Environmental Science & Technology

pubs.acs.org/est

Article

Combining Group Contribution Method and Semisupervised Learning to Build Machine Learning Models for Predicting Hydroxyl Radical Rate Constants of Water Contaminants

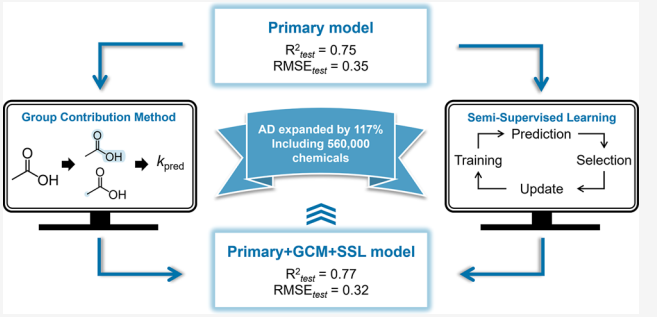
Zhao Liu, Lanyu Shang, Kuan Huang, Zhenrui Yue, Alan Y. Han, Dong Wang, and Huichun Zhang*

Cite This: https://doi.org/10.1021/acs.est.4c11950

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning is an effective tool for predicting reaction rate constants for many organic compounds with the hydroxyl radical (HO•). Previously reported models have achieved relatively good performance, but due to scarce data (<1400 records), the applicability domain (AD) has been significantly limited. To address this limitation, we curated a much larger experimental data set (Primary data set), which contains 2358 kinetic records. We then employed both the group contribution method (GCM) and a semisupervised learning (SSL) strategy to add new data points, aiming to effectively expand the model's AD while improving model performance. The results indicated that GCM improved the model's performance for chemicals outside the AD, while SSL expanded the model's AD. The final model, after incorporating 147,168 new data points, achieved an R^2 = 0.77, root-mean-square-error = 0.32, and mean-absolute-error = 0.24 on the test set. Importantly, the AD was expanded by 117% compared to the model developed solely based on the Primary data set, and the final model can be reliably applied to more than 560,000 chemicals from the DSSTox database. Further model interpretation results indicated that the model made predictions based on a correct "understanding" of the impact of key substituents and reactive sites toward HO•. This research provides an effective method for augmenting data sets, which is important in improving ML model performance and expanding AD. The final model has been made widely accessible through a free online predictor.



KEYWORDS: *applicability domain, group contribution method, hydroxyl radical, machine learning, reaction rate constant, semisupervised learning*

INTRODUCTION

Machine learning (ML) is a powerful tool for addressing many real-world problems. However, its application in various fields, including the environmental field, is often constrained by the lack of sufficient data.[1−3] Specifically, the sample size for modeling environmental chemical reactivity usually ranges from dozens to thousands. For example, an ML model designed to predict the bioaccumulation of organic pollutants from soil to plant roots relied on 341 data points for 72 chemicals.[4] Recent ML models for predicting chemical reactivity across different processes included just 117 organic and 10 inorganic species for abiotic reduction by Fe(II)-associated reductants,[5] 1978 chemicals for adsorption,[6] 978 records for 206 chemicals for anaerobic biodegradation,[7] 12,750 records spanning 6032 chemicals in aerobic biodegradation,[8] and 195, 191, 759, and 557 records for oxidation by HClO, chlorine dioxide, ozone, and sulfate radicals, respectively.[9] Even in the extensively studied field of organic chemical degradation by hydroxyl radicals (HO•), the latest model for predicting reaction rate constants (k) used only 1374 data points.[10] This data scarcity inherently limits the diversity of chemical structures in the data set, so the corresponding ML models face at least three main limitations. First, the model applicability domain (AD) is limited by the chemical structures present in the training data sets. Therefore, it is challenging to extend these models to many environmentally relevant compounds, such as those within the Distributed Structure-Searchable Toxicity (DSSTox) database with over 840,000 chemicals.[11] In fact, only 38% of the chemicals in the DSSTox database fall within the model AD for the Fe(II)-associated reductants.[5] Second, with smaller data sets, the risk of overfitting increases, where models may capture noise rather than underlying principles. This limits the models' ability to generalize to new compounds, reducing their

Received: November 2, 2024
Revised: December 9, 2024
Accepted: December 19, 2024

predictive accuracy and robustness. Lastly, small sample sizes restrict the depth of mechanistic insights and hinder the discovery of new knowledge from the models. Therefore, the obtained ML models capture only partial trends within the data sets, and these trends may even sometimes be biased.[9]

Common methods for obtaining environmental data include literature reviews, database extractions, experiments, and computational techniques.[5,12,13] While literature and databases are commonly used data sources for ML models, existing models have often already summarized the available data. Therefore, expanding the data sets typically requires additional experiments, which can be time-consuming. For example, the OECD 301 tests take 28 days to assess chemical biodegradability.[14] Moreover, the absence of commercial standards for many chemicals makes experimental evaluation challenging. As a result, it is necessary to develop an efficient strategy to significantly expand the data beyond the aforementioned methods. For image data, common augmentation methods include geometric transformations, color space transformations, noise injection, and affine transformations.[15,16] For tabular data, augmentation methods involve synthetic data generation, noise injection into numerical features, and feature transformations.[17] However, these approaches have significant limitations because they do not introduce new, meaningful data points or chemicals. For example, phenol remains phenol regardless of how its images are flipped or rotated.[18] Synthetic data generation methods, such as the synthetic minority oversampling technique (SMOTE) and generative adversarial networks (GANs), only create new data points that resemble the original data to address data imbalance issues,[19] but these synthetic data may lack practical chemical meaning. Furthermore, image data augmentation methods are not applicable to molecular descriptors or molecular fingerprints, which are crucial for describing chemical information. Consequently, the corresponding model ADs remain limited to the data obtained by traditional methods. Therefore, developing alternative data augmentation methods that are suitable for different types of ML models and capable of generating a significant amount of new data is critically needed.

To this end, there are two potentially promising strategies: the group contribution method (GCM) and semisupervised learning (SSL). The GCM refers to a set of techniques used primarily in the field of chemistry and chemical engineering to estimate the properties of molecules and mixtures.[20,21] In terms of chemical reactivity, GCM decomposes a chemical structure into several groups, each of which contributes to a portion of the overall reactivity.[22,23] It performs well when applied to chemical substances with specific structures. For example, Minakata et al. 2009 developed a GCM model based on 310 chemicals to predict the log k between organic compounds and $HO^{\bullet}$, which performed well for chemicals that typically undergo hydrogen-atom abstraction and aromatic compound addition.[23] Therefore, as long as appropriate chemicals are selected for prediction using GCM, the prediction results can serve as new data points. Besides, unlike traditional quantitative structure−activity relationship (QSAR) models that require extensive features such as molecular and quantum chemical descriptors to accurately represent chemical reactions,[9,24−26] GCM models do not require such complex computational data, which can often be challenging to obtain.[27,28] For SSL, it is useful when there are some labeled data—data points with output values—but many more unlabeled data—data points without output values,[29] a situation that closely mirrors our current predicament. Self-training is a specific SSL technique where a model uses its own predictions (pseudolabels) to teach itself. It trains iteratively on a labeled data set, predicts labels for unlabeled data, and then retrains on the combined data set.[30,31] Therefore, SSL is an effective method for addressing data scarcity and the high costs of labeling. Common SSL applications include text classification and image recognition. For example, Meng et al. 2018 trained a text classification model using a pseudodocument generator and self-training module with only 20 labeled documents per class and 500−1000 pseudo documents of each class, resulting in an increase in the micro-F1 score from 0.668 to 0.823.[32] Lee et al. 2013 used SSL to train an image classification model with 600 labeled data points and 70,000 unlabeled data points, improving the classification error from 8.57% to 5.03%.[33] Although there are no relevant environmental applications of SSL so far, we could employ SSL to obtain new data based on the existing unlabeled database. More information on GCM and SSL can be found in Texts S1−S2.

