Chemical Engineering Journal 405 (2021) 126627

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Shedding light on "Black Box" machine learning models for predicting the reactivity of HO· radicals toward organic compounds

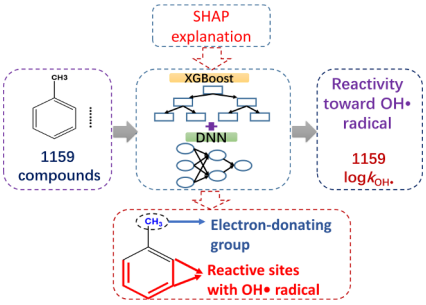
Shifa Zhong[a], Kai Zhang[a], Dong Wang[b], Huichun Zhang[a,*]

[a] Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Cleveland, OH 44106, USA
[b] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

HIGHLIGHTS

- MF-ML assisted-QSAR model was developed for 1089 compounds toward HO· reactivity.
- An ensemble model that combined XGBoost and DNN was developed.
- The SHAP method was used to interpret all the obtained models.
- The model made predictions based on the chemical knowledge correctly "learned".

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Deep neural network
Machine learning
Model interpretation
HO· radical
QSARs
XGBoost

ABSTRACT

Developing quantitative structure-activity relationships (QSARs) is an important approach to predicting the reactivity of HO radicals toward newly emerged organic compounds. As compared with molecular descriptors-based and the group contribution method-based QSARs, a combined molecular fingerprint-machine learning (ML) method can more quickly and accurately develop such models for a growing number of contaminants. However, it is yet unknown whether this method makes predictions by choosing meaningful structural features rather than spurious ones, which is vital for trusting the models. In this study, we developed QSAR models for the $\log k_{HO}$. values of 1089 organic compounds in the aqueous phase by two ML algorithms—deep neural networks (DNN) and eXtreme Gradient Boosting (XGBoost), and interpreted the built models by the SHapley Additive exPlanations (SHAP) method. The results showed that for the contribution of a given structural feature to $\log k_{HO}$. for different compounds, DNN and XGBoost treated it as a fixed and variable value, respectively. We then developed an ensemble model combining the DNN with XGBoost, which achieved satisfactory predictive performance for all three datasets: Training dataset: R-square (R^2) 0.89–0.91, root-mean-squared-error (*RMSE*) 0.21–0.23, and mean absolute error (*MAE*) 0.15–0.17; Validation dataset: R^2 0.63–0.78, *RMSE* 0.29–0.32, and *MAE* 0.21–0.25; and Test dataset: R^2 0.60–0.71, *RMSE* 0.30–0.35, and *MAE* 0.23–0.25. The SHAP method was further used to unveil that this ensemble model made predictions on $\log k_{HO}$. based on a correct 'understanding' of the impact of electron-withdrawing and -donating groups and of the reactive sites in the compounds that can be attacked by HO·. This study offered some much-needed mechanistic insights into a ML-assisted

Abbreviations: BPSO, binary particle swarm optimization; DTB, decision tree boost; GCM, Group contribution method; MDs, Molecular descriptors; ML, Machine learning; MSE, mean-squared-error; QSARs, Quantitative structure-activity relationships; SAR, structure-activity relationship; XGBoost, eXtreme Gradient Boosting; DNN, Deep neural network; GA, genetic algorithm; MAE, mean-absolute-error; MFs, Molecular fingerprints; MLR, multiple linear regression; PCA, principal component analysis; RMSE, root-mean-squared-error; SHAP, SHapley Additive exPlanations

* Corresponding author.
E-mail address: hjz13@case.edu (H. Zhang).

https://doi.org/10.1016/j.cej.2020.126627
Received 12 May 2020; Received in revised form 7 August 2020; Accepted 8 August 2020
Available online 12 August 2020
1385-8947/

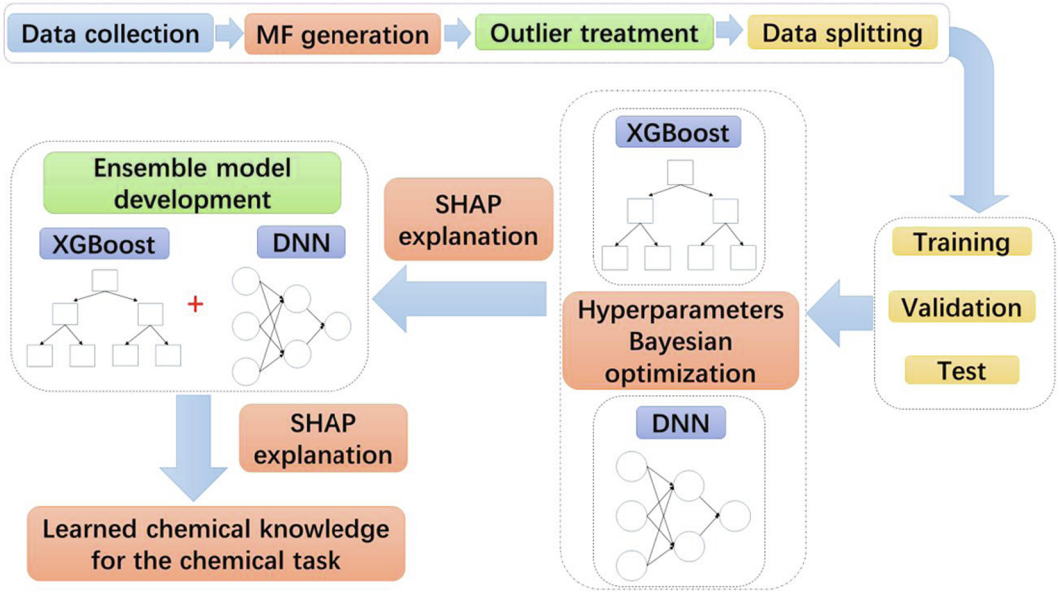
environmental task, which are important for evaluating the trustworthiness of the ML-based models, further improving the models for specific applications, and leveraging the implicit knowledge the models carry.

1. Introduction

Quantitative structure − activity relationships (QSARs) correlate chemical reactivity with chemical and/or structural features of many chemicals [1,2] and can be used to predict the reactivity of new compounds that otherwise needs labor-intensive and expensive experiments. The development of QSARs for HO· radicals has been a long-standing interest [3–5] because the HO· radical, a ubiquitous strong oxidant, plays important roles in natural and engineered waters, the atmosphere, biological systems, and even interstellar space by unselectively and instantaneously reacting with numerous organic and inorganic compounds [6–18]. With more and more toxic organic contaminants released into water environments, a simple, fast and accurate approach is desirable for developing QSAR models to predict the second-order rate constants ($k_{HO\cdot}$) for thousands of organic contaminants.

However, traditional approaches including molecular descriptors (MDs)-based QSARs [3] and the group contribution method (GCM) [19–21] have their own limitations. MDs represent certain molecular physicochemical properties, such as Hammett constants, reduction potential, and topological polar surface area. MDs have to be calculated by advanced computation that relies greatly on one's sophisticated physicochemical knowledge and ability to use different software. For example, Borhani et al. believed that a two-step optimization of the chemical structures should be completed before calculating the MDs, and Density Functional Theory calculations were necessary for obtaining quantum-chemically calculated descriptors [3]. MD-based QSARs often have satisfactory predictive performance when the number of compounds involved is small. For instance, Wang et al. used 4 MDs with multiple linear regression to develop a QSAR for 55 compounds and achieved a low *RMSE* of 0.139 [22]. However, the obtained QSARs often have a small applicability domain that is only applicable to a limited number of compounds. With more and more contaminants involved, the calculation of MDs becomes time consuming while, generally, involving more MDs but this still achieves less accurate QSAR models [23]. For example, Borhani et al. used 8 MDs to develop a binary particle swarm optimization-multiple linear regression model for 457 compounds and achieved a higher *RMSE* of 0.356 [3]. Also, to develop QSAR models with a satisfactory prediction performance, a number of MDs should be manually selected among thousands of MDs but some of them are difficult to understand, that is, the physicochemical meanings of these MDs cannot be easily linked to the HO radical reactivity. Nevertheless, one major advantage of MD-based QSAR models is that they are interpretable because of the chemical meanings carried by the MDs. The word "interpretable" means one usually knows why a model predicts high log$k_{HO\cdot}$ values for some compounds but not others.

