https://doi.org/10.1038/s42004-024-01166-7

Machine learning insights into predicting biogas separation in metal-organic frameworks

Check for updates

Isabel Cooley, Samuel Boobier, Jonathan D. Hirst & Elena Besley

Breakthroughs in efficient use of biogas fuel depend on successful separation of carbon dioxide/methane streams and identification of appropriate separation materials. In this work, machine learning models are trained to predict biogas separation properties of metal-organic frameworks (MOFs). Training data are obtained using grand canonical Monte Carlo simulations of experimental MOFs which have been carefully curated to ensure data quality and structural viability. The models show excellent performance in predicting gas uptake and classifying MOFs according to the trade-off between gas uptake and selectivity, with R^2 values consistently above 0.9 for the validation set. We make prospective predictions on an independent external set of hypothetical MOFs, and examine these predictions in comparison to the results of grand canonical Monte Carlo calculations. The best-performing trained models correctly filter out over 90% of low-performing unseen MOFs, illustrating their applicability to other MOF datasets.

Effective purification of the biogas stream obtained from decomposition of agricultural and industrial waste remains a challenging but promising goal towards a renewable source of biomethane fuel[1] and a more sustainable alternative to fossil fuels[2]. Biogas is composed predominantly of a CH_4/CO_2 mixture which must be separated, along with trace contaminants, to obtain biomethane of increased purity for use in internal combustion engines[2]. Multiple established approaches[2–4] are routinely used for upgrading the calorific content of biogas by removal of CO_2, although these technologies can be costly and energy intensive. Among them, adsorptive[5] and membrane[2] separations by porous materials remain attractive options, subject to improvement in yield, efficiency, and sustainability[2–4]. Optimal porous materials provide a clear route towards improved performance; however, such a material must be selective of CO_2 over CH_4 while also exhibiting CO_2 uptake sufficiently high for practical use. Many gas separations are characterised by a trade-off between selectivity and uptake[6,7], rendering search for high-performing materials challenging. Meanwhile, certain separation processes present further complexity. For example, membrane separations additionally rely on effective diffusivity[2].

Among prominent candidate materials for gas separation are porous metal-organic frameworks (MOFs)[8–11]. High surface area complexes of metal-containing nodes and organic linkers, MOFs have shown excellent performance for a range of chemical processes[12–16] including several gas separations[15]. Structural variety of MOFs occupies a vast and diverse chemical space of reported synthesised MOFs[17] and many more proposed hypothetical structures[18,19], making prediction and tuning of all relevant separation properties inaccessible to experiment. High-throughput use of computational force-field methods provides reasonable uptake and selectivity predictions and fundamental insight into structure-property relationships. Screenings of thousands of MOFs for separation of gas mixtures are readily available in the literature[20,21], including a recent search of nearly 7000 MOFs for biogas upgrading properties with a focus on membrane separation[22].

Computational MOFs screenings may be radically changed by development of machine learning (ML) models suitable for predicting gas sorption properties using only features that are cheap to calculate. This can expand the size of databases that can be screened and further reduce computational cost. Similar application of ML has become prominent in every field of materials and chemistry research, with recent advances in the prediction of material properties[23], solubility[24], protein structure[25], and reaction pathway[26]. ML is increasingly being applied to investigate and optimise MOFs for CO_2 adsorption[27,28] and gas separation[29–37], including for CO_2/CH_4 separation[38–40]. There are several recent reviews of this growing area[41].

Some studies have considered the relative performance of different ML methods reporting higher predictive accuracy for non-linear over linear methods[30]. When it comes to feature selection, structural descriptors, which

School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
e-mail: Elena.Besley@nottingham.ac.uk

can be readily and cheaply calculated, are favoured, as in the CO_2/CH_4 study of ref. 38 Linear regression models developed by ref. 32 to predict CH_4 uptake and working capacity appear to perform well using only three features, all of them structural. However, features which capture chemical information such as Henry constants[42], binding energy or the Voronoi energy introduced by ref. 29 can improve the quality of ML models.

Previous ML studies of MOFs tend to be trained on large databases composed primarily of hypothetical MOFs without significant curation[30,38]. While this is a useful way to employ large amounts of data, it risks using unviable MOF structures which display issues relating to structural determination procedures[43,44] or, in the case of hypothetical MOFs, synthesisability[45]. The importance of data curation is becoming increasingly recognised in the material and chemical domains, with studies showing that well-curated datasets give more accurate and insightful models[46–49]. In this light, curation procedures are gaining traction within high-throughput MOF workflows[43,44,50–52]. Understanding of the capabilities of ML models trained using experimental MOF structures curated for viability is therefore essential.

In this work, we develop a ML model to predict performance of MOFs for biogas upgrading using a well-curated and high quality dataset. One should not underestimate the importance of care in preparation of the dataset. A small number of carefully selected features cover key structural and chemical information. We focus on established (rather than newly invented) descriptors that can provide chemical insight, even if they are, in some cases, relatively expensive to compute. The resultant models are highly accurate and subsequent analysis of the models provides insight into the features of high-performing MOFs for biogas separation.

Results

Preparation of a high-quality dataset

A high-quality dataset is an essential prerequisite for ML-assisted high-throughput screening. Here, we use an experimental MOF subset which was obtained from the Cambridge Structural Database (CSD)[53] as part of a previous high-throughput search for MOFs with biogas upgrading properties by Glover and Besley[22] in which the dataset was algorithmically stripped of solvents and filtered according to geometric criteria. Visualisation of the initial dataset comprising 6768 stripped MOF structures revealed significant issues necessitating further curation (see Supplementary Note 1). Limitations of datasets which may negatively affect high-throughput and machine learning MOF studies are gaining attention. These include persistence of unfeasible structures[43,44] as well as persistence of duplicate structures which may affect diversity and lead to data leakage in machine learning[54,55]. Curation procedures improve the quality of MOF datasets, and are beginning to be applied in contemporary studies[43,44,50–52].

A strict curation workflow, as detailed in the Methods Section 4.1 and shown in Fig. S1 of Supplementary Note 1, was applied to remove unfeasible structures as well as duplicates of the same structure, after which only 1910 MOFs remained in the dataset. The distribution of metal centres present in this dataset is illustrated in Fig. S2 of Supplementary Note 1. The most abundant metal in the dataset is Zn, followed by Cu and Co. The large number of removed MOF structures highlights the critical prevalence of unfeasible structures in MOF databases, which should not be ignored. During the curation stage, overlapping and missing atoms were encountered in nearly 10% of MOFs structures, and the proportion of unrealistic structures increased dramatically when the oxidation state of metal centres was considered (at this stage, more than half of the remaining structures were flagged for removal). The CSD MOF dataset is well-known and commonly utilised, and it is far from the only database which suffers from the issues encountered.

For the remaining MOFs, Grand Canonical Monte Carlo (GCMC) simulations were used to calculate key metrics of biogas separation performance under working conditions of 10 bar pressure and 298 K temperature. Details of the GCMC setup are given in the Methods Section 4.2; the accuracy and reliability of the setup has been validated previously[56] by comparing its performance for CH_4 and CO_2 uptake in the MFM family of copper paddlewheel-based MOFs with existing experimental data[57]. Gas uptake simulations were separately performed to obtain the loading values for both single component (SC) gases and a binary mixture (BM) of 50/50 CO_2/CH_4 gas. Selectivity, *S*, of CO_2 over CH_4 was calculated from the binary mixture CO_2 and CH_4 uptake values, namely, from the loading of CO_2 and the loading of CH_4 when a MOF system was simulated in equilibrium with a 50/50 binary mixture of CO_2 and CH_4 gases (see Methods section 4.2). Simulation conditions were selected for relevance to the separation within the landscape of industrial conditions which vary depending on specifics of process and materials and can be optimised for a given setup. In particular, the common pressure swing adsorption (PSA) technique tends to require adsorption at high pressures at or above 10 bar[58,59] and desorption at pressures of 1 bar or below[38,58,60]. Temperature swing adsorption (TSA) can be carried out at ambient pressures[61] with adsorption temperatures between 273 K and room temperature and desorption temperatures elevated by a margin on the order of 100 K[62]. Membrane separation, meanwhile, uses pressures of a few bar with no demand for elevated temperature[2]. With regard to the selected gas composition, biogas feedstocks vary, with CH_4 composing 50−65% and CO_2 composing 35-50% of a mixture (and trace gases also present)[2,22,63]. The 50/50 mixture was selected as an example of a realistic biogas composition comparable to previous work[22].

The ranges of the uptake metrics in the curated dataset are illustrated in Fig. 1c, d, g, h which show that single component CH_4 uptake reaches nearly 12 mol kg^{-1} (Fig. 1g) whilst single component uptake of CO_2 is notably higher, approaching 25 mol kg^{-1} (Fig. 1h). Binary mixture uptake (Fig. 1c, d) is lower, particularly for the weaker adsorbent CH_4, although binary mixture CO_2 loading above 17 mol kg^{-1} is observed, which compares well to previously reported values[64].

The range of values which selectivity takes in the curated dataset may be seen in Fig. 1b, f, in which a binary mixture CO_2 loading is plotted against CO_2 selectivity. Very high values of selectivity above 10^4 are observed within the curated dataset, although high selectivity can be a consequence of very low CH_4 loading in MOFs whose CO_2 loading is not usefully high. Indeed, Fig. 1b, f illustrate the existence of a trade-off relationship between uptake and selectivity. None of the MOFs which display exceptionally high selectivity also have very high loading. An ideal MOF would possess both, and be found in the upper right portion of the plots. Instead, there is a significant population of MOFs in the lower left, branching into the lower right (high loading) and upper left (high selectivity). High uptake and selectivity of the dominant gas are both of importance to gas separations; it is desirable to identify structures displaying both. In the relative absence of MOFs which unite both metrics, it is instructive to identify those presenting a useful compromise between the two, as quantified here by a metric known as trade off between selectivity and uptake, TSN_{CO_2/CH_4}[6,7] (see Methods Section 4.2), which is herein referred to as TSN. The range of values which TSN takes within the curated dataset is illustrated in Fig. 1i.

