RESEARCH ARTICLE

ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling

Tailong Lei1, Youyong Li3, Yunlong Song4, Dan Li1, Huiyong Sun1 and Tingjun Hou1,2*

Abstract

Background: Determination of acute toxicity, expressed as median lethal dose (LD50), is one of the most important steps in drug discovery pipeline. Because in vivo assays for oral acute toxicity in mammals are time-consuming and costly, there is thus an urgent need to develop in silico prediction models of oral acute toxicity.

Results: In this study, based on a comprehensive data set containing 7314 diverse chemicals with rat oral LD50 values, relevance vector machine (RVM) technique was employed to build the regression models for the prediction of oral acute toxicity in rate, which were compared with those built using other six machine learning approaches, including k-nearest-neighbor regression, random forest (RF), support vector machine, local approximate Gaussian process, multilayer perceptron ensemble, and eXtreme gradient boosting. A subset of the original molecular descriptors and structural fingerprints (PubChem or SubFP) was chosen by the Chi squared statistics. The prediction capabilities of individual QSAR models, measured by q2 ext for the test set containing 2376 molecules, ranged from 0.572 to 0.659.

Conclusion: Considering the overall prediction accuracy for the test set, RVM with Laplacian kernel and RF were recommended to build in silico models with better predictivity for rat oral acute toxicity. By combining the predictions from individual models, four consensus models were developed, yielding better prediction capabilities for the test set (q2 ext = 0.669–0.689). Finally, some essential descriptors and substructures relevant to oral acute toxicity were identified and analyzed, and they may be served as property or substructure alerts to avoid toxicity. We believe that the best consensus model with high prediction accuracy can be used as a reliable virtual screening tool to filter out compounds with high rat oral acute toxicity.

Background

Determination of acute toxicity in mammals (e.g. rats or mice) is one of the most important tasks for the safety evaluation of drug candidates. Acute toxicity is usually expressed as median lethal dose (LD50), which is the dose amount of a tested molecule to kill 50 % of the treated animals within a given period. According to the regulations and guidelines for the toxicity testing of pharmaceutical substances established by the Organization for Economic Co-operation and Development (OECD), the U.S. Food and Drug Administration (FDA), the National Institutes of Health (NIH), the European Agency for the Evaluation of Medicinal Products (EMEA), etc., the use of alternative in vitro or in silico toxicity assessment methods that avoid the use of animals are strongly recommended [1–4]. Moreover, in vivo testing for acute toxicity is time-consuming and costly, and therefore extensive efforts have been devoted to the development of in silico methods for toxicity.

Over past decades, a number of quantitative structure–activity relationship (QSAR) models have been developed to predict rodent acute toxicity [5–7]. It is Open Access

© 2016 Lei et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

*Correspondence: tingjunhou@zju.edu.cn; tingjunhou@hotmail.com
2 State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, People’s Republic of China
Full list of author information is available at the end of the article
well-known that acute toxic effect results from multiple potential modes of action (MOA), and it is quite difficult to develop a universal model with reliable prediction accuracy to an extensive data set. Therefore, most QSAR models were built from small data sets of congenic compounds [8–10] and thus had limited application domains. Recently, several theoretical models were developed based on relatively large-scale data sets with diverse compounds [9–12]. For example, Zhu et al. [10] developed five QSAR models for 7385 compounds with rat oral acute toxicity data, and the two models developed by kNN and RF achieved comparable performance for the test set ($r^2 = 0.66$ and 0.70, respectively) to TOP-KAT. However, in Zhu’s study, 997 molecules were identified as outliers and eliminated from the training set. Another study reported by Raevsky [13] and coworkers proposed a so-called Arithmetic Mean Toxicity (AMT) modelling approach, which produced local models based on a $k$-nearest neighbors approach. This approach gave correlation coefficients ($r^2$) from 0.456 to 0.783 for 10,241 tested compounds, but the prediction accuracy for a molecule depended on the number and structural similarity of its neighbors with experimental data in the training set [13]. Recently, Lu et al. [14] employed local lazy learning (LLL) method to develop $LD_{50}$ prediction models, and the rat acute toxicity of a molecule could be predicted by the experimental data of its $k$ nearest neighbors. A consensus model by integrating the predictions of individual LLL models yielded a correlation coefficient $r^2$ of 0.712 for the test set containing 2896 compounds. Similar to Raevsky’s approach [13], Lu’s approach relied on the priori knowledge of the experimental data of a query’s neighbors, and therefore, the actual prediction capability of this method was associated with the chemical diversity and structural coverage of the training set [15].

