Environmental Science & Technology

pubs.acs.org/est

Article

Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning

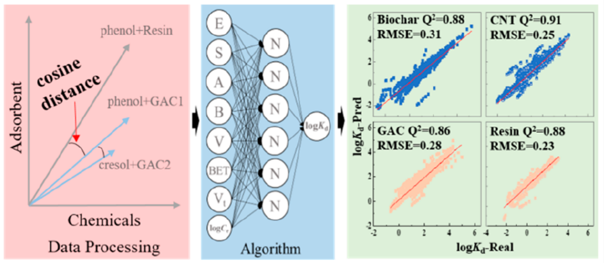
Kai Zhang, Shifa Zhong, and Huichun Zhang*

Cite This: https://dx.doi.org/10.1021/acs.est.0c02526

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Predictive models are useful tools for aqueous adsorption research; existing models such as multilinear regression (MLR), however, can only predict adsorption under specific equilibrium concentrations or for certain adsorption isotherm models. Also, few studies have discussed data processing beyond applying different modeling algorithms to improve the prediction accuracy. In this research, we employed a cosine similarity approach that focused on mining the available data before developing models; this approach can mine the most relevant data concerning the prediction target to build models and was found to considerably improve the prediction accuracy. We then built a machine-learning modeling process based on neural networks (NN), a group-selection data-splitting strategy for grouped adsorption data for adsorbent−adsorbate pairs under different equilibrium concentrations, and polyparameter linear free energy relationships (pp-LFERs) for aqueous adsorption of 165 organic compounds onto 50 biochars, 34 carbon nanotubes, 35 GACs, and 30 polymeric resins. The final NN-LFER models were successfully applied to various equilibrium concentrations regardless of the adsorption isotherm models and showed less prediction deviations than the published models with the root-mean-square errors 0.23−0.31 versus 0.23−0.97 log unit, and the predictions were improved by adding two key descriptors (BET surface area and pore volume) for the adsorbents. Finally, interpreting the NN-LFER models based on the Shapley values suggested that not considering equilibrium concentration and properties of the adsorbents in the existing MLR models is a possible reason for their higher prediction deviations.



INTRODUCTION

Aqueous adsorption, a long-standing purification/separation process, has been continuously investigated for decades; however, challenges still exist for this technology. Traditional batch and column experiments are time-consuming and inefficient toward a growing number of chemicals and adsorbents, and this has led to scarcity in data for adsorption of (new) compounds on (new) adsorbents.[1−4] Also, although the adsorption of many model chemicals like phenols on new or modified adsorbents has been explored extensively,[5−8] these abundant data have not been fully utilized other than for the comparison or selection of a small number of adsorbents. Mining published data to build broad predictive models will be a promising solution to fill the gaps. A model with high prediction accuracy can replace some labor-intensive adsorption experiments; even moderately accurate models are valuable, as they can facilitate the design of adsorption experiments by quickly estimating the adsorbed amounts of different chemicals by given adsorbents (e.g., help to determine the mass of the adsorbents used in the experiments).

The adsorbed amount (Q_e) under an equilibrium concentration (C_e) is essentially a function of three key sets of properties: the properties of the chemical, the properties of the adsorbent, and the equilibrium concentration C_e of the chemical with respective to the adsorbent.

$$\log K_d = \log \frac{Q_e}{C_e} = f(\text{chemical, adsorbent, } C_e) \tag{1}$$

where the adsorption coefficient log K_d has been commonly employed to quantify the extent of adsorption.

Most studies have relied on two approaches to develop predictive models. The first approach examines the adsorption of many chemicals on a small number of similar adsorbents (referred to as the "chemical-based approach" hereafter), among which polyparameter linear free energy relationships (pp-LFERs) are one of the most widely used.[9−11] In the pp-LFERs approach, multilinear regression (MLR) is established between log K_d and the Abraham descriptors (E, S, A, B, and

Received: April 21, 2020
Revised: May 6, 2020
Accepted: May 8, 2020
Published: May 8, 2020

V)[12] of different chemicals under a selected equilibrium concentration level C_e:

$$\log K_d(C_e) = e \cdot E + s \cdot S + a \cdot A + b \cdot B + v \cdot V + c \quad (2)$$

where the equilibrium concentration level C_e is equal to a fraction of a chemical's water solubility (S_w) (e.g., $C_e = 0.01 \times S_w$), e, s, a, b, v, and c are the fitting parameters, and $\log K_d(C_e)$ is the adsorption coefficient of a chemical under a given C_e. The Abraham descriptors E, S, A, B, and V can capture nonspecific interactions arising from induced dipoles, stable polarity (i.e., dipole−dipole interactions), overall H-bonding acidity and basicity (electron-accepting and -donating capacities), and cavitation energy and part of London dispersive forces beyond what is captured by the E term, respectively. With the obtained fitting parameters, the adsorption coefficients of new chemicals (with known Abraham descriptors) at the same C_e can then be calculated through the regression equations.[13,14] Besides the pp-LFERs, quantitative structure−property relationships (QSPRs) have been used to predict aqueous adsorption as either Q_e or K_d,[15−17] where different QSPR descriptors such as those for hydrogen bonding have been used as the independent variables.[18−20]

The pp-LFERs approach, however, generally needs to build one MLR model per C_e, and the predictions based on the established MLR models are only limited to the concentration levels involved in the modeling, which is not helpful when predictions under different equilibrium concentrations are needed. To expand the prediction ability to multiple concentration levels, a few studies have attempted to unify several MLR models into one equation by treating the fitting parameters (e, s, a, b, v, and c) in eq 2 as concentration-dependent.[21,22] Unfortunately, the concentration levels are still limited by the available experimental data (typically covering several concentration levels such as $C_e = 0.001 \times S_w$ or $0.01 \times S_w$) and are not able to cover wide concentration ranges in real applications. This is because some chemicals are much less soluble than others such that we cannot obtain their experimental $\log K_d$ under high equilibrium concentrations; as a result, developing pp-LFERs at multiple concentration levels will greatly reduce the number of chemicals (and hence the amount of data available) that can be included in the modeling process, which further limits the applicability and accuracy of the MLR models.

Another limitation of the MLR modeling is the implicit simplification that the Abraham descriptors are the only variables leading to different adsorption coefficients for chemicals under a given C_e so that the impact of the adsorbent properties is not considered. For various adsorbents belonging to the same class but with different properties (such as surface area (BET) and total pore volume (V_t)), they are treated the same even though they have different adsorption patterns toward the same chemicals. Not including descriptors to capture differences in the adsorbents can inevitably lead to larger prediction deviations by MLR models. Meanwhile, this simplification prevents a possible improvement in the prediction accuracy by incorporating descriptors for adsorbents.

Treating adsorbents that belong to the same class (e.g., GAC) but with significantly different properties equally can lead to other problems. In the MLR models, the adsorption data of one chemical on one class of adsorbents is only considered once (the MLR models require only one dependent variable value per chemical under every C_e), even when the adsorption data of the chemical on several adsorbents are available. This allows the utilization of only a small portion of the abundant adsorption data. Considering the adsorption of phenol on different GACs as an example, although there are plenty of reported data for phenol adsorption on different GACs, only one or an average of the adsorption coefficients per C_e may be used in the MLR modeling; this makes much of the phenol's adsorption data seem repetitive. It is however known that the adsorption of phenol on different GACs can show distinctively different patterns,[23−25] so the seemly repetitive data actually contain important information about the interactions between phenol and the different GACs, which has barely been utilized by the MLR models.

The second approach correlates the adsorption of one or a small number of chemicals under a specific C_e with key adsorbent properties, such as BET, V_t, and particle size of different adsorbents (referred to the "adsorbent-based approach" hereafter).[26−28] The relationships can be used to predict the adsorption of that chemical on new similar adsorbents. In this approach, the established models are mostly confined to predicting the adsorption of a small portion of chemicals.[26,29,30]

To address the limitations in the existing models, a modeling strategy combining the chemical-based and adsorbent-based approaches can provide a helpful solution, as it can not only allow predictions for varying C_e or adsorption isotherm models but also utilize the abundant adsorption data of various chemicals as well as include properties of adsorbents to improve the prediction accuracy. However, no established simple regression methods can combine these two approaches to build such predictive models.

