Environmental Science & Technology

pubs.acs.org/est

Article

# Machine Learning-Assisted QSAR Models on Contaminant Reactivity Toward Four Oxidants: Combining Small Data Sets and Knowledge Transfer

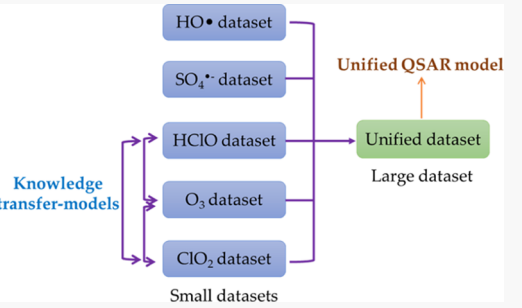
Shifa Zhong, Yanping Zhang,* and Huichun Zhang*

Cite This: https://doi.org/10.1021/acs.est.1c04883

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** To develop predictive models for the reactivity of organic contaminants toward four oxidants—$SO_4^{\bullet-}$, HClO, $O_3$, and $ClO_2$—all with small sample sizes, we proposed two approaches: combining small data sets and transferring knowledge between them. We first merged these data sets and developed a unified model using machine learning (ML), which showed better predictive performance than the individual models for HClO ($RMSE_{test}$: 2.1 to 2.04), $O_3$ (2.06 to 1.94), $ClO_2$ (1.77 to 1.49), and $SO_4^{\bullet-}$ (0.75 to 0.70) because the model "corrected" the wrongly learned effects of several atom groups. We further developed knowledge transfer models for three pairs of the data sets and observed different predictive performances: improved for $O_3$ ($RMSE_{test}$: 2.06 to 2.01)/HClO (2.10 to 1.98), mixed for $O_3$ (2.06 to 2.01)/$ClO_2$ (1.77 to 1.95), and unchanged for $ClO_2$ (1.77 to 1.77)/HClO (2.1 to 2.1). The effectiveness of the latter approach depended on whether there was consistent knowledge shared between the data sets and on the performance of the individual models. We also compared our approaches with multitask learning and image-based transfer learning and found that our approaches consistently improved the predictive performance for all data sets while the other two did not. This study demonstrated the effectiveness of combining small, similar data sets and transferring knowledge between them to improve ML model performance.




**KEYWORDS:** *$ClO_2$, HClO, knowledge transfer, multitask learning, oxidation rate constants, ozone, QSARs, sulfate radicals*



## 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 such QSAR models, different chemical representations, such as molecular descriptors (MDs, physicochemical and structural properties),[16] molecular fingerprints (MFs, binary vectors),[13] or molecular images,[14] can be combined with different modeling 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 complex, nonlinear relationships may exist between the contaminant representations (e.g., MDs) and the reaction rate constants. To handle nonlinear relationships and increasingly diverse contaminants, ML has received increasing attention because of its powerful modeling ability. For example, Huang et al. reported a better performance of a support vector machine-based model in 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 data set of •OH reactivity.[15]

Received: July 28, 2021
Revised: December 3, 2021
Accepted: December 7, 2021

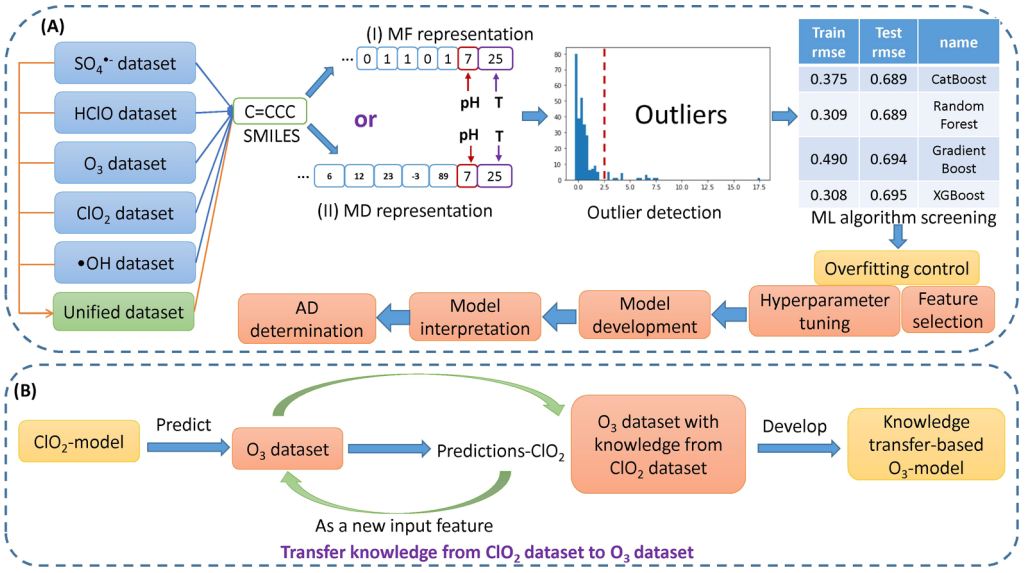



**Figure 1.** Workflow of this study. (A) Single and unified model development based on MFs or MDs. (B) Illustration of how KT is achieved by an example of transferring knowledge from the $ClO_2$ data set to the $O_3$ data set. Briefly, we used the individual model trained on the $ClO_2$ data set ($ClO_2$-model) to predict the reactivity of the chemicals in the $O_3$ data set toward $ClO_2$ (predictions-$ClO_2$), added these predictions to the $O_3$ data set as a new feature, and developed a KT-based model on this modified $O_3$ data set. More details are in Section 2.4. Additional comparison between the proposed KT and the image-based transfer learning approaches is shown in Text S1 in the Supporting Information.