GCM and structure–activity relationships (SAR) are another approach to developing QSARs by hypothesizing that a rate constant of a given organic compound is a combined rate constant of all the structural features [19–21]. Specifically, GCM quantifies HO radical rate constants based on the contributions of four reaction mechanisms (when applicable): (1) H-atom abstraction, (2) HO· addition to alkenes, (3) HO· addition to aromatic compounds, and (4) HO· interaction with sulfur (S)-, nitrogen (N)-, or phosphorus (P)-atom-containing compounds [19,20]. Monod et al. [21] combined SAR with linear regression to develop a QSAR model to predict the HO-oxidation rate constants for 72 aliphatic compounds. The results showed a satisfactory performance as 60% of the estimated values were within the range of 80% of the experimental values. Note that they did not employ metrics such as *RMSE* to evaluate the performance so it is difficult to compare it with MD-based QSARs. Later, they expanded this method to 102 carbonyl compounds with 252 experimental rate constants and the accuracy of this updated SAR was such that 58% of the rate constants were calculated within ± 20% of the experimental data and 76% within ± 40% [24]. Minakata et al. used GCM with genetic algorithms to develop a QSAR model for 434 compounds and 62% of the test compounds were predicted within 0.5–2 times of the experimental values [20]. The drawbacks of GCM or SAR are that: (1) For compounds whose reaction mechanisms are beyond the four mechanisms, GCM may be unreliable and inaccurate [20]; (2) GCM only linearly combines the contributions of different groups for a compound. When thousands of compounds are involved, non-linear correlations may exist, that is, the contribution of a structural group to the reactivity of different compounds may differ, but



Scheme 1. The workflow of the data construction and preprocessing, model development, SHAP explanation, and the final goal.

GCM cannot well capture the differences. Nevertheless, the reaction mechanisms-based GCM is also interpretable.

We have recently used molecular fingerprints (MFs) as the inputs for a deep neural network (DNN) to develop a QSAR model for the $k_{HO\cdot}$ of 500+ compounds [23]. It showed a comparative prediction performance with the MDs- and GCM-based QSARs but can be more readily built for a large number of compounds ($\log k_{HO\cdot}$ for new compounds can be predicted within a millisecond). As compared with MDs and the GCM methods, MFs can be more easily obtained and understood than MDs; obtaining the MFs for thousands of compounds is also faster than obtaining atom groups in GCM. There is no need to manually choose MDs with complex physicochemical meanings. However, MFs carry little mechanistic information and only encode the structural information of the molecules including atoms, bonds, and functional groups as binary vectors containing 0 s and 1 s (more details of MFs in Text S1 and Fig. S1 in the Supplementary Material) [23,25,26]. Although the DNN-MF-based approach significantly simplifies and accelerates the development of QSARs for a large amount of organic compounds with a satisfactory prediction performance, the built QSARs are not yet interpretable, that is, one usually has little idea about how a DNN makes its predictions based on the MF binary vectors. The motivation question for this work therefore was "Does a DNN combined with MFs make predictions for rate constants based on a correct 'understanding' of important structural features for the reactivity?" which has never been addressed. Answering this question can offer a theoretical support for MF-based QSAR models. If a model makes predictions based on spurious features, then we cannot fully trust the model even though it can be easily built with high accuracy. For example, McCloskey et al. used the attribution method to interpret that DNN still learned spurious binding logics despite its perfect classification accuracy on the protein–ligand binding dataset [27], while Belzen et al. revealed that the perfect classification accuracy achieved on a well-trained DNN was based on the correct identification of the specific protein sequence associated with the biological functions [28].

In this study, we employed a large dataset that covered the reported $k_{HO\cdot}$ values for 1159 organic compounds in the aqueous phase (Scheme 1). Based on this dataset, the effects of ML algorithms, data splitting approaches, and outliers on the prediction performance were investigated. XGBoost and DNN were then chosen to develop QSAR models (Scheme 1). Their hyperparamters and the radius and length of MFs were optimized by the Bayesian optimization algorithm. Next, we used the recently developed SHapley Additive exPlanations (SHAP) method [29] to interpret the two ML models about what features (i.e., atom groups) were selected to make the predictions. A brief introduction to DNN, XGBoost, and SHAP is provided in Text S2. The SHAP method is theoretically sound as compared with other interpretation methods (Text S3). It has been applied to interpret the predictions made in gene expression, the concentrations of polycyclic aromatic hydrocarbons in the high arctic, and wet deposition of toluene, ethylbenzene and xylene [30–32]. We then developed an ensemble ML model that combined the DNN with XGBoost and interpreted it by the SHAP method. Interestingly, the developed ensemble model acquired the most relevant chemical knowledge to accurately predict $\log k_{HO\cdot}$, including the influence of electron-donating and -withdrawing groups and the reactive sites. The performance of the ensemble model was also compared with the previously well-established QSARs.

2. Methods

2.1. Dataset, preprocessing and MF generation

A dataset containing 1159 organic compounds and their HO· radical rate constants ($k_{HO\cdot}$, $M^{-1}s^{-1}$) obtained under standard conditions (i.e., 25 °C) in the aqueous phase was created after reviewing relevant journal articles [3,23,33–41]. A detailed description of the dataset was supplied in Text S4. MFs of all the compounds were generated by the RDKit program within milliseconds by converting their "SMILES" strings obtained by the ChemDraw program, which is impossible for their MDs to be calculated in such a short time.

Instead of randomly splitting the dataset into the training, validation, and test datasets, we relied on the compound structures to split the dataset. This was to maximize the diversity in the compound structures in the training dataset and ensure similarities among the training, validation, and test datasets. A detailed explanation of why we should value these two properties was provided in Text S5. According to this principle, we grouped all the organic compounds based on their functional groups. It should be noted that this grouping method was not based on single functional groups because more than 1 functional group exists in most compounds. Instead, it was based on a combination of different functional groups. For instance, the class of "OH" represents simple alcohol species while the class of "OH-X" represents the species that contain both OH and halogen groups. Another important reason for this functional groups-based data splitting was to identify outliers. Without grouping, it was difficult to identify whether the reactivity of a compound was "abnormal" or not. However, with grouping the compounds with the same functional groups should have similar reactivity. Hence, we can have more confidence in identifying outliers when some compound show abnormal reactivity as compared with others in the same group.

Based on all the functional groups, the 1159 organic compounds were first divided into 357 classes. However, 250 of the 357 classes contained less than 3 compounds and could not be divided into three subsets of training, validation, and test datasets. Hence, we merged the classes containing less than 3–4 compounds with other larger groups based on similarity in the functional groups to form 98 classes. For every class, we identified the outliers based on the corresponding boxplot of $\log k_{HO\cdot}$ (not shown). Any compounds with a $\log k_{HO\cdot}$ value outside the range of (Q1 − 1.5 × IQR) to (Q3 + 1.5 × IQR) were excluded as the outliers (Q1 and Q3: 25% and 75% quartile, interquartile range IQR: (Q3 − Q1)). In this way, 70 compounds (6%) were removed from the dataset and the final number of compounds was 1089. These outliers might result from experimental errors ('real' outliers) or the merging operation as stated above ('false' outliers). For example, 1,4-diaminobutane was identified as an outlier in the class of amine. This outlier was 'real' because it was assigned to the correct class. However, for 1,3-dithiolane-2-thione, we merged it into sulfoxides, which is obviously not reasonable, so this outlier might result from the incorrect grouping ('false outliers') and/or experimental errors. To investigate the effect of these outliers, three datasets containing no outliers ($D_{FG\text{-}98}$), 'false outliers' ($D_{FG\text{-}98\text{-}fo}$), and all outliers ($D_{FG\text{-}98\text{-}all}$) were established and compared. The final dataset including the names, structures and $\log k_{OH\cdot}$ values of all the compounds is summarized in the excel file named "Dataset.xlsx" in the Supplementary Material. $D_{FG\text{-}98}$ was used in all modeling processes unless otherwise specified.

After grouping, for each dataset ($D_{FG\text{-}98}$, $D_{FG\text{-}98\text{-}fo}$, and $D_{FG\text{-}98\text{-}all}$) we randomly chose compounds from every class with a ratio of 8:1:1 and combined them to form the training, validation, and test datasets. The specific information about which compounds were in the training, validation or test dataset can be found in the excel file "Dataset.xlsx". A validation dataset is necessary to control the overfitting problem and optimize the hyperparameters; otherwise, DNN can perfectly fit the data points in the training dataset but poorly predict the samples in the test dataset, i.e., overfitting. In this way, the diversity of the compound structures was similar in the training, validation and test datasets. Following this approach, each of the three datasets was split five times to form five different sets of training-validation-test sub-datasets, which were each used independently to train, validate, and test the ML models. In other words, we obtained 5 models for each dataset. This was to check if the performance of the ML models was a coincidental result of the data splitting or not.

To prove that the above grouping approach was reasonable, the dataset that contained no outliers ($D_{FG\text{-}98}$) was split in three additional

ways. "D_{random}" refers to a random split of the dataset into training-validation-test sets (8:1:1). "D_{FG-28}" and "D_{FG-63}" refer to splitting the compounds into 28 and 63 classes, respectively, also based on functional groups but more broadly. Each dataset was further split five times to form five sets of training-validation-test sub-datasets, each containing at least one compound for each functional group. Obviously D_{FG-98} grouped the organic compounds more finely than D_{FG-28} and D_{FG-63}. All the constructed datasets are summarized in Table S1.

