Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

Cite this: J. Mater. Chem. A, 2020, 8, 3228

Machine-learning-assisted screening of pure-silica zeolites for effective removal of linear siloxanes and derivatives†

Shiru Lin,[a] Yekun Wang,[b] Yinghe Zhao,[a] Luis R. Pericchi,[b] Arturo J. Hernández-Maldonado *[c] and Zhongfang Chen *[a]

Received 29th October 2019
Accepted 21st December 2019

DOI: 10.1039/c9ta11909d

rsc.li/materials-a

As emerging organic contaminants, siloxanes have severe impacts on the environment and human health. Simple linear siloxanes and derivates, trimethylsilanol (TMS), dimethylsilanediol (DMSD), monomethylsilanetriol (MMST), and dimethylsulfone ($DMSO_2$), are four persistent and common problematic compounds (PCs) from the hydroxylation and sulfuration of polydimethylsiloxanes. Herein, through a two-step computational process, namely Grand Canonical Monte Carlo (GCMC) simulations and machine learning (ML), we systematically screened 50 959 hypothetical pure-silica zeolites and identified 230 preeminent zeolites with excellent adsorption performances with all these four linear siloxanes and derivates. This work vividly demonstrates that the collocation of data-driven science and computational chemistry can greatly accelerate materials discovery and help solve the most challenging separation problems in environmental science.

Introduction

Siloxanes refer to a class of silicone derivatives containing Si–O bonding[1] and are classified into linear and cyclic compounds. Among others, siloxanes are widely used in medicine, cosmetics, personal care products, and industrial applications such as lubricants, paints, biomedical products and anti-foaming agents.[2–4] In 2018, the world sale volume of siloxanes reached *ca.* 2.8 million tonnes. However, siloxanes are also emerging organic contaminants.[5–11] Due to their high vapor pressure ranging from 124.5 Pa (octamethylcyclotetrasiloxane) to 2.26 Pa (dodecamethylcyclohexasiloxane),[12] siloxanes are persistent and prone to bioaccumulation,[13–18] thus it remains a grand challenge to remove them from various environmental media.[19–23] Meanwhile, the release of siloxanes has severe impacts, for instance, potential toxic effects, namely oestrogen mimicking, connective tissue disorder, adverse immunologic effects, and eventually fatal liver or lung damage in exposed animals.[24,25] Even worse, siloxanes could mask the presence of other contaminants in the detection systems, which hinders the effective removal of other pollutants.

Developing suitable sorbents is a cost-effective solution[26,27] to the notorious siloxane removal problem.[28] Along this line, various adsorbents, such as ion exchange resin[29,30] and activated carbon,[3,31,32] have been explored, but their adsorption abilities are far from satisfactory due to the low affinity.[33] Therefore, it is of paramount importance to search for high-performance sorbents to remove siloxanes effectively.[29,34–36]

Pure-silica zeolites (PSZs) exhibit outstanding structural advantages as adsorbent materials.[37,38] As a type of microporous material consisting of merely silicon and oxygen atoms, PSZs are hydrophobic and without any acid site. Thus, the competitive adsorption of water, which contains high concentrations of cations, can be significantly reduced.[39] Moreover, PSZs are thermally stable and can be easily regenerated when their pores are blocked.[40–42] These unique features make PSZs potential sorbents for siloxane removal; however, to the best of our knowledge, no systematic investigation has been performed so far. Note that there are millions of possible PSZs,[43–45] screening these PSZs one by one as promising candidates for siloxane removal is not practical, if not impossible.

A paradigm shift is now underway, and machine learning (ML) offers us a powerful tool to solve such complex problems. Machine learning has been a kind of flourishing statistical methodology that has been widely used in interdisciplinary studies

[a]Department of Chemistry, University of Puerto Rico, Rio Piedras, San Juan, PR 00931, USA. E-mail: zhongfangchen@gmail.com

[b]Department of Mathematics, University of Puerto Rico, Rio Piedras, San Juan, PR 00931, USA

[c]Department of Chemical Engineering, University of Puerto Rico, Mayagüez Campus, Mayagüez, PR 00681, USA

† Electronic supplementary information (ESI) available: The number, average adsorption loading, and adsorption energy of 500 zeolites for $DMSO_2$, TMS, DMSD, and MMST; the scatter matrix of five features for adsorption of TMS, DMSD, and MMST; the adsorption energy, number, pore diameters (p), surface area (s), crystal parameters (a, b, and c) of top 10 adsorption performance zeolites for $DMSO_2$, TMS, DMSD, and MMST screened by GCMC simulations; the structures of 230 four-class-1 zeolites; the adsorption energies and loading of the second set of randomly chosen 10 zeolites from the predicted 230 four-class-1 zeolites; Github website link for the training data, prediction data and the well-trained models. See DOI: 10.1039/c9ta11909d

recently.[46–50,68] Different from rule-based systems that require much experience, time, and efforts, ML generates mathematical models[51,52] from experimental[53,54] and computational data[55–57] at speeds and scales that far exceed human capabilities. Moreover, those ML models can help recognize the neglected and potential connections and accelerate the ability to predict reactions and materials performances of unknown systems. Recently, ML techniques have been applied in materials discovery in environmental science and technology.[58–62] Among others, Lu *et al.* combined machine learning techniques and density functional theory calculations to predict undiscovered hybrid organic–inorganic perovskites for photovoltaics.[63] Chan *et al.* introduced a set of machine-learned coarse-grained (CG) models that successfully described the structure and thermodynamic anomalies of both water and ice at mesoscopic scales, all at two orders of magnitude cheaper computational cost than existing atomistic models.[64]

Here, we designed a two-step computational framework (Scheme 1) combining Grand Canonical Monte Carlo (GCMC) simulations and the machine learning method to investigate the adsorption performances of pure-silica zeolites. Four representative linear siloxanes and derivatives[65–67] namely trimethylsilanol (TMS), dimethylsilanediol (DMSD),[68,69] monomethylsilanetriol (MMST), and dimethylsulfone ($DMSO_2$)[67,70,71] were considered. We obtained essential features and screened out 230 preeminent zeolites from 50 959 hypothetical PSZs (picking ratio ≈ 0.0045) with excellent adsorption performance towards all these four PCs. Our best models achieved a test score of 0.91 for performance classification, and our further GCMC simulations verified that all 20 randomly chosen ML-recommended PSZs have excellent adsorption performance towards four problematic compounds. This work highlights the promise of combining data-driven modelling with traditional computations to predict the performance of complex zeolite systems.

