Identification of Optimal Metal-Organic Frameworks by Machine Learning: Structure Decomposition, Feature Integration, and Predictive Modeling

Zihao Wang[a], Yageng Zhou[b], Teng Zhou[a,b,*], Kai Sundmacher[a,b]

[a] Process Systems Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany

[b] Process Systems Engineering, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany

[*] Corresponding Author: zhout@mpi-magdeburg.mpg.de (Teng Zhou)

Abstract

A novel integrated machine learning (ML) framework, consisting of structure decomposition, feature integration and predictive modeling, is proposed to correlate MOF structures with gas adsorption capacities. First, metal nodes, organic linkers, and underlying topologies are identified from MOF structures. Numerical features of the organic linker are generated by molecular graph convolution. Later, the metal node is embedded and integrated with organic linker's features to capture the MOF chemical information. In addition, embedded MOF topology and geometric descriptors are considered as additional structure-level features. Finally, using all the chemical and geometric features as inputs, ML models are trained to predict MOF adsorption capacities. Through two case studies on hydrogen storage and ethylene/ethane separation, the proposed ML framework is demonstrated to be reliable and efficient for MOF property modeling. The resulting ML models can provide a fast and reliable prediction of MOF performance indicators, which thereby significantly accelerate the discovery of novel MOFs. The major novelty of the present work is the incorporation of full chemical information of MOF for ML modeling, which largely increases the prediction accuracy of adsorption performance and meanwhile facilitates the subsequent model-based optimal MOF discovery.

Keywords: machine learning; material discovery; molecular graph convolution; metal-organic framework; gas storage and separation

1. Introduction

Metal-organic frameworks (MOFs), which are constructed with building blocks (i.e., metal nodes and organic linkers) to form periodic structures following certain topologies, have attracted considerable attention due to their structural diversity, high porosity, and tailorable functionality. So far, MOFs have been widely applied in energy and environmental areas including gas storage [1, 2] and separation [3 – 5]. Their modular building blocks allow us to tailor MOF structures with desirable properties for specific applications, which has been confirmed by tens of thousands of novel MOFs successfully synthesized over the past decade [6]. In principle, these available building blocks can be assembled to form a huge number of possible MOFs, while experimentally synthesized MOFs reported to date represent only a tiny fraction. The sheer size of the structure space makes finding an optimal MOF for a particular application a great challenge.

With the rapid development of powerful computational techniques and the availability of high-performance computing resources, researchers are able to use classical molecular simulation approaches to identify potential materials from a large database in a brute-force manner. This method is known as high-throughput computational screening. Using this method, the experimental chemists and material scientists can efficiently focus on potential candidates screened out from the database. Through experimental validation, high-performance MOFs can be finally identified [6]. Targeting at different applications, many contributions on high-throughput computational screening have been carried out [7 – 14]. However, one should note that using a brute force screening approach to discover promising MOFs can be prohibitively expensive for very large databases, and manually analyzing such a large amount of data is impractical.

In molecular and materials science, machine learning (ML) is capable of inferring mathematical models to capture complex structure-property relationships from available data [15 – 19]. It provides a more direct and convenient way to estimate adsorption properties from MOF structures, with the computational cost decreased by several orders of magnitude compared to high-throughput molecular simulation [20]. Over the past few years, many researchers have applied different ML methods (e.g., decision tree, support vector machine, and neural network) to predict the uptakes of various gases (e.g., CH_4, CO_2, and H_2) in MOFs [21 – 29]. The established ML models can both accelerate the material screening process and provide deeper physical insights for MOF synthesis and discovery [30, 31].

So far, most of the ML models have employed MOF geometric descriptors (e.g., void fraction and surface area) to correlate their adsorption properties. Although geometric descriptors largely affect the adsorption performance, chemical diversity of building blocks can also play a crucial role [30, 32]. Currently, various chemical descriptors of MOFs have been proposed and applied in ML-based predictive modeling, such as atomic charges [21], number of atoms [22], hybridization and connectivity [23], percentage of metal relative to carbon [26], dipole moment [27], and molecular fingerprint [33]. It is challenging to identify a small and most informative set of features from the massive chemical descriptors for different applications *via* the feature engineering approach [34]. Fortunately, as an alternative, representation learning (a special deep learning technique) can efficiently tackle this issue by automatically learning features (without physical inputs) from material chemical and geometric information.

In this work, we introduce an integrated ML framework able to identify MOF building blocks by structure decomposition, couple the geometric characteristics and chemical features of the building blocks, and learn the gas adsorption properties. By using the proposed ML method integrated with representation learning-based MOF features, the adsorption capacities of MOFs can be directly predicted, thereby supporting fast MOF screening and discovery. Targeting at gas storage and separation, we first apply molecular simulations to generate adsorption data for a number of selected MOFs with diverse structures, then construct predictive ML models, and finally perform large-scale screening on a much larger pool of MOFs to identify potential candidates. Two case studies on hydrogen storage and ethylene/ethane separation are conducted to demonstrate the reliability and efficiency of the proposed ML method, and to confirm the great potential of the identified MOFs.

2. Computational Details

Data is of central importance in discovering relationships between material structures and their process-relevant properties. Experimental data is always preferred for predictive modeling. Unfortunately, they are very scattered and limited. In this context, the grand canonical Monte Carlo (GCMC) simulation is recognized as a powerful tool to generate data for MOF property modeling and discovery, because it has a high efficiency in simulating the adsorption performance of MOFs with a satisfying accuracy [35, 36].

2.1. MOF Database Pre-treatment

The hypothetical MOF (hMOF) database used in this work consists of 137,953 different MOFs [37]. We introduce a multi-step strategy to automatically extract MOF building blocks, identify topologies, and analyze MOF chemical information from their Crystallographic Information Files (CIFs). This data extraction and pre-treatment strategy is based on the deconstruction algorithm proposed by Bucior et al. [38], which stores the information of building blocks and topologies by string-form identifiers, namely MOFid and MOFkey.

First, MOF structures are transferred into MOFid and MOFkey identifiers; second, the building blocks and underlying topological networks are extracted from the identifiers following predefined rules; finally, data cleaning is performed to refine the MOF database. The final steps are as follows: (1) remove MOFs sharing duplicate MOFkey to ensure the uniqueness, and 45,254 MOFs remain; (2) remove MOFs with incomplete MOFkey to ensure that both the chemical information and topology have been successfully identified, and 39,676 MOFs remain; (3) remove MOFs with invalid organic linker molecules, and 33,480 MOFs remain; (4) retain only MOFs consisting of at most two types of organic linkers to reduce the complexity of MOF structures, and 9156 MOFs remain. These 9156 MOFs with identified chemical and topological information are subsequently employed in the GCMC simulation and ML-based property modeling.

In the following sections, firstly details of the GCMC simulation are presented. Then, the proposed integrated ML framework is introduced.

2.2. Molecular Simulation

Adsorption performance of MOFs is quantified by GCMC simulations using RASPA [39]. Each simulation run consists of 5,000 equilibration cycles followed by 20,000 cycles for production. Interactions between non-bonded atoms are modeled by the Lennard-Jones (LJ) potential [40] with a cutoff distance of 12 Å. The LJ parameters for MOF atoms are taken from the DREIDING [41] and Universal Force Fields [42], as listed in Table S1, and the LJ parameters between atoms of different types are calculated using the Lorentz-Berthelot mixing rule. The number of unit cells is adjusted so that each dimension of the simulation cell is at least twice of the cutoff distance, and MOF atomic charges are assigned using the extended charge equilibration (EQeq) method [43]. C_2H_4 and C_2H_6 are modeled using the united atom model of the transferable potentials for phase equilibria (TraPPE) force field [44], where both molecules are described by the two-site LJ potential

[45, 46]. For several popular MOFs, a good agreement between GCMC simulations and experimental measurements (collected from References [47 – 51]) is observed in Figure S1 of the Supporting Information, confirming the reliability of the applied GCMC simulations. In addition, MOF geometric properties including void fraction and surface area are computed by RASPA as well [39], and the pore diameter is calculated by Zeo++ [52].

With the data obtained from GCMC simulations, performance metrics such as deliverable capacity and selectivity can be calculated using Eqs. (1) and (2), respectively. The deliverable capacity ΔN is defined as the difference between the adsorption uptakes at the adsorption and desorption conditions, and the selectivity S is a key metric indicating the efficacy of an adsorbent for gas separation.

$$\Delta N_i = N_{i,ads} - N_{i,des} \tag{1}$$

$$S_{i/j} = \frac{N_i}{N_j} / \frac{y_i}{y_j} \tag{2}$$

where i and j are indexes of gas species, and y_i is the mole fraction of the gas species i in the bulk phase.