In this work, we coupled GCM with SSL for the first time to significantly expand the sample size, and examined the accuracy and generalization ability of the corresponding ML model. Our study focused on the oxidation of organic chemicals by the $HO^{\bullet}$, a process with extensive experimental data and previously reported GCM models.[34−36] While many studies have developed machine learning models to predict the log k between $HO^{\bullet}$ and chemicals, these models are limited by the small size and diversity of the data sets.[10,18,36,37] For example, Zhong et al. 2021 reported a model combining deep neural networks and extreme gradient boosting (XGBoost) that achieved R^2 = 0.60−0.71 with a data set comprising 1089 points.[37] Similarly, Sanches-Neto et al. 2021 developed an XGBoost model to attain an R^2 = 0.82, based on a data set containing 1374 data points.[10] In our approach, as illustrated in Figure 1, we first significantly expanded the Primary data set to include 2358 data points and developed the ML-based Primary model. Then, we identified chemicals suitable for prediction using the GCM model and integrated the prediction results into the training set to develop an improved ML model—the Primary + GCM model. Next, we predicted the log k for all chemicals in the DSSTox data set using the Primary model, selected predictions that met a specific confidence threshold as pseudolabels, and added these predictions and the corresponding chemicals into the training set to update the Primary model. We iteratively refined the model as described above, resulting in the development of the Primary + SSL model. Finally, SSL was applied to the Primary + GCM model to obtain the final Primary + GCM + SSL model. To evaluate the models' applicability, we defined the AD by applying these models to the DSSTox database. For model validation, we utilized the SHAPley Additive exPlanations (SHAP) method, alongside comparisons with known mechanisms, to interpret the resulting models. Finally, the final model has been made accessible through a free, user-friendly online predictor, available at https://envmodel-cwru.streamlit.app/.

METHOD

Kinetic Data Set. In this research, all kinetic data records were collected from published journal articles and the National Institute of Standards and Technology (NIST). A data set, named "Primary dataset", which comprises 2358 data points, is summarized in the Excel file provided in the Supporting

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

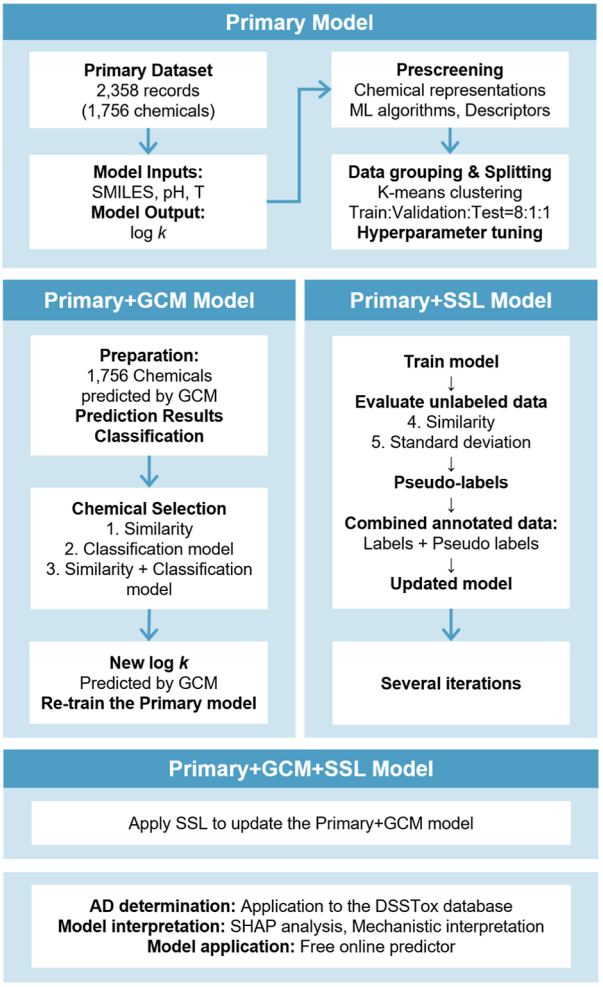


Figure 1. Flowchart of this research.

Information. The data collection process followed four steps: (1) gathering data records from both published papers and NIST, including chemical information, pH, temperature (T), and log k values; (2) acquiring the SMILES notation for each chemical; (3) for data points with the same SMILES, pH, and T, removing outliers based on the log k values and averaging the remaining values to obtain the final log k value; and (4) preprocessing the data, including calculating species acid/base fractions (α notation) and standardizing data format. For more information on the data collection and processing methods, please refer to Texts S3−S4.

Model Development. *Primary Model.* The development of the Primary model for predicting log k included three key stages: (1) Initial screening of 13 ML algorithms, 5 chemical representations, and 5 descriptors (molecular fingerprint, pH, T, pKa, and α notation). This stage was essential for identifying the most effective combination for our predictive models; (2) data grouping and splitting: to ensure that each type of data records was evenly distributed across training, validation, and test sets, chemicals were meticulously grouped by three distinct strategies: K-means clustering, agglomerative hierarchical clustering, and spectral clustering. Then, the data points of each group were divided into training, validation and test sets in a ratio of 8:1:1; (3) hyperparameter tuning using Bayesian optimization, following the selection of the optimal ML algorithm and chemical representation. This step was critical to enhancing the model's performance. More details of Primary model development can be found in Texts S5−S7.

Primary + GCM Model. The Primary + GCM model was built based on the Primary model after selecting suitable chemicals from the DSSTox database (details below). GCM was then used to predict the log k of these chemicals (details in Text S8). To accurately assess the impact of incorporating GCM predictions on model performance, we strictly used experimental data for the validation and test sets. This approach ensures that these sets accurately represent real-world experimental data and unseen scenarios, thereby maintaining the integrity of our model evaluation. After adding the prediction results to the training set, we retrained the Primary model to obtain the Primary + GCM model. One critical aspect of using GCM to acquire new data points was discerning which chemicals were amenable to GCM prediction to ensure the prediction quality. In this research, three methods were deployed for chemical selection: similarity analysis; classification model; a combination of similarity analysis and classification model. Prior to employing these methods, preliminary steps were undertaken, as detailed below and in Figure S2.

Preparation. This step aimed to establish the foundation for the three methods. The Primary data set contained 2358 experimental data points, encompassing 1756 distinct chemicals. GCM was utilized to predict log k for all 1756 chemicals. Based on the prediction results, the chemicals were categorized into three types: type I, where the prediction fell within 0.5−2 times the experimental value; type II, where the prediction fell outside the 0.5−2 times range; and type III, where GCM was incapable of predicting the log k values of the chemicals. The number of chemicals in each category was 476, 417, and 863 in type I, II, and III, respectively.

Similarity Analysis (Method 1). We conducted similarity calculations between all chemicals in DSSTox and the 476 type I chemicals. Details of the similarity calculation method can be found in Text S9. Chemicals from DSSTox with the highest similarity exceeding 0.7 were selected for GCM prediction, resulting in a total of 116,233 chemicals. We chose a threshold of 0.7 because previous research suggested that a high similarity between the query data and the training data increased the reliability of predictions.[8] A lower threshold might result in inaccurate predictions from GCM; whereas a higher threshold could make the added chemicals too similar to the original ones, limiting the model's AD expansion.

Classification Model (Method 2). A classification model was developed based on the 1756 chemicals and their corresponding types. The input was the SMILES of the chemicals, and the output was its type (I, II or III). This classification model was then applied to the whole DSSTox database to identify chemicals that belonged to type I, resulting in a total of 183,759 chemicals. More details of the classification model can be found in Text S10.

Combination of Similarity Analysis and Classification Model (Method 3). The third method combined the first two selection methods. Specifically, the classification model was applied to the preselected 116,233 chemicals, resulting in a total of 67,181 chemicals that not only exhibited a similarity over 0.7 to type I chemicals from the Primary data set but also were classified as type I by the classification model.