Methods

Grand Canonical Monte Carlo (GCMC) simulations

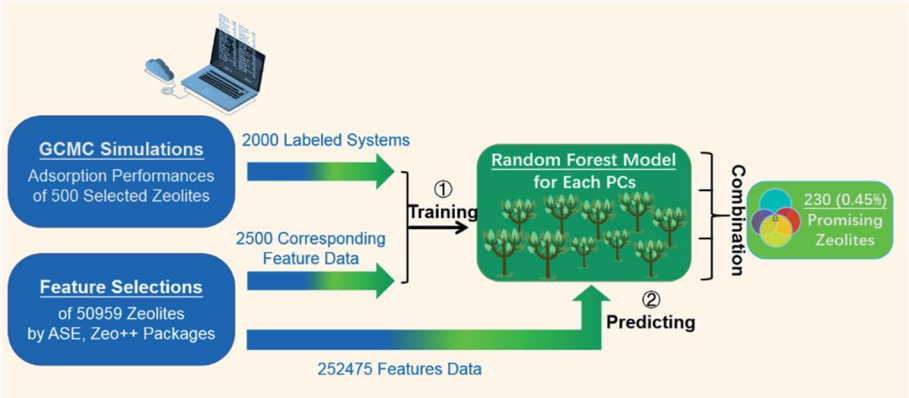
Grand Canonical Monte Carlo (GCMC) simulations in the sorption module of Materials Studio 8.0 were conducted to evaluate the absorption performance of 500 randomly selected zeolites towards each of the four PCs. GCMC is a statistical-mechanical approach, in which the adsorption process is investigated relying on random sampling and probabilistic interpretation in the sorbent framework. We calculated the average adsorption loading (mol^{-1}) and adsorption energy (kcal mol^{-1}) and identified 10 lowest-energy adsorption geometries of each adsorption system, where more substantial adsorption loading and higher adsorption energy indicate better adsorption performances. The GCMC simulations were carried out in a zeolite cell containing a section of the model pore, and each cell had a length of around 4.0 nm, which was sufficiently large to make the effect of finite system size negligible. The systems were equilibrated for 100 000 GCMC steps, and data were collected for another 1000 000 production steps to get the average amount adsorbed. All the GCMC simulations were carried out at a temperature of 298 K and a fixed pressure of 101.33 kPa with the Metropolis Monte Carlo method[72] and COMPASS forcefield.[73,74]

Feature selection

Five key features were used to train the models, which are three crystal parameters (a, b and c/Å), pore diameter (p/Å), and probe-accessible surface area (s, $Å^2$ per unit cell). Crystal parameters of 50 959 zeolites were obtained through the Atomic Simulation Environment (ASE) package,[75] and the pore diameters and probe-accessible surface areas were obtained using the Zeo^{++} package.[76] In Zeo^{++}, the number of sample points specified in the input is randomly displaced in the unit cells. Herein, the pore diameters are the largest inscribed spheres, which were obtained by setting the radius of the spherical probe to 0.[77] The probe-accessible surface area (per unit cell) was obtained by the Monte Carlo (MC) sampling approach, in which the radius of a probe (1.2 Å) was used.[77]

Machine learning models

Random Forest,[78,79] an integrated algorithm of decision trees[80] as implemented in scikit-learn software,[81] was used to train four models. Different super-parameters have been studied for optimal ML models, such as numbers of trees, max depth of



Scheme 1 The flow chart of two-step computational screening to achieve prominent zeolites for adsorbing four linear siloxanes and derivates.

branch and max feature for each branch. The parameters which help model achieve high training and test scores were retained. According to the test computations, the maximum depth for $DMSO_2$ was set as 9, while that for TMS, DMSD, and MMST was set as 10. The number of trees modeled was fixed as 250 for MMST, while for $DMSO_2$, TMS, and DMSD, the numbers of trees were all set as 200. Additionally, the number of maximum features for MMST was set as 4, while that for the other PCs was 3.

Results and discussion

GCMC simulations of 500 randomly selected pure-silica zeolites (PSZs)

As a big zeolite database, the hypothetical zeolite database has more than several hundred thousand zeolite structures with reasonable energy and framework density.[82,83] The family we chose (ABC-6 16-layered structures) has 50 959 pure-silica zeolites (PSZs), which have the same numbers of silicon and oxygen atoms, the same symmetry, and similar unit volume. Subsequently, the differences between these PSZs are defined in the atom positions, pore diameters, surface area, and crystal shape, so that the fundamental rules between zeolite structures and adsorption performances could be simplified, and the predictions of this family would be more reliable.

We first randomly chose 500 zeolites from 50 959 PSZs and computed the average adsorption loading (mol^{-1}) and adsorption energy (kcal mol^{-1}) by Grand Canonical Monte Carlo (GCMC) simulations (Table S1†). Correlating average adsorption loading (mol^{-1}) and adsorption energy (kcal mol^{-1}) of these 500 zeolites (Fig. 1), we found that the relationship between adsorption energy and adsorption loading for TMS is roughly linear, and that for MMST is most dispersive, while that for $DMSO_2$ and DMSD is in between (Fig. 1).

Noting that common sorbents suffer from low adsorption energies towards linear siloxanes, we decided to use the adsorption energy as the standard to classify PSZs in this work. Based on this standard, zeolites with adsorption energy in the top 20% are classified as class-1 (great zeolites, triangle points in Fig. 1), while the rest are classified as class-0 (bad zeolite, square points in Fig. 1). The borderlines of classification are 22.00, 19.20, 18.23, and 17.45 kcal mol^{-1} for $DMSO_2$, TMS, DMSD, and MMST, respectively. Such high adsorption energies also lead to high adsorption loading of these class-1 PSZs: more than 90% class-1 zeolites are in top 39%, 24%, 51%, and 66% loadings for $DMSO_2$, TMS, DMSD, and MMST, respectively. These results make our following screening approach valid for finding zeolites with not only high adsorption energies but also rather high adsorption loadings.

Furthermore, scrutinizing the lowest-energy adsorption geometries of each adsorption system (Fig. 2), we found that the PCs prefer different pores depending on their molecular sizes.

