MSDE

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5me00060b

Discovery of metal–organic frameworks for inverse CO_2/C_2H_2 separation by synergizing molecular simulation and machine learning†

Daohui Zhao, Mao Wang, Zhiming Zhang and Jianwen Jiang *

Received 14th April 2025,
Accepted 21st July 2025

DOI: 10.1039/d5me00060b

rsc.li/molecular-engineering

Separation of carbon dioxide (CO_2) from acetylene (C_2H_2) represents a significant challenge in the petrochemical industry, primarily due to their similar physicochemical properties. By synergizing molecular simulation (MS) and machine learning (ML), in this study, we aim to discover top-performing metal–organic frameworks (MOFs) for inverse CO_2/C_2H_2 separation. Initially, the adsorption of a CO_2/C_2H_2 mixture in MOFs from the Cambridge Structural Database (CSD) is evaluated through MS, structure–performance relationships are constructed, and top-performing CSD MOFs are shortlisted. Subsequently, ML models are trained by utilizing pore geometry, framework chemistry, as well as adsorption heat and Henry's constant as descriptors. The significance of these descriptors is quantitatively assessed through Gini impurity measures and Shapley additive explanations. Finally, the transferability of the ML models is evaluated through out-of-sample predictions for CO_2/C_2H_2 separation in the computation-ready experimental (CoRE) MOFs. Notably, a handful of CoRE MOFs are found to outperform the best CSD MOFs and their performance is further compared with existing literature. The synergized MS and ML approach in this study is anticipated to accelerate the discovery of MOFs in a large chemical space for CO_2/C_2H_2 separation and other important separation processes.

Design, System, Application

As a unique class of nanoporous materials, metal–organic frameworks (MOFs) have attracted tremendous interest. With readily tuneable structures and chemical functionalities, they have been considered promising for a wide variety of applications, such as storage, separation and catalysis. The nearly infinite combinations of metal nodes and organic linkers have led to the synthesis of over 120 000 experimental MOFs and the construction of millions of hypothetical counterparts. It is infeasible to identify the best candidates in the immense chemical space of MOFs for a specific application *via* trial-to-error experiments. In this work, MOFs are rapidly screened *via* a hierarchical approach for CO_2/C_2H_2 separation, which is a crucial and challenging separation process in the chemical industry. First, the adsorption properties of CO_2/C_2H_2 mixture in CSD MOF dataset are predicted *via* molecular simulation. The relationships between separation performance metrics and structural factors are established. Then, machine learning (ML) models are developed and physically interpreted. Finally, the ML models are applied to predict and screen top-performing MOFs from another database (CoRE MOF). Good transferability is found for the ML models from the CCD to CoRE MOFs. The hierarchical approach in this study is useful for the rapid screening and rational design of MOFs for many other important separation processes.

1. Introduction

As a crucial chemical feedstock in the petrochemical industry, acetylene (C_2H_2) is predominantly produced by the partial combustion of natural gas and the steam cracking of hydrocarbons, with carbon dioxide (CO_2) being a main impurity.[1] Separation of CO_2 from C_2H_2 is of paramount importance, however, it presents a significant challenge due to their close boiling points (189.3 K for C_2H_2 and 194.7 K for CO_2) and nearly identical molecular sizes (both with a kinetic diameter of 3.3 Å, as detailed in Table S1†).[2] Currently, C_2H_2/CO_2 separation is primarily achieved through solvent extraction and cryogenic distillation, both of which are energy intensive. Therefore, there is considerable interest in developing environmentally friendly and cost-effective methods for this separation. In this context, adsorption separation utilizing porous materials is regarded as a more economically feasible and energy-efficient alternative to conventional heat-driven separation technologies.

Traditional adsorbents including zeolites, silicas and activated carbons have been tested for C_2H_2/CO_2 separation, yet their effectiveness remains limited. Constructed from inorganic metal nodes and organic linkers, metal–organic

Department of Chemical and Biomolecular Engineering, National University of Singapore, 117576, Singapore. E-mail: chejj@nus.edu.sg

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5me00060b

frameworks (MOFs) have garnered tremendous interest for many potential applications. By meticulously optimizing pore accessibility and chemistry to enhance host–guest interactions through crystal engineering and reticular chemistry, MOFs have demonstrated great potential for the separation of gas mixtures, such as C_2H_2/C_2H_4, C_2H_4/C_2H_6, and C_3H_6/C_3H_8. Furthermore, the periodic networks of MOFs provide an excellent platform for elucidating structure–performance relationships, thereby promoting the advancement of new functional MOFs for specific applications.

Since the first report of a prototypical C_2H_2-selective MOF in 2005,[3] a multitude of MOFs have been synthesized and investigated for C_2H_2/CO_2 separation. In general, most MOFs preferentially adsorb C_2H_2 over CO_2, attributed to strong acid–base interaction or hydrogen bonding between C_2H_2 and frameworks. This preference is further supported by the fact that C_2H_2 possesses a larger quadrupole moment (20.5×10^{-40} C m^2) than CO_2 (13.4×10^{-40} C m^2).[2] Practically, CO_2 as an impurity is a minor component in a C_2H_2/CO_2 mixture. C_2H_2-selective separation necessitates a subsequent desorption process to obtain C_2H_2, which consequently leads to increased energy consumption and secondary CO_2 pollution. By contrast, adsorbents with large CO_2 adsorption capacity and high CO_2/C_2H_2 selectivity are practically more advantageous, as they can effectively remove CO_2 in a single adsorption process. Unfortunately, MOFs with the capability of such inverse CO_2/C_2H_2 separation are scarce. To date, only a limited number of CO_2-selective porous materials have been reported, such as [Mn(bdc)(dpe)],[4] SIFSIX-3-Ni,[5] CD-MOF-1, CD-MOF-2,[6] AlFFIVE-1-Ni,[7] PCP-NH_2-ipa, PCP-NH_2-bdc,[8] Cd-NP,[9] and Ce(IV)-MIL-140-4F.[10] Notably, [Mn(bdc)(dpe)] was the first MOF with CO_2-selective behavior *via* guest discriminatory gating effect, achieving a selectivity of 8.8 for an equimolar CO_2/C_2H_2 mixture (v/v = 1 : 1) at 273 K and 10 bar.[4] A selectivity of 7.7 was observed in SIFSIX-3-Ni for a CO_2/C_2H_2 (v/v = 1 : 2) mixture at 298 K and 1 bar.[5] Based on the ideal-adsorption solution theory (IAST), selectivity values of 6.6 and 16.0 were reported for CO_2/C_2H_2 (v/v = 1 : 2) separation in CD-MOF-1 and CD-MOF-2, respectively, at 298 K and 1 bar.[6] Additionally, Ce(IV)-MIL-140-4F exhibited remarkable inverse CO_2/C_2H_2 (v/v = 1 : 2) separation, with selectivity values reaching 9.5 and 41.5 at 298 K and 273 K, respectively.[10] Despite high selectivity reported, a persistent trade-off exists between adsorption capacity and selectivity. For instance, PCP-NH_2-bdc was found to have a CO_2 uptake of 68 cm^3 g^{-1} at 298 K and 1 bar, but with a low selectivity of 4.4.[8] Currently, there are no established guidelines for the rational design of pore microenvironments that would yield ideal inverse CO_2/C_2H_2 separation. Most research efforts have focused on physisorption-based molecular sieving effect for CO_2/C_2H_2 separation, in relation to pore size and electrostatic interaction.

With readily chemical and topological tunability, thousands of MOFs (>120 000) have been synthesized to date. This vast material space presents an optimal platform for optimizing and identifying promising MOFs for targeted applications. However, traditional trial-and-error experiments are impractical to test such a large number of possible candidates. While molecular simulation (MS) has been applied to computationally screen top-performing MOFs and establish quantitative structure–property relationships for many gas separation applications, the exhaustive brute-force computations are often time-consuming. Recently, data-driven machine learning (ML) has emerged as a disruptive tool for design, screening and development of new materials, and it has been increasingly used for separation in MOFs.[11,12] For example, over 670 000 MOFs were screened for xenon/krypton separation using a random forest (RF) regressor.[13] CH_4/H_2 separation was examined *via* deep learning in a dataset of 134 185 hypothetical MOFs.[14] With various ML algorithms, C_2H_2 adsorption performance was assessed in MOFs on the basis of architectural, chemical and structural features.[15] ML models were also trained for C_2H_6/C_2H_4 separation in hypothetical MOFs.[16] A hierarchical strategy synergizing MS and ML was proposed for the rapid screening of MOFs for C_3H_8/C_3H_6 separation.[17] A genetic algorithm-based inverse design approach was proposed for C_2H_4/C_2H_6 separation in MOFs.[18] By integrating discarded experimental data with structural descriptors, an efficient ML model was develop to predict the separation performance of CO_2/C_2H_2 and C_2H_2/C_2H_4 in anion-pillared MOFs.[19]

Despite the above-mentioned studies, there has been limited ML research on inverse CO_2/C_2H_2 separation in MOFs. By synergizing MS and ML, our primary objective in this study is to discover top-performing MOFs for such an important application. As depicted in Fig. 1, a hierarchical workflow is adopted. First, CSD MOFs were analyzed, their performance for inverse CO_2/C_2H_2 separation was evaluated by Monte Carlo (MC) simulation, and top-performing structures were shortlisted. Then, ML models were trained, validated and interpreted. Finally, the ML models were used to predict the performance in CoRE MOFs for CO_2/C_2H_2 separation. Following this introduction, section 2 describes the MOF datasets, MS method, MOF featurization, ML training and prediction. In section 3, the structure–property relationships for CO_2/C_2H_2 separation in CSD MOFs are first discussed, top CSD MOFs are identified; then the predictive accuracy of the ML models trained upon CSD MOFs is examined and feature importance is analyzed; finally, the transferability of the ML models is examined by predicting CO_2/C_2H_2 separation in CoRE MOFs, and top CORE MOFs are identified and compared with top CSD MOFs, as well as experimentally reported MOFs. The concluding remarks are summarized in section 4.

