Knowledge transfer between small datasets for boosting the predictive performance of machine learning assisted QSAR models on contaminant oxidative reactivity

Journal:	*Environmental Science & Technology*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Zhong, Shifa; Case Western Reserve University, Civil Engineering Zhang, Yanping; Hebei University of Technology Zhang, Huichun; Case Western Reserve University Case School of Engineering, Civil and Environmental Engineering

Knowledge transfer between small datasets for boosting the predictive performance of machine learning assisted QSAR models on contaminant oxidative reactivity

Shifa Zhong[1], Yanping Zhang[2]* and Huichun Zhang[1]*

[1] Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Cleveland, OH 44106-7201, United States

[2] School of Civil Engineering and Transportation, Hebei University of Technology, Tianjin 300401, China

***Corresponding Authors**

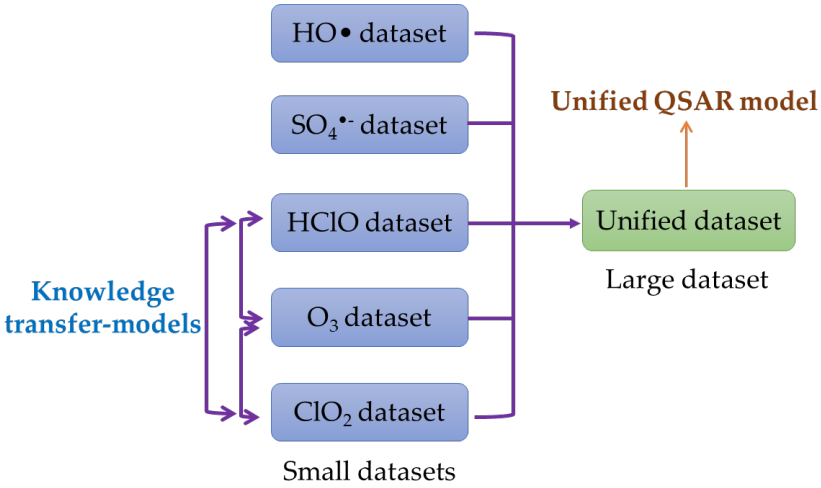
E-mail: hjz13@case.edu, zyphit@hebut.edu.cn

Abstract: Using machine learning (ML) to develop quantitative structure—activity relationship (QSAR) models for contaminant reactivity has emerged as a promising approach because it can effectively handle non-linear relationships. However, ML is often data-demanding, whereas data scarcity is common in QSAR model development. Here, we proposed two approaches to address this issue: combining small datasets and transferring knowledge between them. First, we compiled four individual datasets for four oxidants, i.e., $SO_4^{\bullet-}$, HClO, O_3 and ClO_2, each dataset containing a different number of contaminants with their corresponding rate constants and reaction conditions (pH and/or temperature). We then used molecular fingerprints (MF) or molecular descriptors (MD) to represent the contaminants; combined them with ML algorithms to develop individual QSAR models for these four datasets; and interpreted the models by the Shapley Additive exPlantion (SHAP) method. The results showed that both the optimal contaminant representation and the best ML algorithm are dataset dependent. Next, we merged these four datasets and developed a unified model, which showed better

predictive performance on the datasets of HClO, O_3 and ClO_2 because the model 'corrected' some wrongly learned effects of several atom groups. We further developed knowledge transfer models based on the second approach, the effectiveness of which depends on if there is consistent knowledge shared between the two datasets as well as the predictive performance of the respective single models. This study demonstrated the benefit of combining small similar datasets and transferring knowledge between them, which can be leveraged to boost the predictive performance of ML-assisted QSAR models.

Synopsis: Two approaches improved the predictive performance of machine learning assisted QSAR models on contaminant oxidative reactivity: combining small datasets for different oxidants and knowledge transfer among them.

Keywords: QSAR; machine learning; knowledge transfer; contaminant oxidation; water treatment



TOC Art

1. Introduction

Oxidative processes play a vital role in removing organic contaminants during water and wastewater treatment.[1] Various oxidants, from •OH, $SO_4^{\bullet-}$,[2-4] and ClO_2 to ozone,[5, 6] can be applied for different organic contaminants, such as personal care products, endocrine disrupting chemicals, pesticides and industrial chemicals. The oxidation rate constant of contaminants is an important parameter for optimizing the treatment process by helping to, for example, estimate the removal efficiency of contaminants or determine the dosage of oxidants or the treatment retention time. Experimentally measuring reaction rate constants is time-consuming and labor-intensive. In comparison, developing quantitative structure—activity relationship (QSAR) models is an effective approach to estimating the rate constants for numerous contaminants, thus receiving increasing attention.[7-15] Built upon previous experimental results, QSAR models can correlate chemical structures with various chemical activities and be further applied to new query compounds to estimate their corresponding activity.

Many QSAR models have been successfully developed to predict the rate constants of various contaminants toward different oxidants, such as •OH, $SO_4^{\bullet-}$ and O_3.[9, 11, 16-23] To develop QSAR models, different chemical representations, such as molecular descriptors (MD),[16] molecular fingerprints (MF)[13] or molecular images,[14] can be combined with different regression methods, including multiple linear regression (MLR)[19, 20] and machine learning (ML).[14, 15] With more and more contaminants involved, traditional MLR has limited applicability because non-linear relationships may exist between contaminants and reaction rate constants. To handle non-linear relationships, ML has received increasing attention because of its powerful fitting ability. For example, Huang et al. reported a better performance of a support vector machine (SVM)-based model on predicting the rate constants of contaminants toward O_3 than MLR-based QSAR models.[20] Our recent study showed that ML-based models can achieve satisfactory predictive performance for a large dataset of •OH reactivity.[15]

ML algorithms, especially deep neural networks, often need a massive amount of data. However, data scarcity is a common issue when developing QSAR models for rate constants toward different oxidants, such as only 85 samples in a dataset of $SO_4^{\bullet-}$ radicals [21] or 136 samples in an O_3 dataset.[20] It is however impractical to experimentally measure rate constants (log*k*) for a large number of contaminants toward different oxidants to increase the sample size. We here propose a simple and effective approach—combining small datasets for different oxidants to form a larger dataset. This combined dataset contains samples for five common oxidants, including •OH, $SO_4^{\bullet-}$, O_3, ClO_2 and HClO. Previous studies treated these small datasets independently and developed separate QSAR models for each of them.[7, 16] However, all the involved reactions are oxidation reactions so they should share some common science. For example, for all the oxidants, we know that electron-donating or -withdrawing groups can increase or decrease the rate constant (*k)* for oxidation reactions, which was indeed correctly learned by our recent QSAR models for •OH radicals.[15] Ye et al. found that for $SO_4^{\bullet-}$ electron-donating groups (except for —N<) exhibit a positive coefficient for *k*, while electron-withdrawing groups (except for —S—) exhibit a negative coefficient for *k*.[19] Lee et al.'s study indicated decreasing *k* values with increasing Hammett constants for both ClO_2 and HClO,[7] which might be attributed to higher bond dissociation energies when electron-withdrawing substituents are present.[24] Huang et al. reported that E_{HOMO} (Energy of the Highest Occupied Molecular Orbital) was one of the most important descriptors in their QSAR model for O_3 because, as a measure of the electron-donating ability of a molecule, E_{HOMO} can be used to characterize the affinity of the molecule toward an electrophile.[20, 25] Compounds with higher E_{HOMO} are oxidized by O_3 with faster rates due to their stronger electron-donating ability. Because the shared science may be transferred from one dataset to another, combining small datasets to form a larger dataset may improve the predictive performance of the obtained model for all the oxidants. To the best of our knowledge,

this approach—developing a unified QSAR model on this large, unified dataset—has never been investigated before in developing QSAR models for contaminant reactivity.