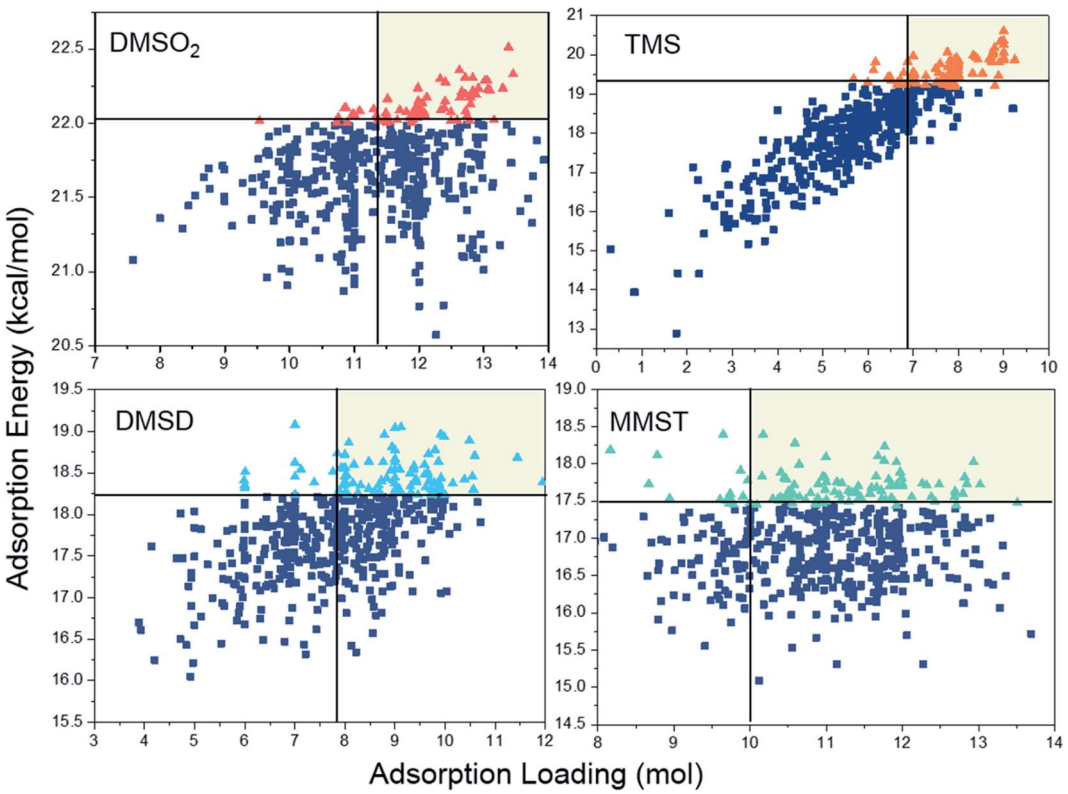


Fig. 1 The average adsorption energy and loading values of 500 PSZs towards four PCs, where triangles represent the top 20% zeolites (class-1) ranked by adsorption energy, while the square points are the other 80% zeolites (class-0), and the light yellow sections denote the top 90% class-1 zeolites ranked by adsorption loading.

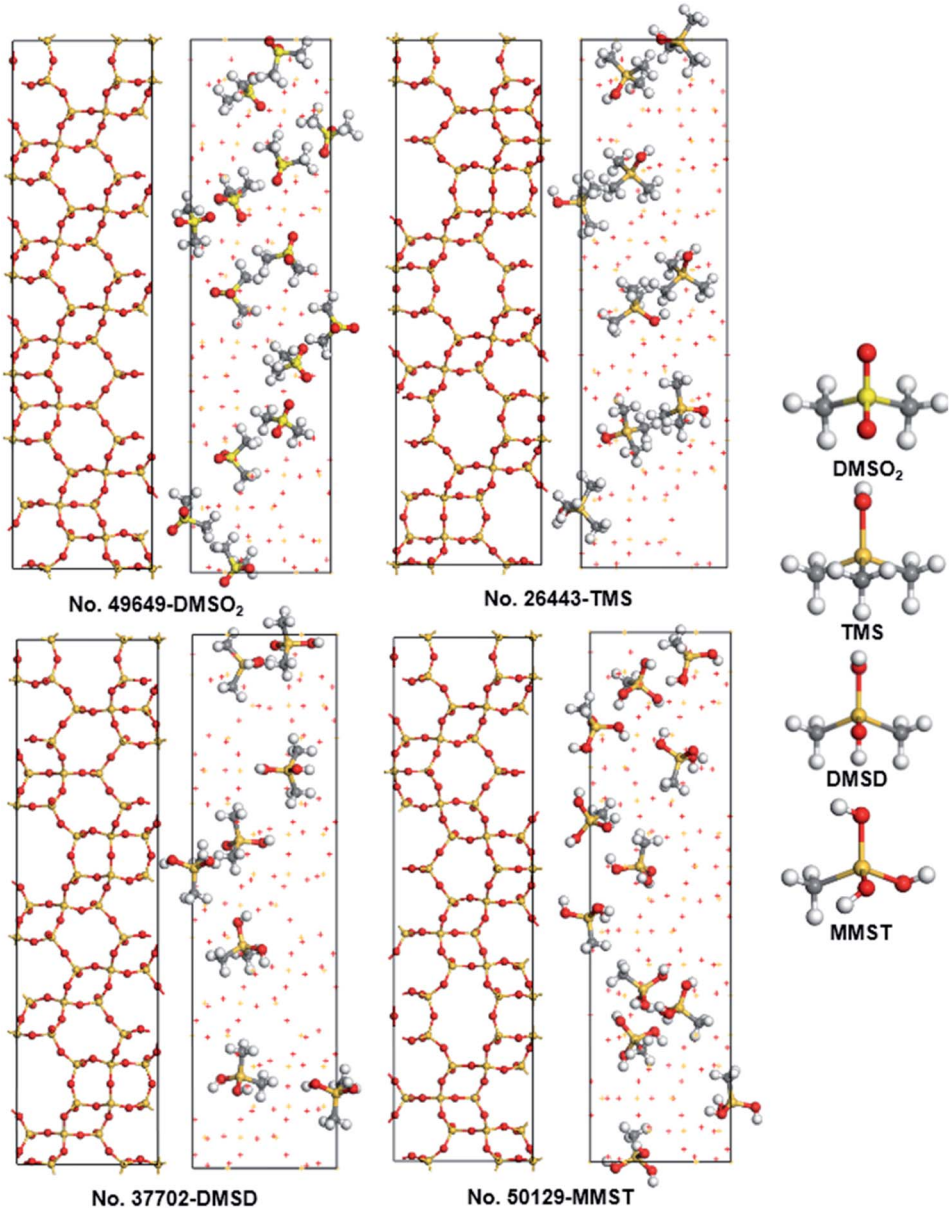


Fig. 2 The structures and the lowest-energy adsorption geometries of the best zeolite for each PC (left); the structures of the four PCs, $DMSO_2$, TMS, DMSD, and MMST (right). The yellow, red, light yellow and white balls represent silicon, oxygen, sulfur and hydrogen atoms, respectively. For clarity, in the adsorption geometries, the PCs are presented as stick-ball models, while silicon and oxygen atoms in zeolite frameworks are presented as yellow and red points.

