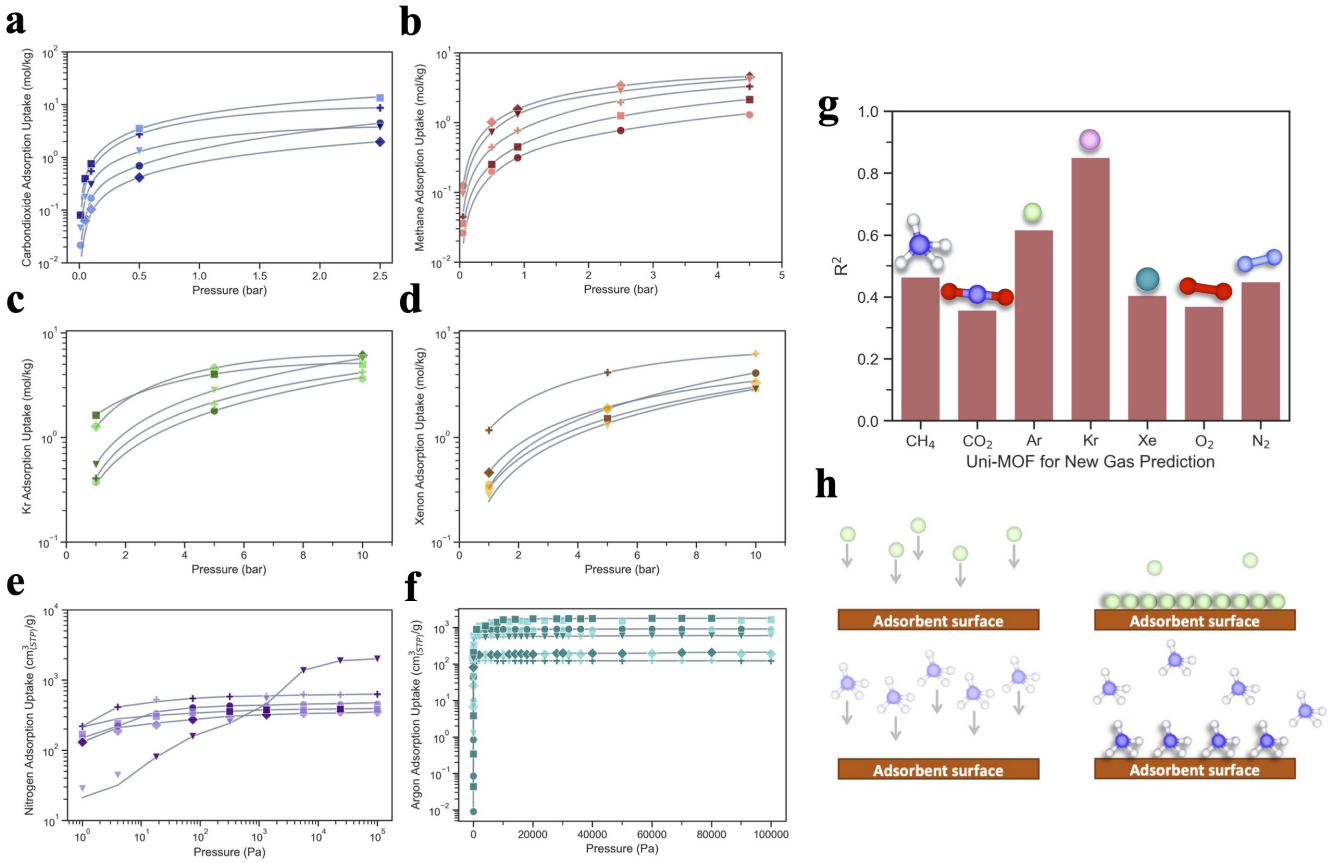

Figure 4. Uni-MOF cross-system prediction cases. The predicted and simulated values of gas adsorption amount versus pressure in CoRE_MOFX_DB database for **a** Carbon-dioxide at 298K, **b** Methane at 298K, **c** Krypton at 273K and **d** Xenon at 273K, and in hMOF_MOFX_DB database for **e** Nitrogen at 77K and **f** Argon at 87K. The lighter color dot represents the predicted adsorption uptake of Uni-MOF, and the darker color dot represents the simulated adsorption uptake in database. The adsorption isotherms are obtained only by simulated value. **g** Prediction results of adsorption uptake with the division of dataset according to adsorbate gases. **h** Schematic diagram of gas adsorption mechanism.

Uni-MOF: A Universal Framework applying Pre-training

In order to verify that the self-supervised learning strategy of pre-training can effectively improve the robustness and downstream prediction performance of the Uni-MOF, we established and compared Uni-MOF w/o pre-training against Uni-MOF on three databases, namely CoRE_MOFX_DB, hMOF_MOFX_DB, and CoRE_MAP_DB. The CoRE_MOFX_DB and hMOF_MOFX_DB databases exhibit a high degree of data concentration, thereby enabling their partitioning into smaller databases containing adsorption data for various materials under identical working conditions (i.e., same gas, temperature, and pressure). We trained both CoRE_MOFX_DB and hMOF_MOFX_DB databases using Uni-MOF and Uni-MOF w/o pre-training, and subsequently calculated the coefficients of determination for each small dataset. In addition, each small dataset is trained separately using the Uni-MOF framework thus to derive the corresponding coefficient of determination.