2.2. Baseline ML algorithm selection

To choose a baseline ML algorithm to compare with DNN, we applied 11 common ML algorithms with their default parameters to the same dataset and obtained their prediction performances (root mean squared errors on the test dataset: $RMSE_{test}$). $RMSE_{test}$ was used as an indicator to evaluate the prediction performance: the lower the $RMSE_{test}$ value is, the better prediction ability the model has. As the results shown in Fig. S3, Random forest, Bagging, Gradient Boosting, and XGBoost showed similar $RMSE_{test}$ and were superior to the other ML methods. Here, we chose XGBoost as the baseline model because of its fast calculation and better design to control overfitting [42].

2.3. Bayes optimization of the hyperparameters of XGBoost, DNN and MFs

Before the training process, there were several parameters called "hyperparameters" whose values needed to be determined. Tuning hyperparameters is a necessary step to optimize the performance of any ML method. Besides the hyperparameters of XGBoost and DNN, the length and radius of MFs should also be carefully tuned. Table S2 summarizes the names and roles of the hyperparameters and the ranges of their values. Because there was a large number of hyperparameters whose values varied over wide ranges, it was impossible to enumerate every value to find out the optimum ones. We thus employed the powerful Bayesian optimization algorithm, which can choose the next optimum hyperparameter candidate value based on the results obtained from the previous ones [43,44]. Hence, the possibility of achieving the optimum values of the hyperparameters was maximized.

For both XGBoost and DNN, we used the same dataset D_{FG-98} to optimize the hyperparameter values. After 500 iterations of the Bayesian optimization (i.e., choosing 500 groups of hyperparameter values), the mean squared errors (MSE) of the training and validation datasets by the XGBoost and DNN were obtained and are plotted in Fig. S4. It should be noted that in each iteration, "early stopping" was used to control overfitting during the training process. The "early stopping" method monitors changes in the MSE of the validation dataset ($MSE_{validation}$). If there was no change or an increase in the $MSE_{validation}$ after a preset number of epochs, the training process stopped. The dotted lines in Fig. S4 indicate the lowest $MSE_{validation}$ position for both the XGBoost and DNN. With decreasing MSE_{train}, the $MSE_{validation}$ decreased, reached the minimum value, and then increased. Hence, underfitting and overfitting problems existed to the right and left hand sides of the dotted line, as shown in Fig. S4. The optimum values of the hyperparameters were the ones that achieved the minimum $MSE_{validation}$, as listed in Table S3.

2.4. Applicability domain

Applicability domain aims to evaluate if a trustful prediction can be made for a query compound. Applicability domain is obtained by comparing the similarity between the query compound and the compounds in the training dataset. The similarity between two compounds A and B was calculated based on the Tanimoto index, T(A, B), according to Eq. (1):

$$T(A, B) = \frac{N_c}{N_a + N_b - N_c} \quad (1)$$

where Nc is the number of 1 s in the MFs of both compounds A and B; Na is the number of 1 s in the MF of compound A and Nb is the number of 1 s in the MF of compound B. Because MFs are binary vectors with a pre-set length filled with 0 s and 1 s, in which only 1 s represent there exist atom groups and the positions of 1 s represent what atom groups are in the compound, the Tanimoto index calculates the percent of the same atom groups that are in both compounds A and B, thus comparing their similarity. If the similarity is above a pre-set threshold, the query compound is within the applicability domain and the prediction made on this compound is reliable [45–47]; otherwise, the query compound is outside the applicability domain and the prediction is not reliable. The threshold value of similarity was determined by comparing the similarity between the compounds in the test dataset and those in the training dataset. To obtain the similarity of one compound in the test dataset to the ones in the training dataset, we calculated the similarity between this compound and every compound in the training dataset. Two similarity metrics were then used: the maximum similarity refers to the maximum value among all the obtained similarity values, while the mean similarity refers to the mean of these similarity values. For every pre-set threshold, the compounds that were outside the applicability domain were removed from the test dataset and the $RMSE_{test}$ was recalculated. The optimum threshold was the one that achieved the lowest $RMSE_{test}$ with the smallest possible number of compounds outside the applicability domain.

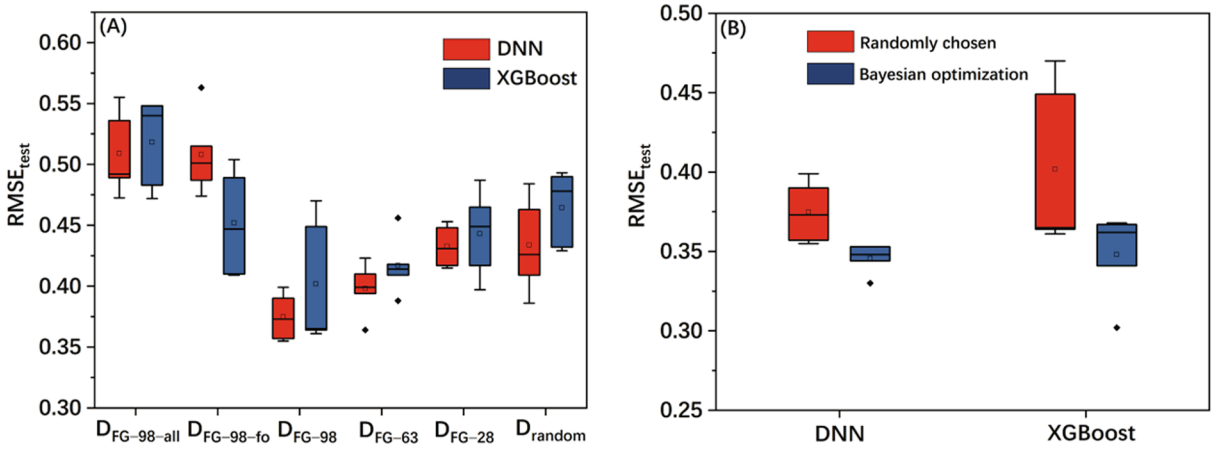


Fig. 1. (A) The effects of outliers and data splitting methods on the predication performance of the DNN and XGBoost. (B) Comparison of the prediction performance of the XGBoost and DNN with the hyperparameter values either randomly selected or obtained by the Bayesian optimization algorithm.

3. Results and discussion

3.1. Effect of outliers and data splitting methods on the performance of models

We first used the randomly chosen hyperparemeters for both DNN and XGBoost, as shown in Table S4. The outliers had a significantly adverse effect on the performance of both XGBoost and DNN, with lower *RMSE* being observed when more outliers were removed (Fig. 1A, from $D_{FG\text{-}98\text{-}all}$ to $D_{FG\text{-}98\text{-}fo}$ and then $D_{FG\text{-}98}$). The prediction performance also relied heavily on the data splitting methods, as the prediction accuracy increased with the chemicals more finely grouped (Fig. 1A, from D_{random} to $D_{FG\text{-}28}$, $D_{FG\text{-}63}$ and then $D_{FG\text{-}98}$). We then optimized the hyperparameters of DNN and XGBoost by the Bayesian algorithm (Table S3), obtained better and more robust prediction performances for both models (Fig. 1B), and confirmed the optimum ratio of training, validation, and test datasets as 8:1:1 (Fig. S5).

Table 1 lists the performance of seven reported models that have been developed based on different algorithms. With an increasing number of chemicals (n from 55 to 526), more molecular descriptors (p from 4 to 13) had to be involved in the reported models, but the model performance still became worse (*RMSE* from 0.079 to 0.356). This may be because there are more complex non-linear correlations between the molecules and their reactivity for a larger dataset. Such correlations cannot be captured by the simple models because of their limited fitting ability. Generally, the predictive performance of all the models is also better for the training datasets than for the test datasets in terms of R^2 and *RMSE*. Based on the DNN-MF model we developed in our previous study for 593 compounds [23], we increased the number of compounds to 1089 but still achieved a comparative or even better prediction performance here in terms of $RMSE_{test}$, demonstrating the effectiveness of our model. The slightly lower R^2_{test} value (Table 1) in this study than in the previous one is likely because the metric of R^2_{test} is highly dependent on the size and distribution of the test dataset [48]. For example, the size of the test dataset in this study is 109, which is even larger than the whole dataset used by Wang et al. [22]. These results suggest that the DNN and XGBoost models had been well trained and were ready to be interpreted.

3.2. SHAP explanation of the XGBoost and DNN results

Fig. 2A and B are the summary SHAP plots of the top-20 most important features of the compounds in the training dataset as learned by the DNN and XGBoost. The greater the absolute SHAP value is, the more influence it has on the $\log k_{OH}$.. The SHAP values of all the features in each compound are summarized in the excel file named "SHAP_values.xlsx" in the Supplementary Material.