2. Methodology

2.1. MOF datasets

Two experimental MOF databases were curated to construct datasets for this study. As listed in Table S2:† (1) the Cambridge Structural Database (CSD),[20] comprising about 10 636 ordered MOFs, was employed to generate simulation data for ML training and validation; (2) the updated

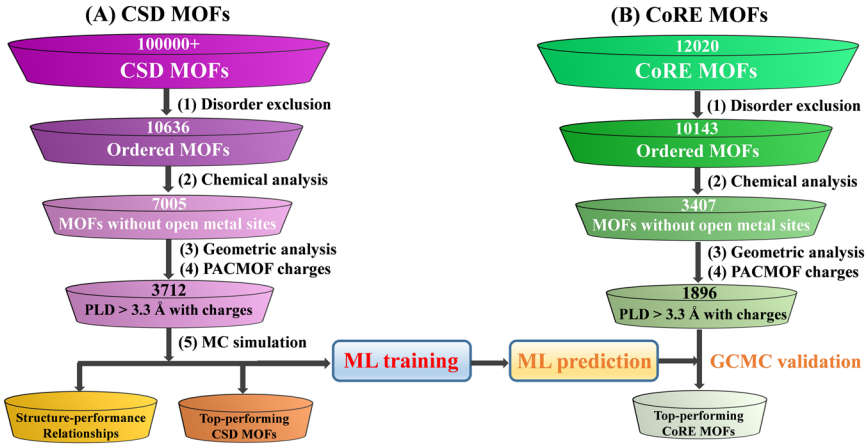

Fig. 1 Hierarchical workflow for screening of CSD MOFs, ML training, and ML prediction of CoRE MOFs.

computation-ready, experimental MOF (CoRE MOF 2019) database,[21] containing 10 143 ordered structures, was utilized for out-of-sample prediction. As illustrated in Fig. 1, MOFs with open metal sites (OMS) tend to interact strongly with the C≡C bond of C_2H_2, thus they were identified *via* open_metal_dectector (**https://github.com/emmhald/open_metal_detector/**) and removed from each database. Then, the geometric features of remaining MOFs, including the largest cavity diameter (LCD), pore-limiting diameter (PLD), pore size distribution (PSD), density (ρ), accessible volumetric surface area (VSA), probe-occupiable pore volume (PV), and void fraction (ϕ), were estimated by using Zeo++.[22] Those with PLD ≤ 3.3 Å (*i.e.*, the kinetic diameter D_k of both CO_2 and C_2H_2) were excluded, as incapable of accommodating CO_2 and C_2H_2. Finally, atomic charges of MOFs were assigned using PACMOF.[23] After these steps, 3712 CSD MOFs and 1896 CoRE MOFs were retained.

2.2. Molecular simulation

Adsorption of a CO_2/C_2H_2 mixture (v/v = 1 : 2) in MOFs at 298 K and 1 bar was simulated by using grand canonical MC (GCMC) method *via* RASPA 2.0 package.[24] The interactions between gas molecules and frameworks were characterized by Lennard–Jones (LJ) potential and electrostatic potential. The potential parameters for framework atoms were derived from the universal force field (UFF),[25] as detailed in Table S3,† while those for CO_2 and C_2H_2 were provided in Table S4.† The cross-interactions were calculated using the Lorentz–Berthelot mixing rules.[26] The LJ interactions were truncated at a distance of 12 Å with tail corrections, while electrostatic interactions were calculated using the Ewald summation method. It was assumed that the influence of framework flexibility was negligible on adsorption, as the pore size exceeded D_k of both gases; therefore, all the MOFs were treated as rigid. For each MOF, the length of simulation cell was extended to at least 24 Å along each dimension, and periodic boundary conditions were imposed in all three dimensions. Each GCMC simulation ran a total of 100 000 cycles, with the initial 25 000 cycles for initialization and the subsequent 75 000 cycles for ensemble averages. A cycle consisted of N steps (N: the number of gas molecules) and it would be equal to 20 if $N < 20$. Five distinct types of trial moves were employed randomly, including translation, rotation, reinsertion, swap, and identity change. Furthermore, the isosteric heats of adsorption ($Q°$) of CO_2 and C_2H_2 at infinite dilution, as well as the Henry's constants (K_H), were determined using the Widom insertion method. CO_2 adsorption capacity (N_{CO_2}, mmol g^{-1}) and CO_2/C_2H_2 selectivity (S_{CO_2/C_2H_2}) were used as metrics to assess the separation performance. S_{CO_2/C_2H_2} was defined as $(N_{CO_2}/N_{C_2H_2})/(y_{CO_2}/y_{C_2H_2})$, with y_{CO_2} and $y_{C_2H_2}$ denoting the mole fractions of CO_2 and C_2H_2 in the mixture, respectively.

2.3. MOF featurization

For ML, MOFs should be featured into machine-readable descriptors. We utilized physically intuitive and computationally viable descriptor types, including pore geometry, framework chemistry, as well as energy-related $Q°$ and K_H. (1) *Pore geometry.* PLD and LCD were commonly used geometric descriptors for MOFs in ML studies. Additionally, pore size distribution (PSD) was incorporated to account for pore heterogeneity, as it was revealed to be important.[17] The PSDs from 3.5 to 12 Å were divided into small bins with a spacing of 0.5 Å, while the PLD < 3.5 Å or >12 Å was amalgamated into a single bin. In such a way, 19 "PSD_bins" were obtained. Together with PLD and LCD, there were 21 geometric descriptors. (2) *Framework chemistry.* To accommodate the various hybridization and connectivity types of framework atoms, the densities of distinct atom types were formulated and enumerated using the lammps_interface[27] under the UFF4MOF nomenclature.[28,29] For carbon, different atom types were identified based on their bonding characteristics, categorized as single, double, triple and aromatic bonds, represented as C_1, C_2, C_3 and C_R, respectively. For metals, Zn_3f_2 and Cu^{4+2} were used to describe the tetrahedrally coordinated Zn^{2+} and paddlewheel-

like Cu^{2+}, respectively. To ensure applicability across diverse MOF datasets, all 221 atom types from the UFF4MOF collection were employed as descriptors. It should be noted that atomic densities were calculated from dividing the quantity of distinct atom types by a unit volume, thereby characterizing an intensive property. Additionally, revised-autocorrelations (RACs)[30] were utilized to characterize framework chemistry. The RACs included metal clusters, organic linkers and functional groups derived from the product or difference of atomic heuristics, such as Pauling electronegativity, connectivity and covalent radii, all of which were computed from molecular or crystal graphs. In total, 156 RAC descriptors were generated using the Molsimplify.[31] (3) *Energy descriptors*. To describe the framework-adsorbate affinity, the isosteric heat of CO_2 adsorption $\left(Q^{\circ}_{CO_2}\right)$ and the ratio of Henry's constants $\left(K^{\circ}_{CO_2/C_2H_2}\right)$ were used as energy descriptors. Table S5† lists the three types of descriptors, used as different descriptor sets in Table S6† to train ML models.