As shown in Figure 5 **a**, the green dots represent the predictive performance of the whole CoRE_MOFX_DB database. It can be seen that the Uni-MOF performs better than Uni-MOF w/o pre-training. The self-supervised learning strategy of pre-training allows the model to learn the three-dimensional configuration of nanoporous materials in depth, thus improving the accuracy of model fine-tuning. No matter Uni-MOF or Uni-MOF w/o pre-training, in most cases, the fine-tuning based on the whole CoRE_MOFX_DB database have greater performance on the whole database than on small data sets, demonstrating the predictive capacity of Uni-MOF on cross-system properties. For fine-tuning of single-system tasks, that is, training individually for each small data set, the predicted performance hardly exceeded the performance of Uni-MOF. Thus, we know that far-ranging data sampling can further promote the prediction capacity of learning model. The same conclusions can be summarized for hMOF_MOFX_DB from Figure 5 **b**. However, due to the scattered sampling of CoRE_MAP_DB database, limited small data sets cannot be divided following the same procedure as previous two databases. Therefore, the correlations between predicted and simulated values from Uni-MOF and Uni-MOF w/o pre-training are compared, shown in Figure 5 **c**.

Similarly, compared with Uni-MOF w/o pre-training, the performance of Uni-MOF was improved from 0.70 to 0.83, which further proved the significance of pre-training strategy for Uni-MOF.

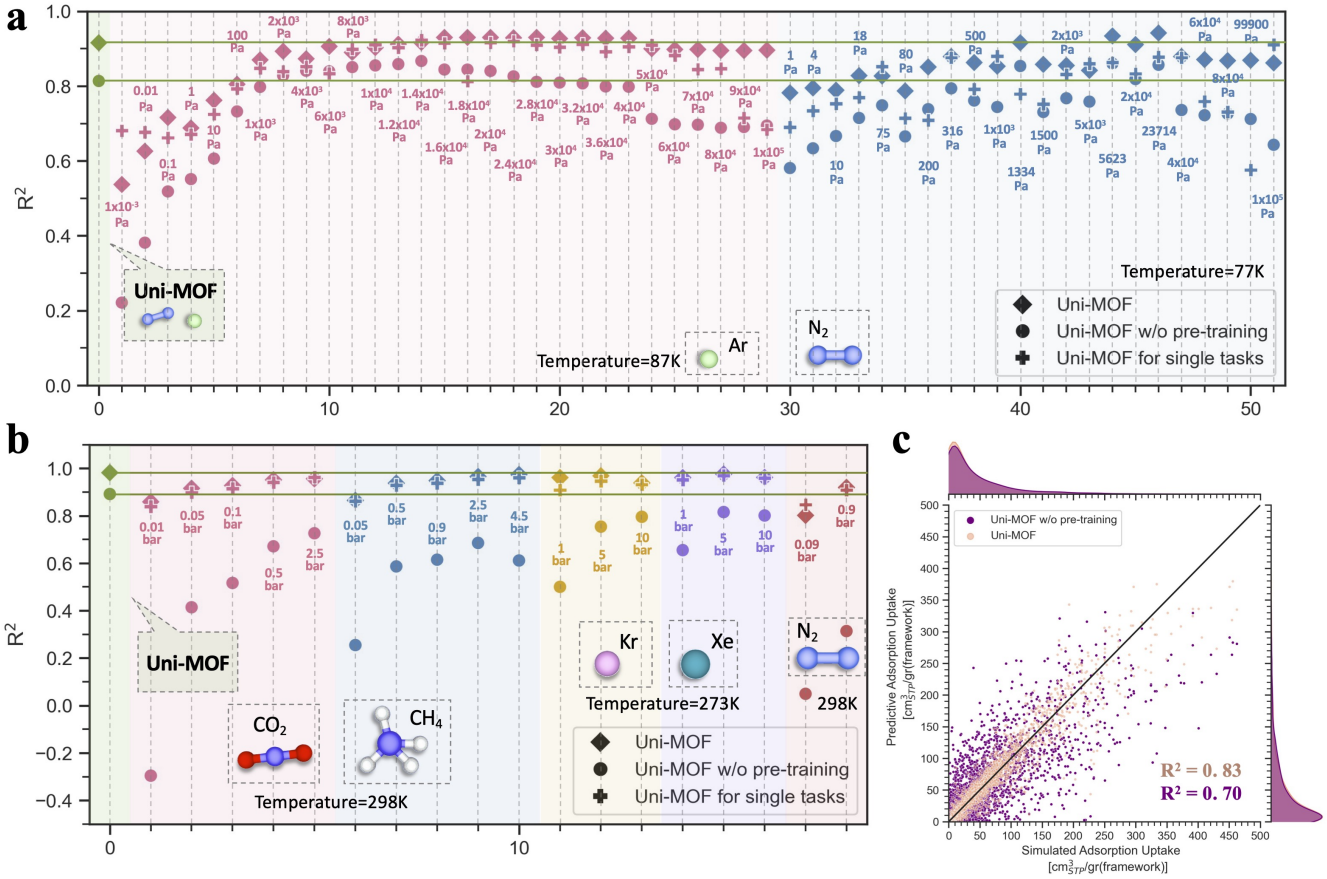

Figure 5. The performance comparison of Uni-MOF and Uni-MOF w/o pre-training in **a** CoRE_MOFX_DB and **b** hMOF_MOFX_DB. The green color represents the prediction performance of the entire database. Other colors represent prediction performance of sub-dataset, with each color for one sub-dataset with identical gas and temperature and varied pressures. **c** The comparison of correlation between predicted and simulated value of gas adsorption amount via Uni-MOF and Uni-MOF w/o pre-training for CoRE_MAP_DB database.