The average diameters of $DMSO_2$, TMS, DMSD, and MMST are 4.00, 4.55, 4.34, and 4.30 Å, respectively. For example, the best zeolite for TMS, zeolite no. 26443 (the order of zeolites is indexed following the database), has larger pores than the best zeolites for other three zeolites, which is consistent with the fact that TMS possesses the most significant size among the four PCs under investigation.

For the top 10 zeolites (ranked by adsorption energies) (ESI, Table S2–S5†), the adsorption energies towards $DMSO_2$ are the largest with the best adsorption energy of 22.51–22.24 kcal mol^{-1}, followed by TMS (20.61–20.03 kcal mol^{-1}), DMSD (19.08–18.78 kcal mol^{-1}), and MMST (21.98–18.05 kcal mol^{-1}). Thus, $DMSO_2$ can be adsorbed more strongly than the other three PCs on these PSZs. Two main factors are responsible for the stronger interaction between $DMSO_2$ and these zeolites. The first is its relatively smaller size: the central S–C bonds in $DMSO_2$ are shorter than the corresponding bonds in the other PCs, and $DMSO_2$ has two methyl groups instead of three as in TMS. The smaller size makes $DMSO_2$ better fit the PSZ pores; the second is the more significant charge differences in $DMSO_2$: the electronegativity difference between S and C atoms is much pronounced than that between Si and C atoms in the other PCs, which enhances the electrostatic interaction between $DMSO_2$ and PSZs.

On the other hand, the adsorption energies of these 500 PSZs towards each PC cover a relatively big range. In detail, the differences between the highest and the lowest adsorption energies for TMS (7.72 kcal mol^{-1}) and MMST (6.90 kcal mol^{-1}) are around two times larger than those for $DMSO_2$ (4.21 kcal mol^{-1}) and DMSD (3.03 kcal mol^{-1}). Such adsorption energy differences strongly suggest that the structures of zeolite frameworks can significantly influence their adsorption performances, especially for TMS and MMST, and demonstrate the importance of finding suitable zeolite frameworks for the effective adsorption and the removal of problematic compounds.

Feature selection

The GCMC simulation results of 500 PSZs make it possible for us to select appropriate features to build the ML models, in which intrinsic features can adequately characterize the differences among PSZs without time-consuming computations. In this study, five features were selected, namely three crystal parameters (a, b and c/Å), pore diameter (p/Å), and probe-accessible surface area (s, $Å^2$ per unit cell). Pore diameters and probe-accessible surface areas can depict the size and the area of common adsorption locations, and crystal parameters (a, b and c) can provide additional information on the overall shape of zeolites.

We carefully checked the numerical values of these five features for the top 10 zeolites for each PC (see the ESI, Table S1–S4†) and found that each PC has its own optimal feature range. For the pore diameters, the ideal values are 5.80–6.00, 6.10–6.40, 5.95–6.25, and 5.90–6.10 Å, for $DMSO_2$, TMS, DMSD, and MMST, respectively. For the probe-accessible surface area, most of the top 10 zeolites for TMS and DMSD have surface areas over 800 $Å^2$ per unit cell, which are consistent with the general expectation that larger surface area leads to more pronounced adsorption loading and energy. However, for adsorption of $DMSO_2$ and MMST, some of the top 10 zeolites have very small accessible surface areas, which are as low as 0 $Å^2$ per unit cell (detected by Zeo^{++} package). This unexpected observation might be rationalized by the relative easiness for the "small" MMST and $DMSO_2$ molecules to insert into the narrow gaps of those zeolites, where are not considered as accessible surface areas according to the Zeo^{++} package. When the crystal parameters (a, b and c/Å) of zeolites are concerned, they can also influence the population and the shapes of pores, but the relationships between zeolite crystal parameters and the adsorption energies are not obvious; thus it is necessary to employ more complex machine learning models to describe such relationships.

Training Random Forest models

The GCMC simulations and feature selection for the 500 randomly selected PSZs as summarized above provided us with deep insights into the relationships between adsorption energies and intrinsic characteristics of PSZs, which serve as the basis for us to train ML models using the Random Forest (RF) algorithm.[78,79] As a widely used algorithm in real-world classification analysis, RF first randomly selects different features and training samples, generates many decision trees, and then averages the results of these decision trees to obtain the final classification. The Random Forest (RF) algorithm is a collection of decision trees (DT). Compared with DT, RF is more general, greatly improves the accuracy of models, and avoids the instability of DT. Moreover, RF can not only depict the underlying pattern of a complicated problem, but also provide feature importance for different features after training, which cannot be obtained by many other algorithms. Therefore, we employed the RF algorithm to train the models.

First, we examined the scatter matrix of the five features (Fig. S1–S4†) of the 500 PSZs investigated above and found that all these five features play significant roles in classifying PSZs, and thus it will be used for ML models.

As pretreatment of data, the balance of data size, division of training sets and test sets, and the data normalization of all features were all carried out. The data normalization of five features was performed for both training and test sets due to the large region of feature values. In order to avoid the influence of the big difference of the data size between class-1 and class-0 (1 : 4), we reproduced the minor class (class-1) two more times to obtain a relatively balanced ratio (3 : 4) of class-1 : class-0. Five-fold cross-validations have been employed to train and optimize the models, where the training set was used to build the model, and the degree of fitting for the test set would also feed back to the model. The input data were randomly split into an 80% training set (605 data) and a 20% test set (145 data). Only when a model achieves both a high test score and a high training score, we settle down the parameters and achieve the final optimal model.

During the training process of classification models, if the predicted labels for a sample match with the true set of labels, the accuracy is 1.0, or otherwise it is 0.0. Therefore, the quality of a model can be evaluated using the training score and test score, which is defined as the following equation:

$$\mathrm{Accurac}(y, \hat{y}) = \frac{1}{n_{\mathrm{samples}}} \sum_{i=0}^{n_{\mathrm{samples}}-1} 1(\hat{y}_i = y_i)$$

where $\hat{y}_i$ is the predicted value of the i-th sample, and y_i is the real value; n_{samples} is the number of samples.[81]

Based on the Random Forest algorithm, the ML models of the adsorption performance of PSZs towards $DMSO_2$, TMS, DMSD, and MMST gained excellent scores for both training and test sets. All the models gained high scores of 0.99 for the training sets, and in the meantime, relatively high scores of 0.91, 0.90, 0.91, and 0.89 were obtained for $DMSO_2$, TMS, DMSD, and MMST in the test sets, respectively (Table 1). The high training and test scores demonstrate that these Random Forest models with five selected features can well describe the effects of structural parameters of PSZs on the adsorption performances, and these models are expected to have outstanding predictive power to classify the adsorption performances of much more PSZs towards the four PCs under consideration.