The SHAP patterns of the training dataset are similar to those of the test dataset (Fig. S6), indicating that the predictions made on the test dataset were indeed based on the knowledge learned from the training dataset. The blue dots mean 0 s in the MFs, which contain no structural information and, thus, theoretically should not affect the predictions (i.e., SHAP = 0). However, for the top-10 features that have 0 s, their contributions to the predictions were non-zero. This is because the XGBoost and DNN "learned" that there were obvious differences in the $\log k_{HO}$. values between the compounds with and without these features (Fig. S7). Nevertheless, the blue dots for most of the features indeed negligibly affected the predictions, as indicated by the top-100 features in Fig. S8. Therefore, the physicochemical meaning that no structural feature means no influence on the reactivity was mostly successfully learned by both the XGBoost and DNN.

The patterns in Fig. 2 also implied that the DNN and XGBoost employed two distinctly different approaches to make predictions. The DNN treated each atom group statically without "considering" the specific bonding environments the atom groups were in, as indicated by the same SHAP values for the same features (vertical bars) in different compounds. In contrast, the XGBoost "considered" every atom group dynamically because it assigned different SHAP values to the same atom groups (vertical and horizontal bars) in different compounds, that is, the contributions of the atom groups to the predictions depended on the specific organic compounds. Obviously, such differential assignments are chemically more meaningful because the same atom groups in different bonding environments can affect the compound reactivity differently.

3.3. Development of ensemble model

To simultaneously achieve the accuracy of the DNN and the chemically meaningful differential assignments of the XGBoost, we developed an ensemble model by linearly combining the predictions from the XGBoost and DNN of certain ratios, as shown in Eq. 2, where the ratio of 0.72:0.28 was obtained by the Bayesian optimization algorithm.

$$\text{Ensemble model} = 0.72 \times \text{DNN} + 0.28 \times \text{XGBoost}$$

Table 1
Comparison among different models for their performance in predicting aqueous k_{HO}. values.

Model	Algorithm	n[a]	p[b]	Training Set		Test Set	
				R^2_{train}	$RMSE_{train}$	R^2_{test}	$RMSE_{test}$
Wang et al. [22]	MLR	55	4	0.905	0.139	0.962	0.079
Kušić et al. [49]	GA-MLR[c]	78	4	0.735	0.174	0.76	0.20
Sudhakaran and Amy [5]	PCA-MLR[d]	83	2	0.918	–	–	–
Jin et al. [50]	MLR	118	7	0.823[h]	0.204	0.772	0.329
Borhani et al. [3]	BPSO-MLR[e]	457	8	0.716	0.347	0.724	0.356
Luo et al. [4]	MLR	526	13	0.805[h]	0.165	0.802	0.232
[23]	DNN	593	0	0.972	0.135	0.747	0.329[f]
Gupta et al. [34]	DTB[f]	995	5	0.954	0.17	0.925	0.14[g]
This study	DNN-MF	1089	0	0.88–0.92	0.20–0.24	0.61–0.69	0.333–0.353
	XGBoost-MF	1089	0	0.80–0.93	0.18–0.29	0.53–0.67	0.302–0.368

[a] n = the total number of chemicals in the dataset;
[b] p = the number of MD or the length of MF;
[c] GA: genetic algorithm; MLR: multiple linear regression;
[d] PCA: principal component analysis;
[e] BPSO: binary particle swarm optimization;
[f] DTB: decision treeboost;
[g] No test dataset was supplied. Instead, $RMSE_{validation}$ is shown here. Hence, the generalization ability of the model cannot be evaluated and this model was not used to compare with other models;
[h] R^2_{adj}

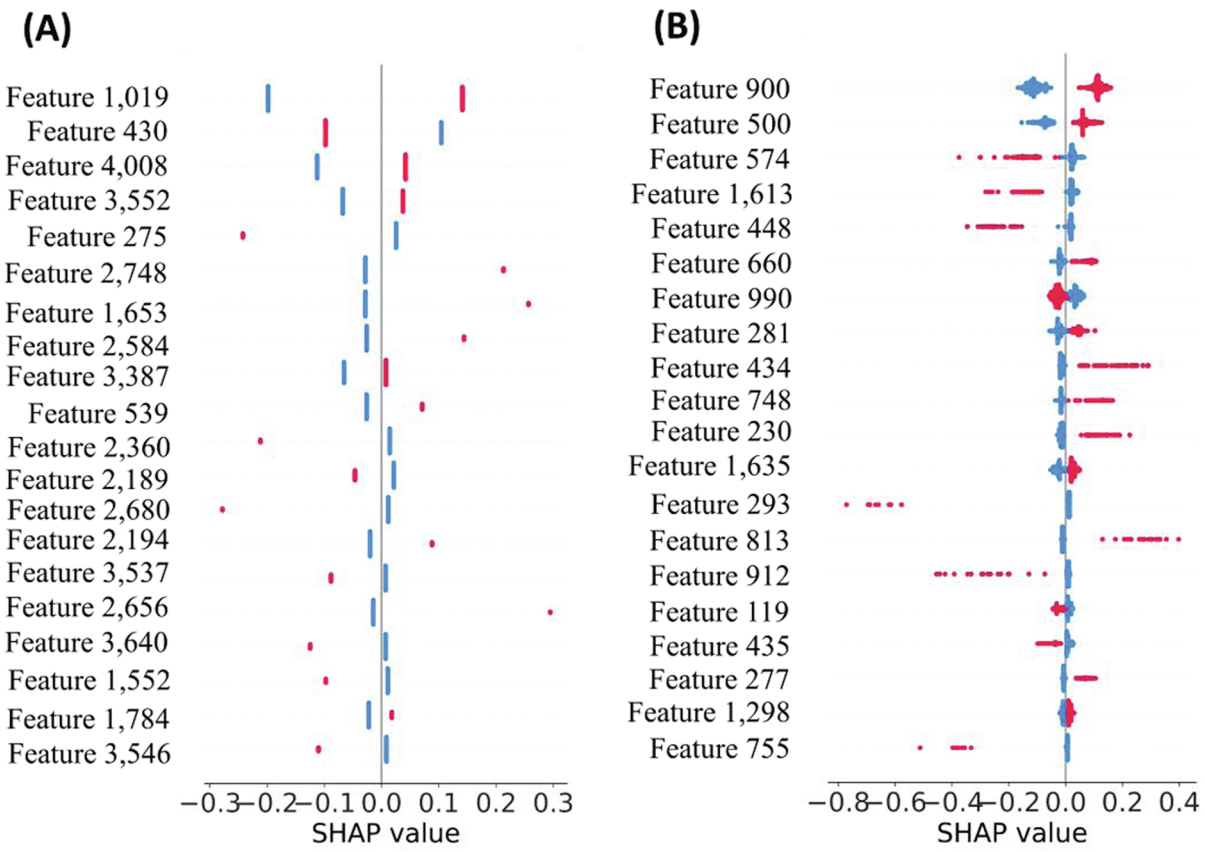


Fig. 2. The summary SHAP plots of the DNN (A) and the XGBoost (B) for the compounds in the training datasets. For each feature, blue means 0 while red means 1 in the MF. The feature numbers on the left label the positions of the features in the MFs, and different positions represent different atoms or substructures (Text S1). The x-axes are the SHAP values where a positive value means that it can increase the logk_{HO}. by the value; whereas a negative value means that it can decrease the logk_{HO} . by the value. The pattern for each feature is composed of the SHAP values for all the chemicals that contain that feature. Note that because the XGBoost and DNN used different lengths of MFs, the same atoms may be represented by different feature numbers. For instance, feature 900 for the XGBoost represents the same atom as feature 1019 for the DNN.

The ensemble model had a lower *RMSE* value than both the XGBoost and the DNN (Fig. 3A), and assigned different contributions to the same feature present in different compounds (Fig. 3B). Fig. 4 showed the scatter plots of the experimental versus predicted logk_{OH}. values by the ensemble model. For all the five groups, the predictive performances for the training, validation and test dataset were respectively similar in terms of R^2, *RMSE* and *MAE*, indicating that the prediction accuracy of the ensemble model was robust, or independent of data splitting. Next, we focused on interpreting this well-established ensemble model by the SHAP method to evaluate whether the model agreed with a few widely accepted chemical reaction principles, as detailed below.

3.4. Learned electron-donating and -withdrawing groups

Table 2 lists the top-10 most important atom groups learned by the ensemble model. The positive or negative effects of these features on logk_{HO}. were obtained based on their positive or negative SHAP values and are also included in Table 2. HO· radicals mainly oxidize electron-rich organic compounds. Any functional group that can increase the electron density of the compound will increase the reactivity, that is, electron-donating groups or conjugated systems can increase the reactivity, whereas electron-withdrawing groups can decrease the reactivity. For example, the carbonyl oxygen (the 2nd feature) and halogen groups (the 5th feature), typically electron-withdrawing, were correctly identified to decrease the reactivity, whereas the electron-donating benzene ring (the 1st feature) and alkenes (the 6th feature) were correctly identified to increase the reactivity. Such chemistry knowledge often needs several years of experience, but our ML model quickly learned this after developing the prediction models based on the MFs and the logk_{HO}.. It should be noted that we did not train our ML model to classify the positive or negative effects but only correlated the logk_{OH}. with the MFs.