Due to the complicated mechanisms involved in acute toxicity, it is a difficult task to build a single QSAR model with reliable prediction accuracy by using traditional statistical approaches, such as multiple linear regression (MLR), partial least squares (PLS), principal components regression (PCR), etc. However, machine learning methods have shown promising potential to establish the complex QSARs for the data sets with diverse ranges of molecular structures and mechanisms. Certainly, each machine learning method has its intrinsic advantages, shortcomings, and practical constraints. Moreover, the performance of different machine learning methods depends on the structural diversity and representativeness of the molecules in the data set. Therefore, it is quite important to choose the most suitable machine learning method to develop the prediction model for a specific toxicity data set.

Among all existed machine learning methods, most of them may have the common problem of overtraining and overfitting in solving high-dimensional and complex nonlinear problems because they usually need to estimate and optimize many hyperparameters. It is well-known that the complexity of a model often grows linearly with the dimension of data, and thus some forms of post-processing are required to reduce the computational complexity. In order to solve this problem, the relevance vector machine (RVM) method introduced the Bayesian criteria into learning process, and it employs a sparse prior to reduce the irrelevant support vectors of the decision boundary in feature space and gets a sparser model accordingly. Contrary to the similar algorithm, support vector machine (SVM), the penalty parameter $C$ and the insensitive-loss parameter $\epsilon$ are automatically valued and error bars are got through covariance function in the RVM regression. Meanwhile, RVM has a comparable generalization ability, and its non-zero weights reflect prototype of sampling more than SVM. Therefore, RVM may be a good choice for QSAR modelling.

In this study, based on a large public data set containing 7385 rat oral acute toxicity data compiled by the previous study [10], RVM was employed to establish the regression models for the prediction of oral acute toxicity in rat, and was compared with the other six machine learning methods, including SVM, $k$-nearest-neighbor regression ($k$NN), random forest (RF), local approximate Gaussian process (laGP), multilayer perceptron ensemble (MPLE), and eXtreme gradient boosting (XGBoost). The performance of all the seven machine learning methods was assessed and compared by the predictive power and application domains of the models to the external test set. Moreover, the possibility to achieve better prediction of rat oral acute toxicity by combining the predictions from multiple QSAR models was explored.

Methods

Data set of rat oral acute toxicity

The rat oral $LD_{50}$ data set with 7385 unique organic molecules reported by Zhu et al. [10] was used in our study. The quality of the data set, originally collected from different sources, was carefully verified. The acute toxicity of each molecule was expressed as $\log(1/(\text{mol/kg}))$ (or $pLD_{50}$).

The SMILES of the 7385 structures in the data set were converted into 3-D structures and optimized in Discovery Studio 2.5 molecular simulation package (DS 2.5) [16]. Here, 68 molecules were eliminated because some molecular descriptors of them could not be successfully generated by Molecular Operating Environment (MOE) 2009 molecular simulation package [17], and 3 molecules...
with pLD_{50} values higher than 7.0 or lower than 0, distantly distributed from the other data, were removed. The final data set contained 7314 molecules, which were randomly re-split into a training set with 4938 (67.5 %) molecules and an external test set with 2376 (32.5 %) molecules by weighing the distribution of their pLD_{50} values.

**Calculation of molecular descriptors and molecular fingerprints**

Originally, 334 descriptors to characterize the physiochemical properties, molecular representations, and drug-like properties of the studied molecules were calculated by using MOE. The descriptors that had zero values or zero variance were removed. Then, the correlations across all pairs of descriptors were calculated, and the redundant descriptors with the correlation \( r \) higher than the predefined threshold (0.95) to any descriptor were removed. Finally, 230 descriptors were chosen for QSAR modeling. In addition, molecular fingerprints, which characterize the substructure features of a molecule, were used. Two sets of fingerprints, including the PubChem fingerprint (PubchemFP) with 881 substructure patterns, and substructure fingerprint (SubFP) with 307 substructure patterns, were generated by PaDEL-Descriptor software [18].