Structural features prediction and high-throughput screening

While the Uni-MOF framework is adept at discerning the spatial arrangement of nanoporous materials, we aim to further explore its potential in forecasting structural attributes. The structural feature prediction results of Uni-MOF for the hMOF and CoRE_MOF materials libraries are presented in Figure 6 **a**-**d**. One have come to recognize that Uni-MOF demonstrates a strong capability in predicting the structural features of materials, owing to the utilization of pre-training on a substantial number of three-dimensional structures. Notably, in materials such as hMOFs, its predictive capability attains the coefficient of determination greater than 0.99, signifying a high level of precision and dependability. Thus, Uni-MOF not only accurately predicts the desired performance of MOF materials but also precisely predicts their structural features, which holds significant importance for material research and application.

Innumerable nanoporous materials can be created with varied secondary building units(SBUs)[44], leading to exceptionally diverse MOF structures, making MOF structure-property analysis a very strategic initiative. In this work, multi-gas adsorption uptakes under various operating conditions were collected and generated, where the argon uptakes at 87 K and 1 bar is representative and analyzable. Argon, an inert gas that accounts for the vast majority of atmosphere (9340 ppm at ambient conditions), has been widely used for insulation and illumination with a commercial value of 3.1 USD/kg.[14] Figure 6 **e** shows a comparison of the kernel density estimate (KDE) for different structural features between the entire CoRE_MOF with argon adsorption values and MOFs in CoRE_MOF with the top 10% performance of argon adsorption. KDE visualizes the distribution of data using continuous probability density curve in a less cluttered way.

As the distribution of some structural features is bounded, it may lead distortions (such as the minus value of surface area

and volume). However, it still presents interpretable and impressive trends in structure of top MOF adsorbents. The typical parameters describing pore sizes are as follows: 1) pore limiting diameter(PLD) is the largest free sphere; 2) largest cavity diameter(LCD) is largest included sphere along the free sphere path. Thus, LCD is intrinsically larger than PLD in the same material. Compared with the entire MOF database, the top 10% of MOFs have larger distribution values of 5-10 Å and 10-12 Å for PLD and LCD, respectively. The void fraction of the whole database is moderately distributed around 0.5, while that of the top 10% MOFs is much larger, mainly distributed around 0.75. Since most of the adsorption occurs on the surface of nanoporous materials, the surface area becomes one of the most critical determinants, and the results show that the top 10% of MOFs possess a very large surface area of about 3000 m^2/g, which is consistent with common sense. The surface area determines the surface adsorption process, while the LCD determines the internal absorption process. Therefore, the top 10% MOFs were classified into three tiers according to their adsorption performance on argon, and the numerical distribution of these two key factors (surface area/LCD) was explored (shown in Figure S1-S4). One can see that at 87 K and 1 bar, the most promising MOF adsorbents for argon mainly possesses an LCD of about 15 Å and a surface area of about 4100 m^2/g.

