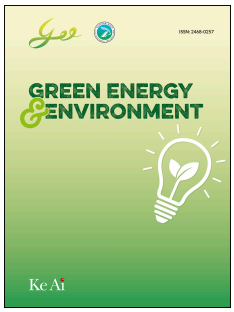
Journal Pre-proof



Machine Learning and High-Throughput Computational Screening of Hydrophobic Metal–Organic Frameworks for Capture of Formaldehyde from Air

Xueying Yuan, Xiaomei Deng, Chengzhi Cai, Zenan Shi, Hong Liang, Shuhua Li, Zhiwei Qiao

PII: S2468-0257(20)30097-2

DOI: https://doi.org/10.1016/j.gee.2020.06.024

Reference: GEE 258

To appear in: *Green Energy and Environment*

Received Date: 16 April 2020

Revised Date: 6 June 2020

Accepted Date: 24 June 2020

Please cite this article as: X. Yuan, X. Deng, C. Cai, Z. Shi, H. Liang, S. Li, Z. Qiao, Machine Learning and High-Throughput Computational Screening of Hydrophobic Metal–Organic Frameworks for Capture of Formaldehyde from Air, *Green Energy & Environment*, https://doi.org/10.1016/j.gee.2020.06.024.

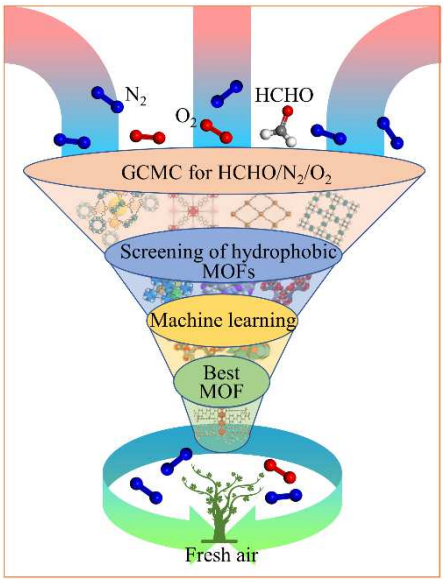
Article

Machine Learning and High-Throughput Computational Screening of Hydrophobic Metal–Organic Frameworks for Capture of Formaldehyde from Air

Xueying Yuan, Xiaomei Deng, Chengzhi Cai, Zenan Shi, Hong Liang, Shuhua Li, and Zhiwei Qiao*

Guangzhou Key Laboratory for New Energy and Green Catalysis, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou 510006, PR China.

**Corresponding author. E-mail: zqiao@gzhu.edu.cn*



The combination of machine learning and high-throughput computational screening were employed to calculate and to identify the top-performing hydrophobic metal–organic frameworks for the removal of formaldehyde from N_2 and O_2.

Abstract

Aiming to efficiently capture the formaldehyde (HCHO) with low content in the air exceeding the standard, 31,399 hydrophobic metal–organic frameworks (MOFs) were first selected from 137,953 hypothetical MOFs to calculate their formaldehyde adsorption performance, namely, adsorption capacity (N_{HCHO}) and selectivity ($S_{HCHO/N2+O2}$) by molecular simulation and machine learning (ML). To combine the $S_{HCHO/N2+O2}$ and N_{HCHO}, a new performance metric, the tradeoff between selectivity and capacity (TSC) was proposed to identify more reasonably the top-performing MOFs. The MOFs were divided into three datasets (i.e., all of the MOFs (AM), MOFs with top 5% of $S_{HCHO/N2+O2}$ (P_S) and MOFs with top 5% of N_{HCHO} (P_N)) to scrutinize and explore the characteristics of different materials capturing formaldehyde from the air (N_2 and O_2). Furthermore, after four ML algorithms (the back propagation neural network (BPNN), support vector machine (SVM), extreme learning machine (ELM), and random forest (RF)) are applied to quantitatively assess the prediction effects of performance indexes in different datasets, RF algorithm with the most accurate prediction revealed that the TSC has strong correlations with the MOF descriptors in P_S dataset. In view of 14.10% of the promising MOFs occupied P_N, the design paths of excellent adsorbents for six MOF descriptors were quantitatively determined, especially for the Henry's coefficient (K_{HCHO}) and heat of adsorption of formaldehyde (Q^0_{st}). Their probabilities of obtaining excellent MOFs could reach 100% and 77.42%, respectively, and both the relative importance and the trends of univariate analysis coherently confirm the important positions of K_{HCHO} and Q^0_{st}. Finally, 20 best MOFs were identified for the single-step separation of formaldehyde with low concentration. The microscopic insights and structure-performance relationship predictions from this computational and ML study are useful toward the development of new MOFs for the capture of formaldehyde from air.

Keywords: molecular simulation, adsorption, metal-organic framework, formaldehyde

1. Introduction

With rapid global industrialization, air pollution is becoming increasingly severe, and volatile organic compounds (VOCs) are considered to be a primary pollution source [1]. As one of the most common and representative poisonous VOCs, formaldehyde (HCHO) almost appears in every corner of human life, including coatings, household items, cosmetics, and clothing. Not only a trace amount of formaldehyde can irritate the respiratory mucosa, skin, and eyes [2]; but also it can even cause tumors and cancer [3]. According to current scientific research, formaldehyde concentration above 0.10 mg/m^3 can stimulate mucosa of the respiratory tract and cause discomfort in the throat. At a concentration of 10 mg/m^3, formaldehyde tends to cause respiratory tract diseases and breathing difficulties. When the content of formaldehyde reaches 30 mg/m^3, it can cause death [4-6]. Therefore, the purification and removal of formaldehyde from the air is vital for human health. At present, the traditional methods of formaldehyde cleansing mainly include photocatalysis [7], catalytic oxidation [2], and adsorption [8]. Photocatalysis has a wide range of applications and a high degradation rate, but the energy and activity of the catalyst are low at ambient temperatures, and toxic outgrowths are easily produced under ultraviolet irradiation. The catalytic oxidation method is simple to conduct, consumes little energy, and produces no pollution. Nevertheless, it requires strict control of the initial temperature, and the catalyst activity is low. In general, the photocatalytic method and catalytic oxidation have remarkable effects for one-off treatments, but they are not appropriate for domestic environments in which formaldehyde is produced for several years. Neither of them is suitable for addressing the continuous release of formaldehyde. So far, owing to its advantages, such as achieving complete reactions, operating in mild conditions, producing no secondary pollution, and being easily regenerated, the adsorption method has been widely used. The adsorbents play a decisive role in the adsorption process, mainly including zeolites, activated carbon, biological-based adsorbents, and metal-based adsorbents. The main reason for the unsatisfactory adsorption effect of these traditional adsorbents is that water molecules with polar functional groups occupy the adsorption sites preferentially, and the strong competitive adsorption of water molecules hinders the adsorption of formaldehyde [9]. To eliminate the competitive adsorption of water [10], in the present work, we examined hydrophobic adsorbents that could continuously absorb formaldehyde.

The emergence of new materials is important for the construction of environmentally friendly communities, and it also promotes socially sustainable development and scientific innovation. In recent years, a class of porous coordination polymers called metal–organic frameworks (MOFs) has become a research focus in the new materials field and has drawn extensive attention in academia. Metal ions or clusters and organic links self-assemble to form MOFs [11], which have flexible frameworks and are also known as "soft zeolites." However, compared with traditional inorganic materials, such as zeolites, they have larger specific surface areas, porosities, and potential for functionalization of their pore structures and diversification of their designs [12]. The structural characteristics of MOFs, including their topologies, pore sizes, shapes, and surface chemistries, can be adjusted to meet actual requirements based on their building block properties [13]. MOFs exhibit excellent performance in various applications, including gas adsorption and storage [14-17], separation [18-20], catalysis [21, 22], sensing [23, 24], and biomedical applications [25], as well as remarkable thermal and chemical stabilities [26]. For their exceptional properties, MOFs have attracted scholars and experts in various fields to conduct in-depth research. More than 300,000 kinds of MOFs with strong CO_2 binding sites in a humid environment were proven to exist by Boyd *et al*. recently, who synthesized two strong hydrophobic MOFs (hMOFs) with high CO_2 capture values [27]. Reddy and co-workers used three kinds of Zn-MOFs as chemical capacitance sensors to detect ammonia, formaldehyde, and ethanol gases with satisfactory responses at room temperature [28]. Wang *et al*. prepared a γ-CD-MOF-K with high-speed formaldehyde adsorption capabilities using γ-cyclodextrin and potassium ions, evaluated the contributions of the hydrogen bonds and host-guest interactions to the adsorption process [29].