**Dimension reduction by Chi squared statistics**

Dimension reduction is essential to the regression analysis of high-dimensional data. The idea in dimension reduction is to find a projection \( p \) to a \( k \)-dimensional subspace, \( k \ll p \), such that less information is lost. Ensemble feature selection is a subset of dimension reduction techniques that apply feature selection algorithms multiple times and combine the results into one decision. The Chi squared statistics, which is one of the most popular ensemble feature selection techniques [19, 20], was employed here to find a subset of the original descriptors. It is a non-threshold based feature selection technique and has been successfully applied in many fields [21, 22]. The Chi squared statistics compared the observed distribution of class-feature value pairs to the distribution predicted by a Chi squared random distribution, and those features distinct from this null distribution were preferred. Then, the Chi squared scores \( \chi^2 \) were transformed into Cramer’s V coefficients (Eq. 1) [23]. At last, the Cramer’s V coefficients were ranked and 120 or 150 features with the highest Cramer’s V coefficients were chosen (Eq. 2).

\[
\chi^2 = \sum_{ij} (O_{ij} - E_{ij})^2 / E_{ij}
\]

\[
V = \sqrt{\frac{\chi^2 / n}{\min(k - 1)}}
\]

where \( \chi^2 \) is Chi square scores, \( O_{ij} \) is the observation values from measurement, \( E_{ij} \) is the expectation values from prediction, \( n \) is the grand total of observations and \( k \) is the number of rows or columns in the contingency table.

**QSAR modeling by machine learning approaches**

A variety of machine learning approaches have been used to develop regression models for the prediction of ADME and toxicities [24–28]. Here, seven machine learning methods, including kNN, RF, SVM, RVM, laGP, MPLE, and XGBoost, were employed for model building. Two main packages of machine learning in R (version 3.1.3 x64), rminer [29] and mlr [30], provide generic and object-oriented interfaces to the employed machine learning methods with good scalability. The important parameters for QSAR modeling are summarized in Table 1.

**Relevance vector machine (RVM)**

Relevance vector machine (RVM), pioneered by Tipping, is a sparse Bayesian learning algorithm for regression and probabilistic classification developed from the standard SVM [31, 32]. It has shown better generalization performance than SVM, but it allows avoiding the set of free parameters that SVM has. RVM acquires relevance vectors and weights \( (w) \) by maximizing a marginal likelihood. The products of weights and kernel functions give the structure of RVM. For a data set of input-target pairs \( \{x_n, t_n\}_{n=1}^N \), we follow the standard probabilistic formulation and assume \( p(t_n|x) \) is Gaussian \( N(t_n|y(x_n), \sigma^2) \). Thus the likelihood of the training data set can be written as:

| Models       | Hyperparameters                                      |
|--------------|-------------------------------------------------------|
| kNN          | The number of predictors at each node = 1–10          |
| RF           | The number of predictors at each node = 105, the number of trees = 230 |
| SVM (RBF)    | The kernel width \( \sigma = 0.03125 \), the penalty parameter \( C = 2 \), and \( \varepsilon \) in the loss function = 0.005 |
| RVM (Laplace) | The kernel width \( \sigma = 0.044 \)                      |
| laGP         | The initial values of lengthscale = 5, the initial values of nugget = 0.1 |
| MPLE         | The number of individual perceptrons = 18, the number of units in the hidden layer = 5–8 |
| XGBoost      | Step size shrinkage = 0.1, maximum depth of a tree = 7, the max number of iterations = 69 |
\[ p(t_n|w, \sigma^2) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{ -\frac{1}{2\sigma^2}t - \Phi w^2 \right\} \] (3)

A proper kernel function is selected to create the \( N \times (N + 1) \) design matrix \( \Phi \) with \( [\phi(x_1), \phi(x_1), \ldots, \phi(x_N)]^T \), wherein \( \phi(x_n) = [1, K(x_n, x_1), K(x_n, x_2), \ldots, K(x_n, x_N)]^T \). To avoid overfitting, starting values for hyper-parameter \( \alpha \) and \( \beta \) (i.e. \( \sigma^{-2} \)) are chosen to build a zero-mean Gaussian prior distribution over the weights as:

\[ p(w|\alpha) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1}) \] (4)

Then the posterior distribution over the weights is calculated as:

\[ p(w|t, \alpha, \sigma^2) = (2\pi)^{-(N+1)/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(w - \mu)^T \Sigma^{-1}(w - \mu) \right\} \] (5)

where the posterior covariance and mean are shown in Eqs. (6) and (7), respectively:

\[ \Sigma = \left( \beta \Phi^T \Phi + A \right)^{-1}, \] (6)

\[ \mu = \beta \Sigma \Phi^T t, \] (7)

