Toxicology Research

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Toxicology Research
www.rsc.org/toxicology
ROYAL SOCIETY OF CHEMISTRY
Chinese Society Of Toxicology
bts
THE BRITISH TOXICOLOGY SOCIETY

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/toxicology

In Silico Prediction of Chemical Aquatic Toxicity with Chemical Category Approaches and Substructural Alerts

Lu Sun[a], Chen Zhang[a], Yingjie Chen[a], Xiao Li[a], Shulin Zhuang[b], Weihua Li[a], Guixia Liu[a], Philip W. Lee[a], Yun Tang[a],*

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

[a]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China;

[b]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China.

*Corresponding Author: Tel: +86-21-64251052; Fax: +86-21-64251033

E-mail address: ytang234@ecust.edu.cn

KEYWORDS: Aquatic toxicity; machine learning method; ternary classification; local model; global model

Toxicology Research Accepted Manuscript

Abstract

Aquatic toxicity is an important endpoint in evaluation of chemical adverse effects on ecosystems. In this study, *in silico* models were developed for prediction of chemical aquatic toxicity on different fish species. At first a large data set containing 6422 data points on aquatic toxicity with 1906 diverse chemicals was constructed. Using molecular descriptors and fingerprints to represent the molecules, local and global models were then developed with five machine learning methods based on three fish species (rainbow trout, fathead minnow and bluegill sunfish). In local models, both binary and ternary classification models were obtained for each of the three fish species. For the global models, data of all the three fish species were used together. The predictive accuracy of both local and global models was around 0.8 for test sets. Meanwhile, data on sheepshead minnow were used as external validation set. For the best local model (model **2**) the predictive accuracy was 0.875 for sheepshead minnow, while for the best global model (model **14**) the predictive accuracy was 0.872 for sheepshead minnow. The FN compounds in model **2** and model **14** were 18 and 10, respectively. Hence, model **14** was the best model, and can predict other fish species' toxicity. Furthermore, information gain and ChemoTyper methods were used to identify toxic substructures which might significantly correlate with chemical aquatic toxicity. This study provided critical tools for early evaluation of chemical aquatic toxicity in environmental hazard assessment.

Toxicology Research Accepted Manuscript

1 Introduction

In the past decades humanity has witnessed unprecedented growth and prosperity, however, this growth has been accompanied by environmental pollution and natural resource depletion. The release of chemicals continues to affect all aspects of natural resources including the atmosphere, water, soil and wildlife. Chemicals are an integral part of daily life in today's world. Therefore, it is very urgent to assess the potential risk of chemicals to our health and environment. For water pollution, fish is usually used as the model species to evaluate chemical aquatic toxicity. Among various fishes, fathead minnow (*Pimephales promelas*) is the most widely used in North America[1]. For example, the U.S. EPA uses fathead minnow toxicity test as one of the management models. Experimental determination of the acute fish toxicity usually contains animal test, resulting in LC_{50} (lethal concentration 50%) values[2]. However, there is an increasing need in reducing or replacing animal test for regulatory purposes. Both *in vitro* assays and *in silico* methods are hence developed as non-animal alternatives[2-6].

To date, a large number of computational methods have been committed to the development of reliable prediction models of toxicity on fathead minnow. Those models can be divided into three categories: local models based on mode of action (MOA)[7-10], local models based on specific functional groups[11-15] and global models[16-27]. Among them, local models based on specific functional groups only could be used for assessment of specific compounds. If a compound contains multiple functional groups, it is difficult to be classified according to functional groups. For the local models based on mode of action, there is a need to know the mode of action of the compounds before assessment. However, such information is very difficult to be obtained, often requires expertise or clues provided by experiments. Therefore, local models have some limitations in application. The most practical model is global model, which need not consider the information of functional group or mode of action in chemicals. However, compared with the local ones, the global models usually apply different

Toxicology Research Accepted Manuscript

toxicity data in model building. Hence, global models are more difficult to be developed with high accuracy than local ones.

In practice, as the first step of hazard risk assessment, we only need to know a compound is toxic or non-toxic, highly toxic or slightly toxic, rather than its exact toxicity value. Chemical category approach is hence suitable for that purpose. The U.S. EPA has defined chemical toxicity categories of aquatic organisms. As shown in Table 1, chemical aquatic toxicity can be divided into five categories, i.e. very highly toxic, highly toxic, moderately toxic, slightly toxic, and practically nontoxic. Since 2001, the OECD Environmental Outlooks have used icons of red, yellow and green traffic lights to highlight the magnitude and direction of pressures on environment and environmental conditions[28]. So we also can use the traffic lights to describe the category of toxicity. Red light indicates the compound is very highly toxic or highly toxic, and yellow light means it is moderately toxic. In the same way, green light means the compound is slightly toxic or non-toxic.

Previously published models only considered chemical aquatic toxicity on one fish species, such as fathead minnow, which seriously limits the application of those models on other fish species. Hence, in this study, we aimed to build both local and global models for the prediction of chemical aquatic toxicity on various fish species. Specifically, three fish species, i.e. fathead minnow (FHM), bluegill sunfish (BS, *Lepomis macrochirus*) and rainbow trout (RT, *Oncorhynchus mykiss*), were used. In local models, both binary classification and ternary classification model were constructed. Only one fish species was used in one local model. The results demonstrated that binary classification models were better than the ternary ones. Hence, binary classification was used for further study. Sheepshead minnow (*Cyprinodon variegatus variegatus*) were used in external validation set to verify the performance of each local model. In global models, three fish species (BS, FHM and RT) were used together. 10-fold cross validation and external validation sets verified that the performance of the

Toxicology Research Accepted Manuscript

global models was better than that of local models. The predictive models built here would be very useful for assessment of chemical aquatic toxicity.

2 Materials and methods

2.1 Data collection and preparation

All chemical acute aquatic toxicity data were obtained from the U.S. EPA ECOTOX database (released at June 14, 2013) [29]. Only the data tested in 96 hours on fresh water fish with LC_{50} values were chosen for this study. In particular, data on warm water fish FHM and BS as well as cold water fish RT were used to develop the predictive models, while data on sheepshead minnow was selected as external set to validate the models.

Chemical 2D structures were obtained from the U.S. EPA Aggregated Computational Toxicology Resource (ACToR) database[30] by CAS Registry Number (CASRN) using in-house scripts. All structures were double checked with the PubChem database[31]. The data were prepared in following steps. At first, compounds containing inorganic and organometallic, salts, and mixtures were removed. Next, based on the U.S. EPA guideline of toxicity categories (Table 1), the compounds were classified into three levels (high, moderate and low toxicity or red-yellow-green) and two levels (red/yellow-green), separately. If one compound had several data points for the same fish species, we followed the rule of "to reduce false negative (FN)", which means the most toxic data is selected if the data points belong to the same category. Otherwise we double check the conflicted compound in other database such as IUPAC, and decide to keep it or not. Finally the data set was randomly divided into training set and test set with the ratio of 80% vs. 20%. For the external validation sets, duplicated compounds with

Toxicology Research Accepted Manuscript

the training and test sets were not removed, because one compound might have different toxicity on different fish species.