The arrangement of organic links and metal nodes, as well as the combination of topological structures, is almost endless. Together with a large number of experimental MOFs, there are also thousands of hypothetical MOFs, which have been designed by computers [30]. So difficult is it to search for MOFs with special properties in such a huge MOF database only through trial and error that helpful theoretical support could guild the screening of candidate materials to save time and resources is quite necessary. High-throughput computational screening (HTCS), which is based on molecular simulation

(MS) and machine learning (ML), has become an efficient method to screen and estimate the performance limits of MOFs in recent decades. The structure–performance behaviors of materials can be analyzed and predicted through data mining technology and ML, and thereby, MOFs with the greatest application prospects can be screened and designed [31]. For example, Bian *et al*. [32] conducted the grand canonical Monte Carlo (GCMC) simulations on 2,932 kinds of MOFs, selected MOFs that possessed excellent adsorption capacities for the adsorption of high concentrations of formaldehyde, and prepared Y-BTC, ZnCar, and Ni-BIC materials. This proved the feasibility and accuracy of HTCS to guide MOFs for adsorbing formaldehyde in the experiment. To separate isomers, Peng *et al*. [33] obtained 13,512 candidate hypothetical MOFs using HTCS. Based on their structural characteristics and the absorption performance score (APS), MOF-163, with the highest APS, was found to be able to separate two dimethylbutanes from hexane isomers efficiently. Using MS, eight MOFs with different compositions and structures were utilized by Gonçalves and co-workers to research the adsorption of methane in natural gas. They found that CO_2 and H_2O in natural gas were likely to affect MOFs with open metal sites, leading to decreased methane adsorption [34]. By applying density functional theory (DFT), GCMC, and ML, Anderson *et al*. illustrated the effects of pore chemical and topology structures in the promotion of CO_2 capture metrics of MOFs. After introducing the molecular building block-based charges, the complex relationship between the GCMC simulations and DFT calculations was established [35]. Bobbitt and Snurr studied the adsorption principles and phenomena of three toxic gases (ammonia, phosphine, and arsine) on 33 metal catecholates in detail. They concluded that the adsorption of ammonia was mainly impacted by the Coulomb effect, and the electron density determined the binding effect of phosphine and arsine with the adsorbent [36].

GCMC simulation, as a powerful tool, has already been applied to simulate and count the six characteristic descriptors and two performance indicators of 31,399 hMOFs in this study. The tradeoff between selectivity and capacity (TSC) was calculated as the third performance standard, the properties of the materials were evaluated by ML, and a firm connection between the structure and performance was established. MOFs as adsorbents are commonly used for the removal of VOCs.

However, there have been few reports on formaldehyde capture. For instances, at ambient temperature, the adsorption capacity of MOF-5 under a formaldehyde atmosphere with a concentration of 0.028 mg/m^3 was 3.89×10^{-6} mol/kg [39]. When the initial formaldehyde concentration was stabilized at 100 mg/m^3, the adsorbed amount of Y-BTC was 0.38 mol/kg [32]. In a formaldehyde environment with a concentration of 0.48 mg/m^3, γ-CD-MOF-K could be absorbed almost completely in 15 min [29]. In a typical environment occupied by humans, the concentration of formaldehyde will not be too high, that may endanger human health. Hence, we set the original concentration of formaldehyde in the air in the GCMC simulations to 13.41 mg/m^3, and the specific molar ratio of N_2:O_2:HCHO was 7.81×10^5:2.19×10^5:1.

2. Models and Methods

2.1. Molecular Models

2.1.1. Models

The 102 building blocks and 6 different topologies were combined into a large crystallographic dataset for the design of 137,953 hypothetical MOFs by Wilmer *et al.* [40]. In this work, to avoid the competitive adsorption of water vapor, 31,399 hMOFs were first selected from Wilmer *et al.*'s MOF database [40]. Six descriptors (the void fraction (ϕ), the volumetric surface area (VSA), the largest cavity diameter (LCD), the Henry's coefficient of formaldehyde (K_{HCHO}), the heat of adsorption (Q^0_{st}), and the density (ρ)) were used to describe the structural/energetic features of MOFs. Using the RASPA [41] software package, ϕ was estimated by detecting the framework using a He atom with a 2.58 Å in diameter. Taking advantage of a simple MC integration technique, the VSA was counted by rolling an N_2 probe with a diameter of 3.64 Å on the framework surfaces of the MOFs. The LCD was estimated using Voronoi network by Zeo++ [42]. The adsorption energies of different gas molecules calculated under infinite dilution conditions to determine K_{HCHO} and Q^0_{st} using the *NVT*-MC (*N* is the number of particles, *V* is the volume of system, *T* is the temperature of the system) in RASPA [41].

2.1.2. Force field parameter

In this work, the hydrophilicity and hydrophobicity of MOFs were determined by the Henry's coefficient of water (K_H) referring to the boundary of ZIF-8. This method was verified in previous works [43, 44]. Fig. S1 illustrates the relationship between ϕ and K_H of 137,953 MOFs. When $\phi < 0.1$, the K_H is characterized by steady growth. Because the increasing of pore volume could enhance the interactions between the water molecules and the frameworks, when the pore MOF is close or smaller than the size of water. However, when the ϕ exceeded 0.1, K_H tended to vary gently as the porous and dense MOFs hindered the intimate contact between the water and the frameworks. Since $K_H = 2.6 \times 10^6$ mmol/g/Pa in the highly hydrophobic ZIF-8, and a low K_H represents a weak binding ability of the framework with water, hMOFs below this limit were screened out. After this, GCMC simulations were used to statically simulate and calculate the absorption capacity (N_{HCHO}) and selectivity ($S_{HCHO/N2+O2}$) values of 31,399 hMOFs for HCHO in N_2 and O_2. Although the structures of the MOFs could be regarded as rigid and hard to deform, the interactions between the formaldehyde and MOFs as well as the potential energy of formaldehyde must be considered. The potential energy interactions between the atoms were described by the Lennard-Jones (LJ) plus electronic potentials:

$$u_{\mathrm{LJ+elec}}(r) = \sum 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum \frac{q_i\, q_j}{4\pi\varepsilon_0 r_{ij}} \qquad (1)$$

where i, j are indices that represent the two atoms interacting in the force field, the minimum distance when they collide with each other is expressed by σ_{ij}, $u_{\mathrm{LJ+elec}}(r)$ is the interaction energy between the atoms, $\varepsilon_o = 8.8542 \times 10^{-12}$ $C^2 \cdot N^{-1}$ is the vacuum dielectric constant, ε_{ij} and r_{ij} represent the well depth and the real distance between the atoms, respectively, and q_i, q_j are the charge numbers of atoms i, j, respectively. The detailed LJ parameters of all the MOFs are listed in Table S1, whose potential energy parameters come from the full periodic table obtained by Rappé *et al*. using a universal force field (UFF) [45]. The MOF electrostatic-potential-optimized charge scheme (MEPO-Qeq) [46] uses a stable and efficient method to generate atomic charges, which could be rapidly calculated and evaluated for big data. Thus, accurate screening can be carried out for finding MOFs to adsorb specific gases.

For the formaldehyde molecule shown in Fig. S2, the three σ bonds connected with the carbon were in the same plane, which was perpendicular to the π bond formed by overlapping a P orbit of C and a P orbit of O. The bond angle of H—C=O was 121.8°, and the bond lengths of H—C and C=O were 1.101 and 1.203 Å, respectively. Due to its central position in the formaldehyde molecular model, the energy parameter σ of C was larger than those of O and H. The charge numbers of C and O were +0.45e and −0.45e, respectively. Hence, there was a dipole moment from the center of the positively charged C to the center of the negatively charged O in the direction of the C=O bond, whose size was 2.6 D. This information was based on the work of Hantal [47], who has listed the force field parameters of formaldehyde in detail. N_2 and O_2, which are the primary components of air, were placed in the space as a three-point model with a partially charged center-of-mass. Their force field parameters were obtained from the transferable potentials for phase equilibria (TraPPE) [48]. The force field arguments of all the adsorbates, namely HCHO, N_2, and O_2, are summarized in Table S2.

2.1.3. GCMC simulation

To simulate the adsorption performance of formaldehyde in the air in a realistic environment, the mimicked mixed gas (HCHO, N_2, and O_2) was assumed to be at 298 K and 1 bar. For each MOF, the GCMC simulation performed an independent calculation to predict their adsorption abilities for the ternary gas mixture. The interactions between the MOFs and formaldehyde, N_2, and O_2 were counted by the Lorentz–Berthelot rules. Periodic boundaries were applied in the spatial coordinate system, and the simulated crystal cells extended to more than 24 Å along the three-dimensional direction. To count the LJ interactions, the spherical truncation radius of the long-range correction was set to 12 Å. The electrostatic interactions between the frameworks and gas molecules and between the gas molecules were calculated by adopting Ewald summation [49]. The GCMC simulation executed 10,000 cycles in each MOF. The purpose of the first 5,000 cycles was to bring the simulation system into balance, and the last 5,000 cycles were used to acquire the average value. Each cycle was made up of *n* simulation movements (*n* is the number of adsorbate molecules), which include translations, rotations, regrowth, and exchanges, where insertions and deletions were regarded as exchange motions. After the

measurement, the impact of additional cycles on the simulation could be ignored.