Transfer learning, widely used in computer vision, is another popular approach to solving the data scarcity issue.[26] Transfer learning refers to pre-training a model on a large dataset and then tuning this pre-trained model on a smaller but similar dataset. We previously employed this concept when developing QSAR models for predicting rate constants for •OH radicals and found that, when employing molecular images to represent contaminants and pre-training a convolutional neural network (CNN) model on the ImageNet dataset, it can considerably increase the generalization ability of the QSAR models.[14] The ImageNet dataset is however quite different from the contaminant image dataset.[27] This transfer learning approach is also only limited to CNN algorithms.

For the datasets of •OH, $SO_4^{\bullet-}$, HClO, O_3 and ClO_2, they are similar to each other in terms of contaminant species and certain reaction mechanisms, as examples discussed above. Therefore, it would be interesting and beneficial to investigate whether the shared knowledge between any two datasets is transferable or not. However, how to effectively transfer knowledge among these different datasets without using CNN algorithms is still challenging. We here proposed a knowledge transfer approach for non-CNN algorithms, such as tree-based ML algorithms (Figure 1b). Our results below showed that the predictive performance of a QSAR model for a specific oxidant can be enhanced by learning from another oxidant without increasing the sample size of either oxidant.

In this study, we compiled the largest four datasets for four common oxidants, namely $SO_4^{\bullet-}$ HClO, O_3, and ClO_2, by including the reaction conditions, i.e., pH and/or temperature. The reaction conditions were seldom considered in previous studies, but including them can significantly increase the sample size. Two chemical representations, i.e., molecular descriptors (MDs) and molecular fingerprints (MFs), were used to combine with different ML algorithms to develop QSAR models. We first developed single QSAR

models for each oxidant. We then combined all of these datasets to form a large dataset and developed a unified QSAR model. The effect of this operation on the predictive performance of each dataset was investigated. We next used the knowledge transfer approach to develop knowledge transfer-based models and compared their predictive performance with the respective single models. The overall workflow of this study is summarized in Figure 1.

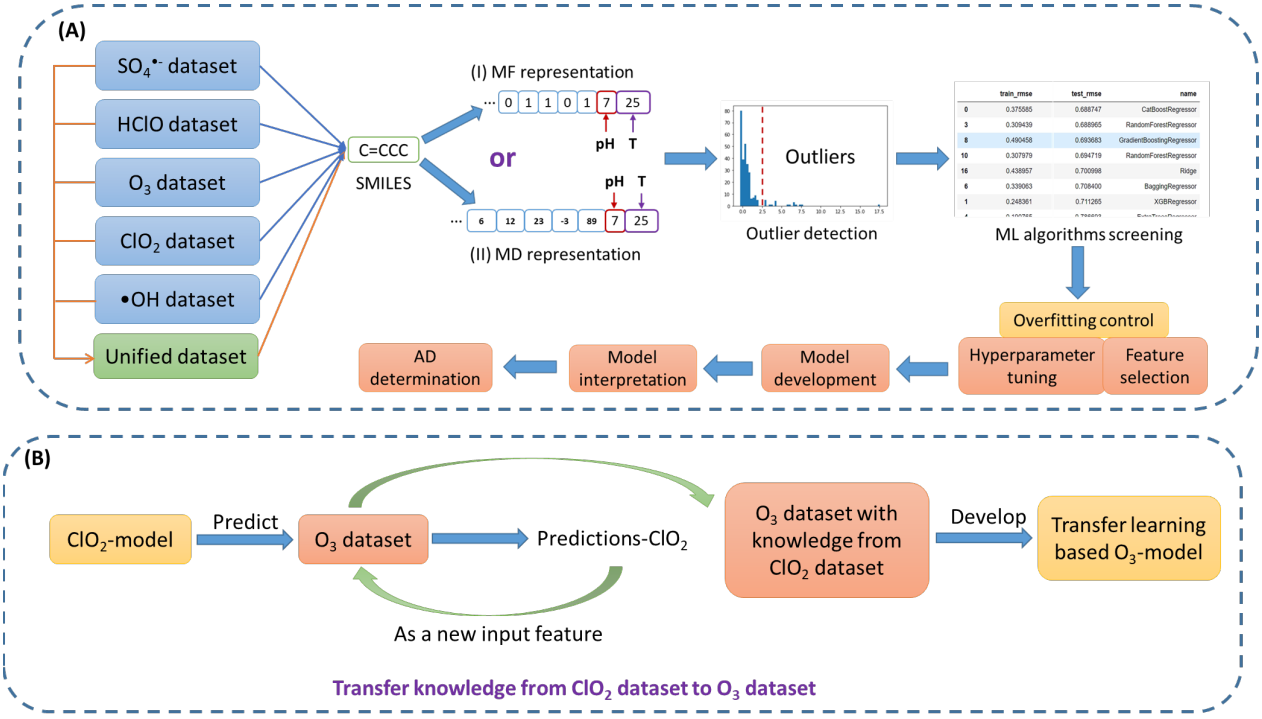


Figure 1. The workflow of this study. (A) The single and unified model development based on MFs or MDs. (B) An illustration of how knowledge transfer is achieved by an example of transferring knowledge from the ClO_2 dataset to the O_3 dataset (More details are in Section 2.4).

2. Materials and methods

2.1 Datasets

The kinetic data for the four oxidants were collected from the published literature, which were mined through Google Scholar (https://scholar.google.com/) by using the keywords: "sulfate radical", "HClO", "O_3" or "ClO_2" + "kinetics". As many as possible samples were collected and the attributes included contaminants, their corresponding rate constants (*k*), and reaction conditions (i.e., pH and/or temperature (T)). The number of studies we collected is listed in Table 1, in which the reported datasets for QSAR studies were directly cited without citing the original sources. Reaction conditions were often not included in previous studies. We here included the reaction conditions because reaction rate constants are condition dependent. For example, pH can affect the dissociation of some contaminants while differently charged contaminant species react with these oxidants at different rates.[6] Moreover, we can increase the sample size by including the reaction conditions. All the *k* values were log-transformed (log*k*) to reduce the range of values. If multiple log*k* values were reported for a contaminant for the same conditions, an average log*k* value was taken. The summary of these four datasets is listed in Table 1 and the details of the datasets are listed in "*data.xlsx*" in the supporting information (SI).

Table 1. Summary of the four datasets used in this study

Oxidant	Number of data points	Number of compounds	Reaction conditions	Number of studies
HClO	195	188	pH	29
ClO_2	191	143	pH	32
O_3	759	484	pH	142
$SO_4^{\bullet-}$	557	342	pH, T	33

2.2 Molecular descriptors (MDs) and molecular fingerprints (MFs)

The simplified molecular-input line-entry system (SMILES) of organic contaminants was obtained by the ChemDraw program. The PaDEL program [28] and the RDKit package in Python® were employed to convert SMILES to MDs and MFs, respectively. The MDs of one contaminant include 1444 physicochemical properties and are represented by a vector with a length of 1444. Each property is one feature or an independent variable. Hence, for the MD representation, the total number of features was 1445 (with pH) or 1446 (with pH and T). The MF is a binary vector that encodes chemical structures into 0s and 1s. Readers are referred to our recent papers for more details on how MFs represent chemicals.[15, 29]

2.3 Model development and interpretation

Before model development, we conducted data preprocessing, including missing value imputation, feature scaling, feature selection and/or outlier treatment; and ML algorithm screening. The details of these procedures can be found in Text S1. For each dataset and each representation, after obtaining the optimum ML algorithm, we tuned their hyperparameters by the Bayesian optimization algorithm, which can efficiently explore a large search space. It will determine the next selection based on the last selection. We have previously used this approach to optimize the hyperparameters of a deep neural network and XGBoost.[15] The working mechanism of this approach has been well documented.[30, 31] A 10-fold cross-validation was also applied to the training dataset and the optimum hyperparameters were the ones that achieved the best validation performance. The root mean squared error (RMSE) and R^2 were used as the evaluation metrics for the predictive performance. Lower RMSE and higher R^2 values mean better predictive performance. After obtaining the optimum hyperparameters, the ML algorithms were retrained on the whole training dataset to obtain the final model. The generalization ability of the final model was evaluated on the test dataset, which was never used during the model development.