We also investigated how the selection methods, and the similarity and quantity of chemicals added impacted model performance and AD. As shown in Figure S3, 116,233,

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

183,759, and 67,181 chemicals were selected by Methods 1, 2, and 3, respectively. The similarity of these three groups of chemicals to the chemicals in the Primary data set was calculated, and each group was classified into different similarity ranges according to the maximum similarity value. Using Method 1 as an example, since the lowest similarity of the chemicals selected must be higher than 0.7, the 116,233 chemicals were divided into three similarity ranges: 0.7−0.8, 0.8−0.9, and 0.9−1.0. The number of chemicals in each range was 14,274, 36,422, and 65,537, respectively. A portion of the chemicals in each range was then randomly selected for prediction using GCM. Methods 1 and 2 investigated the impact of adding high and low similarity chemicals on model performance, respectively; while Method 3 focused on the comprehensive impact of both methods. Specifically, Methods 1 and 3 were designed to select chemicals within similarity ranges of 0.7−0.8, 0.8−0.9, and 0.9−1.0, while Method 2 targeted chemicals with lower similarities, within the ranges of 0.4−0.5, 0.5−0.6, and 0.6−0.7. For each similarity range, we randomly selected 100, 200, 300, or 400 chemicals and used GCM to predict their log k values, which were then added to the training set to retrain the Primary model. It should be noted that the GCM predictions were assumed to represent log k under standard conditions (pH = 7.0, T = 25 °C). Additionally, chemicals with a similarity equal to 1 were eliminated.

Primary + SSL Model. The Primary + SSL model was also built based on the Primary model. Specifically, we used the Primary model to predict log k's (referred to as pseudolabels) for the unlabeled chemicals in the DSSTox database and selected the most confident predictions (details below). Then, we added the selected chemicals with their pseudolabels to the training set and retrained the model. The prediction-augmentation-retrain process was repeated until a stopping criterion was met to obtain the Primary + SSL model. In SSL for regression tasks, confidence filtering was used to assess the level of trust we could place in the model's predictions for unlabeled data. As shown in Figure S2, two strategies were used to determine the confidence of the predictions: similarity analysis and standard deviation.

Similarity Analysis (Method 4). First, we calculated the similarity between each chemical in DSSTox and all chemicals in the Primary data set. When the similarity was higher than a certain threshold, the quarry chemical and its prediction results were added to the training set to retrain the model. In the next iterations, the similarity between each chemical in the DSSTox database and the chemicals in the updated training set was recalculated, and then new chemicals were selected again based on the same threshold.

Standard Deviation (Method 5). By generating multiple models with the same training set and comparing their predictions, we could evaluate the variability of these predictions, which reflected uncertainty. Specifically, we created five models using cross-validation in each iteration. These models concurrently predicted the log k for chemicals in DSSTox, allowing us to compute the predictions' standard deviation for each chemical. A standard deviation below a specified threshold signified a high confidence level in the chemical's prediction. Chemicals with high confidence predictions were then added to the training set for model retraining.

In the process of establishing the Primary + SSL model, we studied the impact of threshold values, the number of added chemicals, and the number of iterations on model performance. For Method 4, we examined the similarity thresholds of 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95. For Method 5, we used standard deviations of 0.0075, 0.01, 0.02, 0.03, or 0.04 as the threshold. Since using the entire DSSTox chemicals was too time-consuming, we only used 10% when testing different thresholds. To analyze the impact of the number of added chemicals, we selected 10% or 100% of the DSSTox chemicals. We conducted a sensitivity analysis to assess the impact on model performance in two ways: (1) by varying pH and temperature to understand how changes in these input assumptions affect predictions, and (2) by introducing different levels of noise into the pseudolabels to evaluate the impact of errors in these labels (details in Text S11). In all cases, the above process was repeated until no new chemicals were selected or a maximum of 9 iterations was reached. Note that if a new iteration selected the same chemical as before, the updated prediction result was added to the training set to replace the previous pseudo label. More details about hyperparameter tuning during SSL can be found in Text S12.

Primary + GCM + SSL Model. For the Primary + GCM model, we identified the most effective method for selecting chemicals and determined the optimal number to add to the training set to enhance model performance. For the Primary + SSL model, we established the most suitable threshold, number of iterations, and the appropriate volume of data from the DSSTox database to use as our selection pool. After determining these parameters, we applied the SSL approach to the Primary + GCM framework, resulting in the final Primary + GCM + SSL model. To address the concern about error propagation, we employed two methods to explore how prediction inaccuracies of GCM could be propagated by the SSL process (details in Text S13).

Model Evaluation and AD. To comprehensively evaluate the practical application potential of the models and the effectiveness of the GCM and SSL strategies, we employed two types of test sets: a similar test set and a dissimilar test set. The similar test set was obtained by grouping all chemicals in the Primary data set, then randomly splitting off 10% from each group and combining them. For the dissimilar test set, the Primary data set was divided into a dissimilar test set (10%) and the remaining set (90%). The similarity between each chemical in the dissimilar test set and all chemicals in the remaining set was below 0.70. After that, the remaining set was grouped and then divided into the training and validation sets. The similar and dissimilar test sets were able to comprehensively reflect the performance of the models against within-AD (similar) and out-of-AD (dissimilar) chemicals, respectively (more below). The test sets were fixed during the model training process to ensure consistent comparison among different models, and the robustness of the models was assessed by different splitting of training and validation sets. It should be noted that since it took a long time for SSL to train a series of iterations, there was only one test set to evaluate model performance during the training process, and the robustness of the model was ensured by dividing the remaining set in five random states. More details of the test sets can be found in Text S14. R^2 (R-squared), RMSE (root-mean-square error) and MAE (mean absolute error) were used to evaluate model performance. More details of the parameters and model evaluation can be found in Text S15.

The AD of a model assessed the model's applicability to a given chemical.[38] This was achieved by calculating the

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

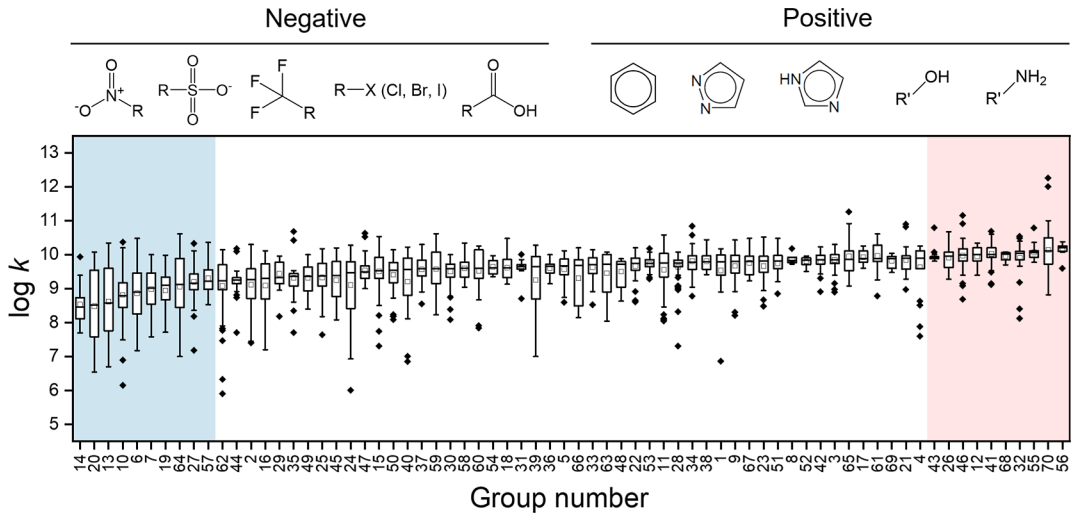


Figure 2. Box plot of log k values for 70 groups using K-means clustering. The left blue area represents the ten groups with the smallest log k. The right red area represents the ten groups with the largest log k. "Negative" and "Positive" mean the representative substructures belonging to the left and right ten groups, respectively. (R indicates the ring structure, while R' indicates the aromatic ring structure).

similarity between the query chemical and each chemical in the training data set. The highest similarity value indicated the similarity level between the target chemical and the training set. If this value surpassed 0.7, the chemical was within the model's AD, indicating a reliable prediction, and the opposite if it did not. After defining the AD for each model, we calculated its coverage on the DSSTox database. The performance of the final model against DSSTox chemicals was determined by testing chemicals across four different ranges of similarity levels (i.e., 0.9−1.0, 0.8−0.9, and <0.7−0.8) under five random states. Their corresponding RMSE and R^2 values were then determined. More details of the AD can be found in Text S16.