2.2 Molecular representation

PaDEL-Descriptor[32] was used to calculate the descriptors (1D and 2D descriptors) of all the compounds. Descriptors with more than 95% zero value or zero variance were removed. The remaining descriptors were used for further chemical feature reduction. F-score and Pearson correlation coefficient were used together to select the descriptors, where F-score calculated the correlation between endpoint and descriptors, and Pearson correlation coefficient calculated the correlation between descriptors. The cutoff value of Pearson correlation coefficient between each descriptor was set to 0.8. Substructure fingerprints (FP4) were also calculated using PaDEL-Descriptor, the detailed description of FP4 can be found on the original literature[32].

2.3 Model building methods

Both local and global models were constructed. The whole workflow was shown in Supporting Information (SI, SI-1 Figure S1). In local models, binary and ternary classification models were built separately. In ternary classification models, red, yellow and green light data belong to different classes, while in binary classification models, the red light and yellow light data were combined as one class, and green light data as the other. The method to build ternary classification models was described in our previous paper[33]. In global models, only binary classification models were built based on data from all the three fish species (BS, FHM and RT). Five machine learning methods, including random forest (RF), naïve Bayes (NB), k-nearest neighbors (k-NN), C4.5 decision tree (C4.5 DT), and support vector machine (SVM), were used to build the models. These methods were performed in Orange Canvas 2.7 (available free of charge at web site: http://www.ailab.si/orange/).

Toxicology Research Accepted Manuscript

Support Vector Machine (SVM). SVM, originally developed by Vapnik[34], is a supervised machine learning method aiming at minimizing the structural risk under the frame of VC theory. SVM can be used for classification and regression analysis. In this study, the Gaussian radial basis function (RBF) kernel was used. RBF is a popular kernel function used in SVM classification. And the parameters C and ϒ for RBF kernel were tuned on the training set by 10-fold cross validation.

k-Nearest Neighbors (k-NN). The k-NN classification method is based on closest training examples in a feature space[35]. In k-NN the value of k must be odd number. Here several k values were used, and k = 9 was the best.

Naïve Bayes (NB). NB[36] is one of the most used methods for classification. In principle NB generates the posterior probabilities. When using NB the most important thing is to select the descriptors, which should be independent each other. The default parameters were used in this study.

C4.5 Decision Tree (C4.5 DT). C4.5 DT is one of the oldest classification methods. It was defined as the possible decision tree by means of a hill climbing search based on the statistical property. The detailed description of C4.5 can be found in the original literature[37]. The parameters used here were also the default values.

Random Forest (RF). RF is a combination of tree predictors, in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest[38]. RF models consist of an ensemble of decision trees, each obtained by splitting object collections until terminal nodes contain only objects of the same class. The output class depends on the mode of classes output by individual trees. In this study the number of trees in forest was set as 100, and stop splitting nodes with 5 or fewer instances. Other parameters were default.

2.4 Evaluation of model performance

Toxicology Research Accepted Manuscript

Models were validated with 10-fold cross validation[33] and external validation set[39]. 10-fold cross validation was used to evaluate the robustness of the models, and external validation set was used to assess the predictive accuracy of the models.

In binary classification, all models were evaluated based on numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The sensitivity, specificity, and the overall predictive accuracy (Q) of the models were calculated as following:

$$\mathit{sensitivit}\ y(SE)=\frac{TP}{TP+FN} \qquad (1)$$

$$\mathit{specificit}\ y(SP)=\frac{TN}{TN+FP} \qquad (2)$$

$$\mathit{predictive}\quad \mathit{accuracy}\ (Q)=\frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

In ternary classification models, the overall predictive accuracy was calculated as following:

$$\mathit{predictive}\quad \mathit{accuracy}\ (Q)=\frac{N_{0-0}+N_{1-1}+N_{2-2}}{N_{Total}} \qquad (4)$$

Herein, $N_{0\text{-}0}$ means non-toxicity predicted as non-toxicity, $N_{1\text{-}1}$ means moderate-toxicity predicted as moderate-toxicity, $N_{2\text{-}2}$ means high-toxicity predicted as high-toxicity, N_{Total} means the total number in the data set.

2.5 Analysis of toxic substructures or substructural alerts

The toxic substructures are defined as molecular functional groups that make compounds toxic, which are hence used as substructual alerts. Substructural alerts were derived directly from mechanistic knowledge[40], so they are important tools to predict toxicity. Information gain (IG) method was used to search substructural fragments. The detailed method was described in our previous papers[41-43]. Another method named ChemoTyper[44] was also used to identify toxic substructures.

Toxicology Research Accepted Manuscript

ChemoTyper, released on November 13, 2013, was developed under a contract from the U.S. FDA, Center for Food Safety and Applied Nutrition (CFSAN), Office of Food Additive Safety.

3 Results

3.1 Data collection and analysis

The total data points in the U.S. EPA ECOTOX database were more than 680,000. After database standardization, the remaining data points were 6422 with 1906 unique compounds. These data were separated into training sets and test sets randomly. As shown in Table 2, the numbers of compounds in training set and test set of local models were 814 and 181 for FHM, 738 and 162 for BS, and 741 and 162 for RT, separately. The data points in these three fish species were combined to develop the global models. The numbers of unique compounds in training set and test set of the global models were 1337 and 320, respectively. The distribution of compounds in different toxic classes of training sets and test sets were balanced.

When building models, the most important thing is the data quality. Hence, Tanimoto coefficient was used to calculate the similarity of compounds in the data sets. The heat map of Tanimoto similarity index and the average Tanimoto similarity index can be seen in SI-1 Figure S2 and Table S1. The average Tanimoto similarity indexes were 0.350, 0.336 and 0.351 for FHM, BS and RT data sets, respectively. These results indicated the good generalization ability of the models. Applicability domain was also an important factor when building a model. In this study, the chemical space distributed by these data sets was defined by molecular weight (MW) and ALogP. As illustrated in Figure 1, for each fish species, data in the training set and test set distributed in the same chemical space, which indicated that these models had reasonable applicability domain.

Toxicology Research Accepted Manuscript

3.2 Molecular descriptors and fingerprints

Totally 770 different molecular descriptors were calculated by software PaDEL-Descriptor. After feature reduction and variable selection by F-score and Pearson correlation coefficient methods, four descriptors with the highest scores commonly occurred in BS, FHM and RT data sets. They were CrippenLogP (Crippen's LogP), ATSm1 (ATS autocorrelation descriptor, weighted by scaled atomic mass), SwHBa (Sum of E-States for weak Hydrogen Bond acceptors), and ETA_dEpsilon_D (a measure of contribution of hydrogen bond donor atoms). The detailed descriptions of these descriptors can be found in the original literature [45-50]. Hence, these four descriptors were used in model building.

In order to identify toxic substances using information gain method, FP4 fingerprint was also calculated for each molecule with software PaDEL-Descriptor.

3.3 Local model building and evaluation

In local models, both binary and ternary classification models were built by five machine learning methods using molecular descriptors or fingerprints to represent the molecules, which led to 20 models for each fish species, including 10 binary classification models and 10 ternary classification models. 10-fold cross validation method was used to evaluate the model robustness. The performance of these models was summarized in SI-2. When using 4 physicochemical descriptors to build models, RF and k-NN algorithms led to better results (shown in SI-2). For example, for FHM, the 10-fold cross validation results demonstrated that the areas under the receiver operating characteristic (AUC) of 4D-RF and 4D-k-NN models were 0.836 and 0.819, respectively. When using FP4 to build models, RF and SVM algorithms got better robustness. For example, for BS, the 10-flod cross validation results showed that the AUC values of FP4-RF and FP4-SVM models were 0.807 and 0.819, respectively. Hence, in the follow study, these four types of models were used for further validation in each data set. Test set and external validation set were used to assess the predictive accuracy of the models. For each

Toxicology Research Accepted Manuscript

data set, the performance of these four types of models was shown in SI-1 Table S2 and Table S3.

Binary classification and ternary classification models were constructed for each single data set (BS, FHM and RT). In each model, the performance for test sets was shown in Figure 2, in which green color means the number of correctly predicted compounds, while yellow and red colors indicate the number of misclassified compounds. The red color was the most priority one (FN compounds). In local models, models **1-12** were binary classification models and models **a-l** were ternary classification models. As shown in Figure 2A-C, for each data set (BS, FHM and RT) the green colored column in binary classification models were higher than that in ternary classification models, which means the performance of binary classification models were better than that of ternary ones. This also indicated that current machine learning methods might not be suitable for ternary classification models. Hence, in the following study only binary classification models were constructed.

To further evaluate the performance of these binary classification models (model ID **1-12**), sheepshead minnow data set was used as external validation set, which contained 367 compounds, including 136 high toxic compounds, 103 moderate toxic compounds and 128 low toxic compounds (Table 2). The results were listed in Table 3. As shown in Table 3, the Q values in models **1-4** ranged from 0.798 to 0.875, in models **5-8** ranged from 0.668 to 0.779, and in models **9-12** from 0.768 to 0.839, which demonstrated that the models built from BS and RT data sets were better than those from FHM data set. The distribution of toxic and non-toxic compounds in BS and RT data sets were more balanced than in FHM data set, which might be one of the reasons to affect the performance of the models. Among these local models, according to Q value, model **2** yielded the best prediction result (Q = 0.875 and SE = 0.925). In model **2** 4 physicochemical descriptors and k-NN algorithms were used to get higher Q value, but FN value in this model was not the best one. The FN values in models **2-4** were 18, 10, and 13, respectively (Table 3). In models **3** and **4** FP4 was used to describe molecules. The model with the lowest FN value was models **9** and **11**, both with the value of 4. But the Q values in

Toxicology Research Accepted Manuscript

these two models were 0.768 and 0.782, which were not good enough.

Comparing the validation results in each data set, we found that models from BS data set had higher Q values, and models from RT data set had lower FN values. In order to get one model with higher Q value and lower FN value, these three data sets (BS, FHM and RT) were combined to build global models.

3.4 Global model building and evaluation

The generalization ability of a model decides the reliability of the model. The data set to build global models contained both warm water fish (BS and FHM) and cold water fish (RT). Hence, the application domain of the global models was wider. Test set and external validation set were used to verify the robust and applicability of the global models. The performance of test set results was summarized in SI-1 Table S2. The performance of external validation set was listed in Table 3 as models **13**-**16**. In these 4 models, the SE values ranged from 0.824 to 0.958, and model **14** yielded the best performance. The Q value and FN value in model **14** were 0.872 and 10, respectively (Table 3). According to the Q value and FN value, model **14** was the best one. The performance of validation set in global models were shown in Figure 2D.

3.5 Comparison with ECOSAR

The Ecological Structure Activity Relationships (ECOSAR) program is a computerized predictive system that estimates aquatic toxicity[51]. To compare the accuracy of our models, ECOSAR was used to predict the chemical aquatic toxicity of our external validation set. As shown in Table 3, ECOSAR led to Q = 0.801, SE = 0.854, SP = 0.703 and FN = 35. The predictive accuracy and FN values were inferior to ours, which indicated that our models are better than ECOSAR.

3.6 Identification of toxic substructures

Toxicology Research Accepted Manuscript

To investigate structural differences between toxic and nontoxic compounds, IG method was performed to identify toxic substructures in FHM, BS and RT data sets based on FP4 fingerprint. According to the values of IG, p(positive) and p(negative), we obtained 25 substructures in FHM data set, 28 substructures in BS data set and 23 substructures in RT data set. The threshold of p(positive)/p(negative) was more than 3. Among these substructures, 6 presented in all the three data sets (see Table 4). They were diaryl ether, quaternary aliph ammonium, chloroalkene, sulfenic derivatives, and phosphoric acid derivatives. Some substructures were found more toxic in FHM (see SI-1 Table S4), including alkene, arylchloride and aryliodide.

Meanwhile, ChemoTyper was also used to find toxic substructures. And the substructures quaternary aliph ammonium and phosphoric acid derivatives in Table 4 were also identified by ChemoTyper.

In Table 4, if an oxygen or sulfur atom bridged two benzene rings, the compound may have para-hydroxyl or halogen just like bisphenol A (BPA), a toxic component in some plastics. Phosphoric acid derivatives are phosphorus fragments. Most of phosphonic acid derivatives, phosphoric trimester and phosphoric acid derivatives are pesticides. They can inhibit the activity of cholinesterase, resulting in the accumulation of acetylcholine, which was the neurotransmitter of cholinergic receptor, and then the cholinergic nerve system function will be disordered. They can also effect on cholinergic receptor directly, leading the next neuron or effector to excessive excitement or inhibition[52]. These substructure alerts were very important in ecological risk assessment and can help us to find toxic compounds.

4 Discussion

4.1 Data set analysis

Toxicology Research Accepted Manuscript

Chemical diversity is a key factor that influences the prediction capability of models. A large data set containing 6422 data points with aquatic toxicity was constructed and used traffic lights (red, yellow and green) to make the data set visible. High toxic, moderated toxic and low toxic molecules were colored in red, yellow and green, respectively. Two warm water fish (FHM and BS) and cold water fish (RT) data were used for model building; meanwhile sheepshead minnow was selected as external validation set. Previously, Martin et al[9] used 924 chemicals to build linear discriminant and random forest models, Singh et al[53] used 505 fish toxicity data to build multispecies models. Most of these models were based on relatively small data sets or only one fish species. Compared with those published models, our models were built on the basis of more data points, more compounds and more fish species with different living environment. These models hence have wider domain of applicability.

In order to explore data distribution in the chemical structural space, the radar chart[43] analysis was performed to explore the applicability domain of the global models. Five physicochemical descriptors, including these four selected descriptors (CrippenLogP, ATSm1, SwHBa, and ETA_dEpsilon_D) and molecular weight, were used in the radar chart. As shown in SI-1 Figure S3, the CrippenLogP ranged from -13.270 to 15.399; ATSm1 ranged from 3.360 to 342.690; SwHBa ranged from -12.659 to 45.429; the ETA_dEpsilon_D ranged from 0 to 0.3; and the molecular weights ranged from 44.053 to 792.848 for global model data set. These data indicated that our models could be used in a large chemical space (All the data sets were provided in SI-3).

4.2 Analysis of different models

Four physicochemical descriptors and FP4 fingerprints combine with five machine learning method were used to build local and global classification models. The data sets in local models only contained one fish species (BS, FHM or RT) and in global models contained all the three species.

Toxicology Research Accepted Manuscript

In local models, both binary classification and ternary classification models were built. As shown in Figure 2, binary classification models performed better than ternary ones, so binary models were studied further. For binary models, 10-flod cross validation verified that four models, i.e. 4D-RF, 4D-k-NN, FP4-RF and FP4-SVM, were better than the others. After external validation, model **2** from BS data set was identified as the best one in terms of Q values. Though the Q value of model **2** was pretty high, its FN value was not as good as models **9** and **11**, which were built from RT data set.

Among global models, as shown in Table 3, model **14** was the best one. The Q, SE and SP values of model **14** were 0.872, 0.958 and 0.711, respectively. Compared with model **2**, the best model in local models, their Q values were quite similar, 0.875 for model **2** vs. 0.872 for model **14**. However, their FN values were pretty different, 18 in model **2** vs. 10 in model **14**. This indicated that model **14** was better than model **2**. Hence, model **14** was the best one and can predict toxicity of other fish species.

Model **14** was built from 4 physicochemical descriptors and k-NN algorithm. k-NN algorithm was a non-parametric method, the input consists of the k closest training examples in the feature space, and the output depends on whether k-NN is used for classification. In model **14**, the 4 physicochemical descriptors were used to calculate the Euclidean distance from the training set for each object. The representation of receiver operating characteristics (ROC) plots also shown that model **14** was a little better than the others from the aspect of AUC. As shown in Figure S4 of SI-1, the AUC value of model **14** was 0.887, a little higher than the others. Hence, when using physicochemical descriptors to represent molecules, k-NN can get better modeling results. And in this study model **14** was the best one.

4.3 Relevance of selected descriptors to aquatic toxicity

Toxicology Research Accepted Manuscript

The selection of molecular descriptors was very important for model building. To increase the interpretability of models, the relationships between eco-toxicity data of the 1337 chemicals and 4 selected physicochemical descriptors, including CrippenLogP, ATSm1, SwHBa and ETA_dEpsilon_D, are present in Figure 3.

Among these 4 descriptors, CrippenLogP is related to hydrophobicity, which distributed between -13.270 and 15.399, with mean of 2.739. The mean values of CrippenLogP were 3.553 and 1.751 in toxic and non-toxic molecules, respectively, which suggests that chemical aquatic toxicity increases as the rise of hydrophobicity. The p-value of CrippenLogP was 2.00e^{-55}, indicating that distributions of toxic and non-toxic are significantly different. As shown in Figure 3A, toxic molecules tend to be more lipophilic then non-toxic molecules. ATSm1 was an ATS autocorrelation descriptor, weighted by scaled atomic mass. ATSm1 values distributed between 3.360 and 342.691, with a mean of 28.340. The mean values of ATSm1 were 34.361 and 21.031 in toxic and non-toxic molecules, respectively. This indicated that toxic molecules were favorable for higher ATSm1 values. The p-value of ATSm1 was 9.21e^{-22}, indicating that distributions of toxic and non-toxic are significantly different (Figure 3B). Hydrogen binding ability is commonly represented by SwHBa and ETA_dEpsilon_D. SwHBa means the sum of E-States for weak hydrogen bond acceptors and ETA_dEpsilon_D means a measure of contribution of hydrogen bond donor atoms. The p-value between SwHBa and ETA_dEpsilon_D were 7.70e^{-23} and 1.76e^{-20}, respectively. This indicated that the distributions of toxic and non-toxic are significantly different. As shown in Figure 3C and D, higher SwHBa value and lower ETA_dEpsilon_D value tend to be toxic molecules.

Actually, chemical aquatic toxicity was a complex process which is related to many factors, such as chemical, biological and environmental conditions. Therefore, it is very difficult to explain their mechanisms using only one or few descriptors.

Toxicology Research Accepted Manuscript

4.4 Analysis of misclassified compounds

The global model using 4 physical descriptors and k-NN algorithm achieved excellent predictive ability (model **14**). Nevertheless, some compounds in external validation set were still predicted incorrectly. As shown in Table 3, there were 10 toxic compounds in validation set predicted as non-toxic (FN compounds) by the model. These 10 compounds were listed in Figure 4, including bromoform (75-25-2), butyl ether (142-96-1), phosmet (732-11-6), amitraz (33089-61-1), bufencarb (8065-36-9), benomyl (17804-35-2), mandipropamid (374726-62-2), pyrimethanil (53112-28-0), flufenpyr-ethyl (188489-07-8), indaziflam (950782-86-2), and flufenpyr-ethyl (188489-07-8). Among them, 8 ones were pesticide. The other two compounds, compared with other fish species, were more toxic, including bromoform and butyl ether. Bromoform was a brominated organic solvent, but in salt water containing up to 1.3 ppm (parts per million) [54]. Sheepshead minnow can live in fresh water and salt water, hence, bromoform had higher toxic in sheepshead minnow and our model misclassified it. Butyl ether was low toxic in BS but higher toxic in sheepshead minnow, just like alkene, arylchloride and aryliodide in FHM were higher toxic (shown in SI-1 Table S4). Among the 8 pesticides, phosmet, amitraz and bufencarb were insecticides, benomyl, mandipropamid and pyrimethanil were fungicides, indaziflam and flufenpyr-ethyl were herbicides. Model **14** used 4 physicochemical descriptors to describe molecules. As we discussed above the mean value of CrippenLogP were 3.553 in toxic. However, 6 of the 8 pesticides had logP value less than 3.553, including phosmet, amitraz, bufencarb, benomyl, mandipropamid and indaziflam. The ATSm1 value in pyrimethanil was 1.608 which even lower than the mean value in non-toxic compounds. The SwHBa value in flufenpyr-ethyl was -1.454 which even lower than the mean value in non-toxic compounds. These are the reasons that our model misclassified them.

Toxicology Research Accepted Manuscript

5 Conclusions

In this study, a large data set containing 6422 data points with aquatic toxicity was constructed. Based on these data, both local and global models were built for the prediction of chemical aquatic toxicity with C4.5 DT, RF, NB, k-NN, and SVM algorithms. In local models, binary classification and ternary classification models were constructed separately. For local models, the ternary classification models can get better false negative (“2-0”) than binary ones, but the overall accuracy of the ternary classification models was not good enough. Sheepshead minnow data set were used to validate all the models, the results showed that global models can predict toxicity of different fish species, which indicates our models have wider domain of applicability. The best local and global models will be integrated as part of our web server admetSAR, which is freely available on http://lmmd.ecust.edu.cn/admetsar1/.

In the study, traffic lights were used to label aquatic toxicity categories, which could make the models vividly and easy to judge the toxicity by color. IG and ChemoTyper methods were used to identify some toxic substructures among toxic chemicals, which would be helpful for understanding mechanism of action and structural modification to reduce or remove the toxicity. This kind of classification strategy might be promoted to other toxicity endpoints such as acute oral toxicity, skin sensitivity, liver toxicity, and so on.

Toxicology Research Accepted Manuscript

This journal is © The Royal Society of Chemistry [year]

Acknowledgments

The authors would like to thank Dr. Chihae Yang of Altamira LLC for providing the software ChemoTyper and helpful comments on the manuscript. This work was supported by the National Natural Science Foundation of China (Grant 81373329), the 863 Project (Grant 2012AA020308) and the Fundamental Research Funds for the Central Universities (Grant WY1113007).

Notes

[a]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. Tel: +86 021-64251052. Fax: +86 21-64251033. E-mail: ytang234@ecust.edu.cn

[b]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China.

† Electronic Supplementary Information (ESI) available: http://pubs.rsc.org/.

Toxicology Research Accepted Manuscript

References

1 Ankley, G. T., Villeneuve, D. L., The fathead minnow in aquatic toxicology: past, present and future. *Aquat. Toxicol.*, 2006, **78**, 91-102.

2 Schuurmann, G., Ebert, R. U., Kuhne, R., Quantitative read-across for predicting the acute fish toxicity of organic compounds. *Environ. Sci. Technol.*, 2011, **45**, 4616-4622.

3 Tunkel, J. M., K., Austin, C., Hickerson, A., Howard, P., Practical considerations on the use of predictive models for regulatory purposes. *Environ. Sci. Technol.*, 2005, **39**, 2218-2219.

4 Lammer, E., Carr, G. J., Wendler, K., Rawlings, J. M., Belanger, S. E., Braunbeck, T., Is the fish embryo toxicity test (FET) with the zebrafish (Danio rerio) a potential alternative for the fish acute toxicity test? Comparative biochemistry and physiology. *Toxicology & pharmacology : CBP,* 2009, **149**, 196-209.

5 Lienert, J. G., K., Escher, B. I., Screening method for ecotoxicological hazard assessment of 42 pharmaceuticals considering human metabolism and excretory routes. *Environ. Sci. Technol.*, 2007, **41**, 4471-4478.

6 Von der Ohe, P. C., Kühne, R., Ebert, R.-U., Altenburger, R., Liess, M., Schüürmann, G., Structural alerts- a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. *Chem. Res. Toxicol.*, 2005, **18**, 536- 555.

7 Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E., Drummond, R. A., Predicting modes of toxic action from chemical structure Acute toxicity in the fathead minnow (Pimephales promelas). *Environ. Toxicol. Chem.*, 1997, **16**, 948-967.

8 Yuan, H., Wang, Y. Y., Cheng, Y. Y., Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. *J. Mol. Graph. Model.*, 2007, **26**, 327-335.

9 Martin, T. M., Grulke, C. M., Young, D. M., Russom, C. L., Wang, N. Y., Jackson, C. R., Barron, M. G., Prediction of aquatic toxicity mode of action using linear discriminant and random Forest models. *J. Chem. Inf. Model.*, 2013, **53**, 2229-2239.

Toxicology Research Accepted Manuscript

10 Lyakurwa, F., Yang, X., Li, X., Qiao, X., Chen, J., Development and validation of theoretical linear solvation energy relationship models for toxicity prediction to fathead minnow (pimephales promelas). *Chemosphere*, 2014, **96**, 188-194.

11 Kulkarni, S. A., Raje, D. V., Chakrabarti, T., Quantitative structure-activity relationships based on functional and structural characteristics of organic compounds. *SAR QSAR Environ. Res.*, 2001, **12**, 565-591.

12 Toropov, A. A., Benfenati, E., QSAR modelling of aldehyde toxicity by means of optimisation of correlation weights of nearest neighbouring codes. *J. Mol. Struc-Theochem*, 2004, **676**, 165-169.

13 Smiesko, M., Benfenati, E., Predictive models for aquatic toxicity of aldehydes designed for various model chemistries. *J. Chem. Inf. Comput. Sci.,* 2004, **44**, 976-984.

14 Smiesko, M., Benfenati, E., Thermodynamic descriptors derived from density functional theory calculations in prediction of aquatic toxicity. *J. Chem. Inf. Model.*, 2005, **45**, 379-385.

15 Lyakurwa, F. S., Yang, X., Li, X., Qiao, X., and Chen, J., Development of in silico models for predicting LSER molecular parameters and for acute toxicity prediction to fathead minnow (Pimephales promelas). *Chemosphere*, 2014, **108**, 17-25.

16 Mazzatorta, P., Vracko, M., Jezierska, A., Benfenati, E., Modeling toxicity by using supervised kohonen neural networks. *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 485-492.

17 Niculescu, S. P., Atkinson, A., Hammond, G., Lewis, M., Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. *SAR QSAR Environ. Res.*, 2004, **15**, 293-309.

18 Gini, G., Craciun, M. V., König, C., Benfenati, E., Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1897-1902.

19 Casalegno, M., Benfenati, E., Sello, G., An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. *Chem. Res. Toxicol.*, 2005, **18**, 740-746.

20 Netzeva, T. I., Aptula, A. O., Benfenati, E., Cronin, M. T. D., Gini, G., Lessigiarska, I., Maran, U., Vracko, M., Schüürmann, G., Description of the electronic structure of organic chemicals using

Toxicology Research Accepted Manuscript

semiempirical and ab initio methods for development of toxicological QSARs. *J. Chem. Inf. Model.*, 2005, **45**, 106-114.

21 Papa, E., Villa, F., Gramatica, P., Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in pimephales promelas (Fathead Minnow). *J. Chem. Inf. Model.*, 2005, **45**, 1256-1266.

22 He, L., Jurs, P. C., Assessing the reliability of a QSAR model's predictions. *J. Mol. Graph. Model.*, 2005, **23**, 503-523.

23 Pavan, M., Netzeva, T. I., Worth, A. P., Validation of a QSAR model for acute toxicity. *SAR QSAR Environ. Res.,* 2006, **17**, 147-171.

24 Amini, A., Muggleton, S. H., Lodhi, H., Sternberg, M. J. E., A novel logic-based approach for quantitative toxicology prediction. *J. Chem. Inf. Model.*, 2007, **47**, 998-1006

25 Hewitt, M., Cronin, M. T. D., Madden, J. C., Rowe, P. H., Johnson, C., Obi, A., Enoch, S. J., Consensus QSAR models do the benefits outweigh the complexity? *J. Chem. Inf. Model.*, 2007, **47**, 1460-1468.

26 Colombo, A., Benfenati, E., Karelson, M., Maran, U., The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity. *Chemosphere*, 2008, **72**, 772-780.

27 Maunz, A., Helma, C., Prediction of chemical toxicity with local support vector regression and activity-specific kernels. *SAR QSAR Environ. Res.*, 2008, **19**, 413-431.

28 The OECD Environmental Outlooks. http://www.oecd.org/environment/outlooks.htm

29 ECOTOX Database, http://cfpub.epa.gov/ecotox/data_download.cfm?sub=main

30 Judson, R., Richard, A., Dix, D., Houck, K., Elloumi, F., Martin, M., Cathey, T., Transue, T. R., Spencer, R., Wolf, M., ACToR--aggregated computational toxicology resource. *Toxicol. Appl. Pharmacol.*, 2008, **233**, 7-13.

31 Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 2009, **37**, 623-633.

Toxicology Research Accepted Manuscript

32 Yap, C. W., PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Coput. Chem.*, 2011, **32**, 1466-1474.

33 Li, X., Chen, L., Cheng, F., Wu, Z., Bian, H., Xu, C., Li, W., Liu, G., Shen, X., Tang, Y., In silico prediction of chemical acute oral toxicity using multi-classification methods. *J. Chem. Inf. Model.*, 2014, **54**, 1061-1069.

34 Cortes, C., Vapnik, V., Support-vector networks. *Mach. Learn.*, 1995, 20, 273-297.

35 Cover, T. M., Hart, P., Nearest neighbor pattern classification. *IEEE. Trans. Inf. Theory*, 1967, **13**, 21-27.

36 Watson, P., Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model*., 2008, **48,** 166-178.

37 Quinlan, J. R., C4.5: Programs for Machine Learning; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1993

38 Breiman, L., Random forests. *Mach. Learn.*, 2001, **45**, 5-32.

39 Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., Tropsha, A., Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.*, 2012, **52**, 2570-2578.

40 Benigni, R., Bossa, C., Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutat. Res-Rev. Mutat.*, 2008, **659**, 248-261.

41 Shen, J., Cheng, F., Xu, Y., Li, W., Tang, Y., Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.*, 2010, **50**, 1034-1041.

42 Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., Tang, Y., In silico Prediction of Chemical Ames Mutagenicity. *J. Chem. Inf. Model.*, 2012, **52**, 2840-2847.

43 Cheng, F., Shen, J., Yu, Y., Li, W., Liu, G., Lee, P. W., Tang, Y., In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere*, 2011, **82**, 1636-1643.

44 ChemoTyper Community Website, https://chemotyper.org/

Toxicology Research Accepted Manuscript

45 Hall, L. H., Kier, L. B., Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1039-1045

46 Liu, R., Sun, H., So, S. S., Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1623-1632.

47 Gramatica, P., Corradi, M., Consonni, V., Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere*, 2000, **41**, 763-777.

48 Roy, K., Ghosh, G., QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes. *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 559-567.

49 Roy, K., Das, R. N., On some novel extended topochemical atom (ETA) parameters for effective encoding of chemical information and modeling of fundamental physicochemical properties. *SAR QSAR Environ. Res.*, 2011, **22**, 451-472

50 Moreau, G., Broto, P., The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau Journal de Chimie*, 1980, 359-360

51 Ecological Structure Activity Relationships (ECOSAR), http://www.epa.gov/oppt/newchems/tools/21ecosar.htm

52 Casida, J. E., Quistad, G. B., Organophosphate toxicology: safety aspects of nonacetylcholinesterase secondary targets. *Chem. Res. Toxicol.*, 2004, **17**, 983-998.

53 Singh, K. P., Gupta, S., Kumar, A., Mohan, D., Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. *Chem. Res. Toxicol.*, 2014, **27**, 741-753.

54 Beech, J. A., Diaz, R., Ordaz, C., Palomeque, B., Nitrates, chlorates and trihalomethanes in swimming pool water. *Am. J. Public Health*, 1980, **70**(1), 79-82.

Toxicology Research Accepted Manuscript

Table 1. Chemical toxicity categories in aquatic organisms

Toxicity category	Aquatic organisms acute concentration (PPM)	Ternary classification	Binary classification
very highly toxic	< 0.1	2 (red light)	1 (red light)
highly toxic	0.1 - 1	2 (red light)	1 (red light)
moderately toxic	> 1 - 10	1 (yellow light)	1 (red light)
slightly toxic	> 10 - 100	0 (green light)	0 (green light)
nontoxic	> 100	0 (green light)	0 (green light)

Toxicology Research Accepted Manuscript

Toxicology Research Accepted Manuscript

Table 2. The data points after database standardization

Toxicity	FHM	BS	RT	others species	sheepshead minnow	total
High	219	329	361	1592	136	2501
Moderate	253	233	251	954	103	1691
Low	523	338	291	1078	128	2230
Total	995	900	903	3624	367	6422

	Training set (80%)				Test set (20%)				Validation set
	FHM	BS	RT	Global	FHM	BS	RT	Global	sheepshead minnow
High	183	260	291	408	36	69	70	93	136
Moderate	202	197	205	325	51	36	46	93	103
Low	429	281	245	604	94	57	46	134	128
Total	814	738	741	1337	181	162	162	320	367

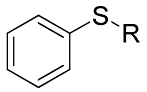
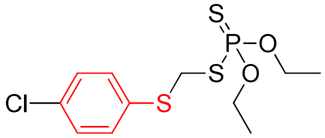
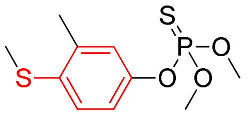
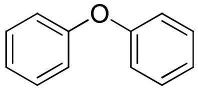
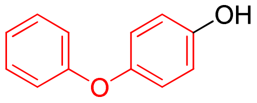
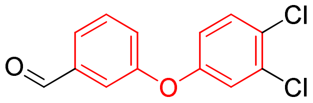
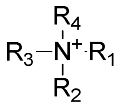
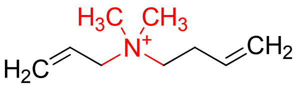
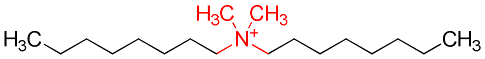
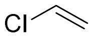
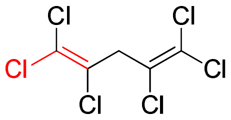
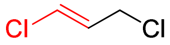
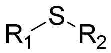
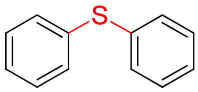
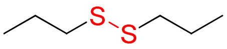
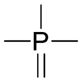
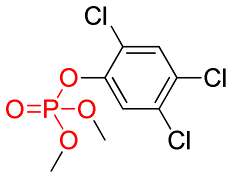
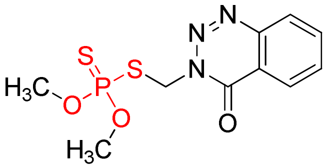
Table 3. The performance of models for external validation set[a]

			Methods	Model ID	Q	SE	SP	FP	FN
sheepshead minnow (367）	Local models	BS	4D-RF	**1**	0.815	0.887	0.680	41	27
			4D-k-NN	**2**	**0.875**	**0.925**	**0.781**	**28**	**18**
			FP4-RF	**3**	0.798	0.958	0.500	64	10
			FP4-SVM	**4**	0.826	0.946	0.602	51	13
		FHM	4D-RF	**5**	0.779	0.782	0.773	29	52
			4D-k-NN	**6**	0.752	0.824	0.617	49	42
			FP4-RF	**7**	0.668	0.615	0.766	30	92
			FP4-SVM	**8**	0.733	0.782	0.641	46	52
		RT	4D-RF	**9**	0.768	0.983	0.367	81	4
			4D-k-NN	**10**	**0.839**	**0.933**	**0.664**	**52**	**19**
			FP4-RF	**11**	**0.782**	**0.983**	**0.406**	**76**	**4**
			FP4-SVM	**12**	0.798	0.971	0.477	67	7
	Global models		4D-RF	**13**	0.798	0.824	0.750	32	42
			4D-k-NN	**14**	**0.872**	**0.958**	**0.711**	**37**	**10**
			FP4-RF	**15**	0.798	0.925	0.563	56	18
			FP4-SVM	**16**	0.831	0.950	0.609	50	12
	ECOSAR				0.801	0.854	0.703	38	35

[a] RF: random forest. k-NN: k-nearest neighbors. SVM: support vector machine. 4D: 4 physicochemical descriptors. FP4: substructure fingerprints.

Toxicology Research Accepted Manuscript

Table 4. The common substructure alerts identified in FHM, BS and RT data sets

Fragments	Examples	
 Alkylarylthioether	 786-19-6 LC_{50}=0.22 PPM	 55-38-9 LC_{50}=1.683 PPM
 Diarylether	 831-82-3 LC_{50}=4.95 PPM	 79124-76-8 LC_{50}=0.3 PPM
 Quaternary aliph ammonium	 26062-79-3 LC_{50}=0.22 PPM	 5538-94-3 LC_{50}=5.2 PPM
 Chloroalkene	 87-68-3 LC_{50}=0.09 PPM	 542-75-6 LC_{50}=0.239 PPM
 Sulfenic derivatives	 882-33-7 LC_{50}=0.11 PPM	 629-19-6 LC_{50}=2.62 PPM
 Phosphoric acid derivatives	 5598-52-7 LC_{50}=0.0021 PPM	 86-50-0 LC_{50}=0.0317 PPM

Toxicology Research Accepted Manuscript

This journal is © The Royal Society of Chemistry [year]

Figure Captions

Figure 1. Diversity analysis of chemicals in the training set and test set. A) BS data sets; B) FHM data sets; C) RT data sets; D) data sets used in global models.

Figure 2. The predictive results of test sets and validation sets with local models and global models. Model ID **1-12**: local binary classification models; **a-l**: local ternary classification models; **13-16**: global models. (A) BS test set with local models; (B) FHM test set with local models; (C) RT test set with local models; (D) external validation set with both local and global models.

Figure 3. Distributions of 4 physicochemical descriptors including CrippenLogP, ATSm1, SwHBa and ETA_dEpsilon_D for toxic and non-toxic classification. p-value: Student's t test was used to determine if two data sets are significantly different from each other.

Figure 4. FN compounds in external validation set.

Toxicology Research Accepted Manuscript

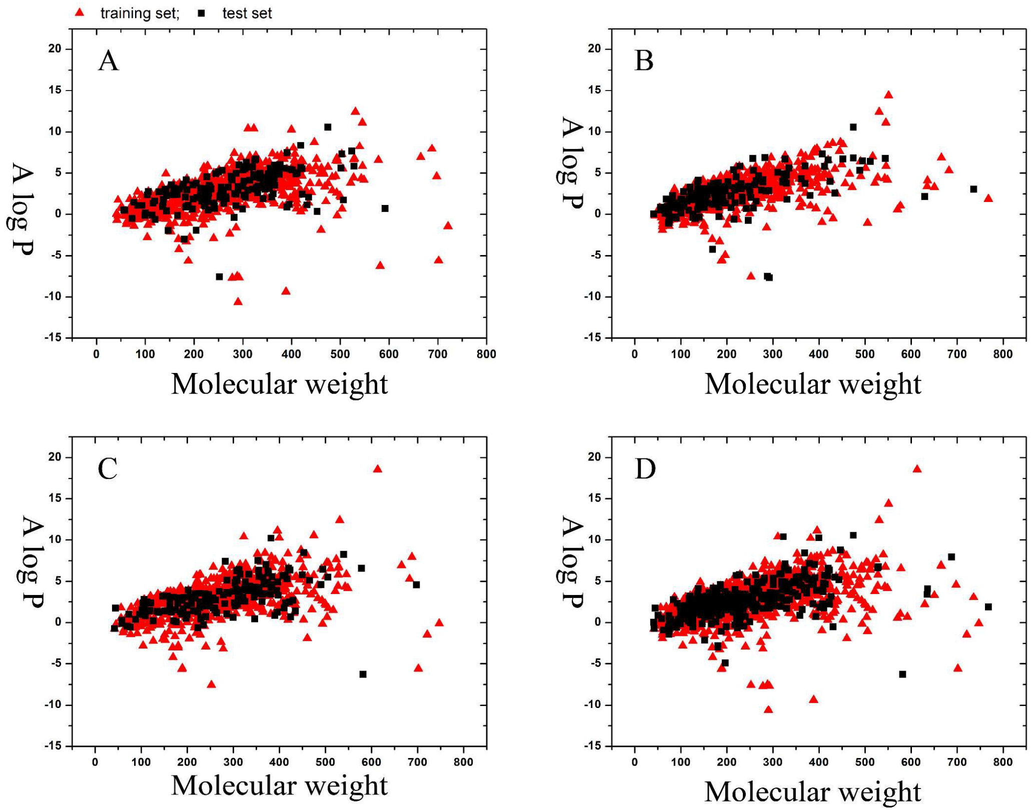


Figure 1. Diversity analysis of chemicals in the training set and test set. A) BS data sets; B) FHM data sets; C) RT data sets; D) data sets used in global models.

Toxicology Research Accepted Manuscript

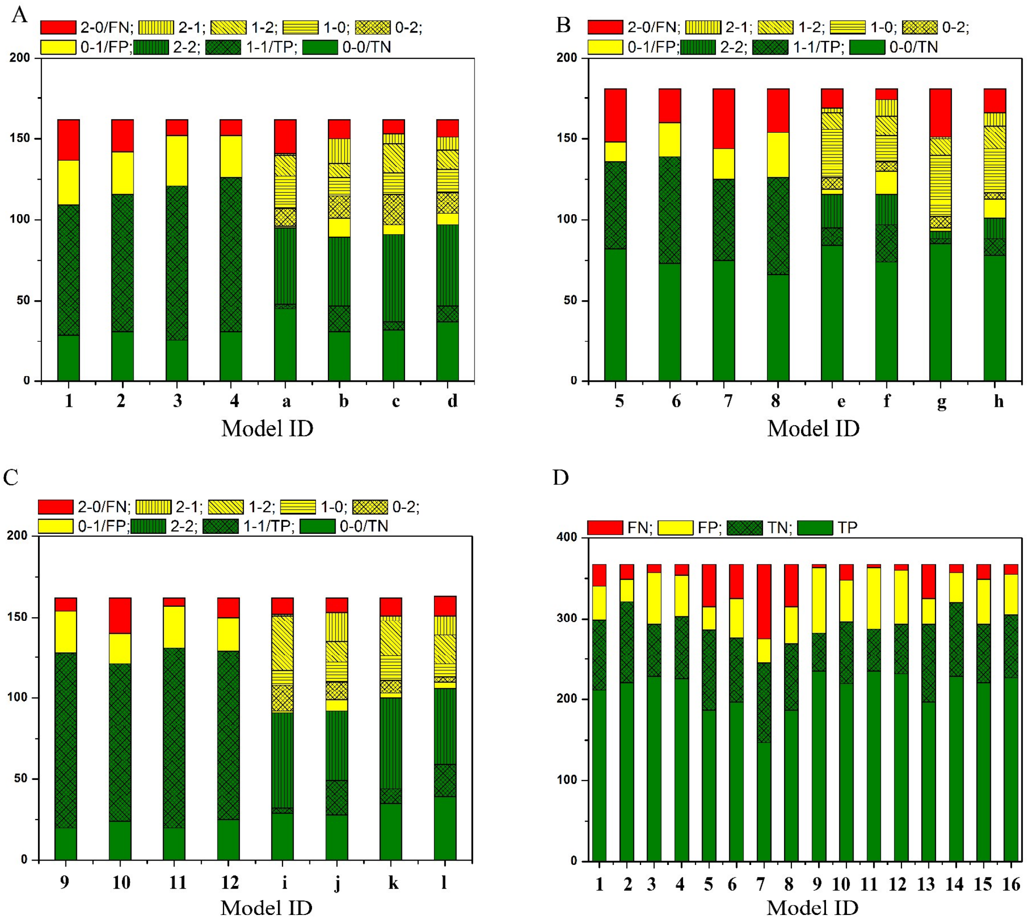


Figure 2. The predictive results of test sets and validation sets with local models and global models. Model ID **1-12**: local binary classification models; **a-l**: local ternary classification models; **13-16**: global models. (A) BS test set with local models; (B) FHM test set with local models; (C) RT test set with local models; (D) external validation set with both local and global models.

Toxicology Research Accepted Manuscript

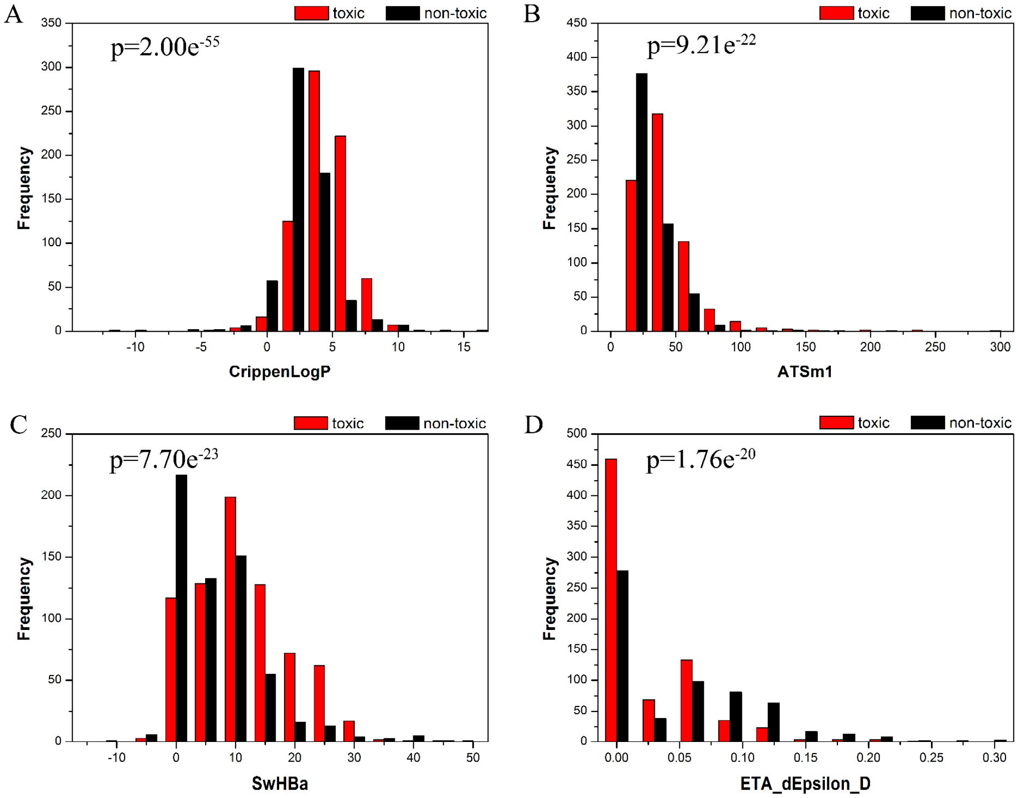


Figure 3. Distributions of 4 physicochemical descriptors including CrippenLogP, ATSm1, SwHBa and ETA_dEpsilon_D for toxic and non-toxic classification. p-value: Student's t test was used to determine if two data sets are significantly different from each other.

Toxicology Research Accepted Manuscript

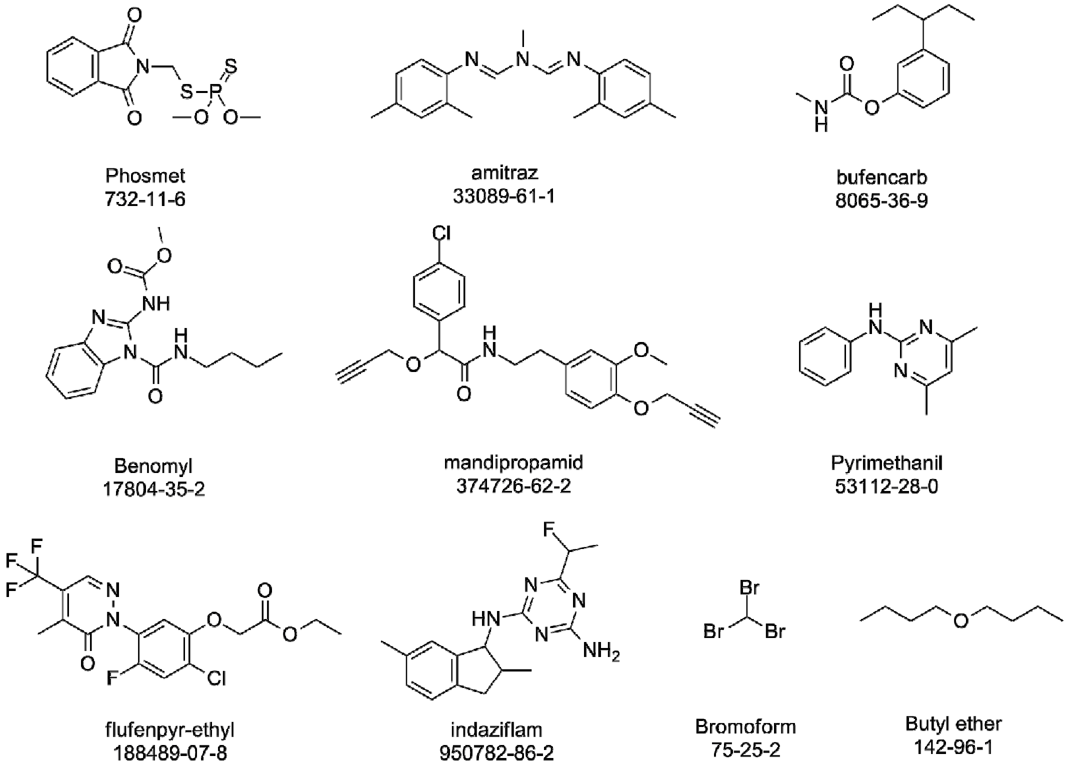


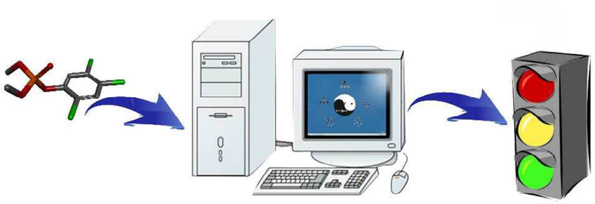
Figure 4. FN compounds in external validation set.

Toxicology Research Accepted Manuscript

Table of Contents Graphic (TOC)

In Silico Prediction of Chemical Aquatic Toxicity with Chemical Category Approaches and Structural Alerts

Lu Sun[a], Chen Zhang[a], Yingjie Chen[a], Shulin Zhuang[b], Weihua Li[a], Guixia Liu[a], Philip W. Lee[a], Yun Tang[a,]*



Toxicology Research Accepted Manuscript