

1     **Prediction of acute toxicity of organic contaminants to fish: model development and a**  
2                                 **novel approach to identify reactive substructures**

3     Shangyu Li <sup>a</sup>, Mingming Zhang <sup>b\*</sup>, Peizhe Sun <sup>a\*</sup>

4

5     <sup>a</sup> School of Environmental Science and Engineering, Tianjin University, Tianjin 300072,  
6     China

7     <sup>b</sup> Heibei Key Laboratory of Metabolic Diseases, Heibei, China

8

9     \*Corresponding Authors

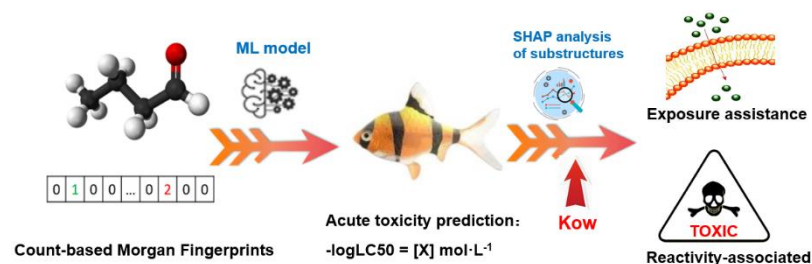
10    Mingming Zhang - Email: zhangmingming123@126.com

11    Peizhe Sun - Email: sunpeizhe@tju.edu.cn

## Abstract

In this study, count-based Morgan fingerprints (CMF) was used to represent the fundamental chemical structures of contaminants, and a neural network model ( $R^2=0.76$ ) was developed to predict acute fish toxicity (AFT) of organic compounds, which surpassed previous models. We found the limitations of in distinguishing homologous compounds may account for the suboptimal performance of binary fingerprints. The principles of generation and collision of CMF was explored and an improved method based on Tanimoto distance was introduced to calculated molecular similarity represented by CMFs as well. Toxic substructures identified by Shapley additive explanation (SHAP) method were substituted benzenes, long carbon chains, unsaturated carbons and halogen atoms. By incorporating  $K_{ow}$  and monitoring shifts in feature importance, the influence of substructures on AFT was further delineated, revealing their roles in facilitating exposure and reactive toxicity. On this basis, we compared the toxicity of similar substructures and the same substructure in different chemical environments. To overcome the limitation of SHAP analysis, this study proposed a new method, toxicity index (TI), to identify substructures that were present in small quantities but highly toxic. With TI, we identified several important substructures, such as parathion and polycyclic substituents. We found that the toxicity of large substructures may be misestimated in the previous studies.

## Graphical abstract



## Keywords:

LC50, QSAR, count-based Morgan fingerprint, machine learning, acute fish toxicity

## 1. Introduction

Chemicals, while contributing to regional economic growth and enhancing livelihoods, may inadvertently be released into the environment throughout their life cycles, becoming pollutants in the process (Wang, H. et al., 2022). It is imperative to conduct environmental risk assessments, as numerous chemicals have been identified as potentially toxic to humans and harmful to the environment. The results of the 96-hour acute fish toxicity (AFT) test, which provides the median lethal concentration (LC50) value, serve as a valuable benchmark for assessing environmental toxicity and ecological risk. However, with over 200 million registered chemicals in the CAS registry and an average daily increase exceeding 30,000, coupled with over 350,000 chemicals and mixtures thereof registered for production and use (Wang et al., 2020), experimental toxicity assessments are both time-consuming and labor-intensive. Furthermore, such laboratory tests are increasingly viewed as unethical and the desire to replace animal experimentation has garnered increased attention (Singh et al., 2013).

Quantitative structure–activity relationships (QSARs) represent promising tools in various research areas, including reaction rate prediction (Zhong et al., 2020; Sun et al., 2022), contaminant screening (Wang et al., 2022; McLachlan et al., 2014), drug discovery (Neves et al., 2018; Bosc et al., 2019) and more. Several QSAR models have been proposed to predict the toxicities of chemicals to fish, demonstrating considerable accuracy (Singh et al., 2013; Samanipour et al., 2023; Sheffield and Judson, 2019; Martin et al., 2001). However, traditional QSARs developed by molecular descriptors (MD) and the group contribution method (GCM) have their own limitations. The development of MD-based models (Singh et al., 2013; Samanipour et al., 2023; Sheffield and Judson, 2019) is highly dependent on a small subset of

59 pre-selected MDs, necessitating the careful selection of the most relevant descriptors by  
60 researchers (Zhong et al., 2020). Moreover, MDs are typically derived from the calculations of  
61 advanced software, and the physical meanings of many MDs are obscure, impeding the  
62 interpretability of the model and the identification of warning structures. GCM (Martin et al.,  
63 2001) is another approach to developing QSARs. It statistically evaluates specific substructures  
64 and assigns varying effect values to them. Consequently, the toxicity value of a molecule is  
65 computed as a linear sum of the contribution values associated with each of its constituent  
66 substructures. Although GCM offers strong interpretability, it often requires manual definition  
67 of important structures and counting, which can be labor-intensive. Therefore, there is an urgent  
68 need for a simple and efficient modeling method to develop QSARs.

69 The identification of toxic substructures is critical step, as it allows us to validate the  
70 reliability of the QSAR model by confirming that it has indeed internalized the relevant patterns.  
71 Furthermore, these toxic groups can serve as substructural alerts and aids in uncovering novel  
72 insights into the mechanisms of toxicity. The traditional GCM (Martin et al., 2001) entails  
73 assigning unique weights to various substructures, which are employed to gauge their  
74 respective toxicities. However, models constructed by GCM is inherently linear and does not  
75 account for the potential interactions between substructures. Consequently, when the activity  
76 or reaction modes of substructures varies due to different chemical environments, GCM may  
77 be unreliable and inaccurate. The Shapley additive explanation (SHAP) method is another  
78 extensively employed technique for identifying toxic substructures within chemical compounds  
79 (Cao et al., 2022; Bo et al., 2023), which can capture non-linear relationships between features  
80 by evaluating the effect of removing a specified molecular fragment on model performance.

However, the investigation into the mutual influence between substructures is relatively scarce using the SHAP method, and this approach may inadvertently overlook substructures that are present in minor concentrations within the chemical compounds being analyzed. Over the last years, more and more techniques have been used to find the relationship between substructure and activity, including but not limited to atom-level coloration (Wang et al., 2022), attention-vector based relevant latent features exploration (Xiong et al., 2020), and masking techniques applied to graph neural networks (Wu et al., 2023). However, these approaches tend to concentrate excessively on substructures in individual molecules, which complicates the identification of common patterns from databases.

The objective of this article is to utilize molecular fingerprints (MF), which contains the most fundamental structural information, to construct a high-precision model, thereby facilitating a more comprehensive analysis of substructural toxicity. An available dataset comprising 908 organic contaminants and their corresponding LC50 values was utilized to build the model (see Fig. 1 for the workflow). Based on this dataset, the effects of MF parameters, ML algorithms and data splitting methods on the prediction performance were investigated. Subsequently, SHAP method was used to interpret the models about what features were selected to make the predictions. Furthermore, we endeavored to distinguish between the lipid solubility and reactive toxicity of molecular structures, as well as to identify highly toxic substructures that were overlooked by SHAP methods. Finally, the model was applied to screen for prevalent toxic substructures and potential highly toxic substances in two datasets: one containing high production volume chemicals (HPV) and another with chemicals approved for production and use by the Food and Drug Administration (FDA). All the datasets are provided

in the [SI Table File](#).