2.4. ML model training

With the above descriptors, ML models were trained for two targets N_{CO_2} and S_{CO_2/C_2H_2} generated from GCMC simulations. As illustrated in Fig. S1,† both N_{CO_2} and S_{CO_2/C_2H_2} exhibit markedly skewed distributions. Thus, we adopted a Box–Cox transformation[32] to transform the two target variables into Gaussian-like distributions. The simulation data were randomly split into two subsets, with 90% for training and 10% for test. The RF regressor, as implemented in the *scikit-learn* toolkit,[33] was applied for the training. As illustrated in Fig. S2,† RF regression is a typical tree-based algorithm that employs an ensemble of decision trees for predictive modeling.[34] In RF regression, hyperparameter optimization was conducted through a random grid search across parameter space (Table S7†), rather than an exhaustive grid search. For different combinations of descriptor sets, the optimal hyperparameters yielding the highest validation score are presented in Table S8.† To mitigate the risk of overfitting, five-fold cross-validation was employed. The VarianceThreshold method was initially applied to eliminate features with zero variance, followed by recursive feature elimination with cross-validation to further reduce descriptors. The model accuracy was quantified by the determination coefficient (R^2), mean absolute error (MAE), and Spearman rank correlation coefficient (SRCC). It should note that R^2 quantifies the linear correlation between predicted and actual values, while MAE indicates the deviation of predicted from actual values. SRCC measures the rank correlation between predicted and actual values, providing insight into rank similarity and model generalization. Higher R^2 and SRCC, along with lower MAE, signify better predictive accuracy of ML models. Additionally, Shapley additive explanations (SHAP)[35] were utilized to quantify the influence of various features on model predictions, both in terms of magnitude (significant or insignificant) and direction (positive or negative).

2.5. ML predictions

To test the transferability of the ML models developed upon CSD MOFs, they were used to predict CO_2/C_2H_2 separation in CoRE MOFs, allowing for rapid screening of CoRE MOFs. These out-of-sample predictions were then validated through GCMC simulations. Concurrently, top-performing CoRE MOFs were compared with CSD MOFs and experimentally reported MOFs.

3. Results and discussion

First, we analyze the simulation results pertaining to inverse CO_2/C_2H_2 separation in CSD MOF dataset, with the aim of uncovering significant structure–property relationships and identifying top-performing CSD MOFs. Subsequently, we assess the accuracy of the trained ML models, interpret feature importance, and elucidate critical factors governing the separation. Finally, the ML predictions for CO_2/C_2H_2 separation in CoRE MOF dataset are presented, with top-performing CoRE MOFs identified.

3.1. Separation in CSD MOFs

Thorough understanding of the relationships between structural features and performance metrics is essential for identifying potential MOFs and aiding in the design of new MOFs for CO_2/C_2H_2 separation. Fig. 2a illustrates the relationship of N_{CO_2} ~ VSA in CSD MOFs with different ϕ. Generally, a positive correlation is observed between ϕ and VSA, though it is not distinctly strong. MOFs with substantial N_{CO_2} (>1.7 mmol g^{-1}) typically possess moderate VSA in a broad range from 110 to 3000 m^2 cm^{-3}, and moderate ϕ ranging from 0.3 to 0.7. Low N_{CO_2} is distinctly noted when VSA reaches 3000 m^3 cm^{-3} or when ϕ exceeds 0.7, suggesting that neither SA nor ϕ serves as the sole determinant of N_{CO_2}. Q° is commonly used to characterize the affinity between adsorbate and framework. As depicted in Fig. 2b, there is a "volcano" in the relationship of N_{CO_2} ~ $Q^{\circ}_{CO_2}$. For MOFs with large pores or high ϕ, the guest–host interactions are relatively weak, resulting in small $Q^{\circ}_{CO_2}$ not conducive to CO_2 adsorption. Conversely, large $Q^{\circ}_{CO_2}$ occurs in MOFs with small pore size or low ϕ. MOFs with high N_{CO_2} > 1.7 mmol g^{-1} exhibit moderate $Q^{\circ}_{CO_2}$ ranging from 18.06 to 91.15 kJ mol^{-1}. However, many MOFs possessing this rang of $Q^{\circ}_{CO_2}$ display low N_{CO_2}. This indicates that, similar to VSA and ϕ, $Q^{\circ}_{CO_2}$ is not the sole determinant of N_{CO_2}. Fig. 2c shows the relationship of S_{CO_2/C_2H_2} ~ LCD. The majority of CSD MOFs are identified as C_2H_2-selective (*i.e.*, $S_{CO_2/C_2H_2} < 1$). For LCD in a small range (4.02–8.72 Å), ϕ varies considerably (0.07–0.64), reflecting diverse structural characteristics. Meanwhile, S_{CO_2/C_2H_2} spans widely (0.0018–19.74), suggesting that the separation performance of these MOFs does not correlate strongly with a single characteristic such as LCD, ϕ, VSA or

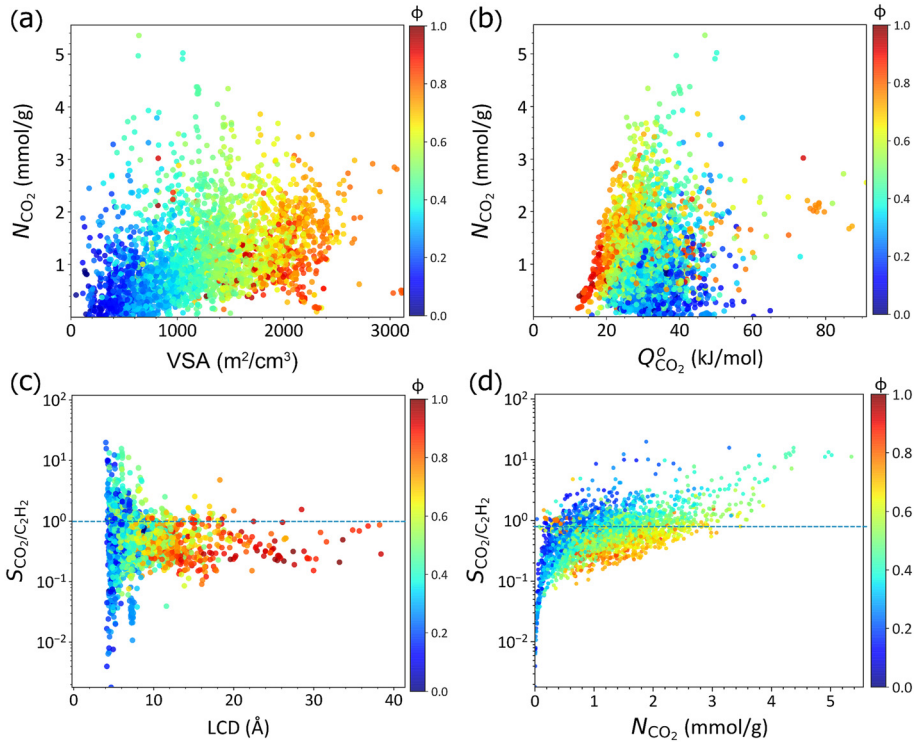

Fig. 2 Relationships of (a) N_{CO_2} ~ VSA, (b) N_{CO_2} ~ $Q^o_{CO_2}$, (c) S_{CO_2/C_2H_2} ~ LCD, and (d) S_{CO_2/C_2H_2} ~ N_{CO_2}. The color scaling in (a–d) denotes different ϕ.

$Q^o_{CO_2}$, but rather with the complex interplay of many factors. High S_{CO_2/C_2H_2} is predominantly associated with small LCD, as small pores/cages impose a strong confinement effect and facilitate selective adsorption. However, MOFs with excessively large LCD can readily adsorb both CO_2 and C_2H_2 molecules, resulting in non-selective adsorption. A total of 15 MOFs are identified with $S_{CO_2/C_2H_2} > 10$, with the highest S_{CO_2/C_2H_2} of 19.74 in borlom_P1. Overall, a single structural descriptor (like VSA or LCD) or energy descriptor (like $Q^o_{CO_2}$) fails to adequately describe the performance of MOFs, as it overlooks the synergy among various descriptors. Fig. 2d shows the relationships of S_{CO_2/C_2H_2} ~ N_{CO_2}. The most promising candidates for C_2H_2/CO_2 separation are located in the upper right quadrant of the plot. If we set $N_{CO_2} > 1.7$ mmol g^{-1} and $S_{CO_2/C_2H_2} > 3.0$, 53 CSD MOFs can be considered top-performing. As tabulated in Table S9,† these MOFs possess 4.02 Å < LCD < 8.51 Å, 112.63 m^2 cm^{-3} < VSA < 1946.46 m^2 cm^{-3}, 0.139 < ϕ < 0.581, and 27.93 kJ mol^{-1} < $Q^o_{CO_2}$ < 57.21 kJ mol^{-1}.

3.2. ML models

The preceding discussion highlights that the performance of MOFs for inverse CO_2/C_2H_2 separation is governed by a multitude of factors, including pore size, framework chemistry, and affinity. The interplay creates a complex multi-dimensional feature space that cannot be intuitively elucidated. To examine the nonlinear multi-dimensional relationships between these features and separation performance, we trained two ML models with N_{CO_2} and S_{CO_2/C_2H_2} as distinct targets based on the simulation data in CSD MOFs. Such models would facilitate more efficient predictive capability. The model accuracy and interpretability are discussed below.