with \( A = \text{diag}(\alpha_0, \alpha_1, \ldots, \alpha_N) \) and \( t = (t_1, t_2, \ldots, t_N) \). The likelihood distribution over the training set can be marginalized by integrating the weights to obtain the marginal likelihood. Every hyper-parameter is iterated by type-II maximum likelihood method to maximize a posterior. In every iteration, the hyper-parameters were updated as \( \alpha^\text{New} = \gamma_i/\mu_i^2 \) and \( \beta^\text{New} = \left(N - \Sigma_i \gamma_iight)/t - \Phi \mu^2 \gamma_i \equiv 1 - \alpha_i \Sigma_{ii} \), where the \( i \)-th diagonal element of the posterior weight covariance from Eq. (6) computed with the current \( \alpha \) and \( \beta \) values. Repeat application of Eqs. (6) and (7) concurrently with updating of the posterior statistics \( \Sigma \) and \( \mu \), until convergence criteria (\( \sigma_i \) is increasing toward infinity) have been satisfied. After that we can make predictions based on the posterior distribution over the weights. In this study, a Laplacian kernel \( k(x_i, x_j) = \exp\left( -\frac{\|x_i - x_j\|}{\sigma} \right) \) was used as the kernel function, and the kernel width \( \sigma \) was 0.044.

**Support vector machine (SVM) regression**

Support vector machine (SVM), under the frame of Vapnik–Chervonenkis theory, [33, 34] is one of the most popular machine learning methods used in QSAR modeling [35]. Although SVM was originally developed for classification, it can also be used for regression (or function approximation). In the case of regression, the objective is to find a hyperplane with small norm while simultaneously minimize the sum of the distances from the data points to the hyperplane. In this study, the Gaussian radial basis function (RBF) was used as the kernel, and grid search was employed for the optimization of the kernel parameter \( \sigma \) [36]. The penalty parameter \( C \) of the error term was set to 2, and the insensitive parameter \( \varepsilon \) in the loss function was set to 0.05.

**k-Nearest-neighbor (kNN) regression**

kNN is a non-parametric learning approach for classification and regression based on the closest training examples in the feature space [37, 38]. The feature selection, the number \( k \) of nearest neighbors, and the shape of the distance weighting function determine the performance of a kNN model. Here, each molecule was eliminated from the training set and its pLD50 value was predicted as the inverse distance weighted average activity of the \( k \) most similar molecules, where the value of \( k \) was optimized as well (\( k = 1–10 \)).

**Random forest (RF)**

Random forest (RF) is an ensemble learning method by combining multiple decision trees and yields the consensus predictions from individual trees [39, 40]. It randomly samples the data from the training set to construct individual trees. Each node of the tree is split using the best subset of total descriptors randomly chosen at that node. Here, a 10-puzzle heuristic searching method was used to determine the most optimal parameters in RF modeling. The number of the predictors sampled for splitting at each node was set to 105, and the number of trees to grow was set to 230.

**Local approximate Gaussian process regression (laGP)**

laGP is a parallel approximate Gaussian Process (GP) regression algorithm for big data [41, 42]. The approximation is based on finding small local designs for independent prediction at particular inputs. A Gaussian process can be used as a prior probability distribution over functions in Bayesian inference, with finite dimensional distributions defined by a mean \( \mu(x) \) and positive definite covariance \( K(x, x') \) for \( p \)-dimensional inputs \( x \) and \( x' \). For smoothing noisy data, a nugget (\( \eta/g \)) can be added to \( K(x, x') \) of the isotropic process. The method involves approximating the predictive equations at the local designs \( X_n(x) \) close to a particular generic location.
\[ \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{\left( \frac{SST}{n-1} \right)^{0.5}} \right)^2 \]

where \(\bar{y}\) is the average of the dependent variable, \(n\) is the number of observations, \(y_i\) is the actual observation, and \(\bar{y}\) is the mean of the observations. The failure to converge, or converge on a different local optimum, would result in the selection of a different set of rules.

The statistical significance of each regression model was assessed by adjusted \(R^2\) and tenfold cross-validation \(q^2\) as shown in Eqs. (8) and (9).

\[
R^2_{adj} = 1 - (1 - R^2) \times \frac{n-1}{n-p-1} = 1 - \left[\frac{SSE/(n-p)}{\left(\frac{SST}{n-1}\right)}\right]
\]

\[
q^2 = \frac{SST-PRESS}{SST}
\]

where \(R^2\) is the square of the Pearson correlation coefficient, \(p\) is the number of the parameters in the regression equation, \(SSE\) is the sum of squares of errors, \(SST\) is the total sum of squared deviations of the dependent variable values from their means, and \(PRESS\) is the predictive residual sum of squares.

The conventional coefficient of determination \(R^2\) was used to evaluate the predictive power of each model on the external test set. The acceptability thresholds of \(q^2\) for the training set and \(q^2_{ext}\) for the test set were both set to \(\geq 0.5\). A model is over-fitted when the difference between \(R^2_{adj}\) and \(q^2_{ext}\) is higher than 0.3 [47, 48].