## 2. Materials and methods

### 2.1. Database and Preprocessing

The 96-hour AFT test results for fathead minnows (*Pimephales promelas*) of 908 chemicals, sourced from OASIS, ECOTOX, and EAT5 were provided by Cassotti et al. (2015). The dataset comprises the concentrations ( $\text{mol}\cdot\text{L}^{-1}$ , measured in  $-\log$  units) that result in death in 50% of the test subjects over a 96-hour period. The data set was initially divided into a training set (80%) and a test set (20%) with a random pattern. The training set was utilized for model development, feature selection and validation, and the independent test set was used to externally evaluate the model performance.

The training set was further subdivided into a pre-training set and a validation set, with an established ratio of 8:2. These datasets were used independently to train and validate the ML frameworks, thereby avoiding bias (Zhao et al., 2023). A similarity-based splitting (SS) method was employed to segment the training set, primarily utilizing the MaxMin algorithm (MaxMinPicker function in the RDKit toolbox, available at <https://www.rdkit.org/>) and calculating molecular similarity using the Tanimoto distance (Sun et al., 2022). Additionally, the random splitting (RS) method (train\_test\_split function in the Scikit-learn toolbox, available at <https://scikit-learn.org/>) was used for model development, serving as a comparator. This comparative approach aimed to elucidate the differences in data distribution and model performance between the two distinct data partitioning methods.

### 2.2. Molecular Representation

As compared with MDs and the GCM methods, MFs can be more easily obtained and

understood than MDs; obtaining the MFs is also faster than obtaining atom groups in GCM (Zhong et al., 2020). Morgan fingerprint is a type of MF which encode structural features of molecules as binary vectors, where the numbers in each bit indicate the presence or absence of specific atom groups in the molecule (Rogers and Hahn, 2010). Binary Morgan fingerprints (BMF) can be generated using the RDKit program with the command “*AllChem.GetMorganFingerprintAsBitVect()*”. BMF can qualify the presence or absence of an atom group, but it cannot quantify the number of occurrences of that group within a molecule. To address this limitation, a novel fingerprint known as count-based Morgan fingerprints (CMF) (Zhong and Guan, 2023), was developed by counting the occurrences of each substructure within the BMF. In this process, all positions with a value of 1 in the BMF were replaced by the corresponding count of each substructure. This process was mainly realized with the command “*AllChem.GetHashedMorganFingerprint()*”. More details can be found in [Text S1](#).

### 2.3. Model training

Initially, we established the parameters for the CMF. Random Forest (RF) is a straightforward and rapid ML technique capable of handling high-dimensional vectors. We employed RF with its default parameters to ascertain the optimal parameters for CMF. The length and radius are two critical parameters for CMF. Typically, increasing the length and radius allows for the storage of more structural features, but it may also diminish computational efficiency and increase the risk of overfitting. To determine the optimal radius and length, we conducted orthogonal tests, varying the radius from 1 to 3 and the vector length across 512, 1024, 2048 and 4096. We then separately built models for each combination of radius and length to identify the most suitable parameters.

Subsequently, we tested various ML methods based on the optimized CMF to identify the most suitable algorithm. Random Forest (RF), Artificial Neural Network (ANN), Supporting Vector Machine (SVM), k-Nearest Neighbor (KNN), etc. are among the mostly commonly applied regression algorithms in ML. We utilized these ML techniques to construct QSAR models and compared their performance to determine the optimal algorithm. The performance of the model was influenced by a wide range of adjustable independent parameters, known as hyperparameters. The GridSearchCV function in Scikit-learn toolbox was employed for hyperparameter optimization. The main hyperparameters of these algorithms were detailed in **Table S1**. Ultimately, the best model was generated and saved.

We primarily evaluated the performance of the developed models by the root mean square error (RMSE) and coefficient of determination ( $R^2$ ). Descriptions of these indexes are provided in **Text S2**.

#### *2.4. Model interpretation*

SHAP, an attribution method used in artificial intelligence, has been widely used to interpret QSAR models (Wang et al., 2022). It was adopted to quantify the influence of different molecular fragments on the output. We employed the SHAP toolbox (<https://github.com/slundberg/shap>) to calculate SHAP values for each bit in MFs generated by each molecule. A larger positive or negative SHAP value of a feature usually indicates a more positive or negative effect on the model output. More details can be found in **Text S3**.

To overcome the limitation of SHAP analysis that tends to overlook low-frequency features, we proposed a new concept of a toxicity index (TI) based on Shapley values and feature matrix. By incorporating the average value of substructure features, this index mitigates



the dominance of high-frequency features, offering a more balanced approach to model interpretation. Detailed discussion are provided in section 3.6.2.

### 2.5. AD Characterization

AD<sub>FP-AC</sub> (Wang et al., 2022), which considers both MF similarity and toxic endpoint of compounds, was proposed to describe the ADs. AD<sub>FP-AC</sub> employs local discontinuity scores ( $S_{LD}$ ) to detect compounds on the activity cliffs.  $S_{LD}$  for a molecule  $m$ ,  $S_{LD}(m)$ , was calculated as

$$S_{LD}(m) = \frac{1}{K} \sum_{\{n | S(m,n) \geq S_{cutoff}, m \neq n\}} S(m,n) \cdot D(m,n)$$

where  $n$  represents a molecule in a community set  $\{n | T_c(m,n) \geq S_{cutoff}, m \neq n\}$  whose similarities with  $m$   $S(m,n)$  is larger than  $S_{cutoff}$ ,  $D(m,n)$  represents absolute difference between the properties of  $m$  and  $n$ , and  $K$  is the number of the set elements. In principle, molecules with small  $S_{LD}$  values, which means the training set contains molecules with similar structures and the properties of these molecules do not differ significantly, are likely to be predicted accurately. When  $S_{LD}(m)$  is smaller than a threshold named  $C_{cutoff}$ , molecule  $m$  is regarded predictable.

## 3. Results and discussion

### 3.1. Effects of the MF length and radius

To determine the optimal radius and length of CMF for the modeling process, a series of CMFs with different radii and lengths were generated and evaluated on a RF algorithm using default parameters. The results were presented in detail in Fig. 2A. Overall, shorter MF length and larger radius led to a weaker precision. This phenomenon was usually attributed to bit collisions (Zhong et al., 2020), yet the underlying principles of these collisions have not been thoroughly investigated. Consequently, we delved into this process to provide a detailed

191 exploration. Briefly, each atom in a distinct environment is assigned a unique identifier at the  
192 very beginning. As the radius expands, new identifiers are produced through an iterative process  
193 of amalgamating the identifiers of neighboring atoms. This procedure is iterated 'radius' times,  
194 enabling the identifiers to encapsulate information about atoms located at increasing distances  
195 from the central atom. As a result, more substructures can be discovered with a larger radius.  
196 The next step is to map these identifiers to CMFs of a certain length. The mapping of  
197 substructures to MF bits is a simple remainder hashing process. For example, substructure with  
198 an identifier of 99 will be mapped to the 9th bit ( $99 \bmod 10$ ) when generating MF with a length  
199 of 10. Thus, when the length of MF is short, the likelihood of collisions goes a step further.  
200 When bit collision occurs, different structures are encoded into the same bit and the accuracy  
201 of the models will be reduced as a result. Unlike collisions in BMF, when CMF bits collide, the  
202 bit value changes to the sum of two different substructures, indicating that CMF-based models  
203 may be more susceptible to bit collisions. Therefore, when using CMF to build models, a larger  
204 length may be necessary. More details are provided in the [Text S4](#).