2.3. Machine Learning

In order to capture both chemical and geometric information of MOFs, an integrated ML architecture is proposed as shown in Figure 1. With the metal nodes and organic linkers identified from MOF structure decomposition, the chemical features are extracted by feature embedding and molecular graph convolution. On the other hand, MOF geometric features include topology and 5 key geometric properties (i.e., void fraction, pore limiting diameter, largest cavity diameter, and volumetric and gravimetric surface areas). Coupling the chemical and geometric features, a feedforward neural network (FNN)-based ML model is trained with GCMC data to determine the relationship between a MOF structure and its adsorption capacity.

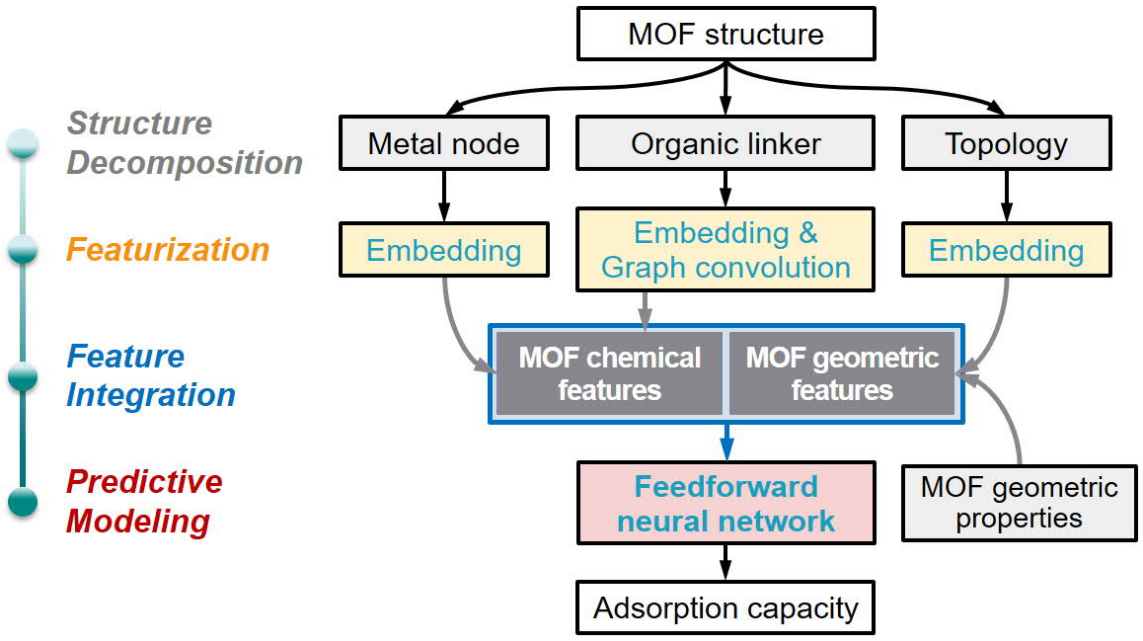

Figure 1. Schematic diagram of the proposed ML framework.

In the featurization stage, metal nodes and topologies are represented with chemical formula and topology identifiers (for example, [Zn][Zn] for the metal node and *pcu* for the topology) and directly embedded with numerical features. The involved organic linkers in our MOF set possess a large structural diversity with the number of heavy atoms in a linker varying from 4 to 102. In order to capture such complex and diverse molecular information, the organic linker is represented as a molecular graph where atoms and chemical bonds are regarded as nodes and edges, respectively. Multiple organic linkers in a MOF are regarded as multiple unconnected subgraphs. The features of each atom are first randomly initialized and iteratively updated with neighboring node's information using the graph convolution strategy, and global pooling on all the atomic features is performed to obtain the feature of the entire organic linker. Finally, concatenating all the chemical (metal and organic linker) and geometric (topology and geometric property) features as input, a feedforward neural network (FNN) is trained to predict the adsorption capacity. Notably, all the embedded features, graph convolution network, and FNN are trained or optimized simultaneously to minimize the overall property prediction error. An example on MOF structure decomposition, feature generation and integration, and property prediction is provided in Figure S2. More detailed descriptions on the ML approach including architecture of the network, atomic features, and graph convolution method are also provided (see Supporting Information).

The ML program is built using PyTorch [53] and PyG (PyTorch Geometric) [54]. The training, validation and test sets account for 80%, 10% and 10% of the employed dataset (corresponding to 7326, 915, and 915 MOFs). These three sets are used for model training, hyper-parameter optimization (the hyper-parameters are listed in Table S2), and model evaluation, respectively. Based on the loss function of mean squared error (MSE), the early-stopping criterion with a patience of 10 epochs is employed to prevent overfitting. That is, if the model performance on the validation set does not improve for 10 epochs, the training process is terminated and the model with the lowest validation loss is saved as the final model. All the computational work is carried out with the Linux Cluster Mechthild at the Max Planck Institute in Magdeburg.

3. Results and Discussion

Two case studies are performed to evaluate and demonstrate the proposed ML approach. One is the hydrogen storage and the other is the ethylene/ethane separation. The adsorption performance is first analyzed using MOF geometric properties only. Then, ML models for the prediction of adsorption capacity are developed and discussed, followed by the ML-assisted large-scale MOF screening.

3.1. Case 1: Hydrogen Storage

Green hydrogen is recognized as one of the most attractive transportation fuels because of its high gravimetric energy density and zero direct carbon emissions. A high-capacity and low-cost method for hydrogen storage would promote the widespread commercialization of fuel cell vehicles. Recently, MOFs have emerged as promising materials for hydrogen storage due to their high surface areas. These materials enable cryo-adsorbed hydrogen storage systems to operate at 100 bar that is much lower than 700 bar for the conventional gas compression-based storage [55].

For hydrogen storage, adsorption capacities of the 9156 MOFs at the pressure-swing adsorption conditions (100 bar/77 K and 2 bar/77 K) have been computed by GCMC simulations [56, 57]. The deliverable capacity (i.e., maximal amount of hydrogen that can be released to vehicles) is calculated from Eq. (1). As shown in Figure 2, most of the investigated MOFs present significantly higher H_2 uptakes at 100 bar than those at 2 bar, revealing their superior hydrogen deliverable capacities. The MOFs close to the diagonal line display similar adsorption capacities at high and low pressures. Hence, they are not recommended for hydrogen storage applications.

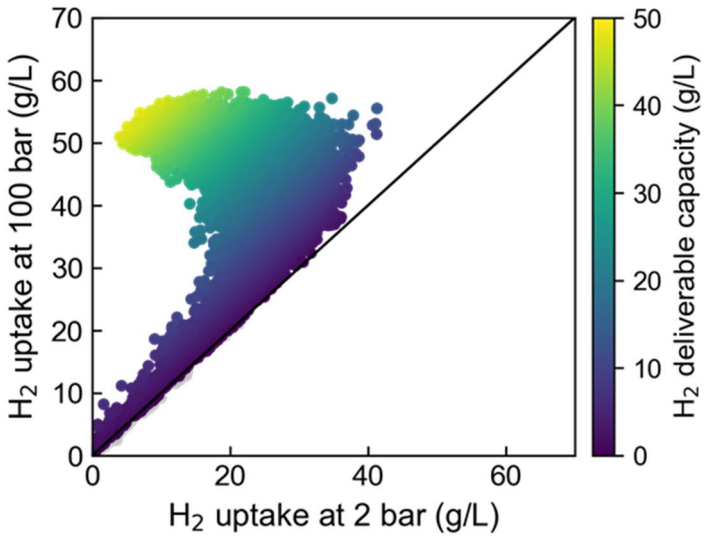

Figure 2. H_2 uptakes for MOFs at 100 bar/77 K and 2 bar/77 K with color indicating the H_2 deliverable capacity.