Table 3 lists 19 well-known aromatic substituents that have positive or negative effects on the oxidation reactivity (the atom groups for all of the substituents are summarized in the file named “All the compounds and their features and representing atoms.txt” in the Supplementary Material). Such effects are the total effect that includes the inductive and resonance effects. Surprisingly, the ensemble model correctly learned the effects of all the substituents (100%) except for the −$CONH_2$ group (5 out 6 chemicals correct). These results suggest that our ensemble model was based on a reasonable casual understanding to make predictions, which made it trustworthy.

We then plot the electronic effects of the substituents as quantified by the Hammett constants [51] against the SHAP values to quantitatively investigate the magnitude of these effects. It should be noted that higher absolute SHAP values mean larger effects while positive and negative SHAP values represent increasing or decreasing effects on the reactivity. Fig. 5A showed the inductive electronic effects on the SHAP values. A higher Hammett constant σ_I represents a stronger inductive effect, which decreases the electron density of compounds more significantly and hence lowers the reactivity toward OH·. Fig. 5A showed the decreasing trend in the SHAP values from positive to negative, that is, from increasing to decreasing the reactivity when σ_I increased from about 0 to 0.65, which is consistent with the chemical knowledge. Fig. 5B showed the resonance effect on the SHAP values. A more negative σ_R represents a stronger resonance effect, leading to a higher

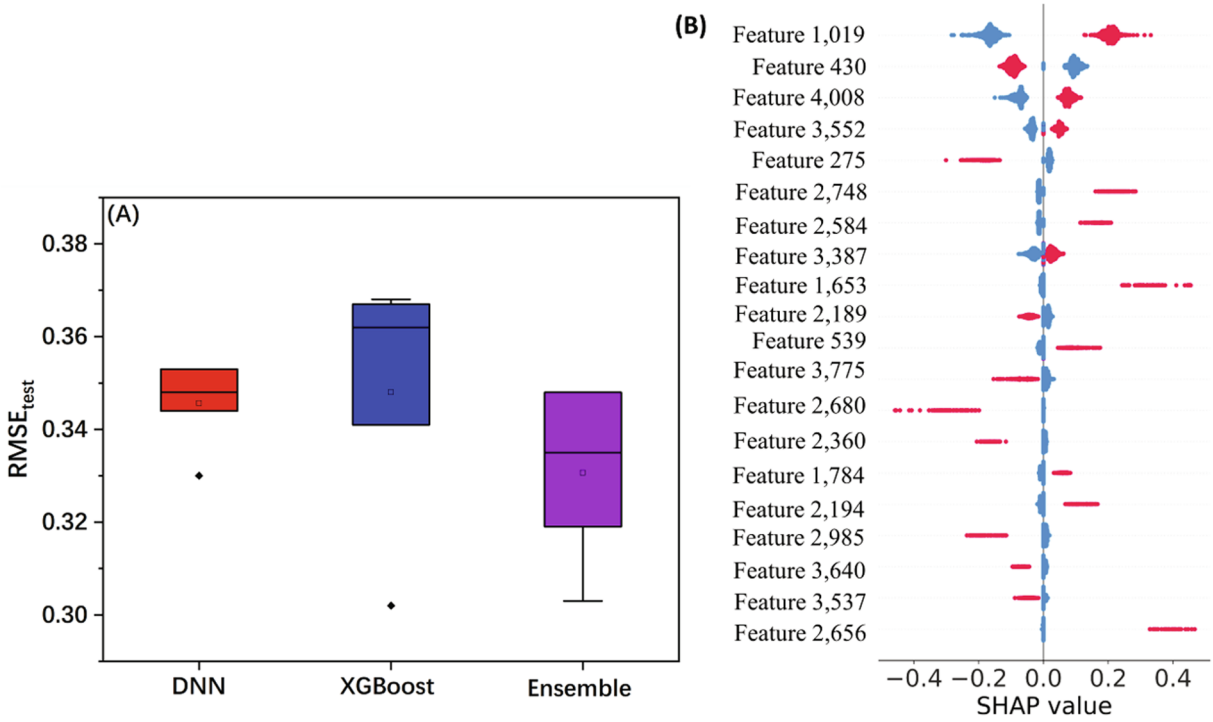


Fig. 3. (A) The comparison among the DNN, XGBoost and ensemble models in terms of the prediction accuracy; (B) the summary plot of the top-20 learned features by the ensemble model. To combine the DNN and XGBoost into the ensemble model, the MF length was set to 4050 and the optimum values of the hyperparameters for XGBoost are in Table S5.

electron density on the compounds, which is beneficial to the oxidative reactivity by HO·. Fig. 5B also showed a similar downward trend between the SHAP values and σ_R. Fig. 5C showed the total electronic effect resulting from both the inductive and resonance effects, with a higher R^2 value (0.548) than in the case of either σ_I (R^2 = 0.371) or σ_R (R^2 = 0.472), indicating a more obvious correlation. This is expected because the reactivity of compounds toward HO· is affected by both the inductive and resonance effects. Given that different compounds have different reactive sites and the effect of a certain substituent toward different reactive sites cannot be the same, the above non-perfect linear correlations are chemically reasonable. Nonetheless, the chemical knowledge learned by the ensemble model contains not only qualitative effects but also quantitative effects on the reactivity, which was not expected.

3.5. Learned reactive sites in the organic compounds toward attack by HO·

The identification of the reactive sites in the organic compounds toward HO· was achieved by examining the largest SHAP values for different features in each compound (an example in Fig. S9). For all 1,089 compounds, we compiled their individual SHAP plots, the chemical structures, and the MFs in the Supplementary Material (the file named "All the compounds and their shap plots.txt"). Fig. 6 summarizes the learned reactive sites in 10 chemical classes whose oxidation mechanisms by HO radicals have been well understood. We started from simple alkanes, whose $-CH_2-$ groups were thought to be the most reactive because of the largest absolute SHAP values (compound No. 41–46). This is interestingly consistent with the experimental observation that HO· prefers to attack $-CH_2-$ groups [8]. When these long-chain alkanes are folded to form cycloalkanes (compound No. 667–671), the ensemble model still identified all the $-CH_2-$ groups as the essential feature, and these groups indeed all react with HO· as the reactive sites [8]. When a single bond in alkanes is replaced by a double bond to form alkenes (compound No. 49–52), the model "smartly" changed its decision from the $-CH_2-$ groups to the carbon atoms in the double bonds, and the latter are indeed attacked by HO· as the reactive sites [9,11]. When one or two H atoms in alkanes are replaced by −OH to form alcohols (compound No. 850–853), the model still underscored the $-CH_2-$ groups rather than the −OH group. This is again consistent with the experimental observation that HO· prefers to abstract H from $-CH_2-$ rather than from −OH [6]. However, the model did not correctly identify the α-C to the −OH group to be more reactive than the β- or γ-C [6]. Different from the case of alcohols, when −OH is replaced by −SH to form thiols (compound No. 1016-1018), the model turned to the −SH group as the reactive site, although the only difference between the alcohols and the thiols is the S and O atoms. Indeed, −SH often reacts with HO· by H-abstraction as the first reactive site [6,11]. Similar cases were found in ethers (compound No. 736–751) and sulfides (compound No. 973–981) in which the methyl and/or methylene groups in the ethers and the sulfur atoms in the sulfides were correctly recognized as the reactive sites [15,18]. The carbon atoms in aldehydes (compound No. 0–5) or the nitrile group (compound No. 810–813), identified as the most reactive features by the model, are also known reactive sites in which they undergo H-abstraction and HO·–addition, respectively [10,13,17]. In the oxidation of amines (compound No. 285–302), they can undergo either H-abstraction from the $-NH_2$ and the neighboring methylene/methyl groups or electron transfer from the tertiary N [7,14]. The top-2 largest SHAP values were correctly assigned to these groups so these reactive sites were also "smartly" identified by the model.