Model Interpretation. To validate the ML models and evaluate the importance and contribution of different features to log k, we employed two methods: (1) the SHAP (SHapley Additive exPlanations) analysis and (2) mechanistic interpretation. SHAP analysis is a method for interpreting complex ML models by attributing the contribution of each feature to the final prediction.[37] To explain the reaction mechanisms learned by the model from the training data, we first correlated the SHAP values of the MACCS fingerprint bits for monosubstituents of aromatic compounds with the reported Hammett constants.[39] As the model development process involved a few times of random splitting of the data, each splitting might result in slightly different SHAP values and their rankings. Therefore, we calculated the mean absolute SHAP values (feature importance scores) as the final SHAP values for each feature. More details of the MACCS fingerprint can be found in Text S17. We predicted log k_{pred} values for selected compounds under experimental conditions and then built linear correlations between the log k_{pred} values and the highest occupied molecular orbital (E_{HOMO}) under selected conditions. A good correlation would indicate that our model was based on a correct understanding of the mechanism.[5,37] We also compared the model's prediction results for aromatic and nonaromatic chemicals in the DSSTox database. More details of SHAP analysis and mechanistic interpretation can be found in Texts S18−S19.

RESULTS AND DISCUSSION

Meta-Analysis of the Primary Data Set. The compiled Primary data set contained a total of 2358 data records for 1756 chemicals. Compared with recent studies on HO• (data points < 1400),[10,18,36,37] the amount of data had been greatly improved. The log k of various chemicals toward HO• were mainly between 10^9 and 10^{10} (Figure S4). When the Primary data set was compared with the DSSTox database, 1126 chemicals were common between the Primary data set and the DSSTox database, representing 0.15% of the 844,412 DSSTox chemicals. Additionally, 710 chemicals were common between the Primary data set and the ANST/POL list (chemicals detected or identified in environmental media,[40] Text S19) (Tables S1−S2), accounting for 3.60% of the total 19,776 chemicals in the list.[40] Among the top 500 chemicals in the ANST/POL list, only 107 were in the Primary data set. As the distribution of the 710 chemicals shown in Figure S5, the top seven categories studied for reaction kinetics with HO•, representing 81.83% of the 710 common chemicals, include pharmaceuticals, pesticides in use, chemicals of emerging concern, plastics additives, gases/VOCs/CFCs/HFCs, disinfection byproducts, and fragrances.

Model Development and Evaluation. *ML Algorithms, Chemical Representation, and Descriptor Prescreening.* The performance of different ML algorithms is shown in Figure S6. The results indicated that Random Forest (RF) and XGBoost (XGB) had better performance than the other ML algorithms. To further compare the performance of RF and XGB, the two models were tuned using Bayesian optimization. After optimization, XGB had better performance and less overfitting than RF (Figure S7).[41] Therefore, XGB was selected as our default algorithm for the regression models. When we compared different chemical representations (Figures S8−S10), the MACCS fingerprint had the best model performance so it was used as the chemical representation below. More discussion about chemical representation can be found in Text S21.

In this study, there are two types of descriptors: compound descriptors, including molecular fingerprint, pKa and α notation; and reaction descriptors, including pH and T. The best model performance was obtained using pH, T and

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

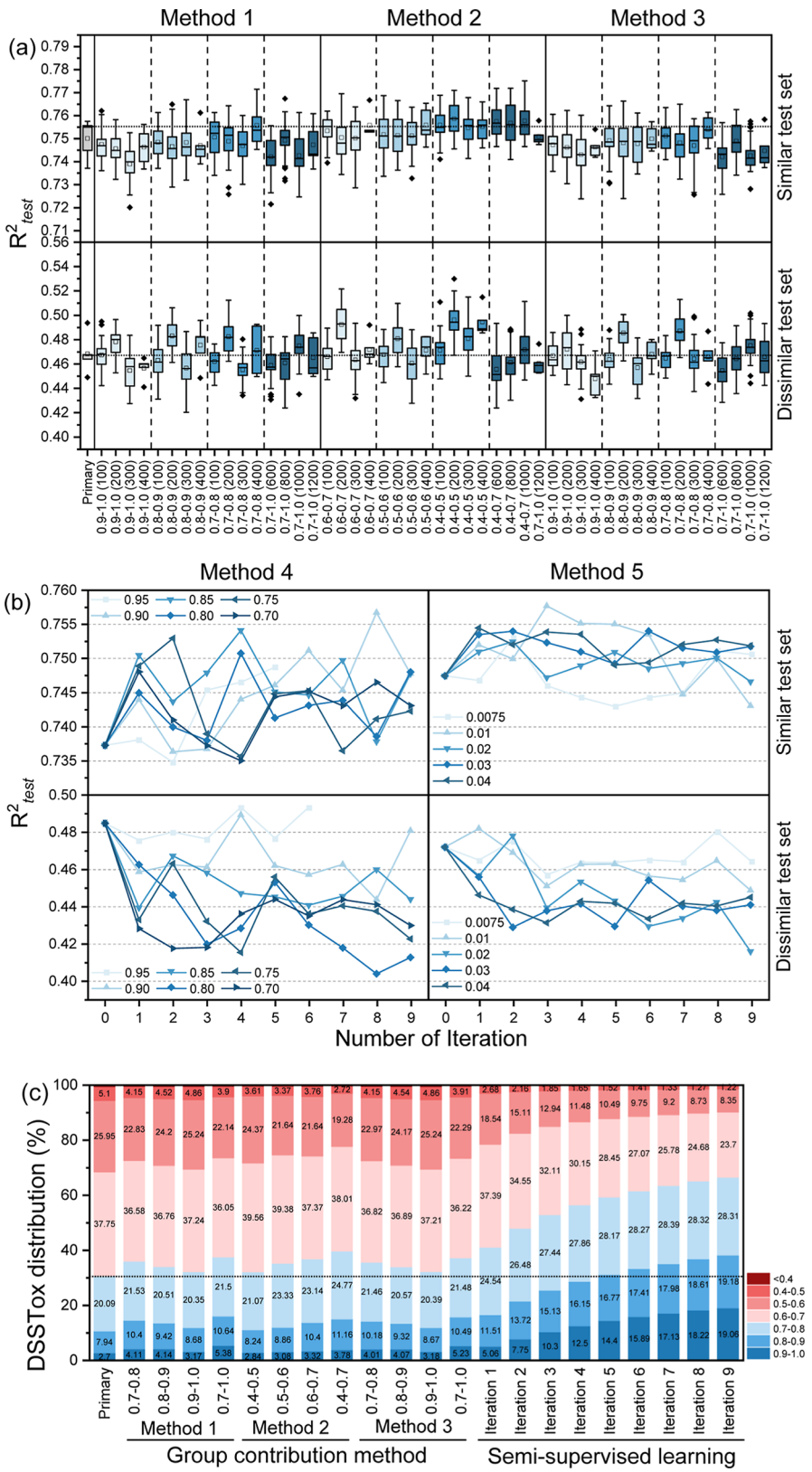


Figure 3. (a) Model performance (R^2) for the similar test set and dissimilar test set before and after the addition of GCM predictions based on Methods 1, 2, and 3. The numerical range represents the similarity of the newly added chemicals to the chemicals in the Primary data set. The number in parentheses indicates the number of newly added chemicals. (Method 1: similarity analysis, Method 2: classification model, Method 3: combination of similarity analysis and classification model). (b) Model performance (R^2) for the similar test set and dissimilar test set after different iterations of self-learning based on Method 4 (similarity analysis) and Method 5 (standard deviation). (c) Comparison of the percentages of DSSTox chemicals in different similarity ranges among different models. The black dotted line, based on the Primary model, divides DSSTox chemicals with a similarity value ≥0.7 from those <0.7. More information can be found in Text S16.