Table 1 The training and test scores of four models of the adsorption performance of the 500 pure-silica zeolites with $DMSO_2$, TMS, DMSD, and MMST

PCs	$DMSO_2$	TMS	DMSD	MMST
Training score	0.99	0.99	0.99	0.99
Test score	0.91	0.90	0.91	0.89

To further check the performance of the above-obtained classification model, we examined the confusion matrices, which can allow clear visualization of the performance of the ML model (Fig. 3). In the matrix table, the data on the upper-left and downright diagonal represent the numbers of accurate predictions, while others are the wrong predictions. Among 145 test PSZs towards adsorbing $DMSO_2$, all the 63 class-1 PSZs were successfully predicted (class-1 error: 0.00); among the 75 class-0 PSZs, 61 were correctly classified, while 13 were wrongly assigned to class-1 (class-0 error: 0.18). A similar phenomenon occurs for the other PCs: the class-1 errors for TMS, DMSD, and MMST are 0.05, 0.09, and 0.09, respectively; and their corresponding class-0 errors are 0.15, 0.09, and 0.12, respectively. Encouragingly, the predictive accuracy for class-1 is higher than that for class-0, which guarantees that our ML models would not miss promising zeolites.

According to our above confusion matrix analyses, our ML model can well classify the PSZs; especially the prediction accuracy for class-1 PSZs (higher than 91%, and that for $DMSO_2$ even reached 100%) is even better than that for class-0. The main errors come from the prediction of class-0, which suggests that the possibility of missing class-1 is even smaller than what we expect from training and test scores.

Unlike a linear regression model, an ML model is hard to interpret directly. Fortunately, RF has the advantage of being able to provide feature importance of each feature and thus provide insights into how the parameters/features affect the properties of

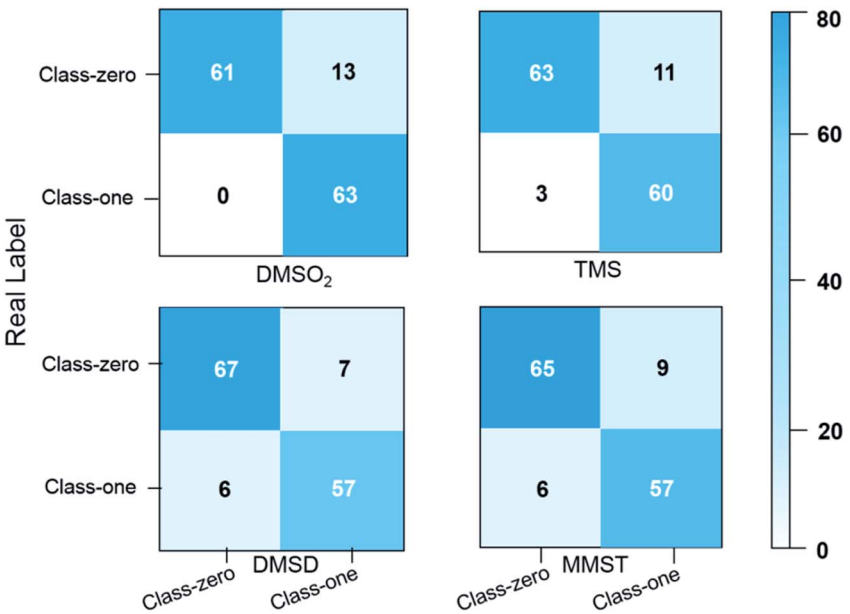


Fig. 3 Confusion matrices of the classification results of the test PSZs adsorbing $DMSO_2$, TMS, DMSD, and MMST.

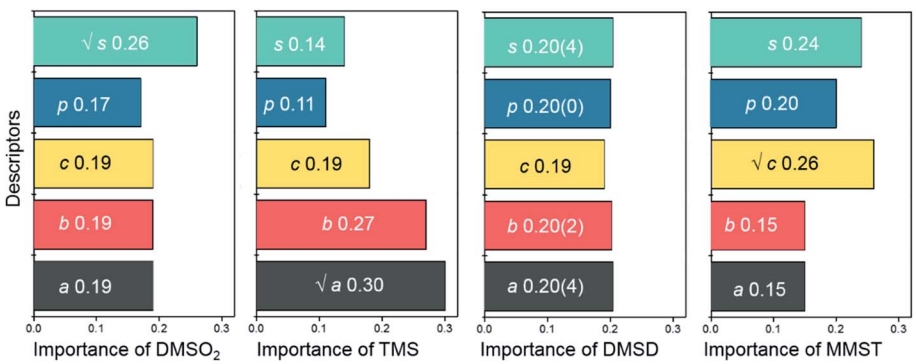


Fig. 4 The feature importance of five features (crystal parameters: *a*, *b*, and *c*; pore diameters: *p*, and surface areas: *s*) in the trained ML models towards adsorption of the four PCs ($DMSO_2$, TMS, DMSD, and MMST).

materials. The sum of feature importance for all features is 1.0; different values indicate different contributions of feature importance, and the feature with larger feature importance affects the output performance more. Thus, we analyzed the feature importance of these five features in the ML models.

Differences in feature importance do exist, as illustrated in Fig. 4. For $DMSO_2$ and MMST, the most critical feature of the 500 PSZs is the available surface area (s), with a feature importance of 0.26 and 0.24, respectively. For TMS, which has the largest size among the four PCs, the horizontal structural parameters, a and b, show more distinguished importance than others (0.30 and 0.27, respectively, Fig. 4). However, all five features of PSZs play nearly equal roles in DMSD adsorption, as indicated by their nearly equivalent values (around 0.2, Fig. 4). These feature importance analyses showed that for a PC with larger size (TMS in this case), the pores and the accessible surface area of zeolites are less important, while the crystal properties (a and b), which control the shape of zeolites, become more critical. The change of feature importance along with the size of the molecules is consistent with a previous study on the adsorption of polycyclic aromatic hydrocarbons on silica nanopores.[84]

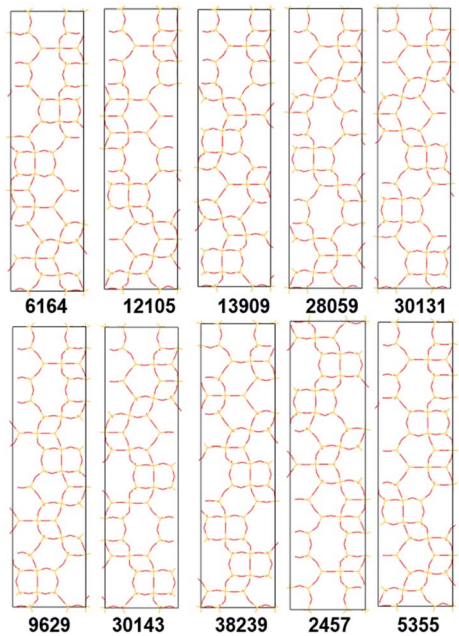


Fig. 5 The structures of the first set of the randomly selected 10 PSZs from the 230 four-class-1 zeolites.