Table 1 gives the names of the six MOFs whose calculated TSN is greater than 22 mol kg^{-1}, along with values of TSN, selectivity and binary mixture CO_2 loading, and ranks of these values within the dataset. It also includes the pore limiting diameter (PLD) and void fraction (VF) of each structure in order to illustrate relationships between structure and uptake properties, as well as the topology of each structure. Structural values were taken from the data of ref. 22 and topology was determined using the CrystalNets.jl[65] web application (https://progs.coudert.name/topology) with the SingleNodes clustering option. To further facilitate detailed structural examination, visualisations of each of the six MOFs are presented in Fig. 2. The six highest-TSN MOFs all rank highly for selectivity, and most also rank highly for binary mixture CO_2 uptake. Interestingly, many of them possess common structural features. The metal centres of the four MOFs with highest TSN are Zn. Meanwhile, the sql square lattice topology features significantly and several of the six presented MOFs possess approximately square channels. The structures of the five MOFs with highest TSN are two dimensional (2D) except for the MOF with reference code UQUVOS which has a three dimensional (3D) structure made up of connected 2D sheets. Further, the six MOFs have small pore limiting diameter, including two

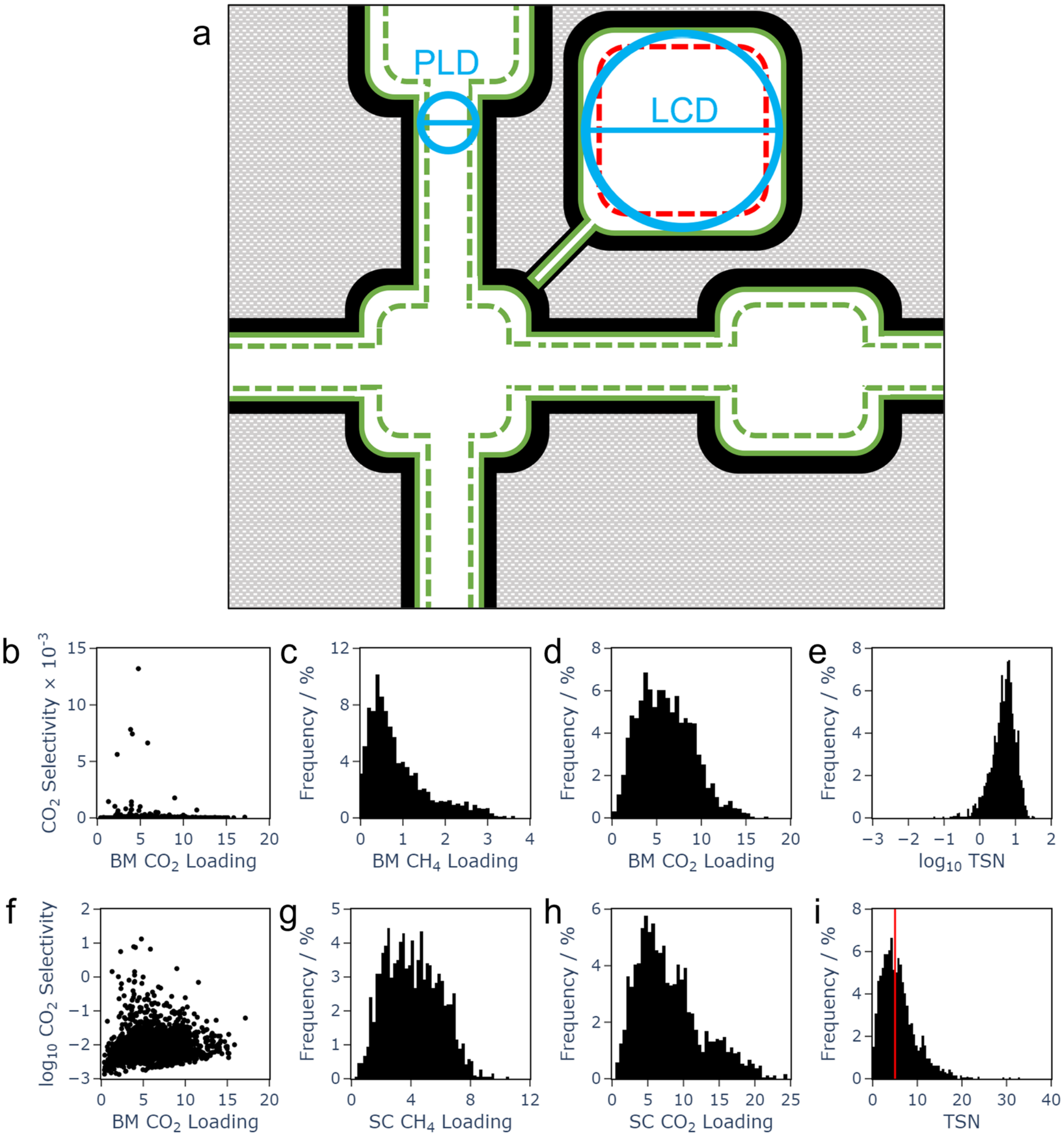


Fig. 1 | Structural features of MOFs used in the ML model and the range of values for uptake metrics, selectivity and TSN used in the curated data set. a An illustration of structural features; blue circles show pore limiting diameter (PLD) and largest cavity diameter (LCD), which describe pore size; green dashed lines show accessible area (ASA) as measured using the centre of a probe (1.86 Å radius), and green solid lines enclose internal volume as measured using a point probe, which is often expressed as a void fraction. Red dashed lines indicate non-accessible area. Plots of binary mixture CO_2 loading against CO_2 selectivity, with a (**b**) linear and (**f**) logarithmic scale; (**c**−**e**) and (**g**−**i**) ranges of the six prediction targets studied in this work. Units for loading and TSN are mol kg^{-1}; Selectivity is unitless. For the trade off between selectivity and uptake metric, TSN, the indicated cutoff of 5 mol kg^{-1} was used to classify low and high TSN MOFs.

among the four highest-TSN MOFs with PLD very close to 3.80 Å, the kinetic diameter of methane and the smallest PLD allowed in the dataset. VF values range from 0.5 to 0.7.

The abundance of Zn centres among the six top-performing MOFs is likely to be at least in part due to the high abundance of Zn-MOFs in the dataset as a whole, in which Zn is the most commonly occurring metal (see Fig. S2). Among 1910 curated MOFs, 542 possess a Zn centre, which amounts to 28% of the dataset. While a significant amount, this is lower than the prevalence of Zn among the top 6 MOFs (67%). It is useful to make similar comparisons to the base dataset for the other structural observations made, to indicate the extent to which they are simply a result of trends in the base data, and to what extent they are specific to the high performers. In terms of dimensionality, 597 of the 1910 MOFs in the base dataset (31%) are 2D, compared to five of the six top performers (83%). For PLD, two of the top performing MOFs have PLD under 3.82 Å, while only 13 of the 1910 MOFs (<1%) have PLD under the same threshold (when rounded). Meanwhile, all of the top performers have PLD under 4.84 Å, compared to 914 (48%) of the curated dataset. For VF, 1230 of the 1910 curated MOFs (64%) fall within the 0.5−0.7 range covered by the top MOFs.

We note that it is possible that some of the identified top MOFs may display flexibility in response to guest adsorption or other external stimuli. There are several modes of flexibility available to MOFs, including the subnetwork displacement mode which applies specifically to 2D layered and 3D interpenetrated structures and involves relocation of separate networks (e.g., layers of 2D MOFs) in relation to each other[66]. While this and other kinds of flexibility can often be useful features of MOF when they are observed, the rigid approximation does not account for them.

The results of the high-throughput GCMC calculations[22] have been able to identify potentially promising structures for biogas upgrading and provide guidance on structural features which may promote strong biogas

Table 1 | Details of the seven MOFs for which TSN is greater than 22 mol kg^{-1} according to the GCMC calculations: refcodes, TSN values, GCMC selectivity values and rank within the dataset, GCMC binary mixture CO_2 loading and rank within the dataset, pore limiting diameter (PLD), void fraction (VF) and topology[65]

MOF Refcode	TSN /mol kg^{-1}	S (rank)	BM CO_2 Loading /mol kg^{-1} (rank)	PLD /Å	VF	Topology
DEPJIR02	32.86	695.89 (15)	11.56 (80)	3.81	0.66	sql
QUDJEF	30.67	61.50 (90)	17.15 (1)	4.27	0.70	sql[a]
UQUVOS	29.19	1765.15 (6)	8.99 (335)	4.42	0.57	hcb
AQOWIN	23.89	244.94 (30)	10.00 (191)	3.80	0.62	sql
SIKYIV	23.55	52.86 (103)	13.67 (27)	4.30	0.69	sql
YOCSEQ	22.36	6633 (4)	5.85 (940)	4.83	0.50	fsc

[a] QUDJEF was assigned a 2-fold catenated sql topology.

upgrading properties. However, acquisition of these results on a large scale is ultimately limited by the computational cost of GCMC calculations. This warrants the development of ML methods able to obtain equivalent information on a shorter timescale. If successful, ML models may allow preselection of the most promising MOFs for a targeted property, bypassing large numbers of GCMC simulations and focusing more costly efforts where they are most relevant. Moreover, while structural guidance was provided from the GCMC results by examination of individual MOFs, ML methods may extend this guidance to a larger scale, providing quantitative information about how the structure of a MOF may affect its performance. Development of ML models for prediction of biogas upgrading performance is addressed in subsequent sections.