3.2.1. Model accuracy. We systematically examine the model accuracy from various combinations of descriptor sets. As listed in Table S6,† "Geo" refers to a descriptor set with six geometric descriptors (PLD, LCD, VSA, PV, ϕ, ρ) that are commonly employed in ML studies of MOFs, albeit providing only coarse understanding of geometric characteristics. "PSD_bins" incorporates pore morphology by specifying upper and lower limits of pore size and their relative distributions, demonstrating sensitivity to minor variations in pore dimensions. The combination of "Geo" and "PSD" is denoted as "Geo + PSD_bins", which can combine with distinct chemical descriptor set "atomic" or "RACs" to become "Geo + PSD_bins + atomic" or "Geo + PSD_bins + RACs". Furthermore, "Geo + PSD_bins + RACs + energy" includes an energy descriptor. Table 1 presents the accuracy of ML models from various combinations of descriptor sets for both training and test sets in CSD MOF dataset. The accuracy was assessed by R^2, MAE and SRCC, as calculated by averaging 50 ML predictions for N_{CO_2} and S_{CO_2/C_2H_2} by randomly selecting training/test sets. The predictive accuracy of "Geo", which relies solely on basic geometric descriptors, is notably inadequate when applied to the test set, as evidenced by the low R^2 values of 0.667 and 0.506, respectively. This observation underscores the difficulty and limitation in differentiating CO_2 and C_2H_2 gases with similar physical properties, through mere adjustment of pore

Table 1 Accuracy of ML models from various combinations of descriptor sets

Target	Combination	Training set			Test set		
		R^2	MAE	SRCC	R^2	MAE	SRCC
N_{CO_2}	Geo	0.937	0.131	0.975	0.667	0.293	0.818
	Geo + PSD_bins	0.944	0.123	0.980	0.692	0.280	0.838
	Geo + PSD_bins + atomic	0.952	0.111	0.984	0.733	0.252	0.865
	Geo + PSD_bins + RACs	0.941	0.121	0.979	0.753	0.239	0.878
	Geo + PSD_bins + RACs + energy	0.967	0.091	0.986	0.827	0.207	0.915
S_{CO_2/C_2H_2}	Geo	0.908	0.164	0.978	0.506	0.404	0.578
	Geo + PSD_bins	0.907	0.166	0.976	0.505	0.405	0.571
	Geo + PSD_bins + atomic	0.925	0.141	0.984	0.637	0.339	0.731
	Geo + PSD_bins + RACs	0.929	0.133	0.985	0.647	0.318	0.755
	Geo + PSD_bins + RACs + energy	0.957	0.103	0.981	0.756	0.260	0.831

geometry. "Geo + PSD_bins" yields only a marginal improvement in accuracy, suggesting that the inclusion of PSD has an insignificant effect on the model performance. Prior ML research has emphasized the significance of chemical features in ML of MOFs,[17,36,37] and notable improvement is also observed here. When combining the atomic densities of "atomic" set into "Geo + PSD_bins", R^2 values in the test set increase from 0.692 to 0.733 for N_{CO_2} and from 0.505 to 0.637 for S_{CO_2/C_2H_2}. There is further improvement when RACs are used instead of "atomic" in "Geo + PSD_bins + RACs", suggesting that RACs more effectively captures pertinent chemistry of MOFs. Moreover, it is found that the inclusion of "energy" descriptors $Q^{\circ}_{CO_2}$ and $K^{\circ}_{CO_2/C_2H_2}$ results in the most accurate predictions, with R^2 of 0.827 and 0.756 for N_{CO_2} and S_{CO_2/C_2H_2} respectively in the test set. This is because that $Q^{\circ}_{CO_2}$ quantifies CO_2-framework interaction and $K^{\circ}_{CO_2/C_2H_2}$ implies the selectivity of CO_2/C_2H_2 at infinite dilution. They provide additional information between gas and framework, which is not available in other descriptors like Geo + PSD_bins + RACs. Additionally, the learning curves for N_{CO_2} prediction are shown in Fig. S3† by varying the size of training set. With increasing the ratio of training/test sets, R^2 rises while MAE drops, and each tends to approach a constant when the ratio is around 90/10.

Fig. 3 shows the parity plots between ML predicted and GCMC simulated N_{CO_2} and S_{CO_2/C_2H_2}. The ML predictions are based on the combination of "Geo + PSD_bins + RACs + energy". Satisfactory predictive accuracy is found. The MAE for N_{CO_2} are 0.09 mmol g^{-1} and 0.21 mmol g^{-1} in the training and test sets, respectively; for S_{CO_2/C_2H_2}, the MAE are 0.10 and 0.26. The R^2 for N_{CO_2} is 0.83 in the test set, while it is 0.76 for S_{CO_2/C_2H_2} as attributed to the imbalanced distribution of training data. Overall, the rankings of MOFs in ML predicted N_{CO_2} and S_{CO_2/C_2H_2} are strongly correlated with simulation

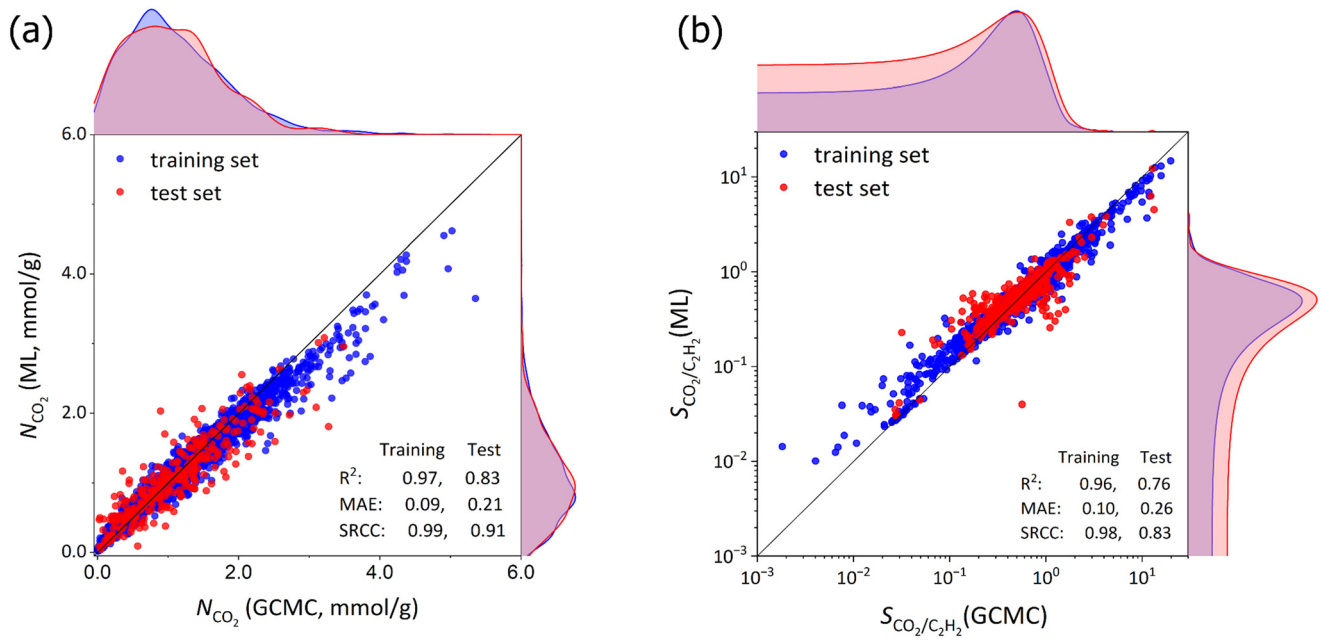

Fig. 3 Parity plots for ML predicted *versus* GCMC simulated (a) N_{CO_2} and (b) S_{CO_2/C_2H_2}. The ML predictions are based on the descriptor combination of "Geo + PSD_bins + RACs + energy".

data, yielding SRCC of 0.91 and 0.83 in the test set, respectively. This indicates that the models based on "Geo + PSD_bins + RACs + energy" are suitable for predicting new MOFs, as elaborated below.