Machine learning (ML) algorithms such as neural networks (NN),[31] the support vector machine (SVM),[32] and Bagging (a tree-based algorithm)[33] have emerged as powerful tools for uncovering hidden relationships. They have drawn much attention and achieved great success in problems related to adsorption prediction,[34] material design, and reaction parameter optimization.[35] Plenty of research has employed ML to model single/multicomponent adsorption with significant improvements compared with traditional regression methods.[36−39] Different research has also used ML to optimize adsorption parameters.[40−42] In several publications, ML has been successfully used to build predictive models based on pp-LFERs or QSPRs.[20,43−51] However, even with versatile ML algorithms, few have ever tried to incorporate properties of the adsorbents into the models, and most models are still only able to predict the adsorption under selected C_e or for certain adsorption isotherm models. One recent study has reported a major advance in this area by incorporating the surface areas as well as the C, H, and O contents of the adsorbents into a deep neural network model (20 hidden layers) to predict the fitting parameters of Freundlich isotherms, by which it largely resolved the limitation on equilibrium concentrations by MLRs.[52] However, plenty of research has documented better isotherm fittings by other models such as the Langmuir, Dubinin−Radushkevich, or Temkin isotherm.[53] So, it is valuable that the predictive models for adsorption are independent of adsorption isotherm models. The descriptors for adsorbents should also be both highly related to the adsorption mechanisms and widely available. BET and V_t are among the most commonly reported properties for porous adsorbents; more importantly, these two parameters are critical

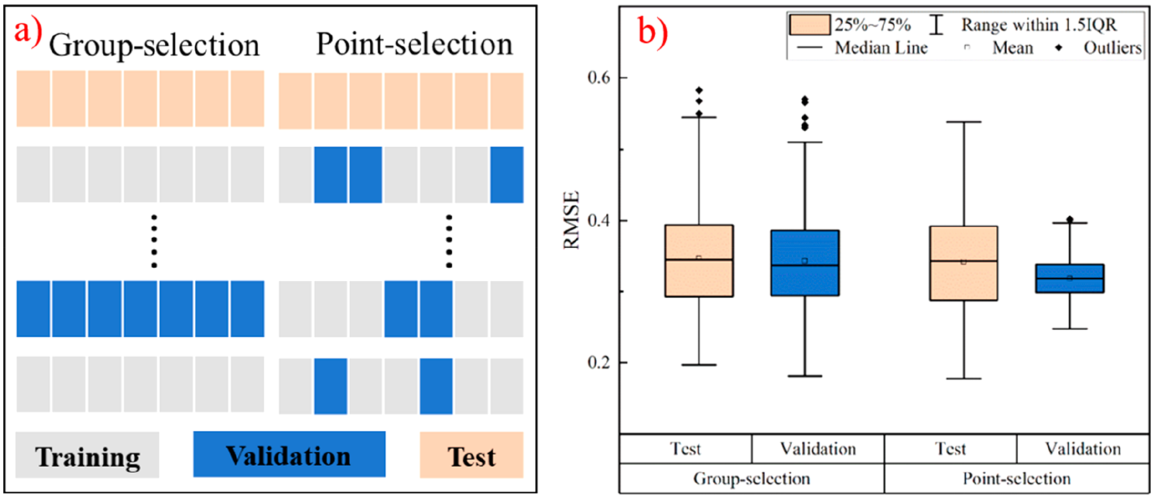


Figure 1. (a) Two different data-splitting approaches, in which each small rectangle is one data point, each row represents one group of data, and the rectangles with gray, blue, and orange colors indicate that the data points were selected into the training, validation, and test sets, respectively. (b) Performance of the two data-splitting approaches for the resin adsorption data (the collected data were split into the training, validation, and test by the ratio of 0.7:0.15:0.15).

for deciding the adsorption of organic compounds through hydrophobic interactions and pore-filling, two key mechanisms for organic compounds to be adsorbed by various adsorbents.[26,29,30] However, no ML models have considered their contributions to adsorption prediction.

Besides the algorithms, the abundance of source data plays a crucial role in improving the prediction accuracy. In other ML research such as natural language processing (NLP)[54] and face recognition,[55] the prediction can be improved by feeding a tremendous amount of data to the models. For experimental research like adsorption, however, the available data are almost always limited, so it will be beneficial to improve the prediction beyond only selecting different algorithms or optimizing the hyperparameters. Unfortunately, how to improve the prediction accuracy for adsorption concerning source data processing has been rarely discussed.

To address the above limitations, we first mined the literature for the adsorption data of 165 organic chemicals on 50 biochars, 34 carbon nanotubes (CNTs), 35 granular activated carbons (GAC), and 30 polymeric resins. We then compared the predictive models using the NN, SVM, Bagging, and three classical regression algorithms and selected NN as the best-performing one (grid search was used to optimize the hyperparameters). We also evaluated two different data-spitting methods to avoid the potential data leakage problem in ML for grouped data (details below). Next, a cosine similarity approach was employed to select the data that are most relevant to a prediction target as the training set to build the NN-LFER models, which helped to further improve the prediction accuracy for a given modeling algorithm. The Shapley values were then calculated to quantify the contribution of each input descriptor to the overall adsorption coefficients and to assess whether the built ML models violated any adsorption rules. Additional comparisons between this and previously published research are in given Text S1 and Figure S1. Lastly, an easy-to-use tool with a Graphical User Interface (GUI) was developed based on the trained ML models.

■ MATERIALS AND METHODS

Four types of widely investigated adsorbents, including three carbon materials (namely biochar, CNTs, and GAC) and polymeric resins, were selected as the target adsorbents. The adsorption isotherms were first collected from the literature, and the data points were then calculated from the isotherms within the reported equilibrium concentration ranges. To obtain good predictive models, high-quality, representative source data are the key. Although there are numerous studies on adsorption, only a small portion has been used to build predictive models because either the data is not of high quality or some key parameters are not reported. For any adsorption isotherm to be selected, it needs to meet the following three requirements: (1) the isotherm fitting coefficient (R^2) needs to be higher than 0.95 with at least 7 experimental data points in the isotherms; (2) the adsorbent properties including at least BET and V_t should be measured and reported (other properties including macropore volume and micropore volume were also collected, if available); and (3) for the adsorption of chemicals onto multiple adsorbents with minor differences, only one adsorbent was chosen for each chemical. For instance, in the adsorption of phenol on two GACs, if the difference in the BET between the two GACs was ≤10 m^2/g, only one of the GACs was chosen because the adsorption capacities were similar. For compounds with acid or base functional groups, the experimental pH of the adsorption data was carefully selected to ensure that the majority of the chemicals existed predominantly in the neutral form (more information about data collection is given in Text S2). Following these requirements, 4102 adsorption data points associated with 586 isotherms were mined from the literature (each isotherm has 7 concentration points, which, together with the corresponding adsorbent properties, will be referred to a "group" hereafter), covering the adsorption of 165 organic chemicals on 50 biochars, 34 CNTs, 35 GACs, and 30 polymeric resins (Tables S5−8).