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

MACCS fingerprint as inputs (Figure S11). Adding p*K*a and α notation did not improve the performance and might introduce unwanted noise.

Chemical Grouping. Model performance for the similar test set, using various grouping strategies, is depicted in Figure S12, with *K*-means of 70 clusters showing optimal results, and chemical distribution across groups is detailed in Table S5. For the dissimilar test, the best performance was achieved with 85 clusters (Figure S13), guiding further optimizations for this test set.

The 70 groups were arranged in order based on their median log *k* values, from smallest to largest. Representative functional groups were identified from the 10 groups with the highest and lowest rate constants. The results in Figure 2 indicated that aromatic compounds generally exhibited higher log *k* than aliphatic compounds, consistent with previous results.[42,43] The conjugated π electron systems of aromatic compounds not only increased electron density but also boosted the stability of the molecules.[44,45] This, in turn, made reactions with $HO^{\bullet}$ more feasible. Functional groups altered a molecule's electron density by donating or withdrawing electrons, affecting its reactivity with $HO^{\bullet}$.[46] For example, electron-withdrawing groups ($-NO_2$, $-SO_3^-$, halogens, $-COOH$) reduced electron density around carbon atoms, making them less reactive toward $HO^{\bullet}$; while the opposite was true for electron-donating groups (e.g., $-OH$, $-NH_2$).[47,48]

Hyperparameter Tuning. As shown in Figure S14a−c, after Bayesian optimization, model performance for the similar test set achieved R_{test}^2 = 0.75, $RMSE_{test}$ = 0.35 and MAE_{test} = 0.24, improved by 13.64%, 17.50% and 7.69%, respectively, after optimization. For the dissimilar test set (Figure S14d−f), the R_{test}^2, $RMSE_{test}$ and MAE_{test} of the model before and after optimization are 0.43, 0.56, 0.42 and 0.47, 0.53, 0.39, respectively, improved by 9.30%, 5.36% and 7.14% respectively. The optimized hyperparameters are shown in Tables S6−S7.

Chemical Selection Methods. *Chemical Selection Methods for GCM.* As shown in Figures 3a and S17, model performance against the similar test set became worse after adding the GCM predicted log *k* by Methods 1 and 3, but improved marginally by Method 2. The chemicals selected by Methods 1 and 3 are highly similar to the chemicals in the Primary data set, which leads to many more chemicals in one category than in others (sample imbalance), causing the model to favor frequently occurring categories, affecting the overall performance. In contrast, the chemicals selected by Method 2 have low similarity to the Primary data set, which augments categories with insufficient data types, thereby improving model performance. For the dissimilar test set, the chemicals added through the three methods all improved model performance to a certain extent. Specifically, 200 chemicals added through Method 2, with a similarity range of 0.4−0.5, most effectively improved the model's performance by 6.00%, 2.79%, and 3.27% in R_{test}^2, $RMSE_{test}$, and MAE_{test}, respectively. These results indicated that adding chemicals with low similarity could enhance the model's performance for out-of-AD chemicals. In other words, for out-of-AD chemicals, it is crucial to select additional (different) data points that can enable the model to learn new information.

The chemical percentages in DSSTox before and after data augmentation by GCM are shown in Figure 3c. Based on the Primary data set, chemicals with a similarity greater than 0.7 comprised 30.73% of the entire DSSTox database. After adding 400 chemicals with a similarity between 0.6 and 0.7 through Method 2, this proportion increased most effectively to 36.86%. Although adding chemicals with similarity between 0.4 and 0.5 through Method 2 improved model performance, it only increased this proportion to 32.15%. The likely reason is that 37.75% of the DSSTox chemicals were in the similarity range of 0.6 to 0.7, while only 5.10% were in the range of 0.4−0.5. Therefore, adding chemicals within the 0.6 to 0.7 similarity range more effectively encompassed a larger number of chemicals to be within the AD. Generally speaking, adding data points from the range with higher density should be more efficient in expanding the AD.

Chemical Selection Methods for SSL. According to Figures 3b and S18, for the similar test set, Method 4 with a 0.90 threshold effectively improved performance, whereas Method 5 across various thresholds did not significantly impact model performance. The likely explanation is that the model's initial performance was sufficiently high (R_{test}^2 = 0.737). Consequently, the selected chemicals and pseudolabels represented "knowledge" the model had learned. Retraining the model using these pseudolabels helped the model to refine its knowledge and improve its performance on data that was similar to the training set. However, using different percentages of the DSSTox database as the selection pool did not significantly affect model performance on the similar test set (Figure S19). For the dissimilar test set, both Methods 4 and 5, regardless of the threshold, generally negatively affected model performance. Given the model's initial poor performance against the dissimilar test set (R_{test}^2 = 0.485), it is evident that the model struggled with these chemicals. Adding pseudolabels further biased the model toward chemicals similar to those in the training set, exacerbating its inability to handle dissimilar data.

When Method 4 was applied to 100% of DSSTox (Figure 3c), the percentage of chemicals added to the training set increased with an increasing number of iterations. Considering both model performance and AD, the model achieved its best performance at iteration 8, where 135,929 chemicals had been added, increasing the proportion of within-AD chemicals (similarity $\geq$0.7) from the initial 30.73% to 65.15%. This increase signified that an additional 34.87% of the chemicals (over 294,000 chemicals) from DSSTox were incorporated into the AD. Therefore, selecting an appropriate threshold and iteration of SSL effectively enhanced the model robustness and expanded the AD. However, SSL did not improve the generalization ability of the model, that is, the predictive ability, for chemicals outside the AD. Therefore, a combined GCM and SSL strategy might both improve model performance and expand AD, as shown below. The sensitivity analysis showed that varying pH (0−13) and temperature (3−85 °C) had minimal impact on the Primary + SSL model. While random GCM prediction errors did not degrade performance, larger systematic errors reduced accuracy, without worsening across iterations. Besides, introducing noise into pseudolabels lowered performance, but the model stabilized with more iterations, highlighting the importance of high-quality inputs. Details on the above sensitivity analysis and error propagation results are provided in Text S22.