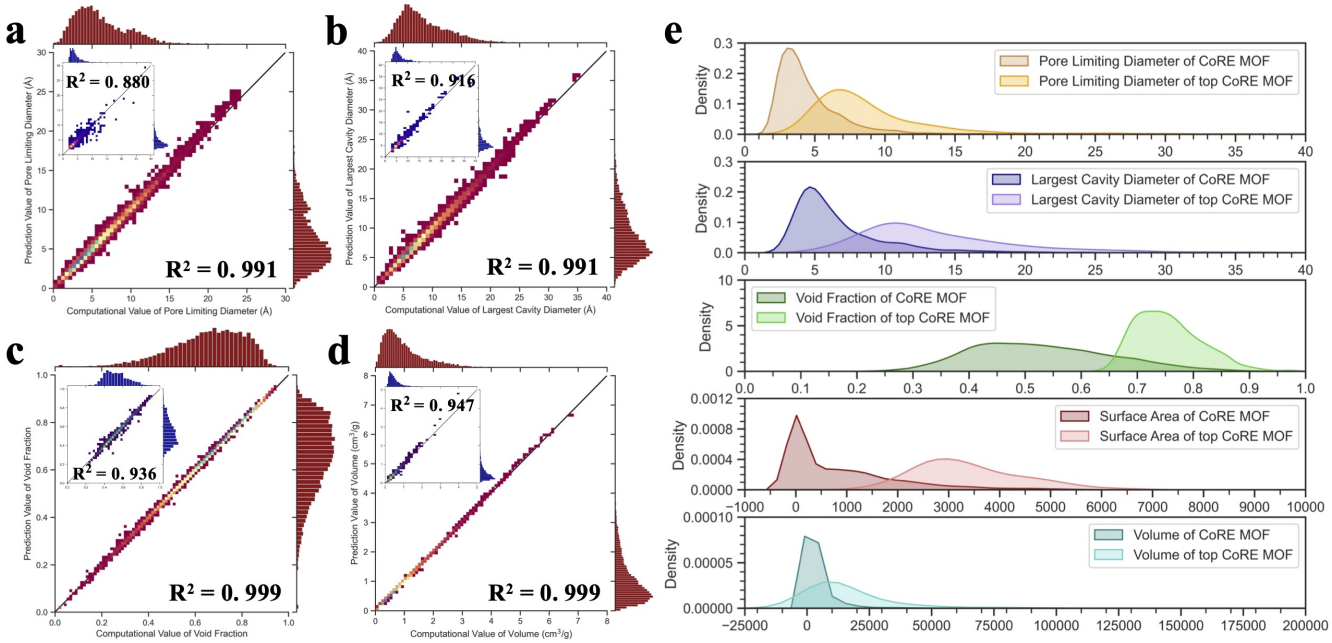

Figure 6. The correlation between predicted and computational value of **a** Pore Limiting Diameter (PLD), **b** Largest Cavity Diameter (LCD), **c** Void Fraction and **d** volume of MOFs in hMOF and CoRE_MOF databases. Red color represents structural features of hMOF, and blue color represents structural features of CoRE_MOF. **e** Comparison of the kernel density estimate (KDE) for different structural features [PLD (Å), LCD (Å), void fraction, surface area (m^2/g), volume ($Å^3$)] between the CoRE_MOF with all argon adsorption values and the CoRE_MOF with the top 10% performance of argon adsorption at 87K and 1bar.

Modeling of material structural representation

To validate that the MOF structures are well learned in both the pre-training and fine-tuning stages, we visualize the structural features, which are 512 dimensional vectors, using the t-distributed stochastic neighbor embedding (t-SNE) method[45]. The results are shown in Figure 7. As illustrated in Figure 7 **a** and **b**, the learned features are capable of classifying MOFs either from CoRE_MOF or hMOF datasets. One may also notice that the boundary between CoRE_MOF and hMOF in the fine-tuned features is much more obvious, suggesting a significant improvement of these features after the fine-tuning. Remarkably, this is clearly demonstrated when we draw the embeddings versus the surface areas of MOFs, as shown in Figure 7 **d**, where the structures with small surface areas are located in the upper-right and lower-left corners, and the structures with large surface areas are in the center. In comparison to the pre-training features, Figure 7 **c**, the fine-tuning can significantly improve the representation quality.

On the other hand, we are more curious about whether the learned structural representations are closely correlated with the adsorption behaviors. To address this issue, we visualize the structural embeddings versus the adsorbate values of both Ar and N_2, and show the result of Ar here as an example, see Figure 7 **e-f**. As one can observe, the representations learned in the pre-training stage are not able to classify the structures with different adsorbate capacities, the structures with various adsorbate

values are grouped together, see Figure 7 **e**. However, after the fine-tuning, the structures with different adsorbate behaviors are well separated, demonstrating a good relationship between the learned representations and the target of adsorbate values, see Figure 7 **f**. This well explained the functions of the fine-tuning stage in further affecting the structural representations as well as other model parameters.