3.2.2. Model interpretability. It is instructive to quantitatively elucidate the significance of various features in ML models for separation. Based on the mean decrease in Gini impurity, Fig. 4 shows the 10 most important features in the optimal ML models for N_{CO_2} and S_{CO_2/C_2H_2}. Interestingly, the relative feature importance differs between N_{CO_2} and S_{CO_2/C_2H_2}. For N_{CO_2}, the geometric descriptor PV contributes the most, followed by $Q^{\circ}_{CO_2}$, $K^{\circ}_{CO_2/C_2H_2}$, ϕ and density. The pore size parameters including LCD and PLD, along with specific chemical properties (f-lig-Z-0, f-lig-chi-0, and mc-Z-2-all), also play a role in N_{CO_2}. For S_{CO_2/C_2H_2}, energy descriptor $K^{\circ}_{CO_2/C_2H_2}$ is as the most influential, as it is indeed equivalent to S_{CO_2/C_2H_2} at the infinite dilution. The second influential factor affecting S_{CO_2/C_2H_2} is LCD, as discussed earlier in the relationships of $S_{CO_2/C_2H_2} \sim$ LCD. Additionally, CO_2-framework affinity (*i.e.*, $Q^{\circ}_{CO_2}$), other chemical and geometric features also contribute to governing S_{CO_2/C_2H_2}. The analysis indicates that N_{CO_2} and S_{CO_2/C_2H_2} predictions are significantly influenced by the incorporation of energy and geometric descriptors in the ML models, alongside with chemical descriptors.

The SHAP analysis is utilized to quantitatively assess how various features affect model outputs.[38] The significance of a feature is represented by the absolute value of SHAP score. Fig. 5a shows the summary plot of top 10 important features for N_{CO_2}. It is evident that PV exerts the most substantial influence. A high PV is generally associated with a positive SHAP score, which correlates with an increased probability of N_{CO_2}. This observation is similarly applicable to other features such as $K^{\circ}_{CO_2/C_2H_2}$, $Q^{\circ}_{CO_2}$ and ϕ. Conversely, a large LCD or PLD corresponds to a negative SHAP score, thus negatively correlated with N_{CO_2}. For illustration, Fig. 5b and c present SHAP force plots in two MOFs namely xewcua_P1_H and fonkea_P1. The former exhibits the highest N_{CO_2} among CSD MOFs, while the latter has the lowest. It is apparent that high PV and $Q^{\circ}_{CO_2}$ in xewcua_P1_H positively affect N_{CO_2}, whereas low PV and $K^{\circ}_{CO_2/C_2H_2}$ in fonkea_P1 possess negative effect. For S_{CO_2/C_2H_2}, as discussed above and shown in Fig. S4a,† $K^{\circ}_{CO_2/C_2H_2}$ is clearly the most significant influential, followed by $Q^{\circ}_{CO_2}$, LCD, and PLD, and f-lig-Z-0. $K^{\circ}_{CO_2/C_2H_2}$, $Q^{\circ}_{CO_2}$ and density demonstrate a positive effect on S_{CO_2/C_2H_2}, while LCD and PLD exhibit a negative effect. Fig. S4b and c† further illustrate the SHAP force plots in borlom_P1 with the highest S_{CO_2/C_2H_2} and in fonkea_P1 with the lowest S_{CO_2/C_2H_2}. Obviously, high $K^{\circ}_{CO_2/C_2H_2}$ and $Q^{\circ}_{CO_2}$ in borlom_P1 have positive effect on S_{CO_2/C_2H_2}, whereas low $K^{\circ}_{CO_2/C_2H_2}$ in fonkea_P1 exhibits a negative effect.

3.3. Predictions in CoRE MOFs

The ML models trained upon CSD MOFs were used to conduct out-of-sample predictions for CO_2/C_2H_2 separation in 1689 CoRE MOFs, with the descriptor combination of "Geo + PSD_bins + RACs + energy". As shown in Fig. 6, the predicted N_{CO_2} and S_{CO_2/C_2H_2} agree fairly well with GCMC simulation data. For N_{CO_2}, the R^2, MAE and SRCC values are 0.81, 0.24 and 0.91, respectively. Notably, many low-ranking CoRE MOFs are located near the parity line; however, a significant disparity is evident among high-ranking CoRE MOFs. For S_{CO_2/C_2H_2}, the R^2, MAE and SRCC values are 0.73, 0.28 and 0.84, respectively. There are discrepancies between predictions and simulations at both low- and high-value regions of S_{CO_2/C_2H_2}. It is evident that high CO_2-selective MOFs exhibit comparable rankings in both predictions and simulations. To a certain extent, the ML models exhibit transferable capability from CSD MOFs to CoRE MOFs. Among 1689 CoRE MOFs, 40 were found to exceed the thresholds ($N_{CO_2} > 1.7$ mmol g^{-1} and $S_{CO_2/C_2H_2} > 3$), which

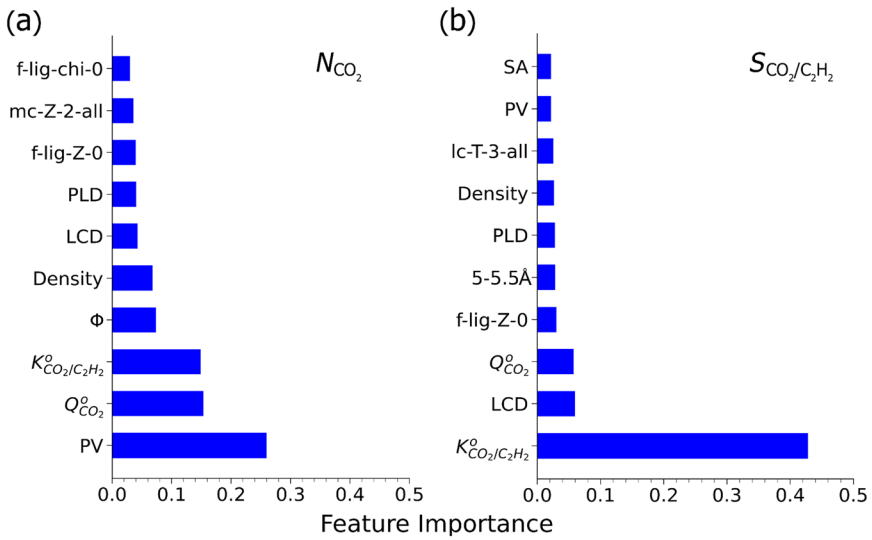

Fig. 4 Top 10 important features in the ML models for (a) N_{CO_2} and (b) S_{CO_2/C_2H_2}.