After the models had been well trained and showed satisfactory predictive performance, we used the SHAP method to interpret the models to check if predictions made by the models are based on a correct understanding of the feature importance. We previously used this method to interpret QSAR models for •OH radicals.[15] The effects of pH, T, and atom groups or MDs on the reactivity (log*k*) were investigated based on the SHAP interpretation results.

2.4 Unified model and knowledge transfer-based model development

To combine the four datasets to form a large dataset, we added a new feature called "Oxidant" to indicate the type of oxidant for a given entry. For these four datasets, their "Oxidant" feature was labeled as "$SO_4^{\bullet-}$", "HClO", "O_3" or "ClO_2". As this new categorical feature should be encoded as a numeric feature, we screened eight encoding methods to select the best one rather than arbitrarily selecting one (Table S1). We then followed the same procedure as described above to develop a unified model (both MF-based and MD-based) on this large dataset, as shown in Figure 1A. It should be noted that we chose not to combine the entire four datasets first and then re-split them. Instead, we combined all the initial training datasets used in developing the single QSAR models to form a combined training dataset. We did the same thing for the individual test datasets to form a combined test dataset, so we can ensure that the generalization ability of the unified model is tested on the same test chemicals as those in the respective single dataset. Hence, any enhancements would be meaningful because the same test chemicals were used. For comparison, in a typical Kaggle competition (https://www.kaggle.com/), even subtle enhancement in the prediction accuracy of a model is desirable and meaningful, which determines if one wins the competition or not, because they are all required to predict the same test dataset.

Figure 1B shows our proposed knowledge transfer approach to developing knowledge transfer-based models. Taking the ClO_2 and O_3 datasets as an example, we

first used the single model developed on the ClO_2 dataset to predict the reactivity of the contaminants in the O_3 dataset toward ClO_2. We then added these predictions as a new input feature to the original O_3 dataset. This modified O_3 dataset thus likely contains some structure-reactivity information from the ClO_2 model. We then developed another model for this new O_3 dataset—referred to as a 'knowledge transfer-based model'—and compared its performance with that of the single model developed on the original O_3 dataset. As described above, the test chemicals remained unchanged when evaluating the performance of the knowledge transfer-based models. Following this approach, we developed a total of 6 knowledge transfer models for three sets of (O_3, ClO_2), (ClO_2, HClO) and (O_3, HClO). The •OH and $SO_4^{\bullet-}$ datasets were not used here because the •OH dataset did not contain reaction conditions while the $SO_4^{\bullet-}$ dataset contains T as a reaction condition.

2.5 Applicability domain (AD) analysis

Because there are reaction conditions in the input, the reported fingerprint-based similarity method cannot be directly applied here.[15] We thus chose a combination of fingerprint-based similarity and range-based methods to determine AD. First, any query chemicals with the reaction conditions (pH and/or T) outside the ranges of pH and/or T of the training dataset were seen as outside of the AD and were not further investigated. For query chemicals whose reaction conditions are within the ranges of pH and/or T of the training dataset, we calculated their similarity to the contaminants in the training dataset based on the Tanimoto index.[32 15] To determine the optimal similarity threshold, we set the chemicals in the test dataset as query chemicals. Any chemicals that were outside the AD (i.e., the similarity values below the threshold) were removed from the test dataset and the $RMSE_{test}$ was recalculated. The optimal threshold is the one that achieved the lowest $RMSE_{test}$.

3. Results and discussion

The detailed results of ML algorithm screening, feature selection and hyperparameter tuning are shown in Text S2. Briefly, different optimum ML algorithms were selected for different datasets, indicating that the optimum ML algorithm is dataset dependent. There is also overfitting in all the ML models with their default hyperparameters, which was alleviated by feature selection and hyperparameter tuning.

3.1 MF versus MD representation and the final individual QSAR models

The statistical comparison between the performances of the two representations are plotted in Figure 2. For all these four oxidants, the training performance for the MD representation is always better than that for the MF-based. However, that is not always the case regarding the generalization ability on the test dataset. For the datasets of $SO_4^{\bullet-}$ and HClO, better predictive performance was achieved on both the training and test datasets for the MD-based models. Hence, the MD-based models were selected as the QSAR models for $SO_4^{\bullet-}$ and HClO. For the datasets of O_3 and ClO_2, the MD-based models showed better predictive performance on the training datasets but worse predictive performance on the test datasets than the MF-based models. This means that overfitting was more serious in the MD-based models. Hence, the final QSAR models for O_3 and ClO_2 were the MF-based models. This result indicated that the optimum chemical representation is dataset-dependent. One possible reason is that the calculated MDs by the commercial PaDEL program might correlate better with the reactivity in the $SO_4^{\bullet-}$ and HClO datasets than with that in the O_3 and ClO_2 datasets. Therefore, it is recommended to screen the optimum chemical representation in future modeling rather than arbitrarily selecting one.

After selecting the appropriate model for each dataset, we compared their performance with previously published ones (results in Text S3). Note that the sizes of our four datasets are much larger. Generally, the predictive performance of a model becomes worse with increasing data size,[15] likely due to the inclusion of more noise

information. For the $SO_4^{\bullet-}$ dataset, our single model showed better predictive performance than previous studies despite the larger sample size. Worse performance was observed for the O_3 dataset because of its significantly larger data size. For the ClO_2 and HClO datasets, we were only able to find one article that reported models for amines[5] and no test performance was provided so it is difficult to make a comparison.