During the ML modeling, the input data had eight descriptors including the equilibrium concentration log C_e (μM), five Abraham descriptors (E, S, A, B, and V) for the chemicals, and two descriptors for the adsorbents (BET in m^2/

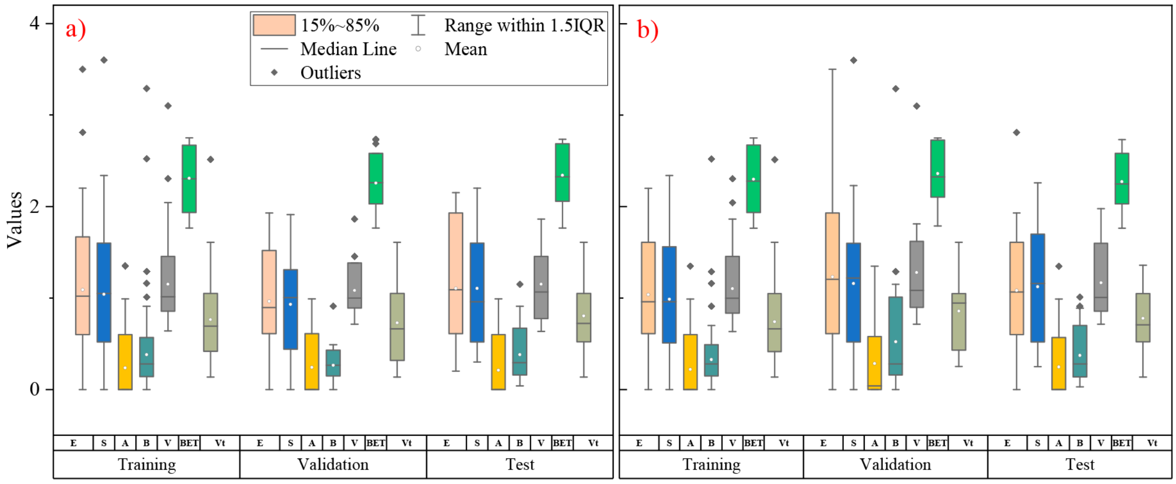


Figure 2. Ranges of the input descriptor (BET in the plot represents log BET) values for 924 training, 196 validation, and 196 test data points for CNTs under two scenarios: (a) a good predictive model (Q^2 = 0.9438 and 0.9486 for the validation and the test set separately) and (b) a poor model (Q^2 < 0.82 for both the validation and test sets).

g and V_t in cm^3/g). The output was the adsorption coefficient log K_d (L/g). The data were then standardized onto the same scale (between −1 and 1) to eliminate possible bias before feeding them to the models because the input values varied within large ranges (e.g., the BET can be as high as 2000 m^2/g, while V_t is less than 2 cm^3/g). This also helped to accelerate the training process (a flowchart for the standardization-involved ML process can be found in Figure S2 and Text S3).

The procedure for the modeling included four steps: (1) collecting published data, (2) selecting the prediction target including chemicals and adsorbents, referred to as the validation set, (3) developing predictive models, and (4) applying the developed models to the prediction target. When training an ML model, data splitting is the first and one of the most critical steps. A good approach is necessary to prevent possible data leakage in the splitting process. This is essential for studies like this because there are grouped data points. Data leakage occurs when some data points in the training and validation sets are from the same group(s). Two different data-splitting approaches (Figure 1a) were tested and compared in this research: (1) group selection that selected the entire group (i.e., one adsorption isotherm plus the adsorbent properties) to the training, validation, or test sets and (2) point selection that randomly split data points from the collected data minus the test set into the training or validation set (the test set still only contained the entire group).[37]

The optimal configuration of the ML models was determined by the grid search method in Matlab R2019a with a deep learning toolbox (details in Text S4 and Figure S3). The performance of the Bagging models with different numbers of base estimators, SVM models with different kernels, and NN models with different hyperparameters (including learning goal, learning rate, activation function, training function, number of hidden layers, and number of neurons in the hidden layers) was evaluated by comparing the root-mean-square error (RMSE, eq S8) and Q^2_{F2} (referred to as Q^2 hereafter, eq S9)[56] values on the validation set. The best NN configuration was selected with the minimal RMSEs and maximal Q^2, and a simpler NN model with fewer hidden layers and neurons in the hidden layers was preferred to avoid possible overfitting (Table S1).[57]

Please see the Supporting Information for additional details on the data splitting (Text S5), cosine similarity (Text S6), outliers (Text S7), and calculations of the Shapley values (Text S8).

■ RESULTS AND DISCUSSION

Different Algorithms and Data-Splitting Approaches. A comparison among different regression algorithms (Table S2) showed that commonly used methods such as Ridge, Lasso, and Elastic Net had a poor performance in dealing with this problem. Specifically, the Ridge provided very poor predictions with high RMSEs and low Q^2; the Lasso and Elastic Net provided acceptable predictions for CNTs, GACs, and resins, but their prediction accuracies for biochars were much lower than those of the SVM or Bagging method. The SVM approach had a slightly lower prediction accuracy than the Bagging method. The Bagging method achieved comparable results with those by the NN method over the four types of adsorbents, but adding another metric mean absolute error (MAE, which is less affected by extreme values) for comparison, the prediction by the Bagging showed significantly higher MAE than the NN method for the biochars and resins (Table S3). Considering both the RMSE and MAE, the NN approach was better and hence was selected.

Four NN models containing different input parameters were built, and their performances were evaluated (Figure S4). Compared with the models considering only the Abraham descriptors as the input, the RMSEs of the prediction were significantly improved by adding either BET or V_t, but the models considering the Abraham descriptors with both BET and V_t as the input gave the best predictions for all four adsorbents. These results indicated the significance of incorporating the key properties of the adsorbents in the prediction models.

The results in Figure 1b and Figures S5−8 showed that for the point-selection approach, the RMSEs of the validation sets maintained at relatively low levels whereas those of the test sets varied within much broader ranges. For instance, there was a significant difference (p value = 3.31 × 10^{-9}) between the RMSEs of the validation and test sets for resins in the point-selection approach (Figure 1b). In contrast, the test and validation sets had comparable RMSEs (p value = 0.5375) in

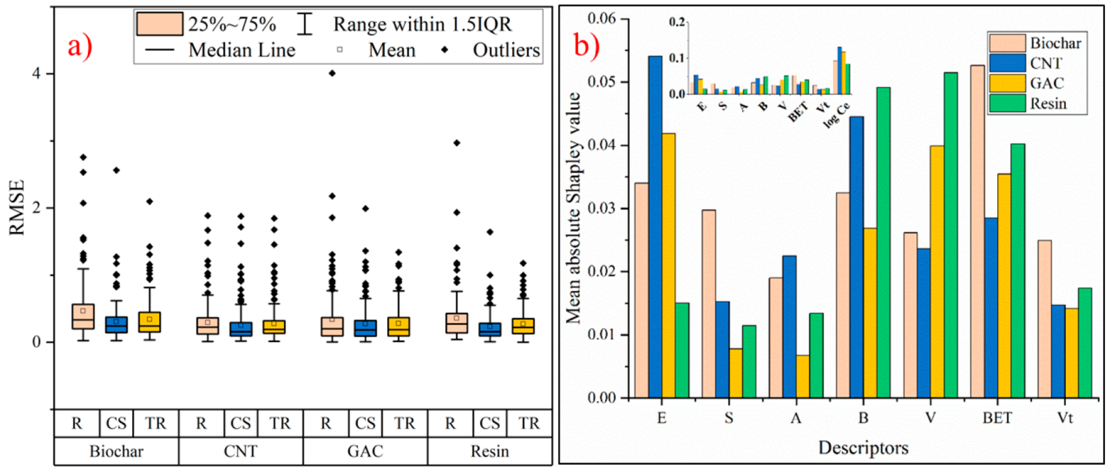


Figure 3. (a) Comparison of the performance of the models based on three different training set selection approaches: R, randomly selected; CS, cosine similarity; TR, leave-one-out cross-validation. Leave-one-out cross-validation means that one group was chosen as the validation set at a time, and the rest was the training set until all of the groups had been selected once as the validation set. The prediction target in each of the three approaches was the same, and the difference was in the training sets. CS had lower RMSE values than TR with the p values of 0.4008, 0.0203, 0.5375, and 0.0178 for biochar, CNTs, GACs, and resins, respectively. TR also had lower RMSE values than R, with the p values of 0.0003, 0.0093, 0.6114, and 0.0002 for biochars, CNTs, GACs, and resin, respectively. (b) Mean absolute Shapely (MAS) values for the input descriptors on the four adsorbents. The inset shows the dominant effect of the log C_e term on the MAS values.

the group-selection approach. Also, there was no significant difference in the RMSEs between the test set in the point-selection approach and the test or validation set in the group-selection approach (p value = 0.4463 and 0.9085, respectively). It means that the validation set in the group-selection can successfully reflect the predicting capability of the NN models on the test set, demonstrating the robustness of the group-selection approach.