3. Results and discussion

3.1. Tradeoff between selectivity and capacity

This work devoted to screen the optimal adsorbents for formaldehyde separation in ternary gas mixtures. Since the performance indicators of MOFs, such as the selective adsorption ability and high adsorption capacity (e.g., $S_{HCHO/N2+O2}$ and N_{HCHO}), have become the criterions to together determine the ideal adsorbent, it is crucial to define a new comprehensive indicator, which strongly correlated with N_{HCHO} and $S_{HCHO/N2+O2}$. Thus, the tradeoff between selectivity and capacity (TSC) in equation 2 was proposed in this work.

$$\mathrm{TSC} = \ln(N_{\mathrm{HCHO}}) \cdot \log_{10}(\log_{10}(S_{\mathrm{HCHO/N_2+O_2}})) \quad (2)$$

where the unit of N_{HCHO} is mol/g. On the basis of Shah *et al.*'s work [50], Yang *et al.* [51], for the first time, presented the TSC for the separation of H_2S and CO_2. By comparing the Pearson coefficients of three tradeoff methods, TSC was verified to possess the strongest correlations with the MOF descriptors [51].

3.2. Machine Learning

After HTCS, 2179 hMOFs are first excluded due their N_{HCHO} = 0, the data of 29,220 hMOFs (31,399 - 2179) are further discussed and performed both the multivariate analysis and univariate analysis. Fig. 1 shows the relationship among N_{HCHO}, $S_{HCHO/N2+O2}$, and TSC. Based on the distribution of the points and the linear fit, we concluded that N_{HCHO} and $S_{HCHO/N2+O2}$ were positively correlated because the initial concentration of formaldehyde was low (13.41 mg/m^3). The N_{HCHO} and $S_{HCHO/N2+O2}$ values of the MOFs in the upper right corner of Fig. 1 are suitable for formaldehyde adsorption, while those in the lower left corner are unsuitable. The TSC values, represented by the color mapping in Fig. 1, were

consistent with the trend of the linear fit, which reflects the distribution of the comprehensive performance of MOFs. The higher the TSC value is, the more likely it is that the MOF possessed large values of N_{HCHO} and $S_{HCHO/N2+O2}$, that is, superior adsorption ability. To explore the commonness of high-performance MOFs by ML, two sub-datasets based on the limits of N_{HCHO} and $S_{HCHO/N2+O2}$ were established, respectively. For N_{HCHO}, top 5% of hMOFs (1461 out of 29,220) were selected and built the P_N sub-dataset, and the 1461 hMOFs have $N_{HCHO} > 16.17\times10^{-5}$ mol/kg. For $S_{HCHO/N2+O2}$, the similar sub-dataset of P_S was built including 1461 hMOFs with top 5% $S_{HCHO/N2+O2}$ ($S_{HCHO/N2+O2} > 1400.45$), as shown in Fig. 1. All of the MOFs (containing P_N, P_S, and the remaining MOFs) were denoted as AM; this group contained 29,220 MOFs. The training and prediction of ML algorithms were executed for these three datasets to determine the structure–performance relationships of the whole MOF database and the excellent MOFs.

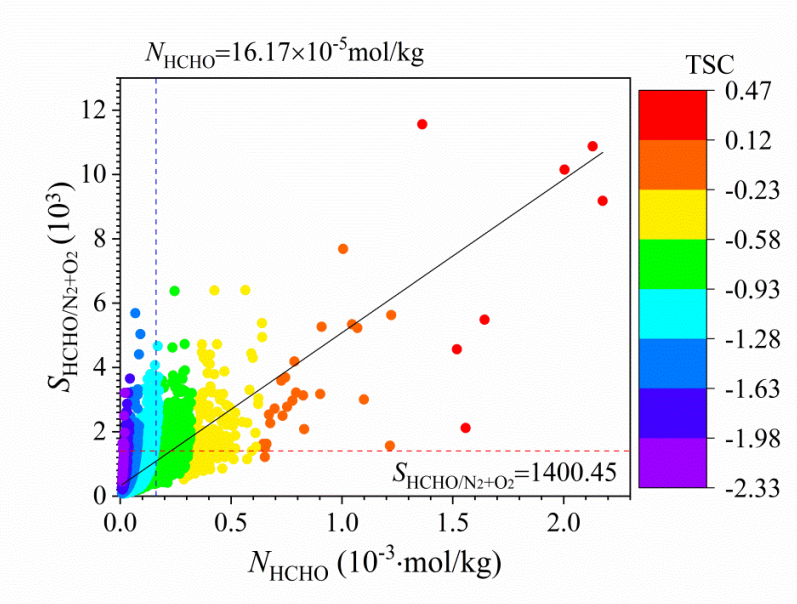


Fig. 1. Selectivity ($S_{HCHO/N2+O2}$) and absorption capacities (N_{HCHO}) of all MOFs (AM dataset). The Ps dataset contained the data above the red line. The P_N dataset contained data to the right of the blue line. The black line shows the overall linear fit line.

Four ML methods (the back propagation neural network (BPNN), support vector machine (SVM), extreme learning machine (ELM), and random forest (RF)) were employed to predict three performance indicators (N_{HCHO}, $S_{HCHO/N2+O2}$, and TSC) of three datasets (AM, P_S, and P_N). The

principles of four MLs are shown in Figs. S3–S6. For the overall predictions of these ML methods, in which the *k* times repeated *k*-fold cross-validation with $k = 5$ was shown in Fig. S7. The final accuracy of the model was evaluated by the Pearson correlation coefficient *R*, mean absolute error (MAE), and root mean square error (RMSE), which come from the average of 25 similar runs with random training and test sets. Note that a higher value of *R* and lower values of MAE and RMSE represent a better prediction in a given ML system.

Table 1 lists the evaluation of 4 ML to predict the TSC of AM and P_S. The predicted outcomes of TSC in the AM are displayed in Fig. 2. Only the RF and BPNN reached the *R* values above 0.790, and their MAE and RMSE values are much smaller than those of SVM and ELM. Fig. 3 shows the prediction effects of four ML methods for TSC in the P_S. Compared with the evaluation indexes (*R*, MAE and RMSE) of AM, the predictions of P_S sub-dataset were more agreement with the simulated results, and the highest *R* by RF algorithm could reach 0.867. Similarly, the RF and BPNN for P_S sub-dataset still exhibited more outstanding prediction effect than SVM and ELM. Given that the data composition of AM and P_S are different, it is the *R* instead of MAE or RMSE as the impartial standard to evaluate the different models more accurate. The evaluation of 4 ML for P_N-TSC, N_{HCHO}, and $S_{HCHO/N2+O2}$ are tabulated in Table S3 and find that, whatever the datasets of MOFs are, the TSC and N_{HCHO} possess better prediction than $S_{HCHO/N2+O2}$.

Table 1. Evaluation of 4 ML algorithms for the TSC of AM and P_S

Performance	Datasets	ML	Cross validation			Test set		
			R	MAE	RMSE	*R*	MAE	RMSE
TSC	AM	BPNN	0.820	0.100	0.128	0.794	0.101	0.139
		SVM	0.778	0.109	0.140	0.775	0.110	0.141
		ELM	0.748	0.115	0.148	0.742	0.116	0.150
		RF	0.886	0.081	0.105	0.817	0.100	0.129
	P_S	BPNN	0.886	0.123	0.157	0.856	0.133	0.177
		SVM	0.856	0.137	0.176	0.846	0.142	0.181
		ELM	0.841	0.144	0.182	0.817	0.148	0.195
		RF	0.922	0.102	0.134	0.867	0.131	0.169