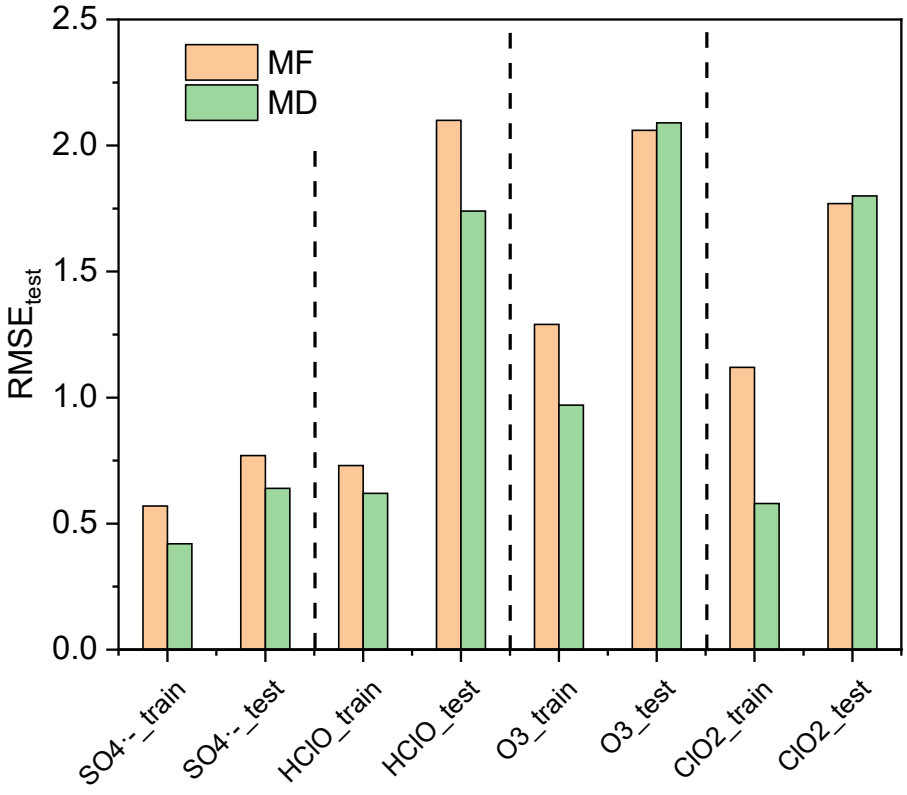


Figure 2. Comparison of the two representations in terms of the predictive performance on the training or the test dataset for the four oxidants.

3.2 Interpretation of the single QSAR models

We interpreted all the single QSAR models to verify (1) if they made predictions based on the correct science and (2) if there are common features used among these models. The latter information may be useful to validate the knowledge transfer strategy.

Figure 3 shows the SHAP interpretation of the MF-based single QSAR models with the top 10 features shown (9 atom groups + pH). The interpretations for the pH effects and the pattern distribution are illustrated in Text S4. The results suggest that all the pH effects were correctly learned, and that different pattern distributions resulted from the employed different ML algorithms. Figure 4 shows the effect of the top 9 atom groups identified in Figure 3 on the log*k*. As shown, the four models share several common atom groups. For example, the 3rd atom group (aromatic carbon) in the $SO_4^{\bullet-}$ model is the same as the 8th atom group in the O_3 model. The number of shared atom groups among these four oxidants is summarized in Table S2. Surprisingly, the learned contributions of some of these atom groups toward log*k* differ significantly among the four datasets. For example, aromatic carbons in the $SO_4^{\bullet-}$ model (3^{rd}) contributed positively to the log*k* while those in the HClO (8^{th}), O_3 (3^{rd}), or ClO_2 (5^{th}) model contributed negatively. The $–NH_2$ group increased the log*k* in the O_3 model (7^{th}) but decreased the log*k* in the ClO_2 model (3^{rd}). However, both aromatic carbons and $–NH_2$ are known electron-donating groups whose presence should lead to higher log*k* values. Therefore, only the $SO_4^{\bullet-}$ model seemed to 'correctly' learn these relationships (thus showing better predictive performance) whereas the HClO, O_3 and ClO_2 models seemed to 'incorrectly' learn some of them (thus showing worse predictive performance).

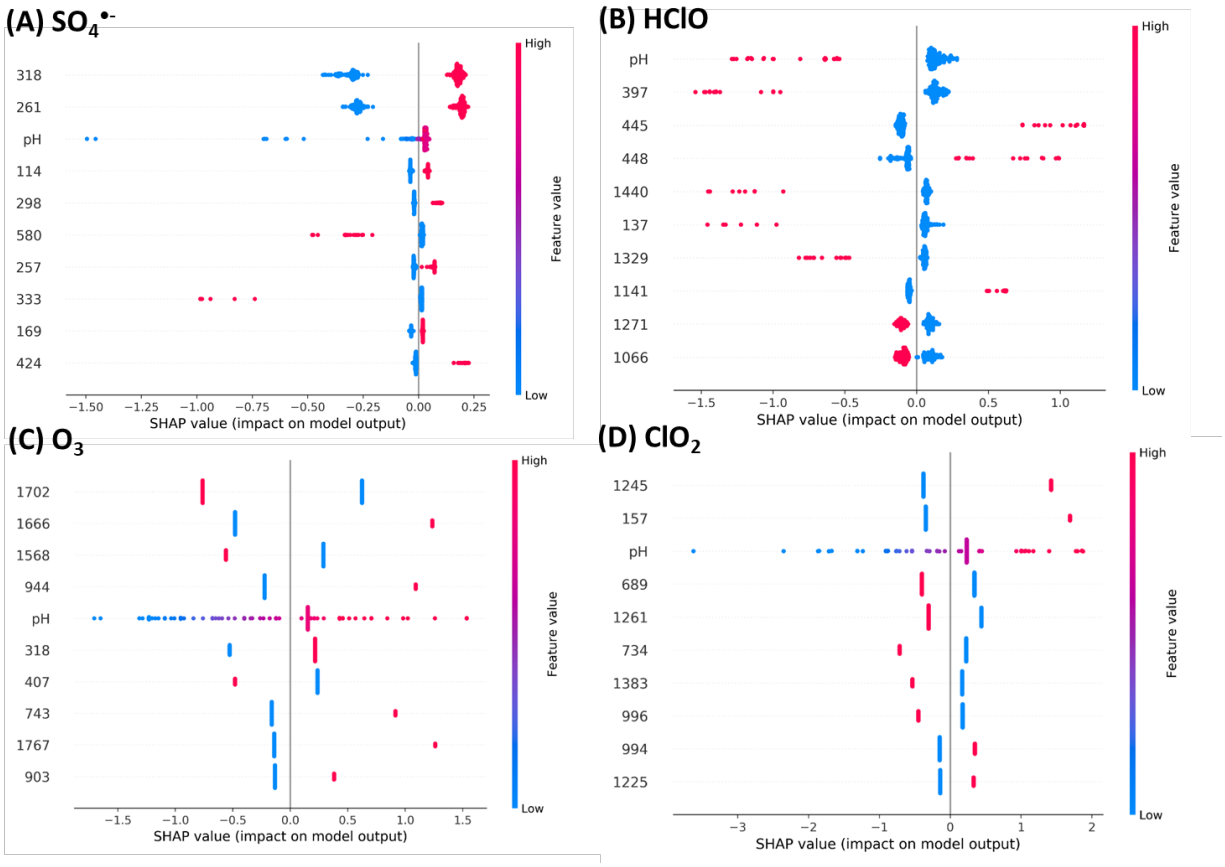
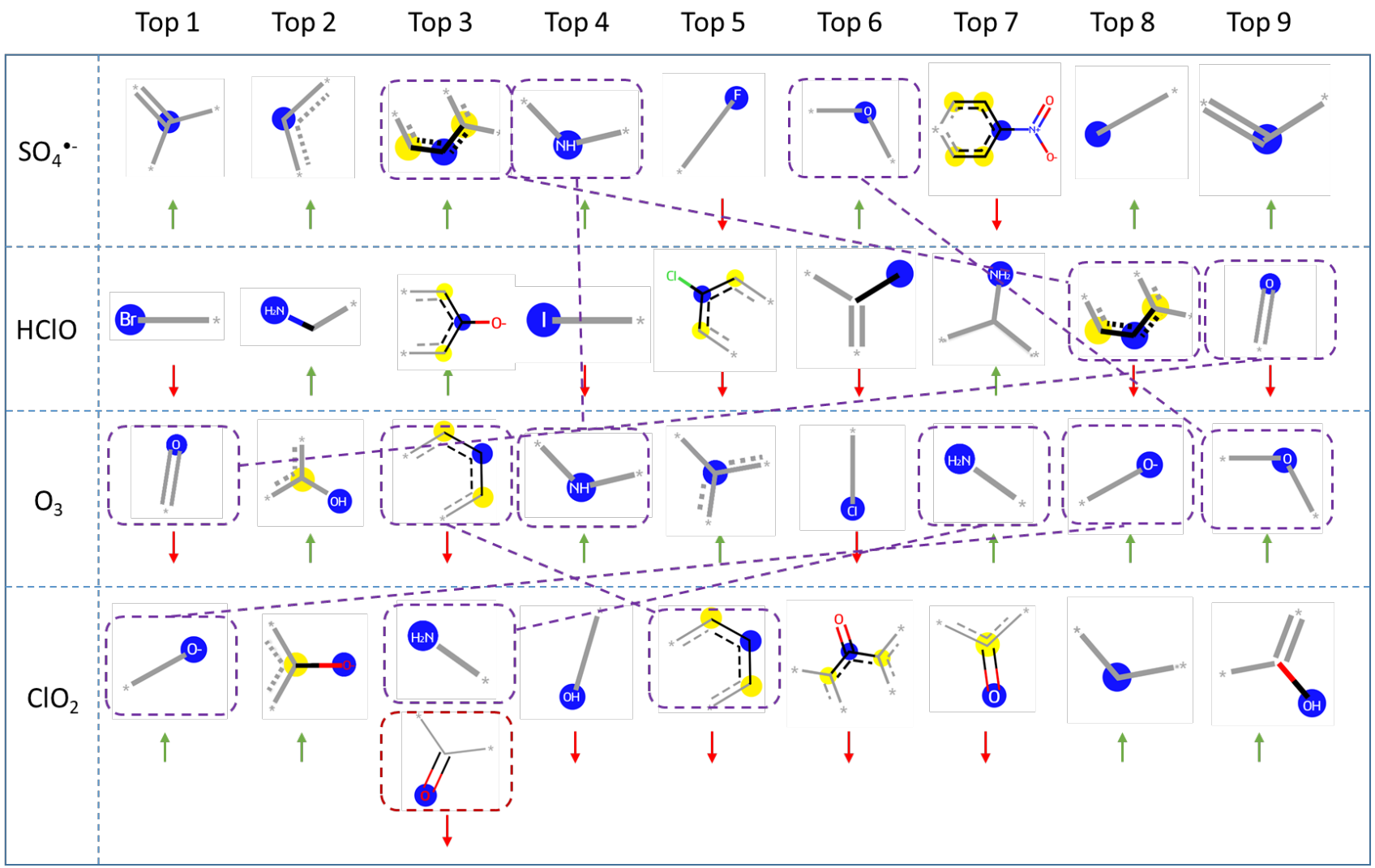