The possible reason for the above results is that data leakage happened in the point-selection. In the MLR modeling, randomly splitting data will not have this problem because there is only one data point per C_e per chemical; no point can appear in both the training and the validation sets and the independence between the two data sets is guaranteed; therefore, no data leakage happens. However, more attention should be paid to the random data-splitting process when the data are grouped because some data points from a group may go into the validation set and the rest may go into the training set. Once data leakage happens, it may lead to overfitting.[58] That is, a model that gives satisfactory predictions on the validation set may not be able to perform well for an external test set. In this case, the training set has already contained some features (e.g., adsorbent properties and/or descriptors of the chemicals) of the validation set by the point-selection, so low RMSEs on the validation set are achieved, but the test set is independent of the training set such that the model performs poorly on the test set. For the group-selection, because all of the data sets are selected by groups, the training, validation, and test sets are independent, so a better consistency in the prediction performance, i.e., similar RMSEs and/or Q^2 values, between the validation and the test sets is achieved. Also, the point-selection is not consistent with real applications because an unknown target will never appear in the training set.[59] Therefore, the group-selection approach was employed. Further discussion of the two data-splitting approaches is given in Text S5.

Cosine Similarity. During the training process, a qualitative trend was observed in the ranges of the descriptor values (Figure 2). That is, the ranges of the descriptor values for the training and validation/test sets were similar when we observed satisfactory predictions (an example is shown in Figure 2a) but were considerably different when we observed relatively poor predictions (e.g., descriptors E and V in Figure 2b). Such a similarity is also physically meaningful; that is, a similarity in the descriptor ranges means that the combination of the chemicals and adsorbents in one group is similar to that in another. Also, following the adsorbent-based approach, when trying to predict the adsorption of phenol on a GAC, it is better to build models based on the adsorption data of phenol or phenol-like chemicals on GACs. To generalize the above finding, it is possible to improve the prediction accuracy by selecting part of the collected data that is similar to the prediction target rather than using all of the collected data. Indeed, further tests suggested that when the training and test/validation sets had similar ranges in the descriptor values, a smaller training set yielded a comparable or slightly better prediction for the test/validation set than that using a much larger training set (Text S6 and Figure S9).

Different methods in ML can be employed to quantify the similarity mentioned above;[60] however, many of them do not always apply to the physicochemical parameters of both the adsorbents and adsorbates in this research. For example, adding the distance (difference) between two pore volumes to the distance between two concentrations may not yield any physicochemically meaningful results. Commonly used distances such as Euclidean or City-block generally need to directly add up the distances of the parameters so they may not be suitable for adsorption prediction (Figure S10). Instead, the cosine distance/similarity that measures similarity by the angles between different data vectors was more suitable and thus selected for further research. To validate the cosine similarity-assisted data-preprocessing method, a cosine similarity cross-validation was performed to assess its effectivenes. Briefly, each group of the collected data was reserved as the validation set once. The cosine similarity between the validation set and the rest of the collected data was then calculated as the criterium to select some of the remaining groups as the training set. The Euclidean and City-block similarity methods were also employed for comparison. The training set was then fed to NN for training, and the prediction for the validation set was obtained based on the trained model. For comparison, the same number of groups as in the above training set was randomly selected as another training set (referred to as "random selection" below), and the obtained model was employed to predict the same validation set. We also employed the leave-one-out cross-validation approach for comparison because it is similar to the cosine similarity approach except that it used the entire data set minus the validation set as the training set (more details can be found in Text S7).

As shown in Figure 3a, the cosine similarity approach generally provided better predictions (the RMSEs were 0.31, 0.25, 0.28, and 0.23 for biochar, CNT, GAC, and resin separately) than the leave-one-out method (RMSEs 0.34, 0.27, 0.28, and 0.27 separately), which in turn was considerably improved compared to the random selection (RMSEs 0.47, 0.29, 0.34, and 0.36 separately), the Euclidean (RMSEs 0.38, 0.28, 0.34, and 0.24 separately), and the City-block (the RMSEs 0.38, 0.28, 0.35, and 0.24 separately) methods. Further discussion of the cosine similarity is in Text S6.

The cosine similarity method can indeed select the adsorption data that are the most relevant to a prediction target. The descriptor ranges of the training set being selected by the cosine similarity method were narrower than those by the leave-one-out and closer to those of the prediction target (Figure S11a), whereas the values of the descriptors by the leave-one-out approach varied within much wider ranges (Figure S11b). For instance, when the prediction target was the adsorption of phenol on MN200 (a polymeric resin with polystyrene backbones), the adsorbates in the training set selected by the cosine similarity approach included 2-nitroaniline, 2-nitrophenol, 4-chloroaniline, 4-chlorophenol, 4-methylephenol, 4-nitroaniline, aniline, catechol, chlorobenzene, naphthalene, nitrobenzene, phenol, resorcinol, and salicylic acid, all of which were aromatic, whereas the adsorbents selected were MN200, XAD-4, CHA-111, MCH-111, NDA-150, NDA-100, ZK-1, AH-1, HJ-01, GQ-06, and HJ-11, all of which were polystyrene-based resins[61−66] and shared similar properties with MN200. As expected, the adsorption on a polyacrylic ester resin XAD-7 was not selected into the training set because of the difference in the backbones of MN200 and XAD-7 and different adsorption behaviors of the chemicals on these two resins. The combination of the resins and the chemicals selected as the training set validated the effectiveness of the cosine similarity method. As for the prediction performance, the RMSE for the adsorption of phenol on MN200 based on the above data set was 0.33 versus 0.44 for the leave-one-out approach.

Model Performance. The optimized NN-LFER models achieved high accuracy when predicting the adsorption of 165 chemicals onto 50 biochars, 34 CNTs, 35 GACs, and 30 resins (Text S7 and Figure 4). Compared with the published results, the overall prediction in this work was significantly improved even though more data points were included for each type of the adsorbents, with the RMSEs ranging from 0.23 to 0.31 log unit and Q^2 ranging from 0.86 to 0.91 (Table 1). Besides, a considerable amount of the prediction deviations was caused by several outliers (Figure 4). When the top 10 deviated predictions were removed (Figure S13), the RMSEs were improved to 0.24, 0.20, 0.22, and 0.18 for biochar, CNTs, GAC, and resins, respectively.

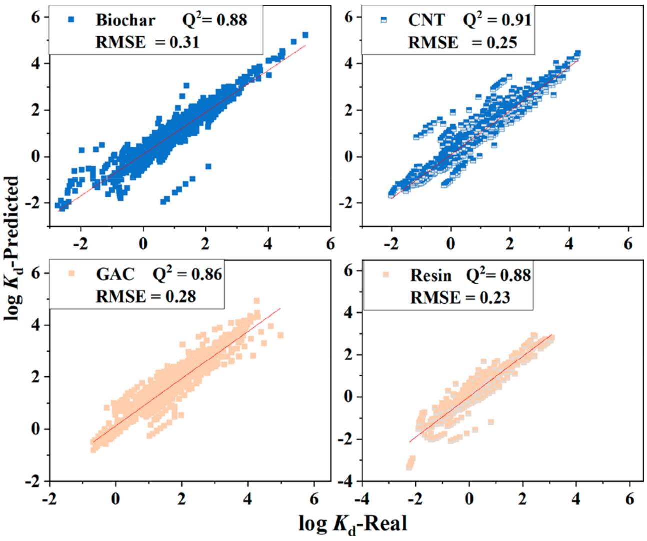


Figure 4. Performance of the NN-LFER models for the selected adsorbents. For each group of input data, the log K_d predicted was obtained from the NN-LFER models based on the cosine similarity approach. There are 952, 1316, 903, and 798 data points for biochar, CNTs, GAC, and resin, respectively, in this plot.

Table 1. Comparison of the Prediction Performance between This Work and Published Results[a]

adsorbent	predicted value	method	Q^2	RSME	ref
biochar	log K_d (N = 128)	pp-LFER	0.85	0.41(SE)	68
	log K_d (N = 11)	pp-LFER	–	0.31	14
CNTs	SW log $K_{d,0.001Sw}$ (N = 30)	pp-LFER	0.88	0.51	13
	MW log $K_{d,0.001Sw}$ (N = 83)	pp-LFER	0.86	0.23	
	log K_{SA} (N = 30)	QSPR + SVM	0.83	0.45	67
GAC	log $K_{d,0.001Sw}$ (N = 89)	pp-LFER	0.83	0.35	13
	log K_d (N = 210)	pp-LFER	0.84	0.28(SE)	22
resin	log C_e(N = 180)	pp-LFER	–	0.97	69
	log C_e(N = 160)		–	0.76	
biochar	log K_d N = 952	NN-LFER	0.87	0.31	this work
CNTs	N = 1316		0.91	0.25	
GAC	N = 903		0.86	0.28	
resins	N = 798		0.88	0.23	

[a] N = number of data points used for building the models; SW = single-walled; MW = multiwalled; S_w = water solubility; SE = standard error; K_{SA} = K_d/BET.