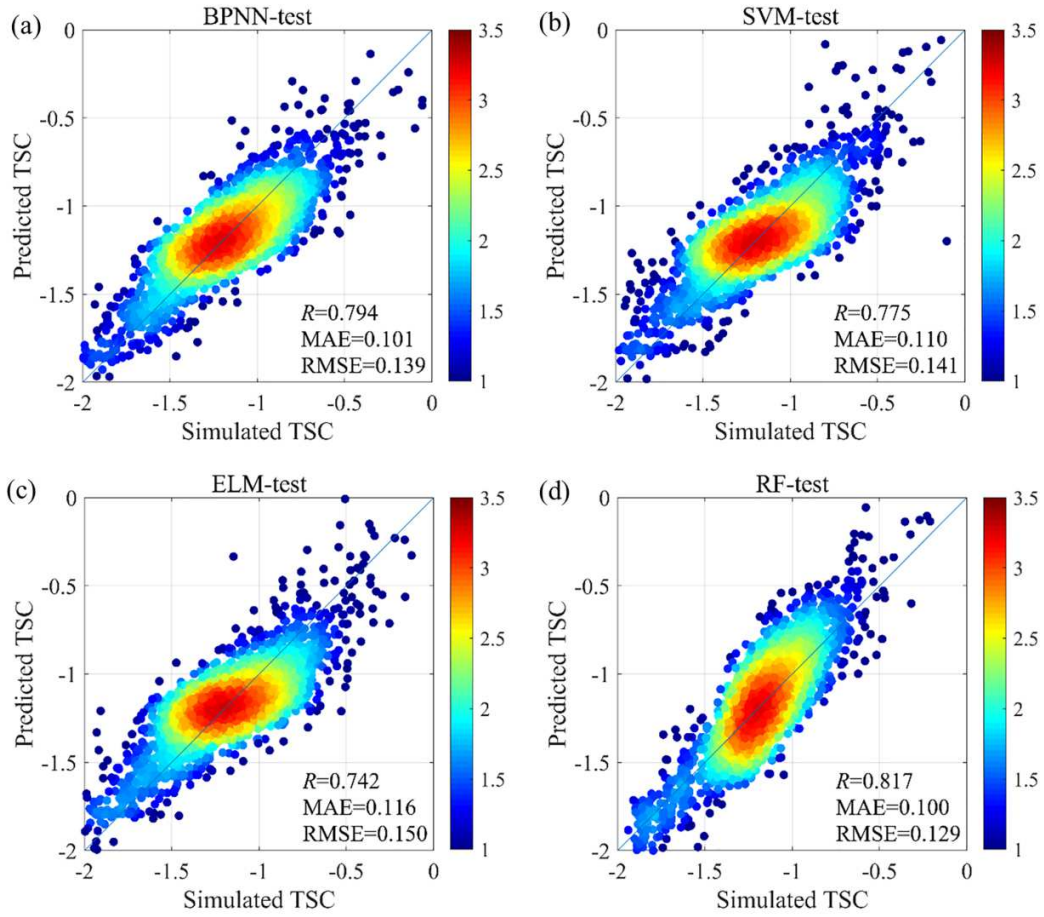


Fig. 2. Predicted TSC values using 4 ML algorithms: (a) BPNN, (b) SVM, (c) ELM, and (d) RF versus the simulated results from the GCMC for the AM dataset. The colors of points represent a base-10 logarithm of the number of MOFs.

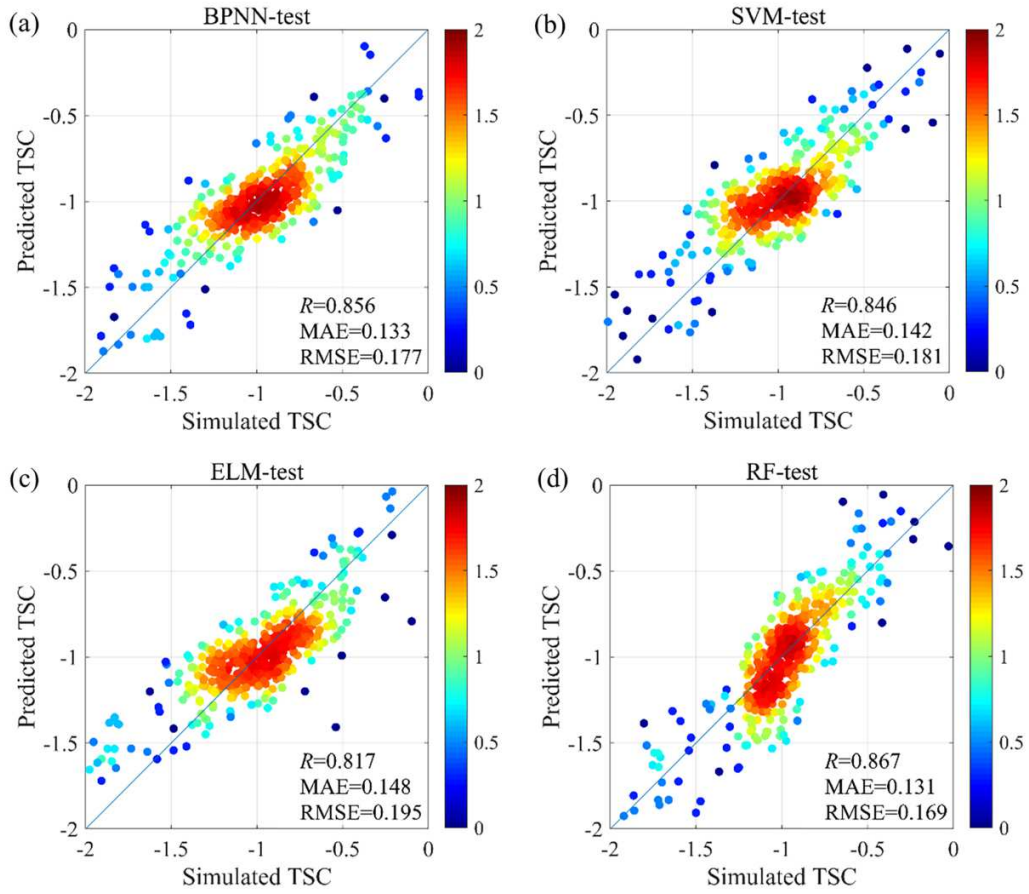


Fig. 3. Predicted values of the TSC using four ML methods: (a) BPNN, (b) SVM, (c) ELM, and (d) RF versus the simulated results from the GCMC simulations for the P_S dataset. The colors of the points correspond to a base-10 logarithm of the number of MOFs.

Based on Figs. S8–S14, which show the prediction effects for N_{HCHO} and $S_{HCHO/N2+O2}$ in each interval, as well as the predicted TSC in the P_N, we found that the *R*-values of the TSC were the greatest. Initially, since the prediction accuracy was greatly disturbed by the widely scattered number distributions of N_{HCHO} and $S_{HCHO/N2+O2}$, as well as the order of magnitude with large differences, we introduced a tradeoff method, TSC, which can successfully balance the N_{HCHO} and $S_{HCHO/N2+O2}$ to make the data more centralized ($-2.5 < TSC < 0.5$). Hence, the prediction of TSC by ML was a suitable method to explore the relationships among MOF descriptors and performance. The sequence of the predicted results was TSC > N_{HCHO} > $S_{HCHO/N2+O2}$, which further demonstrated the need to introduce the tradeoff variable TSC.

For the three datasets (AM, P_S, and P_N), AM contained all the MOFs, P_S contained those with optimal $S_{HCHO/N2+O2}$ values and the corresponding N_{HCHO} values, and P_N contained all the MOFs with the highest N_{HCHO} values and the corresponding $S_{HCHO/N2+O2}$ values. Taking the logarithm of the $S_{HCHO/N2+O2}$ twice in the TSC calculation formula weakened the impact of the $S_{HCHO/N2+O2}$ on the design results. Consequently, the high $S_{HCHO/N2+O2}$ values of the MOFs in the P_S group did not affect the TSC calculation significantly, and the influence of the N_{HCHO} on the TSC value was magnified. Benefiting from the influence of the high N_{HCHO} values, P_N corresponded to higher TSC values overall, but the better TSC values also bring the high dispersity to data that promotes the prediction effects of ML on the P_N group relatively poor. The TSC values of the AM and P_S were concentrated between −1.5 to −0.5, of which 19,385 MOFs were in the AM group (66.34% of 29,220), while 1284 were in the P_S group (87.89% of 1461). Thus, P_S was more concentrated than AM. The ML exhibited the best prediction effects of the TSC for the P_S group with high $S_{HCHO/N2+O2}$ values, and the ranking of the prediction effects of the datasets was P_S > AM > P_N.

Among the four ML approaches (BPNN, SVM, ELM, and RF) applied in three datasets, BPNN was a useful algorithm in the stable ML model for TSC prediction provided accurate prediction results; SVM was not the standout in the TSC prediction, but it can even obtain a better prediction effect than RF in the model of $S_{HCHO/N2+O2}$ prediction with a poor fitting degree; RF has played the stable prediction effect in different datasets since RF has robust generalization abilities. Shi *et al.*'s review [52] also obtained that the RF was generally appropriate for predicting the adsorption and separation performance of various gas mixture systems in MOFs. The highest predictive accuracy was obtained from the prediction of TSC in the P_S dataset, the order of algorithms was RF > BPNN > SVM > ELM.

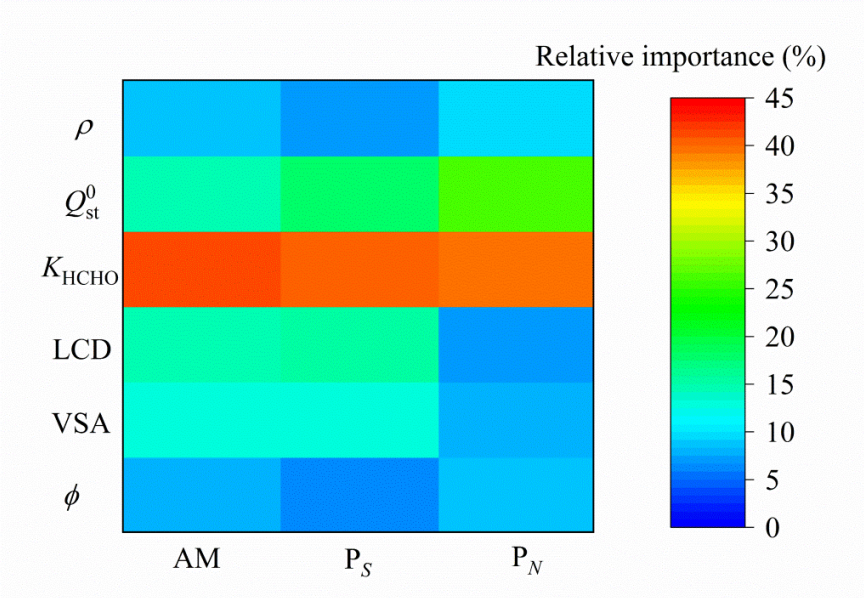


Fig. 4. Relative importance of six MOF descriptors by RF.