Figure 3 exhibits the dependence of MOF H_2 uptakes on four material geometric properties: void fraction, pore limiting diameter, volumetric surface area, and gravimetric surface area. Figure 3(a) shows that the H_2 uptake monotonically increases with the rising of void fraction, and high deliverable capacities are achieved with void fractions around 0.9. In contrast, the H_2 uptake displays a non-monotonic dependence on the pore limiting diameter as presented in Figure 3(b), initially increasing with the rising diameter and then decreasing for larger diameters. Maximal H_2 uptakes are achieved with MOFs showing pore limiting diameters between 5 and 10 Å, while high deliverable capacities are widely observed for MOFs with pore limiting diameters ranging from 5 to 25 Å. The influence of surface area on the H_2 uptake is illustrated in Figures 3(c) and (d). As depicted, the H_2 uptake generally increases with both volumetric and gravimetric surface areas. MOFs with volumetric surface areas in the range of 1500 – 2500 m^2/cm^3 and gravimetric surface areas larger than 5000 m^2/g are likely to achieve a high H_2 deliverable capacity. For completeness, H_2 uptakes at 2 bar are plotted in Figure S3.

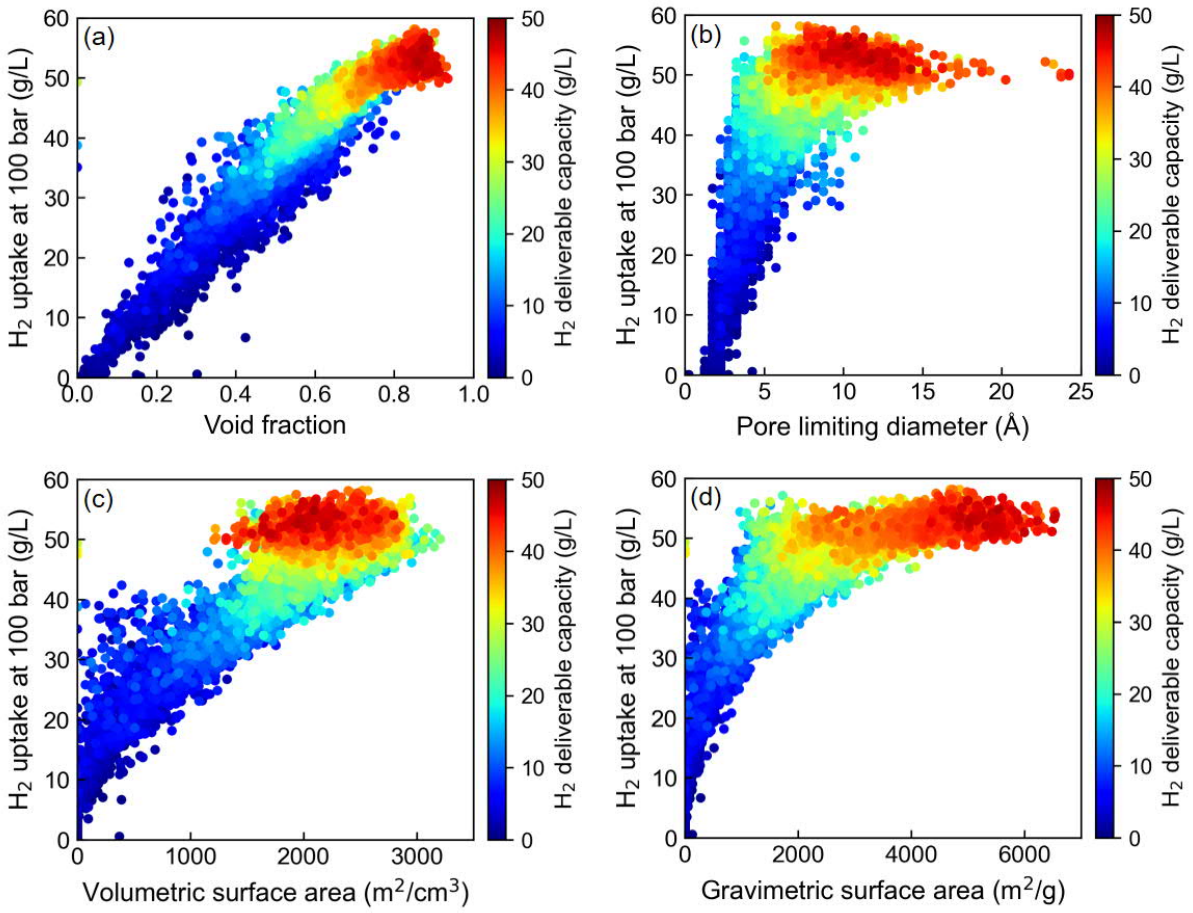

Figure 3. Relations between H_2 uptake at 100 bar/77 K and MOF geometric properties: (a) void fraction, (b) pore limiting diameter, (c) volumetric surface area, and (d) gravimetric surface area. Data points are colored by the H_2 deliverable capacity.

According to the framework in Figure 1, ML model is trained to predict H_2 uptakes using both chemical and geometric features of MOFs. Considering all possible hyper-parameter combinations, the optimal ML configuration (Table S3) is determined by the grid search method. Model performance for the prediction of H_2 uptakes is evaluated with mean absolute error (MAE) and coefficient of determination (R^2), as summarized in Table S4. The parity plot of the obtained ML models is visualized in Figure 4. In general, the two ML models achieve accurate predictions for H_2 uptakes at both 100 bar and 2 bar, with an MAE of 1.04 g/L (1.25 g/L) and 1.03 g/L (1.29 g/L) for the training set (test set), respectively. Additionally, the two ML models show an MAE of 1.24 g/L and 1.29 g/L for the validation set.

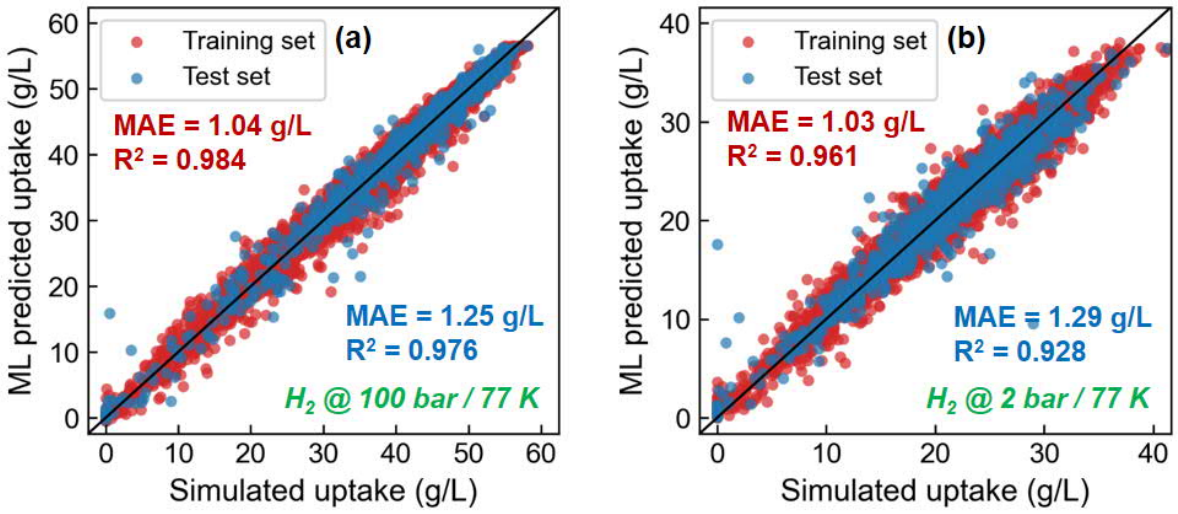

Figure 4. Parity plots of simulated and ML predicted H_2 uptakes at (a) 100 bar/77 K and (b) 2 bar/77 K.

Our proposed ML method considers both chemical and geometric information of MOFs for model development. For comparison, two new ML models are trained using only MOF geometric features to validate the importance of chemical features for the prediction of H_2 adsorption capacity. Re-optimized hyper-parameters of these models are provided in Table S5. It is observed that the removal of chemical features leads to a significant decrease of model parameters (96.4% for 100 bar and 87.1% for 2 bar). Performance of these models is evaluated and summarized in Table S4. When considering only geometric features, the MAE on the test set is 1.33 g/L and 2.02 g/L at 100 bar and 2 bar, respectively. The comparison shown in Figure 5 demonstrates that incorporating MOF chemical information in ML modeling provides more accurate predictions, especially for H_2 uptakes at 2 bar. In addition, our model presents an overall MAE of 1.08 g/L for H_2 uptakes at 100 bar/77 K. It is better than the ML model reported by Anderson et al. [28], with an MAE of 1.57 g/L.