For a fair comparison based on the same data set, the reported SVM-assisted prediction[67] was repeated with our NN-LFER model for the CNTs (Text S7), and a better prediction was obtained with the RMSE decreased from 0.45 to 0.30 (Q^2 increased from 0.83 to 0.85, Figure S14).

Among the adsorbents, the biochars had a slightly higher RMSE than the CNTs, GACs, and resins. A possible reason is that biochars contain larger fractions of impurities including ash and organic matter,[70] and the portions of the impurities vary widely from one biochar to another, which complicates the interactions between the adsorbates and the adsorbents. Biochars in different studies have also been produced with different raw materials under varying conditions (such as temperature, time, and activation methods); as a result, they may have more drastically different properties such as pore size

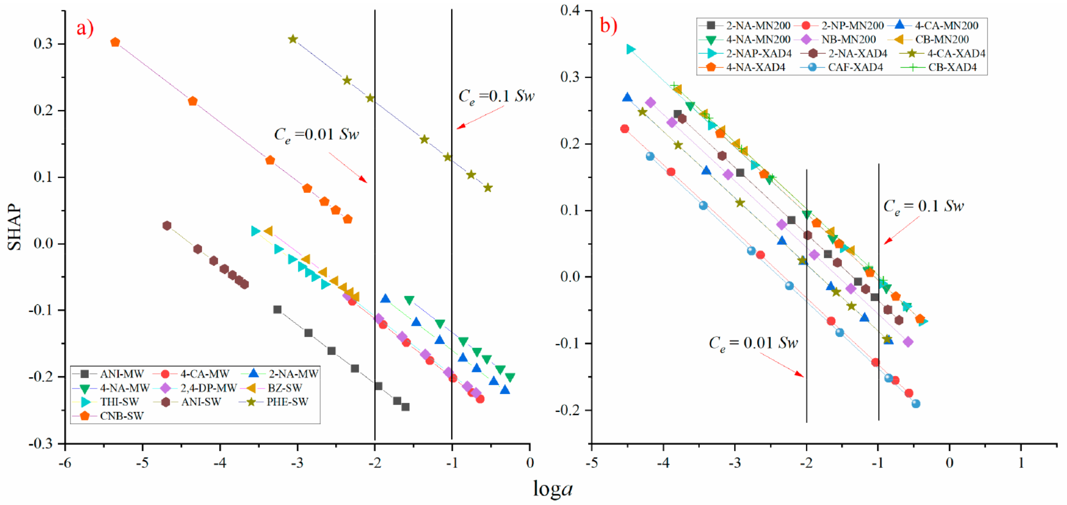


Figure 5. Shapely values at different log *a* (*a* = C_e/S_w) values for the adsorption of different compounds on (a) CNTs and (b) resins. ANI, aniline; CA, 4-chloroaniline; BZ, benzene; DP, 2,4-dichlorophenol; NA, 4-nitroaniline; CB, chlorobenzene; CAF, caffeine; NAP, 2-naphthol; NP, 3-nitrophenol; SW, single-walled; MW, multiwalled; MN200 and XAD-4 are two types of commercial resins. The vertical black lines indicate two concentration levels of 0.01 × and 0.1 × S_w. The *y* values at the intercepts of the vertical lines and the sloped lines are the Shapley values for the chemicals under the concentration levels of 0.01 × or 0.1 × S_{w}.

distribution and surface functional groups, many of which were not included in the modeling because they had not been reported in the literature.

In this research, only two of the most widely used physical properties, namely BET and V_t, were included in the models so we could build the data set as large as possible. Although these properties could only partly reflect all of the behaviors of the adsorbents, the predictions were significantly improved. Also, additional properties of the adsorbents, if available, can be added to the models to further improve the prediction accuracy. For example, the RMSE of the prediction for resins was reduced from 0.224 to 0.198 (the Q^2 increased from 0.91 to 0.92) by dividing the total pore volume into micro- and macropore volumes. In short, the modeling approach in this research provided a powerful, adaptable framework for building predictive models, which can easily incorporate more descriptors for adsorbents and adsorbates to improve the prediction accuracy. In contrast, such an adaptability cannot be readily achieved by the MLR models or by simply applying different ML algorithms.

Mechanistic Interpretation. To help identify the most influential factors in the adsorption of chemicals on those adsorbents, the Shapley values (details in Text S8) were calculated to quantify the contributions of the input factors. The Shapley theory is based on the coalitional game theory and tells us how to fairly distribute the "payout" among the descriptors,[71] where the payout refers to the predicted log K_d in this case. This is especially suitable for quantifying the contributions that are not equal for all factors like the eight input descriptors in this case.[72] A more positive or negative Shapley value means that the descriptor has a larger positive or negative contribution to the log K_d value, and vice versa. The Shapely values were calculated for each descriptor at every data point (each point is the adsorption of one chemical to an adsorbent at a given C_e).[73−75] For each data point, there were eight Shapley values, one for each descriptor (examples are shown in Figure S15, and the collection of all of them is shown in Figure S16). Then, all the Shapley values for a descriptor were used to calculate its mean absolute Shapley (MAS) value (Figure 3b). The overall impact of each descriptor on the adsorption can now be quantified by the MAS values; the larger the MAS value, the more significant the factor in influencing the adsorption.

As shown in Figure 3b, the overall pattern is that the *E*, *B*, *V*, and BET are the most influential factors because they have the largest MAS values, but there are significant differences in the MAS values between the resins and the three types of carbon materials. For the resins, the *B*, *V*, and BET are the most critical factors for the predicted log K_d values. This result is consistent with the adsorption mechanisms for resins. Most resins are hydrophobic polymers; due to the hydrophobic and porous nature of the resins, the adsorption on resins generally happens through hydrophobic interactions and pore-filling, although the hydrogen bond-donating ability of chemicals, as described by the *B* term, cannot be neglected.[76−78] Indeed, our previous pp-LFER models for adsorption on resins have shown that the hydrogen bond-donating (*B*) and cavity energy (described by the *V* descriptor, which is related to pore-filling and hydrophobic interactions) play crucial roles in the adsorption of organic chemicals onto three neutral resins (XAD-4, XAD-7, and MN200).[76]

Among the carbon materials, GACs and CNTs share similar patterns that the *E*, *B*, *V*, and BET matter the most to the predicted log K_d values. This is because the majority of the GACs and CNTs consist of carbon with minor impurities, so similar interactions between the adsorbents and the adsorbates are expected. The published predictive models based on pp-LFERs have also revealed that the *E*, *B*, and *V* are the most important descriptors for adsorption prediction, as suggested by their largest regression coefficients in the MLR equations.[79,80] Note that the adsorbent properties are only implicitly considered in the published MLR models, where the same class of adsorbents (such as GACs) but with different properties are treated the same, and the properties of the adsorbents such as BET and V_t are not directly involved in the MLR models.

Almost all of the input descriptors can significantly influence the adsorption on the biochars. As discussed earlier, the biochar and GAC/CNT differ in that biochar generally contains a large amount of organic/inorganic impurities, which complicates the interactions between the biochars and the adsorbed chemicals. Among the descriptors, the E, S, B, V, and BET are the dominant ones for biochars, agreeing with the reported MLR models in which the S, B, and V are the most important (BET and V_t are not included in the reported MLR models).[11,68,81] The coefficient for E is not recognized as significant; this may be because the adsorption data are only available for 1 biochar and 14 chemicals.[68] The dependency on E may emerge if the data set is to be expanded in both chemicals and adsorbents.

The agreement between this and previous work indicates that the NN-LFER models are both robust and chemically meaningful for adsorption prediction. Besides, compared with the contribution of log C_e, the contributions of the other seven descriptors are much less (Figure 3b, inset), suggesting that log C_e is an important input descriptor for adsorption prediction.