The relative importance of the six descriptors (ϕ, VSA, LCD, K_{HCHO}, Q^0_{st}, and ρ) on the TSC for the three data groups (AM, P_S, and P_N), which predicted by the RF method, are shown in Fig. 4 and Table S4. Whatever the data group was, the K_{HCHO} was undoubtedly the most crucial descriptor in each data set of MOFs (AM: 40.913%, P_S: 40.315%, and P_N: 39.519%). For P_S and P_N, Q^0_{st} possesses the second highest relative importance in Fig. 4; for AM dataset, Q^0_{st} (14.742%) still has the strong influence on the MOF performance, its relative importance was close to that of LCD (14.886%). Thus, Q^0_{st} is also useful in describing the performance of MOFs. This implies that the energetic descriptors of MOFs play a more important role than textural properties in governing formaldehyde capture. K_{HCHO} and Q^0_{st} appear to be the key descriptors due to the extremely low concentration of formaldehyde in the whole adsorption environment. This formaldehyde with very low partial pressures in the mixture of air is similar as the infinite dilution of formaldehyde. Considering the different types of adsorption isotherms, it is generally true that K_{HCHO} and Q^0_{st} could accurately represent the adsorption ability of adsorbents in low adsorption pressure, especially for infinite dilution conditions, as illustrated in some previous works of HTCS [44, 53]. Among the six descriptors of AM and P_S group, LCD and VSA besides K_{HCHO} and Q^0_{st} have larger relative importance compared with ϕ and ρ. However, for P_N, the formaldehyde adsorption was slightly influenced by LCD and VSA. This indicated that, compared to

the whole hMOF database, the LCD and VSA may not have a significant influence on the formaldehyde adsorption in some hMOFs with high N_{HCHO}. The RF method yielded the best prediction for P_S, revealing that K_{HCHO} and Q^0_{st} had the highest relative importance values of 40.315% and 18.321%, respectively. Consequently, K_{HCHO} and Q^0_{st} had the greatest influence on the production results and cannot be overlooked during the prediction of the TSC values. These parameters are useful for guiding the design of excellent MOFs with high formaldehyde capture performance.

3.3 Univariate Analysis

ML algorithms are powerful tools have already yielded predictions of the material properties and the relative importance of the descriptors. The univariate analysis as a means to verify the influence of structural and energetic descriptors of MOFs on the performance was used to explore the related details. Based on the results predicted by ML, the four most significant feature descriptors (VSA, LCD, K_{HCHO}, and Q^0_{st}) and three performance indicators (N_{HCHO}, $S_{HCHO/N2+O2}$, and TSC) were selected for further analysis. To explore the overall distributions and effects of the four key descriptors in the MOF intervals, their correlations to the AM data set, which contained the information about all the MOFs, are shown in Fig. 5.

Fig. 5a shows that in the MOFs with smaller VSA values (VSA < 500 m^2/cm^3), N_{HCHO} and $S_{HCHO/N2+O2}$ had high values, and the points with larger TSC values were concentrated in the upper left corner of Fig. 5a because the MOFs with small VSA values contained pore channels, which allowed formaldehyde molecules to contact the pore walls of the MOFs intimately. However, when the VSA was too large (VSA > 2000 m^2/cm^3), the adsorption space and the number of adsorption sites of N_2 and O_2 as the main parts of the mixed gas increased, while the relative contact area between the MOF pore walls and formaldehyde decreased with the increase in the VSA. This tended to limit the adsorption of formaldehyde, thereby decreasing the values of N_{HCHO} and $S_{HCHO/N2+O2}$. Fig. 5b shows that when the LCD < 3 Å, the steric hindrance between the formaldehyde molecules and hole walls limited the adsorption of the frameworks. When LCD > 5 Å, the interactions between the frameworks and formaldehyde molecules decreased, which intensified the desorption of formaldehyde in the pores

and reduced the adsorption efficiency. The TSC, N_{HCHO}, and $S_{HCHO/N2+O2}$ values of formaldehyde reached the maximum value only if the LCD was about 5 Å. This size was approximately equal to the kinetic diameter (4.5 Å) of formaldehyde. Similar results were also reported by Bian *et al.* [32].

Because the interactions between formaldehyde and the frameworks determine the adsorption ability of hMOFs, that were not conducive to adsorption when $\phi < 0.04$ or $\phi > 0.6$. The relationship between ϕ and performance indexes is shown in Fig. S15a. The appropriate contact yielded the best adsorption when ϕ was around 0.2. Also, Fig. S15b depicts the tendency of the performance indicators to decrease after rising to the peak values as ρ increases, in which the turning point is ρ = 2700 kg/m^3. Although a proper rise in ρ caused the MOFs to have a large number of adsorption sites, for excessively dense MOFs, the forces between the MOFs and formaldehyde molecules changed from attraction to repulsion, hindering the adsorption effects.

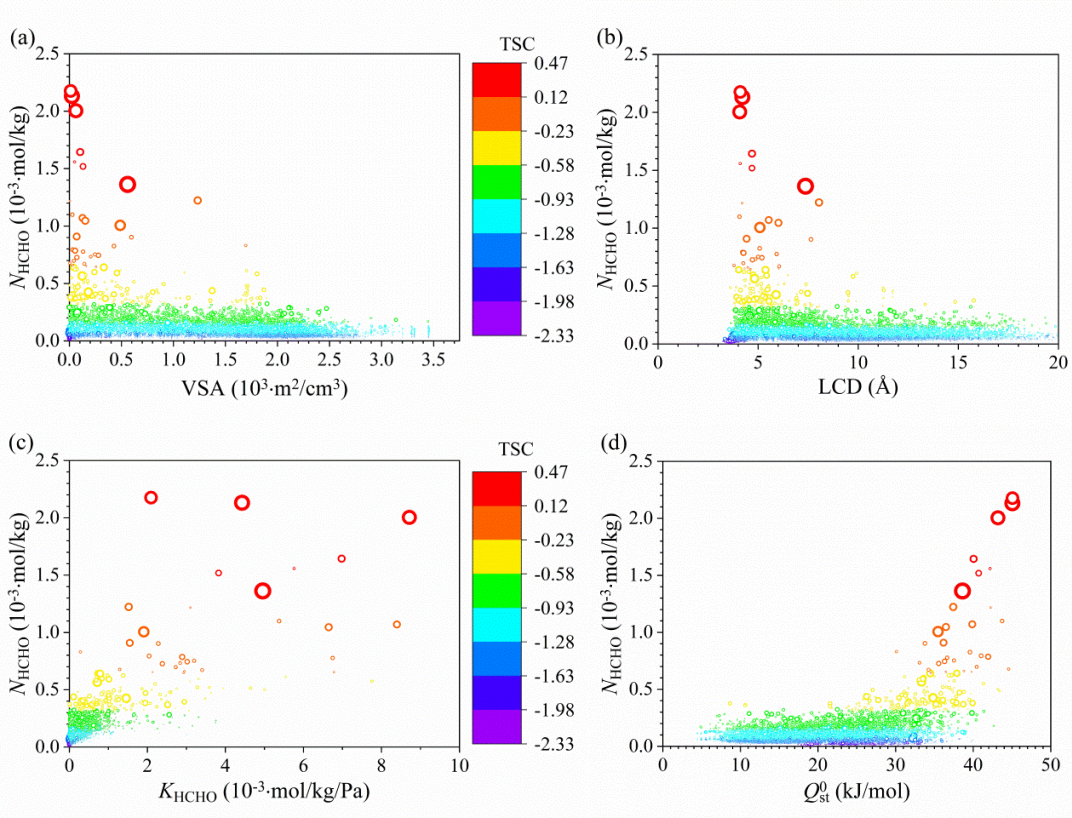


Fig. 5. Relationships among (a) VSA, (b) LCD, (c) K_{HCHO}, (d) Q^0_{st} and N_{HCHO}, $S_{HCHO/N2+O2}$, TSC. The bubble size represents the selectivity of the MOF.