Target and descriptor selection

ML methods were developed to make predictions regarding target values relevant to biogas upgrading using only cheaply calculated descriptors of MOF structures and using the GCMC dataset as training data. The target values to be predicted were introduced in the previous section: single component and binary mixture uptakes of CO_2 and CH_4, which relate to the affinity of a MOF for each gas, including in the presence of its competitor in the mixture, as well as TSN, which describes the trade-off between uptake and selectivity and is particularly relevant for gas separations. Further details on selection of target values can be found in Supplementary Note 2; Table S1 presents an analysis of the minimum, maximum, mean, and median of the target values, and Figs. S3 and S4 show the distribution of the target values. Initial analysis of the data revealed that the distribution of TSN was highly skewed and likely to present problems when training ML methods. Therefore, $\log_{10}$(TSN) was additionally used to reduce the skew and improve the distribution of the data.

Descriptors used as features for a ML model must be cheaply obtained and relevant to the target being predicted. As a starting point for this work, descriptors were taken from the dataset of Glover and Besley[22] previously used in the early stages of a screening for biogas upgrading with a focus on membrane separation. The 21 descriptors include a mixture of energetic and structural information which has been shown to be desirable for models of this kind[29,42]. Features selected from this set are listed in Table 2 along with the method and software used to calculate them (see other details in Table S6), and the full set of 21 features is similarly described in Table S2 of Supplementary Note 3. Through an analysis of the descriptor values (See Figs. S5–S10 of Supplementary Note 3) it was noted that the distributions of many descriptors were skewed. Hence, to allow the model to discriminate between different descriptor values, a $\log_{10}$ scale was applied to pore limiting diameter (PLD), largest cavity diameter (LCD), pore volume (PV) and some chemical descriptors at infinite dilution, namely, Henry constant (K_0), diffusion coefficient (D_0^C), and permeability (P_0). For any pair of correlated descriptors (Fig. 3a, also Table S5 and Fig. S15), only one was retained. This led to the removal of K_0 descriptors in favour of heat of adsorption, Q_0^{st}, and the removal of pore volume and GSA surface area in favour of void fraction (VF). Analysis and distribution of the $\log_{10}$ scaled descriptors can be found in Table S4 and Figs. S11−S14 of Section 3.2 in Supplementary Note 3. Full correlation analysis is presented in Section 3.3 of Supplementary Note 3 (Figs. S16−S21).

In summary, only a small number of relevant descriptors, given in Table 2, was selected to allow the ML model to be transparent and easily interpreted[24,67], and we used our understanding of gas selectivity and uptake to remove the less appropriate descriptors. Diffusion coefficients of gases and permeability at infinite dilution were removed as these descriptors relate specifically to membrane separations. The values used as targets in this study are thermodynamic properties, which, although relevant to several industrial applications, do not alone describe industrial workflows like membrane separation. Furthermore, the diffusion coefficients and permeability descriptors are obtained using molecular dynamics simulations and come with an additional computational cost (and, hence, reduced training efficiency) as compared to the other descriptors.

Random forest results

The complete ML protocol used in this study can be found in Fig. S22 of Supplementary Note 4. Plots of predictions using each ML method for each target can be seen in Figs. S23−S25. The corresponding plots of error against target range can be seen in Figs. S26−S28. Machine learning results using random forest are shown in Fig. 3 using 10-fold cross validation. Additional models were built using Multiple Linear Regression (MLR), Support Vector Machine (SVM), and k-Nearest Neighbours (kNN); results using these methods can be found in Supplementary Note 5. For the six regression targets, RF and SVM gave comparable results, outperforming MLR. RF models gave mean absolute error in the range of 0.097 to 1.224, and in all cases smaller than the standard deviation of the target values, showing the utility of the models. The best models were built for single component and binary mixture CO_2 loading and RF (Fig. 3e, f) with $R^2 > 0.9$. The predictions for TSN (Fig. 3b, c) are very skewed, giving large underpredictions for higher TSN MOFs.

Thus, we see that the regression models are able to predict uptakes more efficiently than TSN (or indirectly selectivity), which is itself ultimately based on uptake values. This likely relates to the fact that for the training set, TSN data are skewed to low values, and also to propagation of errors. Specifically, the equation for calculating selectivity (Eq (1) of the Methods section) involves division by the quantity of adsorbed CH_4 in a binary mixture simulation. Calculated CH_4 loading values are generally small, and prediction errors, which may be small when predicting only CH_4 loading, are amplified, lowering performance of the model. Meanwhile, a small number of CH_4 loading values are particularly small, resulting in a select few unusually high selectivity and TSN values and the observed low-TSN skew in the training data. In order to resolve this, additional classification models were built for TSN with the aim of identifying high and low TSN MOFs, an important task in MOFs screening. The benchmark TSN value of 5 mol kg^{-1} was selected to differentiate between high and low TSN MOFs. The selection of the benchmark value is discussed in Section S2.1 of Supplementary Note 2.

The classification models for TSN gave excellent results for kNN, SVM and RF, with RF giving the best accuracy of 0.881 and area under the receiver-operator curve (AUC) of 0.957 (Fig. 3l, o; also see Figs. S29 and S30). The few misclassifications for RF were localised around the interface between high TSN and low TSN, as shown in Fig. 3n where the class prediction is presented against the actual TSN. Almost all the incorrect classifications are within ± 2 mol kg^{-1} TSN of the interface. The probability of assigning to each class was examined against the real class (Fig. 3k). It was observed that the more confident the model, the more likely the the correct class had been predicted, with very few cases of high confidence resulting in an incorrect prediction. The complete ML protocol used in this study can be found in Fig. S22 of Supplementary Note 4, and the full metrics for

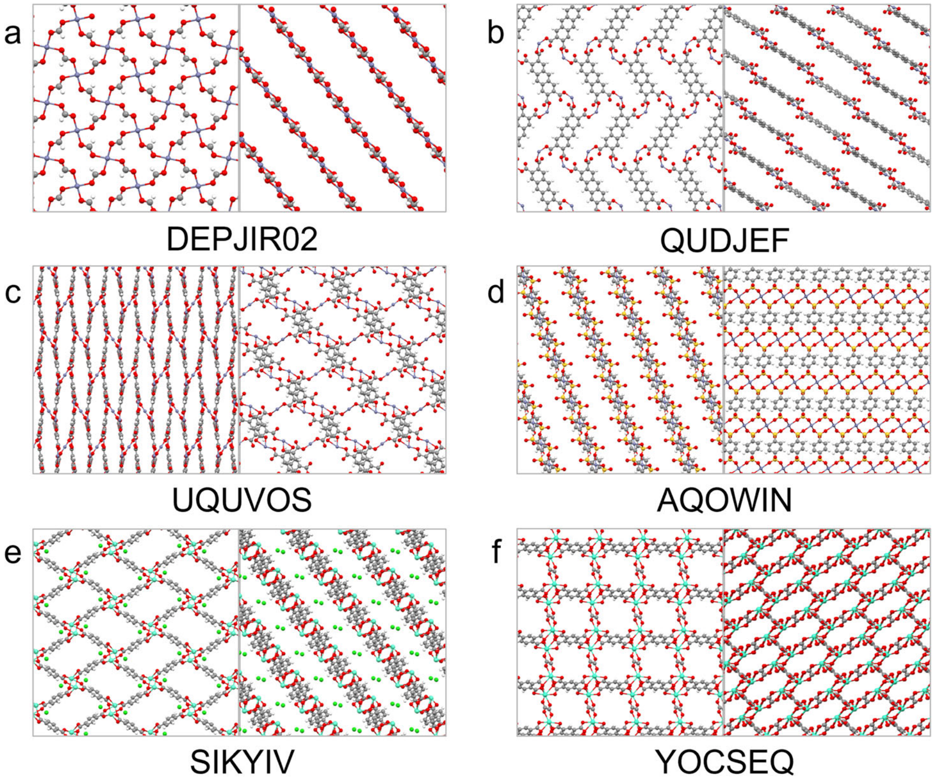


Fig. 2 | Visualisations of the structures of the six MOFs for which TSN at 10 bar is greater than 22 as predicted by GCMC simulations, visualised along two different axes. a DEPJIR02, metal is Zn (**b**) QUDJEF, metal is Zn (**c**) UQUVOS, metal is Zn (**d**) AQOWIN, metal is Zn (**e**) SIKYV, metal is Eu (**f**) YOCSEQ, metal is Cd.

regression and classification models can be seen in Table S7 of Supplementary Note 5.

Discussion

Interpretability of machine learning predictions

To interpret the RF models, the average descriptor importance was calculated from the models (Fig. 3j, m). For the regression models, density of the MOF was important for predicting single component and binary mixture CH_4 loading, whilst volume fraction available for gas uptake in the MOF was the key descriptor for single component and binary mixture CO_2 loading, and heat of adsorption of water in MOF at infinite dilution, Q_0^{st} (H_2O), for TSN. The strong influence of volume fraction on single component CO_2 loading was examined by plotting void fraction against single component CO_2 loading (Fig. 3g), indicating that single component CO_2 loading increases with void fraction up to a certain point, but for the few MOFs with very high void fraction, single component CO_2 loading begins to decrease again. Although the correlation between this descriptor and the target is strong, the model that contains 9 descriptors (Fig. 3f) gives more accurate predictions than using VF alone.