With the results from the NN-LFER models, higher deviations of the published MLR models can be at least partly explained. For single-concentration-level MLRs, one simplification is that the contribution of log C_e to the log K_d values for different chemicals is the same so that the C_e term is eliminated from the regression equations. The contributions of different adsorbents to log K_d are also treated as identical, so the properties of the adsorbents are not included either. These assumptions have clear limitations, and their validity can be tested with the newly developed NN-LFER models. Toward this goal, the contribution of log C_e to the predicted log K_d was quantified as the Shapley values. If the contribution of log C_e is the same for different chemicals under a certain C_e (such as at $C_e = 0.01 \times S_w$ or $0.1 \times S_w$), the Shapley values for different chemicals under the same C_e should be the same. The results in Figure 5 instead indicated that the Shapley values at the same C_e mostly varied from chemical to chemical. These results, therefore, disapproved the first simplification that the contributions of log C_e for different chemicals under the same C_e were identical. Instead, it supports the modeling strategy in this research that considers the equilibrium concentration as a variable when building predictive models.

For the simplification that adsorbents with different BET and V_t are treated the same, it is only correct when the contributions of BET or V_t for different adsorbents have a single Shapley value. However, the calculations showed that their contributions varied within a broad range (Figure S16), so the adsorbent properties cannot be neglected in the adsorption prediction. This can explain part of the higher deviations in commonly used MLRs as the absence of adsorbent properties has failed to capture the differences in the adsorption among the adsorbents, even if they belong to the same class.[22,67,68]

Graphical User Interface Tool and Limitations of the Built Models. Generally, some basic knowledge about NN and coding in Matlab/Python is necessary to use a NN model, and additional time will be required to learn the necessary knowledge and codes on the terminal windows, which do not favor applications of these trained models. To solve this problem and make the models widely accessible and usable, a simple tool with a graphical user interface (GUI) that is run under the Matlab environment was developed based on the well-trained NN-LFER models (Text S9, Figure S17, and the source code in the Supporting Information). With such a tool, predicting aqueous adsorption on the four categories of the adsorbents becomes straightforward. Also, even for chemicals that undergo substantial acid/base dissociation, such as pentachlorophenol (pK_a = 4.7, at pH = 5.0, around 33% of the pentachlorophenol is in the neutral form),[82] we were still able to use the equilibrium concentration of the neutral species instead of the total concentration as the input to achieve a satisfactory prediction, e.g., an RMSE of 0.25 for the adsorption of pentachlorophenol onto three different biochars at pH 5.0.

Despite the major advances in the built models, they (like any model does) have three major limitations in their applications: (1) We need to know the Abraham descriptors for a given chemical to predict its adsorption on one of the adsorbents; however, many emerging chemicals do not have those descriptors reported. Thus, future work should focus on obtaining the descriptors for these chemicals. (2) Considering both the application scope of the built models and the availability of the experimental data, only BET and V_t were selected to describe the adsorbents. As a result, three separate models were built for biochars, CNTs, and GACs to achieve low prediction deviations. With more adsorption data becoming available, future research can try to include additional adsorbent descriptors such as surface elemental composition, micropore volume, and macropore volume to unify those three models so that a more broadly applicable model can be built for all carbon materials. (3) The built models are primarily for adsorption of neutral and partially ionizable chemicals. However, a large number of chemicals are ionic under environmental conditions, and electrostatic interactions are the major adsorption forces. Therefore, the built models cannot predict their adsorption. Future work may include two additional Abraham descriptors J^+ and J^- to account for the contributions of electrostatic interactions to extend the application of the models to ionic chemicals.

Environmental Implications. In this work, we have demonstrated that data preprocessing coupled with NN and pp-LFERs can build accurate, chemically meaningful, and tunable models for predicting aqueous adsorption. The established data-preprocessing approaches that can improve the model performance include adding descriptors of the adsorbents to the models, employing the group-selection approach in the model training, and relying on the cosine similarity approach for source data preprocessing. For a given ML modeling process, it is a common practice to improve the prediction by increasing the volume of the data, which may lead to many irrelevant data and is also hard to achieve if labor-intense experiments such as adsorption experiments are needed; however, the cosine similarity approach provides another option from the data-preprocessing perspective. This is especially useful when there are limited adsorption data available, which is also commonly the case in many other experimental sciences, and we can design a minimum number of new experiments to achieve better prediction accuracy.

By interpreting the built NN models based on the Shapley values, we further showed that it was necessary to include both the equilibrium concentration and the properties of the adsorbents to build predictive models in future research. The absence of these descriptors is a possible reason for high deviations by MLR models. Overall, this research has not only built easy-to-use models for aqueous adsorption prediction but also provided a powerful modeling approach that will make further improvement in prediction accuracy possible, both of which will substantially facilitate predictive modeling research on aqueous adsorption.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.0c02526.

More detailed explanation of the methods and training process in this study, the figures and tables mentioned in the main text, and additional figures and tables to support the training process (PDF)

Source data compiled for this research (XLSX)

Source code for the GUI prediction tool (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Huichun Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0002-5683-5117; Phone: (216) 368-0689; Email: hjz13@case.edu

Authors

Kai Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0003-4058-6512

Shifa Zhong − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.0c02526

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant CBET-1804708.

■ REFERENCES

(1) Guirado, G.; Ayllón, J. A. A Simple Adsorption Experiment. *J. Chem. Educ.* **2011**, *88* (5), 624−628.

(2) Piergiovanni, P. R. Adsorption kinetics and isotherms: a safe, simple, and inexpensive experiment for three levels of students. *J. Chem. Educ.* **2014**, *91* (4), 560−565.

(3) Bunmahotama, W.; Hung, W.-N.; Lin, T.-F. Predicting the adsorption of organic pollutants from water onto activated carbons based on the pore size distribution and molecular connectivity index. *Water Res.* **2015**, *85*, 521−531.

(4) Bunmahotama, W.; Lin, T. F.; Yang, X. Prediction of adsorption capacity for pharmaceuticals, personal care products and endocrine disrupting chemicals onto various adsorbent materials. *Chemosphere* **2020**, *238*, 124658.

(5) Saleh, T. A.; Adio, S. O.; Asif, M.; Dafalla, H. Statistical analysis of phenols adsorption on diethylenetriamine-modified activated carbon. *J. Cleaner Prod.* **2018**, *182*, 960−968.

(6) Ma, L.; Chen, Q.; Zhu, J.; Xi, Y.; He, H.; Zhu, R.; Tao, Q.; Ayoko, G. A. Adsorption of phenol and Cu (II) onto cationic and zwitterionic surfactant modified montmorillonite in single and binary systems. *Chem. Eng. J.* **2016**, *283*, 880−888.

(7) Jiang, N.; Shang, R.; Heijman, S. G.; Rietveld, L. C. Adsorption of triclosan, trichlorophenol and phenol by high-silica zeolites: Adsorption efficiencies and mechanisms. *Sep. Purif. Technol.* **2020**, *235*, 116152.

(8) Fu, Y.; Shen, Y.; Zhang, Z.; Ge, X.; Chen, M. Activated bio-chars derived from rice husk via one-and two-step KOH-catalyzed pyrolysis for phenol adsorption. *Sci. Total Environ.* **2019**, *646*, 1567−1577.

(9) Ersan, G.; Apul, O. G.; Karanfil, T. Predictive models for adsorption of organic compounds by Graphene nanosheets: comparison with carbon nanotubes. *Sci. Total Environ.* **2019**, *654*, 28−34.

(10) Zhu, T.; Chen, W.; Cheng, H.; Wang, Y.; Singh, R. P. Prediction of polydimethylsiloxane-water partition coefficients based on the pp-LFER and QSAR models. *Ecotoxicol. Environ. Saf.* **2019**, *182*, 109374.

(11) Su, P.-H.; Kuo, D. T. F.; Shih, Y.-h.; Chen, C.-y. Sorption of organic compounds to two diesel soot black carbons in water evaluated by liquid chromatography and polyparameter linear solvation energy relationship. *Water Res.* **2018**, *144*, 709−718.

(12) Abraham, M. H. Hydrogen bonding: XXVII. Solvation parameters for functionally substituted aromatic compounds and heterocyclic compounds, from gas—liquid chromatographic data. *Journal of Chromatography A* **1993**, *644* (1), 95−139.