Moreover, other two parameters, mean absolute error (MAE) and root mean square error (RMSE), were used to evaluate the quality of each model.

**Analysis of application domain (AD)**

It is well known that the training set for QSAR modeling might only covers a limited fraction of the entire chemical space and the applicability of any model to the query chemicals is limited, and thus the AD for any model should be defined [49]. As a result, only a certain fraction of the chemicals in any external data set is expected to fall within the AD, and this fraction is therefore referred as the data set coverage. In this study, the Standard Deviation Distance to Model (STD-DM) approach was used to estimate the AD of each model. The detailed description of the algorithm to define AD shown in Eq. (10) has been described in previous literatures [50–53].

\[
STD-DM(J) = STDEV(y(J)) = \sqrt{\frac{\sum (y(J) - \bar{y}(J))^2}{N-1}}
\]

where \(y(J)\) is a quantitative value of prediction for molecule \(J\), and \(N\) is the total number of the molecules in the test set. The margin range of AD was defined as three times of the STD-DM value [50]. When a molecule is outside the AD, the STD-DM value is high and accordingly the margin range is also high.

**Scaffold analysis of molecules with large prediction errors**

The scaffolds for the 249 molecules with large prediction errors (MAE > 1.0) were examined systematically. The scaffolds for each molecule were characterized by four representations, including Murcko frameworks, ring assemblies, bridge assemblies, and the side chains attached to Murcko frameworks. Murcko frameworks developed by Bemis [54] were primarily used to characterize cyclic substructures of molecules. The definitions of these four scaffold representations have been described in previous studies [55, 56]. The scaffolds were
generated by using the Generate Fragments component in Pipeline Pilot 7.5. The frequency of each scaffold architecture was counted, and the scaffolds were sorted by the scaffold frequency. Finally, for each scaffold with frequency equal or larger than 2, its numbers present in the training and test sets were counted.

**Results and discussions**

**Property distributions of rat oral acute toxicity data**

In our study, 7314 organic molecules collected from the previous literature [10] were used for model development and validation. The training and test sets contained 4938 and 2376 molecules, respectively (Fig. 1). The chemical space was characterized by the scattered distributions of the first two principal components derived from the principal component analysis (PCA) for the 334 molecular descriptors and by the scattered distributions of molecule weight and Wildman and Crippen's octanol–water partition coefficient (SlogP) [57]. As shown in Fig. 2, the chemical space of the external test set was roughly within the scope of the training set, and therefore it was feasible to predict the acute toxicity of the molecules in the test set with reasonable reliability by using the QSAR models built from the training set.

The distributions of eight molecular properties for the training and test sets were shown in Fig. 3, and the correlations between rat oral acute toxicity and molecular descriptors were shown in Fig. 4. The eight molecular properties studied here, including molecular weight (MW), H-bond acceptor count (a_acc), flexible rotatable bond count (b_rotN), octanol–water partitioning coefficient (SlogP), intrinsic solubility (logS), topological polar surface area (TPSA), van der Waals volume (vdw_vol), molecular flexibility index (KierFlex), have been widely used in the prediction of ADME and toxicity [58–65].

SlogP and logS were both related to hydrophobicity. As shown in Fig. 3, the SlogP and logS values for 90% of the compounds in the data set were less than 8 and 2, respectively. They did not show any correlation with rat oral toxicity ($R^2 = 0.039$ and 0.057). Meanwhile, 90% of the compounds in the database had a MW smaller than 500, and the correlation analysis showed that MW had a relatively high impact on rat oral toxicity, indicated by the slightly higher correlation ($R^2 = 0.108$). a_acc and TPSA were usually used to represent hydrophilicity, and as shown in Fig. 4, they had worse correlations with rat oral toxicity ($R^2 = 0.029$ and 0.031) than those related to hydrophobicity. The parameter vdw_vol accounted for the size or bulk of a molecule, and it had low correlation with rat oral toxicity ($R^2 = 0.045$). KierFlex and b_rotN characterized the flexibility of a molecule, and both of them had no correlation with rat oral toxicity ($R^2 = 0.022$ and 0.005). Apparently, no single descriptor showed high correlation with rat oral toxicity, and therefore rat oral toxicity could not be reliably predicted by a single or several molecular descriptors.