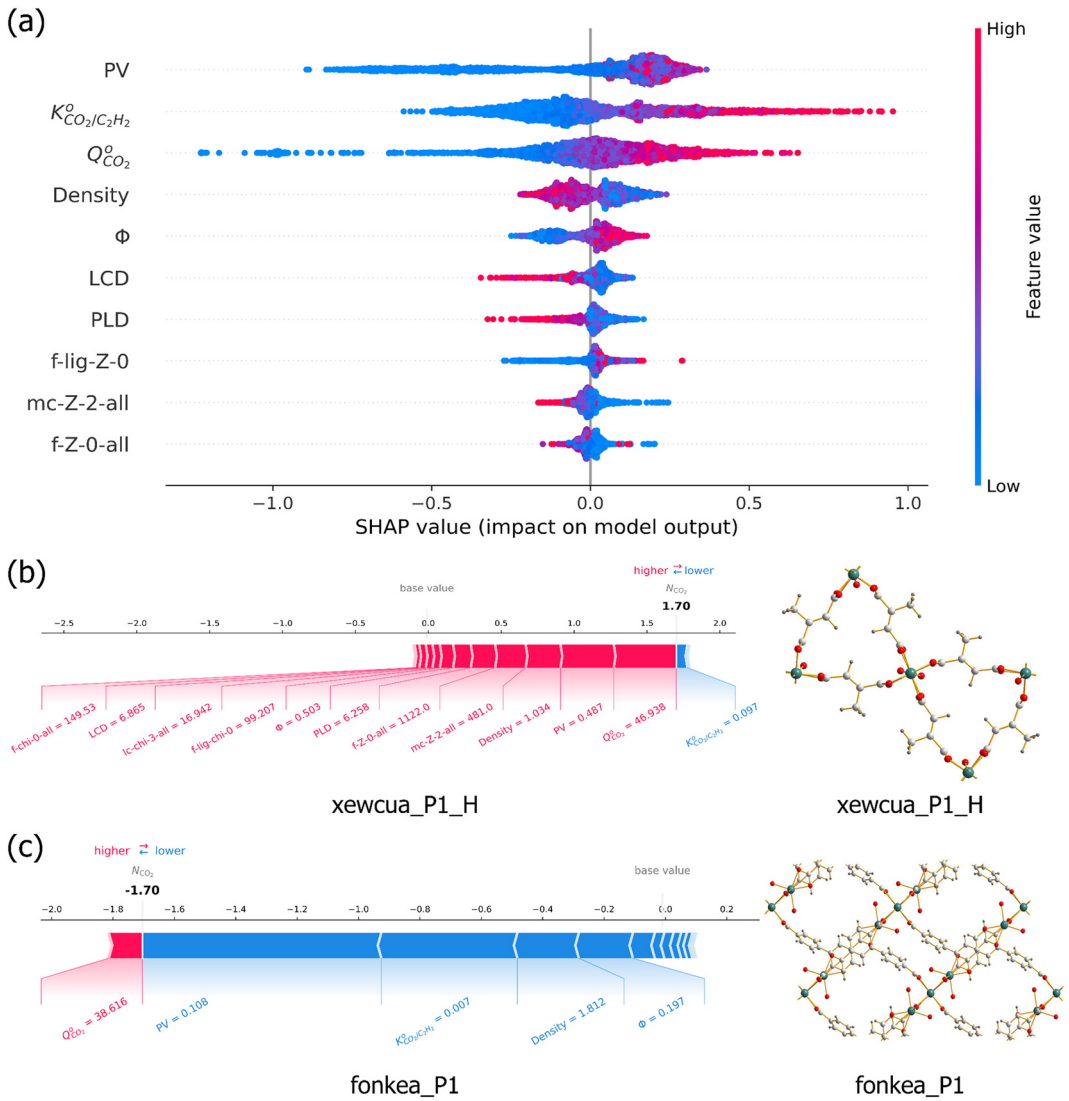

Fig. 5 SHAP analysis for N_{CO_2}. (a) Summary plot of top 10 important features. (b) Force plot for N_{CO_2} in xewcua_P1_H (the highest N_{CO_2} among CSD MOFs), (c) force plot for N_{CO_2} in fonkea_P1 (the lowest N_{CO_2} among CSD MOFs).

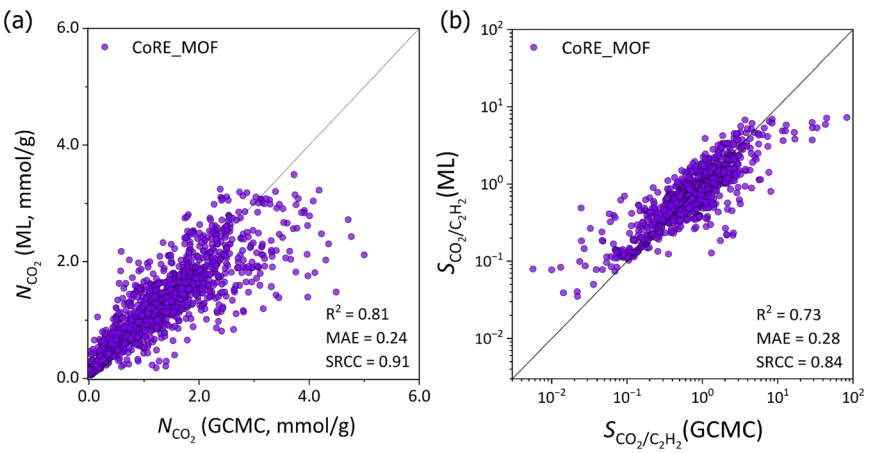

Fig. 6 ML predicted *versus* GCMC simulated (a) N_{CO_2} and (b) S_{CO_2/C_2H_2} in CoRE MOFs.

were also used above to identify best CSD MOFs. After comparing with GCMC simulations, 26 out of these 40 were verified to truly surpass the thresholds. The top seven (as listed in Table 2) possess S_{CO_2/C_2H_2} values ranging from 28.60 to 82.36, which is significantly higher than that in top CSD MOFs (Table S9†). Notably, XISLAM_clean has a S_{CO_2/C_2H_2} of 82.36, along with a substantial N_{CO_2} of 3.76 mmol g^{-1}, positioning it as a promising MOF for CO_2/C_2H_2 separation.

The predictions reveal that the ML models trained upon CSD MOFs have a fairly good transferability to CoRE MOFs. To fundamentally understand, we explore the structural similarity and diversity between the two datasets by using the t-distributed stochastic neighbor embedding (t-SNE) technique. Fig. 7a and b illustrate the t-SNE maps based on the descriptor combination "Geo + PSD_bins + RACs + energy" for the two datasets. The t-SNE maps provide a clear visualization of the distribution of MOFs. Each MOF is represented by a separate point and a cluster of proximate points possesses similar structures. We can observe significant similarity between CSD MOF and CoRE MOF datasets across a broad spectrum of the t-SNE maps. However, notable diversity is seen based on the feature spaces of performance metrics N_{CO_2} and S_{CO_2/C_2H_2} in Fig. 7c and d. The CSD MOFs encompass a slightly wider range of N_{CO_2} than CoRE MOFs, because the number of MOFs differs. Specifically, there are 3229 CSD MOFs and more than 1689 CoRE MOFs, suggesting that the former may possess greater geometric and chemical diversity.

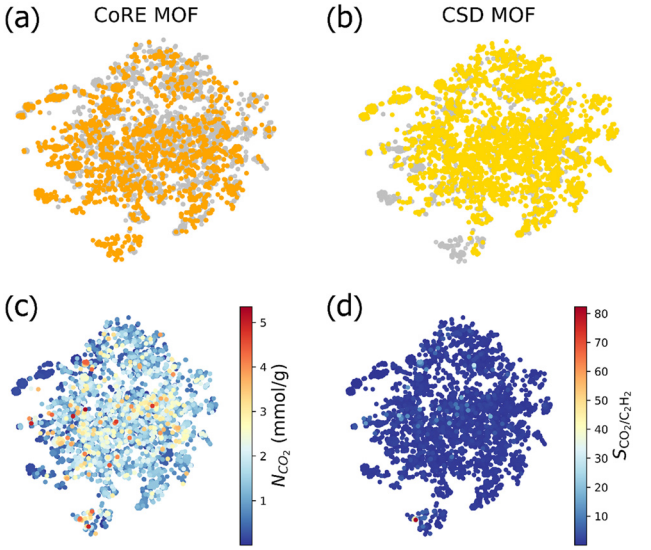

Fig. 7 T-SNE maps based on the descriptor combination "Geo + PSD_bins + RACs + energy". (a) CoRE MOFs in orange and (b) CSD MOFs in yellow. The gray color illustrates the overall feature space of CoRE MOFs and CSD MOFs; each point on the map corresponds to a MOF. The feature spaces are based on (c) N_{CO_2} and (d) S_{CO_2/C_2H_2}.

3.4. Comparison with experiments

In the existing body of literature, several CO_2-selective MOFs were experimentally reported for inverse CO_2/C_2H_2 separation, as detailed in Table 3. These studies typically measure the adsorption capacity of pure CO_2 and estimate CO_2/C_2H_2 selectivity using the IAST. Fig. 8 compares the top-performing CSD and CoRE MOFs identified in this work with experimental MOFs. The well-established trade-off is observed between adsorption capacity and selectivity. Based on experimental measurements, MUF-16,[39] MUF-16(Mn),[39] MUF-16(Ni),[39] Cd-NP,[9] and Mg-NH_4-ZM[40] show good performance and located in the upper right quadrant of the plot, indicating a favorable balance between adsorption capacity and selectivity. At 298 K and 1 bar, notably Cu-F-pymo[41] yields IAST selectivity of 10^5 for equimolar CO_2/C_2H_2 mixture, which is substantially highest as listed in Table 3 and outperforms benchmark CO_2-selective materials such as MUF-16 (ref. 39) (510 at 293 K), PMOF-1(irra)[42] (694 at 273 K) and Cu(hfipbb)(H_2hfipbb)$_{0.5}$ (ref. 43) (696 at 298 K). Nevertheless, its N_{CO_2} is only 1.19 mmol g^{-1}. The 53 CSD MOFs predominantly exhibit $S_{CO_2/C_2H_2} < 20$, rendering them not optimal despite their N_{CO_2} ranging from 1.71 to 5.35 mmol g^{-1}. In contrast, the seven top-performing CoRE MOFs have $S_{CO_2/C_2H_2} > 20$ and high N_{CO_2} ranging from 2.8 to 4.3 mmol g^{-1}. Overall, the CoRE MOFs exhibit superior separation performance as predicted by the ML models and surpass many experimentally reported MOFs in terms of N_{CO_2} and S_{CO_2/C_2H_2}, indicating their significant potential for inverse CO_2/C_2H_2 separation. This underscores the advantage of applying ML to identify top-performing MOFs.