(13) Yu, X.; Sun, W.; Ni, J. LSER model for organic compounds adsorption by single-walled carbon nanotubes: Comparison with multi-walled carbon nanotubes and activated carbon. *Environ. Pollut.* **2015**, *206*, 652−60.

(14) Davis, C. W.; Di Toro, D. M. Modeling Nonlinear Adsorption to Carbon with a Single Chemical Parameter: A Lognormal Langmuir Isotherm. *Environ. Sci. Technol.* **2015**, *49* (13), 7810−7817.

(15) Brusseau, M. L. The Influence of Molecular Structure on the Adsorption of PFAS to Fluid-Fluid Interfaces: Using QSPR to Predict Interfacial Adsorption Coefficients. *Water Res.* **2019**, *152*, 148−158.

(16) Ling, Y.; Klemes, M. J.; Steinschneider, S.; Dichtel, W. R.; Helbling, D. E. QSARs to predict adsorption affinity of organic micropollutants for activated carbon and *β*-cyclodextrin polymer adsorbents. *Water Res.* **2019**, *154*, 217−226.

(17) Metivier-Pignon, H.; Faur, C.; Cloirec, P. L. Adsorption of dyes onto activated carbon cloth: Using QSPRs as tools to approach adsorption mechanisms. *Chemosphere* **2007**, *66* (5), 887−893.

(18) Muliadi, Y. K.; Huang, S.; Zhang, D.; Shi, T.; Chen, L.; Mei, H. Accurate Prediction of the Adsorption Capabilities of Synthetic Organic Contaminants by Single-Walled Carbon Nanotubes. *ChemistrySelect.* **2019**, *4* (8), 2449−2452.

(19) Ghosh, S.; Ojha, P. K.; Roy, K. Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. *Chemosphere* **2019**, *228*, 545−555.

(20) Roy, J.; Ghosh, S.; Ojha, P. K.; Roy, K. Predictive quantitative structure−property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). *Environ. Sci.: Nano* **2019**, *6* (1), 224−247.

(21) Plata, D. L.; Hemingway, J. D.; Gschwend, P. M. Polyparameter linear free energy relationship for wood char−water sorption coefficients of organic sorbates. *Environ. Toxicol. Chem.* **2015**, *34* (7), 1464−1471.

(22) Zhao, Y.; Lin, S.; Choi, J.-W.; Bediako, J. K.; Song, M.-H.; Kim, J.-A.; Cho, C.-W.; Yun, Y.-S. Prediction of adsorption properties for ionic and neutral pharmaceuticals and pharmaceutical intermediates on activated charcoal from aqueous solution via LFER model. *Chem. Eng. J.* **2019**, *362*, 199−206.

(23) Brasquet, C.; Bourges, B.; Le Cloirec, P. Quantitative Structure−Property Relationship (QSPR) for the Adsorption of Organic Compounds onto Activated Carbon Cloth: Comparison between Multiple Linear Regression and Neural Network. *Environ. Sci. Technol.* **1999**, *33* (23), 4226−4231.

(24) Hsieh, C.-T.; Teng, H. Influence of mesopore volume and adsorbate size on adsorption capacities of activated carbons in aqueous solutions. *Carbon* **2000**, *38* (6), 863−869.

(25) Zhang, S.; Liu, X.; Karanfil, T. Applicability of the linear solvation energy relationships in the prediction for adsorption of aromatic compounds on activated carbons from aqueous solutions. *Sep. Purif. Technol.* **2013**, *117*, 111−117.

(26) Oleszczuk, P.; Pan, B.; Xing, B. Adsorption and Desorption of Oxytetracycline and Carbamazepine by Multiwalled Carbon Nanotubes. *Environ. Sci. Technol.* **2009**, *43* (24), 9167−9173.

(27) Yang, K.; Zhu, L.; Xing, B. Adsorption of Polycyclic Aromatic Hydrocarbons by Carbon Nanomaterials. *Environ. Sci. Technol.* **2006**, *40* (6), 1855−1861.

(28) Huang, X.; Yu, J.; Shi, B.; Hao, H.; Wang, C.; Jia, Z.; Wang, Q. Rapid prediction of the activated carbon adsorption ratio by a regression model. *Chemosphere* **2020**, *245*, 125675.

(29) Apul, O. G.; Karanfil, T. Adsorption of synthetic organic contaminants by carbon nanotubes: A critical review. *Water Res.* **2015**, *68*, 34−55.

(30) Lillo-Ródenas, M. A.; Cazorla-Amorós, D.; Linares-Solano, A. Behaviour of activated carbons with different pore size distributions and surface oxygen groups for benzene and toluene adsorption at low concentrations. *Carbon* **2005**, *43* (8), 1758−1767.

(31) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323* (6088), 533−536.

(32) Suykens, J. A. K. Support Vector Machines: A Nonlinear Modelling and Control Perspective. *European Journal of Control* **2001**, *7* (2), 311−327.

(33) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24* (2), 123−140.

(34) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels−Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58* (14), 4515−4519.

(35) Sahu, H.; Yang, F.; Ye, X.; Ma, J.; Fang, W.; Ma, H. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *J. Mater. Chem. A* **2019**, *7* (29), 17480−17488.

(36) Fan, M.; Hu, J.; Cao, R.; Xiong, K.; Wei, X. Modeling and prediction of copper removal from aqueous solutions by nZVI/rGO magnetic nanocomposites using ANN-GA and ANN-PSO. *Sci. Rep.* **2017**, *7* (1), 18040.

(37) Zhang, Z.; Schott, J. A.; Liu, M.; Chen, H.; Lu, X.; Sumpter, B. G.; Fu, J.; Dai, S. Prediction of Carbon Dioxide Adsorption via Deep Learning. *Angew. Chem.* **2019**, *131* (1), 265−269.

(38) Panapitiya, G.; Avendaño-Franco, G.; Ren, P.; Wen, X.; Li, Y.; Lewis, J. P. Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140* (50), 17508−17514.

(39) Franco, D. S.; Duarte, F. A.; Salau, N. P. G.; Dotto, G. L. Adaptive neuro-fuzzy inference system (ANIFS) and artificial neural network (ANN) applied for indium (III) adsorption on carbonaceous materials. *Chem. Eng. Commun.* **2019**, *206* (11), 1−11.

(40) Zhang, Y.; Pan, B. Modeling batch and column phosphate removal by hydrated ferric oxide-based nanocomposite using response surface methodology and artificial neural network. *Chem. Eng. J.* **2014**, *249*, 111−120.

(41) Chittoo, B. S.; Sutherland, C. Adsorption Using Lime-Iron Sludge−Encapsulated Calcium Alginate Beads for Phosphate Recovery with ANN-and RSM-Optimized Encapsulation. *J. Environ. Eng.* **2019**, *145* (5), 04019019.

(42) Abdel Rahman, R. O.; Abdel Moamen, O. A.; Abdelmonem, N.; Ismail, I. M. Optimizing the removal of strontium and cesium ions from binary solutions on magnetic nano-zeolite using response surface methodology (RSM) and artificial neural network (ANN). *Environ. Res.* **2019**, *173*, 397−410.

(43) Hassanzadeh, Z.; Kompany-Zareh, M.; Ghavami, R.; Gholami, S.; Malek-Khatabi, A. Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. *J. Mol. Struct.* **2015**, *1098*, 191−198.

(44) Rahimi-Nasrabadi, M.; Akhoondi, R.; Pourmortazavi, S. M.; Ahmadi, F. Predicting adsorption of aromatic compounds by carbon nanotubes based on quantitative structure property relationship principles. *J. Mol. Struct.* **2015**, *1099*, 510−515.

(45) Brasquet, C.; Le Cloirec, P. QSAR for organics adsorption onto activated carbon in water: what about the use of neural networks? *Water Res.* **1999**, *33* (17), 3603−3608.

(46) Sabour, M. R.; Movahed, S. M. A. Application of radial basis function neural network to predict soil sorption partition coefficient using topological descriptors. *Chemosphere* **2017**, *168*, 877−884.