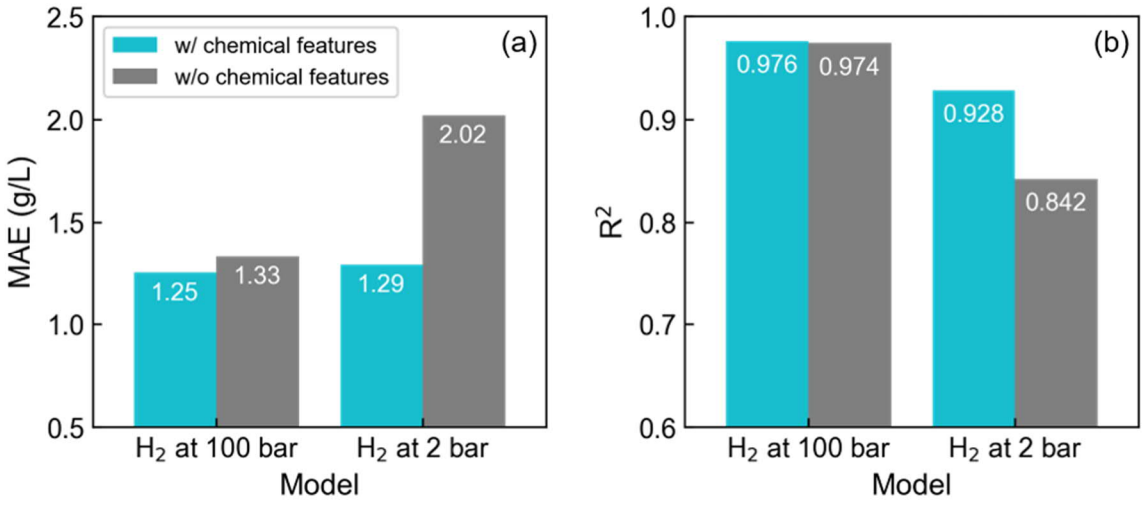

Figure 5. Performance of ML models (on the test set) developed with and without considering MOF chemical features for the H_2 adsorption uptake.

To adopt our ML models for MOF discovery, we extracted another much larger MOF database (named as Database-2) consisting of 21,384 MOFs, which is completely disjoint from the database previously used for model development (named as Database-1). The MOF structures in Database-2 are more complicated, because they contain three distinct organic linkers while those in Database-1 contain no more than two types of organic linkers. Subsequently, ML-assisted large-scale screening is performed by employing our developed ML models to predict H_2 uptakes (at both 100 bar and 2 bar) for all the MOFs in Database-2. Afterwards, H_2 deliverable capacities are calculated to discover potential MOF candidates for efficient H_2 storage. Top 100 candidates are identified, whose deliverable capacities are between 45.44 g/L and 47.20 g/L. Verification on the top 100 MOFs by GCMC simulations results in similar H_2 deliverable capacities ranging from 41.44 – 47.30 g/L (Figures S4 and S5). Based on the GCMC results, top four MOFs with the highest H_2 deliverable capacities are finally selected. Their computed H_2 deliverable capacities are higher than the best experimentally investigated MOFs for hydrogen storage (~ 42 g/L) [58], demonstrating the high potential of these MOFs for practical applications. Figure 6 summarizes the ID numbers, structural details, and GCMC-derived H_2 deliverable capacities for all the four MOFs. As indicated, the MOFs have similar structures. For instance, the same topology ***pcu*** is shared and the metal node is either copper or zinc. Three types of organic linkers are found, which also share high similarities.

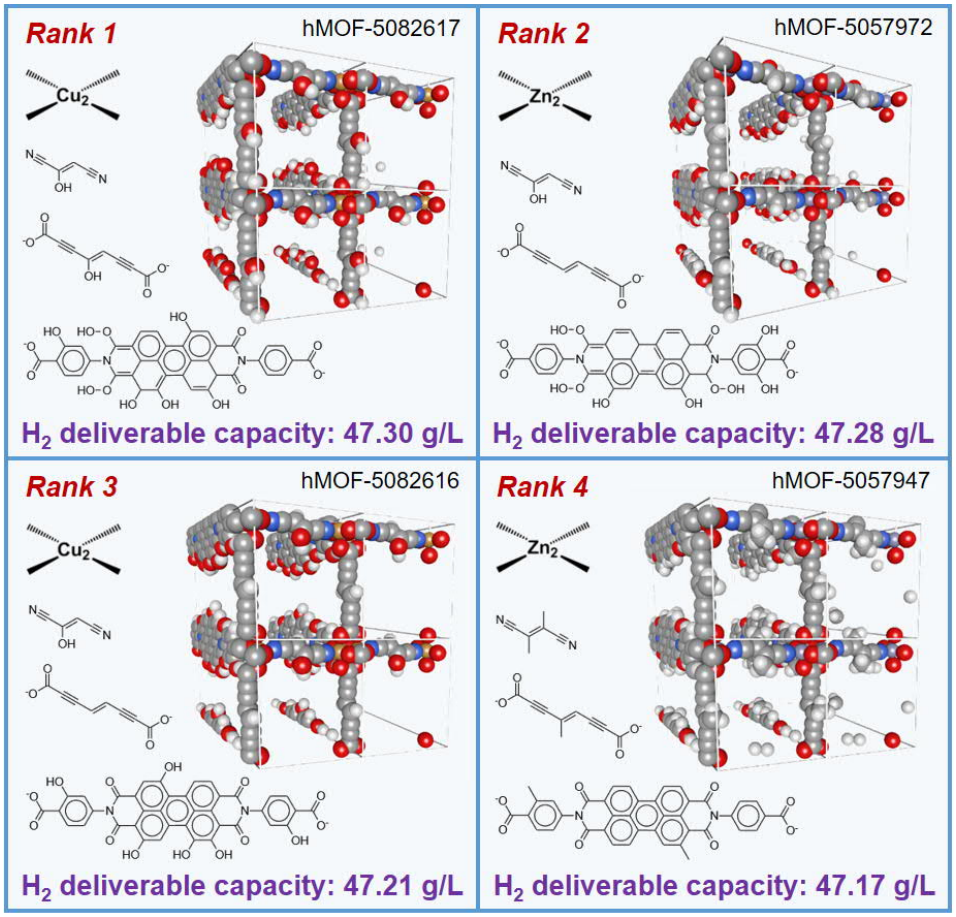

Figure 6. Top MOF candidates identified for hydrogen storage.

3.2. Case 2: C_2H_4/C_2H_6 Separation

Ethylene is one of the major industrial chemicals, primarily produced by thermal cracking of various feedstock such as natural gas. In ethylene production, the separation of ethylene (C_2H_4) and ethane (C_2H_6) is crucial, but challenging and energy-intensive due to their similar properties. Traditionally, cryogenic distillation is employed for the industrial separation of C_2H_6/C_2H_4 mixtures under high pressures and low temperatures (typically 7 – 28 bar and 183 – 258 K) with high distillation towers, which consumes a large amount of energy and capital cost [49]. In contrast, porous materials-based adsorptive separation is a promising alternative due to its high energy-efficiency and operational simplicity.