205 Taking into account both model accuracy and computational efficiency, the optimal radius  
206 and length were determined to be 1 and 2048, respectively. When radius was set to 1, the central  
207 atom captured information up to one atom away. The resulting substructures appeared quite  
208 small, yet they were sufficient to encode the structural information of common functional  
209 groups, such as hydroxyl, carboxyl, ester and halogen atoms. As the radius increased, larger  
210 substructures were discovered, but these structures were found to degrade model performance  
211 and were deemed unnecessary. The inclusion of irrelevant features can indeed diminish model  
212 accuracy (Zhang et al., 2021). Consequently, various techniques are employed to restrict and

eliminate irrelevant features, such as pruning, dropout and regularization penalties. A larger radius may be advantageous for more intricate research, such as capturing the relative positions of substituents on complex ring systems.

### 3.2. Model performance

In order to assess the efficacy of various algorithms, we established models using different MF methods based on the optimal CMF and refined the hyperparameters for each model to enhance performance. As illustrated in Fig. S2, ANN method achieved the highest accuracy. In contrast to traditional ML models, the hyperparameters of ANN are more flexible and adjustable, encompassing the number of neurons, the number of hidden layers, the activation function, and the learning rate, among others. A greater number of hidden layers and neurons endow the ANN with a stronger capability to model or abstract complex phenomena, thereby enabling it to simulate more intricate models.

The optimal model performance is shown in Fig. 2B, with  $R^2$  and RMSE being 0.76 and 0.53 on the test set (n=115) in AD, respectively. The architecture of the ANN model is provided in Fig. S3. These results significantly surpass those obtained by regression models based on MDs using the same dataset, as reported by Samanipour et al. (2023) ( $R^2=0.7$ ) and Cassotti et al. (2015) ( $R^2=0.62$ ), proving the feasibility of using CMF to establish a high-precision model. Samanipour et al. calculated 2757 MDs and Cassotti et al. calculated 3582 MDs, with 8 and 6 filtered out through complex feature engineering, respectively, to establish their respective models. Our model circumvents the need for complex feature engineering and the use of difficult-to-interpret MDs (i.e.: AATS0p, SpMax3\_Bhm, VP-0), which not only simplifies the model development process but also enhances model interpretability.

### 3.3. BMF vs. CMF

To ascertain the benefits of CMF in the representation of molecular structures, we assessed the performance of different ML methods based on BMF and CMF, and the results are presented in Fig. 2C. The default values for the hyperparameters of these ML methods were employed, rather than those adjusted based on CMF, to maintain fairness competition in the testing process. The results consistently demonstrated that CMF outperformed BMF across all ML methods. This discrepancy may be attributed to the fact that BMF is unable to quantify the number of atom groups within a molecule, thereby limiting its capacity to accurately describe molecular structures. For example, n-pentane (SMILES: CCCCC) and n-hexane (SMILES: CCCCCC) are structurally similar straight alkanes differing by a single methylene group. When the radius of the BMF is less than 2, identical BMFs are generated for both compounds, indicating that BMF fails to distinguish between these two molecules. Consequently, identical features lead to identical predicted values. In contrast, CMF can account for the variations in the number of substructures, generating distinct characteristics for compounds with similar structures and enabling diverse predictions. The generation of CMF aligns with the concept of the group contribution method, which has been demonstrated to be capable of constructing high-precision AFT prediction models (Martin et al., 2001).

It is important to note that CMF can have a very large value on a single bit, which may preclude the use of certain distance measurement methods that are sensitive to individual features. For example, distances calculation methods such as Euclidean distance and Chebyshev distance could be influenced by a single bit significantly, which is not reasonable for the calculation of similarities of molecules represented by CMF. Furthermore, the traditional

Tanimoto similarity for BMFs (B-Tanimoto), which is based on statistics of 1s in BMFs, is not suitable for calculating the similarity of CMFs as well. To address this limitation, we used a new calculation method, referred to as C-Tanimoto, specifically tailored for CMF. This method was detailed in [Text S5](#).

Several groups of molecules with a similarity calculated by B-Tanimoto greater than 0.8 and difference of AFT values greater than 2 is provided in [Table S2](#). These groups possess similar features but significant differences, which complicates the learning process for the model. The high proportion of homologues (86.2%) in the dataset may contribute to the suboptimal performance of BMF-based models. The similarity calculated by C-Tanimoto was also counted. The average similarity in the list decreased from 0.95 to 0.51, suggesting that CMF can fully recognize the differences among these compounds and facilitate diverse predictions.

#### *3.4. Random splitting vs. similarity-based splitting*

In order to test the impact of dataset partitioning methods on prediction results, the initial training set was divided by different proportions and methods. Random parameters ('seed' in the 'MaxMin' function and 'random\_state' in the 'train\_test\_split' function) were varied to generate different pre-training and validation sets. Two partitioning methods with ten distinct random parameter values were tested on a RF regressor using default hyperparameters. As shown in [Fig. 2D](#), for models based on the random splitting (RS) method, the  $R^2$  and RMSE of the validation set (20%) varied from 0.42 to 0.63, and from 0.92 to 1.11, respectively, across the ten experiments. For models based on the similarity-based splitting (SS) method, the  $R^2$  and RMSE of the test set ranged from 0.71 to 0.75 and from 0.57 to 0.61, respectively. These results

indicate that the MaxMin algorithm, which relies on Tanimoto similarity, aids the model in achieving higher accuracy and stability. The subsets generated by the SS method exhibit an average size of 727 features, while conversely, the subsets extracted by the RS method are characterized by a smaller average feature count, amounting to only 676. Lower Tanimoto similarity values indicate greater chemical dissimilarity. The SS method prioritizes the selection of molecules with significant structural differences for training, ensuring the diversity of features in the training set. Consequently, the model can learn more complex relationships between features and become more robust, enhancing its performance. We also calculated the similarity between each molecule in the validation set and the closest molecule in the pre-training set under both partitioning methods. As shown in Fig. S4, the mean of this indicator for the SS method was higher ( $0.78 > 0.64$ ), which was consistent with our conjecture. Moreover, the diverse structures in the pre-training set generated by the SS method can lead to a broader range of ADs.

Furthermore, we incrementally adjusted the proportion of the pre-training set used to train a RF regressor with default hyperparameters. The remaining portion of the training set was utilized for validating the model performance. As shown in Fig. 3, model performance gradually enhanced as the training set size increased. The SS-based models consistently outperformed the RS-based models on the validation set, indicating that the SS method is more effective in predicting the remaining dataset when the training set size remains constant. This suggests that prioritizing the selection of compounds with diverse structures based on the MaxMin method is advantageous for establishing models and evaluating the overall dataset. It is interesting that SS method often performed poorly on the training set. This may be due to the low similarity

and diverse compound structures within the pre-training set, which poses a more challenging task for accurate predictions.

### 3.5. Model interpretation by SHAP

We further used the SHAP method to interpret the result analyzed by the final model. Since the ‘DeepExplainer’ in SHAP toolbox used an approximate approach to reduce computational costs, we used the average of 10 operations for all of the following model explanations. As shown in the Fig. 4A and 4C, we identified the top 12 features that exert the most significant influence. Red on the right and blue on the left indicate that the structure will increase toxicity, while conversely, it will reduce toxicity. SHAP images based on BMF only have red and blue colors, which is because the bits are binary. CMF incorporates the counts of atom groups, and the numbers in each bit can exceed 1, resulting in a gradient of colors ranging between red and blue (Zhong and Guan, 2023). As shown in the figure, the middle section exhibits purple, and the color variation is generally consistent with the color legend. This indicates that the increase in substructures leads to a continuous increase or decrease in predicted values, which is consistent with our domain knowledge or experience. CMF’s ability to represent features with multiple values contributes to the generation of a broader range of activation values in ANN and Shapley values in the SHAP plot, thereby enhancing the performance of CMF-based models.