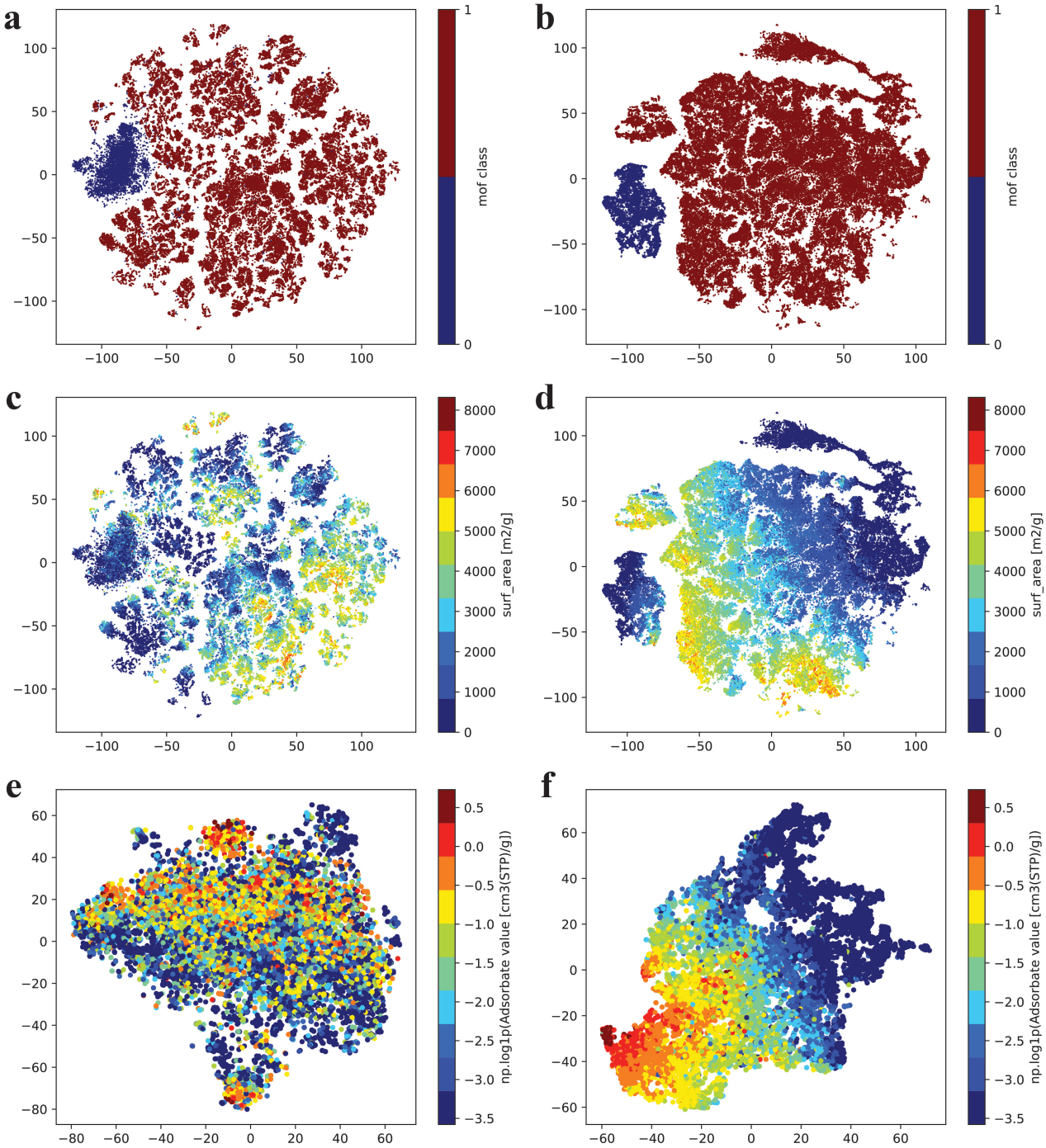

Figure 7. Visualization of structural representations of MOF in the hMOF and CoRE_MOF datasets, the low-dimensional embeddings are computed by t-SNE. The representations retrieved after **a, c, e** pre-training and **b, d, f** fine-tuning versus the other properties. **a** and **b** the low-dimensional embeddings *vs* dataset labels, where 0 represents CoRE_MOF and 1 represents hMOF. **c** and **d** illustrate the representations *vs* surface area in m^2/g for hMOF and CoRE_MOF combined. **e** and **f** show the relationship of representations with respect to the adsorbate values of Ar at 87K, 0.01Pa for CoRE_MOF dataset only.

Conclusion

In this study, we introduced Uni-MOF, a universal framework that can accurately predict gas adsorption in MOF materials. We also generated, collected, and organized relevant databases of nanoporous materials and gas adsorption datasets. The self-supervised learning of a database containing over 631,000 MOFs and COFs was performed, resulting in a high prediction accuracy of 0.98. This indicates that the representation learning framework based on three-dimensional pre-training effectively learns the complex structural information of MOFs while avoiding over-fitting. We applied Uni-MOF to predict the gas adsorption performance of three major databases and achieved a high prediction accuracy of up to 0.98 in database with sufficient data. In the case of a sufficiently sampled dataset, Uni-MOF not only maintains a predictive accuracy above 0.83, but also accurately selects high-performance adsorbents under high pressure by solely predicting adsorption under low pressure, consistent with experimental screening results. Thus, Uni-MOF represents a significant breakthrough in the field of material science with regards to the application of machine learning techniques. Furthermore, our Uni-MOF framework shows superior performance on cross-system datasets compared to single-system tasks and can accurately predict the adsorption properties of unknown gases with a high prediction accuracy of up to 0.85, demonstrating its strong predictive ability and generality. Through extensive pre-training of three-dimensional structures, Uni-MOF effectively learns the structural features of MOFs, achieving a high coefficient of determination of 0.99 for hMOFs. Additionally, t-distributed stochastic neighbor embedding (t-SNE) analysis confirms that the fine-tuning stage can further learn structural features, and structures with different adsorbate behaviors are well identified, indicating a strong correlation between learned representations and gas adsorption targets. In summary, Uni-MOF framework serves as a versatile predictive platform for gas adsorption in MOF materials, functioning as a "gas adsorption detector" for MOFs, as it exhibits high precision in predicting gas adsorption under diverse operating conditions and has broad applications in the field of material science.

Method

Materials and data generation

Apart from exploring nanoporous materials in the materials library, we employed the ToBaCCo.3.0 program to generate over 306,773 MOF structures. In addition, we conducted Grand Canonical Monte Carlo (GCMC)[46] simulations on RASPA[47] software to produce another 10,000+ gas adsorption uptake dataset, with 5×10^4 steps used as initialization cycles and an additional 5×10^4 steps employed for adsorption capacity samples.

Material analysis

The key of high-throughput computational materials science is robust software tools, here we use pymatgen[48] (Python Materials Genomics), a robust and open-source python library, to derive useful material properties from raw crystallographic structural data and conduct comprehensive materials analysis.