However, ML algorithms, especially deep neural networks (DNNs), often need a massive amount of data,[24] whereas data scarcity is a common issue in chemistry data,[25] such as when developing QSAR models for rate constants toward different oxidants, for example, only 85 samples in a data set of $SO_4^{\bullet-}$ radicals[21] or 136 samples in an $O_3$ data set.[20] Note that the data scarcity (or small data sets) is in comparison to big data, and the involved sample size may be large in reference to classical QSARs. Yet, it is impractical to experimentally measure rate constants (log $k$) for a large number of contaminants toward different oxidants to increase the sample size. To still take advantage of ML algorithms when developing predictive models[26] for small data sets of contaminant oxidative reactivity toward common oxidants, we here propose a simple and effective approach—combining small data sets for different oxidants to form a larger data set. This combined data set contains samples for five common oxidants, including •OH, $SO_4^{\bullet-}$, $O_3$, $ClO_2$, and HClO. Previous studies treated these small data sets 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 demonstrated 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.[27] Huang et al. reported that the energy of the highest occupied molecular orbital ($E_{HOMO}$) 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.[20,28] Compounds with higher $E_{HOMO}$ are oxidized by $O_3$ with faster rates due to their stronger electron-donating ability. Furthermore, oxidants oxidize contaminants primarily using three mechanisms: (1) hydrogen abstraction, (2) electron transfer, and (3) addition or substitution reactions. For each oxidant, at least one mechanism is applicable. For example, both $SO_4^{\bullet-}$ and •OH radicals follow all three mechanisms for different contaminants,[13,29,30] while at least two of them (hydrogen abstraction and electron transfer) are involved in the oxidation by $ClO_2$ or $O_3$.[6,31] Because the shared science may be transferred from one data set to another, combining small data sets to form a larger data set 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 data set—has never been investigated before in developing QSAR models for contaminant reactivity.

Transfer learning, widely used in computer vision, is another popular approach to addressing the data scarcity issue.[32] Transfer learning refers to pretraining a model on a large data set and then fine-tuning this pretrained model on a smaller but similar data set. 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 pretraining a convolutional NN (CNN) model on the ImageNet data set, it can considerably increase the generalization ability of the QSAR models.[14] The ImageNet data set is, however, quite different from the contaminant image data set.[33] This transfer learning approach (referred to as image-based transfer learning) is also only limited to CNN algorithms. Recently, Goh et al. developed a ChemNet which was pretrained on images of ∼1 million chemicals. ChemNet also involved a CNN and image data but changed the pretrained data set from the ImageNet to a chemical-based data set.[25] However, the ChemNet was only trained on chemical images, so it cannot


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

be directly applied to data sets that contain other information such as reaction conditions.

For the data sets of $^{\bullet}OH$, $SO_4^{\bullet-}$, 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. To address the data scarcity issue when developing ML models for each individual data set, it would be interesting and beneficial to investigate whether the shared knowledge between any two similar, smaller data sets is transferable or not. Moreover, tabular data are common in the environmental field but rarely handled by CNN.[26] Hence, how to effectively transfer knowledge without using CNN algorithms and image data is still challenging. We here proposed a knowledge transfer (KT) approach for non-CNN algorithms (e.g., tree-based algorithms) and non-image data (e.g., tabular data) (Figure 1B).

In this study, we compiled the largest four data sets for four common oxidants, namely $SO_4^{\bullet-}$, HClO, $O_3$, and $ClO_2$, by including the reaction conditions, that is, pH and/or temperature (T). The reaction conditions were seldom considered in previous studies, but including them can significantly increase the sample size. For example, we can collect several samples of phenol under different pH conditions, but we only have one sample of phenol if we fix the pH. Two chemical representations, that is, MDs and 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 data sets to form a large data set and developed a unified QSAR model. The effect of this operation on the predictive performance of each data set was investigated. We next used the KT approach to develop KT-based models (Figure 1B, more details in Section 2.4) and compared their predictive performances with that of the respective single model. We also compared the model performance between the two proposed approaches and the widely used multitask learning and transfer learning approaches. The overall workflow of this study is summarized in Figure 1.

## 2. MATERIALS AND METHODS

**2.1. Data Sets.** 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 $T$). The number of studies we collected is listed in Table 1, in which if there were QSAR studies they were considered but the original sources were not. All these studies, including the original sources for the QSAR studies, are listed in the excel file "*data.xlsx*" in the Supporting Information. 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] Text S2 shows in detail how the pH effects on the log $k$ of ionizable compounds were modeled. Moreover, we can increase the sample size by including the reaction conditions. This approach allows more data to be collected than the traditional approach, where only data under certain conditions can be collected. There are no strict criteria for what data should be collected. Rather, all articles that reported contaminant reactivity toward any of these four oxidants were included. 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 to smooth noise in the samples. The outliers, that is, abnormal reactivity, were detected by seven outlier detection algorithms (Table S1) and removed. A summary of these four data sets is listed in Table 1 and the details of the data sets are listed in "*data.xlsx*" in the Supporting Information.

**Table 1. Summary of the Four Data Sets Used in This Study**

| oxidant | # of data points | # of compounds | reaction conditions | # 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 and Molecular Fingerprints.** The simplified molecular-input line-entry system (SMILES) of organic contaminants was obtained using the ChemDraw program. The PaDEL program[34] 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 used here is the Morgan Fingerprint, which is a binary vector that encodes chemical structures into 0s and 1s. The MF was obtained using the RDKit package in Python with the command "*AllChem.GetMorganFingerprintAsBitVect()*". Readers are referred to our recent papers for more details on how MFs represent chemicals.[15,35] Both MD and MF are one-dimensional (1D) vectors of a certain length. The conditions, such as pH and T, can be directly attached to the end of the vector to form a longer 1D vector.