We find that for the most important features, particularly for the TSN targets, a mixture of structural and energetic descriptors is required, supporting evidence from previous studies[29,42]. For the gas uptake targets specifically, structural features dominate whilst energetic features contribute relatively little to the machine learning models. Among the structural features, void fraction stands out particularly as having very high importance for CO_2 loading and strong contribution to a number of the other models, most prominently to the TSN classification model and the TSN regression models (CO_2 loading contributes directly to TSN). The dependence of CO_2 loading on void fraction indicates that under the 10 bar conditions studied, available pore space is relevant to how readily new gas molecules may enter the MOF as unfavourable arrangements of guest molecules are necessitated. Indeed, some of the smaller-VF MOFs may have already reached saturation by this point, meaning that only structures with sufficient VF are able to continue to adsorb. Meanwhile, Fig. 3g indicates that at the highest void fraction values in the dataset, adsorption is reduced compared to at the most favourable void fractions. Here, interactions with pore walls are decreased and adsorption is reduced.

It is interesting that the energetic feature displaying the most overall importance for predicting the target values is heat of adsorption of water, Q_{st} (H_2O), which provides a description of the hydrophilicity and polarity of a MOF structure. Such a description accounts for the selectivity contribution to TSN, with highly polar MOFs likely to unite affinity for CO_2 with lack of affinity for CH_4. For both the RF regression and the classification models, Q_{st} (H_2O) provides a more important description of TSN than the heat of adsorption of either of the two individual gases for both the regression and the classification models.

The models were further analysed by examining the best- and worst-predicted MOFs for structural similarities. This was to determine the types of MOFs for which the models provided excellent and poor predictions. The analysis for the regression model with the best overall performance, random forest single component CO_2 loading, and the TSN random forest classification model can be found in Supplementary Notes 5 and 6 (see analysis of the outliers in Tables S10 and S11). For the CO_2 regression model, the majority of the largest errors correspond to an overprediction of the loading in MOFs displaying very high void fraction. Meanwhile, the majority of the best-predicted structures possess narrow pores, with smaller pore limiting diameter than the poorly predicted MOFs. For the classification model, void fraction also seems to display a relationship with prediction quality, with particularly well-predicted MOFs having void fraction falling largely within defined ranges, and low-performing MOFs which were incorrectly classified as high-performing displaying larger void fraction values. Pore limiting diameter varies among both correctly classified and incorrectly classified MOFs, but higher maximum pore limiting diameter is seen for the poorly classified MOFs than for the well-classified structures.

In the context of a MOF screening, it is instructive to ascertain the performance of a model for the MOFs displaying the best performance. In this case, therefore, the six MOFs identified as most promising for biogas upgrading by GCMC and identified in Fig. 2 are considered. Since these MOFs were identified based on TSN, the results of the RF TSN classification model were checked. It was confirmed that the TSN classification model correctly classified all six identified top-performers, all with high-

Table 2 | The 9 descriptors used to train the ML model in this work, selected from among the 21 descriptors of Glover and Besley[22]

Descriptor	Description	Method	Software
PLD/Å	Diameter of the largest sphere that can percolate through the MOF	Voronoi network	Zeo++[69]
LCD/Å	Diameter of the largest sphere that fits inside the MOF	Voronoi network	Zeo++[69]
Density/g cm^{-3}	Mass of MOF per unit volume		Zeo++[69]
VSA/m^2 cm^{-3}	Surface area accessible to the centre of a probe (r = 1.86 Å) per unit volume	Voronoi network MC sampling	Zeo++[69]
VF	Fraction of the volume not occupied by MOF atoms (calculated with point probe)	Voronoi network MC sampling	Zeo++[69]
Q_0^{st} (CH_4)/kJ mol^{-1}	Heat of adsorption of CH_4 in the MOF at infinite dilution	Force fields GCMC	RASPA[72]
Q_0^{st} (CO_2)/kJ mol^{-1}	Heat of adsorption of CO_2 in the MOF at infinite dilution	Force fields GCMC	RASPA[72]
Q_0^{st} (H_2S)/kJ mol^{-1}	Heat of adsorption of H_2S in the MOF at infinite dilution	Force fields GCMC	RASPA[72]
Q_0^{st} (H_2O)/kJ mol^{-1}	Heat of adsorption of H_2O in the MOF at infinite dilution	Force fields GCMC	RASPA[72]

Where relevant: MC = Monte Carlo, GCMC = grand canonical Monte Carlo, MD = molecular dynamics, r = probe radius.

performing probability above 0.9 with the exception of YOCSEQ (probability=0.63), suggesting that the model is successful in correctly classifying the very high-performing MOFs most desired in a screening.

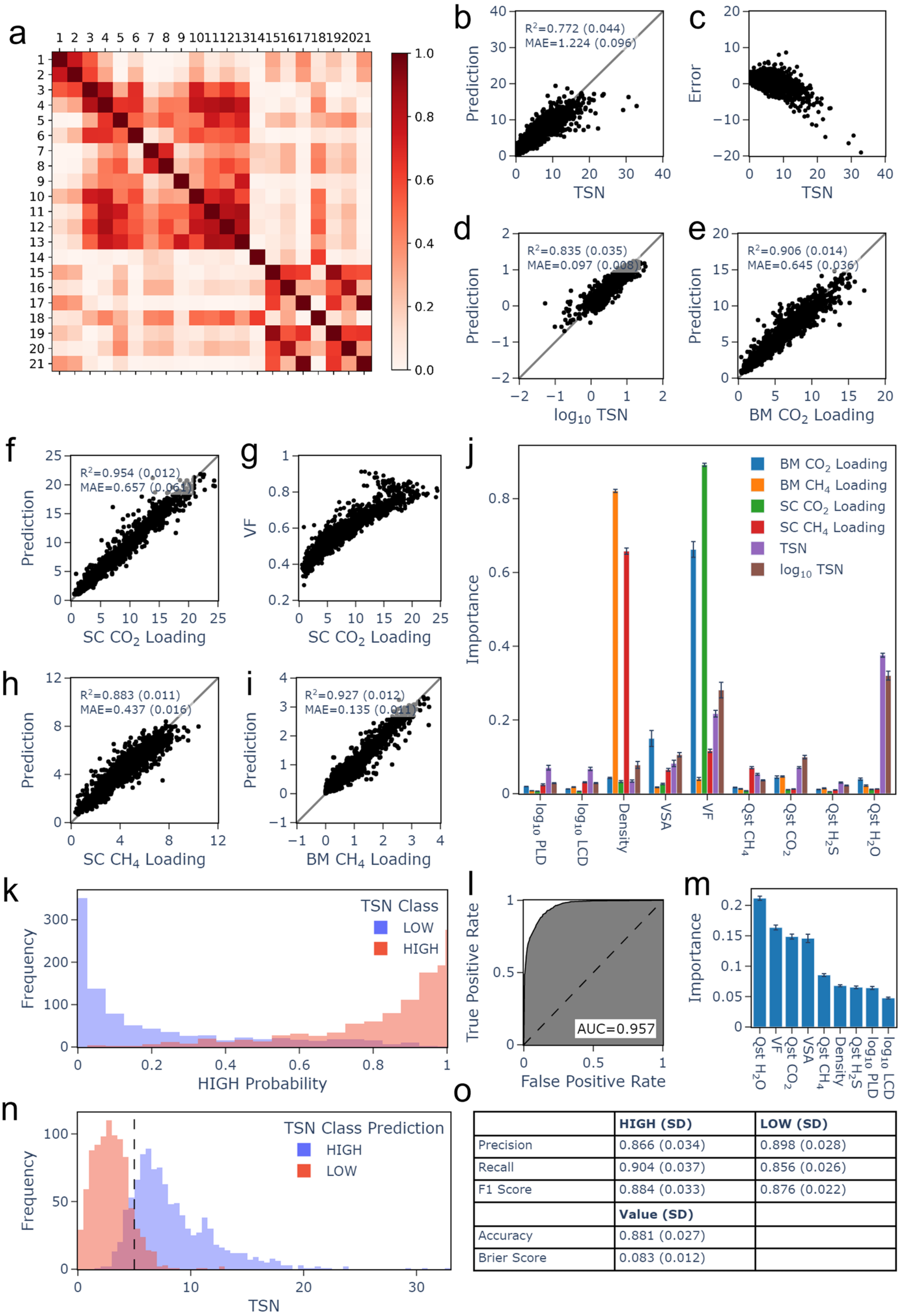
External test data

The best-predicted regression target, single component CO_2 loading, and the TSN classification model were retrained using the full training set and their performance was tested on a dataset of unseen MOFs originating from an external source. To generate this dataset, 1000 structures were selected at random from the Northwestern hypothetical database[18] and subjected to the same curation procedure applied to the training data, including the application of geometrical and charge criteria used by ref. 22 and with the addition of a check for similarity to the training set to avoid data leakage[54,55]. No MOFs were found to be identical to training set structures. The curation left a total of 330 structures in the external test dataset. The distribution of metals present in these MOFs is shown in Fig. S2 Supplementary Note 1; the dataset is heavily dominated by Zn and Cu MOFs. The nine selected descriptors were calculated for each MOF as described in the Methods section to allow predictions to be made using the ML models. Relevant targets were also calculated using GCMC as described in the Methods section for comparison to predictions. The unseen test set descriptors and targets were analysed by the same statistics as the training set, as shown in Table S8. The ranges distributions of the descriptors and targets are shown in Figs. S33−S37. The correlation of the unseen test set descriptors can be found in Fig. S38. The full metrics for test set predictions are shown in Table S9.