Figure 3. The SHAP interpretation of the MF-based single QSAR models for the four oxidants. The x-axes are the SHAP values and the y-axes are the identified top 10 most influential features. The numbered features, such as 318, 261 and 1702, represent the feature positions in the MFs, with each position representing a certain atom group (see below). MFs are vectors of 1s and 0s; the red color represents 1s in those positions—the presence of an atom group—while blue means 0s—no atom groups in those positions. pH values are continuous values from the minimum to the maximum for different datasets so they are colored from blue to red. A feature with a positive SHAP value means that it can increase the log*k* value; whereas a feature with a negative SHAP value means that it can decrease the log*k* value. The pattern for each feature is composed of the SHAP values for all the chemicals in the dataset that contain that feature. All other SHAP plots in this work follow the same interpretation.


Top 1
Top 2
Top 3
Top 4
Top 5
Top 6
Top 7
Top 8
Top 9
$SO_4^{\bullet-}$
HClO
O_3
ClO_2

300

Figure 4. The effect of the top 9 atom groups shown in Figure 3 on the log*k* values, in which the up and down arrows mean increasing and decreasing the log*k* values, respectively. The same atom groups in different datasets are marked by squares and connected by dotted lines. The $-NH_2$ and carbonyl groups are overlapped at the 3rd position for the ClO_2 dataset. Note that the length of the MFs has been optimized by the Bayesian algorithm to achieve the best predictive performance, but the overlap still happened, indicating the intrinsic limitation associated with the MFs. The blue dots 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.

For the $-NH_2$ group in the ClO_2 dataset, its negative effect on log*k* resulted from its overlap with the electron-withdrawing carbonyl group in the MFs, that is, the position of 689 in the MFs (Figure 3D) is assigned to two atom groups ($—NH_2$ and carbonyl) while carbonyl is a strong electron-withdrawing group that decreases the log*k*. To understand the reason for the observed negative effect of aromatic carbons, we plotted the distribution of experimental log*k* values for the compounds with or without aromatic carbons. Figure S1 shows that the average log*k* value for the compounds containing aromatic carbons in the $SO_4^{\bullet-}$ dataset is greater than that for the compounds not containing aromatic carbons in the same dataset, whereas this trend is reversed in the datasets of HClO, O_3 and ClO_2. This explains why the developed models learned different effects of aromatic carbons on the log*k*. This finding suggests that the average effect of a specific atom group on the chemical reactivity is dataset-dependent, which is expected. For example, when ClO_2 reacts with aliphatic amines, the log*k* value decreases in the following order: tertiary amine > secondary amine > primary amine.[6] If an ML-based QSAR model is developed based on this dataset, a primary amine will be 'learned' to be

an atom group that decreases log*k* because the average experimental log*k* for primary amines is smaller than that for all amines in the dataset, although –NH_2 is a well-known electron-donating group. In other words, the types of chemicals involved in a dataset affects the model-derived positive or negative contribution of an atom group to log*k*. To illustrate the above idea for our datasets, we took the ClO_2 dataset as an example, where it has 36 chemicals containing aromatic carbons (5th atom group for ClO_2 in Figure 4). Among these 36 chemicals, 28 of them (77%) contain electron-donating groups, such as –O^-, —NH_2, or —OH (Table S3), that are stronger in their electron-donating effects than aromatic carbons. As a result, aromatic carbons in the ClO_2 dataset were 'learned' to have negative effects on log*k*. The same explanation can be applied to the HClO and O_3 datasets. We believe that if a dataset is large enough and contains a diverse range of chemical structures, the corresponding ML model should be able to learn the correct effects of various atom groups that match the known chemistry. In other words, the quality of a dataset determines the quality of the corresponding ML model, which is similar to that of traditional QSAR models.

Figure S2 shows the SHAP interpretation of the MD-based single QSAR models. Detailed explanation for them was provided in Text S5. Compared with the MF-based models, fewer MDs (only 1 – 2) were shared among these four models. It is not easy to examine how some of these MDs affected the log*k* because their physicochemical meanings are not readily interpretable.

3.3 Unified models based on the MFs or MDs

To improve the model performance, we combined the four datasets to form a large unified dataset and developed a MF-based unified model (refer to as "MF-UN-1"), following the same procedure as for the single MF-based models. Figure 5A shows better predictive performance of MF-UN-1 on the test dataset (R^2_{test} = 0.76) than all the single models (Text S3) (the RMSE values depend on the ranges of the log*k* values, so they were

not used for comparison), indicating the effectiveness of the unified approach. We then examined its predictive performance on the four single datasets, as shown in Figure 5B. Except for the $SO_4^{\bullet-}$ dataset, the performance of MF-UN-1 is better than that of the respective single models for the other three datasets. Figure 5C plots the distribution of the log*k* values in the four datasets, demonstrating the range of log*k* values for the $SO_4^{\bullet-}$ dataset deviating substantially from that for the other three datasets. This may be the reason that the performance of MF-UN-1 on the $SO_4^{\bullet-}$ dataset was worse.