**2.3. Model Development and Interpretation.** Briefly, ML model development is to use ML algorithms to link the MDs or MFs of organic contaminants—the inputs—with their corresponding reactivity—the output. The obtained models can then make predictions for the reactivity of organic contaminants based on their MDs or MFs. Before model development, we conducted data preprocessing, including missing value imputation, feature scaling, feature selection, outlier treatment, and ML algorithm screening. The details of these procedures can be found in Text S3. The number of MDs and their names are listed in Texts S4 and S5 in the Supporting Information. For each data set and each representation, after obtaining the optimum ML algorithm, we tuned their hyperparameters using 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 DNN and XGBoost.[15] The working mechanism of this approach has been well documented.[36,37] A 10-fold cross-validation was also applied to the training data set (not the entire data set) and the optimum hyperparameters were the ones that achieved the best validation performance (average performance on the 10 validation sets). The root


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 2. Final Models Used for Different Data Sets and Their Performance[a]

| models | data sets | $RMSE_{train}$ | $R_{train}^2$ | $RMSE_{test}$ | $R_{test}^2$ | ML algorithm | scaler | encoder |
|---|---|---|---|---|---|---|---|---|
| MF-based | $SO_4^{\bullet-}$ | 0.52 | 0.81 | 0.75 | 0.62 | CatBoost | | |
| | HClO | 0.73 | 0.93 | 2.10 | 0.43 | CatBoost | | |
| | **$O_3$** | **1.29** | **0.76** | **2.06** | **0.46** | **Ridge** | | |
| | **$ClO_2$** | **1.12** | **0.88** | **1.77** | **0.49** | **Ridge** | | |
| MD-based | **$SO_4^{\bullet-}$** | **0.42** | **0.88** | **0.64** | **0.72** | **RF** | **MAS** | |
| | **HClO** | **0.62** | **0.94** | **1.74** | **0.60** | **ET** | **MAS** | |
| | $O_3$ | 0.97 | 0.87 | 2.09 | 0.45 | CatBoost | MMS | |
| | $ClO_2$ | 0.58 | 0.97 | 1.80 | 0.47 | CatBoost | RS | |
| MF-UN-1 | combined data sets[b] | 1.03 | 0.90 | 1.62 | 0.76 | XGBoost | | BDE |
| MD-UN | combined data sets[b] | 0.73 | 0.95 | 1.67 | 0.75 | CatBoost | MAS | HE |
| MF-UN-2 | combined data sets[c] | 0.41 | 0.82 | 0.58 | 0.68 | CatBoost | | OH |
| MF-UN-3 | combined data sets[d] | 0.70 | 0.95 | 1.48 | 0.82 | RF | | SE |
| KT-1 | $O_3$/HClO | | $RMSE_{test}$: $O_3$ 2.01, HClO 1.98 | | | Ridge/CatBoost | | |
| | | | $R^2$: $O_3$ 0.49, HClO 0.49 | | | | | |
| KT-2 | $O_3$/$ClO_2$ | | $RMSE_{test}$: $O_3$ 2.01, $ClO_2$ 1.95 | | | CatBoost/Ridge | | |
| | | | $R^2$: $O_3$ 0.49, $ClO_2$ 0.38 | | | | | |
| KT-3 | $ClO_2$/HClO | | RMSEtest: $ClO_2$ 1.77, HClO 2.1 | | | Ridge/CatBoost | | |
| | | | $R^2$: $ClO_2$ 0.49, HClO 0.42 | | | | | |

[a]The bolded rows are for the final individual QSAR models for that data set. [b]Combining the data sets of $SO_4^{\bullet-}$/$O_3$/$ClO_2$/HClO. [c]Combining the data sets of $^{\bullet}OH$/$SO_4^{\bullet-}$. [d]Combining the data sets of $^{\bullet}OH$/$SO_4^{\bullet-}$/$O_3$/$ClO_2$/HClO. RF: random forest; ET: extra trees; MAS: MinAbsScaler; MMS: MinMaxScaler RS: RobustScaler; BDE: BackwardDifferenceEncoder; HE: HelmerEncoder; OH: OneHotEncoder; SE: SumEncoder.

mean squared error (RMSE) and $R^2$ values 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 (Table S2), the ML algorithms were retrained on the whole training data set (not cross-validation anymore) to obtain the final model (hence, we only have one RMSE or one $R^2$ value for each training set). The generalization ability of the final model was evaluated on the test data set, which was never used during the model development (we also only have one RMSE or $R^2$ value for each test set).

After the models had been well trained to show 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 $^{\bullet}OH$ 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. Text S6 shows how the atom groups from the MFs were extracted.

**2.4. Unified Model and KT-based Model Development.** To combine the four data sets to form a large data set, we added a new feature called "Oxidant" to indicate the type of oxidant for a given entry. For these four data sets, 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 (Figure 1A) to develop a unified model (both MF-based and MD-based) on this large data set. It should be noted that we chose not to combine the entire four data sets first and then re-split them. Instead, we combined all the initial training data sets used in developing the single QSAR models to form a combined training data set. We did the same thing for the individual test data sets to form a combined test data set, so we could ensure that the generalization ability of the unified model was tested on the same test chemicals as those in the respective single data set. 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 data set.

Figure 1B shows our proposed KT approach to developing KT-based models. Taking the $ClO_2$ and $O_3$ data sets as an example, we first used the single model developed on the $ClO_2$ data set to predict the reactivity of the contaminants in the $O_3$ data set toward $ClO_2$. We then added these predictions as a new input feature to the original $O_3$ data set. This modified $O_3$ data set thus likely contains some structure-reactivity information from the $ClO_2$ model. We then developed another model for this new $O_3$ data set—referred to as a "KT-based model"—and compared its performance with that of the single model developed on the original $O_3$ data set. As described above, the test chemicals remained unchanged when evaluating the performance of the KT-based models. Following this approach, we developed a total of six KT models for three sets of ($O_3$, $ClO_2$), ($ClO_2$, HClO), and ($O_3$, HClO). The $^{\bullet}OH$ and $SO_4^{\bullet-}$ data sets were not used here because the $^{\bullet}OH$ data set did not contain reaction conditions while the $SO_4^{\bullet-}$ data set 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 the AD. First, any query chemicals with the reaction conditions (pH and/or $T$) outside the ranges of pH and/or $T$ of the training data set 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 data set, we calculated the similarity between their MFs and those of the contaminants in the training data set based on the Tanimoto index.[15,38] Please refer to our recent paper about how to calculate the Tanimoto


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

index.[15] To determine the optimal similarity threshold, we set the chemicals in the test data set as the query chemicals. Any chemicals that were outside the AD (i.e., the similarity values below the threshold) were removed from the test data set 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 the ML algorithm screening, feature selection, and hyperparameter tuning are shown in Text S4. Briefly, different optimum ML algorithms were selected for different data sets, indicating that the optimum ML algorithm is data set-dependent. There is also overfitting in all the ML models with their default hyperparameters, which was alleviated by feature selection and/or hyperparameter tuning. For convenience, we summarized all the models used in this work in Table 2, which were all developed following a similar procedure, such as ML screening, feature selection, and/or 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