The results are shown in Fig. 4 for RF. The mean absolute error for regression models was lower for RF (2.275) than SVM or MLR, and much lower than the standard deviation of the external test data (3.796). This model had a modest R^2 of 0.332. However, an examination of Fig. 4a, b shows that the modest R^2 is due to underprediction of mid and high loading MOFs. Interestingly, the correlation between single component CO_2 loading and void fraction is different to the training set (Fig. 3g) and, due to the predictions' strong dependence on void fraction, the predictions are affected accordingly. For the TSN classification models, accuracy (0.712) and area under the curve (0.850) was found with RF, with comparable results for RF and kNN. The recall for HIGH MOFs was much lower than the cross validation results, partially due to the low number of HIGH MOFs in the external test set. The model was particularly good at identifying LOW MOFs (Fig. 4d, e). However, several MOFs with very high TSN were predicted LOW. The model was good at assigning the correct label when the probability was < 0.2 or > 0.8 and less so at intermediate values. To further assess the importance of each descriptor in the model, random forest models were rebuilt for the regression targets with 10-fold cross validation leaving out each descriptor in turn (see Section 5.4 of Supplementary Note 5). Figs. S31 and S32 show the mean metric and the standard deviation error across the 10 folds. Models were also trained with the full training dataset and tested on the unseen test set. The full prediction and error plots can be seen in Figs. S39−S44. The classification models were retrained using the full training set and tested on the unseen test set. The analysis for the cross validation was repeated and can be seen in Figs. S45 and S46.

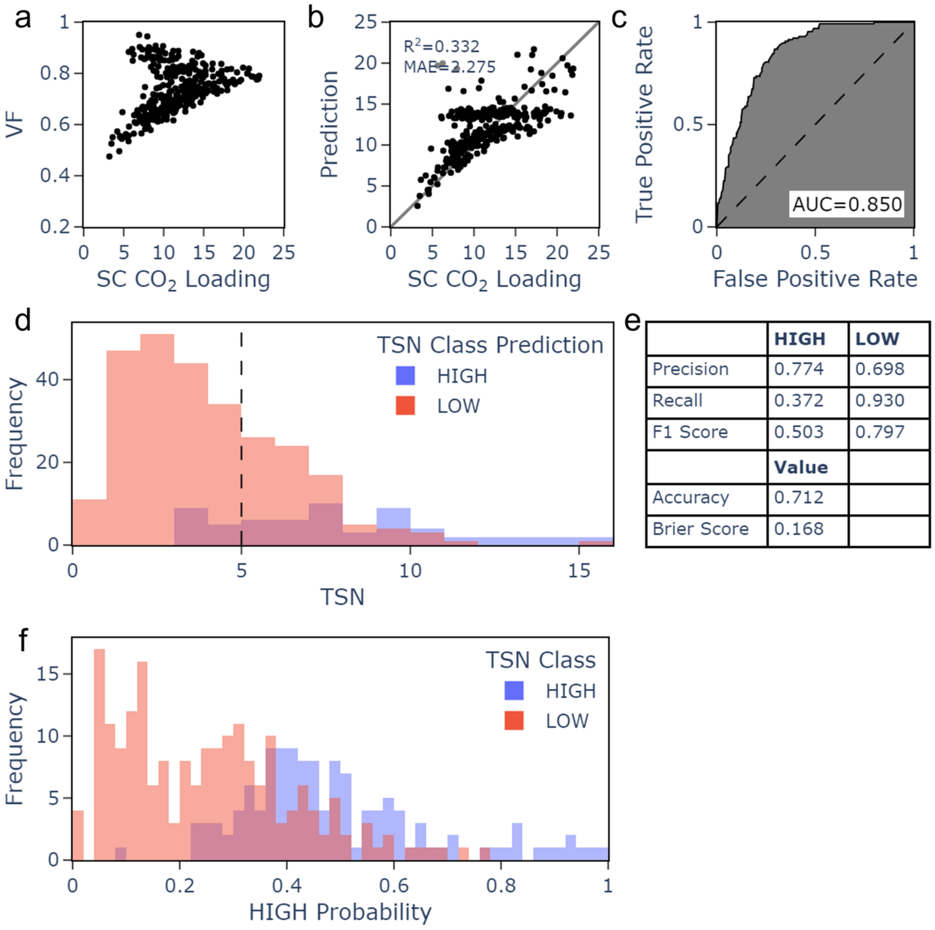
The underprediction of the performance of HIGH MOFs displaying high void fraction illustrated in Fig. 4a, b for the external test set appears to be a more extreme case of similar underprediction observed for high-void fraction MOFs in the training and validation sets. Comparison of Fig. 4a alongside Fig. 3g reveals that for both the validation set and the test set there is a peak in single component CO_2 loading at a void fraction of around 0.8, and that the test set contains proportionally many more points above this peak than the validation set does. This offers an explanation for the difference between the model performance for the training set and for the external test set: the test set contains more MOFs with particularly high void fractions leading to comparatively low loading. With fewer of these kinds of MOFs seen in training, model predictions are poorer. The feature space covered by the external test set, which is composed of hypothetical MOFs, is not the same as the feature space the models were trained on. In addition to providing valuable information about the performance of the trained models, this highlights an interesting point about the merits of detailed study of the similarities between commonly used hypothetical and real sets of MOFs. The benefits of hypothetical MOF databases are defined by the fact that they contain MOFs which have not been produced experimentally, so some level of dissimilarity between the two is advantageous. However, a database which significantly departs from structures similar to those known to be experimentally synthesisable risks containing sets of coordinates which are not useful to pursue. It is vital to understand the strengths and limitations of hypothetical MOF databases. This was partially addressed in work by ref. 68 comparing the feature space covered by MOF databases. The work identified differences between hyopthetical and real datasets and a lack of diversity in hypothetical MOFs (including, as seen in this work, a limited number of metal centres present in hypothetical MOFs), with resulting implications on screening conclusions. Consistent scrutiny of hypothetical MOF databases is needed as they continue to be widely utilised in material design and selection.

To identify the MOFs likely to be best performing according to the predictions of the random forest ML model, the six MOFs from the external test set classified as high TSN with the highest probability are detailed in Table 3 and visualised in Fig. 5. Table 3 also includes relevant geometrical features PLD and VF, and includes topology[65] determined as described in section 2.1, as well as degree of interpenetration, also determined using CrystalNets.jl[65]. Interpenetration is widely observed among the identified promising hMOFs. Similarly, the six MOFs predicted by the random forest regression model to display the highest single component CO_2 loading are detailed in Table S12 and visualised in Fig. S47 of Supplementary Note 7. In



	HIGH (SD)	LOW (SD)
Precision	0.866 (0.034)	0.898 (0.028)
Recall	0.904 (0.037)	0.856 (0.026)
F1 Score	0.884 (0.033)	0.876 (0.022)
	Value (SD)	
Accuracy	0.881 (0.027)	
Brier Score	0.083 (0.012)	

Fig. 3 | Descriptor selection and random forest results using 10-fold cross validation and nine final descriptors. (**a**) Correlation heat map for descriptors proposed in ref. 22 (numbered as in Table S3) (**b**, **d**−**f**) and (**h**, **i**) Regression target predictions (**c**) Error of TSN predictions compared to TSN range (**g**) Relationship between VF, which was identified as the most important descriptor in this model, and SC CO2 Loading. (**j**) Relative importance of each descriptor for each regression target. (**k**) Frequency of HIGH TSN prediction probability compared to the real class. (**l**)TSN classification model receiver operating characteristic curve. (**m**) Relative importance of each descriptor for TSN classification. (**n**) Frequency of HIGH and LOW predictions compared to TSN range. **o** Metrics for TSN classification model. Throughout, the standard deviation across 10 folds is shown in parentheses as a measure of model variability.



	HIGH	LOW
Precision	0.774	0.698
Recall	0.372	0.930
F1 Score	0.503	0.797
	Value	
Accuracy	0.712	
Brier Score	0.168	

Fig. 4 | External test set results using random forest retrained on the full training set and 9 final descriptors. (**a**) Relationship between VF and SC CO_2 Loading (**b**) Predictions for SC CO_2 Loading (**c**) Receiver-operator characteristic curve for TSN classification. (**d**) Frequency of HIGH and LOW predictions compared to TSN range. (**e**) TSN classification metrics. (**f**) Frequency of HIGH TSN prediction probability compared to the real class.

general, the predictions are fairly successful, with the predicted top TSN MOFs displaying high GCMC TSN, clearly above the threshold of 5 mol kg^{-1} and the predicted top loading MOFs displaying predicted loading close to their GCMC values. Several of the top MOFs for both properties display structural similarities: all top-TSN MOFs and most top-SC CO_2 loading MOFs possess interpenetrated frameworks based on the pcu (primitive cubic) lattice. Higher degrees of interpenetration are observed for the top-TSN MOFs than for the top-uptake MOFs. For the uptake model, the top MOFs have void fraction very close to the value of 0.8 which corresponds to the loading peak previously mentioned. All 6 of the MOFs predicted to display the highest single-component CO_2 loading have VF between 0.79 and 0.87, in common with 86 (26%) of the 330 MOFs in the external test set. Meanwhile, 5 of these 6 (83%) have VF between 0.79 and 0.82, in common with 47 (14%) of the 330 external test set structures. For the TSN model, the top MOFs generally possess VF lower than 0.8. This is true of all six predicted top-TSN MOFs, in common with 242 of the 330-strong test set (73%). Five of the six top predictions have VF lower than 0.72, in common with 145 (44%) of the 330 test MOFs. A range of values for pore limiting diameter is observed, in line with the lower feature importance of this metric, although top MOFs for loading generally have higher pore limiting diameter than top MOFs for TSN. The six top-SC-CO_2 loading MOFs have PLD ranging from 6.69 Å to 11.5 Å, a range covered by 109 (33%) of the full curated test set, while the six top-TSN MOFs have PLD ranging from 4.40 Å to 6.69 Å, a range covered by 154 (47%) of the full curated test set. None of the top predicted MOFs are two dimensional; the external dataset of hypothetical MOFs contained only 3D structures. All of the top six TSN MOFs and the majority of the top six MOFs for single component CO_2 loading (4 of the 6, 67%) possess Zn centres. This is in line with similar observations seen in the training and validation set, but must largely be attributed to the significant dominance of Zn centres in the test set (247 of 330, or 75%, see Figure S2 of Supplementary Note 1).