Final Model Performance and AD. In the final Primary + GCM + SSL model, the Primary + GCM model used the 200 chemicals in the 0.4−0.5 similarity interval selected by Method 2; the SSL utilized Method 4, employing a threshold of 0.9 and a DSSTox percentage of 100%, to train the final model based

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

on the Primary + GCM model. According to Figure S25a,b and Table S9, the model's performance on the similar test set improved marginally, with the R^2 increasing from 0.75 in the Primary model to 0.77 in the Primary + GCM + SSL model at iteration = 9. The performance of the Primary + GCM + SSL model for chemicals in different similarity intervals is shown in Table 1. For chemicals with a similarity value greater than 0.7, the model achieved an R^2 of 0.69−0.83 and a RMSE of 0.27−0.32. In contrast, chemicals with similarity below 0.7 are considered outside the AD, resulting in an R^2 = 0.51 and RMSE = 0.45. The analysis of the MACCS fingerprint bit distribution for the Primary data set (Figure S26a) revealed that nearly all bits were represented, indicating that the data set effectively captured the essential structural features necessary for predicting log k. The similar distribution of log k values in the Primary data set versus the added chemicals (Figure S25c) suggested that the log k values obtained for the added chemicals would enhance the data density at the peak of the Primary data set. Furthermore, the distribution of chemicals in the Primary data set was relatively uniform (Figure S26b), with the added chemicals increasing density in several regions. However, compared with the DSSTox database, some sparse areas remained (Figure S26c,d), likely due to the structural similarity-based chemical selection methods. As shown in Figure S25d, the model's AD significantly expanded from 30.73% to 66.90% after integrating GCM and SSL, such that over 560,000 chemicals from DSSTox are now included in the

Table 1. AD of the Primary + GCM + SSL Model and the Model Applicability toward the DSSTox Database

	expected prediction		percentages of DSSTox chemicals	
similarity	R^2	RMSE	each level (%)	cumulative (%)
0.9−1.0	0.83 ± 0.02	0.27 ± 0.02	19.24	19.24
0.8−0.9	0.76 ± 0.01	0.34 ± 0.01	19.25	38.49
0.7−0.8	0.69 ± 0.04	0.41 ± 0.03	28.41	66.90
<0.7	0.51 ± 0.02	0.45 ± 0.02	33.10	100

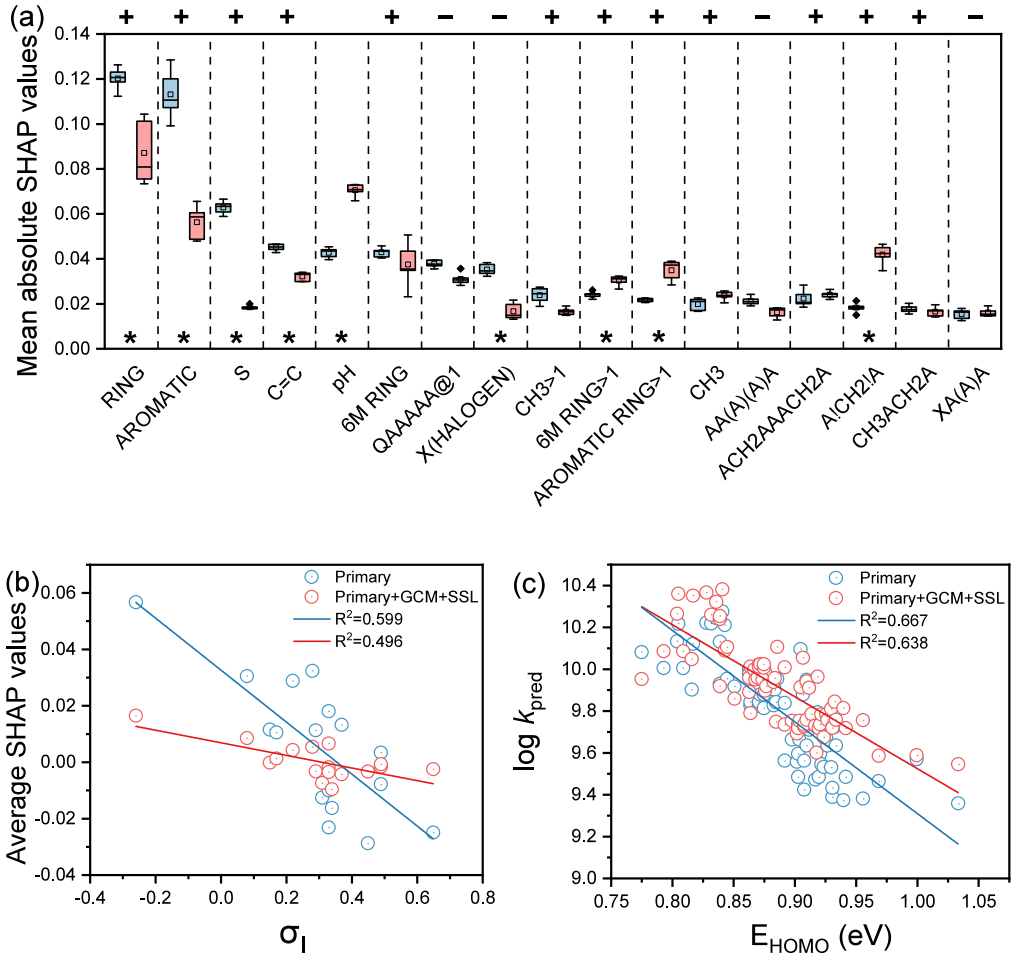


Figure 4. (a) Common features of the top 20 most important features based on their mean absolute SHAP values for the Primary model (blue) and Primary + GCM + SSL model (red). The box plots show their importance scores; while the "+" and "−" symbols on top indicate positive and negative contributions to the log k, respectively. The "*" symbol on the bottom indicates there is statistical difference (p-values <0.05) between the SHAP values. A: any valid periodic table element symbol; Q: hetro atoms, any non-C or non-H atom; X: halogens; =: double bond; $: ring bond; !: chain or nonring bond. More information about MACCS fingerprints can be found in Text S17 and Table S3. Note: there is no symbol above pH because the effect of pH on log k is 2-fold. (b) Correlations between average SHAP values of electron donating/withdrawing groups and Hammett constant of inductive effect. The specific functional groups, MACCS fingerprint bits, and Hammett constant values (σ_I, σ_R, and σ_P) are shown in Table S11. (c) Correlations of log k_{pred} with E_{HOMO}.

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

AD, allowing our model to provide reliable predictions for them.

Model Interpretation. To validate the Primary + GCM + SSL model, we first identified the most influential factors on log k by calculating the SHAP values for all 168 features—166 MACCS fingerprint bits, pH, and T. According to Figures 4a and S27 and S28, the top 20 important features of the Primary and Primary + GCM + SSL models largely overlapped, indicating that the integration of GCM and SSL did not significantly alter the prediction mechanism of the models. Specifically, for both models, aromatic functional groups and ring structures had a strong positive impact on log k, consistent with the meta-analysis results (Figure 2). S and C=C double bonds also had positive effects on log k, likely because they are electron-donating.[49−51] In fact, the H−S group typically reacts first, compared to H−C and H−N groups.[34] Features that had a negative impact on log k mainly included heterocyclic ring, chain structures, and halogens functional groups (i.e., QAAAAA@1, X(HALOGEN), AA(A)(A)A, XA(A)A). This is also consistent with the meta-analysis results. pH significantly influenced log k, exhibiting both positive and negative effects, although the positive effects were more predominant due to the pH-dependent reduction potential of $HO^{\bullet}$ and deprotonation of functional groups.[52,53] Meanwhile, T was not highlighted due to limited data, revealing a limitation of SHAP analysis.[8] Further discussion of pH and T is in Text S23. Note that for the Primary + GCM + SSL model, the order of the top 20 features slightly differed from that in the Primary model, and their importance values were also slightly lower. This variation may be due to the expanded training set, allowing the model to learn more about chemical structures and better assess the impact of structural features.

Figure S30a shows the distribution of the final model predicted log k values for the DSSTox database. While the values for both aromatic and nonaromatic chemicals largely fall between 9 and 10, the nonaromatic chemicals show a broader distribution in the 7−9 range. A statistically significant difference (p-value <0.05) between the two groups confirms that the model has correctly learned the effect of aromatic rings on log k values. For aromatic compounds, the effects of substituents are evident in the significant correlation between log k and the Hammett constant.[54] As shown in Figures 4b and S30b,c, important electron-donating and -withdrawing functional groups (based on the SHAP values) had a good linear relationship with their corresponding Hammett σ values for the Primary and Primary + GCM + SSL models. However, the fitting performance of the Primary + GCM + SSL model was slightly inferior to that of the Primary model. A possible reason is that the training set of the Primary + GCM + SSL model has been significantly expanded, allowing for a comprehensive consideration of multiple features and feature interactions during predictions.[37] As a result, the fitting performance for individual features and the Hammett constant has decreased.

To further validate and interpret the models, we analyzed log k_{pred} values for selected compounds under experimental conditions similar to those in the literature. While E_{HOMO} has been proven by many studies to have a strong linear correlation with log k,[55−57] it was not used as input because not many chemicals had this value available. The results in Figure 4c showed that there was a good linear relationship between log k_{pred} and E_{HOMO} for the Primary model (R^2 = 0.667) and the Primary + GCM + SSL model (R^2 = 0.638), demonstrating that the models predicted a chemical's log k based on correct knowledge.

ENVIRONMENTAL SIGNIFICANCE

This study developed enhanced ML models to predict the reaction rate constants of $HO^{\bullet}$ with organic compounds in water, as a case study to explore the impact of two data augmentation approaches—GCM and SSL. A Primary data set comprising 2358 experimental data points was collected, which is significantly larger than those previously reported (typically less than 1500 data points). However, the model's AD based solely on this data set covered only 30.73% of the DSSTox database. To significantly enhance the model's applicability, we integrated GCM with SSL for the first time to effectively expand the data set to include more than 140,000 chemicals. The AD of the final Primary + GCM + SSL model covered 66.90% of the DSSTox database, encompassing approximately 560,100 chemicals. Additionally, the model exhibited good predictive performance for chemicals within the AD (R^2 = 0.69−0.83 and RMSE = 0.27−0.32). To enhance the accessibility of the final model developed in this study, we created a free online predictor (https://envmodel-cwru.streamlit.app/) (a user guide in Text S24). This tool simplifies the model's usage, even for users with limited knowledge of ML. Compared to previous models, the AD of our model is significantly improved. Industries can leverage the model to predict the reactivity of chemical intermediates with $HO^{\bullet}$, optimizing synthetic pathways to reduce undesirable by-products and enhance yield and safety in manufacturing. The model's expanded AD ensures accurate assessment of a wide range of chemicals, enabling more efficient and cost-effective production. Additionally, understanding how contaminants react with $HO^{\bullet}$ allows water treatment facilities to improve purification systems, ensuring more effective removal of contaminants and advancing treatment strategies. Pollutants with low reaction rate constants tend to persist in natural and engineered systems. Our model can prioritize these persistent pollutants, help assess their environmental risks, and guide their classification to support regulatory and remediation efforts.

Our study goes beyond simply building a predictive model for the reaction rate constants of hydroxyl radicals. The primary goal of this research is to develop an effective method for expanding data sets in a meaningful way. This approach offers potential benefits not only for environmental research but also for the broader chemical sciences, where limited data sets often constrain research progress. This new approach is not only applicable to $HO^{\bullet}$ but also to other reactive species, such as ozone, sulfate radical, and chlorine, by leveraging existing QSAR models to generate new data points.[58−60] Furthermore, this study is the first to apply SSL in the environmental field to address data scarcity, and this methodology can be extended to other fields where large sets of unlabeled data are available. However, the effectiveness of SSL is dependent on the underlying model performance. Poor initial predictions could introduce noise in pseudolabels, limiting model reliability. Additionally, some of the findings from this study offer practical guidance for future experimental work. Notably, there are currently few experimental studies on chemicals with low similarity. Targeted experiments in this area would not only advance our understanding of these substances but also broaden the applicability of the proposed model. In this way, our work bridges the gap between data-

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

driven modeling and experimental research, laying a foundation for more comprehensive future studies.

■ ASSOCIATED CONTENT

Data Availability Statement

The code associated with this research available on GitHub: https://github.com/ZhaoLiu0919/hydroxyl-radical-reaction-rate-constant-prediction.git.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.4c11950.

More details about the mechanism of GCM for the hydroxyl radical, introduction of SSL, data collection, pK_a and α notation, ML algorithms, chemical representations and descriptors, data set grouping, Bayesian optimization, GCM prediction method, similarity calculation, classification model development and evaluation, sensitivity analysis, hyperparameter tuning and validation during SSL, error propagation, test sets, model evaluation, AD and the DSSTox database, MACCS fingerprint, SHAP analysis, mechanistic interpretation, ANST/POL list, Web site instructions, additional results and discussion, and supplementary figures and tables (PDF)

Data sets used in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

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

Authors

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

Lanyu Shang − *School of Information Sciences, University of Illinois Urbana−Champaign, Champaign, Illinois 61820, United States*

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

Zhenrui Yue − *School of Information Sciences, University of Illinois Urbana−Champaign, Champaign, Illinois 61820, United States*

Alan Y. Han − *Department of Computer Science, Cornell University, Ithaca, New York 14850, United States*

Dong Wang − *School of Information Sciences, University of Illinois Urbana−Champaign, Champaign, Illinois 61820, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c11950

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the US National Science Foundation Grant # CHE-2105005.

■ REFERENCES

(1) Liu, X.; Lu, D.; Zhang, A.; Liu, Q.; Jiang, G. Data-driven machine learning in environmental pollution: gains and problems. *Environ. Sci. Technol.* **2022**, *56* (4), 2124.

(2) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine learning in environmental research: common pitfalls and best practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671.

(3) Xia, D.; Chen, J.; Fu, Z.; Xu, T.; Wang, Z.; Liu, W.; Xie, H.-b.; Peijnenburg, W. J. Potential application of machine-learning-based quantum chemical methods in environmental chemistry. *Environ. Sci. Technol.* **2022**, *56* (4), 2115.

(4) Gao, F.; Shen, Y.; Sallach, J. B.; Li, H.; Liu, C.; Li, Y. Direct prediction of bioaccumulation of organic contaminants in plant roots from soils with machine learning models based on molecular structures. *Environ. Sci. Technol.* **2021**, *55* (24), 16358.

(5) Gao, Y.; Zhong, S.; Zhang, K.; Zhang, H. Abiotic reduction of organic and inorganic compounds by Fe (II)-associated reductants: comprehensive data sets and machine learning modeling. *Environ. Sci. Technol.* **2023**, *57* (46), 18026.

(6) Zhang, K.; Zhang, H. Predicting solute descriptors for organic chemicals by a deep neural network (DNN) using basic chemical structures and a surrogate metric. *Environ. Sci. Technol.* **2022**, *56* (3), 2054.

(7) Cheng, Y.; Zhang, K.; Huang, K.; Zhang, H. Meta-analysis and machine learning models for anaerobic biodegradation rates of organic contaminants in sediments and sludge. *Environ. Sci. Technol.* **2024**, *58* (29), 12976−12988.

(8) Huang, K.; Zhang, H. Classification and regression machine learning models for predicting aerobic ready and inherent biodegradation of organic chemicals in water. *Environ. Sci. Technol.* **2022**, *56* (17), 12755.

(9) Zhong, S.; Zhang, Y.; Zhang, H. Machine learning-assisted QSAR models on contaminant reactivity toward four oxidants: combining small data sets and knowledge transfer. *Environ. Sci. Technol.* **2022**, *56* (1), 681.

(10) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. py SiRC": machine learning combined with molecular fingerprints to predict the reaction rate constant of the radical-based oxidation processes of aqueous organic contaminants. *Environ. Sci. Technol.* **2021**, *55* (18), 12437.

(11) US EPA. Distributed Structure-Searchable Toxicity (DSSTox) Database. https://www.epa.gov/chemical-research/distributedstructure-searchable-toxicity-dsstox-database (accessed 06 10, 2022).

(12) Ai, H.; Zhang, K.; Sun, J.; Zhang, H. Short-term lake erie algal bloom prediction by classification and regression models. *Water Res.* **2023**, *232*, 119710.

(13) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J. Machine learning: new ideas and tools in environmental science and engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741.

(14) Reuschenbach, P.; Pagga, U.; Strotmann, U. A critical comparison of respirometric biodegradation tests based on OECD 301 and related test methods. *Water Res.* **2003**, *37* (7), 1571.

(15) Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Global Transit. Proc.* **2022**, *3* (1), 91.

(16) Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258.

(17) Machado, P.; Fernandes, B.; Novais, P. Benchmarking data augmentation techniques for tabular data. *International Conference on Intelligent Data Engineering and Automated Learning*, 2022, 104−112.

(18) 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.

(19) Tanaka, F. H. K. D. S.; Aranha, C. Data augmentation using GANs. *arXiv* **2019**, arXiv:1904.09135.

(20) Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE J.* **1994**, *40* (10), 1697.
(21) Kehiaian, H. Group contribution methods for liquid mixtures: a critical review. *Fluid Phase Equilib.* **1983**, *13*, 243.
(22) Jankowski, M. D.; Henry, C. S.; Broadbelt, L. J.; Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **2008**, *95* (3), 1487.
(23) Minakata, D.; Li, K.; Westerhoff, P.; Crittenden, J. Development of a group contribution method to predict aqueous phase hydroxyl radical (HO•) reaction rate constants. *Environ. Sci. Technol.* **2009**, *43* (16), 6220.
(24) 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.
(25) 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.
(26) Zhang, P.; Sun, M.; Zhou, C.; He, C.-S.; Liu, Y.; Zhang, H.; Xiong, Z.; Liu, W.; Zhou, P.; Lai, B. Origins of selective oxidation in carbon-based nonradical oxidation processes toward organic pollutants: Quantitative structure-activity relationships (QSARs). *Environ. Sci. Technol.* **2024**, *58* (10), 4781.
(27) 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* (3), 1111.
(28) Yang, Z.; Luo, S.; Wei, Z.; Ye, T.; Spinney, R.; Chen, D.; Xiao, R. Rate constants of hydroxyl radical oxidation of polychlorinated biphenyls in the gas phase: A single- descriptor based QSAR and DFT study. *Environ. Pollut.* **2016**, *211*, 157.
(29) Yang, X.; Song, Z.; King, I.; Xu, Z. A survey on deep semi-supervised learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35* (9), 8934.
(30) Van Engelen, J. E.; Hoos, H. H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109* (2), 373.
(31) Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; Wang, X. A survey of deep active learning. *ACM Comput. Surv.* **2022**, *54* (9), 1.
(32) Meng, Y.; Shen, J.; Zhang, C.; Han, J. Weakly-supervised neural text classification. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*; CIKM, 2018; pp 983−992.
(33) Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning*; ICML, 2013; Vol. 3, p 896.
(34) 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* (2), 513.
(35) Cheng, M.; Zeng, G.; Huang, D.; Lai, C.; Xu, P.; Zhang, C.; Liu, Y. Hydroxyl radicals based advanced oxidation processes (AOPs) for remediation of soils contaminated with organic compounds: a review. *Chem. Eng. J.* **2016**, *284*, 582.
(36) 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.
(37) 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.* **2021**, *405*, 126627.
(38) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2* (10), 573.
(39) Hansch, C.; Leo, A.; Taft, R. A survey of Hammett substituent constants and resonance and field parameters. *Chem. Rev.* **1991**, *91* (2), 165.
(40) Muir, D. C.; Getzinger, G. J.; McBride, M.; Ferguson, P. L. How many chemicals in commerce have been analyzed in environmental media? A 50 year bibliometric analysis. *Environ. Sci. Technol.* **2023**, *57* (25), 9119−9129.
(41) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*; ACM Digital Library, 2016, pp 785−794.
(42) Wojnárovits, L.; Tóth, T.; Takács, E. Critical evaluation of rate coefficients for hydroxyl radical reactions with antibiotics: a review. *Crit. Rev. Environ. Sci. Technol.* **2018**, *48* (6), 575.
(43) Wojnárovits, L.; Takács, E. Rate coefficients of hydroxyl radical reactions with pesticide molecules and related compounds: A review. *Radiat. Phys. Chem.* **2014**, *96*, 120.
(44) Hou, M.; Li, F.; Liu, X.; Wang, X.; Wan, H. The effect of substituent groups on the reductive degradation of azo dyes by zerovalent iron. *J. Hazard. Mater.* **2007**, *145* (1−2), 305.
(45) Xie, Z.-H.; He, C.-S.; Zhou, H.-Y.; Li, L.-L.; Liu, Y.; Du, Y.; Liu, W.; Mu, Y.; Lai, B. Effects of molecular structure on organic contaminants' degradation efficiency and dominant ROS in the advanced oxidation process with multiple ROS. *Environ. Sci. Technol.* **2022**, *56* (12), 8784.
(46) Anbar, M.; Meyerstein, D.; Neta, P. The reactivity of aromatic compounds toward hydroxyl radicals. *J. Phys. Chem.* **1966**, *70* (8), 2660.
(47) Jeong, K. M.; Kaufman, F. Kinetics of the reaction of hydroxyl radical with methane and with nine chlorine-and fluorine-substituted methanes. 1. Experimental results, comparisons, and applications. *J. Phys. Chem.* **1982**, *86* (10), 1808.
(48) Jeong, K. M.; Kaufman, F. Kinetics of the reaction of hydroxyl radical with methane and with nine chlorine-and fluorine-substituted methanes. 2. Calculation of rate parameters as a test of transition-state theory. *J. Phys. Chem.* **1982**, *86* (10), 1816.
(49) 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.
(50) Albarran, G.; Bentley, J.; Schuler, R. H. Substituent effects in the reaction of OH radicals with aromatics: Toluene. *J. Phys. Chem. A* **2003**, *107* (39), 7770.
(51) Gligorovski, S.; Strekowski, R.; Barbati, S.; Vione, D. Environmental implications of hydroxyl radicals (• OH). *Chem. Rev.* **2015**, *115* (24), 13051.
(52) Wardman, P. Reduction potentials of one-electron couples involving free radicals in aqueous solution. *J. Phys. Chem. Ref. Data* **1989**, *18* (4), 1637.
(53) Zhang, P.; Sun, M.; Liang, J.; Xiong, Z.; Liu, Y.; Peng, J.; Yuan, Y.; Zhang, H.; Zhou, P.; Lai, B. pH-modulated oxidation of organic pollutants for water decontamination: A deep insight into reactivity and oxidation pathway. *J. Hazard. Mater.* **2024**, *471*, 134393.
(54) Peres, J. A.; Domínguez, J. R.; Beltran-Heredia, J. Reaction of phenolic acids with Fenton-generated hydroxyl radicals: Hammett correlation. *Desalination* **2010**, *252* (1−3), 167.
(55) Lee, Y.; Von Gunten, U. Quantitative structure-activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. *Water Res.* **2012**, *46* (19), 6177.
(56) Kušić, H.; Rasulev, B.; Leszczynska, D.; Leszczynski, J.; Koprivanac, N. Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: A QSAR study. *Chemosphere* **2009**, *75* (8), 1128.
(57) Yin, R.; Guo, W.; Ren, N.; Zeng, L.; Zhu, M. New insight into the substituents affecting the peroxydisulfate nonradical oxidation of sulfonamides in water. *Water Res.* **2020**, *171*, 115374.
(58) 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.

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

(59) Ding, H.; Hu, J. Prediction of Second-Order Rate Constants of sulfate radical with aromatic contaminants using quantitative structure-activity relationship model. *Water* **2022**, *14* (5), 766.
(60) Lei, Y.; Cheng, S.; Luo, N.; Yang, X.; An, T. Rate constants and mechanisms of the reactions of Cl• and Cl2•- with trace organic contaminants. *Environ. Sci. Technol.* **2019**, *53* (19), 11170.

https://doi.org/10.1021/acs.est.4c11950
Environ. Sci. Technol. XXXX, XXX, XXX−XXX