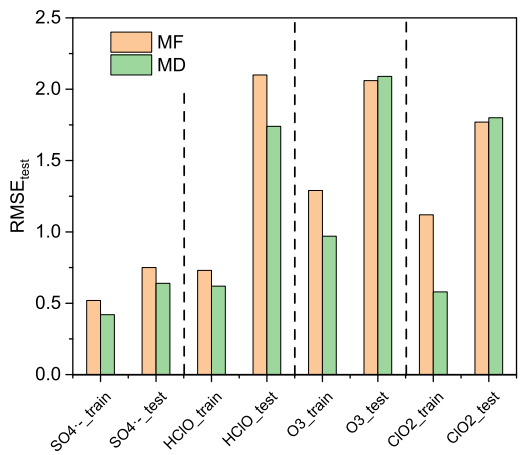



**Figure 2.** Comparison of the two representations in terms of the predictive performance ($RMSE_{test}$) on the training and the test data sets for the four oxidants. The corresponding $R^2$ values are listed in Table 2 and showed similar patterns. Comparison of the performance of the single ML models with previously published ones is shown in Text S7.

performance for the MD representation is always better than that for the MF-based models. However, that is not always the case regarding the generalization ability on the test data set. For the data sets of $SO_4^{\bullet-}$ and HClO, better predictive performance was achieved on both the training and test data sets for the MD-based models. Hence, the MD-based models were selected as the QSAR models for $SO_4^{\bullet-}$ and HClO. For the data sets of $O_3$ and $ClO_2$, the MD-based models showed better predictive performance on the training data sets but worse predictive performance on the test data sets 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 data set-dependent. One possible reason is that the types of MDs calculated by the PaDEL program are fixed for all contaminants. These MD-represented properties might correlate more with the contaminant reactivity toward $SO_4^{\bullet-}$ and HClO than with those toward $O_3$ and $ClO_2$. Therefore, it is recommended to screen the optimum chemical representation in future modeling rather than arbitrarily selecting one.

**3.2. Interpretation of the Single QSAR Models.** We interpreted all the single QSAR models (Table 2) to verify (1) if they made predictions based on the correct science and (2) if there were common features extracted among these models. The latter information may be useful to validate the KT strategy. Figure 3 shows the SHAP interpretation of the MF-based single QSAR models (Table 2) with the top 10 features shown (nine atom groups + pH). The interpretations of the pH effects and the pattern distribution are illustrated in Text S8. The results suggest that all the pH effects were correctly learned, and that different pattern distributions resulted from the different ML algorithms employed. Figure 4 shows the effect of the top nine atom groups identified in Figure 3 on the log $k$. As shown, the four models share several common atom groups. For example, the fourth atom group (aromatic carbon) in the $SO_4^{\bullet-}$ model is the same as the eighth atom group in the HClO model. The number of shared atom groups among these four oxidants is summarized in Table S3. Surprisingly, the learned contributions of some of these atom groups toward log $k$ differ significantly among the four data sets. For example, aromatic carbons in the $SO_4^{\bullet-}$ model (fourth) contributed positively to the log $k$ while those in the HClO (eighth), $O_3$ (third), or $ClO_2$ (fifth) model contributed negatively. The $-NH_2$ group increased the log $k$ in the $O_3$ model (seventh) but decreased the log $k$ in the $ClO_2$ model (third). 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).

For the $-NH_2$ group in the $ClO_2$ data set, its negative effect on the 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 $SO_4^{\bullet-}$ data set is greater than that for the compounds not containing aromatic carbons in the same data set, whereas this trend is reversed in the data sets 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 data set-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 data set, 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 data set, although $-NH_2$ is a well-known electron-donating group. In other words, the types of chemicals involved in a data set