Table 3 | Details of the six MOFs predicted to display high TSN with the highest probability among the external test set according to the RF classification model: the identifier, probability, calculated (GCMC) TSN, pore limiting diameter (PLD), void fraction (VF), topology, and degree of interpenetration (DI)

Numerical Identifier	High TSN probability	GCMC TSN /mol kg^{-1}	PLD /Å	VF	Topology	DI
13060	1.00	11.0	4.71	0.62	pcu	4
5030899	0.994	9.60	5.11	0.67	pcu	4
5022978	0.966	6.33	5.26	0.71	pcu	3
3886	0.956	12.2	6.69	0.79	pcu	2
5030750	0.938	10.9	4.40	0.65	pcu	4
5032593	0.930	7.04	6.42	0.71	pcu	3

Methods

Dataset curation and cleaning

The curation workflow used to refine the dataset to contain only physically viable MOFs is outlined below. Further details are given in Supplementary Note 1. Structures containing no metal, no carbon or only one or two elements were first identified and removed from the initial dataset taken from ref. 22 of 6768 MOFs, leaving a total of 6663 structures. To combat overlapping atoms, any structures with any atom-atom distances less than 0.5 Å were next removed, leaving 6,638 structures in the dataset. Any D atoms present in any crystallographic information files were replaced with H

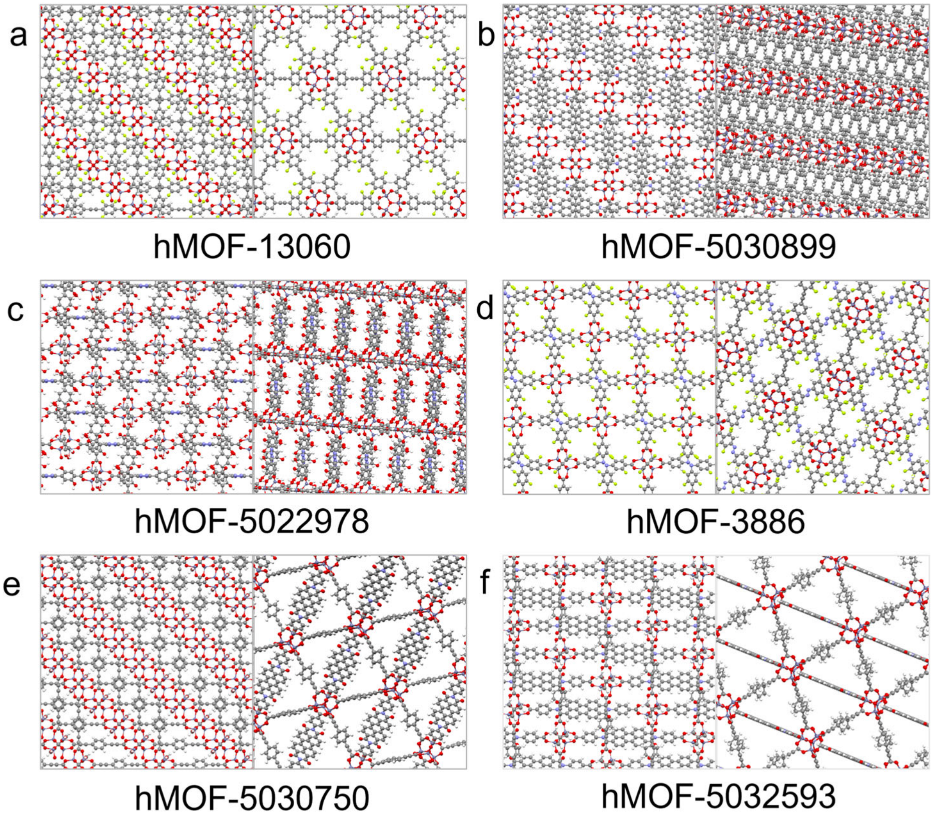


Fig. 5 | Visualisations of the structures of the six MOFs predicted to have high TSN with the highest probability among the external test set by the random forest classification model. The MOFs are (**a**) hMOF-13060, metal is Zn (**b**) hMOF-5030899, metal is Zn (**c**) hMOF-5022978, metal is Zn (**d**) hMOF-3886, metal is Zn (**e**) hMOF-5030750, metal is Zn (**f**) hMOF-5032593, metal is Zn.

and then all files with no H were removed, leaving 6359 structures. Issues dealt with up to this stage would be clear from visualisation.

Oxidation state counting was then used to remove structures with unviable oxidation states using a published MOF Oxidation State And Electron Count (MOSAEC) code[44]. The MOSAEC code was applied to all 6359 remaining structures and any that were flagged with any features indicating problematic oxidation states were removed. A total of 3553 structures were flagged as problematic, leaving only 3086 MOFs in the dataset. A filter for dimensionality was then applied: the dimensionality of each of the 3086 structures was determined algorithmically using the Zeo++ software package[69]. Of these, 1715 3D structures, 686 2D structures, 611 1D structures and 74 0D structures were identified. All 1D and 0D structures were removed from the database, leaving a total of 2401 MOFs. Of these 2401 MOFs, charge equilibration calculations necessary for the simulations were unable to complete for 20 structures, leaving a total of 2381.

A final step was removal of duplicate structures[54,55]. Weisfeiler-Lehman structure hashes for undecorated (agnostic to atom type) structural graphs of each MOF were obtained using the Pymatgen materials analysis python library[70] and the NetworkX package[71]. For any groups of structures with identical hashes, only one MOF was retained. Additionally, groups of structures whose CSD refcode is based on the same six-letter string are based on the same MOF, possibly with changes to the synthesis procedure, and are therefore expected to by highly similar. Only one MOF was retained for each six-letter string. Following deduplication, and therefore following the full curation procedure, 1910 MOFs remained in the CSD dataset. An equivalent curation procedure was applied to the external test set, yielding 330 curated structures.

Calculation of descriptors and labels

Computational details of the Monte Carlo setup used within the RASPA software package[72] to determine absolute gas uptake are outlined below. Interactions between components of the system were modelled using Lennard-Jones potentials, considering host-guest and guest-guest interactions, but no host-host interactions. The guest molecules CO_2 and CH_4 were modelled using the TraPPE formalism, in which CO_2 is treated with a 3-site model with each site possessing a partial charge and Lennard-Jones parameters, and CH_4 is treated using a single site model with no charge but with Lennard-Jones parameters. Lennard-Jones parameters for the framework were taken from the Universal Force Field (UFF)[73]. Framework partial charges necessary to model interactions of framework atoms with CO_2 were determined prior to the uptake simulations using the extended charge equilibration (eQeq) method available within the RASPA software package (version 2.0.35)[72].

During GCMC simulations, van der Waals interactions were calculated with a cutoff of 12.8 Å, above which the potential was shifted to zero without tail corrections. Often, tail corrections can be a convenient choice to make the adsorption results less sensitive to the details of the truncation[50]. Unlike atoms were treated using Lorentz-Berthelot mixing rules. Sufficient replicas of the unit cell of each MOF were used so that all perpendicular distances were at least 25.6 Å (two times the cutoff). Electrostatic interactions where required were calculated using the Ewald method with precision of 10^{-6}. The MOF system was in equilibrium with an imaginary reservoir of gas molecules, either pure CO_2, pure CH_4 or a CO_2/CH_4 mixture with each gas present with a mole fraction of 0.5. Gas phase fugacities of the components were calculated using the Peng-Robinson equation of state. The temperature used was 298 K and the external pressure modelled was 10 bar. Monte Carlo moves available to the system were translation of a guest, rotation of a guest, insertion or deletion of a guest (grouped together by the software under a move called swap) and in the binary mixture case changing a guest's identity from CO_2 to CH_4 or the reverse. Simulations were run for 10,000 equilibration cycles followed by 10,000 production cycles, where the number of steps in a cycle is equal to the number of molecules in the system, or to 20 if there are fewer than 20 molecules in the system. The computational cost of these GCMC calculations of gas adsorption under relevant conditions varies depending on the MOF considered. The time for an uptake calculation to complete for a single MOF ranges from several hours to multiple days when running on a high-performance compute cluster possessing an Intel Xeon Gold 6138 20C 2.0 GHz CPU.

Further separation metrics are determined from binary mixture CO_2 and CH_4 loading. Selectivity, S, quantifies the preferential adsorption of CO_2 over CH_4 and is given by Eq. (1), where q_i is the quantity of gas i in the adsorbed phase, and y_i is the mole fraction of gas i in the gas phase.

$$S_{CO_2/CH_4} = \frac{q_{CO_2} y_{CH_4}}{q_{CH_4} y_{CO_2}} \qquad (1)$$

TSN_{CO_2/CH_4} quantifies the trade-off between S_{CO_2/CH_4} and binary mixture CO_2 uptake. It is given in equation (2), and is referred to in this work as TSN.

$$TSN_{CO_2/CH_4} = N_{CO_2} \log\left(S_{CO_2/CH_4}\right) \qquad (2)$$

Descriptors for the training data used in this study were taken from a previous high-throughput screening of the training MOFs for membrane separation of biogas mixtures[22]. Equivalent descriptors were calculated for the external test set as follows, mirroring the setup used in the previous screening for consistency.

Structural descriptors, PLD, LCD, density, VSA, GSA, VF and PV, were calculated using the Zeo++ software package using high accuracy settings in all cases[69]. Surface area was calculated with a probe radius of 1.86 Å and with 5000 Monte Carlo cycles. This is the size of a spherical probe conventionally used to represent the N_2 molecule, which is commonly used for experimental surface area measurements[74]. It is also consistent with the methods used by Glover and Besley to generate other data used in this work[22]. Volume was calculated with a point probe and 50,000 Monte Carlo cycles. Structural features for a single MOF can be calculated within seconds. The energetic features, infinite dilution heats of adsorption, were calculated using the Widom insertion method with 100,000 Monte Carlo cycles at 298 K. Moves available for selection were translation, rotation and reinsertion. A heat of adsorption value can be obtained in under half an hour.