Figure 5D shows the SHAP interpretation of this unified model and the identified top 6 atom groups (among the top 10 features in Figure 5D, only 6 of them are atom groups). Table S4 shows these atom groups as well as their effects on the logk, in which all of these effects were correctly learned. Although aromatic carbons were not among the top 6 atom groups, we still examined them here because their effects in the HClO, ClO_2 and O_3 datasets, as well as the effect of the $–NH_2$ group in the ClO_2 dataset, were previously learned to decrease the log*k*. For MF-UN-1, interestingly, the effect of $-NH_2$ was 'learned' to be increasing the log*k*, although the aromatic carbons still decreased the log*k* in this unified dataset. For the $SO_4^{\bullet-}$ dataset, the effect of aromatic carbons changed from increasing the log*k* in the individual model to decreasing the log*k* in MF-UN-1, which should be the reason for the worse predictive performance of the unified model on the $SO_4^{\bullet-}$ dataset. In contrast, the effect of the $-NH_2$ group in the ClO_2 dataset changed from decreasing the log*k* in the individual model to increasing the log*k* in MF-UN-1, so the predictive performance improved (Figure 5B). The effects of these two groups on the log*k* are the same for HClO and O_3 datasets before and after combining the datasets, but the predictive performance became better, which may be due to some unknown synergetic effects or similar "correction" effects of atom groups that are not among top 9.

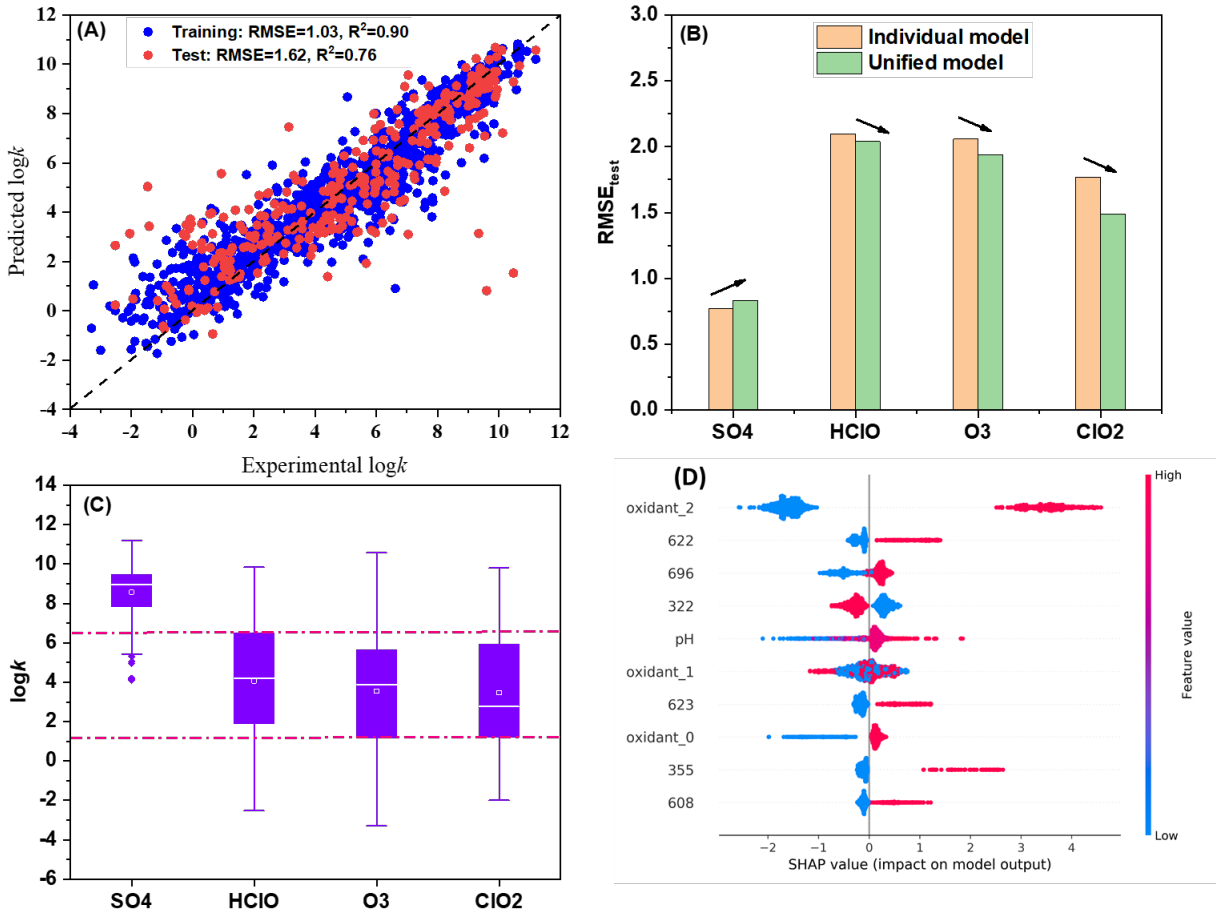


Figure 5. The predictive performance of the unified model on the training and test datasets for the unified dataset (A) and the single datasets (B); (C) the ranges of log*k* values for the single datasets; and (D) the SHAP interpretation of the unified model, in which the x-axis is the SHAP value and the y-axis is the features. The features of 'Oxidant_1', 'Oxidant_2', and 'Oxidant_3' are the encoded features for these four oxidants and they can take only values of 0 or 1. Their different combinations (i.e., ['Oxidant_1', 'Oxidant_2', 'Oxidant_3']) represent different oxidants, such as [0, 0, 1] for HClO or [0, 1, 0] for O_3. Other features represent atom groups and are listed in Table S4

Figure S3 shows the performance of the unified model based on the MD representation. This unified model was developed following the same procedure as for the single MD-based models. The $RMSE_{test}$ (1.67, Figure S3A) is slightly higher than that

of MF-UN-1 (1.62); the predictions made by the MD-based unified model marginally improved for O_3 and ClO_2, but marginally decreased for $SO_4^{\bullet-}$ and HClO (Figure S3B). This improvement was less than by MF-UN-1 (Figure 3B) and the overfitting trend was more obvious, as suggested by the larger difference in the RMSE values between the training and test datasets (Figure S3C). The SHAP patterns in Figure S5D are similar to those of the single MDs-based models (Figure S2). Overall, the MD representation did not perform as well as the MF representation when developing the unified model, so we only focus on the MF representation below.

As mentioned above, the range of log*k* values for $SO_4^{\bullet-}$ is quite different from those of the other three datasets (Figure 3C), which may be one reason that the predictive performance of MF-UN-1 did not improve for $SO_4^{\bullet-}$. To test this idea, we combined the $SO_4^{\bullet-}$ dataset with a reported $OH^{\bullet}$ dataset to form a large dataset because their log*k* values fall in the same range (Figure S4C). The $OH^{\bullet}$ dataset contains 1086 chemicals and was previously used successfully to develop ML-based QSAR models.[14, 15] We then developed another MF-based unified model (refers to as "MF-UN-2") on this dataset and Figure S4A shows the R^2_{test} = 0.68. Figure S4B suggests that the predictive performance of MF-UN-2 for $SO_4^{\bullet-}$ became much better than the single model while that for $OH^{\bullet}$ became worse. As the SHAP interpretation of MF-UN-2 shown in Figure S4D, the effect of the identified top 8 atom groups on the log*k* were all correctly learned (only 8 of the top 10 features are atom groups) (Table S5). This worse performance for •OH was probably because the additional fixed T (25 ^{o}C) and pH (7) conditions were added into the •OH dataset to combine with the $SO_4^{\bullet-}$ dataset, which might have introduced noise information to the model, although future work is needed to understand the exact reason. For prediction purposes, MF-UN-2 can be used for $SO_4^{\bullet-}$ while the reported MF-based single model is still recommended for •OH.