Fig. 5c illustrates that when K_{HCHO} tended 0, the weak interactions between the formaldehyde molecules and MOF frameworks resulted in poor performance of MOFs. Compared with the distribution trends of MOFs for other descriptors (VSA, LCD, ϕ, ρ, and Q^0_{st}), the distribution extents of MOFs according to the K_{HCHO} were unprecedented concentrated. Fig. 5c shows that with the increase in K_{HCHO}, all of the N_{HCHO}, $S_{HCHO/N2+O2}$ and TSC always keep the upward trend. The robust positive correlations with the three performance indices enabled the alteration of the K_{HCHO} value to affect the quality of the MOF. Thus, K_{HCHO} can be used to indicate the properties of MOFs. This high centralization of MOFs and positive relationships between K_{HCHO} and performance indicators seemed to be the primary causes for the impact of K_{HCHO} on the TSC predictions, which agreed with the conclusion obtained by ML of the relative importance of K_{HCHO} to TSC. The increase of Q^0_{st} in the range of low values does not cause marked fluctuation of performance indexes, as shown in Fig. 5d, the performance indexes rise sharply with a high Q^0_{st} value, indicating a good correlation between TSC and Q^0_{st} that was consistent with the evaluation of ML. When Q^0_{st} was less than 4 kJ/mol, the adsorption reactions of formaldehyde on the MOFs were weak, because the interactions between the formaldehyde and MOFs were weak in the environment with a low Q^0_{st}, and there was not a sufficient driving force to promote the adsorption. The maximum TSC values occurred for Q^0_{st} > 37 kJ/mol, it is similar to the findings for low-concentration mercaptan captured in air [44].

3.4 Best MOF Adsorbents

The studies of ML and univariate analysis has already proved that K_{HCHO} and Q^0_{st} are the key factors in the formaldehyde adsorption. However, the qualitative analysis could not provide accurate quantitative standards. Therefore, we screened the materials twice based on their performance information. All the data of the original groups (AM, P_S, and P_N) were placed in the first layer. MOFs with TSC > −0.58 in each group were defined as promising MOFs, and it was used as the condition for a material to be screened for the second time. The descriptors of promising MOFs and those in the original MOF groups were classified and measured individually, as shown in Fig. 6. The purpose of this practice was to obtain excellent MOF candidates in the specified descriptor interval based on the

results of the quantitative analysis.

The number of promising MOFs of each original group was as follows: 206 (14.10% of 1461) were screened from P_N (Fig. 6), 144 (9.86% of 1461) were screened from P_S (Fig. S16), and 206 (0.70% of 29,220) were screened from AM (Fig. S17). Due to the particularity of TSC, dimensionality reduction by taking twice logarithms in succession weakened the role of $S_{HCHO/N2+O2}$ and magnified the importance of N_{HCHO}. Thus, the TSC values in P_N were higher as a whole, which was verified by the conclusion that promising MOFs had the highest probability (14.10%) of occurrence in P_N. Consequently, we assume that MOFs with high N_{HCHO} were more likely to be candidate materials for the efficient adsorption of formaldehyde.

Next focused on the promising and original MOFs in each descriptor interval. The descriptors of all promising MOFs belonging to three original groups had the same distributions overall. Most of the promising MOFs were found in small ranges of VSA and K_{HCHO}: 0 < VSA < 750 m^2/cm^3 and 0 < K_{HCHO} < 0.004 mol/kg/Pa. However, their ϕ (0.2–0.4), LCD (3–7 Å), Q^0_{st} (20–37 kJ/mol), and ρ (1200–2200 kg/m^3) values were mainly concentrated in the middle parameter ranges. Since the AM dataset including the whole information contained most of MOFs with bad performance, its distribution characteristics were very different from those of promising MOFs. In addition to the similar distributions of LCD and K_{HCHO}, in the AM group, ϕ (0.6–0.95), VSA (1500–2250 m^2/cm^3), Q^0_{st} (4–20 kJ/mol), and ρ (200–1200 kg/m^3) were mostly filled with inefficient MOFs.

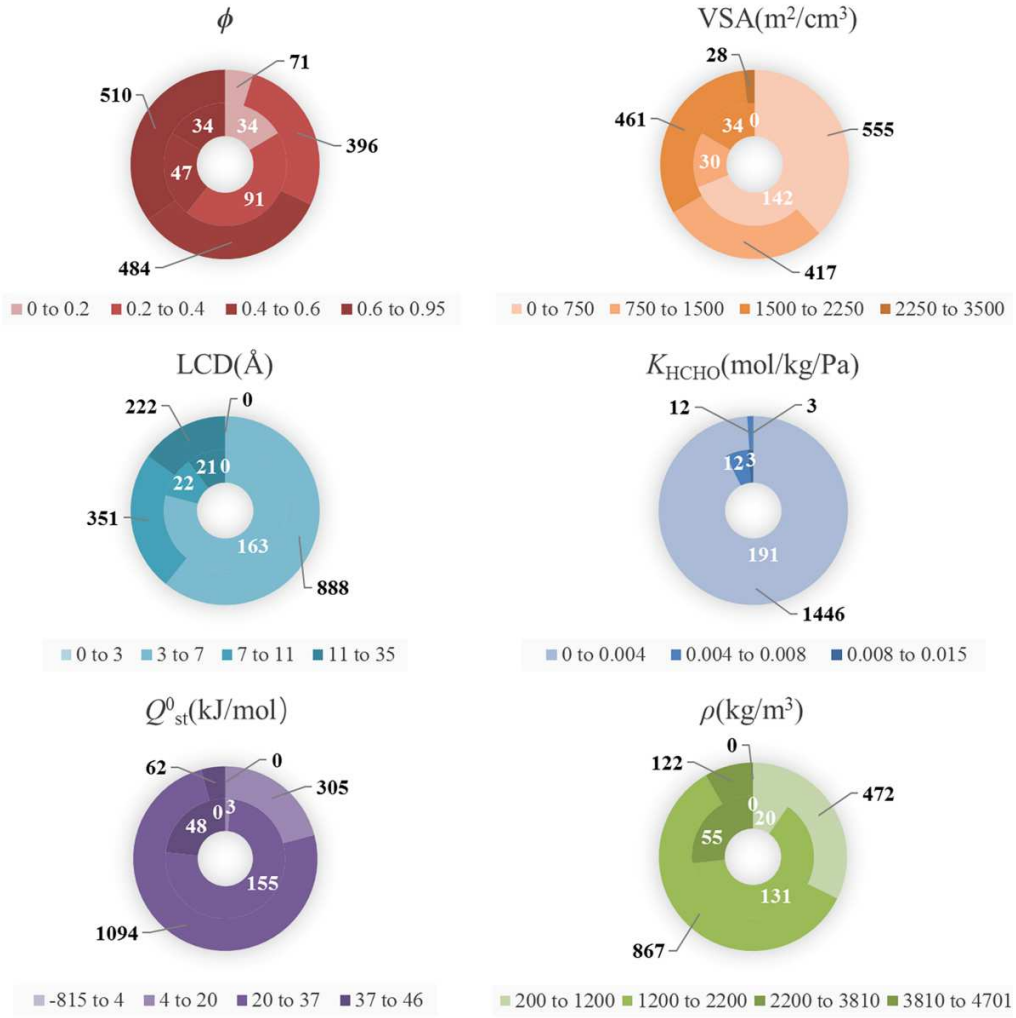


Fig. 6. Distributions of six descriptors on the HCHO/N_2+O_2 separation performance of MOFs. Numbers on the circles represent the number of MOFs. The outer circle represents all of the MOFs considered in the P_N (1461 MOFs), and the inner circle represents the 206 promising MOFs (TSC>-0.58) of P_N.

For P_N and P_S, the values of N_{HCHO} or $S_{HCHO/N2+O2}$ could help to screen out most of the undesirable MOFs, and their descriptors had the same highest frequency distribution intervals. Except that more than half of ϕ existed in 0.6–0.95, this range differed from that of the promising MOFs (0.2–0.4), the other five descriptors (VSA, LCD, K_{HCHO}, Q^0_{st}, and ρ) of groups P_N and P_S had the same distributions as those of the promising MOFs. Based on the univariate graph (Fig. S15a) of the ϕ and three performance indicators (N_{HCHO}, $S_{HCHO/N2+O2}$, and TSC), the ϕ of majority of MOFs with inferior performance were distributed from 0.6 to 0.95, while that of a small number of MOFs with superior

performance were near 0.2. For the concentration ranges of ϕ, it is the promising MOFs, rather than original MOFs, were closer to the distribution ranges of excellent MOFs. However, this did not mean that excellent MOFs were completely found in the descriptor intervals of the VSA (0–750 m^2/cm^3), LCD (3–7 Å), K_{HCHO} (0–0.004 mol/kg/Pa), Q^0_{st} (20–37 kJ/mol), and ρ (1200–2200 kg/m^3, which with the largest numbers of promising MOFs. In view of the classification information of the original MOFs (P_N and P_S) and the corresponding top-performing MOFs, we could distinguish that their five descriptors had the densest distributions in the specific ranges. However, excellent and bad MOFs coexisted in these regions, and the existence of bad MOFs interfered with the search for target materials, which confuse the probabilities of obtaining excellent MOFs using these intervals. Therefore, we decided to use the number ratio of promising MOFs to original MOFs in the corresponding range of descriptors as the probability of finding excellent MOFs.