Uni-MOF framework

Pre-training

We employed Uni-Mol as the pre-training framework, which is a dedicated pure three-dimensional pre-training framework designed for molecules. Uni-Mol has shown outstanding performance in various downstream tasks in the field of drug discovery. However, due to the completely different structure and three-dimensional spatial distribution of MOF materials compared to small molecule drugs, as well as the periodic boundary conditions of porous structures and the much larger number of heavy atoms in crystals, we made necessary modifications to Uni-Mol based on these facts:

1) We leverage an extra head of lattice matrix to preserve cell geometric information. Presence of periodic boundary conditions (PBC) is considered in MOF representation learning as PBC is natural for MOF materials. Besides masked atoms prediction and coordinates recovery in Uni-Mol pretraining, a regression head to prediction lattice 3x3 matrix is used to learn PBC information:

$$\mathscr{L}_{lattice} = MSE(A,\hat{A}) = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - FFN(CLS^i_{repr})\right|^2 \tag{1}$$

$$\mathscr{L} = \mathscr{L}_{lattice} + \mathscr{L}_{Uni-Mol} \tag{2}$$

Where CLS_{repr} indices Uni-Mol's representation for classification token, CLS(classification token) refers to the special token in the input sequence of atoms, it is used to represent the entire moelcule in the Uni-Mol's output. A is the lattice matrix, we use FFN(Feedforward Neural Network) of CLS_{repr} to predict lattice matrix with optimizing MSE loss directly. $\mathscr{L}$ in Uni-MOF pretraining is a summation of $\mathscr{L}_{lattice}$ and original $\mathscr{L}_{Uni-Mol}$. In $\mathscr{L}_{Uni-Mol}$, atom masked prediction and coordinates recovery is major component of loss items. More details pleas refer to Uni-Mol's original paper.

2) We propose a edge gated kernel for geometric spatial positional encoding. MOFs share totally different arrangement of atoms compared to drug molecules in three-dimensional space, with porous structures and average above 1000 heavy atoms in single cell. Uni-Mol use Gaussian kernel to encode spatial positional information, while in MOF pretraining suffers from training instability with Gaussian kernel of much larger pair distance and atom counts. To address this, we propose a edge gated distance kernel:

$$\mathscr{A}(d,r;a,b) = a_r d + b_r \tag{3}$$

$$p_{ij} = p_{ij}^{proj} + p_{ij}^{emb} = LN(FFN(d_{ij})) + \sigma(\mathscr{A}(d_{ij},t_{ij};a,b)) \cdot Embedding(t_{ij}) \tag{4}$$

Where d_{ij} is the Euclidean distance of atom pair ij, and t_{ij} is the edge type of atom pair ij. Please note the edge here is not the chemical bond, and edge type is determined by the atom types of pair ij. $\mathscr{A}(\cdot,\cdot;a,b)$ is the affine transformation with parameters a and b, it affines d_{ij} corresponding to its edge type. p_{ij} indices the edge gated distance kernel, which is a summation of distance projection and edge gated embedding. FFN is Feedforward Neural Network about non-linear transformation of distance d_{ij} and LN is LayerNorm operation. A sigmoid gated is used as the affine transformation of $\mathscr{A}$ to weight edge pair type edmbedding.

Fine-tuning

Prediction of multi-gas adsorption under different operating conditions, the fine-tuning model should be fed with not only the three-dimensional spatial structure but also the gas and operating conditions (i.e., temperature and pressure). Therefore, gas block and temperature/pressure blocks are proposed in Uni-MOF to form a cross-system performance prediction module:

Gas block The gas representation is an combination of gas id and gas intrinsic property related descriptors:

$$x_g = concat(Eembedding(g_i); FFN(g_x)) \tag{5}$$

where g_i indices the gas id of gas g, g_x represents the gas g descriptors(listed in Table S2). The gas representation $\widehat{x}_g$ is a concatenation of gas id embedding and FFN layer mapping of gas descriptors.

Num block We use EDD(Equal Distance Discretization) and LD(Logarithm Discretization) for temperature and pressure encoding respectively. EDD frist maps the numerical value into correspond bucket with equal width then applied embedding mapping, LD utilizes the logarithm transform with EDD to accommodate with logarithmic likely features:

$$EDD(x_j) = Eembedding(\lfloor((x_j - x_j^{min})/w)\rfloor) \tag{6}$$

$$LD(x_j) = EDD(\log_{10}(x_j)) \tag{7}$$

$$x_{num} = concat(EDD(x_{temp}); FFN(x_{temp}); LD(x_{pressure}); FFN(x_{pressure})) \tag{8}$$

Where the interval width is noted as $w = (x_j^{max} - x_i^{min})/N_j$. x denotes to numerical features, x_{temp} and $x_{pressure}$ is the original temperature and pressure value of correspond environment. For temperature feature we use EDD and FFN to embedding, and for pressure feature we choose to use LD and FNN with consideration of logarithmic likely transformation in pressure.

High-throughput analysis of large MOF database

To further investigate the effect of material structure on gas adsorption, Zeo++[49], a software package for crystalline porous materials analysis, was used to perform analysis of structure and topology of the material geometry. Structural features include Largest Cavity Diameter (LCD), Pore Limiting Diameter (PLD), void fraction, void volume and specific surface area.

Statistical data visualization

In this work, Python data visualization library such as seaborn[50] and matplotlib[51] were used for informative statistical graphics. Three-dimensional structures in Figure 1 are drawn using the web service Bohrium™ at https://bohrium.dp.tech.

Data Availability

The MOF/COF structures used for pertaining are either collected from the currently available database or generated using the corresponding program. There is a wealth of existing MOF/COF databases, including computer-synthesized databases of hMOFs[52], ToBaCCo (Topologically Based Crystal Constructor) MOFs, and experimental-level databases of CoRE (Computation-Ready Experimental) MOFs[53], CoRE COFs[54] and CCDC (The Cambridge Crystallographic Data Centre), etc. One integrated database

online is MOFXDB, where more than 168,000 MOF/COF structures are available. Additionally, we used the ToBaCCo.3.0 program to generate over 300,000 MOF structures.