**Comparison of various regression models for rat oral acute toxicity**

The statistical results for the training and test sets given by the QSAR models based on the MOE descriptors
and two different sets of fingerprints were summarized in Tables 2, 3, 4, and 5. According to the tenfold cross validations for the training set and the predictions for the test set, the performances of the seven machine learning approaches were quite different. Apparently, among these models, the RVM models always gave the best predictions for both the training and test sets ($q^{2}_{\text{ext}} = 0.640–0.659$). The prediction capability ($q^{2}_{\text{ext}} = 0.639–0.646$) of the RF models was slightly worse than that of the RVM models but obviously better than those of the other models. The good performance of RF was not surprising because other recent studies showed that RF models generally outperformed other comparable machine learning approaches for QSAR modeling based on extensive data sets [39, 66–68]. When considering the overall statistics and prediction accuracy, RVM and RF were recommended to build the in silico models for the prediction of rat oral acute toxicity.

As shown in Tables 2, 3, 4, and 5, the MPLE models gave the lowest $q^{2}_{\text{ext}}$ (0.572–0.596) and the highest RMSE (0.729–0.754) and MAE (0.558–0.580) values for the test set, suggesting that they had the worst prediction capabilities. Meanwhile, their $R^{2}_{\text{adj}}$ (0.633–0.656) for the training set were always the lowest. As far as we know, our study was the first application of MPLE in QSAR modeling, and therefore we could not give our judgment to the predictive power of MPLE to different QSAR problems. However, according to our results, MPLE was not a good choice for this specific toxicity data set.

laGP is a parallelized version of the approximate Gaussian Process algorithm. Based on the molecular descriptors and PubchemFP fingerprint, the predictive power
Fig. 3 Distributions of eight studied properties
of the laGP models ($q^2_{\text{ext}} = 0.605$ or 0.614) was slightly better than that of the $k$NN models ($q^2_{\text{ext}} = 0.585$ or 0.602) while slightly worse than that of the SVM models ($q^2_{\text{ext}} = 0.606$ or 0.627). However, based on the molecular descriptors and SubFP fingerprint, the predictive power of the laGP models ($q^2_{\text{ext}} = 0.634$ or 0.635) was slightly worse than that of the $k$NN models ($q^2_{\text{ext}} = 0.636$ or 0.642) while slightly better than that of the SVM models ($q^2_{\text{ext}} = 0.617$ or 0.638). Therefore, overall, laGP, $k$NN and SVM performed similarly to this specific toxicity endpoint.

The RVM method is quite similar to the SVM algorithm in many aspects, but it can provide a fully probabilistic output. However, up to now, little information on RVM applications in QSAR modeling has been reported in the literature. According to the data shown in Tables 2, 3, 4, and 5, we observed that the RVM models ($q^2_{\text{ext}} = 0.640$ or 0.659) were obviously better than the SVM models ($q^2_{\text{ext}} = 0.606$ or 0.638). Moreover, we found that the RVM modeling was more computationally efficient than the SVM modeling because RVM did not need to estimate the error/margin tradeoff parameter $C$, which might reduce the computational cost. Due to better prediction accuracy and higher computational efficiency compared with SVM, we believed that RVM should have a promising potential for the practical use in QSAR modeling in the future.

The AD coverages for the established models were summarized in Tables 2, 3, 4, and 5. The $k$NN models always showed the smallest AD coverage for the test set. Compared with the other models, the MPLE and RF models showed relatively larger AD coverages, but the RF models could give better predictions to the test set than the MPLE models. Therefore, according to the $q^2_{\text{ext}}$ and AD coverage, the RF models would give the best predictions for this data set.

In this study, two well-defined substructural fingerprints (SubFP and PubchemFP) were used. According to the predictions to the test set, the models based on the

### Table 2 Statistical results for the QSAR models based on 120 descriptors and Pubchem fingerprints for the test set

|                          | $R^2_{\text{adj}}$ | $q^2$ | $q^2_{\text{ext}}$ | RMSE_{train} | MAE_{train} | RMSE_{test} | MAE_{test} | AD coverage (%) |
|--------------------------|--------------------|-------|-------------------|--------------|-------------|-------------|------------|-----------------|
| $k$NN                    | 0.783              | 0.774 | 0.602             | 0.413        | 0.299       | 0.707       | 0.398      | 51.4            |
| RF                       | 0.949              | 0.922 | 0.639             | 0.242        | 0.171       | 0.707       | 0.544      | 81.7            |
| SVM                      | 0.923              | 0.915 | 0.627             | 0.253        | 0.119       | 0.688       | 0.507      | 58.6            |
| RVM                      | 0.936              | 0.935 | 0.644             | 0.221        | 0.172       | 0.680       | 0.511      | 62.9            |
| laGP                     | 0.775              | 0.756 | 0.614             | 0.430        | 0.322       | 0.713       | 0.550      | 72.2            |
| MPLE                     | 0.716              | 0.693 | 0.580             | 0.482        | 0.349       | 0.743       | 0.572      | 78.4            |
| XGBoost                  | 0.920              | 0.903 | 0.624             | 0.271        | 0.205       | 0.700       | 0.533      | 74.5            |
| Consensus                | 0.923              | NA    | 0.676             | 0.278        | 0.208       | 0.666       | 0.504      | 71.7            |
| Consensus (Except MPLE)  | 0.933              | NA    | 0.678             | 0.257        | 0.194       | 0.661       | 0.499      | 68.9            |
SubFP fingerprint (Tables 4, 5) were better than those based on the PubchemFP fingerprint (Tables 2, 3). It is possible that some fragments in SubFP were more closely related to acute toxicity than those in PubchemFP.