Several features were identified to have the largest negative SHAP values, including $-NO_2$ (compound No. 820–833), −COR (compound

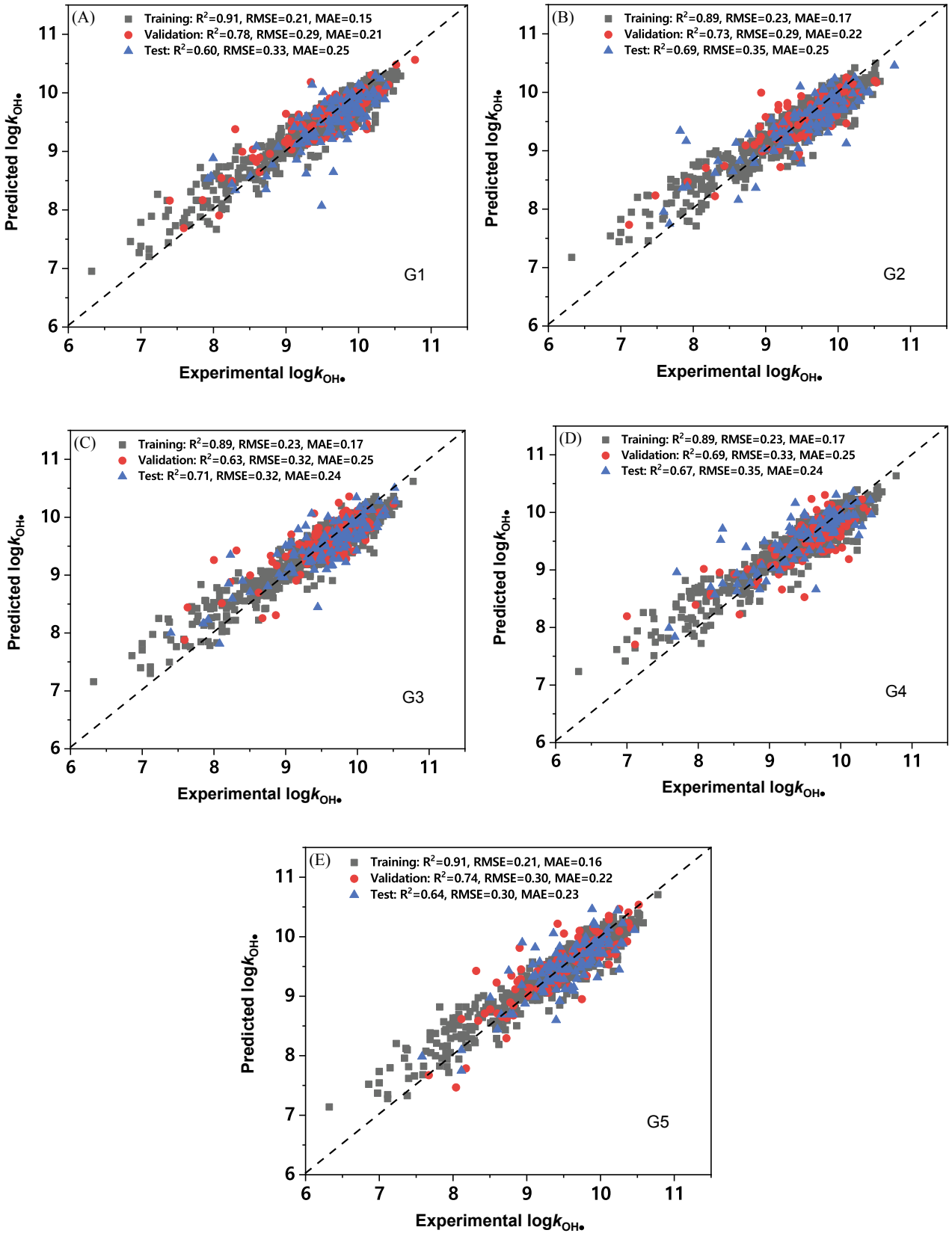
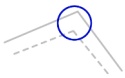
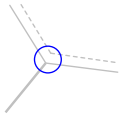
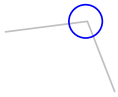
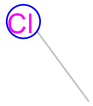
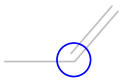
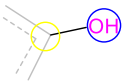
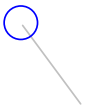
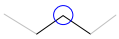
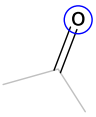


Fig. 4. The scatter plots of the experimental versus the predicted log$k_{HO·}$ by the ensemble model for the five groups: (A) G1, (B) G2, (C) G3, (D) G4, and (E) G5.

No. 800–807), −COOR (compound No. 718–727), and −N=O (compound No. 834–845), but they are not the reactive sites toward HO· [6,11,52]. This is because all of these features are electron-withdrawing groups and can significantly lower the reactivity. Based on the obtained large negative SHAP values, the ensemble model exactly identified their significant negative effects on the predicted reactivity.

Table 2
The top-10 most important features and their positive and negative effects on the reactivity as learned by the ensemble model. The values in the parentheses are the sum of the absolute SHAP values for that feature in all compounds.

Ranking	1st (201.51)	2nd (102.26)	3rd (83.14)	4th (45.03)	5th (43.43)
atoms					
Feature #	1019	430	4008	3522	275
Atom group	Aromatic carbon	Carbonyl oxygen	Aromatic carbon	Methylene carbon	Chlorine
Effect on logk_{HO}.	positive	negative	positive	positive	negative
Ranking	6th (35.47)	7th (32.02)	8th (30.18)	9th (26.34)	10th (24.12)
atoms					
Feature #	2748	2584	3387	1653	2189
Atom group	Ethylene carbon	Aromatic OH	Methyl	Methylene	Carbonyl
Effect on logk_{HO}.	positive	positive	positive	positive	negative

Note: One compound may have multiple identical features. For example, feature 1019 represents aromatic carbon atoms that are only attached to one H atom; toluene has five of them (Fig. S1c). The blue circles represent the center atoms, the black solid lines represent the bonds in the feature, the grey lines represent the neighboring bonds not in the feature, the dotted lines represent conjugated structures, e.g., aromatic, and the yellow color represents an aromatic atom in the feature. All heavy atoms except for C, such as O and Cl, are shown in different colors, while the C atoms are not shown.

3.6. Applicability domain

Table 4 shows that with an increasing threshold of the maximum similarity, more compounds (0 to 5) in the test dataset were outside the applicability domain, but the $RMSE_{test}$ first decreased and then increased. A threshold of 0.35 led to the minimal $RMSE_{test}$ (0.317), indicating the best prediction performance of the model. Nevertheless, there is not much difference among the $RMSE_{test}$ when the threshold was above 0.25. With an increasing threshold for the mean similarity, more compounds (0 to 5) in the test dataset were also determined to be outside the applicability domain. Different from the case of maximum similarity, the $RMSE_{test}$ first increased and then decreased. Although lower $RMSE_{test}$ can be achieved by further increasing the threshold value, the number of compounds outside the applicability domain increased. Gadaleta et al. pointed out that lower $RMSE_{test}$ and less compounds outside of applicability domain are preferred when choosing an appropriate applicability domain [45]. When we compared the number of compounds outside the domain at the $RMSE_{test}$ = 0.317 based on the maximum and mean similarity, 4 and 3 compounds were excluded, respectively. Hence, the optimum similarity metric and threshold value are the mean similarity and 0.028. We can thus conclude that if the mean structural similarity of a given compound to the ones in the training dataset is higher than 0.028, our ensemble model can make a reliable prediction. Such a low similarity level indicates that the ensemble model is robust and can be applied to a broad range of organic compounds.

4. Conclusions

In this study, we combined two ML algorithms, DNN and XGBoost,

Table 3
Common aromatic substituents with positive or negative electronic effects. [a]

Effect	Functional Group	Feature #	No. of compounds with the feature	No. of correctly identified compounds	Percentage (%)
Positive	−NH_2	3359	62	62	100
	−NHR	2292	10	10	100
	−NR_2	2054	7	7	100
	−OH	2584	121	121	100
	−OR	2258	53	53	100
	−NHCOR	1733 + 430 + 1568[b]	6	6	100
	−C_6H_5	1019	19	19	100
	−CH=CH_2	2748	6	6	100
	−CH_2CH_3	1979	15	15	100
Negative	−NO_2	3115 + 430 + 495[b]	36	36	100
	−CN	3408	23	23	100
	−CHO	3749 + 430[b]	11	11	100
	−$COCH_3$	2800 + 430[b]	9	9	100
	−$CONH_2$	1685 + 430 + 539[b]	6	5	83.3
	−CONHR	1790 + 430 + 566[b]	9	9	100
	−$CONR_2$	162 + 1747 + 430[b]	1	1	100
	−N=O	2514 + 430[b]	2	2	100
	−Cl	275	63	63	100
	−Br	2680	16	16	100

[a] These groups were chosen because (1) their positive or negative effects on the oxidation reactivity are well established; and (2) these groups are in our dataset.
[b] Some groups were too large to be described by one feature when the MF radius was 1, such as −$CONR_2$. In these cases, their SHAP values were calculated by summing the SHAP values of all of the features involved.