Table S5 lists the probabilities of finding excellent MOFs in each descriptor path from the AM, P_S, and P_N databases, this approach was to find the optimal paths to obtain excellent MOFs with highest probabilities. The optimal paths of the five descriptors (VSA, LCD, K_{HCHO}, Q^0_{st}, and ρ) of the AM, P_S, and P_N datasets were consistent, the intervals for obtaining ideal MOFs were as follows: VSA was 0–750 m^2/cm^3, LCD was 3–7 Å, K_{HCHO} was 0.004–0.015 mol/kg/Pa, Q^0_{st} was 37–46 kJ/mol, and ρ was 2200–3810 kg/m^3. For the P_S and P_N subgroups, the maximum probability extent of ϕ was 0 to 0.2. The information for the excellent MOFs could be locked more quickly and conveniently through probability distributions. Compared with those of the AM group, the distributions of P_S and P_N were more accurate, since the numbers of MOFs with low performance were much lower in these groups.

Based on the results of Table S5, Fig. 7 compares the possibilities of three databases acquiring excellent MOFs under their optimal paths. K_{HCHO} and Q^0_{st} were the descriptors with the highest probabilities for any database, which agreed with the conclusion of the relative importance from the ML prediction. Although there were only 15 MOFs with K_{HCHO} values in the 0.004–0.015 mol/kg/Pa range, the probabilities for the three databases in this optimal interval were 100%, which indicates that a MOF can become a potential candidate if its K_{HCHO} breaks a definite threshold value. Based on the optimal interval (37–46 kJ/mol) of Q^0_{st}, Fig. 7 shows that the probabilities of the three databases were

between 55%–78%, indicating that it is also very likely to search for top MOFs by using Q^0_{st}. The next most important parameters were ϕ and ρ, which yielded moderate probabilities of finding excellent MOFs. If these parameters were combined with each other or Q^0_{st} to jointly screen MOFs, the possibilities of locating target materials would be greatly promoted. The descriptors with the lowest probabilities of obtaining excellent MOFs were the LCD and VSA. Therefore, it is relatively difficult to judge the adsorption of formaldehyde by MOFs only through the LCD and VSA.

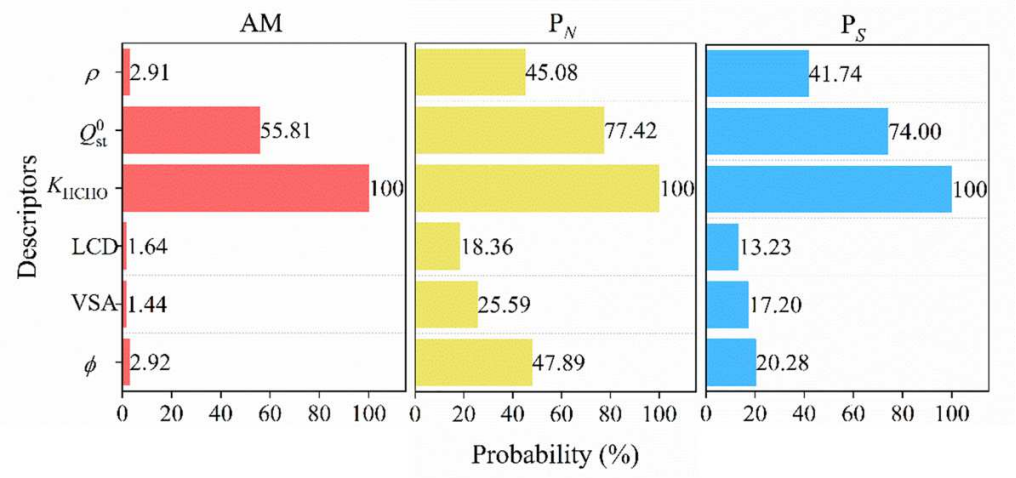


Fig. 7. Probabilities of excellent MOFs using six descriptors for three datasets.

The probabilities of screening for top MOFs from the P_N group was higher than the other datasets using any descriptor confirmed the prior conclusions that N_{HCHO} had the greatest impact on the performance of hMOFs. Therefore, the 20 best MOFs based on the N_{HCHO} values in the P_N dataset were sorted and listed in Table 2. At least one of the descriptors of most of these MOFs fell within the optimal interval presented above. The exception was the sixteenth MOF (ID 2000792), which had the largest ϕ, VSA, and LCD values and the smallest K_{HCHO}, Q^0_{st}, and ρ values. The first MOF (ID 37557) with the largest Q^0_{st} of 20 MOFs had the highest formaldehyde adsorption capacity of 2.18×10^{-3} mol/kg, was the best potential candidate for adsorption for an extremely low initial formaldehyde concentration (13.41 mg/m^3), whose descriptors were in or close to the optimal ranges. This low initial concentration was comparable to a real environment with formaldehyde in the air, and thus, the results provide a reliable reference and guidance for the selection of MOFs to use in such an environment.

Table 2. Best MOFs.

No.	ID[a]	TSC	N_{HCHO} (10^{-3} mol/kg)	$S_{HCHO/N2+O2}$ (10^3)	ϕ	VSA (m^2/cm^3)	LCD (Å)	K_{HCHO} (10^{-3} mol/kg/Pa)	Q^0_{st} (kJ/mol)	ρ (10^3 kg/m^3)
1	37557	0.46	2.18	9.18	0.13	12.21	4.11	2.09	45.06	2.61
2	27102	0.46	2.13	10.88	0.16	24.19	4.21	4.43	45.06	2.69
3	35481	0.42	2.00	10.15	0.19	61.62	4.08	8.72	43.19	2.93
4	22616	0.28	1.64	5.49	0.25	103.4	4.69	6.98	40.06	2.49
5	5066904	0.24	1.52	4.57	0.19	130.55	4.69	3.82	40.71	3.55
6	27094	0.23	1.56	2.12	0.15	49.63	4.10	5.76	42.17	1.95
7	1004322	0.19	1.36	11.56	0.36	559.74	7.36	4.96	38.62	2.73
8	7000407	0.12	1.22	5.63	0.52	1233.09	8.04	1.52	37.43	1.83
9	6003154	0.10	1.22	1.57	0.14	4.75	4.20	3.10	42.24	1.57
10	37554	0.05	1.10	3.01	0.15	28.11	4.07	5.38	43.73	2.03
11	7000801	0.04	1.07	5.23	0.14	126.44	5.53	8.40	39.89	2.40
12	7001316	0.03	1.05	5.34	0.16	152.65	6.01	6.65	36.49	1.81
13	1003754	0.00	1.01	7.69	0.43	487.82	5.08	1.91	35.47	2.39
14	1003126	-0.05	0.91	5.26	0.24	70.09	4.43	1.55	36.18	1.68
15	1004229	-0.06	0.90	3.17	0.36	596.59	7.64	2.28	33.78	1.75
16	2000792	-0.10	0.83	2.08	0.87	1695.69	22.41	0.28	30.10	0.40
17	1004231	-0.10	0.83	3.13	0.34	427.68	5.08	14.77	39.58	3.41
18	16879	-0.13	0.79	3.22	0.21	32.07	4.22	2.04	41.03	2.48
19	5066448	-0.14	0.79	4.19	0.14	54.87	4.27	2.89	41.93	3.13
20	7000855	-0.14	0.78	2.97	0.15	131.35	5.93	6.75	36.65	1.80

[a]IDs for hypothetical MOFs [40].

4. Conclusions

In this work, 31,399 hMOFs with good hydrophobicity were acquired through the screening of Henry's coefficient of water, and their adsorption abilities to adsorb formaldehyde from ternary air mixtures were screened by the high-throughput simulations. Four ML methods (BPNN, SVM, ELM, and RF) were employed to calculate the three performance indicators (N_{HCHO}, $S_{HCHO/N2+O2}$, and TSC) of the three data sets (AM, P_S, and P_N), then we found that the RF yielded the best predictions for the TSC in the P_S dataset, because of the highest *R*, the smallest MAE and RMSE. The relative importance values of K_{HCHO} and Q^0_{st}, which were the most influential, were determined to be 40.315% and 18.321%, respectively. For the six descriptors (ϕ, VSA, LCD, K_{HCHO}, Q^0_{st}, and ρ), we obtained the

probabilities of finding advanced hMOFs under the specific paths of them by calculating the ratios of promising hMOFs in three datasets. In the P_N dataset, the probabilities of obtaining excellent hMOFs in the optimal ranges of K_{HCHO} (0.004–0.015 mol/kg/Pa) and Q^0_{st} (37–46 kJ/mol) were 100% and 77.42%, respectively, further confirmed K_{HCHO} and Q^0_{st} are key factors governing the formaldehyde adsorption at low concentrations. Comparing the probability distributions of the promising hMOFs obtained for the three data sets, we found that excellent MOFs fell preferentially in the P_N group with remarkable N_{HCHO} values. Finally, 20 hMOFs with the best qualities to capture formaldehyde from the atmosphere were identified. This work revealed the structure–performance relationships of hMOFs, the theoretical guidance and evidence are insightful for the development of new MOFs for the capture of formaldehyde from the air.