4. Conclusions

We have synergized MS and ML to identify top-performing MOFs for the inverse separation of a CO_2/C_2H_2 mixture. Initially, 3712 CSD MOFs are evaluated for their separation performance

Table 2 Predicted top-performing CoRE MOFs

CoRE MOF	N_{CO_2} (mmol g^{-1})	S_{CO_2/C_2H_2}	PLD (Å)	LCD (Å)	VSA (m^2 cm^{-3})	ϕ	$Q^{\circ}_{CO_2}$ (kJ mol^{-1})
XISLAM_clean	3.76	82.36	3.64	4.34	367.54	0.27	51.51
TAWHOO_clean	3.98	44.48	3.64	4.39	93.43	0.33	47.75
TUYLIJ_clean	4.30	42.86	3.71	4.52	501.42	0.36	49.04
QOLVET_clean	2.80	33.19	3.97	4.18	446.70	0.30	53.78
MORZID_clean	2.96	28.85	3.97	4.20	452.74	0.30	53.12
CAVSUP_clean	3.66	28.61	5.47	6.03	926.71	0.35	61.76
IRAHUF_clean	2.95	28.60	3.66	4.41	402.13	0.35	44.55

Table 3 Experimentally reported MOFs for inverse CO_2/C_2H_2 separation

MOF	Condition (T/K, P/kPa)	N_{CO_2} (mmol g^{-1})	$N_{C_2H_2}$ (mmol g^{-1})	$Q^{\circ}_{CO_2}$ (kJ mol^{-1})	$Q^{\circ}_{C_2H_2}$ (kJ mol^{-1})	Selectivity (50/50)*
MUF-16 (ref. 39)	293, 100	2.13	0.18	32.3	25.8	510
MUF-16 (Mn)[39]	293, 100	2.26	0.44	36.6	N.A.	31
MUF-16 (Ni)[39]	293, 100	2.13	0.32	37.3	N.A.	46
[Mn(bdc)(dpe)][4]	273, 91	2.08	0.32	29.5	27.8	8.8
SIFSIX-3-Ni (ref. 5)	298, 100	2.80	3.30	50.9	36.7	7.5[a]/7.7[b]
$K_2[Cr_3O(OOCH)_6(4\text{-ethylpyridine})_3]_2$ [α-$SiW_{12}O_{40}$] (ionic crystal)[44]	278, 100	0.50	0.10	38.0	30	5.6[a]
CD-MOF-1 (ref. 6)	298, 100	2.87	2.23	41.0	17.6	5.7[a]/6.6[b]
CD-MOF-2 (ref. 6)	298, 100	2.65	2.03	67.2	25.8	12.8[a]/16.0[b]
Cd-NP (ref. 9)	298, 100	2.59	0.43	27.7	N.A.	85
$[Tm_2(OH\text{-}bdc)_2(\mu_3\text{-}OH)_2(H_2O)_2]$[45]	298, 100	5.83	2.10	45.2	17.8	18.2[a]/17.5[b]
$[Tm_2(OH\text{-}bdc)_2(\mu_3\text{-}OH)_2]$[45]	298, 100	6.21	5.25	32.7	26.0	1.65
$[Zn(odip)_{0.5}(bpe)_{0.5}(CH_3OH)]\cdot 0.5NMF\cdot H_2O$ (ref. 46)	298, 100	118.7[d]	39.8[d]	42.3	35.0	13.2
Cu-F-pymo[41]	298, 100	1.19	0.10	28.8	N.A.	>10^5
Ce^{IV}-MIL-140-4F[10]	298, 100	4.92	1.85	39.5	27.4	9.5[b]
PCP-NH2-ipa[8]	298, 100	3.21	1.93	36.6	26.8	6.4[a]
PCP-NH2-bdc[8]	298, 100	3.03	1.90	34.57	25.6	4.4[a]
PMOF-1(bef)[42]	273, 100	47.5[d]	9.5[d]	N.A.	N.A.	50
PMOF-1(irra)[42]	273, 100	53.3[d]	7.5[d]	N.A.	N.A.	694
$[Zn(atz)(BDC\text{-}Cl_4)_{0.5}]$[47]	285, 100	34.6[c]	18[c]	32.7	25.4	2.4[a]
$[Cu(hfipbb)(H_2hfipbb)_{0.5}]$[43]	298, 100	0.74	0.10	25.5	N.A.	696[a]
$Co(HL^{dc})$[48]	195, 100	239.5[d]	140[d]	N.A.	N.A.	N.A.
AlFFIVE-1-Ni[7]	298, 100	2.76	4.6	47	38	N.A.
NbOFFIVE-1-Ni[7]	298, 100	2.16	2.4	54.6	34	N.A.
en-Mg_2(dobpdc)[49]	298, 100	4.48	2.43	71.2	22.3	N.A.
nmen-Mg_2(dobpdc)[49]	298, 100	4.73	2.38	62.32	23.9	N.A.
een-Mg_2(dobpdc)[49]	298, 100	4.85	1.84	68.77	23.4	N.A.
MFU-4 (ref. 50)	300, 100	3.17	N.A.	24.2	NA	3363*
ZU-610a[51]	298, 100	1.51	0.12	27.3	NA	207/1840*
Y-bptc[52]	298, 100	2.48	1.17	31.5	NA	4.1/114*
Mg-NH_4-ZM (ref. 40)	298, 100	3.12	0.44	65	NA	98.0

[a] IAST selectivity for CO_2/C_2H_2 mixture (v/v = 1 : 1). [b] IAST selectivity for CO_2/C_2H_2 mixture (v/v = 1 : 2). [c] Uptake in unit of cm^3 cm^{-3}. [d] Uptake in unit of cm^3 g^{-1}. Value marked with * is the kinetic selectivity (calculated from the ratio of diffusive time constants), while unmarked is the IAST selectivity for CO_2/C_2H_2 mixture (v/v = 1 : 1).

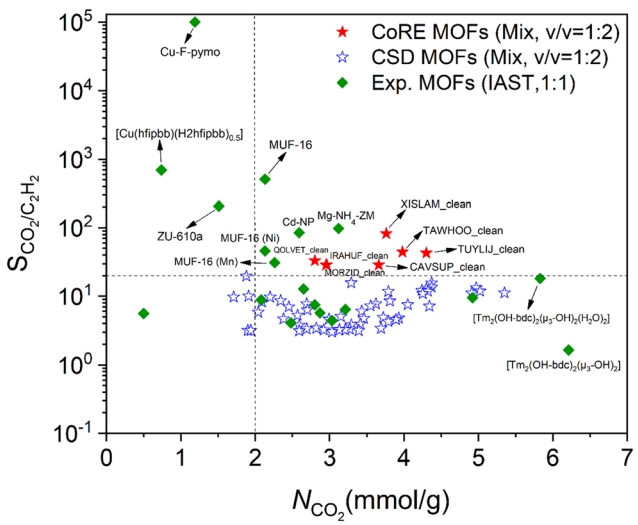

Fig. 8 Comparison of top-performing MOFs identified in this work for CO_2/C_2H_2 mixture (v/v = 1 : 2) with experimental MOFs reported in the literature for CO_2/C_2H_2 mixture (v/v = 1 : 1).

through GCMC simulations. The structure–performance relationships are established for geometric and energetic descriptors (LCD, ϕ, VSA and $Q^{\circ}_{CO_2}$) with performance metrics (N_{CO_2} and S_{CO_2/C_2H_2}). It is found that the separation performance is governed by a complex interplay of multiple factors. With the thresholds of $N_{CO_2} > 1.7$ mmol g^{-1} and $S_{CO_2/C_2H_2} > 3.0$, 53 top-performing CSD MOFs are identified, with their respective LCD, VSA, ϕ, and $Q^{\circ}_{CO_2}$ values ranging from 4.02 to 8.51 Å, 112.63 to 1946.46 m^2 cm^{-3}, 0.139 to 0.581, and 27.93 to 57.21 kJ mol^{-1}. Subsequently, ML models are developed upon the simulation data in CSD MOFs, utilizing geometrical, chemical and energetic features. The model performance is evaluated in terms of both accuracy and interpretability. The inclusion of pore size distribution in the ML models yields only marginal improvement in predictive accuracy; however, large improvement is observed if chemical and energy-related descriptors are included. Feature importance analysis indicates that energy-related $Q^{\circ}_{CO_2}$ and $K^{\circ}_{CO_2/C_2H_2}$ are more significant for accurate predictions than geometrical and chemical descriptors. The transferability of the ML models is assessed by predicting the performance for CO_2/C_2H_2 separation in CoRE MOFs. The out-of-sample predictions in CoRE MOFs agree fairly well with

simulation results, with seven CoRE MOFs predicted to surpass the performance of top CSD MOFs and comparable to many experimentally reported MOFs. The ML models developed in this study are useful for the development of new MOFs for inverse CO_2/C_2H_2 separation and many other important applications. On the other hand, we are aware of the key limitations in our simulations for constructing datasets. First, the widely used universal force field is adopted here; however, its generalizability and transferability to the adsorption of CO_2/C_2H_2 mixture needs to be further verified. Second, the frameworks are assumed to be rigid and hence possible structural flexibility is not taken into account. In this regard, it is desired to develop more accurate force fields (*e.g.*, ML potentials) and incorporate structural flexibility for future improvement.