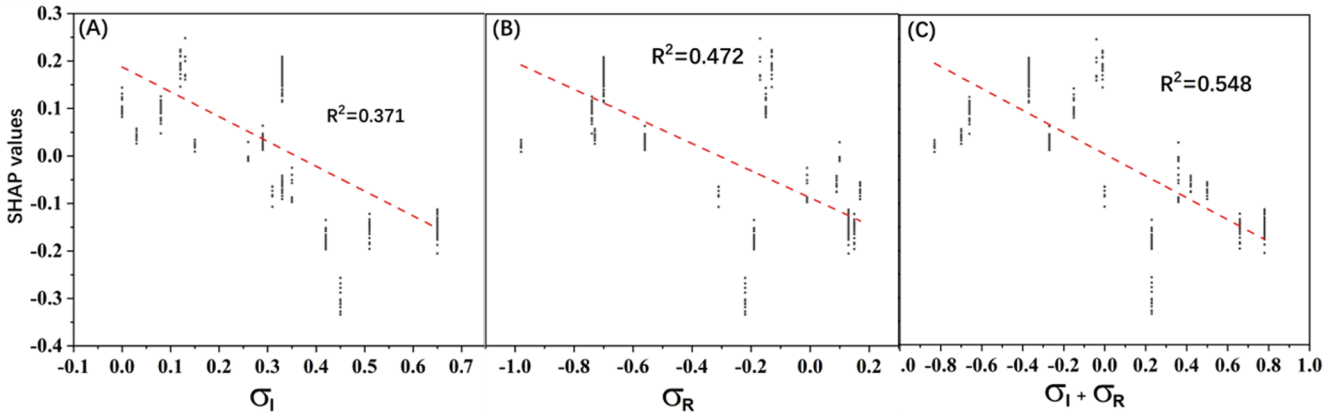


Fig. 5. Correlations between the Hammett constants of the aromatic substituents in Table 2 and their SHAP values: (A) the inductive effect, (B) the resonance effect, and (C) the total electronic effect.

with the MFs of organic compounds to develop an accurate prediction model for the logk_{HO}. of 1089 compounds ($RMSE_{test}$ for DNN-MF: 0.333–0.353, $RMSE_{test}$ for XGBoost-MF: 0.302–0.368). How the two algorithms made predictions on logk_{HO}. were interpreted by the SHAP method. The results showed that for the contribution of a given structural feature to the predicted logk_{HO}. values for different compounds, the DNN and XGBoost treated it as a fixed and variable value, respectively. Based on this interpretation, we recognized that the DNN may not use all the correct chemical knowledge to develop models, even though the models' predictive performance was acceptable. Then, we developed the ensemble model by combining the DNN with the XGBoost, which achieved more accurate and robust prediction performance ($RMSE_{test}$: 0.30–0.35) while relying on the same working mechanism as the XGBoost. Given the fact that more and more contaminants may arise in the future, the ML-MF based approach will show great applications in the environmental field. We then interpreted the ensemble model and found that it surprisingly "learned" the reaction patterns after linking the reactivity to the MFs, including "knowing" which atom groups can decrease or increase the reactivity, "quantifying" the decreasing or increasing effect of structural features, and "locating" the reactive sites in the compound structures toward HO· radicals. Specifically, aromatic carbons and carbonyl groups can increase and decrease the reactivity the most for these 1,089 compounds, and common well-known atom groups that can decrease or increase the reactivity, such as $-NH_2$, $-Br$, $-CH_2{=}CH_2$ and $-N{=}O$, were correctly identified. The decreasing or increasing effect was quantitatively analyzed by the plots of Hammett constants versus the SHAP values. The reactive sites for 10 common chemical classes were correctly located based on the highest absolute SHAP values. The applicability domain of the ensemble model was determined to be 0.028 of the mean similarity, that is, the HO· reactivity of any query compounds with a similarity value of over 0.028 can be reliably predicted, indicating the wide application potential of the built model to a broad range of organic compounds. These results demonstrated that ML can implicitly learn the chemical knowledge when it helps perform a chemical task. Therefore, although the ML methods used here have a "black box"

Table 4
The thresholds of similarity, the number of compounds outside the applicability domain for each threshold value, and the corresponding $RMSE_{test}$.

Similarity metric	Threshold value	# of Compounds outside the applicability domain	$RMSE_{test}$
Maximum	0.25	0	0.319
	0.30	2	0.318
	0.35	4	0.317
	0.40	5	0.318
Mean	0.015	0	0.319
	0.020	2	0.321
	0.028	3	0.317
	0.030	5	0.313

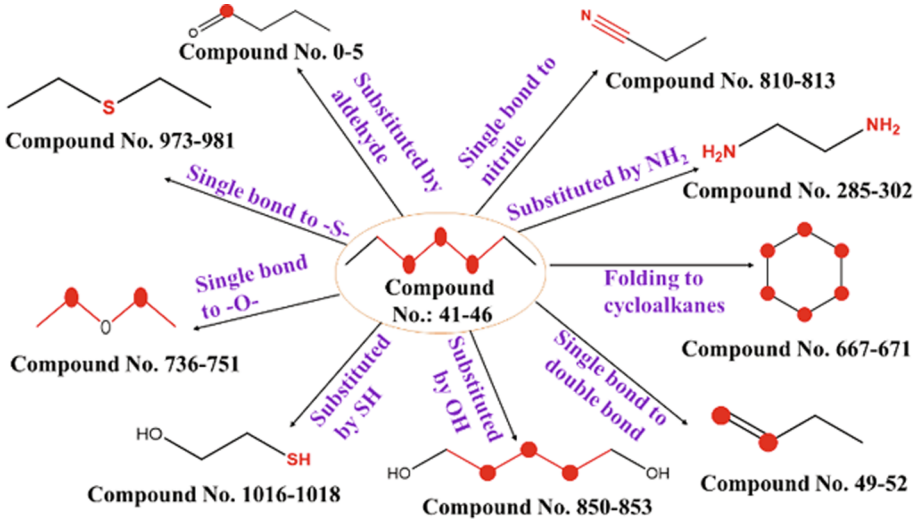


Fig. 6. The important features (highlighted in red) with the largest SHAP values for each compound learned by the ensemble model smartly shifted among the 10 chemical classes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

nature, we can still trust the ensemble models. This finding offers a theoretical support for the MF-ML method in developing QSAR models in general. The SHAP method can also be widely used as an interpretation method to reveal "black box" ML algorithms and to validate the developed models. Moreover, after unveiling the knowledge contained in these "black box" ML, it is possible to learn if there is any new knowledge. In brief, by uncovering the underlying chemical knowledge, we can trust a ML model, modify the model to make it more powerful, and leverage the implicit knowledge the model carries.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant CBET-1804708 and the U. S. Department of Agriculture under Grant 1022123.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2020.126627.