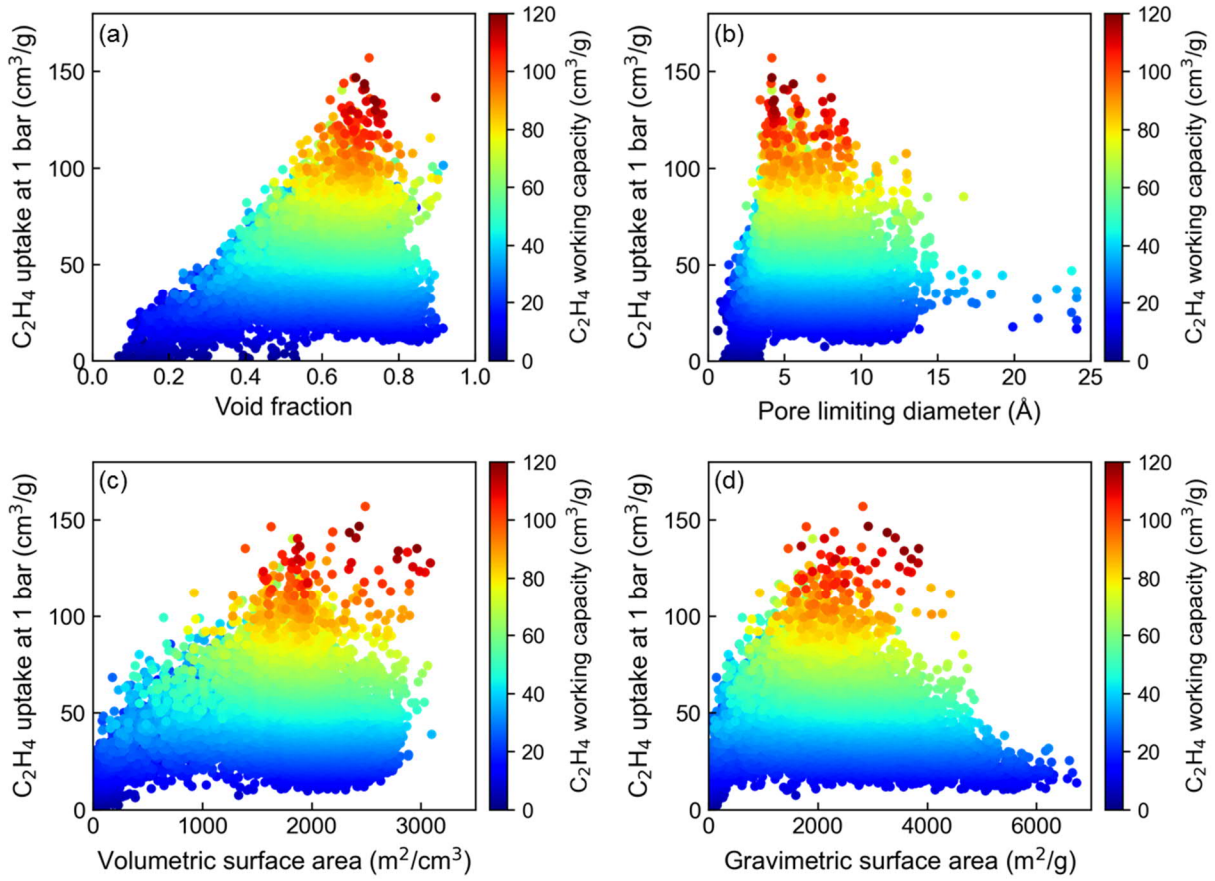

Figure 7. Relations between MOF geometric properties and the C_2H_4 uptake at 1 bar and 298 K: (a) void fraction, (b) pore limiting diameter, (c) volumetric surface area, and (d) gravimetric surface area. Data points are colored by the C_2H_4 working capacity.

The adsorptive separation of a typical gas product mixture from thermal cracking (C_2H_4/C_2H_6 = 15:1) is studied by GCMC simulations at 1 bar and 298 K to discover potential MOFs that selectively adsorb the small amount of C_2H_6 over abundant C_2H_4. Figure 7 shows the relations between the MOF geometric properties and the C_2H_4 uptake. As shown, the maximal C_2H_4 adsorption loading occurs at void fractions of 0.6 – 0.8, pore limiting diameters of 4 – 8 Å, volumetric surface areas of 1500 – 2500 m^2/cm^3, and gravimetric surface areas of 2000 – 3500 m^2/g. Looking at Figure 8, very similar trends are observed for C_2H_6 uptakes at the same conditions.

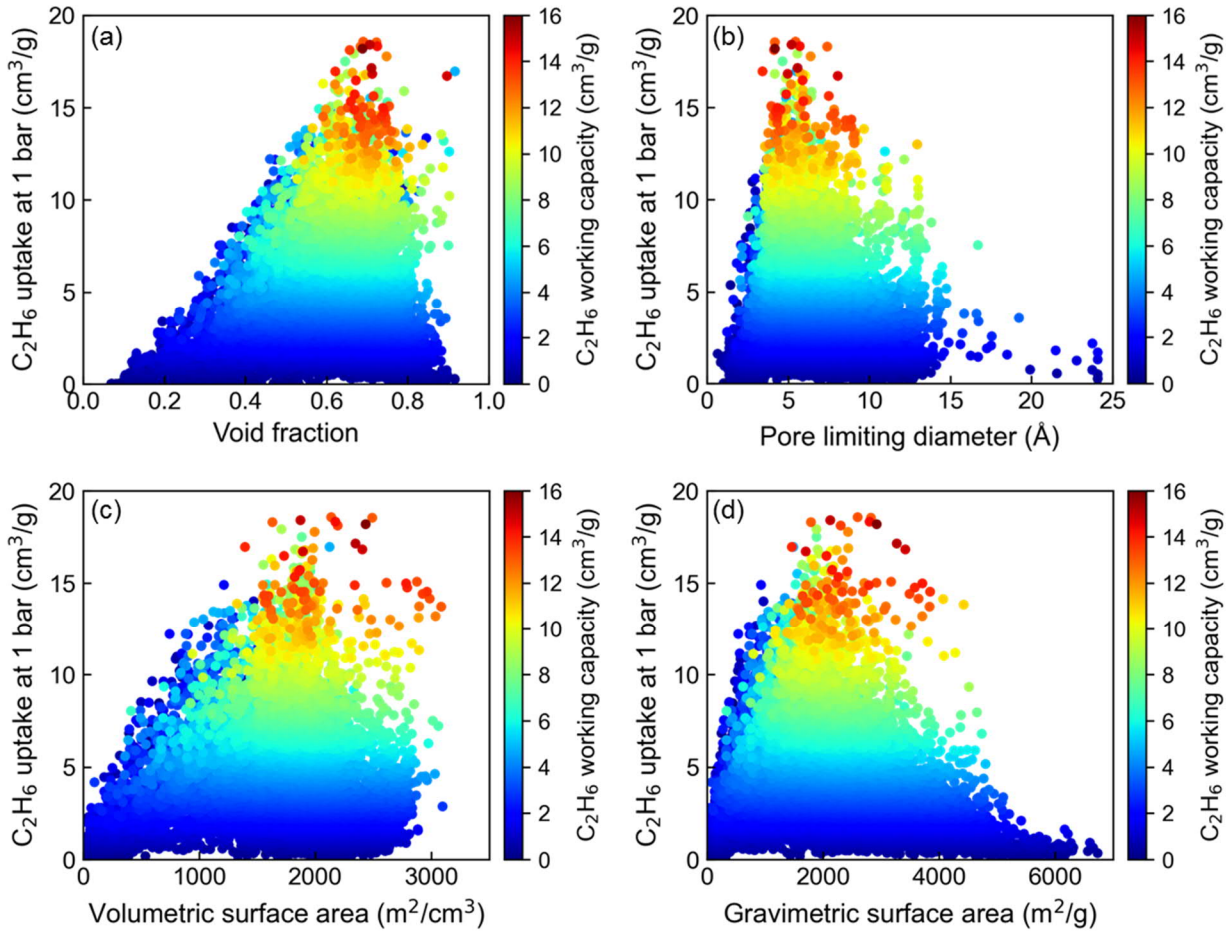

Figure 8. Relations between MOF geometric properties and the C_2H_6 uptake at 1 bar and 298 K: (a) void fraction, (b) pore limiting diameter, (c) volumetric surface area, and (d) gravimetric surface area. Data points are colored by the C_2H_6 working capacity.

To evaluate the performance of MOFs for separating C_2H_4 and C_2H_6, the C_2H_6/C_2H_4 selectivity is calculated using Eq. (2) considering both the uptake gap and the gas composition difference. The relations between the C_2H_6/C_2H_4 selectivity and MOF geometric properties are visualized in Figure S6. The top five MOFs with the highest selectivity (larger than 3.5) have an average void fraction of 0.41, a pore limiting diameter of 3.33 Å, and a volumetric surface area of 639 m^2/cm^3 and a gravimetric surface area of 361 m^2/g. High separation selectivity can be achieved by MOFs with relatively low pore limiting diameters and surface areas. However, the opposite is not always true, as indicated in Figure S6. Considering the implicit and ambiguous relations between MOF performance and their geometric properties, quantitative models to predict the adsorption uptakes and further calculate the selectivity from both geometric and chemical features of MOF are highly desirable.

Employing the framework in Figure 1, two ML models are trained to predict the C_2H_6 and C_2H_4 equilibrium uptakes, from which the selectivity is determined, using MOF chemical and geometric features. Considering all possible hyper-parameter combinations, the optimal ML configuration is determined by the grid search method (Table S3). The model performance is evaluated with the MAE and R^2, as shown in Table S6. Moreover, adsorption uptakes predicted by the ML models are compared with simulated results, as visualized in Figure 9. In general, ML models achieve satisfying predictions, with an MAE of 5.00 cm^3/g (5.79 cm^3/g) and 0.62 cm^3/g (0.77 cm^3/g) on the training set (test set) for C_2H_4 and C_2H_6, respectively. Additionally, the two ML models show an MAE of 6.03 cm^3/g and 0.80 cm^3/g for the validation set. Notably, the accuracy of the ML models for the C_2H_4 and C_2H_6 uptakes is slightly lower than those for H_2, which is easy to understand. The adsorption capacities of C_2H_4 and C_2H_6 in the binary mixture are more difficult to predict comparing to the case of pure gas adsorption (e.g., H_2 storage).