(47) Shao, Y.; Liu, J.; Wang, M.; Shi, L.; Yao, X.; Gramatica, P. Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods. *Atmos. Environ.* **2014**, *88*, 212−218.

(48) Brasquet, C.; Bourges, B.; Le Cloirec, P. Quantitative structure− property relationship (QSPR) for the adsorption of organic compounds onto activated carbon cloth: Comparison between multiple linear regression and neural network. *Environ. Sci. Technol.* **1999**, *33* (23), 4226−4231.

(49) Özdemir, U.; Özbay, B.; Veli, S.; Zor, S. Modeling adsorption of sodium dodecyl benzene sulfonate (SDBS) onto polyaniline (PANI) by using multi linear regression and artificial neural networks. *Chem. Eng. J.* **2011**, *178*, 183−190.

(50) Timofei, S.; Kurunczi, L.; Suzuki, T.; Fabian, W. M.; Mureşan, S. Multiple Linear Regression (MLR) and Neural Network (NN) calculations of some disazo dye adsorption on cellulose. *Dyes Pigm.* **1997**, *34* (3), 181−193.

(51) Wang, B.; Chen, J.; Li, X.; Wang, Y. n.; Chen, L.; Zhu, M.; Yu, H.; Kühne, R.; Schüürmann, G. Estimation of Soil Organic Carbon Normalized Sorption Coefficient (Koc) Using Least Squares-Support Vector Machine. *QSAR Comb. Sci.* **2009**, *28* (5), 561−567.

(52) Sigmund, G.; Gharasoo, M.; Hüffer, T.; Hofmann, T. Deep Learning Neural Network Approach for Predicting the Sorption of Ionizable and Polar Organic Pollutants to a Wide Range of Carbonaceous Materials. *Environ. Sci. Technol.* **2020**, *54* (7), 4583−4591.

(53) Ayawei, N.; Ebelegi, A. N.; Wankasi, D. Modelling and Interpretation of Adsorption Isotherms. *J. Chem.* **2017**, *2017*, 1−11.

(54) Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493−2537.

(55) Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. *A discriminative feature learning approach for deep face recognition*; European Conference on Computer Vision, Amsterdam, The Netherlands, October 11−14, 2016; Springer: Cham, Switzerland, 2016; pp 499−515.

(56) Todeschini, R.; Ballabio, D.; Grisoni, F. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf. Model.* **2016**, *56* (10), 1905−1913.

(57) Mignan, A.; Broccardo, M. One neuron versus deep learning in aftershock prediction. *Nature* **2019**, *574* (7776), E1−E3.

(58) Porumb, M.; Iadanza, E.; Massaro, S.; Pecchia, L. A convolutional neural network approach to detect congestive heart failure. *Biomedical Signal Processing and Control* **2020**, *55*, 101597.

(59) Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D. C.; Kording, K. P. Voodoo machine learning for clinical predictions. *Biorxiv* **2016**, 059774.

(60) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **2019**, *10* (1), 539.

(61) Pan, B.; Zhang, X.; Zhang, W.; Zheng, J.; Pan, B.; Chen, J.; Zhang, Q. Adsorption of phenolic compounds from aqueous solution onto a macroporous polymer and its aminated derivative: isotherm analysis. *J. Hazard. Mater.* **2005**, *121* (1−3), 233−241.

(62) Huang, J.; Jin, X.; Deng, S. Phenol adsorption on an N-methylacetamide-modified hypercrosslinked resin from aqueous solutions. *Chem. Eng. J.* **2012**, *192*, 192−200.

(63) Zheng, K.; Pan, B.; Zhang, Q.; Zhang, W.; Pan, B.; Han, Y.; Zhang, Q.; Wei, D.; Xu, Z.; Zhang, Q. Enhanced adsorption of p-nitroaniline from water by a carboxylated polymeric adsorbent. *Sep. Purif. Technol.* **2007**, *57* (2), 250−256.

(64) Long, C.; Li, A.; Wu, H.; Liu, F.; Zhang, Q. Polanyi-based models for the adsorption of naphthalene from aqueous solutions

onto nonpolar polymeric adsorbents. *J. Colloid Interface Sci.* **2008**, *319* (1), 12−18.

(65) Sun, Y.; Chen, J.; Li, A.; Liu, F.; Zhang, Q. Adsorption of resorcinol and catechol from aqueous solution by aminated hypercrosslinked polymers. *React. Funct. Polym.* **2005**, *64* (2), 63−73.

(66) Pan, B.; Zhang, H. Reconstruction of adsorption potential in Polanyi-based models and application to various adsorbents. *Environ. Sci. Technol.* **2014**, *48* (12), 6772−6779.

(67) Wang, Q. L.; Apul, O. G.; Xuan, P.; Luo, F.; Karanfil, T. Development of a 3D QSPR model for adsorption of aromatic compounds by carbon nanotubes: comparison of multiple linear regression, artificial neural network and support vector machine. *RSC Adv.* **2013**, 3 (46), 23924−23934.

(68) Plata, D. L.; Hemingway, J. D.; Gschwend, P. M. Polyparameter linear free energy relationship for wood char-water sorption coefficients of organic sorbates. *Environ. Toxicol. Chem.* **2015**, *34* (7), 1464−71.

(69) Jadbabaei, N.; Zhang, H. Sorption Mechanism and Predictive Models for Removal of Cationic Organic Contaminants by Cation Exchange Resins. *Environ. Sci. Technol.* **2014**, *48* (24), 14572−14581.

(70) Lehmann, J.; Rillig, M. C.; Thies, J.; Masiello, C. A.; Hockaday, W. C.; Crowley, D. Biochar effects on soil biota − A review. *Soil Biol. Biochem.* **2011**, *43* (9), 1812−1836.

(71) Winter, E. The shapley value. *Handbook of game theory with economic applications* **2002**, *3* (2), 2025−2054.

(72) Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games* **1953**, *2* (28), 307−317.

(73) Stojić, A.; Stanić, N.; Vuković, G.; Stanišić, S.; Perišić, M.; Šoštarić, A.; Lazić, L. Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. *Sci. Total Environ.* **2019**, *653*, 140−147.

(74) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv, Neur. Inf. Proc. Syst.* **2017**, 4765−4774.

(75) Zhao, Y.; Wang, L.; Luo, J.; Huang, T.; Tao, S.; Liu, J.; Yu, Y.; Huang, Y.; Liu, X.; Ma, J. Deep Learning Prediction of Polycyclic Aromatic Hydrocarbons in the High Arctic. *Environ. Sci. Technol.* **2019**, *53* (22), 13238−13245.

(76) Pan, B.; Zhang, H. Interaction Mechanisms and Predictive Model for the Sorption of Aromatic Compounds onto Nonionic Resins. *J. Phys. Chem. C* **2013**, *117* (34), 17707−17715.

(77) Pan, B.; Zhang, H. A modified Polanyi-based model for mechanistic understanding of adsorption of phenolic compounds onto polymeric adsorbents. *Environ. Sci. Technol.* **2012**, *46* (12), 6806−6814.

(78) Pan, B.; Zhang, H. Reconstruction of adsorption potential in Polanyi-based models and application to various adsorbents. *Environ. Sci. Technol.* **2014**, *48* (12), 6772−6779.

(79) Shih, Y. H.; Gschwend, P. M. Evaluating activated carbon-water sorption coefficients of organic compounds using a linear solvation energy relationship approach and sorbate chemical activities. *Environ. Sci. Technol.* **2009**, *43* (3), 851−857.

(80) Apul, O. G.; Wang, Q.; Shao, T.; Rieck, J. R.; Karanfil, T. Predictive model development for adsorption of aromatic contaminants by multi-walled carbon nanotubes. *Environ. Sci. Technol.* **2013**, *47* (5), 2295−2303.

(81) Lu, Z.; MacFarlane, J. K.; Gschwend, P. M. Adsorption of organic compounds to diesel soot: frontal analysis and polyparameter linear free-energy relationship. *Environ. Sci. Technol.* **2016**, *50* (1), 285−293.

(82) Peng, P.; Lang, Y.-H.; Wang, X.-M. Adsorption behavior and mechanism of pentachlorophenol on reed biochars: pH effect, pyrolysis temperature, hydrochloric acid treatment and isotherms. *Ecological Engineering.* **2016**, *90*, 225−233.