References

[1] S.M. Free, J.W. Wilson, A mathematical contribution to structure-activity studies, J. Med. Chem. 7 (4) (1964) 395–399.
[2] H. Kubinyi, QSAR and 3D QSAR in drug design Part 2: applications and problems, Drug Discov. Today 2 (12) (1997) 538–546.
[3] T. Borhani, M. Saniedanesh, M. Bagheri, J. Lim, QSPR prediction of the hydroxyl radical rate constant of water contaminants, Water Res. 98 (2016) 344–353.
[4] S. Luo, Z. Wei, R. Spinney, F.A. Villamena, D.D. Dionysiou, D. Chen, C.-J. Tang, L. Chai, R. Xiao, Quantitative structure–activity relationships for reactivities of sulfate and hydroxyl radicals with aromatic contaminants through single–electron transfer pathway, J. Hazard. Mater. 344 (2018) 1165–1173.
[5] S. Sudhakaran, G.L. Amy, QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification, Water Res. 47 (3) (2013) 1111–1122.
[6] G.V. Buxton, C.L. Greenstock, W.P. Helman, A.B. Ross, Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals (·OH/·O− in Aqueous Solution, J. Phys. Chem. Ref. Data 17 (2) (1988) 513–886.
[7] S. Das, C. von Sonntag, The oxidation of trimethylamine by OH radicals in aqueous solution, as studied by pulse radiolysis, ESR, and product analysis. The reactions of the alkylamine radical cation, the aminoalkyl radical, and the protonated aminoalkyl radical, Zeitschrift für Naturforschung B 41 (4) (1986) 505–513.
[8] W.B. DeMore, K.D. Bayes, Rate constants for the reactions of hydroxyl radical with several alkanes, cycloalkanes, and dimethyl ether, J. Phys. Chem. A 103 (15) (1999) 2649–2654.
[9] E.J. Feltham, M.J. Almond, G. Marston, K.S. Wiltshire, N. Goldberg, Reactions of hydroxyl radicals with alkenes in low-temperature matrices, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 56 (13) (2000) 2589–2603.
[10] A. Galano, Mechanism of OH radical reactions with HCN and CH_3CN: OH regeneration in the presence of O2, J. Phys. Chem. A 111 (23) (2007) 5086–5091.
[11] S. Gligorovski, R. Strekowski, S. Barbati, Environmental implications of hydroxyl radicals (•OH), Chem. Rev. 115 (2015) 13051–13092.
[12] A. Hatipoglu, D. Vione, Y. Yalçın, C. Minero, Z. Çınar, Photo-oxidative degradation of toluene in aqueous media by hydroxyl radicals, J. Photochem. Photobiol., A 215 (1) (2010) 59–68.
[13] A.J. Kerr, D.W. Sheppard, Kinetics of the reactions of hydroxyl radicals with aldehydes studied under atmospheric conditions, Environ. Sci. Technol. 15 (8) (1981) 960–963.
[14] K.N. Leitner, P. Berger, B. Legube, Oxidation of amino groups by hydroxyl radicals in relation to the oxidation degree of the α-carbon, Environ. Sci. Technol. 36 (14) (2002) 3083–3089.
[15] C. Schoeneich, A. Aced, K. Asmus, Mechanism of oxidation of aliphatic thioethers to sulfoxides by hydroxyl radicals. The importance of molecular oxygen, J. Am. Chem. Soc. 115 (24) (1993) 11376–11383.
[16] L. Sjöberg, T.E. Eriksen, R.-L research, The reaction of the hydroxyl radical with glutathione in neutral and alkaline aqueous solution, Radiat. Res. 89 (1982) 255–263.
[17] G.S. Tyndall, J.J. Orlando, T.J. Wallington, M.D. Hurley, M. Goto, M. Kawasaki, Mechanism of the reaction of OH radicals with acetone and acetaldehyde at 251 and 296 K, PCCP 4 (11) (2002) 2189–2193.
[18] C. Zavala-Oseguera, J.R. Alvarez-Idaboy, G. Merino, A. Galano, OH radical gas phase reactions with aliphatic ethers: a variational transition state theory study, J. Phys. Chem. A 113 (50) (2009) 13913–13920.
[19] E.S. Kwok, R. Atkinson, Estimation of hydroxyl radical reaction rate constants for gas-phase organic compounds using a structure-reactivity relationship: an update, Atmos. Environ. 29 (14) (1995) 1685–1695.
[20] D. Minakata, K. Li, P. Westerhoff, J. Crittenden, Development of a group contribution method to predict aqueous phase hydroxyl radical (HO•) reaction rate constants, Environ. Sci. Technol. 43 (16) (2009) 6220–6227.
[21] A. Monod, J. Doussin, Structure-activity relationship for the estimation of OH-oxidation rate constants of aliphatic organic compounds in the aqueous phase: alkanes, alcohols, organic acids and bases, Atmos. Environ. 42 (33) (2008) 7611–7622.
[22] Y.n. Wang, J. Chen, X. Li, S. Zhang, X. Qiao, Estimation of aqueous-phase reaction rate constants of hydroxyl radical with phenols, alkanes and alcohols, QSAR Comb. Sci. 28 (11–12) (2009) 1309–1316.
[23] S. Zhong, J. Hu, X. Fan, X. Yu, H. Zhang, A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants, J. Hazard. Mater. 383 (2020) 121141.
[24] J.-F. Doussin, A. Monod, Structure-activity relationship for the estimation of OH-oxidation rate constants of carbonyl compounds in the aqueous phase, Atmos. Chem. Phys. 13 (23) (2013) 11625.
[25] R.C. Glen, A. Bender, C.H. Arnby, L. Carlsson, B.-S. Idrugs, Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME, IDrugs 9 (2006) 199–204.
[26] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754.
[27] K. McCloskey, A. Taly, F. Monti, M.P. Brenner, L.J. Colwell, Using attribution to decode binding mechanism in neural network models for chemistry, Proc. Natl. Acad. Sci. 116 (2019) 11624–11629.
[28] J. Upmeier zu Belzen, T. Bürgel, S. Holderbach, F. Bubeck, L. Adam, C. Gandor, M. Klein, J. Mathony, P. Pfuderer, L. Platz, M. Przybilla, M. Schwendemann, D. Heid, M.D. Hoffmann, M. Jendrusch, C. Schmelas, M. Waldhauer, I. Lehmann, D. Niopek, R. Eils, Leveraging implicit knowledge in neural networks for functional dissection and engineering of proteins, Nature Mach. Intelligence 1 (5) (2019) 225–235.
[29] S.M. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions (2017), pp. 4765-4774.
[30] D.F. Read, K. Cook, Y.Y. Lu, K.G. Le Roch, W.S. Noble, Predicting gene expression in the human malaria parasite Plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features, PLoS Comput. Biol. 15 (9) (2019) e1007329.
[31] A. Stojić, N. Stanić, G. Vuković, S. Stanišić, M. Perišić, A. Šoštarić, L. Lazić, Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition, Sci. Total Environ. 653 (2019) 140–147.
[32] Y. Zhao, L. Wang, J. Luo, T. Huang, S. Tao, J. Liu, Y. Yu, Y. Huang, X. Liu, J. Ma, Deep learning prediction of polycyclic aromatic hydrocarbons in the high arctic, Environ. Sci. Technol. 53 (2019) 13238–13245.
[33] W.-W. Cai, T. Peng, J.-N. Zhang, L.-X. Hu, B. Yang, Y.-Y. Yang, J. Chen, G.-G. Ying, Degradation of climbazole by UV/chlorine process: kinetics, transformation pathway and toxicity evaluation, Chemosphere 219 (2018) 243–249.
[34] S. Gupta, N. Basant, Modeling the aqueous phase reactivity of hydroxyl radical towards diverse organic micropollutants: an aid to water decontamination processes, Chemosphere 185 (2017) 1164–1172.
[35] J. Khan, X. He, N.S. Shah, M. Sayed, H.M. Khan, D.D. Dionysiou, Degradation kinetics and mechanism of desethyl-atrazine and desisopropyl-atrazine in water with •OH and $SO_4^{\bullet-}$ based-AOPs, Chem. Eng. J. 325 (2017) 485–494.
[36] D. Lee, M. Kwon, Y. Ahn, Y. Jung, S.-N. Nam, I.-H. Choi, J.-W. Kang, Characteristics of intracellular algogenic organic matter and its reactivity with hydroxyl radicals, Water Res. 144 (2018) 13–25.
[37] H. Lin, N. Oturan, J. Wu, V.K. Sharma, H. Zhang, M.A. Oturan, Removal of artificial sweetener aspartame from aqueous media by electrochemical advanced oxidation processes, Chemosphere 167 (2017) 220–227.
[38] E.V. Ortiz, D.O. Bennardi, D.E. Bacelo, S.E. Fioressi, P.R. Duchowicz, The conformation-independent QSPR approach for predicting the oxidation rate constant of water micropollutants, Environ. Sci. Pollut. Res. 24 (35) (2017) 27366–27375.
[39] Z. Wu, K. Guo, J. Fang, X. Yang, H. Xiao, S. Hou, X. Kong, C. Shang, X. Yang, F. Meng, L. Chen, Factors affecting the roles of reactive species in the degradation of micropollutants by the UV/chlorine process, Water Res. 126 (2017) 351–360.
[40] P. Xie, J. Ma, W. Liu, J. Zou, S. Yue, X. Li, M.R. Wiesner, J. Fang, Removal of 2-MIB and geosmin using UV/persulfate: contributions of hydroxyl and sulfate radicals, Water Res. 69 (2015) 223–233.
[41] R. Zhang, Y. Yang, C.H. Huang, L. Zhao, P. Sun, Kinetics and modeling of sulfonamide antibiotic degradation in wastewater and human urine by UV/H_2O_2 and UV/PDS, Water Res. 103 (2016) 283–292.
[42] T. Chen, C. Guestrin. XGBoost: A Scalable Tree Boosting System. arXiv, 785-794 (2016).
[43] I. Dewancker, M. McCourt, S. Clark. Bayesian Optimization for Machine Learning : A Practical Guidebook. arXiv:1612.04858 (2016).
[44] J. Snoek, H. Larochelle, H. and in neural information, A.-R.P. Practical bayesian optimization of machine learning algorithms. Advances in neural information Processing Systems 25 (NIPS 2012) (2012).
[45] D. Gadaleta, G.F. Mangiatordi, M. Catto, A. Carotti, O. Nicolotti, Applicability domain for QSAR models: where theory meets reality, Int. J. Quantitative Struct.-Property Relationships (IJQSPR) 1 (1) (2016) 45–63.

[46] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (5) (2012) 4791–4810.
[47] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 48 (9) (2008) 1733–1746.
[48] R. Todeschini, D. Ballabio, F. Grisoni, Beware of unreliable Q 2! A comparative study of regression metrics for predictivity assessment of QSAR models, J. Chem. Inf. Model. 56 (10) (2016) 1905–1913.
[49] H. Kušić, B. Rasulev, D. Leszczynska, J. Leszczynski, N. Koprivanac, Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: A QSAR study, Chemosphere 75 (8) (2009) 1128–1134.
[50] X. Jin, S. Peldszus, P.M. Huck, Predicting the reaction rate constants of micropollutants with hydroxyl radicals in water using QSPR modeling, Chemosphere 138 (2015) 1–9.
[51] C. Hansch, A. Leo, R.W. Taft, A survey of Hammett substituent constants and resonance and field parameters, Chem. Rev. 91 (2) (1991) 165–195.
[52] M. Wollenhaupt, J.N. Crowley, Kinetic studies of the reactions CH_3 + NO_2 → Products, CH_3O + NO_2 → products, and OH + $CH_3C(O)CH_3$ → $CH_3C(O)OH$ + CH_3, over a range of temperature and pressure, J. Phys. Chem. A 104 (27) (2000) 6429–6438.