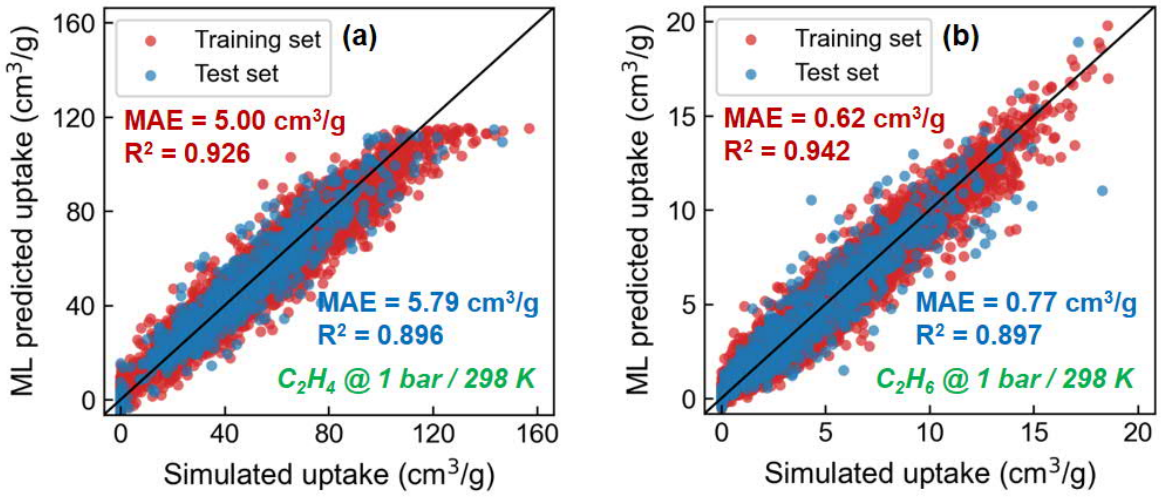

Figure 9. Parity plots of simulated and ML predicted (a) C_2H_4 and (b) C_2H_6 uptakes at 1 bar and 298 K.

To validate the importance of MOF chemical features in the prediction of C_2H_4 and C_2H_6 adsorption uptakes, two new ML models are trained using only geometric features, whose optimal hyper-parameters and model performances are presented in Tables S5 and S6, respectively. As expected, the removal of chemical features leads to a significant decrease of model parameters. When using only the geometric features as inputs, the MAE of the model on the same test set is 9.00 cm^3/g and 1.17 cm^3/g for C_2H_4 and C_2H_6, respectively. The comparison presented in Figure 10 clearly indicates that the incorporation of MOF chemical information significantly improves the

ML prediction accuracy, proving the significance of chemical features of MOF for the C_2H_4/C_2H_6 separation.

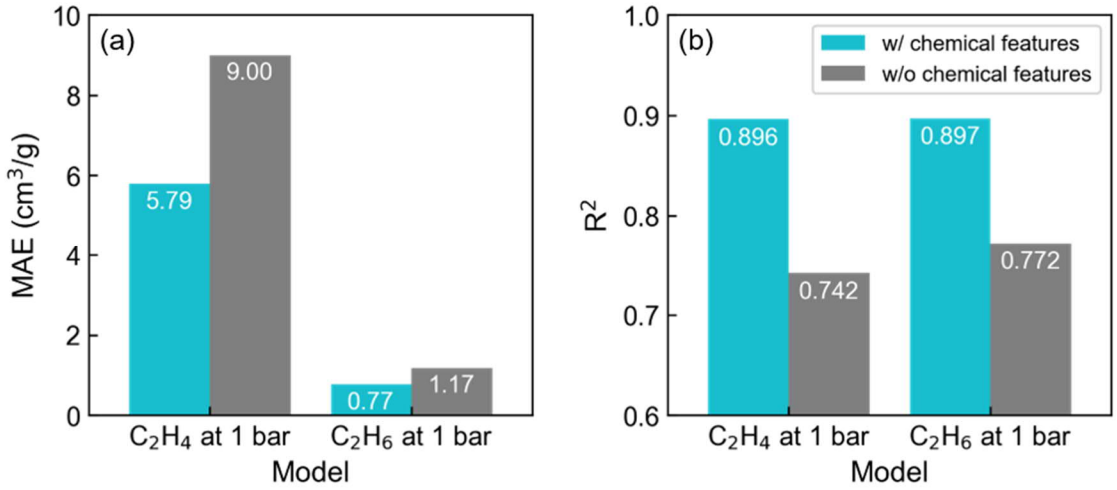

Figure 10. Performance of ML models (on the test set) developed with and without considering MOF chemical features for the C_2H_4 and C_2H_6 adsorption uptakes.

Finally, we employed the ML models to directly predict the performance of the 21,384 new MOFs in Database-2. The top 100 MOFs with the highest C_2H_6/C_2H_4 selectivity are identified, and the GCMC simulation is performed to validate their practical performance (Figure S7). Although the ML models tend to overestimate the selectivity, a decent C_2H_6/C_2H_4 selectivity of 5.52 is confirmed by GCMC, which is still higher than the largest selectivity (5.06) in the original training dataset. Importantly, the GCMC simulation for the top 100 MOFs takes over 140 node-hours on the high-performance computing resources, while the ML-assisted large-scale screening on the 21,384 MOFs is completed within 2 minutes. This again proves the high efficiency of our proposed method for fast MOF discovery. Figure 11 summarizes the ID numbers, metal nodes, and organic linkers of the top four MOFs (all share the same topology *pcu*) with the highest GCMC-derived selectivities ranging from 3.90 to 5.52. An extensive literature review is performed to find potential benchmark materials for comparison. It is found that PCN-250 shows an experimental selectivity around 2.0 at the same gas composition, temperature and pressure as used in the present work [29]. In the future, the superior performance of our identified MOF candidates should be validated by dedicated adsorption experiments prior to possible large-scale applications.

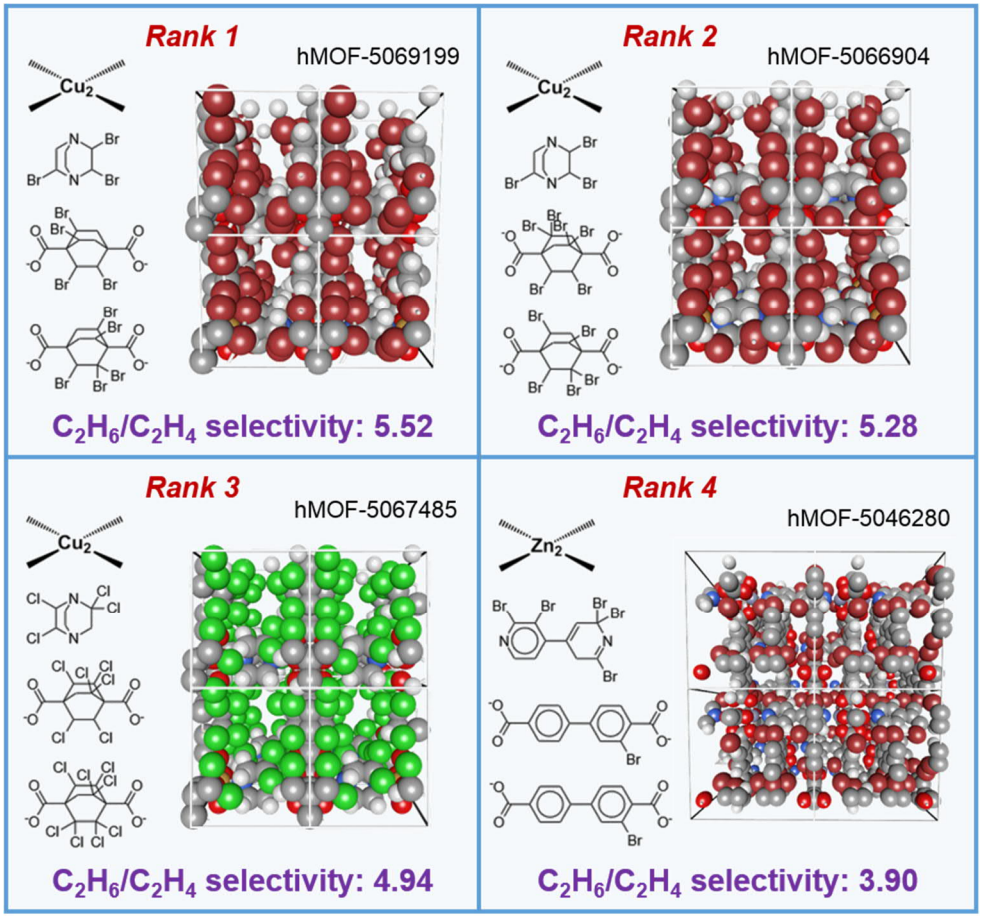

Figure 11. Top MOF candidates identified for the C_2H_4/C_2H_6 separation.