Classifying 50459 PSZs by a trained ML model

The strongest motivation for developing ML models is to accelerate the materials discovery process. With the well-trained ML models for each PC based on GCMC simulations of 500 PSZs, we can quickly classify the adsorption performance of a much larger number of unexplored PSZs. Note that the remaining 50459 PSZs in the database (totally 50 959 PSZs) have similar compositions to the 500 training structures, which makes our ML models well suited to classify these PSZs.

The values of the five features (254795 in total) for these 50459 PSZs were also abstracted by ASE[75] and the Zeo^{++} package.[76] The classification of the adsorption performance towards each PC can be easily performed at super high efficiency, costing only several seconds in a desktop machine. Due to the differences in geometries and functional groups of the four PCs under consideration, the requirements of optimal zeolites for them are also different. Our ML model classified 10 347 (20.49%), 19 219 (38.06%), 15 437 (30.57%), and 12 550 (24.63%) as class-1 zeolites towards $DMSO_2$, TMS, DMSD, and MMST, respectively.

For practical applications, "omnipotent zeolites", which can strongly adsorb all four PCs studied in this work, are highly desired. Thus, we searched for PSZs with high adsorption energy and loading (class-1, ranked top 20%) towards all four PCs (named four-class-1 zeolites) and identified 230 four-class-1 zeolites (Fig. S5–S11†). These 230 "omnipotent zeolites" are merely 0.45% of the 50 959 PSZs examined in this study, and these powerful adsorbents make it possible to cost-effectively and collectively remove all four PCs. Note that the screening process was tremendously accelerated by ML methodology, and the

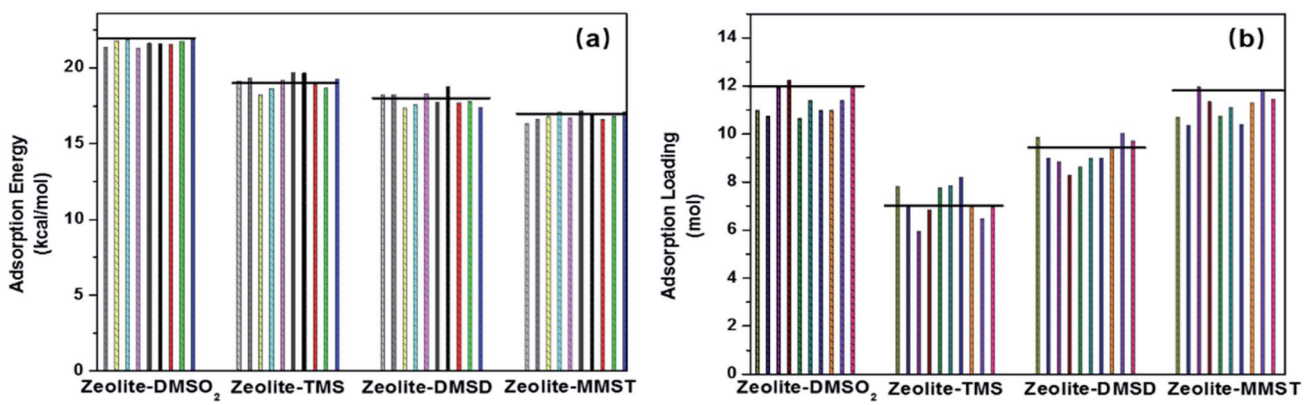


Fig. 6 (a) The adsorption energies and (b) adsorption loading of the first set of the randomly chosen 10 zeolites from the predicted 230 four-class-1 zeolites, where the upper lines correspond to the top 20% adsorption energies (a) and top 30% adsorption loading in the training data of 500.

candidate list of PSZs was dramatically reduced, which provides good guidance for future experimental and theoretical investigations on developing potent materials for siloxane removal.

To reconfirm the accuracy of ML models and the superior adsorption performance of the selected 230 PSZs, we randomly chose two sets (10 each, List 2 in the ESI†) of zeolites from these 230 four-class-1 zeolites and computed their average adsorption energies and adsorption loading towards four PCs by GCMC simulations. Fig. 5 and 6 present the structures and adsorption performance for the first set of 10 PSZs (the corresponding data for the second set are given in List S2 and Fig. S12 in the ESI†). Encouragingly, these 20 zeolites all have class-1 adsorption energies, which are comparable to or better than the corresponding value of the top 20% in the training data of 500 (Fig. 6a and S12a†). Moreover, all these 20 zeolites have average loading values higher than or close to the top 30% zeolites in the training data of 500 (Fig. 6b and S12b†).

Conclusion

Removal of siloxanes is a critical challenge due to their widespread, persistent, and toxic nature. Considering the weak interactions between siloxanes with typical sorbents, it is highly desirable to enhance the capacity of sorbents significantly. Pure-silica zeolites (PSZs) are promising sorbents for siloxane removal due to their unique characteristics, such as high thermal stability and hydrophobicity. In this work, by means of GCMC simulations and machine learning (ML) techniques, we captured essential structural features of PSZs and unveiled structure–property relationships by identifying the relative importance of pore diameters, surface areas, and crystal frameworks for different PCs. By screening the database of 50 959 PSZs using ML models, we discovered 230 promising zeolites with enhanced adsorption performance towards four important and representative linear siloxanes and derivates ($DMSO_2$, TMS, DMSD, and MMST). Our best models achieved 9.0% test error for classification of PSZs, and all selected PSZs (20/20) were verified to be excellent for all four problematic compounds by GCMC simulations.

Note that our process of screening sorbents by ML methodology can be extended to other sorbents such as Al-containing zeolites, metal-doped zeolites, and metal–organic frameworks; in addition, the contaminants can also be other organic/inorganic compounds which are difficult to be removed in nature. Therefore, once again, we vividly demonstrate the power of ML methodologies to accelerate materials discovery, which can greatly help conquer grand challenges facing the world today.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

This work was financially supported by NASA (Grant 80NSSC17M0047). A portion of the calculations used the resources of the Compute and Data Environment for Science (CADES) at ORNL and of the National Energy Research Scientific Computing Center, which are supported by the Office of Science of the U. S. DOE under contract no. DE-AC05-00OR22750 and DE-AC02-05CH11231, respectively. LR Pericchi is supported by awards #P20GM103475 from NIGMS and #U54CA096297 from NCI of the NIH.

References

1 M. Pedrouzo, F. Borrull, R. M. Marcé and E. Pocurull, *TrAC, Trends Anal. Chem.*, 2011, **30**, 749–760.
2 T. Dudzina, N. von Goetz, C. Bogdal, J. W. Biesterbos and K. Hungerbühler, *Environ. Int.*, 2014, **62**, 86–94.
3 V. T. L. Tran, P. Gélin, C. Ferronato, P. Mascunan, V. Rac, J.-M. Chovelon and G. Postole, *Chem. Eng. J.*, 2019, **371**, 821–832.
4 N. Ratola, S. Ramos, V. Homem, J. Silva, P. Jiménez-Guerrero, J. Amigo, L. Santos and A. Alves, *Environ. Sci. Pollut. Res.*, 2016, **23**, 3273–3284.
5 A. Divsalar, H. Divsalar, M. N. Dods, R. W. Prosser and T. T. Tsotsis, *Ind. Eng. Chem. Res.*, 2019, **58**(36), 16502–16515.
6 B. Tansel and S. C. Surita, *Waste Manag.*, 2019, **96**, 121–127.
7 C.-u. Bak, C.-J. Lim, J.-G. Lee, Y.-D. Kim and W.-S. Kim, *Sep. Purif. Technol.*, 2019, **209**, 542–549.
8 C. Rauert, T. Harner, J. K. Schuster, A. Eng, G. Fillmann, L. E. Castillo, O. Fentanes, M. n. Villa Ibarra, K. S. Miglioranza and I. Moreno Rivadeneira, *Environ. Sci. Technol.*, 2018, **52**, 7240–7249.
9 E. Santos-Clotas, A. Cabrera-Codony, B. Ruiz, E. Fuente and M. J. Martín, *Bioresour. Technol.*, 2019, **275**, 207–215.
10 A. S. Calbry-Muzyka, A. Gantenbein, J. Schneebeli, A. Frei, A. J. Knorpp, T. J. Schildhauer and S. M. Biollaz, *Chem. Eng. J.*, 2019, **360**, 577–590.
11 Y. Lu, T. Yuan, W. Wang and K. Kannan, *Environ. Pollut.*, 2011, **159**, 3522–3528.
12 Y. D. Lei, F. Wania and D. Mathers, *J. Chem. Eng. Data*, 2010, **55**, 5868–5873.
13 M. M. Coggon, B. C. McDonald, A. Vlasenko, P. R. Veres, F. o. Bernard, A. R. Koss, B. Yuan, J. B. Gilman, J. Peischl and K. C. Aikin, *Environ. Sci. Technol.*, 2018, **52**, 5610–5618.
14 L. Xu, S. Xu, L. Zhi, X. He, C. Zhang and Y. Cai, *Environ. Sci. Technol.*, 2017, **51**, 12337–12346.
15 L. Zhi, L. Xu, Y. Qu, C. Zhang, D. Cao and Y. Cai, *Environ. Sci. Technol.*, 2018, **52**, 12235–12243.
16 X. Wang, J. Schuster, K. C. Jones and P. Gong, *Atmos. Chem. Phys.*, 2018, **18**, 8745–8755.
17 I. S. Krogseth, X. Zhang, Y. D. Lei, F. Wania and K. Breivik, *Environ. Sci. Technol.*, 2013, **47**, 4463–4470.
18 J. Sanchís, A. Cabrerizo, C. Galbán-Malagón, D. Barceló, M. Farré and J. Dachs, *Environ. Sci. Technol.*, 2015, **49**, 4415–4424.
19 A. A. Bletsou, A. G. Asimakopoulos, A. S. Stasinakis, N. S. Thomaidis and K. Kannan, *Environ. Sci. Technol.*, 2013, **47**, 1824–1832.

20 C. Sparham, R. Van Egmond, S. O'Connor, C. Hastie, M. Whelan, R. Kanda and O. Franklin, *J. Chromatogr. A*, 2008, **1212**, 124–129.

21 W.-J. Hong, H. Jia, C. Liu, Z. Zhang, Y. Sun and Y.-F. Li, *Environ. Pollut.*, 2014, **191**, 175–181.

22 S. Genualdi, T. Harner, Y. Cheng, M. MacLeod, K. M. Hansen, R. van Egmond, M. Shoeib and S. C. Lee, *Environ. Sci. Technol.*, 2011, **45**, 3349–3354.

23 C. Sánchez-Brunete, E. Miguel, B. Albero and J. L. Tadeo, *J. Chromatogr. A*, 2010, **1217**, 7024–7030.

24 D.-G. Wang, W. Norwood, M. Alaee, J. D. Byer and S. Brimble, *Chemosphere*, 2013, **93**, 711–725.

25 J. Velicogna, E. Ritchie, J. Princz, M.-E. Lessard and R. Scroggins, *Chemosphere*, 2012, **87**, 77–83.

26 J. A. Willemsen, S. C. Myneni and I. C. Bourg, *J. Phys. Chem. C*, 2019, **123**(22), 13624–13636.

27 D. Shan, S. Deng, C. He, J. Li, H. Wang, C. Jiang, G. Yu and M. R. Wiesner, *Chem. Eng. J.*, 2018, **332**, 102–108.

28 Y.-H. Liu, Z.-Y. Meng, J.-Y. Wang, Y.-F. Dong and Z.-C. Ma, *Pet. Sci.*, 2019, 1–9.

29 L. Carter, J. Perry, M. J. Kayatin, M. Wilson, G. J. Gentry, E. Bowman, O. Monje, T. Rector and J. Steele, *45th International Conference on Environmental Systems*, 2015.

30 M. Ajhar, M. Travesset, S. Yüce and T. Melin, *Bioresour. Technol.*, 2010, **101**, 2913–2923.

31 D.-G. Wang, M. Aggarwal, T. Tait, S. Brimble, G. Pacepavicius, L. Kinsman, M. Theocharides, S. A. Smyth and M. Alaee, *Water Res.*, 2015, **72**, 209–217.

32 A. Cabrera-Codony, M. A. Montes-Morán, M. Sánchez-Polo, M. J. Martín and R. Gonzalez-Olmos, *Environ. Sci. Technol.*, 2014, **48**, 7187–7195.

33 D. R. Ortega and A. Subrenat, *Environ. Technol.*, 2009, **30**, 1073–1083.

34 X. Wei, Y. Wang, A. J. Hernández-Maldonado and Z. Chen, *Green Energy & Environment*, 2017, **2**, 363–369.

35 R. T. Yang, A. J. Hernández-Maldonado and F. H. Yang, *Science*, 2003, **301**, 79–81.

36 S. M. Rivera-Jiménez and A. J. Hernández-Maldonado, *Microporous Mesoporous Mater.*, 2008, **116**, 246–252.

37 S. Li, X. Wang, D. Beving, Z. Chen and Y. Yan, *J. Am. Chem. Soc.*, 2004, **126**, 4122–4123.

38 D. S. Wragg, R. E. Morris and A. W. Burton, *Chem. Mater.*, 2008, **20**, 1561–1570.

39 T. D. Pham, R. Xiong, S. I. Sandler and R. F. Lobo, *Microporous Mesoporous Mater.*, 2014, **185**, 157–166.

40 M. Palomino, A. Cantín, A. Corma, S. Leiva, F. Rey and S. Valencia, *Chem. Commun.*, 2007, 1233–1235.

41 D. H. Olson, X. Yang and M. A. Camblor, *J. Phys. Chem. B*, 2004, **108**, 11044–11048.

42 W. Zhu, F. Kapteijn, J. Moulijn, M. Den Exter and J. Jansen, *Langmuir*, 2000, **16**, 3322–3329.

43 M. Treacy, S. Rao and I. Rivin, *A combinatorial method for generating new zeolite frameworks*, Elsevier, 1993, pp. 381–388.

44 M. Foster and M. Treacy, *A Database of Hypothetical Zeolite Structures*, 2010, vol. 5, http://www.hypotheticalzeolites.net.

45 M. Treacy, I. Rivin, E. Balkovsky, K. Randall and M. Foster, *Microporous Mesoporous Mater.*, 2004, **74**, 121–132.

46 F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

47 S. Chmiela, H. E. Sauceda, K.-R. Müller and A. Tkatchenko, *Nat. Commun.*, 2018, **9**, 3887.

48 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, **2**, 16028.

49 H. Sahu, W. Rao, A. Troisi and H. Ma, *Adv. Energy Mater.*, 2018, **8**, 1801032.

50 W. Li, K. G. Field and D. Morgan, *npj Comput. Mater.*, 2018, **4**, 36.

51 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.

52 H. Trevor, T. Robert and F. JH, *Journal*, 2009, **77**, 482.

53 T. K. Patra, F. Zhang, D. S. Schulman, H. Chan, M. J. Cherukara, M. Terrones, S. Das, B. Narayanan and S. K. Sankaranarayanan, *ACS Nano*, 2018, **12**, 8006–8016.

54 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.

55 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.

56 X. Zhang, Z. Zhang, D. Wu, X. Zhang, X. Zhao and Z. Zhou, *Small Methods*, 2018, **2**, 1700359.

57 X. Zhang, Z. Zhang and Z. Zhou, *J. Energy Chem.*, 2018, **27**, 73–85.

58 Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duke and J. Brgoch, *Nat. Commun.*, 2018, **9**, 4377.

59 T. H. Miller, M. D. Gallidabino, J. I. MacRae, C. Hogstrand, N. R. Bury, L. P. Barron, J. R. Snape and S. F. Owen, *Machine Learning for Environmental Toxicology: A Call for Integration and Innovation*, ACS Publications, 2018, vol. 52, pp. 12953–12955.

60 Q. Xiao, H. H. Chang, G. Geng and Y. Liu, *Environ. Sci. Technol.*, 2018, **52**, 13260–13269.

61 Z. Ban, Q. Zhou, A. Sun, L. Mu and X. Hu, *Environ. Sci. Technol.*, 2018, **52**, 9666–9676.

62 T. Cordier, P. Esling, F. Lejzerowicz, J. Visco, A. Ouadahi, C. Martins, T. Cedhagen and J. Pawlowski, *Environ. Sci. Technol.*, 2017, **51**, 9118–9126.

63 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, *Nat. Commun.*, 2018, **9**, 3405.

64 H. Chan, M. J. Cherukara, B. Narayanan, T. D. Loeffler, C. Benmore, S. K. Gray and S. K. Sankaranarayanan, *Nat. Commun.*, 2019, **10**, 379.

65 D. L. Carter, B. Tobias and N. Y. Orozco, *Status of ISS Water Management and Recovery*, 2013.

66 T. Rector, C. Metselaar, B. Peyton, J. Steele, W. Michalek, E. Bowman, M. Wilson, D. Gazda and L. Carter, *An Evaluation of Technology to Remove Problematic Organic Compounds from the International Space Station Potable Water*, 2014.

67 L. Carter, J. Perry, M. J. Kayatin, M. Wilson, G. J. Gentry, E. Bowman, O. Monje, T. Rector and J. Steele, presented in part at the *45th International Conference on Environmental Systems*, Bellevue, Washington, 12–16 July 2015.

68 J. A. Rutz, J. R. Schultz, C. M. Kuo, M. Curtis, P. R. Jones, O. D. Sparkman and J. T. McCoy, *Discovery and Identification of Dimethylsilanediol as a Contaminant in ISS Potable Water*, 2011.
69 D. L. Carter, E. M. Bowman, M. E. Wilson and T. J. Rector, *43rd International Conference on Environmental Systems*, 2013, p. 3510.
70 J. A. Rutz, J. R. Schultz, C. M. Kuo, H. E. Cole, S. Manuel, M. Curtis, P. R. Jones, O. D. Sparkman and J. T. McCoy, presented in part at the *41st International Conference on Environmental Systems*, Portland, Oregon, 2011.
71 T. Rector, C. Metselaar, B. Peyton, J. Steele, W. Michalek, E. Bowman, M. Wilson, D. Gazda and L. Carter, presented in part at the *44th International Conference on Environmental Systems*, Tucson, Arizona, 13–17 July 2014.
72 W. K. Hastings, *Biometrika*, 1970, **1**, 97–109.
73 H. Sun, *J. Phys. Chem. B*, 1998, **102**, 7338–7364.
74 H. Sun, P. Ren and J. Fried, *Comput. Theor. Polym. Sci.*, 1998, **8**, 229–246.
75 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer and C. Hargus, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.
76 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
77 D. Ongari, P. G. Boyd, S. Barthel, M. Witman, M. Haranczyk and B. Smit, *Langmuir*, 2017, **33**, 14529–14538.
78 A. Liaw and M. Wiener, *R news*, 2002, vol. 2, pp. 18–22.
79 R. Díaz-Uriarte and S. A. De Andres, *BMC Bioinf.*, 2006, **7**, 3.
80 L. Rokach and O. Maimon, in *Data mining and knowledge discovery handbook*, Springer, 2009, pp. 149–174.
81 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
82 Y. Li and J. Yu, *Chem. Rev.*, 2014, **114**, 7268–7316.
83 J. Li, A. Corma and J. Yu, *Chem. Soc. Rev.*, 2015, **44**, 7112–7127.
84 H. Sui, L. Li, X. Zhu, D. Chen and G. Wu, *Chemosphere*, 2016, **144**, 1950–1959.