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

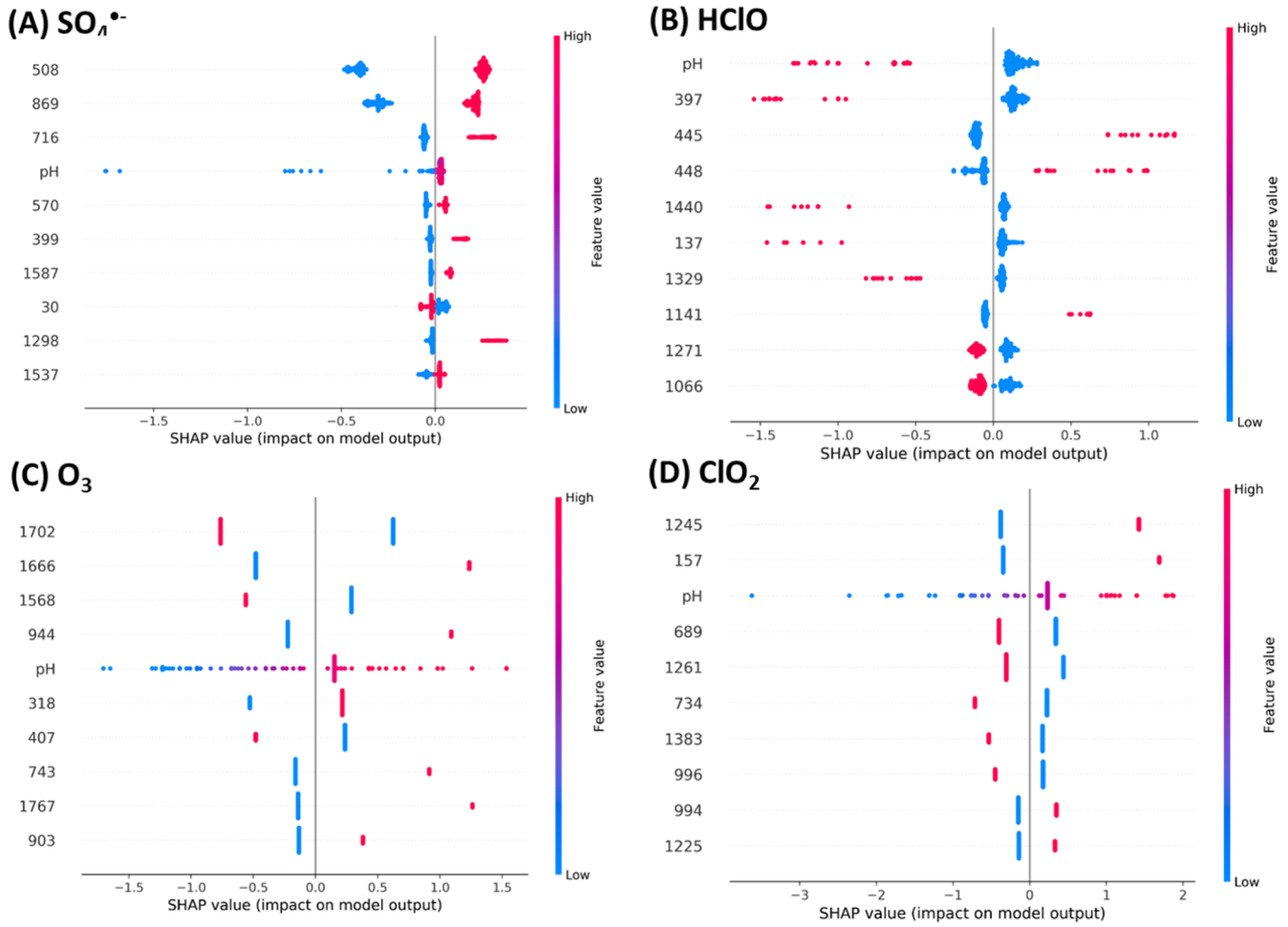



**Figure 3.** SHAP interpretation of the MF-based single QSAR models for the four oxidants. The *x*-axes are the SHAP values and the *y*-axes are the identified top 10 most influential features. The numbered features, such as 508, 1702, and 1666, represent the feature positions in the MFs, with each 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 data sets, 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 data set that contain that feature. All other SHAP plots in this work follow the same interpretation.

affects the model-derived positive or negative contribution of an atom group to log *k*. To illustrate the above mentioned idea for our data sets, we took the $ClO_2$ data set as an example, which has 36 chemicals containing aromatic carbons (fifth 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 S4), that are stronger in their electron-donating effects than aromatic carbons. As a result, aromatic carbons in the $ClO_2$ data set were "learned" to have negative effects on log *k*. The same explanation can be applied to the HClO and $O_3$ data sets. We believe that if a data set 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 data set 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 (Table 2). Detailed explanation for them is 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 data sets to form a large unified data set 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 data set ($R_{test}^2 = 0.76$) than all the single models (Table 2) (the RMSE values depended 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 data sets, as shown in Figure 5B. Except for the $SO_4^{\bullet-}$ data set, the performance of MF-UN-1 is better than that of the respective single models for the other three data sets. Figure 5C plots the distribution of the log *k* values in the four data sets, demonstrating the range of log *k* values for the $SO_4^{\bullet-}$ data set deviating substantially from that for the other three data sets. This may be the reason that the performance of MF-UN-1 on the $SO_4^{\bullet-}$ data set was worse.

Figure 5D shows the SHAP interpretation of this unified model and the identified top six atom groups (among the top


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

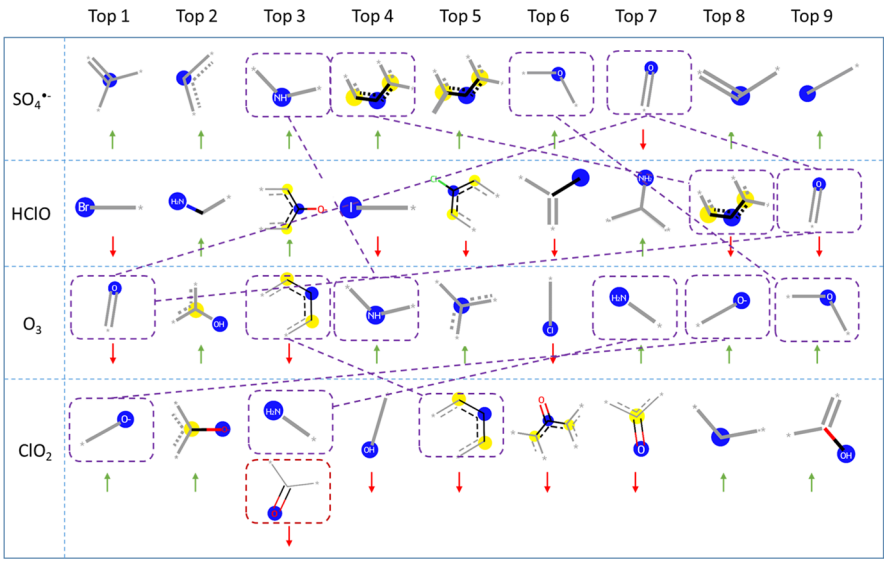



**Figure 4.** Effect of the top nine atom groups shown in Figure 3 on the log $k$ values, in which the up and down arrows mean increasing and decreasing the log $k$ values, respectively. The same atom groups in different data sets are marked by squares and connected by dotted lines. The $-NH_2$ and carbonyl groups are overlapped at the third position for the $ClO_2$ data set. Note that the length of the MFs has been optimized using 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 gray lines represent the neighboring bonds not in the feature; the dotted lines represent conjugated structures, for example, 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.