Machine learning

Machine learning was implemented with *Python/Scikit-learn*[75]. Separate models were built for the six regression targets and one classification target using Support Vector Machine and Random Forest. Additionally, for regression Multiple Linear Regression was used and for classification k-Nearest Neighbours were also built. The number of trees used for Random Forest models was 500, with the default values used for all other parameters. Hyperparameters for the Support Vector Machine models were optimised using the training set and a 5-fold cross validation grid search within the 10-fold cross validation protocol (nested cross validation) to avoid data leakage. From a wide range of values, the optimal values for γ and C were found for all models, in addition to the optimal ϵ value for regression models. C is the regularisation parameter and specifies the penalty for an incorrect prediction. γ is the kernel coefficient, a term in the kernel which transforms the data in the algorithm. ϵ is a term in the regression model which specifies the allowed error between the data and function for which no penalty is applied. The kernel was set to the radial basis function (rbf) and the default values were used for all other hyperparameters. The optimal number of neighbours for k-Nearest Neighbours was optimised between 1-99. Models were initially tested using 10-fold cross validation using the full dataset. By analysing the predictions for each fold, average metrics were calculated with the error as the standard deviation of the metrics across the 10 folds. For regression models, Mean Absolute Error (MAE) and R^2 score (coefficient of determination) were used as metrics. For the TSN classification model, both the predicted class and probability of that class were obtained, where HIGH was assigned the positive label, i.e., a probability of 0.9 denoted 90% confidence the instance belonged to the HIGH class and 10% confidence the instance belonged to the LOW class. Models were assessed with the precision, recall, F1 score of the classes and the overall accuracy of the models. In addition, the Brier score, the receiver operating characteristic (ROC), and the resultant area under the curve (AUC) were calculated from the class probabilities. For full explanation of metrics used in this work, see Supplementary Note 4. The average importance of each descriptor was assessed using the standard protocol employed in the feature_importances_ attribute of each RF model in the 10-fold cross validation. Again the error was the standard deviation of the importance across the 10 folds. The optimal hyperparamters were recalculated using 5-fold cross validation for the full dataset. These parameters were used to train models using the full dataset. The resultant models were used to predict targets for the external test data. These predictions were analysed using the same methods mentioned above.

Conclusions

In this work, GCMC simulations have been used to obtain computational measures of biogas separation metrics in a range of curated MOF structures with a view to identify MOFs promising for biogas upgrading. Highly accurate machine learning models have been trained to make predictions regarding these metrics efficiently and with a lower computational cost than GCMC. Regression models were trained to predict CO_2 and CH_4 loading, and classification models were trained to classify MOFs as high-performing or low-performing. Models were trained on a carefully curated set of real structures, ensuring the data was high quality. An advantage of machine learning is that after the model has been trained and the descriptors have been calculated, predictions are almost instantaneous. The descriptors used can be computationally obtained several times more cheaply than the biogas upgrading metrics predicted. The machine learning protocol therefore represents a clear cost reduction for high-throughput screening compared to a conventional GCMC-only approach.

The models were extensively tested both on validation data and on an external test set of hypothetical MOFs. Models displayed strong performance, especially for CO_2 and CH_4 loading regression models and TSN classification models, with random forest models showing the best overall performance. Predicting an external test set of hypothetical MOFs was challenging. This can be rationalised by the distribution of the training and test sets, which do not cover precisely the same areas of target and feature space. There is therefore scope for future improvement of supplementing the training data to increase the coverage of the models. The difference between the performance of the models on data from different sources also highlights the diversity of data between different kinds of MOF databases, and the differences between hypothetical and real MOF datasets.

The results of both the GCMC simulations and the machine learning models were used to identify selected MOFs which may be promising for biogas upgrading, and also to identify structural features which are common to high-TSN MOFs. It was seen that 2D MOFs with narrow separation between sheets, as well as interpenetrated frameworks based on square or cubic topology, are among those which may be useful for the application. It was also seen that void fractions around 0.8 facilitate optimum uptake of CO_2 while slightly lower void fractions around 0.7 facilitate an optimum trade-off between uptake and selectivity.

Model availability

In order to use the model, first descriptors should be calculated by following the steps described in Section 4.2. All the data and code required to run the models presented in this work is available on GitHub at https://github.com/samuel-boobier/ML-MOFs. In addition, simple instructions of how to reproduce the models or make predictions for your own test sets are given. Finally, sample RASPA input files and curation procedures are provided.

Data availability

All correspondence and material requests should be made to Professor Elena Besley, School of Chemistry, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom. Email: Elena.Besley@nottingham.ac.uk. Full details of the code and datasets used in this work can be accessed at https://github.com/samuel-boobier/ML-MOFs. Further analysis of the data and models presented in this work are provided as Supporting Information.

Received: 15 November 2023; Accepted: 2 April 2024;
Published online: 08 May 2024

References

1. Bluemling, B., Mol, A. P. & Tu, Q. The social organization of agricultural biogas production and use. *Energy Policy* **63**, 10–17 (2013).
2. Chen, X. Y., Vinh-Thang, H., Ramirez, A. A., Rodrigue, D. & Kaliaguine, S. Membrane gas separation technologies for biogas upgrading. *RSC Adv.* **5**, 24399–24448 (2015).
3. Struk, M., Kushkevych, I. & Vítězová, M. Biogas upgrading methods: recent advancements and emerging technologies. *Rev. Environ. Sci. Biotechnol.* **19**, 651–671 (2020).
4. Ahmed, S. F. et al. Biogas upgrading, economy and utilization: a review. *Environ. Chem. Lett.* **19**, 4137–4164 (2021).
5. Canevesi, R. L., Andreassen, K. A., Da Silva, E. A., Borba, C. E. & Grande, C. A. Pressure swing adsorption for biogas upgrading with carbon molecular sieve. *Ind. Eng. Chem. Res.* **57**, 8057–8067 (2018).
6. Shah, M. S., Tsapatsis, M. & Siepmann, J. I. Identifying optimal zeolitic sorbents for sweetening of highly sour natural gas. *Angew. Chem. Int. Ed.* **55**, 5938–5942 (2016).
7. Qiao, Z., Xu, Q. & Jiang, J. Computational screening of hydrophobic metal-organic frameworks for the separation of H_2S and CO_2 from natural gas. *J. Mater. Chem. A* **6**, 18898–18905 (2018).
8. Hoskins, B. F. & Robson, R. Infinite polymeric frameworks consisting of three dimensionally linked rod-like segments. *J. Am. Chem. Soc.* **111**, 5962–5964 (1989).
9. Yaghi, O. & Li, H. Hydrothermal synthesis of a metal-organic framework containing large rectangular channels. *J. Am. Chem. Soc.* **117**, 10401–10402 (1995).
10. Chui, S. S., Lo, S. M., Charmant, J. P., Orpen, A. G. & Williams, I. D. A chemically functionalizable nanoporous material $[Cu_3(TMA)_2 (H_2O)_3]_n$. *Science* **283**, 1148–1150 (1999).
11. Li, H., Eddaoudi, M., O'Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
12. Long, J. et al. Metal-organic framework materials as catalysts. *Chem. Soc. Rev.* **38**, 1450–1459 (2009).
13. Kreno, L. E. et al. Metal-organic framework materials as chemical sensors. *Chem. Rev.* **112**, 1105–1125 (2012).
14. Horcajada, P. et al. Metal-organic frameworks in biomedicine. *Chem. Rev.* **112**, 1232–1268 (2012).
15. Li, H. et al. Porous metal-organic frameworks for gas storage and separation: status and challenges. *EnergyChem* **1**, 100006 (2019).
16. Aniruddha, R., Sreedhar, I. & Reddy, B. M. MOFs in carbon capture - past, present and future. *J. CO2 Util.* **42**, 101297 (2020).
17. Moghadam, P. Z. et al. Targeted classification of metal-organic frameworks in the Cambridge Structural Database (CSD). *Chem. Sci.* **11**, 8373–8387 (2020).
18. Wilmer, C. E. et al. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
19. Gómez-Gualdró, D. A. et al. Evaluating topologically diverse metal-organic frameworks for cryo-adsorbed hydrogen storage. *Energy Environ. Sci.* **9**, 3279–3289 (2016).
20. Avci, G., Velioglu, S. & Keskin, S. High-throughput screening of MOF adsorbents and membranes for H_2 purification and CO_2 capture. *ACS Appl. Mater. Interfaces* **10**, 33693–33706 (2018).
21. Sikora, B. J., Wilmer, C. E., Greenfield, M. L. & Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137 000 hypothetical metal-organic frameworks. *Chem. Sci.* **3**, 2177–2396 (2012).
22. Glover, J. & Besley, E. A high-throughput screening of metal-organic framework based membranes for biogas upgrading. *Faraday Discuss.* **231**, 235–257 (2021).
23. Li, X. et al. Machine learning-assisted crystal engineering of a zeolite. *Nat. Commun.* **14**, 3152 (2023).
24. Boobier, S., Hose, D. R., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **11**, 5753 (2020).
25. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
26. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**, 1186 (2022).
27. Orhan, I., Le, T., Babarao, R. & Thornton, A. Accelerating the prediction of CO_2 capture at low partial pressures in metal-organic frameworks using new machine learning descriptors. *Commun. Chem.* **6**, 214 (2023).
28. Cao, Z., Magar, R., Wang, Y. & Farimani, A. MOFormer: self-supervised transformer model for metal-organic framework property prediction. *J. Am. Chem. Soc.* **145**, 2958–2967 (2023).
29. Simon, C. M., Mercado, R., Schnell, S. K., Smit, B. & Haranczyk, M. What are the best materials to separate a xenon/krypton mixture? *Chem. Mater.* **27**, 4459–4475 (2015).
30. Fernandez, M., Woo, T. K., Wilmer, C. E. & Snurr, R. Q. Large-scale Quantitative Structure-Property Relationship (QSPR) analysis of methane storage in metal-organic frameworks. *J. Phys. Chem. Lett.* **117**, 7681–7689 (2013).
31. Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *J. Phys. Chem. Lett.* **5**, 3056–3060 (2014).
32. Suyetin, M. The application of machine learning for predicting the methane uptake and working capacity of MOFs. *Faraday Discuss.* **231**, 224–234 (2021).
33. Shi, Z. et al. Machine-learning-assisted high-throughput computational screening of high performance metal-organic frameworks. *Mol. Syst. Des. Eng.* **5**, 725–742 (2020).
34. Pétuya, R. et al. Machine-learning prediction of metal-organic framework guest accessibility from linker and metal chemistry. *Ang. Chem. Int. Ed.* **61**, e202114573 (2022).
35. Bennett, S. & Jelfs, K. E. Porous molecular materials: exploring structure and property space with software and artificial intelligence. in *AI-Guided Design and Pproperty Prediction for Zeolites and Nanoporous Materials,* (eds. Sastre, G. & Daeyaert, F.), 251–282 (Wiley, Chichester, UK, 2023).
36. Nandy, A., Duan, C. & Kulik, H. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal-organic frameworks. *J. Am. Chem. Soc.* **143**, 17535–17547 (2021).
37. Bailey, T. et al. Gradient boosted machine learning model to Predict H_2, CH_4, and CO_2 uptake in metal-organic frameworks using experimental data. *J. Chem. Inf. Model.* **63**, 4545–4551 (2023).
38. Aghaji, M. Z., Fernandez, M., Boyd, P. G., Daff, T. D. & Woo, T. K. Quantitative structure-property relationship models for recognizing metal organic frameworks (MOFs) with High CO_2 working capacity and CO_2/CH_4 selectivity for methane purification. *Eur. J. Inorg. Chem.* **2016**, 4505–4511 (2016).
39. Yulia, F., Chairina, I. & Zulys, A. Multi-objective genetic algorithm optimization with an artificial neural network for CO2/CH4 adsorption prediction in metal-organic framework. *Therm. Sci. Eng. Prog.* **25**, 100967 (2021).
40. Cheng, X. et al. Multi-scale design of MOF-based membrane separation for CO2/CH4 mixture via integration of molecular simulation, machine learning and process modeling and simulation. *J. Membr. Sci.* **672**, 121430 (2023).
41. Demir, H., Daglar, H., Gulbalkan, H., Aksu, G. & Keskin, S. Recent advances in computational modeling of MOFs: From molecular simulations to machine learning. *Coord. Chem. Rev.* **484**, 215112 (2023).
42. Wu, X., Xiang, S., Su, J. & Cai, W. Understanding quantitative relationship between methane storage capacities and characteristic properties of metal-organic frameworks based on machine learning. *J. Phys. Chem. C.* **123**, 8550–8559 (2019).