4. Conclusions

An integrated ML framework is proposed for the prediction of gas adsorption capacities coupling both chemical information and geometric characteristics from decomposed MOF structures. It has been proven that the consideration of chemical features significantly improves the accuracy of the ML model compared to the conventional studies using only geometric features as model inputs. Two case studies on hydrogen storage and ethylene/ethane separation are investigated. Based on the obtained ML models coupling chemical and geometric features, a number of high-potential MOFs are successfully identified by large-scale database screening. Their superior performances have been validated by GCMC simulations and should be further verified by adsorption experiments. Notably, the efficiency of large-scale MOF screening has been significantly improved as the performance evaluation on over 20,000 MOFs is completed within 2 minutes using the ML models. To the best of our knowledge, this is the first attempt in the process

systems engineering community to propose an integrated ML framework incorporating both chemical and geometric features of MOF for adsorption performance prediction and optimal MOF discovery.

Despite the remarkable progress achieved, some limitations also exist. The proposed ML method can discover MOFs with novel organic linkers. However, the design space is limited to MOFs containing the same metal nodes as covered by the MOF training set. Additionally, in order to avoid complex and task-specific feature engineering, representation learning is incorporated in the ML framework to automatically generate features for property prediction. This is efficient, reliable, and can help to discover promising MOFs. However, it is difficult to draw useful conclusions on how MOF chemical and geometric characteristics influence their performance. This knowledge can be acquired by using the so-called interpretable ML technique, which deserves future studies. Finally, our MOF discovery is achieved by large-scale screening on existing databases. An alternative and probably more efficient way for targeting of optimal MOFs is to formulate and solve an optimization-based reverse design problem based on the ML models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the International Max Planck Research School for Advanced Methods in Process and Systems Engineering (IMPRS ProEng), Magdeburg, Germany. Moreover, Teng Zhou acknowledges the financial support from the Max Planck Society, Germany, to his Junior Professorship at OvGU Magdeburg dedicated to Computer-Aided Material and Process Design (CAMPD).

References

[1] Kaye, S.S., Dailly, A., Yaghi, O.M., Long, J.R., 2007. Impact of preparation and handling on the hydrogen storage properties of Zn4O(1, 4-benzenedicarboxylate)3 (MOF-5). J. Am. Chem. Soc. 129 (46), 14176-14177.

[2] Gándara, F., Furukawa, H., Lee, S., Yaghi, O.M., 2014. High methane storage capacity in aluminum metal-organic frameworks. J. Am. Chem. Soc. 136 (14), 5271-5274.

[3] Chen, Y., Qiao, Z., Wu, H., Lv, D., Shi, R., Xia, Q., Zhou, J., Li, Z., 2018. An ethane-trapping MOF PCN-250 for highly selective adsorption of ethane over ethylene. Chem. Eng. Sci. 175, 110-117.

[4] Bao, Z., Wang, J., Zhang, Z., Xing, H., Yang, Q., Yang, Y., Wu, H., Krishna, R., Zhou, W., Chen, B., Ren, Q., 2018. Molecular sieving of ethane from ethylene through the molecular cross-section size differentiation in gallate-based metal-organic frameworks. Angew. Chem., Int. Ed. 130 (49), 16252-16257.

[5] Cui, X., Chen, K., Xing, H., Yang, Q., Krishna, R., Bao, Z., Wu, H., Zhou, W., Dong, X., Han, Y., Li, B., Ren, Q., Zaworotko, M.J., Chen, B., 2016. Pore chemistry and size control in hybrid porous materials for acetylene capture from ethylene. Science 353 (6295), 141-144.

[6] Witman, M., Ling, S., Anderson, S., Tong, L., Stylianou, K.C., Slater, B., Smit, B., Haranczyk, M., 2016. In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. Chem. Sci. 7 (9), 6263-6272.

[7] Aksu, G.O., Daglar, H., Altintas, C., Keskin, S., 2020. Computational selection of high-performing covalent organic frameworks for adsorption and membrane-based CO2/H2 separation. J. Phys. Chem. C 124 (41), 22577-22590.

[8] Avci, G., Erucar, I., Keskin, S., 2020. Do new MOFs perform better for CO2 capture and H2 purification? Computational screening of the updated MOF database. ACS Appl. Mater. Interfaces 12 (37), 41567-41579.

[9] Altintas, C., Keskin, S., 2016. Computational screening of MOFs for C2H6/C2H4 and C2H6/CH4 separations. Chem. Eng. Sci. 139, 49-60.

[10] Qiao, Z., Peng, C., Zhou, J., Jiang, J., 2016. High-throughput computational screening of 137953 metal-organic frameworks for membrane separation of a CO2/N2/CH4 mixture. J. Mater. Chem. A 4 (41), 15904-15912.

[11] Zhou, Y., Zhou, T., Sundmacher, K., 2020. In silico screening of metal-organic frameworks for acetylene/ethylene separation. In Comput.-Aided Chem. Eng. (Vol. 48, pp. 895-900). Elsevier.

[12] Braun, E., Zurhelle, A.F., Thijssen, W., Schnell, S.K., Lin, L.C., Kim, J., Thompson, J.A., Smit, B., 2016. High-throughput computational screening of nanoporous adsorbents for CO2 capture from natural gas. Mol. Syst. Des. Eng. 1 (2), 175-188.

[13] Mohamed, A.M., Bicer, Y., 2021. A comprehensive methodology to screen metal-organic frameworks towards sustainable photofixation of nitrogen. Comput. Chem. Eng. 144, 107130.

[14] Ga, S., Lee, S., Kim, J., Lee, J.H., 2020. Isotherm parameter library and evaluation software for CO2 capture adsorbents. Comput. Chem. Eng. 143, 107105.

[15] Alshehri, A.S., Gani, R., You, F., 2020. Deep learning and knowledge-based methods for computer-aided molecular design-toward a unified approach: State-of-the-art and future directions. Comput. Chem. Eng. 141, 107005.

[16] Zhang, L., Mao, H., Liu, L., Du, J., Gani, R., 2018. A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. Comput. Chem. Eng. 115, 295-308.

[17] Radhakrishnapany, K.T., Wong, C.Y., Tan, F.K., Chong, J.W., Tan, R.R., Aviso, K.B., Janairo J.I., Chemmangattuvalappil, N.G., 2020. Design of fragrant molecules through the incorporation of rough sets into computer-aided molecular design. Mol. Syst. Des. Eng. 5 (8), 1391-1416.

[18] Zhang, L., Mao, H., Zhuang, Y., Wang, L., Liu, L., Dong, Y., Du, J., Xie, W., Yuan, Z., 2021. Odor prediction and aroma mixture design using machine learning model and molecular surface charge density profiles. Chem. Eng. Sci. 245, 116947.

[19] Dev, V.A., Datta, S., Chemmangattuvalappil, N.G., Eden, M.R., 2017. Comparison of tree based ensemble machine learning methods for prediction of rate constant of Diels-Alder reaction. In Comput.-Aided Chem. Eng (Vol. 40, pp. 997-1002). Elsevier.

[20] Zhou, T., Song, Z., Sundmacher, K., 2019. Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design. Engineering 5 (6), 1017-1026.

[21] Cho, E.H., Deng, X., Zou, C., Lin, L.C., 2020. Machine learning-aided computational study of metal-organic frameworks for sour gas sweetening. J. Phys. Chem. C 124 (50), 27580-27591.

[22] Wu, X., Xiang, S., Su, J., Cai, W., 2019. Understanding Quantitative relationship between methane storage capacities and characteristic properties of metal-organic frameworks based on machine learning. J. Phys. Chem. C 123 (14), 8550-8559.

[23] Fanourgakis, G.S., Gkagkas, K., Tylianakis, E., Froudakis, G.E., 2020. A universal machine learning algorithm for large-scale screening of materials. J. Am. Chem. Soc. 142 (8), 3814-3822.