Acknowledgements

We gratefully thank the National Natural Science Foundation of China (Nos. 21978058 and 21676094) and the Natural Science Foundation of Guangdong Province (2020A1515010800) for financial support.

Conflict of interests

The authors declare no conflict of interests.

References

[1] V.K. Saini, J. Pires, J. Environ. Sci. 55 (2017) 321-330.

[2] L. Nie, J. Yu, M. Jaroniec, F.F. Tao, Catal. Sci. Technol. 6 (2016) 3649-3669.

[3] C.-J. Na, M.-J. Yoo, D.C. Tsang, H.W. Kim, K.-H. Kim, J. Hazard. Mater. 366 (2019) 452-465.

[4] E.Y. Nakanishi, M.R. Cabral, P. De Souza Gonçalves, V. Dos Santos, H.S. Junior, Journal of Cleaner Production 195 (2018) 1259-1269.

[5] K. Vikrant, M. Cho, A. Khan, K.-H. Kim, W.-S. Ahn, E.E. Kwon, Environ. Res. 178 (2019) 108672.

[6] K.-W. Zhou, Y. Zhou, Y. Sun, X.-J. Tian, Acta Chim. Sinica 66 (2008) 943-946.

[7] H. Dou, D. Long, X. Rao, Y. Zhang, Y. Qin, F. Pan, K. Wu, ACS Sustain. Chem. Eng. 7 (2019) 4456-4465.

[8] S.-C. Hu, Y.-C. Chen, X.-Z. Lin, A. Shiue, P.-H. Huang, Y.-C. Chen, S.-M. Chang, C.-H. Tseng, B. Zhou, Environ. Sci. Pollut. R. 25 (2018) 28525-28545.

[9] K.J. Lee, J. Miyawaki, N. Shiratori, S.-H. Yoon, J. Jang, J. Hazard. Mater. 260 (2013) 82-88.

[10] C. Montoro, F. Linares, E. Quartapelle Procopio, I. Senkovska, S. Kaskel, S. Galli, N. Masciocchi, E. Barea, J.A. Navarro, J. Am. Chem. Soc. 133 (2011) 11888-11891.

[11] O.M. Yaghi, M. O'keeffe, N.W. Ockwig, H.K. Chae, M. Eddaoudi, J. Kim, Nature 423 (2003) 705-714.

[12] H. Li, M. Eddaoudi, M. O'keeffe, O.M. Yaghi, Nature 402 (1999) 276-279.

[13] P.Z. Moghadam, S.M. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D.A. Gomez-Gualdron, V. Van Speybroeck, Matter 1 (2019) 219-234.

[14] F. Rezaei, S. Lawson, H. Hosseini, H. Thakkar, A. Hajari, S. Monjezi, A.A. Rownaghi, Chem. Eng. J. 313 (2017) 1346-1353.

[15] Y. He, F. Chen, B. Li, G. Qian, W. Zhou, B. Chen, Coord. Chem. Rev. 373 (2018) 167-198.

[16] B. Wang, L.-H. Xie, X. Wang, X.-M. Liu, J. Li, J.-R. Li, Green Energy Environ. 3 (2018) 191-228.

[17] G. Xu, Z. Meng, Y. Liu, X. Guo, K. Deng, R. Lu, Int. J. Hydrogen Energy 44 (2019) 6702-6708.

[18] X. Zhao, Y. Wang, D.S. Li, X. Bu, P. Feng, Adv. Mater. 30 (2018) 1705189.

[19] M. Kang, D.W. Kang, C.S. Hong, Dalton Trans. 48 (2019) 2263-2270.

[20] K. Vikrant, Y. Qu, J.E. Szulejko, V. Kumar, K. Vellingiri, D. Bukhvalov, T.J. Kim, K.-H. Kim, Nanoscale 12 (2020) 8330-8343.

[21] Q. Wang, D. Astruc, Chem. Rev. 120 (2020) 1438-1511.

[22] P. Leo, G. Orcajo, D. Briones, F. Martínez, G. Calleja, Catal. Today 345 (2020) 251-257.

[23] W.P. Lustig, S. Mukherjee, N.D. Rudd, A.V. Desai, J. Li, S.K. Ghosh, Chem. Soc. Rev. 46 (2017) 3242-3285.

[24] B. Yan, J. Mater. Chem. C 7 (2019) 8155-8175.

[25] G. Chedid, A. Yassin, Nanomaterials 8 (2018) 916.

[26] H. Li, L. Li, R.-B. Lin, W. Zhou, S. Xiang, B. Chen, Z. Zhang, EnergyChem 1 (2019) 100006.

[27] P.G. Boyd, A. Chidambaram, E. García-Díez, C.P. Ireland, T.D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S.M. Moosavi, M.M. Maroto-Valer, Nature 576 (2019) 253-256.

[28] A.J.M. Reddy, N. Katari, P. Nagaraju, S.M. Surya, Mater. Chem. Phys. 241 (2020) 122357.

[29] L. Wang, X.-Y. Liang, Z.-Y. Chang, L.-S. Ding, S. Zhang, B.-J. Li, ACS Appl. Mater. Interfaces 10 (2018) 42-46.

[30] N.S. Bobbitt, M.L. Mendonca, A.J. Howarth, T. Islamoglu, J.T. Hupp, O.K. Farha, R.Q. Snurr, Chem. Soc. Rev. 46 (2017) 3357-3385.

[31] Y.J. Colón, R.Q. Snurr, Chem. Soc. Rev. 43 (2014) 5735-5749.

[32] L. Bian, W. Li, Z. Wei, X. Liu, S. Li, Acta Chim. Sinica 76 (2018) 303-310.

[33] L. Peng, Q. Zhu, P. Wu, X. Wu, W. Cai, PCCP 21 (2019) 8508-8516.

[34] D.V. Gonçalves, R.Q. Snurr, S.M. Lucena, Adsorption 25 (2019) 1633-1642.

[35] R. Anderson, J. Rodgers, E. Argueta, A. Biong, D.A. GóMez-GualdróN, Chem. Mater. 30 (2018) 6325-6337.

[36] N.S. Bobbitt, R.Q. Snurr, Ind. Eng. Chem. Res. 56 (2017) 14324-14336.

[37] X. Sun, X. Gu, W. Xu, W.-J. Chen, Q. Xia, X. Pan, X. Zhao, Y. Li, Q.-H. Wu, Front. Chem. 7 (2019) 652.

[38] L.N. Mchugh, A. Terracina, P.S. Wheatley, G. Buscarino, M.W. Smith, R.E. Morris, Angew. Chem. Int. Ed. 58 (2019) 11747-11751.

[39] Z.-Y. Gu, G. Wang, X.-P. Yan, Anal. Chem. 82 (2010) 1365-1370.

[40] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Nat. Chem. 4 (2012) 83.

[41] D. Dubbeldam, S. Calero, D.E. Ellis, R.Q. Snurr, Mol. Simul. 42 (2016) 81-101.

[42] T.F. Willems, C.H. Rycroft, M. Kazi, J.C. Meza, M. Haranczyk, Microporous Mesoporous Mater. 149 (2012) 134-141.

[43] P.Z. Moghadam, D. Fairen-Jimenez, R.Q. Snurr, J. Mater. Chem. A 4 (2016) 529-536.

[44] Z. Qiao, Q. Xu, A.K. Cheetham, J. Jiang, J. Phys. Chem. C 121 (2017) 22208-22215.

[45] A.K. Rappé, C.J. Casewit, K. Colwell, W.A. Goddard Iii, W.M. Skiff, J. Am. Chem. Soc. 114 (1992) 10024-10035.

[46] E.S. Kadantsev, P.G. Boyd, T.D. Daff, T.K. Woo, J. Phys. Chem. Lett. 4 (2013) 3056-3061.

[47] G. Hantal, P. Jedlovszky, P.N. Hoang, S. Picaud, J. Phys. Chem. C 111 (2007) 14170-14178.

[48] M.G. Martin, J.I. Siepmann, J. Phys. Chem. B 102 (1998) 2569-2577.

[49] P.P. Ewald, Ann. Phys. 369 (1921) 253-287.

[50] M.S. Shah, M. Tsapatsis, J.I. Siepmann, Angew. Chem. Int. Ed. 55 (2016) 5938-5942.

[51] W. Yang, H. Liang, Z. Qiao, Acta Chim. Sinica 76 (2018) 785-792.

[52] Z. Shi, W. Yang, X. Deng, C. Cai, Y. Yan, H. Liang, Z. Liu, Z. Qiao, Mol. Syst. Des. Eng. 5 (2020)

725-742.

[53] T. Watanabe, D.S. Sholl, Langmuir 28 (2012) 14114-14128.

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: