Science of the Total Environment 859 (2023) 160228

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Universal machine-learning algorithm for predicting adsorption performance of organic molecules based on limited data set: Importance of feature description

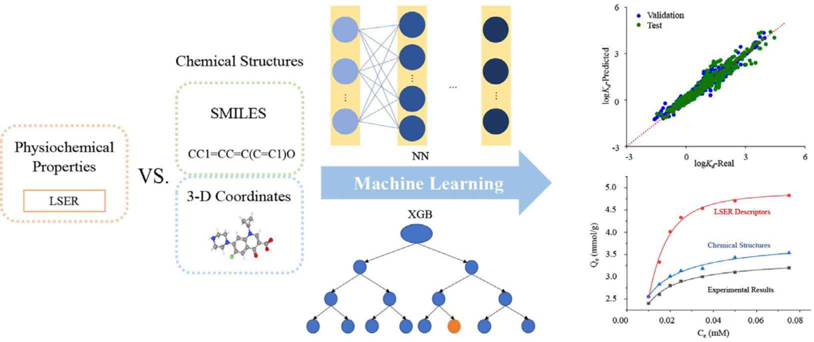
Chaoyi Huang [a], Wenyang Gao [b], Yingdie Zheng [a], Wei Wang [a], Yue Zhang [b], Kai Liu [a,*]

[a] *Division of Environment and Resources, College of Engineering, Westlake University, Hangzhou, Zhejiang 310024, China*
[b] *Division of Artificial Intelligence and Data Science, College of Engineering, Westlake University, Hangzhou, Zhejiang 310024, China*

HIGHLIGHTS

- Structural descriptors produce better adsorption isotherm prediction accuracy compared with LSER descriptors.
- XGB is better than NN-based models in predicting adsorption isotherm of organic molecules.
- Structure of organic molecules are the most important feature for constructing XGB models.
- 3D coordinate is more accurate in predicting adsorption isotherm of similar organic molecules.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jay Gan

Keywords:
Water treatment
Organic contaminants
Artificial intelligence
Adsorption isotherm

ABSTRACT

Adsorption of organic molecules from aqueous solution offers a simple and effective method for their removal. Recently, there have been several attempts to apply machine learning (ML) for this problem. To this end, polyparameter linear free energy relationships (pp-LFERs) were employed, and poor prediction results were observed outside model applicability domain of pp-LFERs. In this study, we improved the applicability of ML methods by adopting a chemical-structure (CS) based approach. We used the prediction of adsorption of organic molecules on carbon-based adsorbents as an example. Our results show that this approach can fully differentiate the structural differences between any organic molecules, while providing significant information that is relevant to their interaction with the adsorbents. We compared two CS feature descriptors: 3D-coordination and simplified *molecular*-input line-entry system (*SMILES*). *We then built CS-ML models based on* neural networks (NN) and extreme *gradient boosting (XGB). They all outperformed* pp-LFERs based models and are capable to accurately predict adsorption isotherm of isomers with similar physiochemical properties such as chiral molecules, even though they are trained with achiral molecules and racemates. We found for predicting adsorption isotherm, XGB shows better performance than NN, and 3D-coordinations allow effective differentiation between organic molecules.

1. Introduction

Today, carbon-based adsorbents remain one of the most simple and versatile adsorbents for the removal of aqueous organic contaminants (Pai and Wang, 2022; Rojas and Horcajada, 2020; Van Duck and van de Voorde, 1984). Application wise, batch experiments are usually required to determine

* Corresponding author.
E-mail address: liukai@westlake.edu.cn (K. Liu).

http://dx.doi.org/10.1016/j.scitotenv.2022.160228
Received 19 September 2022; Received in revised form 9 November 2022; Accepted 12 November 2022
Available online 17 November 2022

the optimum dosage of adsorbents, which can be tedious and inefficient (Luo et al., 2022; Zhang et al., 2020; Zhu et al., 2022). Besides, such effort can be prohibitive for synthesized organic compounds such as chiral chemicals. Therefore, the construction of a predictive model of adsorption property is highly desired.

Traditionally, quantitative structure-property relationships (QSPRs) were used to correlate adsorption property of organic molecules with its representative descriptor (Apul et al., 2013; Dickenson and Drewes, 2010; Kennicutt et al., 2016). To that end, multi-linear regression (MLR) models were typically developed in conjunction with polyparameter linear free energy relationships (pp-LFERs) (Su et al., 2022; Xu et al., 2021); more specifically, linear solvation energy relationship (LSER) descriptors were used (Yu et al., 2015). However, several obstacles significantly reduce the accuracy of QSPRs, including the limited availability of experimentally derived LSER descriptors. Since the accuracy of predicted LSER descriptors is substantially lower than that of experimentally derived ones (Ulrich et al., 2017), new prediction methods independent to LSER descriptors need to be developed.

Due to the rapid development of artificial intelligence, predictive models based on machine learning (ML) have already been reported for the optimization of various remediation technologies such as disinfection and filtration (Lowe et al., 2022). Since only a few operation parameters such as flow rate and dosage are involved, basic ML models such as Artificial Neural Network (ANN) and Random Forest (RF) were directly employed with good prediction result (Cordero et al., 2021; Li et al., 2017). Usually, these models were trained using data collected from a certain wastewater treatment plant (WWTPs) and thus cannot be applied to other WWTPs due to the lack of data diversity. Several studies have also reported the application of ML in predicting the adsorption performance of organic contaminants, which is substantially more complicated than the former applications due to the diversity in the physiochemical properties of organic molecules. Existing ML based adsorption prediction heavily relies on the physiochemical properties of the adsorbates, which are conveniently described by pp-LFERs (Qi et al., 2020; Zhao et al., 2022). However, for organic molecules with subtle differences such as chiral compounds, the experimental value of LSER descriptors is scare. In addition, structure of organic molecule reported in adsorption study is extremely diverse. To solve that, data preprocessing such as selecting similar data for ML training has been reported (Zhang et al., 2020). Unfortunately, such strategy suffers from the lack of data diversity, and cannot be applied to a broad range of organic molecules.

In this study, we proposed an improved ML strategy for predicting adsorption performance of organic molecules, by incorporating feature descriptors based on their chemical structures as well as adsorbents. This feature is clearly different from the previous ML methods, by which pp-LFERs were used as organic molecule descriptors. We have compared the prediction performance obtained from a combination of different structural descriptors and ML algorithms. Thus, diversities in ML along with their effects in prediction performance have been examined in detail. This information should be valuable for the future application of ML to solve environmental problems with limited but diverse experimental data.

2. Materials and methods

Chemicals involved in the current study are summarized in Table S1. Our objective is to develop a universal ML algorithm based on features to describe organic molecules as well as adsorbents. To this end, we have employed two systems to describe chemical structures: 1) 3-D chemical structures of organic molecules (CS); 2) isomeric simplified *molecular*-input line-entry system (*SMILES*) (Weininger, 1988). For the former, Cartesian coordination of all atoms from each organic molecules were collected from PubChem Database. These structures were calculated based on MMFF94s force field, and represent energetically accessible and biologically relevant conformations, rather than energy-minimal forms (Halgren, 1999; Kim et al., 2013). For the later, isomeric *SMILES was chosen since it* allows specifying isotopism and stereochemistry of organic molecules. A list of SMILES and the structure of each organic molecules used in our study is summarized in Table S2 and Fig. S1.

2.1. Data set construction

Large-scale database is necessary to develop and validate ML algorithm. Unfortunately for environmental applications, such database is usually scarce due to the limited availability of experimental data, even for a well-studied subject such as the adsorption of organic molecules using carbon-based adsorbents. In both traditional polyparameter linear free energy relationships (pp-LFER) method and recently published NN-LFER method (Zhang et al., 2020), the adsorption capacity at certain equilibrium concentration could be calculated by the LSER descriptors (E, S, A, B, and V) according to the following equation (Zhang et al., 2020):

$$\log K_d = eE + sS + aA + bB + vV + c \tag{1}$$

where the adsorption coefficient log K_d represents the extent of adsorption, e, s, a, b and v are fitting parameters, and c is the constant form.

After comparing data sets used in relevant studies, we decided to adopt an adsorption data set reported by a recent study (Zhang et al., 2020), denoted as the Original Data Set, due to the relatively large number of organic molecules with diverse structure that were included. It includes 4102 data points associated with 586 isotherms and 165 organic molecules collected from literatures on biochars, CNTs, GACs, and polymeric resins. Each data point from this data set includes 5 Abraham descriptors for organic molecules (E, S, A, B, and V), 2 descriptors for adsorbents including surface area (BET) and total pore volume (V_t), log C_e (aqueous concentrations at the adsorption equilibrium), and log K_d (adsorption coefficient). 7 data points were used to describe each adsorption isotherm. It also has to be noted that since the original data set did not specify adsorbate and adsorbent for each data point, to make use of this data set, we have determined the identity of adsorbate by matching each data point to organic molecules of known Abraham descriptors. Unfortunately, 203 data points cannot be identified due to unmatching, and were thus removed. Furthermore, since the major adsorption mechanisms of organic molecules on polymeric resins are different to that of carbon-based adsorbents (Caetano et al., 2009; Zhao et al., 2019), we have decided to omit data for resin adsorption. These lead to a smaller data set consisting of 130 organic molecules, which is denoted as the Master Data Set in LSER Descriptors Data Sets. Since more than one adsorption isotherm for each organic molecule have been reported, the Master Data Set consists of 924, 1288, 714 data points for biochar, carbon nanotubes (CNTs), and granular activated carbons (GACs), respectively. In addition, we have expanded this data set by incorporating data from recently published studies to form the Expanded Master Data Set in LSER Descriptors Data Sets. We have followed the data collection rule reported for the construction of the Original Data Set (Zhang et al., 2020). In addition, since experimentally derived LSER descriptors are more accurate than predicted ones, for added data, we have only chosen organic molecules without conflicting experimental LSER descriptors. The Expanded Master Data Set involves a total of 135 organic molecules (Table S1), or 966, 1323, 833 data points for biochar, CNTs, and GACs, respectively.

2.2. Data preprocessing

We have described the origin of descriptors that were used to describe structures of organic molecules. Briefly, Cartesian coordination of all atoms from each organic molecules, which is calculated by MMFF94s force field, were collected from PubChem Database. Due to the large number of different atoms included in the organic molecules listed in the Structural Descriptors Data Set, an array of 210 Cartesian coordinates is necessary to describe possible positions of all atoms from the 135 organic molecules. Since most organic molecules in this data set contains lesser atoms than the maximum number of atoms possible, this results in many empty coordinates, especially for C and H atoms, which were filled with zeros. However, for such small data set, high-dimensional data may lead to over-fitting and poor prediction performance. Therefore, dimension reduction was performed using principal component analysis (PCA). A number of new features were generated through PCA-CS in order to capture the variability in the data with fewer features (Fig. S2). We combined these new features with adsorbent properties such as BET surface area and total pore volume (V_t), along with adsorption equilibrium concentration (C_e) as input features for the ML modeling.

Alternatively, isomeric SMILES was used to represent the structure of organic molecules involved in this study (Table S2). Since it is a linear representation, it needs to be preprocessed to provide information regarding atoms and their chemical environment that are necessary to build ML models using GNN (graph neuron network), which is based on learning representations of fingerprints (or r-radius subgraphs) (Tsubaki et al., 2018). As the first step to create molecular vectors, isomeric SMILES was transformed into uniformed 2-D molecular graph (Fig. S3). Meanwhile, atom types and bond information were extracted from SMILES through “RDKit” package in Python (Tsubaki et al., 2018). Then, GNN encodes 2-D molecular graphs to form features as nodes and edges, corresponding to atomic features and bond features, respectively. Since each organic molecule contains a large number of atoms, and not all atoms participate in adsorption, existing nodes and edges were grouped to form fingerprints as new nodes based on the chemical structure. These fingerprints were built by setting central molecule with certain radius. Subsequently, new edges would describe bond features between these fingerprints. Finally, fingerprints were non-linearly transformed into atom vectors by GNN. These atom vectors were combined into a single molecular vector to represent organic molecule's structure.

To increase the data diversity and generalizability of ML models, the preprocessed data were first grouped before been randomly shuffled, to ensure the element of randomness for the construction of representative models. Typically, all experimental data points extracted from the same adsorption isotherm were grouped together, since they shared the same adsorbate and absorbent, and the only variable was C_e.

Before training ML models, data splitting is necessary. In general, the split ratio of training, validation, and test set is 8:1:1. Considering the relatively small data set used in this study, larger validation and test set may avoid fortuity and reduce error. Therefore, we have split the preprocessed data into training, validation, and test set with the ratio of 7:1.5:1.5. This is also consistent with reported practice using similar data set.(Zhang et al., 2020).

2.3. Modeling methods

Gradient Boosting Decision Tree (GBDT) and Neural Network (NN) are two of the most widely employed machine learning methods, from which six ML algorithms including NN-LFER, NN-PCA-CS, NN-GNN-SMILES, XGB-LFER, XGB-PCA-CS, and XGB-GNN-SMILES were proposed and evaluated for the prediction performance of adsorption isotherm of organic chemicals on carbon-based adsorbents (Fig. 1).

Neuron network (NN) constructed by three layers (input, hidden, and output layers) was applied to different feature description methods including LFER, PCA-CS, and GNN-SMILES. In general, NN was optimized though the adjustment of hidden layers and activation functions. For NN-LFER, there is no need to preprocess the data since only eight features (LFER descriptors, BET, V_t, and log C_e) were involved. Similarly, for NN-PCA-CS model, new features obtained from data preprocessing were used in the input layer. For all NN models, the number of hidden layers was set to 4, while the output layer was composed of log K_d values. Besides, rectified linear units (ReLU), which is a piecewise linear function, was selected as activation function between hidden layers, and Sigmoid, an alias for the logistic function, was applied before output layers.

To prevent overfitting for such small data set, *k*-fold cross-validation (CV) was applied during the model construction, it's known to improve model performance and generalization. We chose to use 5 as *k* value, since this produces validation and test sets with similar size. In each cycle of the 5-fold cross-validation, 70 % of the data were selected to train and build the NN model, while the rests were applied for validation and test. Since input values varied with large range, they were normalized to the same scale between −1 and 1. As the result, the output data of log K_d has to be denormalized in order to compare with experimental values during validation or test process.

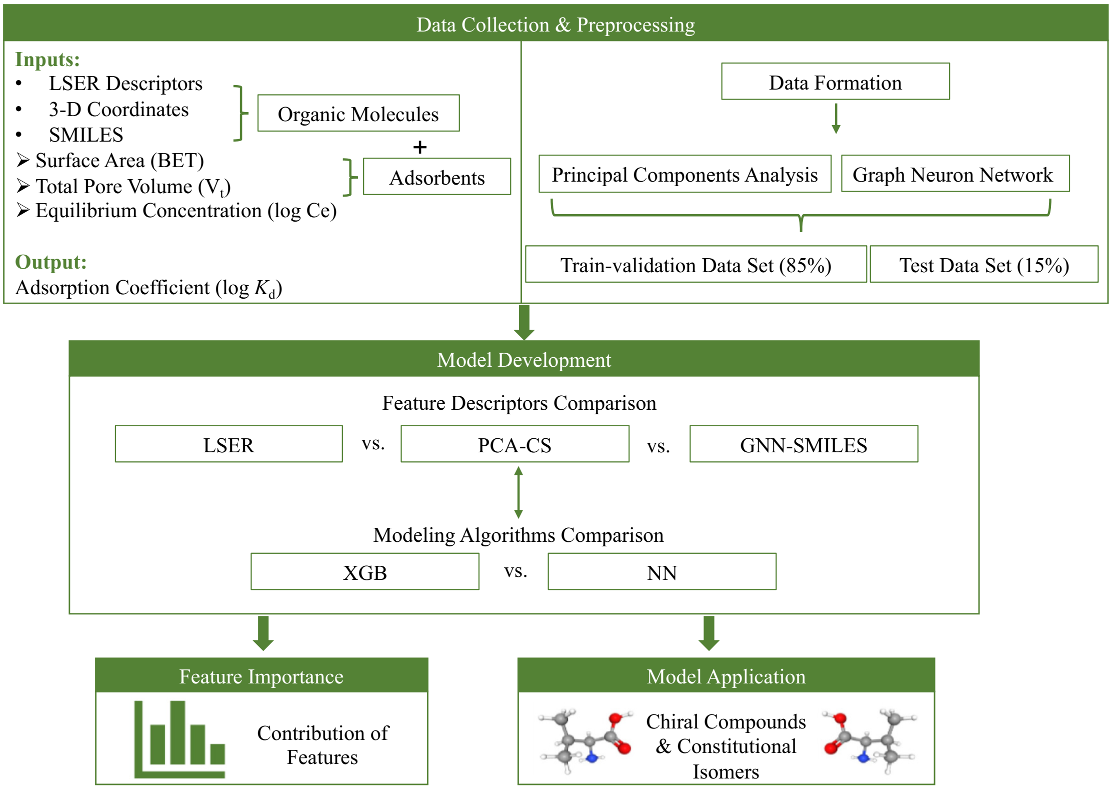


Fig. 1. Flowchart detailing the application of ML in predicting the adsorption of organic molecules on carbon-based adsorbents.

Since different algorithms impact the ML model prediction, extreme *gradient boosting* (XGB) was compared with NN. XGB is a scalable machine learning system for gradient tree boosting (Chen and Guestrin, 2016). It has been reported that XGB is more proficient in solving tabular and small-scale data sets by building tree structures (Ching et al., 2022; Liu et al., 2022; Sagi and Rokach, 2021). The basic components of XGB include small trees, root nodes, numerical weight, internal nodes, and leaf nodes. It can be mathematically expressed as the following (Chen and Guestrin, 2016):

$$\hat{y}_i^{(t)} = \Sigma_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

where $\hat{y}_i$ is the predicted value with respect to input x_i; t is the total number of regression trees being used; and f_k represents the predicted value of each independent regression tree. Different from NN models, XGB can optimize the model by adjusting the number of trees, learning rate, and tree depth. Besides, early stopping, an optimization technique, was employed to reduce overfitting without compromising model accuracy. We used feature importance analysis to visualize contribution of features to the model performance.

2.4. Error metrics

The model performance of all algorithms (NN-LFER, NN-PCA-CS, NN-GNN-SMILES, XGB-LFER, XGB-PCA-CS, and XGB-GNN-SMILES) was evaluated in terms of correlation coefficient (R^2), Q_{F2}^2value (referred as Q^2), and the root mean square error (RMSE), as described in Eqs. (3), (4) and (5), respectively.

$$R^2 = 1 - \frac{\Sigma_1^n (y_i - y)^2}{\Sigma_1^n (y_i - \bar{y})^2} \tag{3}$$

$$Q_{F2}^2 = 1 - \frac{\Sigma_1^n (y - y_i)^2}{\Sigma_1^n (y - \bar{y})^2} \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{\Sigma_1^n (y_i - y)^2}{n}} \tag{5}$$

where y_i, y are the predicted and real values of the target labels, respectively; $\bar{y}$ is the mean value of the real target labels; and n is the total number of data points. To allow comparison with results from previous study, both R^2 and Q^2 were calculated for all models. Generally, higher R^2/Q^2 accompanied with lower RMSE suggest better prediction performance of the model.

2.5. Model selection

The performance of ML algorithms is typically evaluated based on the prediction results of their respective models using test set (Yuan et al., 2021; Zhu et al., 2022). For a large data set with high diversity, the prediction performance of test set is consistent, so each model is representative for the ML algorithm which it was derived. However, for many environmental problems, such as prediction of adsorption isotherm, the data set is of high diversity but relatively small size. As the result, Q^2/R^2 as well as RMSE of the test set may fluctuate within a wide range, rendering difficulty in choosing the best model that is representative for the ML algorithm. So, standard to choose representative model needs to be imposed in order to properly compare the performance of ML algorithms for environmental problems with diverse but limited data.

To this end, we proposed the Medium-Selection Cross-Validation Method, which is an improved version of cross-validation method. It is specifically designed for small data set to avoid fortuity during ML training while ensuring statistical significance. Specifically, we first perform cross-validation method to generate 5 ML models, and the ML model with the best prediction performance was chosen. Subsequently, we repeat this process for 8 more times, and summarized the result (Table S3 shows an example). Then, ML model with median value of R^2/Q^2 for test set is selected as the representative model. It represents a comprehensive assessment of model performance which is trained with small and diverse data set.

3. Results and discussion

3.1. Effect of data size on prediction accuracy

For NN-LFER, two data sets with different sizes in LSER Descriptors Data Sets including Master Data Set and Expanded Master Data Set were first compared to distinguish whether data expansion could improve the model performance. The joint scattered plots including experimental and predicted values of log K_d are shown in Figs. 2 and 3. The results are summarized in Table S4. As expected, the goodness-of-fit, which is represented by Q^2, substantially increased when the larger data set was used. For example, Q^2 were 0.64, 0.72, 0.59, and 0.74 for biochar, CNTs, GACs, and all adsorbents using Master Data Set, while Q^2 increased to 0.71, 0.80, 0.70, and 0.84 separately after Expanded Master Data Set was used. In addition, better prediction performance was also achieved, which is evident from the reduction in RMSE, except for GACs. This result suggests the prediction accuracy of NN-LFER can be improved with additional experimental data, which is unfortunately very tedious to obtain. Of course, the inaccuracy in predicted LSERs values used in the model training also aids the deviation between predicted and experimental adsorption isotherms.

3.2. Effect of feature descriptors on prediction accuracy

We evaluated the data size effect on prediction accuracy using NN-LFER as model algorithm, log K_d at given C_e was used as the label, our results show limited improvement can be achieved when a larger data set was used (Figs. 2 and 3). So, we emphasized on the feature descriptor effect on prediction accuracy. Three NN based ML models, namely NN-LFER, NN-PCA-CS, and NN-GNN-SMILES were evaluated for the prediction of adsorption isotherm of organic molecules. The results are summarized in Tables S4 and S5. The latter two models showed better prediction results compared to NN-LFER using same data set (Figs. 4, S4 and S5). As we can see, NN-PCA-CS outperforms the other models, with Q^2 reached to 0.89, 0.91, 0.89 and 0.90, for biochar, CNT, GAC, and all carbon-based adsorbents, respectively, and RMSE reduced to 0.25, 0.32, 0.32, and 0.34 respectively for four adsorbents. The prediction performance of NN-PCA-CS is followed by NN-GNN-SMILES. Slightly reduced Q^2 values (0.82, 0.83, 0.80 and 0.85 for biochar, CNT, GAC, and all adsorbents, respectively) were observed, along with increased RMSE values, ranging from 0.32 to 0.35 for all adsorbent types.

The possible reason of reduced prediction accuracy for NN-GNN-SMILES might be the incomplete expression of molecular features, which results from information extraction via GNN preprocessing. For example, benzene ring from organic molecules might be separated into different fingerprints, which cannot express the structure of benzene appropriately. In addition, GNN ignored the bonds inside fingerprints, which omits important features that maybe involved in the adsorbent-adsorbate interaction. Albeit, both NN-PCA-CS and NN-GNN-SMILES models outperform NN-LFER model in terms of prediction accuracy.

We analyzed the LSER descriptors involved in this study, and found they are a mixture of experimental and predicted values. Among which conflicting values of experimental LSERs descriptors can be found for 82 molecules (Fig. S6, Table S6), and 12 molecules' LSER descriptors have never been experimentally determined (Fig. S6). Using QSPR prediction tool provided by the UFZ-LSER database, the predicted LSER descriptors for these 12 molecules all fell outside the application domain (Ulrich et al., 2017). It has to be noted that these LSER descriptors were adopted from the reported data set without any modification, in order to allow performance comparison (Zhang et al., 2020), therefore the selection of these LSER descriptors is beyond the scope of this paper. But nevertheless, it shows the proportion of reliable LSER descriptors that were employed in the NN-LFER model

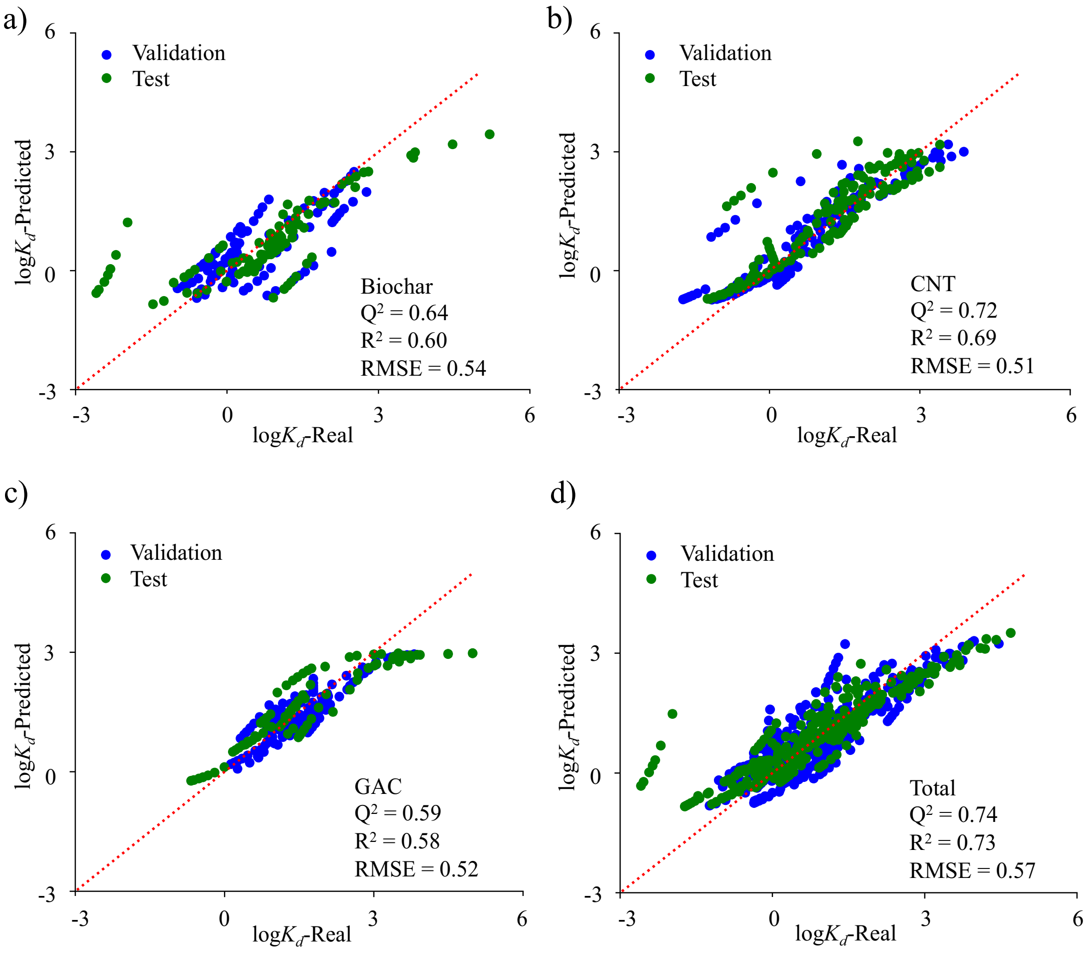


Fig. 2. Prediction performance of NN-LFER models for a) biochar, b) CNT, c) GAC, and d) all carbon-based adsorbents using the Master Data Set from LSER Descriptors Data Sets with total sample size of 2926.

development is as low as 30.37 % (Fig. S6). Therefore, poor prediction accuracy can be expected using this approach. In a word, the selection of proper feature descriptors, such as structural descriptors vs. LSER descriptors, can significantly impact the model performance with limited experimental data.

Overfitting is a modeling error in statistics, it occurs when the model properly learns the data during the training process, but the test performance is poor (Dietterich, 1995; Ying, 2019). NN models are prone to overfitting when large number of features are used in the input layer. To reduce the possibility of overfitting, we employed the strategy of *k*-fold CV adjustment. Rationales are provided in Materials and Methods Section. The possibility of overfitting can be visualized using the training-validation loss curves. Typically, during the training process, the loss value of both training and validation data set will decrease and tend to be stable. Overfitting usually occurs when the training or validation loss suddenly increased as the model continues to learn the data. As evident from our result (Figs. S7, S8 and S9), the 5-fold CV strategy is successful to prevent over-fitting in all models including NN-LFER, NN-PCA-CS, and NN-GNN-SMILES.

3.3. Effect of machine learning algorithms on prediction accuracy

We have also investigated the effects of machine learning algorithms on prediction accuracy. To this end, we have compared XGB with NN. Tree-based algorithms have been widely applied in solving environmental problems with good performance (Ching et al., 2022; Wang et al., 2022). Among tree-based algorithms, XGB has the advantages to avoid overfitting and optimize the models (Chen and Guestrin, 2016). Besides, early stopping can be easily employed to prevent overfitting. We evaluated the application of XGB algorithm using different descriptor approach, namely LFER, PCA-CS, and GNN-SMILES. It has to be noted that the inputs of XGB models were identical to that of NN models discussed earlier, in order to allow performance comparison. After optimization of XGB-based models, we found the prediction performance of these models were superior to that of NN-based models, evidenced by the enhanced test R^2 and reduced RMSE (Tables S4 and S5). Our results show a clear advantage of XGB over NN-based algorithms (Figs. 4, 5, S10 and S11). It is notable that for prediction of all carbon-based adsorbents, both structural descriptors based XGB models exhibited similar prediction performance, and RMSE of 0.34 and 0.32, Q^2 of 0.93 and 0.94 were observed for XGB-PCA-CS and XGB-GNN-SMILES, respectively. This may be due to the fact that only 135 different organic molecules were involved in the current study, and both PCA-CS and GNN-SMILES can effectively describe these molecules. Nevertheless, XGB-PCA-CS and XGB-GNN-SMILES showed significant improvement in prediction accuracy compared to XGB-LFER. Our result is also significantly improved compared with published results (Table S7), although they are primarily based on LSERs.

Feature importance analysis has been frequently applied to XGB models to understand the distribution of each feature in model construction (Asare et al., 2021; Yuan et al., 2021). Such information can help us to further improve the ML algorithm by selecting features with high contribution. For XGB models employed in the current study, several factors related to adsorption isotherm were described by features, including structures and concentration of organic molecules, and physical properties of carbon-based adsorbents. Some features such as BET area, pore volume of adsorbents, and concentration of organic molecules are straightforward to describe.

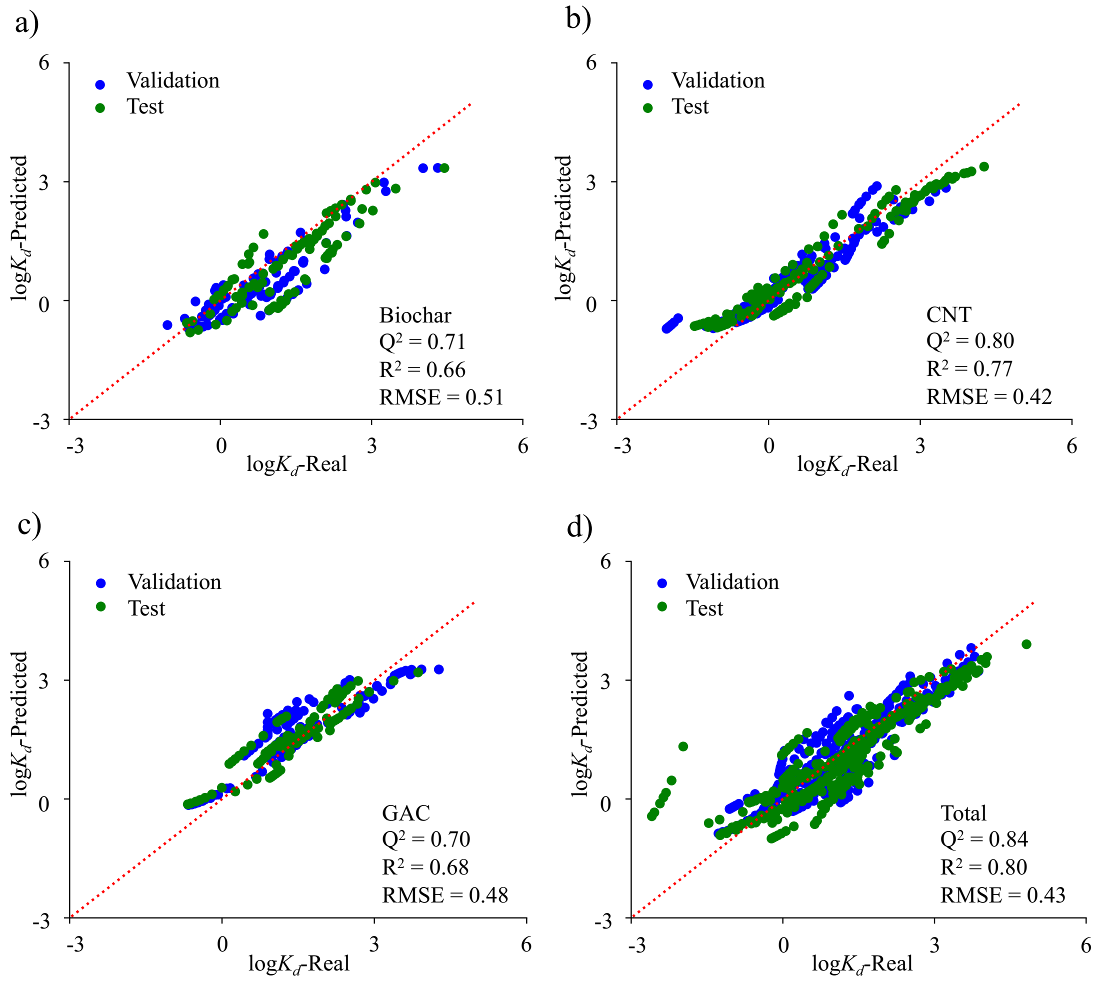


Fig. 3. Prediction performance of NN-LFER models for a) biochar, b) CNT, c) GAC, and d) all carbon-based adsorbents using Expanded Master Data Set from LSER Descriptors Data Sets with total sample size of 3122.

However, instead of directly employing 3-D coordinates and SMILES as features to describe chemical structures of organic molecules, new structural features were generated by PCA and GNN in terms of principal components and molecular vectors, in order to properly describe them. For example, for XGB-PCA-CS model, 30 features related to structural information of organic molecules were generated by PCA (Fig. 6a). Separately, each of these feature does not contain complete information on bond length nor bond distance. So, we combined all PCA generated structural features in order to consider the importance of chemical structures in XGB modeling. We performed feature importance analysis for XGB-PCA-CS, the results are shown in Fig. 6a, and are summarized in Fig. 6b. Our results show that among all features used for XGB-PCA-CS modeling, the chemical structure of organic molecules (a combination of 30 structural features) accounts for the highest proportion (36.69 %) in feature importance graph, followed by concentration of organic molecules (26.01 %), BET area of adsorbents (19.84 %), and pore volume of adsorbents (14.47 %). With a total feature importance of 62.7 %, this result exemplifies the importance to properly describe organic molecules in ML models for predicting their adsorption isotherm. Of course, the physical properties of adsorbents, although only accounts for 34.31 % of feature importance, is also important to the prediction accuracy.

In order to further enhance prediction accuracy of structural descriptor-based ML models, as well as understand their application domain, we analyzed the identity of outliers in the predicted vs. actual values plots (Figs. 7 and S12). The results are summarized in Tables S8 and S9. For prediction of all carbon-based adsorbents, we noticed there is no outliers for XGB-PCA-CS, while outliers have been detected for 2 data groups for XGB-GNN-SMILES, however, most data points of these data groups show excellent prediction result. In contrast, NN-PCA-CS shows 3 data groups with outliers, for non-outlier data within those groups, the prediction results are similar to real values. Further, for NN-GNN-SMILES, not only it contains 3 data groups with outliers, but all data points from these groups are also poorly predicted. This result is consistent with the order of prediction performance (Q^2, R^2, and RMSE) of these models. In addition, it shows XGB-PCA-CS has the greatest application domain in terms of organic molecules and carbon-based adsorbents, since no outlier was found. We further examined outliers for NN-PCA-CS for different carbon-based adsorbents (Fig. S12), our results show no outlier was found for GAC prediction, while both biochar and CNT predictions contain 2 data groups with outliers. This suggests for NN-PCA-CS, the current adsorbent description is adequate for GAC with limited pore volume and BET surface area, but inadequate for biochar and CNT, where higher BET surface area and pore volume, as well as additional functional group other than carboxyl group and hydroxyl group, can be expected, leading to limited application domain for NN-PCA-CS. It has to be noted that outliers share no similarity in their chemical structure nor physical property (Table S9). Therefore, in order to increase the application domain of ML models with outliers, it is necessary to further increase data diversity, although it is constrained by the limited experimental studies available.

3.4. Application of chemical structure-based machine learning on isomers

Isomers include constitutional isomers and stereoisomers. However, the selective adsorption of isomer can be difficult depending on the difference

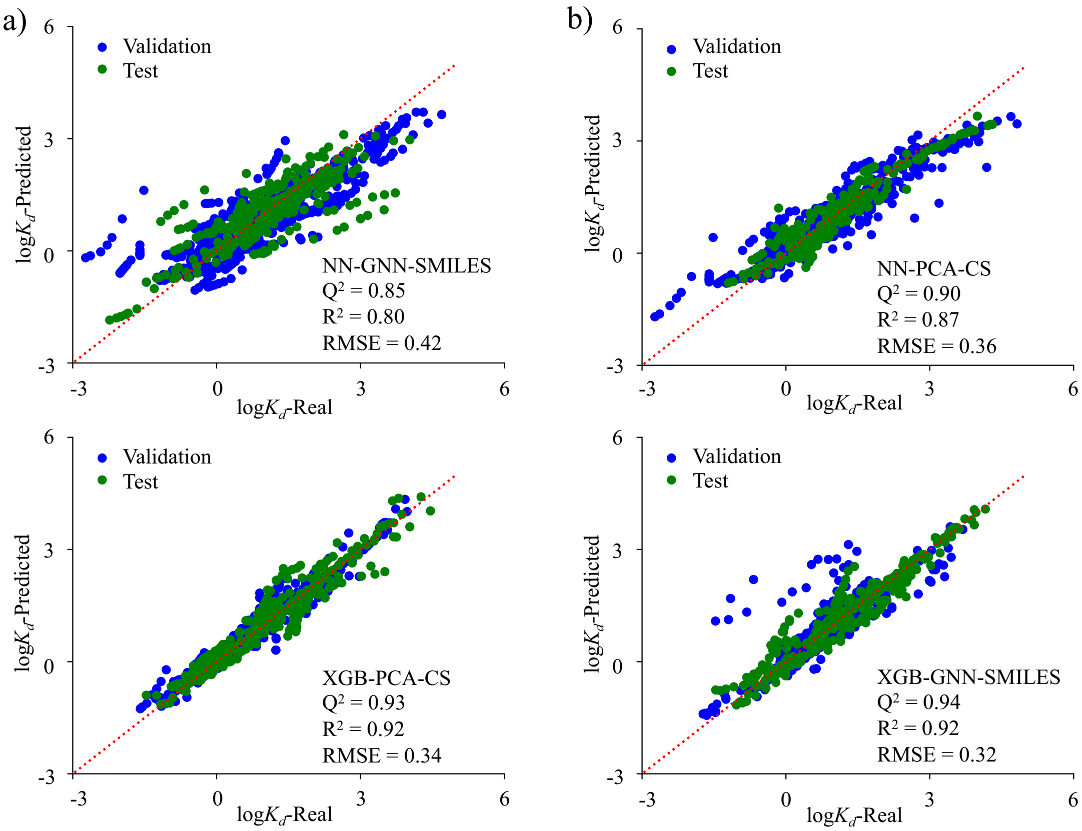


Fig. 4. Prediction performance of four ML models for all carbon-based adsorbents using Structural Descriptors Data Set: a) NN-GNN-SMILES, b) NN-PCA-CS, c) XGB-PCA-CS, and d) XGB-GNN-SMILES.

between their physiochemical properties. For example, cresol is a mixture of constitutional isomers including ortho-cresol (o-cresol), meta-cresol (m-cresol), and para-cresol (p-cresol). Since they share extremely similar boiling point, pKa, density, and even dipole moment, recrystallization is typically employed in the industry for their separation. Further, stereoisomers such as chiral chemicals often require chiral adsorbents for their separation (Casado et al., 2012; Chang et al., 2012), which comes in high cost and the selectivity is often difficult to predict. Because these isomers often behave differently in bioavailability and toxicity (Nikolai et al., 2006; Sanganyado et al., 2017), simple yet effective method is greatly desired for their selective removal from the environment. Numerous studies have reported the selective adsorption of constitutional isomers onto activated carbons (Ravi et al., 1998; Suresh et al., 2012). The primary reason for their differentiation is often attributed to the difference in binding mechanisms between the constitutional isomers and adsorbents, such as π-π interaction, hydrogen bonding, etc. Further, it has been shown that enantiomeric excess (ee) enrichment can also be achieved using achiral adsorbents (Farhadian et al., 2015; Gomis-Berenguer et al., 2020). Notably, a number of studies have shown that achiral carbon-based adsorbents can selectively adsorb chiral chemicals (Belhamdi et al., 2016; Gomis-Berenguer et al., 2020; Huang and Garcia-Bennett, 2021), albeit the exact mechanism for their differentiation is beyond the scope of this study. Nevertheless, such strategy offers a simple but efficient approach for the selective removal of isomers.

We first evaluated the applicability of CS-ML strategy on constitutional isomers and stereoisomers with similar physiochemical properties. Cresol isomers and a number of chiral compounds were chosen for this purpose, their physiochemical properties are listed in Table S10. We used their experimentally determined adsorption isotherm as test set and applied CS-ML models including NN-PCA-CS, NN-GNN-SMILES, XGB-PCA-CS, and XGB-GNN-SMILES. The results are shown in Fig. S13. It shows for the prediction of isomers with similar physiochemical properties, XGB-PCA-CS is better than other CS-ML based methods, followed by XGB-GNN-SMILES, and trailed by NN based models. This result is different to that of our previous observation, by which XGB-GNN-SMILES shows the best predict performance for all organic molecules (Table S8), in which only 18.5 % were isomers. This result suggests that the combination of chemical structure as feature descriptor and PCA as pre-processing method is more adequate to capture the subtle structural difference between isomers. In contrast, more structural information is lost when the combination of SMILES and GNN were employed. In terms of ML algorithm, XGB based models produced better prediction result for these isomers, which is consistent with our previous observation for all organic molecules (Table S8).

In order to visualize the difference in prediction performance between CS and LFER based models, we plotted the Langmuir adsorption isotherm curves of constitutional isomers (cresol isomers) and chiral organic molecules obtained from ML models against that of experimental values (Belhamdi et al., 2016; Gomis-Berenguer et al., 2020; Ravi et al., 1998). A close fit between XGB-PCA-CS predicted and experimentally derived isotherms can be seen in Figs. 8 and S14, while the adsorption isotherm predicted by NN-LFER significantly deviates from the experimental curve. This result is remarkable since all CS-ML models were trained using data set that does not include enantiomers. As discussed earlier, such good prediction result on chiral organic molecules can be primarily attributed to the more accurate description of 3D-structures of organic molecules.

It is beneficial to extract additional chemical and physical information from the predicted data. To this end, we have also calculated Langmuir constant (K_L) and maximum adsorption capacity (q_{max}) from the predicted isotherm curves. The results are summarized in Table S11. Unfortunately, both K_L *and* q_{max} calculated from the predicted isotherms are significantly different from those of experimentally derived values. The reason for such deviation may be due to the fact that our models were trained using adsorption data within a narrow range of C_e. For example, many experimental isotherms employed in the current study have not reached maximum adsorption capacity even for highest C_e reported. Therefore, fitting of the

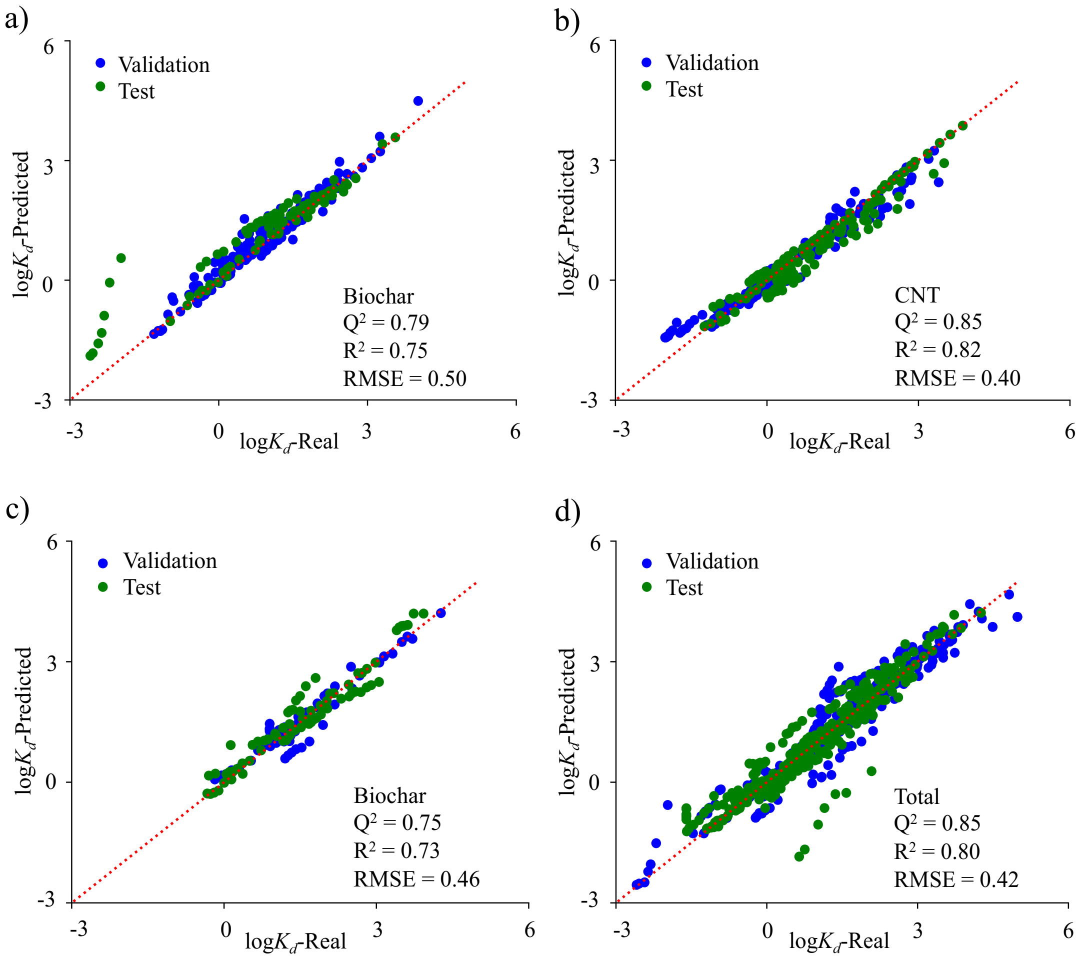


Fig. 5. Prediction performance of XGB-LFER models for a) biochar, b) CNT, c) GAC, and d) all carbon-based adsorbents using Expanded Master Data Set from LSER Descriptors Data Sets with total sample size of 3122.

isotherm data is necessary to calculate K_L and q_{max}, which often occur at C_e that is beyond the prediction range of our CS-ML models. As the result, in order to improve the prediction performance of K_L and q_{max}, it is necessary to achieve a much higher prediction accuracy of adsorption isotherm, which is unfortunately very difficult based on the limited data set that can be collected from the literatures.

4. Conclusion

In summary, we have developed ML-based models to predict adsorption isotherm of organic molecules on carbon-based adsorbents. We found ML models developed using structural descriptor are superior to models developed using LSER descriptors. In addition, we found XGB can produce more accurate prediction results compared with NN, and both GNN-SMILES and PCA-CS are capable to effectively describe the difference in organic molecules (RMSE of 0.34 and 0.32, Q^2 of 0.93 and 0.94 for XGB-PCA-CS and XGB-GNN-SMILES, respectively). However, for chiral and structural related isomers with similar chemical structure, XGB-PCA-CS is more adequate to differentiate them and is thus able to produce better prediction results (RMSE of 0.30, R^2 of 0.89) even when achiral molecules were used as training set. To our knowledge, this is the most accurate model in predicting adsorption isotherm of structural isomers up to date. Thus, our strategy is

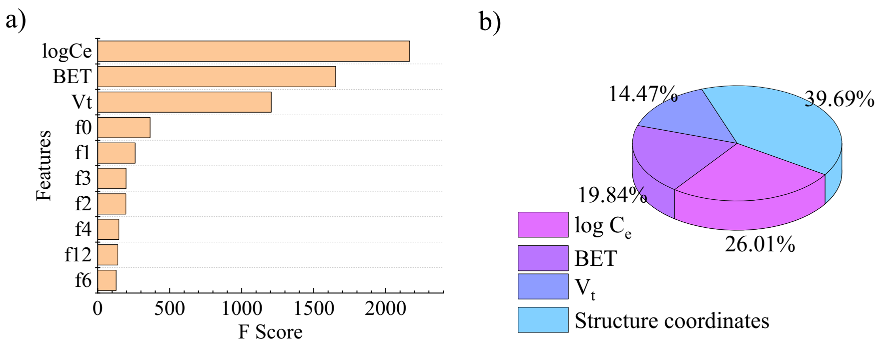


Fig. 6. Feature importance analysis of the XGB model: a) Top 10 features with highest F score; b) feature importance proportion for four main features.

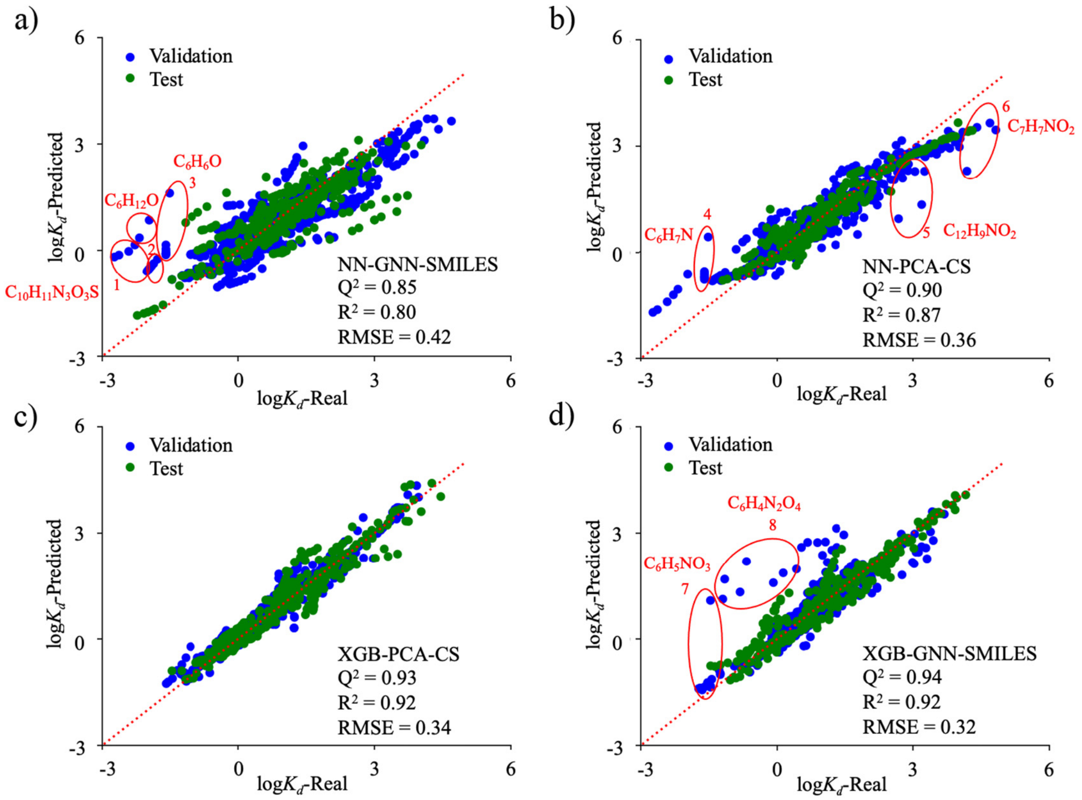


Fig. 7. The outliers in the scatter plots with structural descriptors and different algorithms for all carbon-based adsorbents.

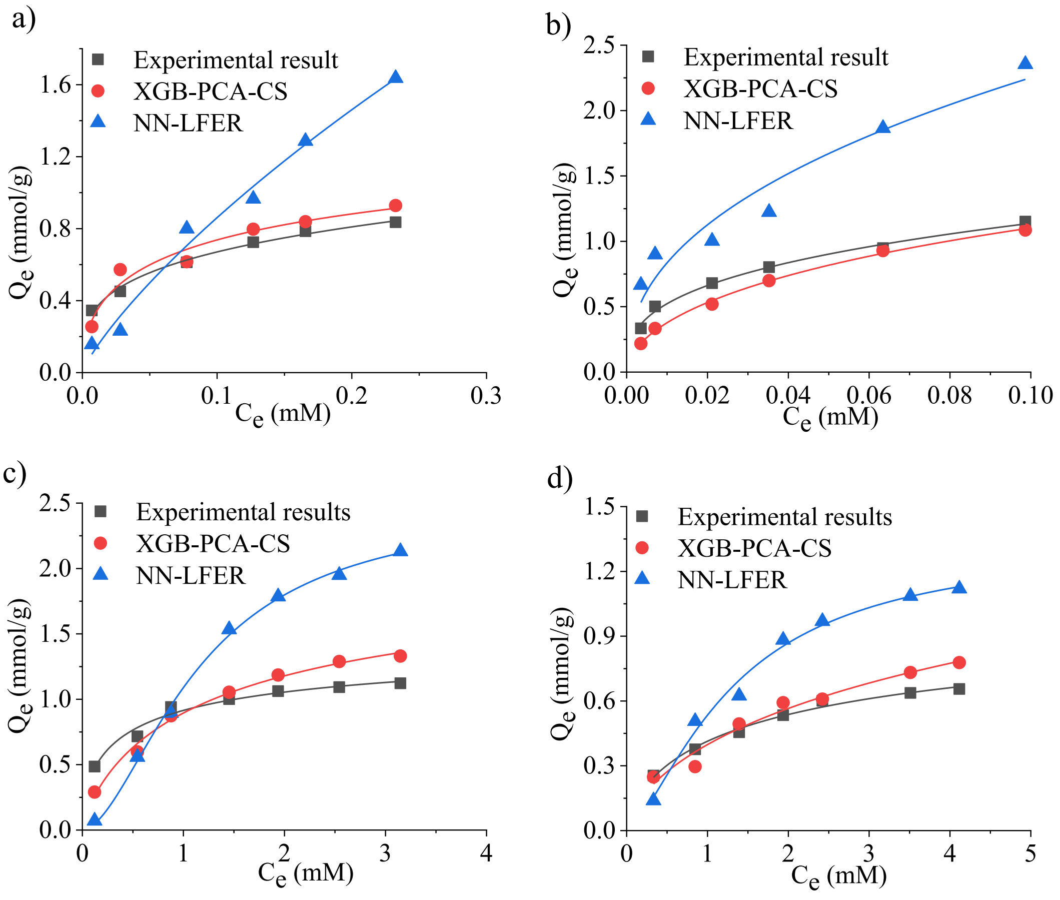


Fig. 8. Prediction of S-Metolachlor adsorption on a) L27 (Gomis-Berenguer et al., 2020) and b) AQ630 (Gomis-Berenguer et al., 2020) activated carbons; *L*-phenylalanine adsorption on c) ACK (Belhamdi et al., 2016) and d) ACZ (Belhamdi et al., 2016) activated carbons.

of broad interest to researches related to environmental applications such as removal of prioritized contaminants, prediction of uptake and toxicity of isomers.

In addition, based on feature importance analysis, our study reveals that the structure of organic molecules plays a major role in determine their adsorption capacity on carbon-based adsorbents, followed by their solution concentration, and characteristics of adsorbents. To accurately evaluate the performance of ML models with limited data set, we proposed a Medium-Selection Cross-Validation Method, from which we found XGB-PCA-CS is a universal ML model to predict the adsorption capacity of organic molecules on carbon-based adsorbents, since no outliers were detected. Our results can be used to guide the design and optimization of high-performance carbon-based adsorbent. It also has to be noted that some of the adsorbents involved in the current study exhibits toxicity at molecular, cellular, and animal levels (Liu et al., 2013), thus it is beneficial to take toxicity into consideration when designing adsorbents; however, due to the limited information on adsorbent toxicity that is available, it is beyond the scope of the current study. In addition, since the majority of adsorption isotherm reported up to date were performed at relatively high concentration of organic contaminants in order to determine maximum adsorption capacity, adsorption performance at low concentration is out of the application domain of the current ML model. But nevertheless, our results provide insight in choosing proper molecular descriptors for environmental applications of ML where limited data set and a diverse range of organic molecules are involved.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.160228.

CRediT authorship contribution statement

Chaoyi Huang: Methodology, Investigation, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Wenyang Gao:** Data curation, Software. **Yingdie Zheng:** Data curation. **Wei Wang:** Methodology. **Yue Zhang** Supervision. **Kai Liu:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kai Liu reports financial support was provided by Zhejiang Provincial Natural Science Foundation of China.

Acknowledgments

This work was supported by the Westlake Education Foundation [Grant No. 103506022001, 2021]; Westlake University-Muyuan Joint Research Institute [Grant No. 206006022007, 2021]; and Zhejiang Provincial Natural Science Foundation of China [Grant No. LQ20C030002, 2021].

References

Apul, O.G., Wang, Q., Shao, T., Rieck, J.R., Karanfil, T., 2013. Predictive model development for adsorption of aromatic contaminants by multi-walled carbon nanotubes. Environ. Sci. Technol. 47, 2295–2303.

Asare, K.O., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., Ferreira, D., 2021. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR mHealth uHealth 9, e26540.

Belhamdi, B., Merzougui, Z., Trari, M., Addoun, A., 2016. A kinetic, equilibrium and thermodynamic study of l-phenylalanine adsorption using activated carbon based on agricultural waste (date stones). J. Appl. Res. Technol. 14, 354–366.

Caetano, M., Valderrama, C., Farran, A., Cortina, J.L., 2009. Phenol removal from aqueous solution by adsorption and ion exchange mechanisms onto polymeric resins. J. Colloid Interface Sci. 338, 402–409.

Casado, C., Castán, J., Gracia, I., Yus, M., Mayoral, Á., Sebastián, V., et al., 2012. L- and d-proline adsorption by chiral ordered mesoporous silica. Langmuir 28, 6638–6644.

Chang, C., Wang, X., Bai, Y., Liu, H., 2012. Applications of nanomaterials in enantioseparation and related techniques. TrAC Trends Anal. Chem. 39, 195–206.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, pp. 785–794.

Ching, P.M.L., Zou, X., Wu, D., So, R.H.Y., Chen, G.H., 2022. Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. Environ. Res. 210, 112953.

Cordero, J.A., He, K., Janya, K., Echigo, S., Itoh, S., 2021. Predicting formation of haloacetic acids by chlorination of organic compounds using machine-learning-assisted quantitative structure-activity relationships. J. Hazard. Mater. 408, 124466.

Dickenson, E.R., Drewes, J.E., 2010. Quantitative structure property relationships for the adsorption of pharmaceuticals onto activated carbon. Water Sci. Technol. 62, 2270–2276.

Dietterich, T., 1995. Overfitting and undercomputing in machine learning. ACM Comput. Surv. 27, 326–327.

Farhadian, N., Sharifi, A., Lashgari, E., 2015. Selective adsorption of metoprolol enantiomers using 2-hydroxypropyl-β-cyclodextrin cross-linked multiwalled carbon nanotube. Biomed. Chromatogr. 29, 366–372.

Gomis-Berenguer, A., Laidin, I., Renoncial, S., Cagnon, B., 2020. Study of enantioselective metolachlor adsorption by activated carbons. RSC Adv. 10, 40321–40328.

Halgren, T., 1999. MMFF VI. MMFF94s option for energy minimization studies. J. Comput. Chem. 20, 720–729.

Huang, Y., Garcia-Bennett, A.E., 2021. Equilibrium and kinetic study of l- and d-valine adsorption in supramolecular-templated chiral mesoporous materials. Molecules (Basel, Switzerland) 26, 338.

Kennicutt, A.R., Morkowchuk, L., Krein, M., Breneman, C., Kilduff, J., 2016. A quantitative structure–activity relationship to predict efficacy of granular activated carbon adsorption to control emerging contaminants. SAR QSAR Environ. Res. 27, 1–24.

Kim, S., Bolton, E.E., Bryant, S.H., 2013. PubChem3D: conformer ensemble accuracy. J. Cheminformatics 5, 1.

Li, H.Y., Osman, H., Kang, C.W., Ba, T., 2017. Numerical and experimental investigation of UV disinfection for water treatment. Appl. Therm. Eng. 111, 280–291.

Liu, Y., Zhao, Y., Sun, B., Chen, C., 2013. Understanding the toxicity of carbon nanotubes. Acc. Chem. Res. 46, 702–713.

Liu, X., Liu, T., Feng, P., 2022. Long-term performance prediction framework based on XGBoost decision tree for pultruded FRP composites exposed to water, humidity and alkaline solution. Compos. Struct. 284, 115184.

Lowe, M., Qin, R., Mao, X., 2022. A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. Water 14, 1384.

Luo, Z., Yao, B., Yang, X., Wang, L., Xu, Z., Yan, X., et al., 2022. Novel insights into the adsorption of organic contaminants by biochar: a review. Chemosphere 287, 132113.

Nikolai, L.N., McClure, E.L., Macleod, S.L., Wong, C.S., 2006. Stereoisomer quantification of the beta-blocker drugs atenolol, metoprolol, and propranolol in wastewaters by chiral high-performance liquid chromatography-tandem mass spectrometry. J. Chromatogr. A 1131, 103–109.

Pai, C.-W., Wang, G.-S., 2022. Treatment of PPCPs and disinfection by-product formation in drinking water through advanced oxidation processes: comparison of UV, UV/Chlorine, and UV/H2O2. Chemosphere 287, 132171.

Qi, X., Li, X., Yao, H., Huang, Y., Cai, X., Chen, J., et al., 2020. Predicting plant cuticle-water partition coefficients for organic pollutants using pp-LFER model. Sci. Total Environ. 725, 138455.

Ravi, V.P., Jasra, R.V., Bhat, T.S.G., 1998. Adsorption of phenol, cresol isomers and benzyl alcohol from aqueous solution on activated carbon at 278, 298 and 323 K. J. Chem. Technol. Biotechnol. 71, 173–179.

Rojas, S., Horcajada, P., 2020. Metal-organic frameworks for the removal of emerging organic contaminants in water. Chem. Rev. 120, 8378–8415.

Sagi, O., Rokach, L., 2021. Approximating XGBoost with an interpretable decision tree. Inf. Sci. 572, 522–542.

Sanganyado, E., Lu, Z., Fu, Q., Schlenk, D., Gan, J., 2017. Chiral pharmaceuticals: a review on their environmental occurrence and fate processes. Water Res. 124, 527–542.

Su, L., Wang, Y., Wang, Z., Zhang, S., Xiao, Z., Xia, D., et al., 2022. Simulating and predicting adsorption of organic pollutants onto black phosphorus nanomaterials. Nanomaterials 12, 590.

Suresh, S., Srivastava, V.C., Mishra, I.M., 2012. Adsorption of catechol, resorcinol, hydroquinone, and their derivatives: a review. Int. J. Energy Environ. Eng. 3, 32.

Tsubaki, M., Tomii, K., Sese, J., 2018. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35, 309–318.

Ulrich, N., Endo, S., Brown, T.N., Watanabe, N., Bronner, G., Abraham, M.H., et al., 2017. UFZ-LSER Database v 3.2 [Internet].

Van Duck, P.J., van de Voorde, H., 1984. Activated charcoal and microflora in water treatment. Water Res. 18, 1361–1364.

Wang, J., He, L., Lu, X., Zhou, L., Tang, H., Yan, Y., et al., 2022. A full-coverage estimation of PM2.5 concentrations using a hybrid XGBoost-WD model and WRF-simulated meteorological fields in the Yangtze River Delta urban agglomeration, China. Environ. Res. 203, 111799.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.

Xu, J., Wang, L., Sun, H., 2021. Adsorption of neutral organic compounds on polar and nonpolar microplastics: prediction and insight into mechanisms based on pp-LFERs. J. Hazard. Mater. 408, 124857.

Ying, X., 2019. An overview of overfitting and its solutions. Journal of Physics: Conference Series 1168, 022022 IOP Publishing.

Yu, X., Sun, W., Ni, J., 2015. LSER model for organic compounds adsorption by single-walled carbon nanotubes: comparison with multi-walled carbon nanotubes and activated carbon. Environ. Pollut. 206, 652–660.

Yuan, X., Suvarna, M., Low, S., Dissanayake, P.D., Lee, K.B., Li, J., et al., 2021. Applied machine learning for prediction of CO2 adsorption on biomass waste-derived porous carbons. Environ. Sci. Technol. 55, 11925–11936.

Zhang, K., Zhong, S., Zhang, H., 2020. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. Environ. Sci. Technol. 54, 7008–7018.

Zhao, Y., Lin, S., Choi, J.-W., Bediako, J.K., Song, M.-H., Kim, J.-A., et al., 2019. Prediction of adsorption properties for ionic and neutral pharmaceuticals and pharmaceutical intermediates on activated charcoal from aqueous solution via LFER model. Chem. Eng. J. 362, 199–206.

Zhao, Y., Fan, D., Li, Y., Yang, F., 2022. Application of machine learning in predicting the adsorption capacity of organic compounds onto biochar and resin. Environ. Res. 112694.

Zhu, X., He, M., Sun, Y., Xu, Z., Wan, Z., Hou, D., et al., 2022. Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning. J. Hazard. Mater. 423, 127060.