For the downstream task, i.e. gas adsorption uptake by MOFs, we collected data from online sources such as MOFXDB, composing datasets of more than 2,400,000 sorptions of hMOFs on five gases (CO_2, N_2, CH_4, Kr, Xe) at $273K/298K$ and $0.01-10Pa$ and over 460,000 sorptions of CoRE MOFs on two gases (Ar, N_2) at $77K/87K$ and $1-10^5Pa$. Another dataset was generated by Grand Canonical Monte Carlo(GCMC)[46] simulations on RASPA[47] software, around 10,000 sorptions were obtained within $150K-300K$ and $1Pa-3bar$, considering seven types of gas molecules (CH_4, CO_2, Ar, Kr, Xe, O_2, He).

Table 1. Structure & Data resources.

		Availability	Software	Size
Structure	collection	hMOF[52]		137,000+
		ToBaCCo		10,000+
		CoRE MOF[53]		12,000+
		CCDC		12,000+
		CoRE COF[54]		600+
		GCOFs[55]		160,000+
	generation			300,000+
Adsorption Uptake Data	collection	hMOF_MOFX_DB		2,400,000+
		CoRE_MOFX_DB		460,000+
	generation		RASPA[47]	9,900+
Other Property Data	generation		Zeo++[49]	149,000+

Code Availability

Code to run the Uni-MOF model is available at https://github.com/dptech-corp/Uni-MOF.
The notebook demo can be found at https://bohrium.dp.tech/notebook/cca98b584a624753981dfd5f8bb79674

Computing environment

Uni-MOF pre-training and fine-tuning are performed on V100/A100 GPUs, and Monte Carlo simulations are performed on CPU cluster of Bohrium™.

References

1. Sholl, D. S. & Lively, R. P. Seven chemical separations to change the world. *Nature* **532**, 435–437 (2016).
2. Kohl, A. L. & Nielsen, R. *Gas purification* (Elsevier, 1997).
3. Yang, R. T. *Gas separation by adsorption processes*, vol. 1 (World Scientific, 1997).
4. Boyd, P. G. *et al.* Data-driven design of metal–organic frameworks for wet flue gas co2 capture. *Nature* **576**, 253–256 (2019).
5. Koytsoumpa, E. I., Bergins, C. & Kakaras, E. The co2 economy: Review of co2 capture and reuse technologies. *The J. Supercrit. Fluids* **132**, 3–16 (2018).
6. Kather, A. & Scheffknecht, G. The oxycoal process with cryogenic oxygen supply. *Naturwissenschaften* **96**, 993–1010 (2009).
7. Jee, J.-G., Kim, M.-B. & Lee, C.-H. Pressure swing adsorption processes to purify oxygen using a carbon molecular sieve. *Chem. Eng. Sci.* **60**, 869–882 (2005).
8. Van Groenestijn, J. & Kraakman, N. Recent developments in biological waste gas purification in europe. *Chem. Eng. J.* **113**, 85–91 (2005).
9. Zhuang, L.-L., Yang, T., Zhang, J. & Li, X. The configuration, purification effect and mechanism of intensified constructed wetland for wastewater treatment from the aspect of nitrogen removal: a review. *Bioresour. technology* **293**, 122086 (2019).
10. Akerib, D. *et al.* Chromatographic separation of radioactive noble gases from xenon. *Astropart. Phys.* **97**, 80–87 (2018).
11. Lu, Z.-T. *et al.* Tracer applications of noble gas radionuclides in the geosciences. *Earth-Science Rev.* **138**, 196–214 (2014).

12. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **341**, 1230444 (2013).

13. Ding, M., Cai, X. & Jiang, H.-L. Improving mof stability: approaches and applications. *Chem. Sci.* **10**, 10209–10230 (2019).

14. Wang, J., Zhou, M., Lu, D., Fei, W. & Wu, J. Virtual screening of nanoporous materials for noble gas separation. *ACS Appl. Nano Mater.* **5**, 3701–3711 (2022).

15. Yang, Q., Liu, D., Zhong, C. & Li, J.-R. Development of computational methodologies for metal–organic frameworks and their application in gas separations. *Chem. Rev.* **113**, 8261–8323 (2013).

16. Li, J.-R., Kuppler, R. J. & Zhou, H.-C. Selective gas adsorption and separation in metal–organic frameworks. *Chem. Soc. Rev.* **38**, 1477–1504 (2009).

17. Knebel, A. & Caro, J. Metal–organic frameworks and covalent organic frameworks as disruptive membrane materials for energy-efficient gas separation. *Nat. Nanotechnol.* **17**, 911–923 (2022).

18. Zhou, M. & Wu, J. Inverse design of metal–organic frameworks for c2h4/c2h6 separation. *npj Comput. Mater.* **8**, 256 (2022).

19. Wang, J., Zhou, M., Lu, D., Fei, W. & Wu, J. Computational screening and design of nanoporous membranes for efficient carbon isotope separation. *Green Energy & Environ.* **5**, 364–373 (2020).

20. Lin, R.-B., Xiang, S., Zhou, W. & Chen, B. Microporous metal-organic framework materials for gas separation. *Chem* **6**, 337–363 (2020).