The SHAP analysis revealed that substructures with exceptionally high toxicity to fish include substituted benzenes and alicyclic ring (e.g., features 1380, 1019), unsaturated carbon (e.g., features 694, 1873), multiple methylene groups (e.g., feature 1911) and halogen atoms (e.g., features 728, 1683). In general, chemicals containing these substructures tend to possess

strong lipophilic properties, making it easier to partition into the hydrophobic phase (cell membrane), enabling them to remain within the biological system for extended periods and interfere with the organism. The presence of methylene groups extends the carbon chain length, contributing to higher lipid solubility. The presence of halogen atoms and halogen-containing substituents in compounds enhances their lipophilicities and finally their bioavailability as well. The strong electronegativity of chlorine and bromine atoms reduces the electron density of the carbon atoms they are bonded to, making nearby carbon atoms more prone to nucleophilic substitution reactions. Unsaturated carbon's unshared electron pairs enable them to readily participate in chemical reactions, potentially leading to changes in the molecular structure of the organism and affecting normal biological functions. Some substructures (e.g., features 1057, 841, 294) were observed to reduce toxicity, a finding that has infrequently been discussed in previous studies. A commonality among these structures is the presence of methyl groups, which may be attributed to the fact that methyl is a common substituent with relatively lower toxicity, occupying substitution sites to prevent the formation of more toxic substances.

### *3.6. Reactive toxicity analysis of substructures*

#### *3.6.1. Exposure assistance vs. reactive toxicity*

This study did not take into account the steric toxic effect. Consequently, we ascribed the acute toxicity observed in fish exclusively to the exposure assistance and reactive toxicity. In other words, the former helps compounds enter the interior of the organism, while the latter has stronger reactivity. Uptake, transport, and distribution of organic toxicants via passive diffusion are best modeled by hydrophobicity, which is typically quantified by log  $K_{OW}$ , indicating its importance in exposure assessment (Singh et al., 2013). We attempted to highlight the



substructure of reaction toxicity by excluding the contribution of  $K_{OW}$ . The values of standardized log  $K_{OW}$ , which can be generated by the RDKit program using the command “*Descriptors.MolLogP()*”, were concatenated to the CMFs as an additional feature. All subsequent analyses were conducted in the model incorporating  $K_{OW}$ .

As shown in Fig. 4B, the  $K_{OW}$  feature obtained the highest score in the importance ranking of all 2049 features, underscoring its substantial influence on toxicity. In fact, some AFT predictions are entirely based on  $K_{ow}$  (Lambert et al., 2022). The alteration in the importance ranking of SHAP features can, to some extent, distinguish between the lipid solubility and biological toxicity of substructures. Overall, the most significant substructures did not undergo substantial changes upon the introduction of  $K_{OW}$ . Among the 12 most important features of the two models, 8 were found to be identical. This coincidence can be explained by the observation that highly lipophilic substructures facilitate the penetration of chemicals into the organism and are also commonly associated with the manifestation of biological toxicity (Zeliger, 2013). However, as we marked in Fig. 4C, the importance ranking of some substructures changed. An increase in ranking typically implies stronger toxicity rather than higher lipid solubility, and vice versa. For example, the importance ranking of carbon chains (feature 1911) decreased, suggesting that its toxicity is predominantly achieved through passive transport. Conversely, sections of the benzene ring devoid of substituents (feature 694) were regarded as exhibiting lower reactivity. On the contrary, substructures with rising rankings exert their impact on toxicity through reactivity. For example, methyl substituents (features 1057, 294), typically classified as low-toxicity groups, tend to occupy substitution positions and avoid the production of more toxic substitutes. As a result, the reduction in reactivity enhanced their overall ranking.

However, the method based on SHAP diagram showed limitations in identifying highly toxic substructures. For example, as depicted in Fig. 4A, features 1873 and 1750 were deemed more significant than feature 728, which may seem incongruous when identifying highly toxic substructures. This discrepancy can be attributed to the influence of eigenvalue distribution on the ranking within the SHAP diagram, where substructures that are more prevalent in the dataset are often regarded as more critical by the model. The ranking in the SHAP diagram can be easily affected by occurrence-frequency, and thus lead to an oversight of substructures that are highly toxic yet less frequent in the dataset. A detailed explanation of this phenomenon is provided in Text S6.

### 3.6.2. Highly toxic substructures

To address this issue and identify such substructures, we adopted an approach where we approximately attribute Shapley values based on the difference between the sample feature and the mean value of that feature in the background dataset. More details can be found in Text S7. This method allowed us to highlight highly toxic substructures that might be underestimated in the model's importance rankings. Therefore, we proposed the definition of TI, and the TI of feature  $j$ ,  $TI_j$ , can be calculated as

$$TI_j = \frac{1}{n} \sum_{i=1}^n \frac{S_{ij}}{F_{ij} - f_{ave\ j}}$$

where  $n$  equals to the number of samples in the explanatory background dataset. To mitigate the effects of randomness,  $S_{ij}$  and  $F_{ij}$  denote the values of the  $i$ -th sample and  $j$ -th feature in the Shapley value matrix and feature matrix, respectively, and  $f_{ave\ j}$  represents the average value of the  $j$ -th feature across the dataset. To ensure statistical significance, we only considered substructures that appeared more than 10 times in the dataset. The calculated results

are detailed in Table S3. The features with the top 12 TI values are shown in Fig. 5.

The identified substructures primarily encompassed those containing sulfur and phosphorus elements (features 97, 116, 192 and 1729), bromine atoms (feature 728), lipid moieties (feature 1386), polycyclic substituents (feature 352), unsaturated carbon linkages (features 916, 1366 and 1645), nitrogen-containing ring substituents (feature 1145) and nitro groups (feature 715). Despite their significant toxicological relevance, these substructures were less prevalent in the dataset, which led to their assignment of lower rankings in the SHAP diagram. This underrepresentation in the dataset may obscure their importance in the model's overall assessment of toxicity.

It is worth noting that when MFs are used as features in the model, the SHAP method assesses the importance of individual feature bit rather than the entire structure. Consequently, the significance of a particular structure should approximately be equated to the cumulative importance of all the substructures it encompasses. This implies that when the radius of a substructure exceeds 1, its contribution may be misestimated. For example, as shown in Fig. 6A, the toxic potential of a parathion bond should be the combined toxicity of the sulfur atom (feature 97), the phosphorus atom (feature 192) and the parathion bond (feature 1729), rather than being attributed solely to the bit representing the parathion bond (feature 1729). This will result in a substantial underestimation of the toxicity associated with parathion-like structures. This concept has often been misinterpreted in previous literature, leading to inaccurate assessments of the impact of certain large substructures. Similarly, as shown in Fig. 6B, the toxicity of polycyclic substitutions was underestimated, since both the benzene ring (feature 1873) and benzene ring substitution (feature 1380) were identified to increase toxicity. The

411 reactivity toxicity of lipid structures (feature 1386) is typically regarded as low, and their  
412 identification as significant contributors to toxicity was unexpected. This may be attributed to  
413 their lipophilic nature, which indicated that the incorporation of the  $K_{OW}$  cannot entirely account  
414 for the effects of lipid solubility on toxicity. Furthermore, oxygen-containing structures were  
415 often perceived by models as reducing toxicity, which could result in an overestimation of the  
416 toxic potential of such substructures.

417 Overall, the definition of TI facilitated the identification of highly toxic substructures that  
418 were previously overlooked in the SHAP diagram. Based on this, we pinpointed two  
419 substructures whose toxicity was severely underestimated: the parathion and polyaromatic  
420 moieties. Parathion-like substructures are commonly employed in pesticides, and polycyclic  
421 aromatic hydrocarbons are frequently associated with high carcinogenicity. Additionally,  
422 halogen was also highlighted by TI analysis. Consequently, we further applied TI to analyze  
423 the reactive toxicity of different organosulfur, halogenated and polyaromatic substructures.

### 424 3.6.3. *Reactive toxicity of organosulfur substructures*

425 Sulfur is a major inorganic element, essential for the entire biological kingdom because of  
426 its incorporation into amino acids, proteins, enzymes, vitamins and other biomolecules. As  
427 shown in Fig. 6A, we examined the toxicity associated with various sulfur-containing  
428 substructures within the dataset. Notably, the dataset does not include sulfones or sulfoxides.  
429 Consequently, feature 350 was attributed to the sulfonic group, feature 116 corresponded to the  
430 thioether linkage, and feature 97 was identified in the parathion bond as well as a minor fraction  
431 of compounds containing a C=S double bond. Overall, the sulfonic group exhibited the lowest  
432 level of toxicity, whereas the sulfur-phosphorus structure demonstrated the highest toxicity. The

sulfonic groups are characterized by their hydrophilicity and stability, and sulfonamide drugs are widely used in aquaculture, such as sulfamethoxazole, trimethoprim. Parathion, widely utilized in organic pesticides, primarily gains entry into the body through the gills. Within the liver, parathion-like structures can be metabolized into the more toxic paraoxon-like structures, which inhibits the activity of cholinesterase, thereby impairing its capacity to hydrolyze acetylcholine. This disruption leads to an accumulation of acetylcholine, which subsequently blocks excitatory receptors, ultimately resulting in neurotoxic effects (Benke et al., 1974).

#### 3.6.4. Reactive toxicity of halogenated substructures

Halogenated organic compounds are organic molecules that contain fluorine (F), chlorine (Cl), bromine (Br) or iodine (I) atoms. These chemicals cause adverse effects on a diverse range of plants and animals as well as humans, and are one of the largest groups of environmental chemical pollutants. Therefore, we conducted an investigation into the potential reaction toxicity of halogenated substructures. The predominant halogen forms in our dataset are the elements F, Cl and Br. As depicted in Fig. 6C, halogens generally exhibited toxicological properties. Overall, F exhibited the lowest level of toxicity, whereas Br demonstrated the highest toxicity. Generally, F has relatively low biological activity and a number of fluorinated hydrocarbons are regarded as inert in toxicological contexts. The modification of drug formulations frequently involves the inclusion of F to modulate the dissociation constant, thereby facilitating the efficient and smooth delivery of the active pharmaceutical ingredient to its target site. Br was deemed to possess the highest toxicity. The largest atomic volume of Br presents the lowest electronegativity and highest lipophilicity. Additionally, the toxicity and metabolic products of brominated compounds are usually stronger. For instance, brominated

compounds that are part of disinfection by-products are typically associated with higher levels of toxicity. However, it was intriguing to observe that certain halogen-containing substructures (features 1453, 1456 and 495) were associated with a decrease in toxicity, which appears counterintuitive. We attributed this phenomenon to a compensatory mechanism within the model. It is known that aromatic halides generally exhibit greater toxicity than their aliphatic counterparts (Han et al., 2021). The model may have inadvertently overestimated the toxicity of aliphatic halide substructures, thereby compensating for the overall halogen-related toxicity predictions.

### 3.6.5. Enhanced reactive toxicity by aromatic moieties

As previously discussed, analogous structures can exhibit varying degrees of toxicity, and moreover, the identical specific substructure may demonstrate distinct toxicological profiles contingent upon the specific chemical context in which it is situated. For instance, aromatic halogenated compounds were generally more toxic than their aliphatic halogenated counterparts, as previously mentioned. Similarly, as illustrated in Fig. 4B, features 1019 and 1 corresponded to methyne (CH) substructure located on a straight chain and within a ring, respectively. This suggested that the substitution of a fatty ring with certain groups results in higher toxicity. Moreover, as depicted in Fig. 6D, we explored the toxicity of hydroxyl groups across various chemical contexts. Notably, the model attributed greater significance to the larger structures harboring hydroxyl groups rather than to the hydroxyl groups themselves. The toxicity associated with phenolic hydroxyl groups was observed to be substantially higher compared to that of alcohol hydroxyl groups. This discrepancy can be attributed to the conjugated electron system of aromatic rings, which effectively disperses electron density

across the hydroxyl groups. The electron distribution facilitates the departure of hydrogen atoms from phenolic hydroxyl groups, thereby enhancing the electrophilic substitution reaction activity of phenols. It is noteworthy that the presence of phenolic hydroxyl groups also leads to an increase in benzene ring-substituted structure (feature 1380), which significantly elevates toxicity.

### 3.7. Application Domain

Hypertoxic and prevalent toxic substructures warrant special attention. We endeavored to identify such structures within the HPV and FDA chemical databases. Prior to this, the AD of this CMF-based model should be determined. As shown in Fig. S5, we calculated the performance of the models under various  $S_{\text{cutoff}}$  and  $C_{\text{cutoff}}$  thresholds, as well as the number of chemicals in test sets within the AD. A notable improvement was observed in the test group where chemicals outside the AD were removed by setting appropriate  $S_{\text{cutoff}}$  and  $C_{\text{cutoff}}$ . The threshold should satisfy both criteria of the lowest RMSE and the least number of chemicals outside the AD. The results indicated that prediction could be stable and effective with  $S_{\text{cutoff}} \geq 0.6$  and  $C_{\text{cutoff}} \leq 1$ , and the thresholds can be employed with the model for AFT prediction. The accuracy improved to  $R^2$  up to 0.81 by further constraining the thresholds, but chemicals within AD decreased to 43.3% at the same time. Hyper-stringent thresholds can be applied to some studies that require higher precision.

However, the definition  $C_{\text{cutoff}}$  under a high similarity may also overlook inherent differences, that is, the overall structure of the two molecules is similar, resulting in significant toxicity differences due to local differences. Significant differences in toxicity may be generated by the presence of active sites or incorrect measurements. For example, as shown in

**Tabel S2**, nitro may greatly affect AFT, which can be verified through further experiments. During this process, suspicious experimental values can also be verified. Overall, the definition of  $C_{\text{cutoff}}$  can be used for the discovery of active sites and the validation of datasets.

### 3.8. Environmental Applications

Beyond the application in predicting acute toxicity, our model was also utilized to identify prevalent toxic substructures, as well as compounds that incorporate highly toxic substructures. Initially, CMFs were generated for the chemicals in HPV and FDA. The proportion of non-zero bits was calculated to quantify the number of features in the dataset, and a threshold of 0.1 was set to indicate a significant presence. This information is detailed in **Table S4**. Subsequently, we identified toxic structures by referring to the SHAP plot. Specifically, substructures represented by bits 1380, 694, 1911, 1873, 1750 and 1683 were found to be prevalent and potentially toxic, and potentially highly toxic substances containing bits of 728, 1729 or 352 (Br, parathion and polycyclic aromatics) are provided in **Table S5**. These identified potential highly toxic substances may necessitate subsequent experimental validation, warrant cautious handling in industrial applications and everyday use, and prompt a proactive search for safer alternatives.

## 4. Conclusion

In summary, this study investigated the feasibility of constructing a high-precision model for AFT using MF. We identified the limitations of BMF and posited that the abundance of homologues in the dataset may contribute to the suboptimal performance of BMF. We successfully developed a high-precision model employing CMF and introduced a novel method for calculating CMF similarity. Transfer learning was demonstrated to discriminate between the



lipophilicity and reactivity of substructures. By utilizing the SHAP method and proposing a new definition of feature importance, TI, we identified key substructures associated with high toxicity but present in small quantities. Finally, we identified abundant toxic substructures and potential highly toxic substances within two external datasets, HPV and FDA. Additionally, we introduced a flexible AD that can accommodate research with varying levels of accuracy. A limitation of our study is the representation of compounds in a fragmented manner, which overlooks the potential toxicity that may arise from their holistic structures. Incorporating more characteristics of the overall structure into the modeling or conducting targeted experiments could be beneficial for further research on the fish toxicity of compounds.

#### **CRedit authorship contribution statement**

**Shangyu Li:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mingming Zhang, Peizhe Sun:** Writing – review & editing, Supervision, Methodology, Conceptualization.

#### **Declaration of competing interest**

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

#### **Data availability**

Data will be made available on request.

## Acknowledgments

This work was supported by the project from National Natural Science Foundation of China (No. 22176141 and 22322608).

## Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/>.

## References

- Wang, H.; Wang, Z.; Chen, J.; Liu, W. Graph attention network model with defined applicability domains for screening PBT chemicals. *Environ. Sci. Technol.* 2022, 56, 6774– 6785.
- Wang, Z.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environ. Sci. Technol.* 2020, 54 (5), 2575– 2584.
- Singh K P; Gupta S; Rai P. Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches. *Ecotox. Environ. Safe.* 2013, 95, 221– 233.
- 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.
- Sun, P.; Ma, H.; Li, S.; Yao, H.; Zhang, R. Prediction of second-order rate constants between carbonate radical and organics by deep neural network combined with molecular fingerprints. *Chin. Chem. Lett.* 2022, 33 (1), 438– 441.
- McLachlan, M. S.; Kierkegaard, A.; Radke, M.; Sobek, A.; Malmvarn, A.; Alsberg, T.; Arnot, J. A.; Brown, T. N.; Wania, F.; Breivik, K.; Xu, S. Using model-based screening to help discover unknown environmental contaminants. *Environ. Sci. Technol.* 2014, 48 (13), 7264– 7271.
- Neves, B. J.; Braga, R. C.; Melo-Filho, C. C.; Moreira-Filho, J. T.; Muratov, E. N.; Andrade, C. H. QSAR-based virtual screening: advances and applications in drug discovery. *Front.*

570 Pharmacol. 2018, 9, 1275.

571 Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large scale comparison  
 572 of QSAR and conformal prediction methods and their applications in drug discovery. J.  
 573 Cheminf. 2019, 11, 1- 16.

574 Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From Molecular  
 575 Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical  
 576 Prioritization. Environ. Sci. Technol. 2023, 57 (46), 17950- 17958.

577 Sheffield, T. Y.; Judson R. S. Ensemble QSAR modeling to predict multispecies fish toxicity  
 578 lethal concentrations and points of departure. Environ. Sci. Technol. 2019, 53 (21), 12793-  
 579 12802.

580 Martin, T. M.; Young, D. M. Prediction of the acute toxicity (96-h LC50) of organic compounds  
 581 to the fathead minnow (*Pimephales promelas*) using a group contribution method. Chem.  
 582 Res. Toxicol. 2001, 14, 1378– 1385.

583 Cao, H.; Peng, J.; Zhou, Z.; Yang, Z.; Wang, L.; Sun, Y.; Wang, Y.; Liang, Y. Investigation of  
 584 the Binding Fraction of PFAS in Human Plasma and Underlying Mechanisms Based on  
 585 Machine Learning and Molecular Dynamics Simulation. Environ. Sci. Technol. 2022, 57  
 586 (46), 17762- 17773.

587 Bo, T., Lin, Y., Han, J., Hao, Z., Liu, J. Machine learning-assisted data filtering and QSAR  
 588 models for prediction of chemical acute toxicity on rat and mouse. J. Hazard. Mater. 2023,  
 589 452, 131344.

590 Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.;  
 591 Zheng, M. Pushing the boundaries of molecular representation for drug discovery with the  
 592 graph attention mechanism. J. Med. Chem. 2020, 63 (16), 8749- 8760.

593 Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C.-Y.;  
 594 Hou, T. Chemistry-intuitive explanation of graph neural networks for molecular property  
 595 prediction with substructure masking. Nat. Commun. 2023, 14 (1), 2585.

596 Cassotti, M.; Ballabio, D.; Todeschini, R.; Consonni, V. A similarity-based QSAR model for  
 597 predicting acute toxicity towards the fathead minnow (*Pimephales promelas*). SAR QSAR  
 598 Environ. Res. 2015, 26, 217- 243.

599 Zhao, J.; Shang, C.; Yin, R. Developing a hybrid model for predicting the reaction kinetics

600 between chlorine and micropollutants in water. *Water Res.* 2023, 247, 120794.

601 Rogers, D; Hahn M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 2010, 50(5), 742-

602 754.

603 Zhong, S.; Guan, X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable

604 Molecular Representation in Developing Machine Learning-Based Predictive Regression

605 Models for Water Contaminants' Activities and Properties. *Environ. Sci. Technol.* 2023,

606 57 (46), 18193- 18202.

607 Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards Better Process

608 Management in Wastewater Treatment Plants: Process Analytics Based on SHAP Values

609 for Tree-Based Machine Learning Methods. *J. Environ. Manage.* 2022, 301, 113941.

610 Zhang, N.; Deng, S.; Cheng, X.; Chen, X.; Zhang, Y.; Zhang, W.; Chen, H. Drop redundant,

611 shrink irrelevant: Selective knowledge injection for language pretraining. *IJCAI.* 2021,

612 4007- 4014.

613 Lambert, F. N.; Vivian, D. N.; Raimondo, S.; Tebes-Stevens, C. T.; Barron, M. G. Relationships

614 between aquatic toxicity, chemical hydrophobicity, and mode of action: Log K<sub>OW</sub> revisited.

615 *Arch. Environ. Contam. Toxicol.* 2022, 83(4), 326- 338.

616 Zeliger, H. I. Lipophilic chemical exposure as a cause of cardiovascular disease. *Interdiscip.*

617 *Toxicol.*, 2013, 6 (2), 55- 62.

618 Benke, G. M.; Cheever, K. L.; Mirer, F. E.; Murphy, S. D. Comparative toxicity,

619 anticholinesterase action and metabolism of methyl parathion and parathion in sunfish and

620 mice. *Toxicol. Appl. Pharmacol.*, 1974, 28(1), 97-109.

621 Han, J.; Zhang, X.; Jiang, J.; Li, W. How Much of the Total Organic Halogen and

622 Developmental Toxicity of Chlorinated Drinking Water Might Be Attributed to Aromatic

623 Halogenated DBPs?. *Environ. Sci. Technol.* 2021, 55, 5906- 5916.

## Figures:

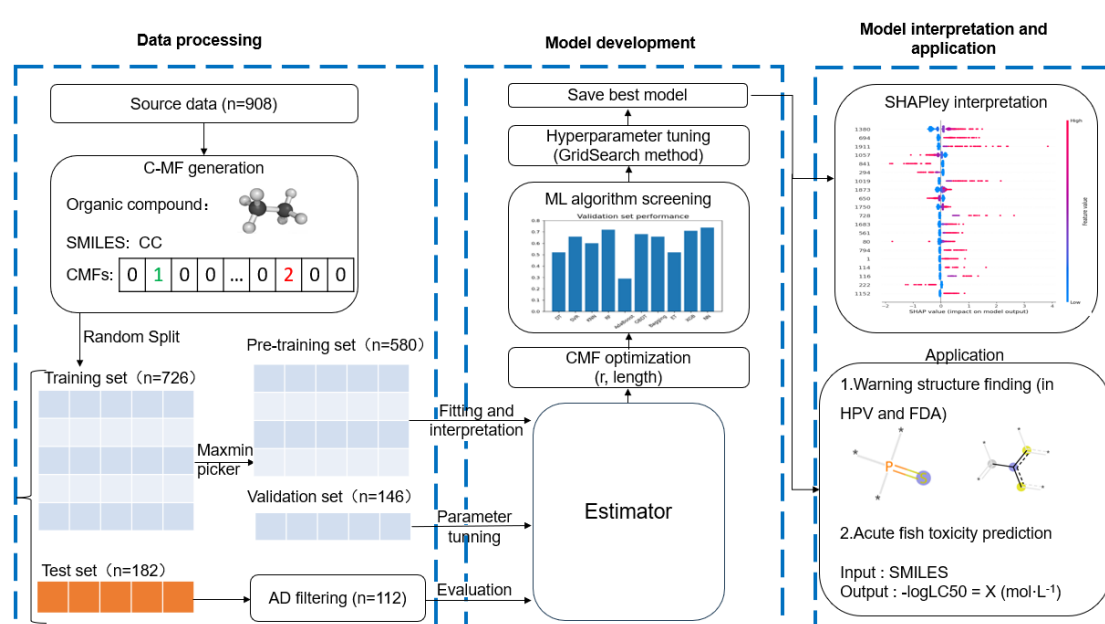


Fig. 1. The working program for the establishment of the hybrid model.

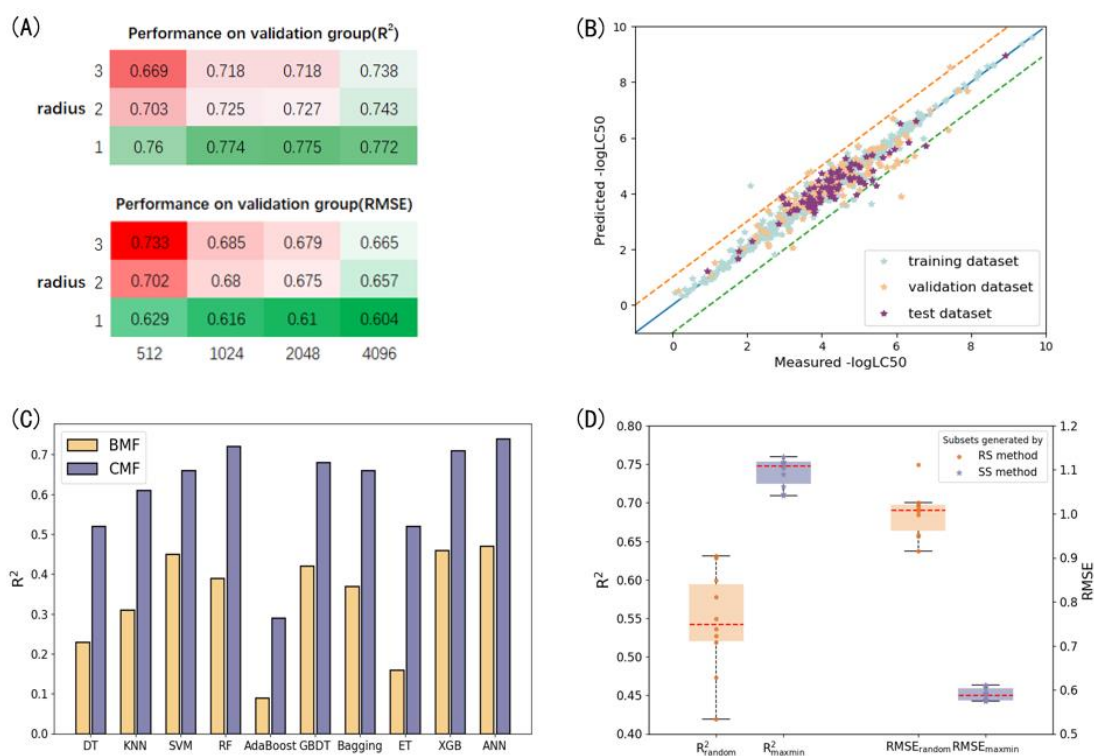


Fig. 2. (A) Performance of RF models on the validation set developed by CMFs with different radii and lengths. (B) The scatterplot of the predicted vs. the experimental values of  $-\log LC50$  for pre-training, validation and test datasets. (C) Plots of  $R^2_{\text{Validation}}$  of different algorithms using default parameters based on BMF and CMF. (D) Model performance on the variable validation sets. The RS and SS methods were used to generate pre-training sets respectively and the remaining sets were used as the validation set.

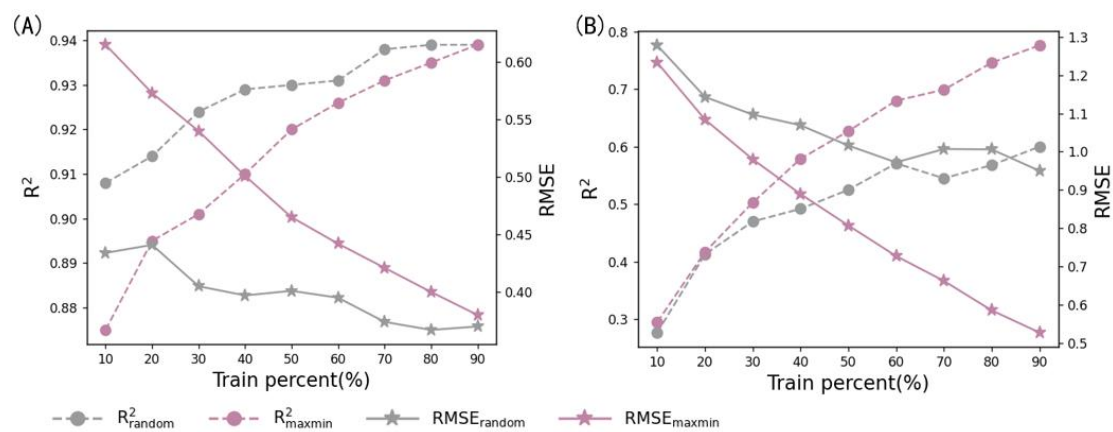


Fig. 3. Model performance of (A) training set and (B) validation set with different size of training set under different dataset partitioning methods.

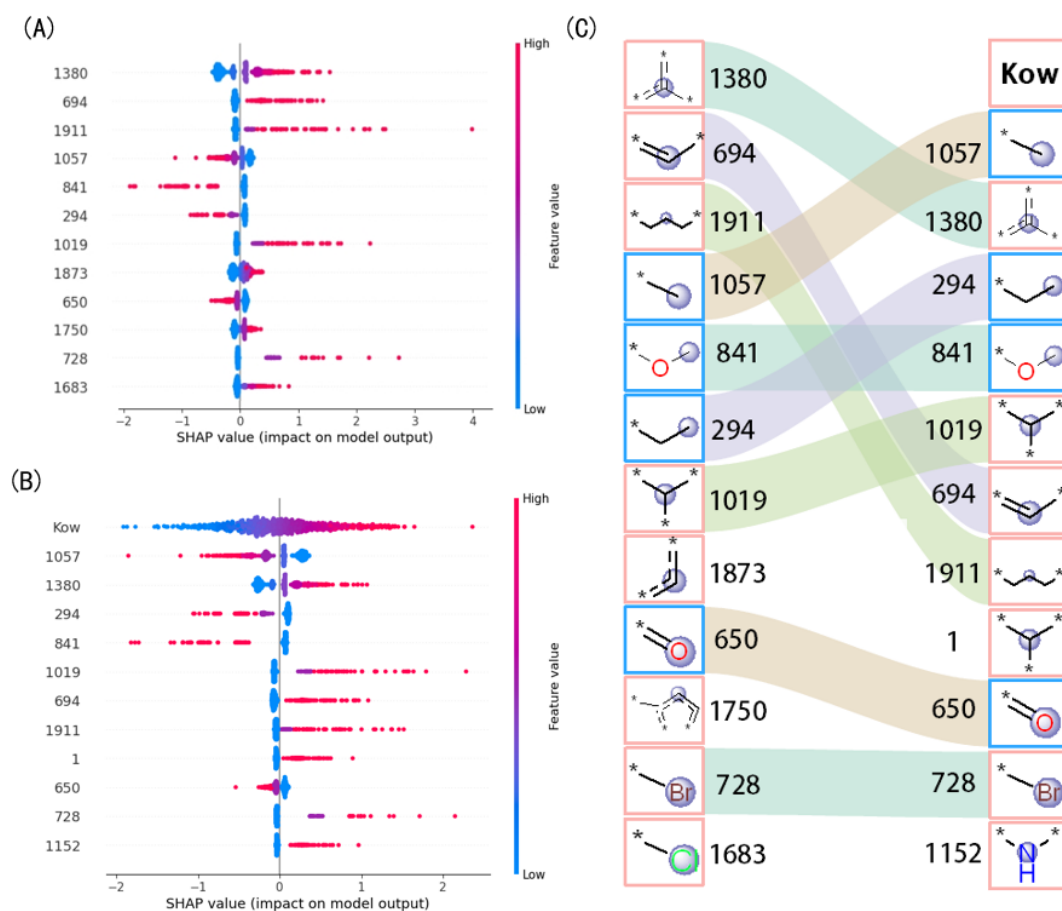


Fig. 4. The summary plot of the top 12 important features (A) before and (B) after introducing the feature  $K_{ow}$  and (C) the effects of top 12 important features on the  $-\log LC_{50}$  values.





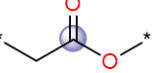
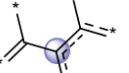

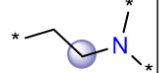




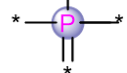
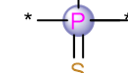
MF #	Top 1 (97)	Top 2 (728)	Top 3 (1386)	Top 4 (352)	Top 5 (915)	Top 6 (1145)
Feature						
Toxicity Index	0.604	0.526	0.499	0.476	0.445	0.402
SHAP Ranking	20	10	60	26	57	53
MF #	Top 7 (1366)	Top 8 (1645)	Top 9 (715)	Top 10 (116)	Top 11 (192)	Top 12 (1729)
Feature						
Toxicity Index	0.395	0.354	0.354	0.352	0.339	0.337
SHAP Ranking	18	27	13	28	34	32

Fig. 5. The summary plot of features with the top 12 toxicity index.

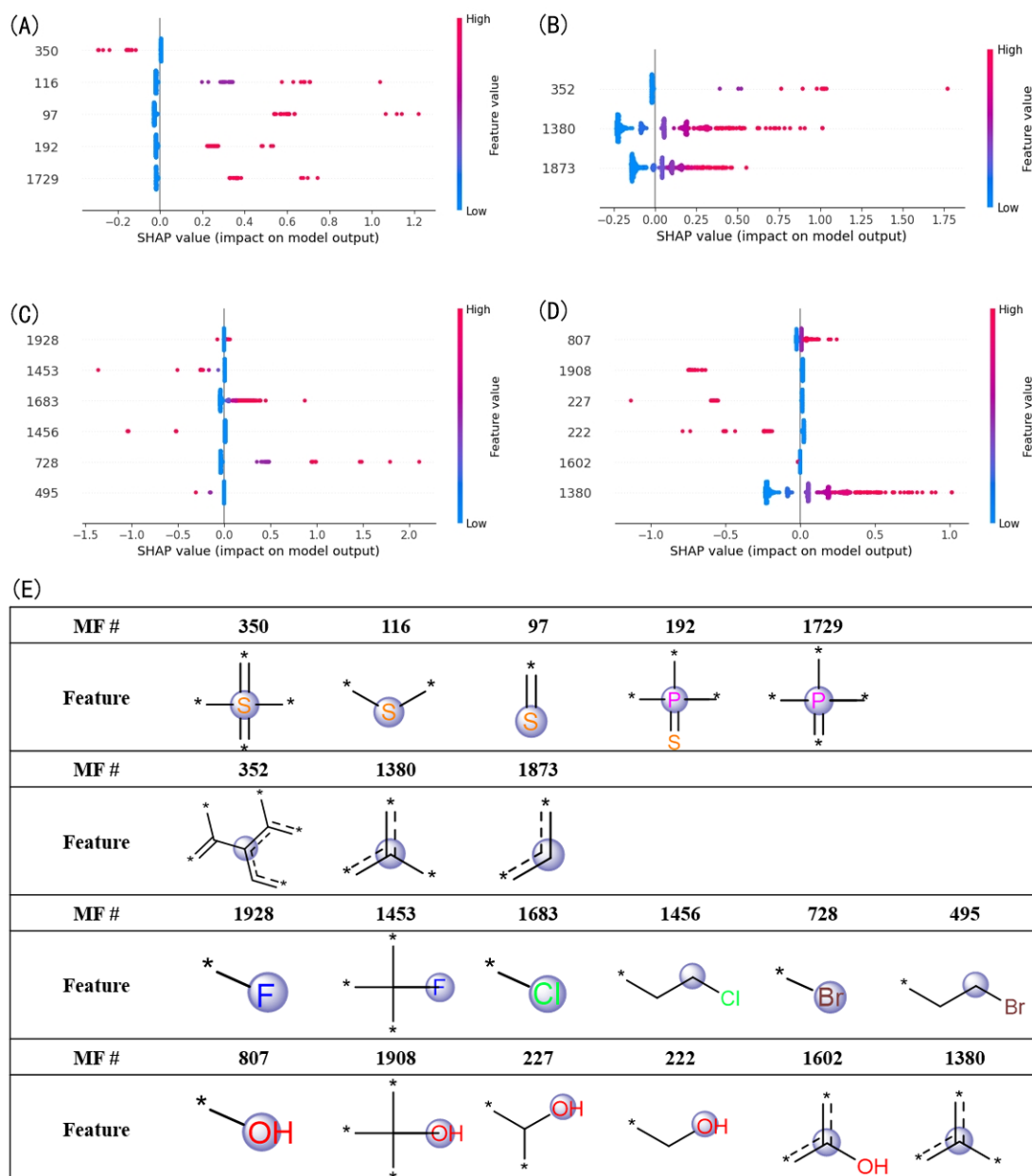


Fig. 6. The summary plot of (A) elemental sulfur and phosphorus related substructures, (B) polycyclic substitution related substructures, (C) halogen related substructures, (D) hydroxyl related substructures and (E) feature meanings.