Finally, we combined all of these five datasets to form the largest dataset to develop another MF-based unified model (refers to as "MF-UN-3"). Figure S5A shows that the R^2_{test} reached 0.82. While the predictions for $SO_4^{\bullet-}$, HClO and ClO_2 became better, those for •OH and O_3 became slightly worse (Figure S5B). Table S6 shows the effects of the identified top 8 atom groups (only 8 of the top 10 features are atom groups) based on the SHAP plot of Figure S5C, and all of them were correctly learned. The marginally worse predictive performance for the •OH dataset is explained above, but the marginally worse predictive performance than MF-UN-1 for the O_3 dataset is unexpected. We do not have a good explanation for this yet. These results suggested that it is not always better to combine datasets to achieve better predictive performance.

3.4 Knowledge transfer models

Figure 6 shows the predictive performance of different knowledge transfer models that were developed based on our proposed approach shown in Figure 1B. The $SO_4^{\bullet-}$ dataset was not used because it contains not only pH but also T, while other three datasets only contain pH. The models developed based on these three datasets cannot make predictions for contaminants under different T.

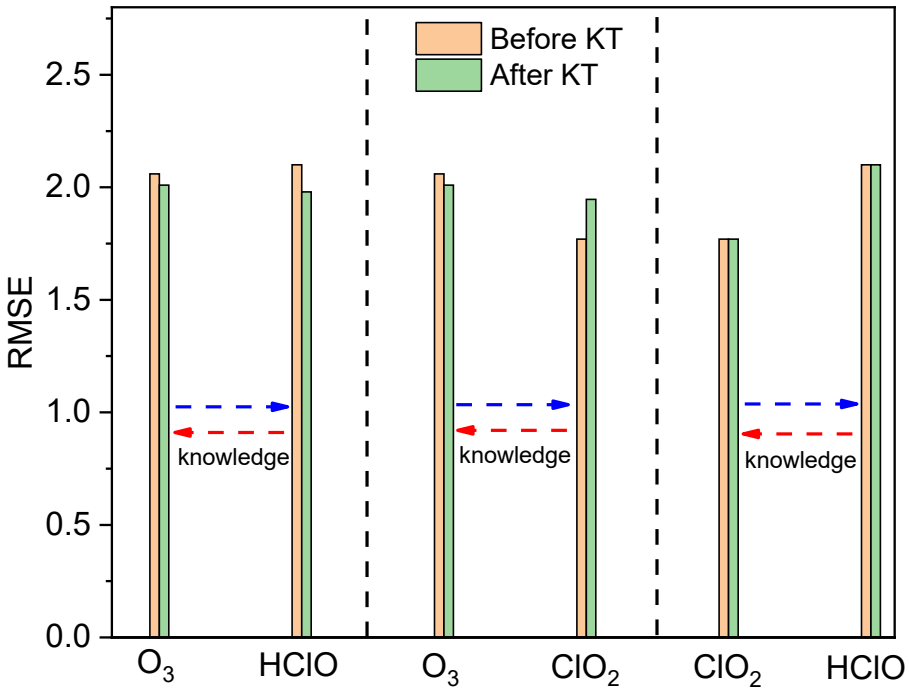


Figure 6. The predictive performance of different ML models before and after knowledge transfer (KT)

There are three distinct scenarios for these knowledge transfer models. For O_3/HClO, both of the knowledge transfer models show better predictive performance than the original models before the transfer. There is one shared atom group (carbonyl) among the top 9 atom groups between O_3 and HClO (Figure 4) and the effect of this atom group was consistent (i.e., decreasing the logk) between the two datasets. Moreover, the predictive performance of the single models for O_3 and HClO was similar ($RMSE_{test}$ 2.06 for O_3 and 2.10 for HClO). Both of these two factors should have contributed to the effectiveness of knowledge transfer. For O_3/ClO_2, the knowledge transfer model for O_3 became better after receiving knowledge from the ClO_2 model, while the knowledge transfer model for ClO_2 became worse. There are 3 atom groups shared between O_3 and ClO_2, but the effects of $–NH_2$ in these two datasets are opposite (Figure 4). Moreover, the predictive performance of the single model for ClO_2 ($RMSE_{test}$ 1.77) is better than that for

446 O_3 ($RMSE_{test}$ 2.06), so the information transferred from O_3 to ClO_2 has more uncertainties,
447 which should have led to the worse performance. For ClO_2/HClO, no change in the
448 predictive performance was observed for both oxidants. This is expected because there
449 are no shared atom groups between these two datasets (Figure 4). These results indicated
450 that the effectiveness of our knowledge transfer approach is determined by if there is
451 consistent knowledge shared by the single models as well as their respective predictive
452 performance.

453

454 3.5 The final QSAR models and their AD determination

455 For the four oxidants, we ranked all the developed models in terms of the
456 predictive performance and finally obtained the optimal QSAR models, shown in Table
457 2. Both the unified models and transfer learning models outperformed all the individual
458 models and were selected as the final models, validating the effectiveness of our
459 proposed two approaches. We next determined their ADs, as shown in Table 2. For each
460 model, with increasing threshold value, more contaminants were identified as outside
461 AD and the recalculated $RMSE_{test}$ first decreased and then increased. The optimal
462 threshold values for these four datasets are bolded in Table 2. For a query compound, if
463 its similarity to the contaminants in the training dataset is above the threshold value, the
464 models can provide a reliable prediction for its reactivity toward one of these four
465 oxidants.

466

467 Table 2. The final selected models for each dataset and their AD determination

Dataset	Best Model	Best $RMSE_{test}$	Threshold value	# of contaminants outside AD	Recalculated $RMSE_{test}$
$SO_4^{\bullet-}$	MF-UN-2	0.703	0.50	0	0.703

			0.60	***1***	***0.699***
			0.70	2	0.700
HClO	Knowledge Transfer model (O_3-HClO)	1.982	0.28	0	1.982
			0.30	1	1.955
			0.42	***2***	***1.895***
			0.43	3	1.906
O_3	MF-UN-1	1.942	0.50	0	1.942
			0.55	1	1.909
			0.56	***3***	***1.906***
			0.62	4	1.912
ClO_2	MF-UN-3	1.465	***0.66***	***0***	***1.465***
			0.67	1	1.468
			0.83	2	1.486

4. Environmental significance

In this study, we investigated QSAR models for datasets that are different but share some similarities (i.e., oxidation reactions). Previous studies viewed these datasets independently, whereas we tried to obtain relationships among them to enhance the predictive performance of the QSAR models. We proposed two approaches—combining individual datasets to form a large, unified dataset and transferring knowledge between individual datasets. When developing single ML models using these single datasets, we found that (1) the optimal ML algorithm is dataset dependent. Screening ML algorithms from several candidate algorithms is recommended and simpler ML algorithms are preferred if they show similar predictive performance as complex ones; and (2) the optimal representation for contaminants is also dataset-dependent because some representations may not capture the key features of the dataset. Combining similar datasets to form a large dataset and developing a unified model can generally improve the predictive performance on the individual datasets, because some 'wrongly' learned knowledge based on a smaller dataset (e.g., bias of the dataset) may be corrected this way. In other words, data bias can be mitigated by increasing the sample size. Knowledge

transfer is effective when there is consistent knowledge shared between the two datasets and when the single models themselves have good predictive performance. Overall, this study provided new insights into developing ML-based QSAR models for small datasets. We demonstrated that there are synergistic effects among similar datasets, which can be leveraged to boost the predictive performance of QSAR models.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant #CHEM-1808406.