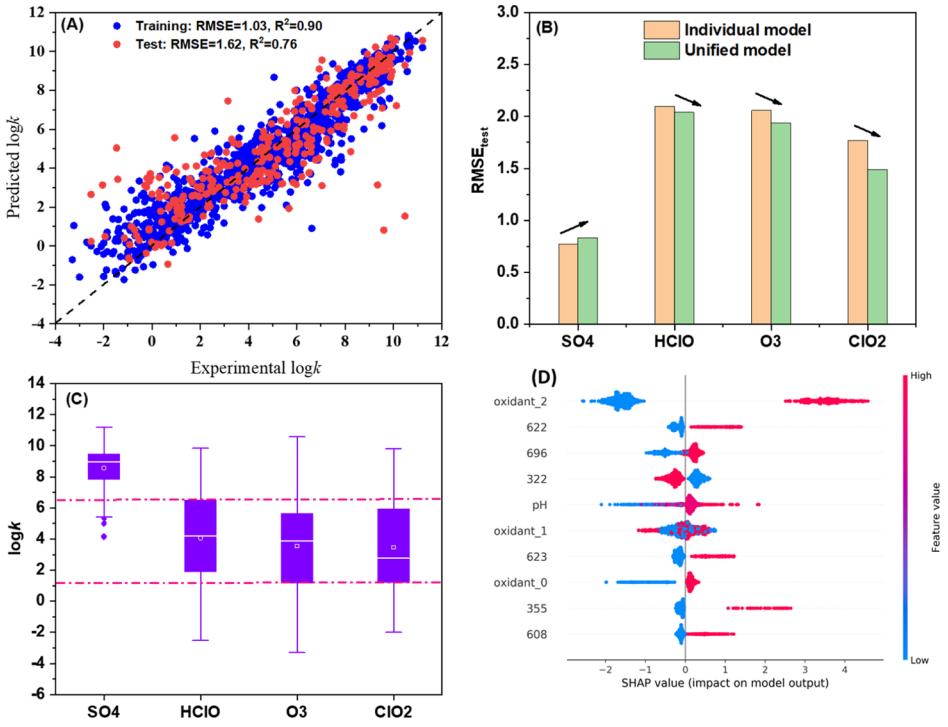



**Figure 5.** Predictive performance of the unified model on the training and test data sets for the unified data set (A) and the single data sets (B); (C) ranges of log $k$ values for the single data sets; and (D) SHAP interpretation of the unified model, in which the $x$-axis is the SHAP value and 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 only take values of 0 or 1. Their different combinations [i.e., ("Oxidant_1", "Oxidant_2", and "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 S5.

10 features in Figure 5D, only 6 of them are atom groups). Table S5 shows these atom groups and their effects on the log $k$, in which all of these effects were correctly learned. Although aromatic carbons were not among the top six atom groups, we still examined them here because their effects in the HClO, $ClO_2$, and $O_3$ data sets, and the effect of the $-NH_2$ group in the $ClO_2$ data set, 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 data set. For the


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

$SO_4^{\bullet-}$ data set, 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 $SO_4^{\bullet-}$ data set. In contrast, the effect of the $-NH_2$ group in the $ClO_2$ data set 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 $O_3$ data sets before and after combining the data sets, 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 the top nine.

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 accuray of 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 data sets (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 outperform the MF representation when developing the unified model, so we only focus on the MF representation below.

As mentioned above, the range of the log $k$ values for $SO_4^{\bullet-}$ is quite different from those of the other three data sets (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 $SO_4^{\bullet-}$ data set with a reported $OH^{\bullet}$ data set to form a large data set because their log $k$ values fall in the same range (Figure S4C). The $OH^{\bullet}$ data set 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 data set and Figure S4A shows the $R_{test}^2$ = 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 eight atom groups on the log $k$ were all correctly learned (only 8 of the top 10 features are atom groups) (Table S6). This worse performance for $^{\bullet}OH$ was probably because the additional fixed $T$ (25 °C) and pH (7) conditions were added into the $^{\bullet}OH$ data set to combine with the $SO_4^{\bullet-}$ data set, 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 $^{\bullet}OH$.

Finally, we combined all of these five data sets to form the largest data set to develop another MF-based unified model (refers to as "MF-UN-3"). Figure S5A shows that the $R_{test}^2$ reached 0.82. Although the predictions for $SO_4^{\bullet-}$, HClO, and $ClO_2$ became better, those for $^{\bullet}OH$ and $O_3$ became slightly worse (Figure S5B). Table S7 shows the effects of the identified top eight 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 $^{\bullet}OH$ data set is explained above, but the marginally worse predictive performance than MF-UN-1 for the $O_3$ data set is unexpected. We do not have a good explanation for this yet. These results suggested that it is not always better to combine data sets to achieve better predictive performance.

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

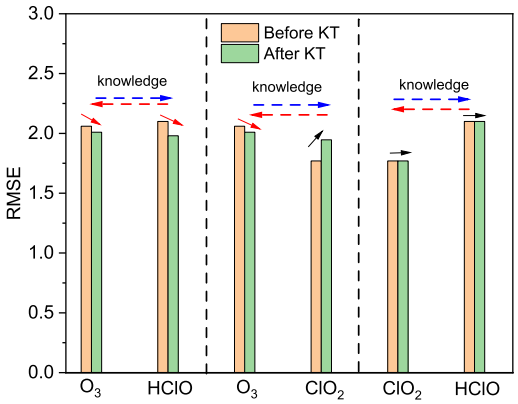



**Figure 6.** Predictive performance of different ML models before and after KT.

There are three distinct scenarios for these KT models. For $O_3$/HClO, both of the KT models show better predictive performance than the original models before the transfer. There is one shared atom group (carbonyl) among the top nine atom groups between $O_3$ and HClO (Figure 4) and the effect of this atom group was consistent (i.e., decreasing the log $k$) between the two data sets. 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 the KT. For $O_3$/$ClO_2$, the KT model for $O_3$ became better after receiving knowledge from the $ClO_2$ model, while the KT model for $ClO_2$ became worse. There are three atom groups shared between $O_3$ and $ClO_2$, but the effects of $-NH_2$ in these two data sets are opposite (Figure 4). Moreover, the predictive performance of the single model for $ClO_2$ ($RMSE_{test}$ 1.77) is better than that for $O_3$ ($RMSE_{test}$ 2.06), so the information transferred from $O_3$ to $ClO_2$ has more uncertainties, which should have led to the worse performance. For $ClO_2$/HClO, no change in the predictive performance was observed for both oxidants. This is expected because there are no shared atom groups between these two data sets (Figure 4). These results indicated that the effectiveness of our KT approach depends on if there is consistent knowledge shared between the single models and on their respective predictive performance. Although the KT models only showed some improvements in the predictive performance, the abovementioned information about when the KT approach will work is important for extending this approach to other data sets.