43. Chen, T. & Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal-Organic Framework Database. *RSC Adv.* **10**, 26944–26951 (2020).
44. Burner, J. et al. ARC-MOF: a diverse database of metal-organic frameworks with DFT-derived partial atomic charges and descriptors for machine learning. *Chem. Mater.* **35**, 900–916 (2023).
45. Anderson, R. & Gómez-Gualdrón, D. A. Large-scale free energy calculations on a computational metal-organic frameworks database: toward synthetic likelihood predictions. *Chem. Mater.* **32**, 8106–8119 (2020).
46. Cooley, I. & Besley, E. Do residual solvent molecules always hinder gas sorption in metal-organic frameworks? *Chem. Mater.* **36**, 219–231 (2024).
47. Velioglu, S. & Keskin, S. Revealing the effect of structure curations on the simulated CO_2 separation performances of MOFs. *Mater. Adv.* **1**, 341–353 (2020).
48. Fan, F. J. & Shi, Y. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorg. Med. Chem.* **72**, 117003 (2022).
49. Artrith, N. et al. Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508 (2021).
50. Jablonka, K. M. et al. Connecting lab experiments with computer experiments: making “routine" simulations routine (ChemRxiv Preprint). https://doi.org/10.26434/chemrxiv-2021-h3381-v2 (2021).
51. Charlambous, C. et al. Chedding Light on the Stakeholders' Perspectives for Carbon Capture. (ChemRxiv Preprint). https://doi.org/10.26434/chemrxiv-2023-sn90q (2023).
52. Fu, X., Xie, T., Rosen, A. S., Jaakkola, T. & Smith, J. MOFDiff: Coarse-Grained Diffusion for Metal-Organic Framework Design (Rxiv Preprint). https://doi.org/10.48550/arXiv.2310.10732 (2023).
53. Moghadam, P. Z. et al. Development of a Cambridge Structural Database Subset: a collection of metal-organic frameworks for past, present, and future. *Chem. Mater.* **29**, 2618–2625 (2017).
54. Jablonka, K. M., Rosen, A. S., Krishnapriyan, A. S. & Smit, B. An ecosystem for digital reticular chemistry. *ACS Cent. Sci.* **9**, 563–581 (2023).
55. Ongari, D., Talirz, L., Jablonka, K. M., Siderius, D. W. & Smit, B. Data-driven matching of experimental crystal structures and gas adsorption isotherms of metal-organic frameworks. *J. Chem. Eng. Data* **67**, 1743–1756 (2022).
56. Cooley, I., Efford, L. & Besley, E. Computational predictions for effective separation of Xenon/ Krypton gas mixtures in the MFM family of metal-organic frameworks. *J. Phys. Chem. C.* **126**, 11475–11486 (2022).
57. Humby, J. D. et al. Host-guest selectivity in a series of isoreticular metal-organic frameworks: observation of acetylene-to-alkyne and carbon dioxide-to-amide interactions. *Chem. Sci.* **10**, 1098–1106 (2019).
58. Shang, J. et al. Separation of CO_2 and CH_4 by pressure swing adsorption using a molecular trapdoor chabazite adsorbent for natural gas purification. *Ing. Eng. Chem. Res* **59**, 7857–7865 (2020).
59. PG&E. Pressure Swing Adsorption Technical Analysis [White paper]. https://www.pge.com/pge_global/common/pdfs/for-our-business-partners/interconnection-renewables/interconnections-renewables/PressureSwingAdsorption_TechnicalAnalysis.pdf (2018).
60. Ko, D., Siriwardane, R. & Biegler, L. T. Optimization of a pressure-swing adsorption process using Zeolite 13X for CO_2 Sequestration. *Ind. Eng. Chem. Res.* **42**, 339–348 (2003).
61. Jiang, L. et al. Comparative analysis on temperature swing adsorption cycle for carbon capture by using internal heat/mass recovery. *Appl. Therm. Eng.* **169**, 114973 (2020).
62. Chen, L. et al. Temperature swing adsorption for CO_2 capture: thermal design and management on adsorption bed with single-tube/three-tube internal heat exchanger. *Appl. Therm. Eng.* **199**, 117538 (2021).
63. Huertas, J. I., Giraldo, N. & Izquierdo, S. *Mass Transfer in Chemical Engineering Processes* (InTech, 2010).
64. Ghanbari, T., Abnisa, F. & Wan Daud, W. M. A. A review on production of Metal Organic Frameworks (MOF) for CO_2 adsorption. *Sci. Total Environ.* **707**, 135090 (2020).
65. Zoubritzky, L. & Coudert, F. X. CrystalNets.jl: identification of crystal topologies. *SciPost Chem.* **1**, 005 (2022).
66. Schneemann, A. et al. 2D framework materials for energy applications. *Chem. Sci.* **12**, 1600–1619 (2021).
67. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 1–7 (2016).
68. Moosavi, S. M. et al. Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **11**, 4068 (2020).
69. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Micropor. Mesopor. Mater.* **149**, 134–141 (2012).
70. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comp. Mater. Sci.* **68**, 314–319 (2013).
71. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference*, 11–15 (Pasadena, CA USA, 2008).
72. Dubbeldam, D., Calero, S., Ellis, D. E. & Snurr, R. Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **42**, 81–101 (2016).
73. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
74. Ongari, D. et al. Accurate characterization of the pore volume in microporous crystalline materials. *Langmuir* **33**, 14529–14538 (2017).
75. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

The authors acknowledge exploratory work carried out by MSci students, Mr. Ian Chester and Mr. Jonathan Lister, as part of “Machine Learning in Science Project" conducted by the University of Nottingham. Computational work has been supported by the University of Nottingham's Augusta High Performing Computing service. J.D.H. is supported by the Department of Science, Innovation and Technology (DSIT) and Royal Academy of Engineering under the Chairs in Emerging Technologies scheme (CiET2021 17). E.B. acknowledges a Royal Society Wolfson Fellowship, a Royal Society Leverhulme Trust Senior Research Fellowship, and EPSRC Programme Grant 'Metal Atoms on Surfaces and Interfaces (MASI) for Sustainable Future' (EP/V000055/1) for financial support.

Author contributions

E.B. devised the concept of the project and supervised it. I.C. and S.B. carried out computations. E.B., I.C., S.B. and J.D.H. discussed the results and contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-024-01166-7.

Correspondence and requests for materials should be addressed to Elena Besley.

Peer review information *Communications Chemistry* thanks Arthur Hardiagon and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints