

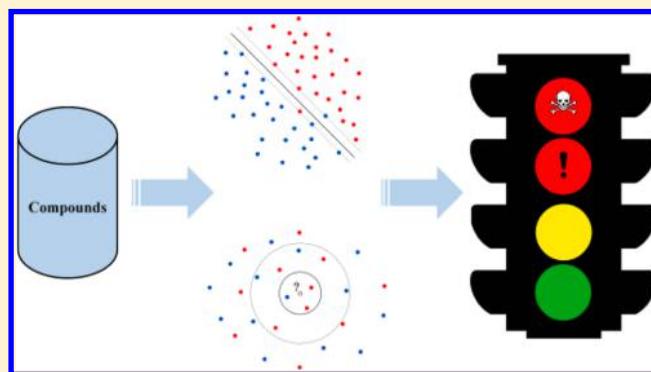
In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods

Xiao Li,^{†,‡,§} Lei Chen,^{†,§} Feixiong Cheng,[†] Zengrui Wu,[†] Hanping Bian,[†] Congying Xu,[†] Weihua Li,[†] Guixia Liu,[†] Xu Shen,[‡] and Yun Tang^{*,†}

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

[‡]Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

ABSTRACT: Chemical acute oral toxicity is an important end point in drug design and environmental risk assessment. However, it is difficult to determine by experiments, and *in silico* methods are hence developed as an alternative. In this study, a comprehensive data set containing 12 204 diverse compounds with median lethal dose (LD_{50}) was compiled. These chemicals were classified into four categories, namely categories I, II, III and IV, based on the criterion of the U.S. Environmental Protection Agency (EPA). Then several multiclassification models were developed using five machine learning methods, including support vector machine (SVM), C4.5 decision tree (C4.5), random forest (RF), κ -nearest neighbor (kNN), and naïve Bayes (NB) algorithms, along with MACCS and FP4 fingerprints. One-against-one (OAO) and binary tree (BT) strategies were employed for SVM multiclassification. Performances were measured by two external validation sets containing 1678 and 375 chemicals, separately. The overall accuracy of the MACCS-SVM_{OAO} model was 83.0% and 89.9% for external validation sets I and II, respectively, which showed reliable predictive accuracy for each class. In addition, some representative substructures responsible for acute oral toxicity were identified using information gain and substructure frequency analysis methods, which might be very helpful for further study to avoid the toxicity.



INTRODUCTION

There are numerous chemicals around us. Some are beneficial to our health, while others might be harmful. Therefore, it is very important to evaluate chemical safety as early as possible in order to reduce the harm of chemicals to our health. Some regulatory agencies, such as the U.S. Environmental Protection Agency (EPA) and European Chemicals Agency, play key roles in this area.

One of the most common toxicity end points is acute toxicity, which describes the adverse effects occurring immediately or in a short time after administration of a single dose of a chemical or multiple doses given within 24 h.¹ The accurate determination of chemical acute toxicity should be performed by *in vivo* experiments. However, this approach is very complicated, costly, and time-consuming, meanwhile it is not practical to screen a large number of compounds, especially for virtual molecules. Therefore, it is very urgent to develop alternative approaches including *in silico* methods to estimate chemical acute toxicity.

Over past decades, many quantitative structure–activity relationship (QSAR) models have been developed to predict acute rodent toxicity of organic chemicals.^{2–4} In those studies statistical methods such as multilinear regression (MLR) and neural network (NN) were used to build models based on

different data sets. A consensus modeling approach was also employed to yield QSAR models which showed superior to individual models.⁵ However, those models only used a relatively small number of compounds and had limited external predictive power. In fact, accurate prediction of acute toxicity is a great challenge because the mechanisms of action are quite diverse.

For precautionary labeling requirements, the U.S. EPA has established toxicity categories on the basis of median lethal dose (LD_{50}) or median lethal concentration (LC_{50} ; Table 1).⁶ For acute oral toxicity, there are four categories (categories I, II, III and IV) to indicate the level of toxicity. From category I to

Table 1. Toxicity Categories Based on the Study Results Defined by U.S. EPA⁶

acute toxicity	category I	category II	category III	category IV
oral (mg/kg)	≤ 50	> 50	> 500	> 5000
		≤ 500	≤ 5000	
dermal (mg/kg)	≤ 200	> 200	> 2000	> 5000
		≤ 2000	≤ 5000	

Received: January 18, 2014

Published: April 4, 2014

Table 2. Statistical Description of Training, Test, and External Validation Sets

	category I	category II	category III	category IV	total
training set	798	1943	4310	1051	8102
test set	225	463	1157	204	2049
external validation set I	92	342	1103	141	1678
external validation set II	57	93	183	42	375
total	1172	2841	6753	1438	12 204

category IV, the corresponding signal words are DANGER/POISON, WARNING, CAUTION, and None Required. DANGER/POISON means the chemical is highly toxic and it is fatal if swallowed, usually marked by a symbol of skull and crossbones. WARNING indicates the chemical is moderately toxic, while CAUTION means it is slightly toxic. Category IV is generally considered to be practically nontoxic. This semi-quantitative description might be more intuitive in toxicity estimation than simple numbers, by which it is hard for some people to understand how toxic it is. Actually the severity of compounds could give us enough information to deal with them properly.

Based on the U.S. EPA definition of toxicity category, in this study a multiclassification model was developed to predict chemical acute toxicity. For that purpose, at first an acute toxicity data set was compiled containing about 12 200 diverse substances with experimental LD₅₀ values for rat by oral exposure, which is believed to be the largest one published so far. Then multiclassification models were developed using several machine learning methods. External validation was used to assess the predictive power of the models. Finally, some privileged substructures responsible for acute oral toxicity were proposed by information gain and substructure analysis methods.

MATERIALS AND METHODS

Data Collection and Preparation. In this study, all the data were collected from three sources, totally containing 12 204 compounds with oral rat acute toxicity in LD₅₀ values.

The first data set was obtained from the admetSAR database in SMILES format.⁷ Compounds containing inorganics and organometallics, salts, and mixtures were removed. Then tautomers and molecules with molecular weight more than 800 were also eliminated. The final data set contained 10 151 compounds with measured LD₅₀ values. The compounds were then randomly divided into a training set and test set with a ratio of 4:1. To validate the prediction ability of the models, two external validation sets were extracted from the MDL Toxicity Database (version 2004.1)⁸ and Toxicity Estimation Software Tool (TEST version 4.1)⁹ program of the U.S. EPA, which were prepared with the same method as the training set. For external validation set I, the duplicate substances with the training set and test set were removed, and for validation set II, the duplicates with training set, test set, and validation set I were eliminated. After removal of duplicates, these two validation sets contained 1678 and 375 compounds, respectively.

According to the U.S. EPA definition of toxicity categories (Table 1), all the compounds were divided into the corresponding four categories with different levels of toxicity. The statistical description of the data set is shown in Table 2.

Calculation of Molecular Fingerprints. The substructure pattern recognition method developed in our group was used to code the molecular structure.¹⁰ Two frequently used

fingerprints, MACCS keys and FP4 fingerprint, were used for substructure dictionaries in this work. MACCS keys contain 166 substructure patterns from MDL Public Keys,¹¹ while FP4 contains 307 ones. The fingerprints were generated by PaDEL-Descriptor software.¹²

Model Building. Five machine learning methods, including support vector machine (SVM), C4.5 decision tree (C4.5), random forest (RF), κ -nearest neighbor (kNN), and naïve Bayes (NB) algorithms, were employed for model building. Except for the SVM algorithm, provided by the open source LIBSVM [LIBSVM3.16 package],¹³ the other four methods were performed in Orange 2.6 (version 2.6.1, freely available at <http://www.ailab.si/orange/>).¹⁴

C4.5 Decision Tree. C4.5 is an algorithm used to generate a decision tree developed by Quinlan.¹⁵ At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The attribute with the highest normalized information gain is chosen to make the decision. Then the algorithm recurses on the smaller sublists.

Random Forest. RF is an ensemble learning method developed by Breiman for classification and regression.¹⁶ In this approach, each tree in the ensemble is formed by first selection at random and a small group of input coordinates (features or variables hereafter) to split on at each node. Then, the best split is calculated based on these features in the training set. The tree is grown up to maximum size without pruning.

Naïve Bayes. NB is a simple probabilistic classifier based on the Bayes rule for the conditional probability. This statistical method allows the user to categorize instances in a data set based on the equal and independent contributions of their attributes.¹⁷ In this study, Orange with the default setting was used to perform the NB classification.

κ -Nearest Neighbor. The kNN algorithm is a nonparametric method for classifying objects based on closest training examples in the feature space.¹⁸ An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In this study, the nearness was measured by hamming distance metrics, and the parameter of $k = 5$ was used.

Support Vector Machine. SVM is a kernel-based tool for binary data classification and regression introduced by Vapnik and Cortes,¹⁹ which was widely used to solve binary classification problems.^{20–23} Each molecule was described as a binary string by a substructure pattern recognition method which worked as an eigenvector for SVM. After training, SVM could give a decision function for classification. In this study, the Gaussian radial basis function (RBF) kernel was used. And the parameters C and γ for RBF kernel were tuned on the training set by 5-fold cross validation.

Multiclass Models. C4.5, RF, kNN, and NB algorithms could be used to develop multiclass models directly. SVM is originally designed for binary classification. Several methods

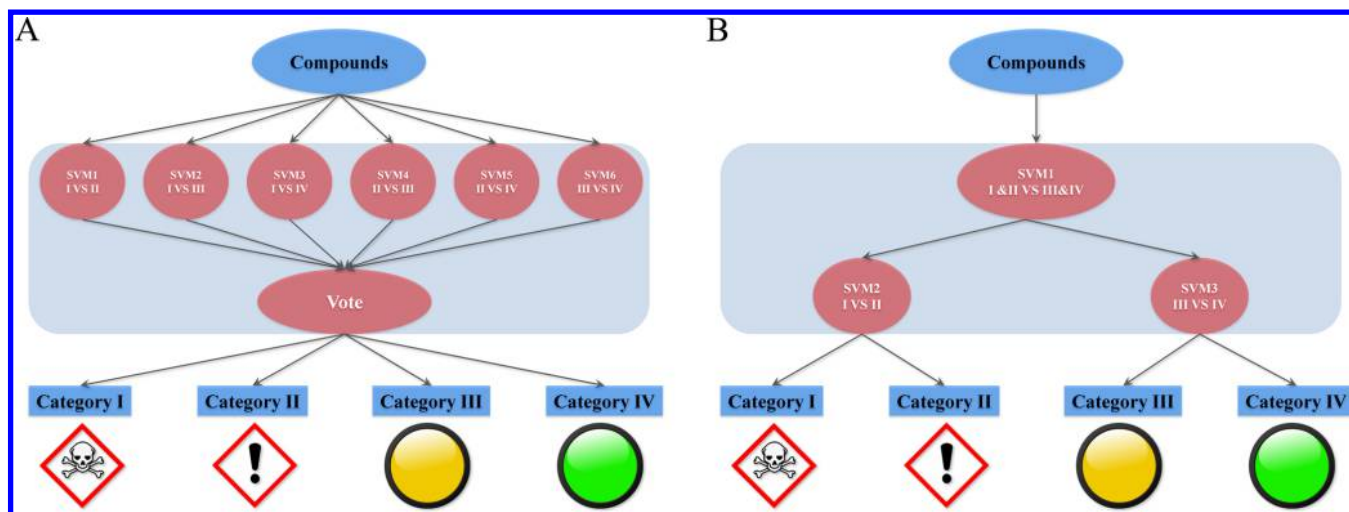


Figure 1. Workflow of one-against-one and binary tree methods for multiclass classification. (A) Workflow of one-against-one method. (B) Workflow of binary tree method.

have been proposed to effectively extend SVM for multiclass classification, such as one-against-all (OAA), one-against-one (OAO), and directed acyclic graph (DAG). Hsu and Lin gave a detailed comparison and concluded that OAO is a competitive approach,²⁴ which was hence implemented in LIBSVM for multiclass classification.¹³ The OAO approach constructs $N(N - 1)/2$ two-class classifiers, using all the binary pairwise combinations of the N classes. Each classifier is trained using the samples of the first class as positive examples and the samples of the second class as negative examples. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes (shown in Figure 1A). Binary tree (BT) is another method for SVM multiclass classification,^{25,26} which was introduced to reduce the number of binary classifiers and to achieve a fast decision. The BT method uses multiple SVM models arranged in a binary tree structure. A SVM in each node of the tree is trained using two of the classes. All samples in the node are assigned to the two subnodes derived from the previously selected classes by similarity. This step repeats at every node until each node contains only samples from one class (shown in Figure 1B). OAO and BT approaches have been used widely in previous research.^{27–31}

Assessment of Model Performance. All models were optimized by a diverse test set and validated by two external validation sets. All the models were assessed by the counts of true positives (TP) of each class. The predictive accuracy (Q_i) of each class, the positive predictive value (precision rate) which is the proportion of positive test results that are true positives, and the total predictive accuracy (Q) which represents the predictive accuracy of all chemicals were calculated with the following equations.

$$Q_i = \frac{TP}{P_i} \quad (1)$$

$$PPV_i = \frac{TP}{PC_i} \quad (2)$$

$$Q_{\text{total}} = \frac{\sum_{i=1}^n TP}{P} \quad (3)$$

Herein, Q_i is the predictive accuracy of class i , TP_i is the correct count of class i , P_i is the total number of class i , PPV_i is the precision rate of class i , PC_i is the number of chemicals which are predicted into class i (positive calls), Q_{total} is the total predictive accuracy of all chemicals in data set, and P is the total number of all chemicals.

Analysis of Privileged Substructures or Structural Alerts. The privileged substructures or structure alerts (SAs) are defined as molecular functional groups that are known to bring the toxicity. Their appearance in a chemical structure alerts the researchers to the potential toxicities of the test compounds.³² Structural alerts are important predictive toxicity tools due to being derived directly from mechanistic knowledge.³³ Here, the privileged substructure fragments were analyzed using the information gain and substructure frequency analysis.³⁴ The privileged structures known to have provided ligands for diverse receptors are capable of illustrating the biological mechanism.³⁵ If a substructure was more frequently presented in category I and category II chemical classes, this substructure could be a privileged substructure involved in chemical toxicity. The frequency of a fragment was defined as follows:

$$F = \frac{N_{\text{fragment_class}} * N_{\text{total}}}{N_{\text{fragment_total}} * N_{\text{class}}} \quad (4)$$

where $N_{\text{fragment_class}}$ is the number of compounds containing the fragment in category I and category II chemicals, N_{total} is the total number of compounds, $N_{\text{fragment_total}}$ is the total number of compounds containing the fragment, and N_{class} is the number of category I and category II chemicals.

RESULTS

Data Set Analysis. The training set and test set contained 8102 and 2049 compounds, separately. The external validation set I was composed of 1678 compounds, and external validation set II contained 375 compounds. According to the U.S. EPA definition of toxicity category, all compounds were classified into four categories based on their LD_{50} values (see Table 2).

To investigate the chemical space distribution, the molecule weight (MW) and Ghose–Crippen LogKow (ALogP) of each class in the database were analyzed. The distribution scatter

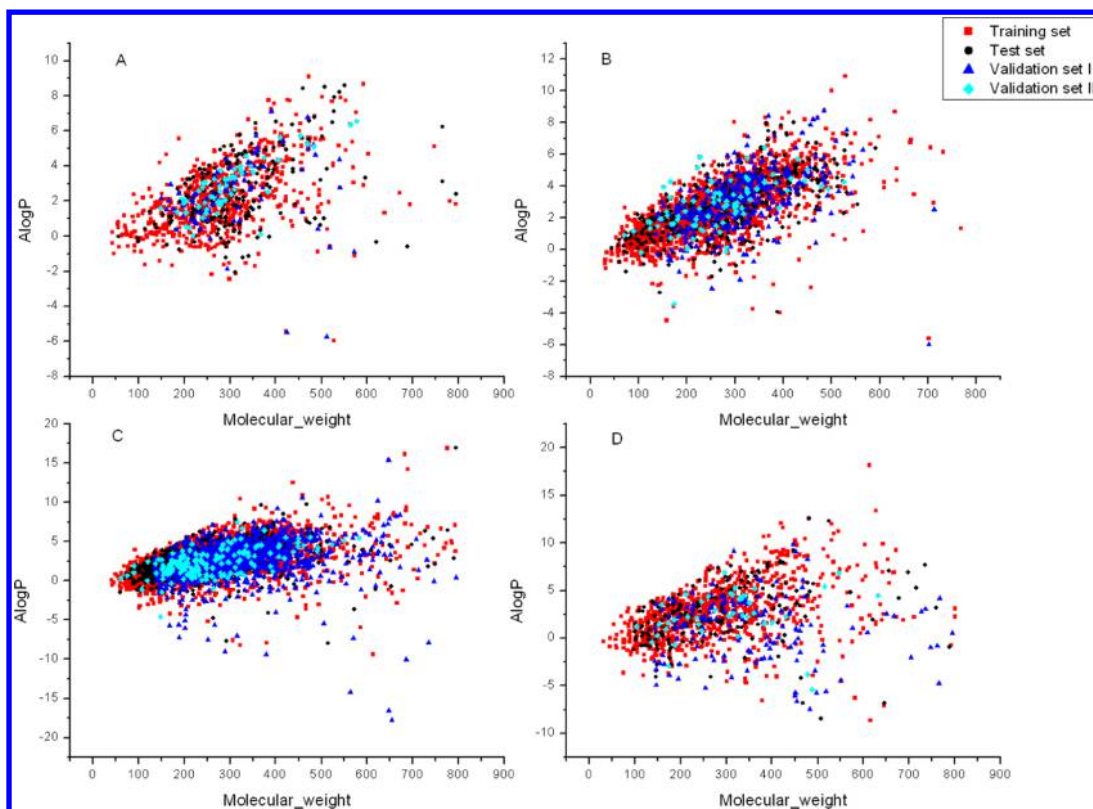


Figure 2. Chemical space defined by molecular weight and AlogP of each class in data sets. (A) Category I; (B) category II; (C) category III; (D) category IV. Red squares stand for the training set, black circles stand for the test set, blue triangles stand for the external validation set I, cyan diamonds stand for the external validation set II.

diagram was presented in Figure 2, which illustrated that the test and validation sets shared similar chemical space with the training set. In addition, the chemical space of a set of 969 FDA approved small molecule drugs from DrugBank was compared with that of the training set. As shown in Figure 3, the chemical space of the data set used in this study was well comparable with that of approved drugs.

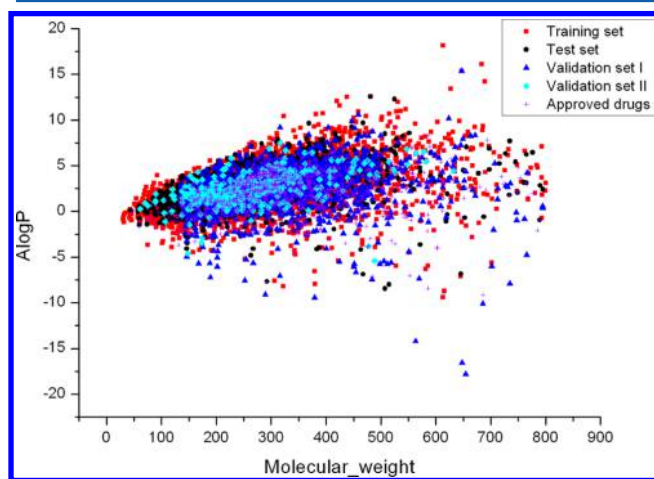


Figure 3. Chemical space defined by molecular weight and AlogP of data sets and approved drugs from Drugbank. Red squares stand for the training set, black circles stand for the test set, blue triangles stand for the external validation set I, cyan diamonds stand for the external validation set II, violet crosses stand for the approved drugs.

Construction of Multiclass Models. Totally, 12 multiclass models were built using different methods. Among them, eight models were generated by C4.5, RF, NB, and kNN algorithms along with MACCS and FP4 fingerprints. OAO and BT approaches were implemented to construct the other four SVM models together with MACCS and FP4 fingerprints. As shown in Figure 1A, six binary SVM classifiers were trained for each pair of categories in OAO method. Classification of an unknown compound was done according to the maximum vote from the six classifiers. Unlike OAO, BT utilized a hierarchical division to construct a multiclass model. It is clear that in Figure 1B this binary tree was made up of three SVM classifiers. The root node made discrimination between categories I and II and categories III and IV at first. The left output was divided into category I and II, while the right output was divided into category II and IV by two subnodes afterward. Each chemical was classified through two SVM binary classifiers.

The test set was used to evaluate the performance of these multiclass models, and the detailed results were presented in Table 3. Most of the models exhibited good overall predictive performance for the test set. Among these models, MACCS-SVM_{OAO} and MACCS-kNN models provided the best results with high predictive accuracy. At the same time, they also showed good performance for each class. The MACCS-SVM_{OAO} model was obtained with the accuracy of 78.2% for category I, 76.7% for category II, 91.2% for category III, and 58.3% for category IV. The accuracies of the MACCS-kNN model for the four classes were 84.0%, 79.3%, 85.5%, and 71.1%, respectively. Moreover, the models that used MACCS

Table 3. Performance of Models for Test Set

models	category I		category II		category III		category IV		Q _{total}
	Q	PPV	Q	PPV	Q	PPV	Q	PPV	
MACCS-SVM _{OAO}	0.782	0.876	0.767	0.785	0.912	0.859	0.583	0.708	0.832
MACCS-SVM _{BT}	0.796	0.829	0.756	0.792	0.951	0.825	0.275	0.966	0.822
SubFP-SVM _{OAO}	0.707	0.710	0.551	0.614	0.825	0.783	0.446	0.476	0.712
SubFP-SVM _{BT}	0.667	0.743	0.581	0.614	0.882	0.769	0.284	0.707	0.731
MACCS-NB	0.587	0.332	0.361	0.350	0.344	0.690	0.632	0.216	0.403
SubFP-NB	0.613	0.496	0.309	0.460	0.765	0.687	0.294	0.349	0.598
MACCS-CT	0.716	0.676	0.633	0.579	0.767	0.797	0.456	0.484	0.700
SubFP-CT	0.698	0.631	0.488	0.585	0.806	0.749	0.382	0.459	0.680
MACCS-RF	0.747	0.757	0.577	0.686	0.912	0.771	0.275	0.800	0.755
SubFP-RF	0.564	0.679	0.413	0.632	0.935	0.710	0.137	0.778	0.697
MACCS-kNN	0.840	0.829	0.793	0.754	0.855	0.893	0.711	0.642	0.825
SubFP-kNN	0.653	0.607	0.577	0.539	0.717	0.778	0.417	0.344	0.648

Table 4. Performance of Models for External Validation Set I

models	category I		category II		category III		category IV		Q _{total}
	Q	PPV	Q	PPV	Q	PPV	Q	PPV	
MACCS-SVM _{OAO}	0.761	0.875	0.740	0.707	0.890	0.872	0.617	0.763	0.830
MACCS-SVM _{BT}	0.739	0.829	0.731	0.689	0.902	0.843	0.326	0.885	0.810
SubFP-SVM _{OAO}	0.663	0.629	0.567	0.545	0.802	0.807	0.447	0.492	0.717
SubFP-SVM _{BT}	0.685	0.733	0.582	0.567	0.860	0.792	0.220	0.721	0.740
MACCS-NB	0.565	0.158	0.409	0.294	0.427	0.749	0.291	0.168	0.420
SubFP-NB	0.500	0.311	0.316	0.352	0.699	0.728	0.284	0.244	0.575
MACCS-CT	0.663	0.526	0.605	0.474	0.700	0.821	0.567	0.432	0.667
SubFP-CT	0.565	0.406	0.453	0.414	0.734	0.763	0.348	0.426	0.635
MACCS-RF	0.598	0.647	0.523	0.530	0.851	0.783	0.291	0.745	0.723
SubFP-RF	0.348	0.561	0.336	0.485	0.890	0.738	0.255	0.679	0.694
MACCS-kNN	0.783	0.699	0.728	0.586	0.768	0.887	0.738	0.533	0.758
SubFP-kNN	0.609	0.519	0.596	0.501	0.726	0.820	0.504	0.382	0.675

Table 5. Performance of Models for External Validation Set II

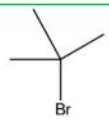
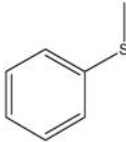
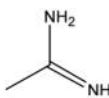
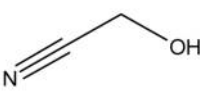

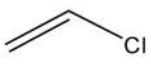
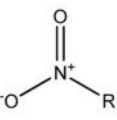
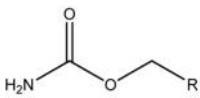
models	category I		category II		category III		category IV		Q _{total}
	Q	PPV	Q	PPV	Q	PPV	Q	PPV	
MACCS-SVM _{OAO}	0.965	0.932	0.860	0.889	0.934	0.900	0.738	0.861	0.899
MACCS-SVM _{BT}	0.860	0.817	0.774	0.828	0.962	0.789	0.119	1.000	0.805
SubFP-SVM _{OAO}	0.789	0.738	0.559	0.703	0.902	0.782	0.500	0.724	0.755
SubFP-SVM _{BT}	0.772	0.786	0.538	0.694	0.945	0.739	0.310	1.000	0.747
MACCS-NB	0.702	0.215	0.398	0.230	0.055	0.667	0.167	0.538	0.251
SubFP-NB	0.579	0.277	0.473	0.278	0.295	0.593	0.071	0.429	0.357
MACCS-CT	0.825	0.770	0.677	0.692	0.847	0.812	0.500	0.656	0.763
SubFP-CT	0.632	0.655	0.398	0.561	0.852	0.681	0.333	0.560	0.648
MACCS-RF	0.684	0.780	0.548	0.654	0.896	0.686	0.119	0.625	0.691
SubFP-RF	0.404	0.657	0.237	0.611	0.956	0.579	0.048	1.000	0.592
MACCS-kNN	0.965	0.917	0.903	0.875	0.885	0.942	0.833	0.745	0.896
SubFP-kNN	0.684	0.813	0.613	0.655	0.858	0.793	0.595	0.595	0.741

fingerprints as attributes performed quite better than those that used FP4 fingerprint.

Performance of External Validation. The generalization abilities of the models were estimated by two external validation sets. The validation sets were totally independent from the training and test sets. Compounds in the validation sets were not used to construct the models, and thus the performance of the models on the validation sets could objectively reflect the predictive ability of the models. The detailed results of the models were illustrated in Tables 4 and 5. Similar with the results for test set, the MACCS-SVM_{OAO} and MACCS-kNN models were the best models with accuracies of 83.0% and

75.8% for external validation set I and 89.9% and 89.6% for external validation set II, respectively. They also yielded high predictive accuracy for each class. Besides, MACCS generated models with higher accuracy than FP4 fingerprint. MACCS molecular fingerprint is based on the well-defined structural fragments dictionary, which is full of structural information.³⁶ Each bit of FP4 corresponds to a particular chemical feature rather than to the general patterns, and therefore, the features defined beforehand are often very restrictive to represent a large number of chemicals. Considering the amount of computation and prediction accuracy, MACCS was recommended for building the *in silico* acute oral toxicity models.

Table 7. Privileged Substructures in Category II Chemicals Using Information Gain (IG) and Frequency Value Analysis

No.	Description	Frequency in individual class				SMARTS	General structure
		I	II	III	IV		
1	Alkylbromide	0.685	1.763	0.791	0.713	[BrX1][CX4]	
2	Alkylarylthioether	0.754	1.617	0.898	0.475	[SX2](c)[CX4;!\$(C([SX2])[O,S,#7,#15])]	
3	Amidine	0.851	2.130	0.654	0.462	[NX3;!\$(NC=[O,S])][CX3;\$([CH]),\$(C([#6])=[NX2;!\$(NC=[O,S])])]	
4	Cyanhydrine	1.729	2.014	0.511	0.600	[NX1]#[CX2][CX4;\$([CH2]),\$(C([CH])([CX2])([#6]),\$(C([CX2])([#6])([#6]))[OX2H])]	
5	Nitrile	1.205	1.808	0.716	0.520	[NX1]#[CX2]	
6	Chloroalkene	2.175	1.573	0.584	0.755	[CIX1][CX3]=[CX3]	
7	Nitro	1.456	1.324	0.834	0.730	[\$([NX3](=O)=O),\$([NX3+](=O)[O-])][!#8]	
8	Urethan	1.917	1.282	0.823	0.483	[#7X3][#6](=[OX1])[#8X2][#6]	

Diversity of Data Set. Chemical diversity is considered to be a key factor that could influence the prediction capability of QSAR models. Many QSAR models have been developed for various toxicity end points. However, most of these models were based on relatively small data sets or homologous compounds. As a result they usually had poor generalization capability on external data sets. To build more reliable models, we collected 10 151 diverse compounds as the training set and test set. And multiclass models were obtained with high predictive accuracy for the external validation sets, which indicated that these multiclass models would have good external prediction power.

Analysis of Substructural Alerts. Fourteen substructures were identified to appear in toxic compounds more frequently than the others, and they were believed responsible for extremely acute oral toxicity. As shown in Table 6, three of them were phosphorus fragments. Phosphonic acid derivative, phosphoric trimester, and phosphoric acid derivative exist in a large class of pesticides as organophosphates. They can inhibit the activity of cholinesterase and result in the accumulation of acetylcholine, a neurotransmitter of the cholinergic receptor, which will make the function of the cholinergic nervous system disordered. They can also affect the cholinergic receptor

directly, leading to the next neuron or effector to excessive excitement or inhibition.³⁷ The toxicity of cyanhydrine compounds was mainly caused by its decomposition with water. The decomposition product cyanide ion could halt cellular respiration by inhibiting cytochrome c oxidase in mitochondria. Nitrile was also a potentially toxic fragment. Many nitrile compounds such as acetonitrile, acrylonitrile, and propionitrile were highly toxic mainly due to the release of cyanide anions through hydrolysis.³⁸ The cyanide anion was an inhibitor of the enzyme cytochrome c oxidase, by which it could affect the central nervous system and the heart. A lot of nitro-compounds, especially nitroaromatics, were hazardous chemicals which displayed several manifestations of toxicity in humans.³⁹ Their toxicity was primarily caused by electrophilic reactivity of the nitro group.⁴⁰ Carbamate insecticides always contained the ethyl carbamate functional group also called urethan. They could inhibit the enzyme acetylcholinesterase leading to high toxicity.

It is interesting that alkylfluoride was also observed as a substructure of highly toxic chemicals. In fact the toxicity did not arise from alkylfluoride alone. After carefully analyzing these alkylfluoride compounds, we learned that some toxic fragments contain alkylfluoride such as 2-(trifluoromethyl)-

benzimidazole that exhibited cytotoxicity, antibacterial, and antifungal activity in previous studies.^{41–43} This exact fragment could not be obtained because 2-(trifluoromethyl)-benzimidazole was not defined in the fingerprint dictionary. This also suggests that specific fingerprints should be developed against different end points to improve QSAR modeling.

Analysis of Misclassified Compounds. The model using MACCS fingerprint and SVM_{OAO} algorithm achieved excellent predictive ability. Nevertheless, some compounds in validation sets were predicted incorrectly. The detailed statistical results of MACCS-SVM_{OAO} model for external validation set II are shown in Table 8. There were, totally, 38 compounds in

Table 8. Detailed Statistical Results of MACCS-kNN Model for External Validation Set II^a

	P_I	P_{II}	P_{III}	P_{IV}	total
O_I	55	2	0	0	57
O_{II}	4	80	9	0	93
O_{III}	0	7	171	5	183
O_{IV}	0	1	10	31	42
total	59	90	190	36	375

^a O_I : number of objective category I chemicals; O_{II} : number of objective category II chemicals; O_{III} : number of objective category III chemicals; O_{IV} : number of objective category IV chemicals; P_I : number of predicted category I chemicals; P_{II} : number of predicted category II chemicals; P_{III} : number of predicted category III chemicals; P_{IV} : number of predicted category IV chemicals.

validation set II predicted incorrectly by the model. Nineteen of them should belong to category III but were misclassified in category II (nine compounds) and category IV (10 ones). This could be explained by that the number of moderately toxic compounds is much larger than the others. It was also illustrated that the incorrectly predicted compounds were mostly predicted into nearest classes. For instance, there were no category I compounds predicted as category IV ones and no category IV compounds predicted as category I ones.

CONCLUSIONS

In this study, multiclassification models were built for the prediction of chemical acute oral toxicity with C4.5, RF, NB, kNN, and SVM algorithms. MACCS and FP4 fingerprints were used for chemical description. The SVM algorithm that employed the OAO approach along with MACCS fingerprint showed reliable ability in generalization, and this MACCS-SVM_{OAO} model has been integrated as part of our Web server admetSAR, which is freely available at <http://www.admetexp.org/>. Compared with regression methods, the semiquantitative model could determine toxic severity of compounds with high accuracy directly. Besides, fingerprints proved effective attributes for acute toxicity models, which could connect the molecular structure and properties.

We also proposed some privileged substructures using information gain and substructure frequency analysis methods. These privileged substructures appear more frequently in compounds with high toxicity, and thus they should be responsible for acute oral toxicity. Our study provided a useful tool to estimate acute toxicity in chemical safety assessment. And the multiclassification strategy in this study might be promoted to other toxicity end points.

AUTHOR INFORMATION

Corresponding Author

*Phone: +86-21-6425-1052. Fax: +86-21-6425-3651. E-mail: ytang234@ecust.edu.cn.

Author Contributions

[§]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the 863 Project (Grant 2012AA020308), the National Natural Science Foundation of China (Grant 81373329), and the Fundamental Research Funds for the Central Universities (Grant WY1113007).

REFERENCES

- (1) Walum, E. Acute oral toxicity. *Environ. Health Perspect.* **1998**, 106 (Suppl. 2), 497–503.
- (2) Guo, J. X.; Wu, J. J.; Wright, J. B.; Lushington, G. H. Mechanistic insight into acetylcholinesterase inhibition and acute toxicity of organophosphorus compounds: a molecular modeling study. *Chem. Res. Toxicol.* **2006**, 19, 209–216.
- (3) Freidig, A. P.; Dekkers, S.; Verweij, M.; Zvinavashe, E.; Bessems, J. G.; van de Sandt, J. J. Development of a QSAR for worst case estimates of acute toxicity of chemically reactive compounds. *Toxicol. Lett.* **2007**, 170, 214–222.
- (4) Toropov, A. A.; Rasulev, B. F.; Leszczynski, J. QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by MLRA and optimal descriptors. *QSAR Comb. Sci.* **2007**, 26, 686–693.
- (5) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, 22, 1913–1921.
- (6) *Label Review Manual*; U.S. EPA: Washington, DC, 2012; Chapter 7: Precautionary Statements.
- (7) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* **2012**, 52, 3099–3105.
- (8) MDL Toxicity Database (presently Accelrys Toxicity Database). <http://accelrys.com/products/databases/bioactivity/toxicity.html> (accessed on February 14th, 2013).
- (9) Quantitative Structure Activity Relationship. <http://www.epa.gov/nrmrl/std/qsar/qsar.html> (accessed on February 14th, 2013).
- (10) 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.
- (11) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1273–1280.
- (12) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, 32, 1466–1474.
- (13) Chang, C. C.; Lin, C. J. LIBSVM -- A library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed on February 14th, 2013).
- (14) Orange, version 2.6.1. <http://www.ailab.si/orange/> (accessed on February 14th, 2013).
- (15) Quinlan, J. R. *C4.5: programs for machine learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1993.
- (16) Breiman, L. Random forests. *Machine Learning* **2001**, 45, 5–32.
- (17) Watson, P. Naïve Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, 48, 166–178.
- (18) Cover, T. M.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, 13, 21–27.

- (19) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
- (20) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of cytochrome P450 inhibitors and non-inhibitors using combined classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996–1011.
- (21) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.
- (22) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J. Chem. Inf. Model.* **2007**, *47*, 92–103.
- (23) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome P450 substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588–2605.
- (24) Hsu, C.-W.; Lin, C.-J. A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks* **2002**, *13*, 415–425.
- (25) Fei, B.; Liu, J. Binary tree of SVM: a new fast multiclass training and classification algorithm. *IEEE Trans. Neural Networks* **2006**, *17*, 696–704.
- (26) Cheong, S.; Oh, S. H.; Lee, S.-Y. Support vector machines with binary tree architecture for multi-class classification. *Neural Inf. Process. Lett. Rev.* **2004**, *2*, 47–51.
- (27) Debnath, R.; Takahide, N.; Takahashi, H. A decision based one-against-one method for multi-class support vector machine. *Pattern Anal. Appl.* **2004**, *7*, 164–175.
- (28) Zhang, H.; Xiang, M.-L.; Ma, C.-Y.; Huang, Q.; Li, W.; Xie, Y.; Wei, Y.-Q.; Yang, S.-Y. Three-class classification models of logS and logP derived by using GA–CG–SVM approach. *Mol. Divers.* **2009**, *13*, 261–268.
- (29) Dejaegher, B.; Dhooghe, L.; Goodarzi, M.; Apers, S.; Pieters, L.; Heyden, Y. V. Classification models for neocryptolepine derivatives as inhibitors of the β -haematin formation. *Anal. Chim. Acta* **2011**, *705*, 98–110.
- (30) Madzarov, G.; Gjorgjevikj, D.; Chorbev, I. A multi-class SVM classifier utilizing binary decision tree. *Informatica* **2009**, *33*, 233–241.
- (31) Qu, D.; Li, W.; Zhang, Y.; Sun, B.; Zhong, Y.; Liu, J.; Yu, D.; Li, M. Support vector machines combined with wavelet-based feature extraction for identification of drugs hidden in anthropomorphic phantom. *Measurement* **2013**, *46*, 284–293.
- (32) Kruhlak, N. L.; Contrera, J. F.; Benz, R. D.; Matthews, E. J. Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv. Drug Delivery Rev.* **2007**, *59*, 43–55.
- (33) 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. Res.* **2008**, *659*, 248–261.
- (34) Jensen, B. F.; Vind, C.; Padkjær, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.
- (35) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.
- (36) 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.
- (37) Casida, J. E.; Quistad, G. B. Organophosphate toxicology: safety aspects of nonacetylcholinesterase secondary targets. *Chem. Res. Toxicol.* **2004**, *17*, 983–998.
- (38) Bhattacharya, R.; Satpute, R. M.; Hariharakrishnan, J.; Tripathi, H.; Saxena, P. B. Acute toxicity of some synthetic cyanogens in rats and their response to oral treatment with alpha-ketoglutarate. *Food Chem. Toxicol.* **2009**, *47*, 2314–2320.
- (39) Katritzky, A. R.; Oliferenko, P.; Oliferenko, A.; Lomaka, A.; Karelson, M. Nitrobenzene toxicity: QSAR correlations and mechanistic interpretations. *J. Phys. Org. Chem.* **2003**, *16*, 811–817.
- (40) Cronin, M. T.; Gregory, B. W.; Schultz, T. W. Quantitative structure-activity analyses of nitrobenzene toxicity to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **1998**, *11*, 902–908.
- (41) Andrzejewska, M.; Yepez-Mulia, L.; Cedillo-Rivera, R.; Tapia, A.; Vilpo, L.; Vilpo, J.; Kazimierczuk, Z. Synthesis, antiprotozoal and anticancer activity of substituted 2-trifluoromethyl- and 2-pentafluoroethylbenzimidazoles. *Eur. J. Med. Chem.* **2002**, *37*, 973–978.
- (42) Stefanska, J. Z.; Graleska, R.; Starosciak, B. J.; Kazimierczuk, Z. Antimicrobial activity of substituted azoles and their nucleosides. *Pharmazie* **1999**, *54*, 879–884.
- (43) Wolinowska, R.; Zajdel-Dabrowska, J.; Starosciak, B. J.; Kazimierczuk, Z. Antimicrobial activity of substituted 2-trifluoromethyl- and 2-pentafluoroethylbenzimidazoles. *Acta. Microbiol. Polym.* **2002**, *51*, 265–273.