**3.5. Comparison of the Two Proposed Approaches with Multitask Learning and Image-Based Transfer Learning.** Because the model performance was not drastically


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

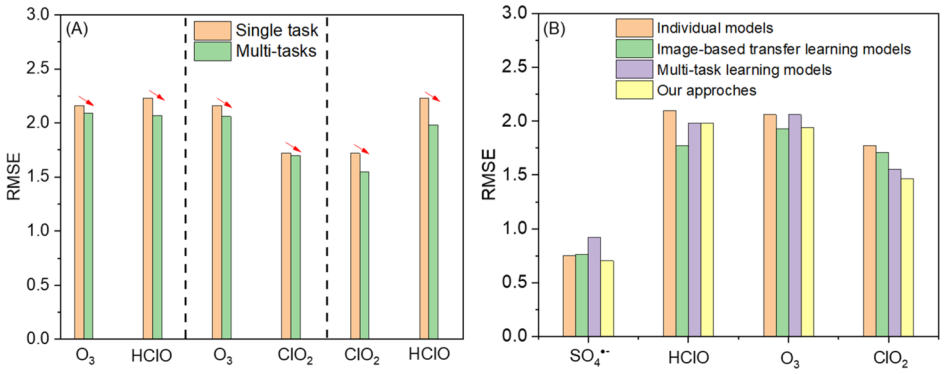



**Figure 7.** (A) Comparison of the predictive performance between the single-task and multitask (two-task) learning models; (B) comparison of the predictive performance among our approaches (i.e., combined-data set and KT approaches), multitask learning and image-based transfer learning. Only the model that achieved better performance (between combined-data set and KT approaches) for that data set was shown in (B) and referred to as "Our approaches".

**Table 3. Final Selected MF-Based Models for Each Data Set and Their AD Determination**[a]

| data set | 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 | KT 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 |

[a]The ADs of our models were not compared with previous published ones because previous studies used MDs to represent contaminants while we used MFs.

improved by the two proposed approaches, we further examined whether there were other approaches that could improve the model performance for these data sets. To this end, we applied single-task and multitask learning on the HClO, $O_3$, and $ClO_2$ data sets. An introduction to multitask learning and how we trained single-task and multitask learning models are in Text S9 in the Supporting Information. Figure 7A shows the comparison of the predictive performance of the single-task and multitask DNN models. First, for the single-task DNN models, most of their predictive performance was worse than the individual non-DNN based models (Table S8). For example, the $RMSE_{test}$ values for the HClO, $O_3$, and $ClO_2$ data sets were 2.23, 2.16, and 1.72 for the single-task DNN models, while those for the respective individual non-DNN based models were 1.74, 2.06, and 1.77. This may be because the sample sizes are too small for DNN (DNN is a data-demanding algorithm). After applying multitask learning, all of the multitask DNN models showed better predictive performance (i.e., generalization ability) on the test sets, indicating the effectiveness of multitask learning. However, these multitask DNN models still underperformed the combined data set-based model, although they outperformed some of the KT-based models (Table 2). We also conducted three-task or four-task learning but did not observe any improvements in the predictive performance (Table S8).

We next applied the image-based transfer learning approach to these four data sets. We previously used this approach on a $^{\bullet}OH$ data set and observed good predictive performance.[14] The details of how we applied this approach have been shown in our previous study.[14] Figure 7B showed that for the $SO_4^{\bullet-}$ data set, only the combined-data set approach was effective at decreasing the $RMSE_{test}$ while the image-based transfer learning and multitask learning increased it. For the $ClO_2$ data set, the combined-data set approach showed the best predictive performance. However, for the HClO data set, the image-based transfer learning outperformed both the KT model and the multitask learning model in terms of $RMSE_{test}$. Hence, the effectiveness of these approaches is also data set-dependent. For the $O_3$ data set, the performance of the combined-data set approach was similar to that of transfer learning and better than that of multitask learning. Overall, our approaches showed consistently good predictive performance for all these four data sets, while the image-based transfer learning and multitask learning approaches failed in some cases, such as for the $SO_4^{\bullet-}$ and $O_3$ data sets.

### 3.6. Final QSAR Models and Their AD Determination.

For the four oxidants, we ranked all the developed models in terms of the predictive performance and finally obtained the optimal QSAR models, as shown in Table 3. Both the unified models and transfer learning models outperformed all the


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

individual models and were selected as the final models, validating the effectiveness of our proposed two approaches. We next determined their ADs, as shown in Table 3. For each model, with an increasing threshold value, more contaminants were identified as outside the AD, and the recalculated $RMSE_{test}$ first decreased and then increased. The optimal threshold values for these four data sets are bolded in Table 3. For a query compound, if its similarity to the contaminants in the training data set is above the threshold value, the models can provide a reliable prediction for its reactivity toward one of these four oxidants.

## 4. ENVIRONMENTAL SIGNIFICANCE

In this study, we investigated ML-assisted QSAR models for data sets that are different but share similarities (i.e., oxidation reactions). Previous studies viewed these data sets independently, whereas we tried to utilize the shared knowledge among them to enhance the predictive performance of the models by four approaches—combining individual data sets to form a large, unified data set; transferring knowledge between individual data sets; applying multitask learning; and employing image-based transfer learning. When developing single ML models using these single data sets, we found that (1) the optimal ML algorithm is data set-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 data set-dependent because some representations may not capture the key features of the data set.