Supporting Information Available

The Supporting Information includes the data preprocessing, 5 candidate scaler methods, 8 encoding methods, 7 outlier detection methods and 16 ML algorithms, the working mechanism of recursive feature selection, the screening results for scaler, encoding and ML algorithms, the overfitting control, the specific number of MDs used and their names, the correlation plots of single QSAR models, the comparison of the single QSAR models with previously published ones, and other related tables and figures for model interpretations, which is available free of charge online.

References

1. von Gunten, U., Oxidation Processes in Water Treatment: Are We on Track? *Environmental Science & Technology* **2018,** *52*, (9), 5062-5075.
2. Deng, Y.; Ezyske, C. M., Sulfate radical-advanced oxidation process (SR-AOP) for simultaneous removal of refractory organic contaminants and ammonia in landfill leachate. *Water research* **2011,** *45*, (18), 6189-6194.
3. Acero, J. L.; Stemmler, K.; Von Gunten, U., Degradation kinetics of atrazine and its degradation products with ozone and OH radicals: a predictive tool for drinking water treatment. *Environmental science & technology* **2000,** *34*, (4), 591-597.
4. Kwon, M.; Kim, S.; Jung, Y.; Hwang, T.-M.; Stefan, M. I.; Kang, J.-W., The impact of natural variation of OH radical demand of drinking water sources on the optimum operation of the UV/H2O2 process. *Environmental science & technology* **2019,** *53*, (6), 3177-3186.
5. Chin, A.; Bérubé, P., Removal of disinfection by-product precursors with ozone-UV advanced oxidation process. *Water research* **2005,** *39*, (10), 2136-2144.

6. Gan, W.; Ge, Y.; Zhong, Y.; Yang, X., The reactions of chlorine dioxide with inorganic and organic compounds in water treatment: kinetics and mechanisms. *Environmental Science: Water Research & Technology* **2020,** *6*, (9), 2287-2312.
7. Lee, Y.; von Gunten, U., Quantitative structure–activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. *Water Research* **2012,** *46*, (19), 6177-6195.
8. Su, H.; Yu, C.; Zhou, Y.; Gong, L.; Li, Q.; Alvarez, P.; Long, M., Quantitative structure–activity relationship for the oxidation of aromatic organic contaminants in water by TAML/H2O2. *Water Research* **2018,** *140*, 354-363.
9. Cheng, Z.; Yang, B.; Chen, Q.; Gao, X.; Tan, Y.; Ma, Y.; Shen, Z., A Quantitative-Structure-Activity-Relationship (QSAR) model for the reaction rate constants of organic compounds during the ozonation process at different temperatures. *Chemical Engineering Journal* **2018,** *353*, 288-296.
10. Luo, S.; Wei, Z.; Spinney, R.; Villamena, F. A.; Dionysiou, D. D.; Chen, D.; Tang, C.-J.; Chai, L.; Xiao, R., Quantitative structure–activity relationships for reactivities of sulfate and hydroxyl radicals with aromatic contaminants through single–electron transfer pathway. *Journal of Hazardous Materials* **2018,** *344*, 1165-1173.
11. Xiao, R.; Ye, T.; Wei, Z.; Luo, S.; Yang, Z.; Spinney, R., Quantitative structure–activity relationship (QSAR) for the oxidation of trace organic contaminants by sulfate radical. *Environmental science & technology* **2015,** *49*, (22), 13394-13402.
12. Li, C.; Wei, G.; Chen, J.; Zhao, Y.; Zhang, Y.-N.; Su, L.; Qin, W., Aqueous OH Radical Reaction Rate Constants for Organophosphorus Flame Retardants and Plasticizers: Experimental and Modeling Studies. *Environmental Science & Technology* **2018,** *52*, 2790-2799.
13. Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H., A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. *Journal of hazardous materials* **2020,** *383*, 121141.
14. Zhong, S.; Hu, J.; Yu, X.; Zhang, H., Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chemical Engineering Journal* **2021,** *408*, 127998.
15. Zhong, S.; Zhang, K.; Wang, D.; Zhang, H., Shedding Light On "Black Box" Machine Learning Models for Predicting the Reactivity of HO• Radicals toward Organic Compounds. *Chemical Engineering Journal* **2020**, 126627.
16. Borhani, T.; Saniedanesh, M.; Bagheri, M.; Lim, J., QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Research* **2016,** *98*, 344-353.
17. Cheng, Z.; Yang, B.; Chen, Q.; Shen, Z.; Yuan, T., Quantitative relationships between molecular parameters and reaction rate of organic chemicals in Fenton process in temperature range of 15.8 °C - 60 °C. *Chemical Engineering Journal* **2017,** *350*, 534-540.
18. Sudhakaran, S.; Amy, G. L., QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. *Water Research* **2013,** *47*, (3), 1111-1122.
19. Ye, T.; Wei, Z.; Spinney, R.; Tang, C.-J.; Luo, S.; Xiao, R.; Dionysiou, D. D., Chemical structure-based predictive model for the oxidation of trace organic contaminants by sulfate radical. *Water Research* **2017,** *116*, 106-115.
20. Huang, Y.; Li, T.; Zheng, S.; Fan, L.; Su, L.; Zhao, Y.; Xie, H.-B.; Li, C., QSAR modeling for the ozonation of diverse organic compounds in water. *Science of The Total Environment* **2020,** *715*, 136816.
21. Gupta, S.; Basant, N., Modeling the reactivity of ozone and sulphate radicals towards organic chemicals in water using machine learning approaches. *RSC advances* **2016,** *6*, (110), 108448-108457.
22. Gerrity, D.; Gamage, S.; Jones, D.; Korshin, G. V.; Lee, Y.; Pisarenko, A.; Trenholm, R. A.; Von Gunten, U.; Wert, E. C.; Snyder, S. A., Development of surrogate correlation models to predict trace

organic contaminant oxidation and microbial inactivation during ozonation. *Water Research* **2012,** *46*, (19), 6257-6272.
23. Lee, Y.; Kovalova, L.; McArdell, C. S.; von Gunten, U., Prediction of micropollutant elimination during ozonation of a hospital wastewater effluent. *Water research* **2014,** *64*, 134-148.
24. dos Santos, D. J. V. A.; Newton, A. S.; Bernardino, R.; Guedes, R. C., Substituent effects on O–H and S–H bond dissociation enthalpies of disubstituted phenols and thiophenols. *Int. J. Quantum Chem* **2008,** *108*, (4), 754-761.
25. Karelson, M.; Lobanov, V. S.; Katritzky, A. R., Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical Reviews* **1996,** *96*, (3), 1027-1044.
26. Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B., Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099* **2017**.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012; 2012; pp 1097-1105.
28. Yap, C. W., PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011,** *32*, (7), 1466-74.
29. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010,** *50*, (5), 742-754.
30. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G., Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021,** *590*, (7844), 89-96.
31. Dewancker, I.; McCourt, M.; Clark, S., Bayesian Optimization for Machine Learning: A Practical Guidebook. *arXiv preprint arXiv:1612.04858* **2016**.
32. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **2015,** *7*, (1), 1-13.