[24] Simon, C.M., Kim, J., Gomez-Gualdron, D.A., Camp, J.S., Chung, Y.G., Martin, R.L., Mercado, R., Deem, M.W., Gunter, D., Haranczyk, M., Sholl, D.S., Snurr, R.Q., Smit, B., 2015. The materials genome in action: identifying the performance limits for methane storage. Energy Environ. Sci. 8 (4), 1190-1199.

[25] Fernandez, M., Woo, T.K., Wilmer, C.E., Snurr, R.Q., 2013. Large-scale quantitative structure–property relationship (QSPR) analysis of methane storage in metal-organic frameworks. J. Phys. Chem. C 117 (15), 7681-7689.

[26] Pardakhti, M., Moharreri, E., Wanik, D., Suib, S.L., Srivastava, R., 2017. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). ACS Comb. Sci. 19 (10), 640-645.

[27] Anderson, R., Rodgers, J., Argueta, E., Biong, A., Gómez-Gualdrón, D.A., 2018. Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: from molecular simulation to machine learning. Chem. Mater. 30 (18), 6325-6337.

[28] Anderson, G., Schweitzer, B., Anderson, R., Gómez-Gualdrón, D.A., 2018. Attainable volumetric targets for adsorption-based hydrogen storage in porous crystals: molecular simulation and machine learning. J. Phys. Chem. C 123 (1), 120-130.

[29] Shi, Z., Liang, H., Yang, W., Liu, J., Liu, Z., Qiao, Z., 2020. Machine learning and in silico discovery of metal-organic frameworks: Methanol as a working fluid in adsorption-driven heat pumps and chillers. Chem. Eng. Sci. 214, 115430.

[30] Chong, S., Lee, S., Kim, B., Kim, J., 2020. Applications of machine learning in metal-organic frameworks. Coord. Chem. Rev. 423, 213487.

[31] Altintas, C., Altundal, O.F., Keskin, S., Yildirim, R., 2021. Machine learning meets with metal organic frameworks for gas storage and separation. J. Chem. Inf. Model. 61 (5), 2131-2146.

[32] Moosavi, S.M., Nandy, A., Jablonka, K.M., Ongari, D., Janet, J.P., Boyd, P.G., Lee, Y., Smit, B., Kulik, H.J., 2020. Understanding the diversity of the metal-organic framework ecosystem. Nat. Commun. 11 (1), 1-10.

[33] Qiao, Z., Li, L., Li, S., Liang, H., Zhou, J., Snurr, R.Q., 2021. Molecular fingerprint and machine learning to accelerate design of high-performance homochiral metal-organic frameworks. AIChE J. 67 (10), e17352.

[34] Jablonka, K.M., Ongari, D., Moosavi, S.M., Smit, B., 2020. Big-data science in porous materials: materials genomics and machine learning. Chem. Rev. 120 (16), 8066-8129.

[35] Moghadam, P.Z., Islamoglu, T., Goswami, S., Exley, J., Fantham, M., Kaminski, C.F., Snurr, R.Q., Farha, O.K., Fairen-Jimenez, D., 2018. Computer-aided discovery of a metal-organic framework with superior oxygen uptake. Nat. Commun. 9 (1), 1-8.

[36] Chung, Y. G., Gómez-Gualdrón, D. A., Li, P., Leperi, K. T., Deria, P., Zhang, H., Vermeulen, N.A., Stoddart, J.F., You, F., Hupp, J.T., Farha, O.K., Snurr, R.Q., 2016. In silico discovery of metal-organic frameworks for precombustion CO2 capture using a genetic algorithm. Sci. Adv. 2 (10), e1600909.

[37] Wilmer, C.E., Leaf, M., Lee, C.Y., Farha, O.K., Hauser, B.G., Hupp, J.T., Snurr, R.Q., 2012. Large-scale screening of hypothetical metal-organic frameworks. Nat. Chem. 4 (2), 83-89.

[38] Bucior, B.J., Rosen, A.S., Haranczyk, M., Yao, Z., Ziebel, M.E., Farha, O.K., Joseph, T., Hupp, Siepmann, J.I., Aspuru-Guzik, A., Snurr, R.Q., 2019. Identification schemes for metal-organic frameworks to enable rapid search and cheminformatics analysis. Cryst. Growth Des. 19 (11), 6682-6697.

[39] Dubbeldam, D., Calero, S., Ellis, D.E., Snurr, R.Q., 2016. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. Mol. Simul. 42 (2), 81-101.

[40] Tee, L.S., Gotoh, S., Stewart, W.E., 1966. Molecular parameters for normal fluids. Lennard-Jones 12-6 Potential. Ind. Eng. Chem. Fundam. 5 (3), 356-363.

[41] Mayo, S.L., Olafson, B.D., Goddard, W.A., 1990. DREIDING: a generic force field for molecular simulations. J. Phys. Chem. 94 (26), 8897-8909.

[42] Rappé, A.K., Casewit, C.J., Colwell, K.S., Goddard III, W.A., Skiff, W.M., 1992. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J. Am. Chem. Soc. 114 (25), 10024-10035.

[43] Wilmer, C.E., Kim, K.C., Snurr, R.Q., 2012. An extended charge equilibration method. J. Phys. Chem. Lett. 3 (17), 2506-2511.

[44] Eggimann, B.L., Sunnarborg, A.J., Stern, H.D., Bliss, A.P., Siepmann, J.I., 2014. An online parameter and property database for the TraPPE force field. Mol. Simul. 40 (1-3), 101-105.

[45] Martin, M.G., Siepmann, J.I., 1998. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. J. Phys. Chem. B 102 (14), 2569-2577.

[46] Wick, C.D., Martin, M.G., Siepmann, J.I., 2000. Transferable potentials for phase equilibria. 4. United-atom description of linear and branched alkenes and alkylbenzenes. J. Phys. Chem. B 104 (33), 8008-8016.

[47] He, Y., Krishna, R., Chen, B., 2012. Metal-organic frameworks with potential for energy-efficient adsorptive separation of light hydrocarbons. Energy Environ. Sci. 5 (10), 9107-9120.

[48] Liao, Y., Zhang, L., Weston, M.H., Morris, W., Hupp, J.T., Farha, O.K., 2017. Tuning ethylene gas adsorption via metal node modulation: Cu-MOF-74 for a high ethylene deliverable capacity. Chem. Commun. 53 (67), 9376-9379.

[49] Gucuyener, C., Van Den Bergh, J., Gascon, J., Kapteijn, F., 2010. Ethane/ethene separation turned on its head: selective ethane adsorption on the metal-organic framework ZIF-7 through a gate-opening mechanism. J. Am. Chem. Soc. 132 (50), 17704-17706.

[50] Mondal, S.S., Hovestadt, M., Dey, S., Paula, C., Glomb, S., Kelling, A., Schilde, U., Janiak, C., Hartmann, M., Holdt, H.J., 2017. Synthesis of a partially fluorinated ZIF-8 analog for ethane/ethene separation. CrystEngComm 19 (39), 5882-5891.

[51] Liao, P., Zhang, W., Zhang, J., Chen, X., 2015. Efficient purification of ethene by an ethane-trapping metal-organic framework. Nat. Commun. 6 (1), 1-9.

[52] Willems, T.F., Rycroft, C.H., Kazi, M., Meza, J.C., Haranczyk, M., 2012. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous Mesoporous Mater. 149 (1), 134-141.

[53] PyTorch. https://pytorch.org/

[54] PyTorch Geometric. https://pytorch-geometric.readthedocs.io/

[55] Gómez-Gualdrón, D.A., Colón, Y.J., Zhang, X., Wang, T.C., Chen, Y.S., Hupp, J.T., Yildirim, T., Farha, O.K., Zhang, J., Snurr, R.Q., 2016. Evaluating topologically diverse metal-organic frameworks for cryo-adsorbed hydrogen storage. Energy Environ. Sci. 9 (10), 3279-3289.

[56] Metal Organic Framework Database. https://mof.tech.northwestern.edu/

[57] Bobbitt, N.S., Chen, J., Snurr, R.Q., 2016. High-throughput screening of metal-organic frameworks for hydrogen storage at cryogenic temperature. J. Phys. Chem. C 120 (48), 27328-27341.

[58] Ahmed, A., Liu, Y., Purewal, J., Tran, L.D., Wong-Foy, A.G., Veenstra, M., Matzger, A.J., Siegel, D.J., 2017. Balancing gravimetric and volumetric hydrogen density in MOFs. Energy Environ. Sci. 10 (11), 2459-2471.