21. Banerjee, D. *et al.* Metal–organic framework with optimally selective xenon adsorption and separation. *Nat. communications* **7**, ncomms11831 (2016).

22. Iftimie, R., Minary, P. & Tuckerman, M. E. Ab initio molecular dynamics: Concepts, recent developments, and future trends. *Proc. Natl. Acad. Sci.* **102**, 6654–6659 (2005).

23. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143 (2018).

24. Rubinstein, R. Y. & Kroese, D. P. *Simulation and the Monte Carlo method* (John Wiley & Sons, 2016).

25. Weinan, E., Ren, W. & Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **66**, 052301 (2002).

26. Zhou, M. & Wu, J. A gpu implementation of classical density functional theory for rapid prediction of gas adsorption in nanoporous materials. *The J. Chem. Phys.* **153**, 074101 (2020).

27. Altintas, C., Altundal, O. F., Keskin, S. & Yildirim, R. Machine learning meets with metal organic frameworks for gas storage and separation. *J. Chem. Inf. Model.* **61**, 2131–2146 (2021).

28. Abdi, J. & Mazloom, G. Machine learning approaches for predicting arsenic adsorption from water using porous metal–organic frameworks. *Sci. Reports* **12**, 16458 (2022).

29. Nandy, A. *et al.* Mofsimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Sci. Data* **9**, 74 (2022).

30. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE transactions on neural networks* **20**, 61–80 (2008).

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

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

33. Kang, Y., Park, H., Smit, B. & Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal-organic frameworks. *Nat. Mach. Intell.* (2023).

34. Cao, Z., Magar, R., Wang, Y. & Barati Farimani, A. Moformer: self-supervised transformer model for metal–organic framework property prediction. *J. Am. Chem. Soc.* **145**, 2958–2967 (2023).

35. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. *Int. Conf. on Learn. Represent.* (2023).

36. Colón, Y. J., Gómez-Gualdrón, D. A. & Snurr, R. Q. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Cryst. Growth & Des.* **17**, 5801–5810, DOI: 10.1021/acs.cgd.7b00848 (2017). https://doi.org/10.1021/acs.cgd.7b00848.

37. Gómez-Gualdrón, D. A. *et al.* Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **9**, 3279–3289, DOI: 10.1039/C6EE02104B (2016).

38. 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).

39. Nath, K., Ahmed, A., Siegel, D. J. & Matzger, A. J. Microscale determination of binary gas adsorption isotherms in mofs. *J. Am. Chem. Soc.* **144**, 20939–20946 (2022).

40. Zhao, Z., Li, Z. & Lin, Y. Adsorption and diffusion of carbon dioxide on metal- organic framework (mof-5). *Ind. & Eng. Chem. Res.* **48**, 10015–10020 (2009).

41. Walton, K. S. *et al.* Understanding inflections and steps in carbon dioxide adsorption isotherms in metal-organic frameworks. *J. Am. Chem. Soc.* **130**, 406–407 (2008).

42. Nugent, P. *et al.* Porous materials with optimal adsorption thermodynamics and kinetics for co2 separation. *Nature* **495**, 80–84 (2013).

43. Duff, D. G., Ross, S. M. & Vaughan, D. H. Adsorption from solution: An experiment to illustrate the langmuir adsorption isotherm. *J. Chem. Educ.* **65**, 815 (1988).

44. Kalmutzki, M. J., Hanikel, N. & Yaghi, O. M. Secondary building units as the turning point in the development of the reticular chemistry of mofs. *Sci. advances* **4**, eaat9180 (2018).

45. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

46. Hammersley, J. *Monte carlo methods* (Springer Science & Business Media, 2013).

47. Dubbeldam, D., Calero, S., Ellis, D. E. & Snurr, R. Q. Raspa: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **42**, 81–101 (2016).

48. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

49. 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 Mesoporous Mater.* **149**, 134–141 (2012).

50. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021, DOI: 10.21105/joss.03021 (2021).

51. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. & Eng.* **9**, 90–95, DOI: 10.1109/MCSE.2007.55 (2007).

52. Chung, Y. G. *et al.* In silico discovery of metal-organic frameworks for precombustion co2 capture using a genetic algorithm. *Sci. advances* **2**, e1600909 (2016).

53. Chung, Y. G. *et al.* Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *J. Chem. & Eng. Data* **64**, 5985–5998 (2019).

54. Tong, M., Lan, Y., Yang, Q. & Zhong, C. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chem. Eng. Sci.* **168**, 456–464 (2017).

55. Lan, Y. *et al.* Materials genomics methods for high-throughput construction of cofs and targeted synthesis. *Nat. Commun.* **9**, 5274 (2018).

Acknowledgements

We thank Yuzhi Zhang, Hang Zheng and many other colleagues in DP Technology for their constructive comments and great help in this project.

Author contributions

J.W. and J.L. contributed equally to this work. D.L., Z.G. and J.W. designed and guided the project. J.W., J.L. and Z.G. performed the data collection, conducted breakthrough experiments and drafted the paper. H.W., G.K. and L.Z. provided important advice and revised the paper. All authors contributed to the discussion of results and commented on the paper.

Competing Interests

The authors declare no competing financial interests.

Additional information

Supplementary information The online version contains supplementary material available at https://github.com/dptech-corp/Uni-MOF.

Correspondence and requests for materials should be addressed to Jingqi Wang.