Combining similar data sets to form a large data set and developing a unified model can generally improve the predictive performance on the individual data sets because some "wrongly" learned knowledge based on a smaller data set (e.g., bias of the data set) may be corrected this way. In other words, data bias can be mitigated by increasing the sample size. KT is effective when there is consistent knowledge shared between the two data sets and when the single models themselves have good predictive performance. These two approaches can also help us understand the involved reaction mechanisms. For example, if the KT models have better predictive performance than the respective single models, it is likely that the two data sets share common reaction mechanisms, and if we already understand one of them, the other one can be more easily understood. We can also test if two data sets share similar reaction mechanisms by checking if the KT approach works. Furthermore, if some knowledge is not correctly learned by a model, there may exist data bias in the data set—for example, having too many highly reactive chemicals. This indicates that we need more data. Combining similar data sets is an effective approach to addressing this issue. In addition, multitask learning deserves some attention in the case of DNN algorithms and when the data sets are related. Better generalization ability was observed for multitask models when compared to the single-task models. Image-based transfer learning may also improve the model performance, as demonstrated here and in our recent work.[14]

Here, we introduced a new perspective on handling different small data sets, that is, some data sets may be interconnected—sharing information—even if they are all small. The traditional perspective of each data set as a separate and individual data set cannot capture this interaction information. This new perspective on data sets can be applied to many other scientific problems, particularly regarding small data sets sharing common knowledge. In fact, because it is much easier to find small data sets with similar ground truth than it is to find a large, similar data set for the transfer learning approach to be used, both approaches can help address the common data scarcity issue. In addition, many researchers tend to rely on more advanced ML algorithms to develop better models, and there indeed are many better ML algorithms developed every year. Instead, our work may inspire researchers to also focus on the data sets themselves because transferring the shared knowledge from one data set to another may yield better models than simply using advanced ML algorithms.

Overall, this study provided new insights into developing ML-based QSAR models for small data sets. The synergistic effects among similar data sets can be leveraged to boost the predictive performance of QSAR models. These findings can also be extended to other fields when small data sets are involved.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c04883.

Collected data (XLSX)

Details for differences between the well-known image-based transfer learning and our proposed KT approach; modeling the effect of pH on the log $k$; data preprocessing, 5 candidate scaler methods, 8 encoding methods, 7 outlier detection methods, and 16 ML algorithms; working mechanism of recursive feature selection; screening results for the scaler, encoding, and ML algorithms; overfitting control; specific number of MDs used and their names; SHAP interpretation of the MD-based single ML models; extracting atom groups from molecular fingerprints; comparison of the single ML models with previously published ones; interpretations of the pH effects and the pattern distributions; single-task and multitask learning model development; correlation plots of single ML models; and other related tables and figures for the model interpretation (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

**Yanping Zhang** − *School of Civil Engineering and Transportation, Hebei University of Technology, Tianjin 300401, China*; Email: zyphit@hebut.edu.cn

**Huichun Zhang** − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106-7201, United States*; orcid.org/0000-0002-5683-5117; Email: hjz13@case.edu

### Author

**Shifa Zhong** − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106-7201, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c04883


### Notes

The authors declare no competing financial interest.



https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

The instruction for using the models can be found on the GitHub: https://github.com/nogoodnameye/SO4-HClO-O3-ClO2.

## ■ ACKNOWLEDGMENTS

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

## ■ REFERENCES

(1) von Gunten, U. Oxidation Processes in Water Treatment: Are We on Track? *Environ. Sci. Technol.* **2018**, *52*, 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 Res.* **2011**, *45*, 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. *Environ. Sci. Technol.* **2000**, *34*, 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. *Environ. Sci. Technol.* **2019**, *53*, 3177−3186.
(5) Chin, A.; Bérubé, P. R. Removal of disinfection by-product precursors with ozone-UV advanced oxidation process. *Water Res.* **2005**, *39*, 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. *Environ. Sci.: Water Res. Technol.* **2020**, *6*, 2287−2312.
(7) Lee, Y.; von Gunten, U. Quantitative structure-activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. *Water Res.* **2012**, *46*, 6177−6195.
(8) Su, H.; Yu, C.; Zhou, Y.; Gong, L.; Li, Q.; Alvarez, P. J. J.; Long, M. Quantitative structure-activity relationship for the oxidation of aromatic organic contaminants in water by TAML/H2O2. *Water Res.* **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. *Chem. Eng. J.* **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. *J. Hazard. Mater.* **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. *Environ. Sci. Technol.* **2015**, *49*, 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. *Environ. Sci. Technol.* **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. *J. Hazard. Mater.* **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. *Chem. Eng. J.* **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. *Chem. Eng. J.* **2020**, *405*, 126627.
(16) Borhani, T. N. G.; Saniedanesh, M.; Bagheri, M.; Lim, J. S. QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Res.* **2016**, *98*, 344−353.
(17) Zhiwen, C.; 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. *Chem. Eng. J.* **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 Res.* **2013**, *47*, 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 Res.* **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. *Sci. Total Environ.* **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 Adv.* **2016**, *6*, 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 Res.* **2012**, *46*, 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 Res.* **2014**, *64*, 134−148.
(24) Najafabadi, M. M.; Villanustre, F.; Khoshgoftaar, T. M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1−21.
(25) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. Using rule-based labels for weak supervised learning: a ChemNet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018; pp 302−310.
(26) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741−12754.
(27) 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*, 754−761.
(28) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1044.
(29) Buxton, G. V.; Greenstock, C. L.; Helman, W. P.; Ross, A. B. 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* **1988**, *17*, 513−886.
(30) Xia, X.; Zhu, F.; Li, J.; Yang, H.; Wei, L.; Li, Q.; Jiang, J.; Zhang, G.; Zhao, Q. A Review Study on Sulfate-Radical-Based Advanced Oxidation Processes for Domestic/Industrial Wastewater Treatment: Degradation, Efficiency, and Mechanism. *Front. Chem.* **2020**, *8*, 592056.
(31) Von Gunten, U. Ozonation of drinking water: Part I. Oxidation kinetics and product formation. *Water Res.* **2003**, *37*, 1443−1467.
(32) Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming Data Scarcity with Transfer Learning. Submission date: Nov 7, 2017. arXiv preprint arXiv:1711.05099 2017. https://arxiv.org/abs/1711.05099 (accessed date: Aug 2, 2020)
(33) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012; pp 1097−1105.


https://doi.org/10.1021/acs.est.1c04883
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(34) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(36) 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*, 89−96.
(37) Dewancker, I.; McCourt, M.; Clark, S. *Bayesian Optimization for Machine Learning: A Practical Guidebook*. https://arxiv.org/abs/1612.04858. (Submission date: Dec 14, 2016, arXiv preprint arXiv:1612.04858 2016, accessed date: Apr 12, 2020)
(38) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 1−13.


https://doi.org/10.1021/acs.est.1c04883
*Environ. Sci. Technol.* XXXX, XXX, XXX−XXX