Data availability

The data supporting this article have been included as part of the ESI.†

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

We gratefully acknowledge A*STAR LCER-FI (LCERFI01-0015 U2102d2004), National Research Foundation Singapore (NRF-CRP26-2021RS-0002), Ministry of Education of Singapore (R-279-000-578-112, R-279-000-598-114, R-279-000-574-114) for financial support.

References

1 A. Granada, S. B. Karra and S. M. Senkan, *Ind. Eng. Chem. Res.*, 1987, **26**, 1901–1905.
2 X. P. Fu, Y. L. Wang and Q. Y. Liu, *Dalton Trans.*, 2020, **49**, 16598–16607.
3 R. Matsuda, R. Kitaura, S. Kitagawa, Y. Kubota, R. V. Belosludov, T. C. Kobayashi, H. Sakamoto, T. Chiba, M. Takata, Y. Kawazoe and Y. Mita, *Nature*, 2005, **436**, 238–241.
4 M. L. Foo, R. Matsuda, Y. Hijikata, R. Krishna, H. Sato, S. Horike, A. Hori, J. Duan, Y. Sato, Y. Kubota, M. Takata and S. Kitagawa, *J. Am. Chem. Soc.*, 2016, **138**, 3022–3030.
5 K.-J. Chen, H. S. Scott, D. G. Madden, T. Pham, A. Kumar, A. Bajpai, M. Lusi, K. A. Forrest, B. Space, J. J. Perry and M. J. Zaworotko, *Chem*, 2016, **1**, 753–765.
6 L. Li, J. Wang, Z. Zhang, Q. Yang, Y. Yang, B. Su, Z. Bao and Q. Ren, *ACS Appl. Mater. Interfaces*, 2019, **11**, 2543–2550.
7 Y. Belmabkhout, Z. Zhang, K. Adil, P. M. Bhatt, A. Cadiau, V. Solovyeva, H. Xing and M. Eddaoudi, *Chem. Eng. J.*, 2019, **359**, 32–36.
8 Y. Gu, J. J. Zheng, K. I. Otake, M. Shivanna, S. Sakaki, H. Yoshino, M. Ohba, S. Kawaguchi, Y. Wang, F. Li and S. Kitagawa, *Angew. Chem., Int. Ed.*, 2021, **60**, 11688–11694.
9 Y. Xie, H. Cui, H. Wu, R. B. Lin, W. Zhou and B. Chen, *Angew. Chem., Int. Ed.*, 2021, **60**, 9604–9609.
10 Z. Zhang, S. B. Peh, R. Krishna, C. Kang, K. Chai, Y. Wang, D. Shi and D. Zhao, *Angew. Chem., Int. Ed.*, 2021, **60**, 17198–17204.
11 H. Tang, L. Duan and J. W. Jiang, *Langmuir*, 2023, **39**, 15849–15863.
12 X. Zhang, T. Zhou and K. Sundmacher, *AIChE J.*, 2021, **68**, e17524.
13 C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chem. Mater.*, 2015, **27**, 4459–4475.
14 R. Wang, Y. Zhong, L. Bi, M. Yang and D. Xu, *ACS Appl. Mater. Interfaces*, 2020, **12**, 52797–52807.
15 P. Yang, G. Lu, Q. Yang, L. Liu, X. Lai and D. Yu, *Green Energy Environ.*, 2022, **7**, 1062–1070.
16 P. Halder and J. K. Singh, *Energy Fuels*, 2020, **34**, 14591–14597.
17 H. J. Tang, Q. S. Xu, M. Wang and J. W. Jiang, *ACS Appl. Mater. Interfaces*, 2021, **13**, 53454–53467.
18 M. Zhou and J. Wu, *npj Comput. Mater.*, 2022, **8**, 256.
19 J. Hu, J. Cui, B. Gao, L. Yang, Q. Ding, Y. Li, Y. Mo, H. Chen, X. Cui and H. Xing, *Matter*, 2022, **5**, 3901–3911.
20 A. Li, R. B. Perez, S. Wiggin, S. C. Ward, P. A. Wood and D. Fairen-Jimenez, *Matter*, 2021, **4**, 1105–1106.
21 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
22 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
23 S. Kancharlapalli, A. Gopalan, M. Haranczyk and R. Q. Snurr, *J. Chem. Theory Comput.*, 2021, **17**, 3052–3064.
24 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Mol. Simul.*, 2015, **42**, 81–101.
25 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. I. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.
26 H. A. Lorentz, *Ann. Phys.*, 1881, **248**, 127–136.
27 P. G. Boyd, S. M. Moosavi, M. Witman and B. Smit, *J. Phys. Chem. Lett.*, 2017, **8**, 357–363.
28 D. E. Coupry, M. A. Addicoat and T. Heine, *J. Chem. Theory Comput.*, 2016, **12**, 5215–5225.
29 M. A. Addicoat, N. Vankova, I. F. Akter and T. Heine, *J. Chem. Theory Comput.*, 2014, **10**, 880–891.
30 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
31 E. I. Ioannidis, T. Z. Gani and H. J. Kulik, *J. Comput. Chem.*, 2016, **37**, 2106–2117.
32 T. Zhang and B. Yang, *Technometrics*, 2017, **59**, 189–201.
33 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
34 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
35 S. M. Lundberg and S.-I. Lee, *arXiv*, 2025, preprint, arXiv:1705.07874, DOI: **10.48550/arXiv.1705.07874**, (accessed 15 Jan. 2025).
36 Z. Zhang, H. Tang, M. Wang, B. Lyu, Z. Y. Jiang and J. W. Jiang, *ACS Sustainable Chem. Eng.*, 2023, **11**, 8148–8160.

37 Z. Wang, Y. Zhou, T. Zhou and K. Sundmacher, *Comput. Chem. Eng.*, 2022, **160**, 107739.
38 Z. Wang, T. Zhou and K. Sundmacher, *Chem. Eng. J.*, 2022, **444**, 136651.
39 O. T. Qazvini, R. Babarao and S. G. Telfer, *Nat. Commun.*, 2021, **12**, 197.
40 B. Ma, D. Li, Q. Zhu, Y. Li, W. Ueda and Z. Zhang, *Angew. Chem., Int. Ed.*, 2022, **61**, e202209121.
41 Y. Shi, Y. Xie, H. Cui, Y. Ye, H. Wu, W. Zhou, H. Arman, R. B. Lin and B. Chen, *Adv. Mater.*, 2021, **33**, e2105880.
42 L.-Z. Cai, Z.-Z. Yao, S.-J. Lin, M.-S. Wang and G.-C. Guo, *Angew. Chem., Int. Ed.*, 2021, **60**, 18223–18230.
43 H. Cui, Y. Xie, Y. Ye, Y. Shi, B. Liang and B. Chen, *Bull. Chem. Soc. Jpn.*, 2021, **94**, 2698–2701.
44 R. Eguchi, S. Uchida and N. Mizuno, *Angew. Chem., Int. Ed.*, 2012, **51**, 1635–1639.
45 D. Ma, Z. Li, J. Zhu, Y. Zhou, L. Chen, X. Mai, M. Liufu, Y. Wu and Y. Li, *J. Mater. Chem. A*, 2020, **8**, 11933–11937.
46 L. N. Ma, G. D. Wang, L. Hou, Z. Zhu and Y. Y. Wang, *ACS Appl. Mater. Interfaces*, 2022, 26858–26865.
47 X.-Y. Li, Y. Song, C.-X. Zhang, C.-X. Zhao and C. He, *Sep. Purif. Technol.*, 2021, **279**, 119608.
48 W. Yang, A. J. Davies, X. Lin, M. Suyetin, R. Matsuda, A. J. Blake, C. Wilson, W. Lewis, J. E. Parker, C. C. Tang, M. W. George, P. Hubberstey, S. Kitagawa, H. Sakamoto, E. Bichoutskaia, N. R. Champness, S. Yang and M. Schröder, *Chem. Sci.*, 2012, **3**, 2993–2999.
49 D. S. Choi, D. W. Kim, D. W. Kang, M. Kang, Y. S. Chae and C. S. Hong, *J. Mater. Chem. A*, 2021, **9**, 21424–21428.
50 Q. Liu, S. G. Cho, J. Hilliard, T. Y. Wang, S. C. Chien, L. C. Lin, A. C. Co and C. R. Wade, *Angew. Chem., Int. Ed.*, 2023, e202218854.
51 J. Cui, Z. Qiu, L. Yang, Z. Zhang, X. Cui and H. Xing, *Angew. Chem., Int. Ed.*, 2022, **61**, e202208756.
52 C. He, P. Zhang, Y. Wang, Y. Zhang, T. Hu, L. Li and J. Li, *Sep. Purif. Technol.*, 2023, **304**, 122318.