**Accurate prediction of rat oral acute toxicity by consensus modeling**

The statistical results showed that the theoretical models using different machine learning methods have different prediction capability and model uncertainty. A useful way to reduce the model uncertainty is consensus modeling by averaging the outputs from multiple models [69–71]. Since the consensus prediction is made based on multiple different but comparable QSAR models, it may be capable of capturing the relationship between the chemical structures of the molecules and the endpoint more efficiently than a single model. Here, four consensus models were first developed by simply averaging the
predictions for the test set given by the individual models shown in Tables 2, 3, 4, and 5. All the contributions of the individual models were equal, and therefore we could avoid the limitation or overemphasis of any machine learning approach. The statistical results clearly illustrated that the consensus models had higher predictive accuracy ($q^2_{\text{ext}} = 0.669–0.689$) than any individual model. In addition, by comparing the MAEs given by the consensus versus individual models using the Wilcoxon test, we found that the improvement of the consensus models compared with all individual models was statistically significant ($p < 0.01$).

Because the predictions given by the MPLE models were significantly worse than those given by the other models, four consensus models were then developed without considering the predictions given by the MPLE models. As shown in Tables 2, 3, 4, and 5, all the consensus models without the MPLE predictions showed obvious performance improvement to the training set and slight performance improvement to the test set. The scatter plot of the experimental pLD50 values versus the predicted values given by the consensus model without the MPLE predictions (Table 5) for the training and test sets was shown in Fig. 5.

**Analysis of molecules with large prediction errors**

As mentioned above, most prediction models had good capability for the test set, but some molecules in the test set could not be well predicted by any model or even by all models. If MAE > 1.0 was used as the criteria, the MAE of chemicals with large prediction errors given by all individual models in Table 5 ranged from 1.002 to 3.486 for the test set. In total, 575 molecules could not be well predicted by any individual model in Table 5, and 249 molecules could not be well predicted by the best consensus model in Table 5. For these 249 molecules with large prediction errors, the average experimental pLD50 value was 3.321, which was obviously higher than that of the molecules in the training set (2.558). Therefore, the prediction for the molecules with higher pLD50 values are worse than those for the molecules with lower pLD50 values.

The 20 molecules in the test set with the largest prediction errors by using the consensus model in Table 5 were shown in Table 6. Some of them (molecules 9, 13 and 17) had complicated structures, some of them (molecules 12, 18 and 19) have few analogues in the data set, and some of them (molecules 3, 4, 5 and 8) have phosphate groups with severe toxicity. Then, the scaffolds of the 249 molecules with large prediction errors were generated and analyzed. For the scaffolds with frequency equal or larger than 2, their numbers present in the training and test sets were counted. The scaffolds and the associated molecules in the test sets with the largest prediction errors were examined, and the representative scaffolds were summarized in Table 7. It could be observed that these scaffolds were not abundant in the data set. For example, the number of the molecules with fragment 1 in the test set was 3, and that in the training set was only 3; the number of the molecules with fragment 2 in the test set was 5, and that in the training set was only 1; the number of the molecules with fragment 10 in the test set was 3, and that in the training set was even 0. Apparently, for these scaffolds shown in Table 7, the associated molecules in the training set were quite limited and thus the established model could not give

---

**Fig. 5** Scatter plot of the experimental pLD50 values versus the predicted values for the molecules in the (a) training and (b) test sets given by the consensus model without the MPLE predictions in Table 5.
Table 6 Experimental and predicted LD$_{50}$ values for the 20 tested molecules with the largest prediction errors

| No. | Structure | Exp. $^a$ | Fingerprints | kNN | RF | SVM | RVM | LaGP | MPLG | XGBoost | Cons. $^b$ |
|-----|-----------|----------|--------------|-----|----|-----|-----|------|------|---------|-----------|
| 1   | ![Structure](image1) | 5.957    | PubchemFP    | 3.433 | 3.349 | 2.772 | 3.177 | 2.962 | 2.889 | 2.780 | 3.052 |
|     |           |          | SubFP        | 2.286 | 2.614 | 2.756 | 2.534 | 2.614 | 2.566 | 2.533 | 2.558 |
| 2   | ![Structure](image2) | 5.513    | PubchemFP    | 2.590 | 2.529 | 2.461 | 2.694 | 2.514 | 2.743 | 2.638 | 2.596 |
|     |           |          | SubFP        | 2.648 | 2.574 | 2.912 | 2.700 | 2.787 | 2.551 | 2.852 | 2.718 |
| 3   | ![Structure](image3) | 5.658    | PubchemFP    | 3.807 | 2.838 | 3.307 | 3.301 | 3.742 | 2.948 | 2.869 | 3.259 |
|     |           |          | SubFP        | 3.340 | 2.777 | 3.070 | 2.848 | 2.867 | 2.733 | 3.008 | 2.949 |
| 4   | ![Structure](image4) | 5.406    | PubchemFP    | 1.819 | 2.830 | 2.371 | 2.727 | 2.481 | 2.801 | 3.047 | 2.582 |
|     |           |          | SubFP        | 2.976 | 2.609 | 2.839 | 2.637 | 2.736 | 2.755 | 2.662 | 2.745 |
| 5   | ![Structure](image5) | 5.446    | PubchemFP    | 3.057 | 2.833 | 2.683 | 2.789 | 2.764 | 2.828 | 2.991 | 2.849 |
|     |           |          | SubFP        | 3.756 | 2.974 | 2.897 | 2.775 | 2.842 | 2.972 | 2.617 | 2.976 |
| 6   | ![Structure](image6) | 5.310    | PubchemFP    | 2.588 | 2.595 | 2.589 | 2.861 | 1.843 | 2.817 | 2.564 | 2.551 |
|     |           |          | SubFP        | 2.536 | 2.609 | 3.427 | 3.001 | 2.943 | 2.912 | 2.487 | 2.845 |
| 7   | ![Structure](image7) | 5.307    | PubchemFP    | 4.251 | 3.110 | 2.705 | 2.941 | 3.333 | 2.797 | 2.692 | 3.118 |
|     |           |          | SubFP        | 3.171 | 3.319 | 3.094 | 2.487 | 2.706 | 2.568 | 3.074 | 2.917 |
| 8   | ![Structure](image8) | 6.402    | PubchemFP    | 3.136 | 3.148 | 2.672 | 2.915 | 3.189 | 3.726 | 2.868 | 3.093 |
|     |           |          | SubFP        | 4.263 | 3.692 | 4.778 | 4.332 | 4.084 | 4.073 | 3.292 | 4.073 |
| 9   | ![Structure](image9) | 6.159    | PubchemFP    | 3.510 | 3.481 | 2.882 | 3.666 | 3.407 | 3.075 | 3.528 | 3.364 |
|     |           |          | SubFP        | 3.693 | 3.885 | 4.097 | 3.825 | 3.771 | 3.496 | 4.417 | 3.883 |
| 10  | ![Structure](image10) | 5.170    | PubchemFP    | 2.641 | 2.758 | 3.202 | 3.004 | 3.119 | 2.816 | 2.821 | 2.909 |
|     |           |          | SubFP        | 2.944 | 2.811 | 3.041 | 2.976 | 2.964 | 2.757 | 2.999 | 2.927 |
### Table 6 continued

| No. | Structure | Exp. \(^a\) | Fingerprints | kNN | RF | SVM | RVM | laGP | MPEL | XGBoost | Cons. \(^b\) |
|-----|-----------|-------------|--------------|-----|----|-----|-----|------|------|---------|-------------|
| 11  | ![Structure 11](image1) | 4.019 | PubchemFP | 1.302 | 2.003 | 2.101 | 2.018 | 1.876 | 2.123 | 1.960 | 1.912 |
|     |           |            | SubFP       | 1.326 | 1.871 | 1.944 | 1.860 | 2.030 | 1.955 | 1.867 | 1.836 |
| 12  | ![Structure 12](image2) | 4.780 | PubchemFP | 2.685 | 2.642 | 2.686 | 2.598 | 2.793 | 2.529 | 2.517 | 2.636 |
|     |           |            | SubFP       | 2.666 | 2.548 | 2.980 | 2.663 | 2.640 | 2.454 | 2.540 | 2.642 |
| 13  | ![Structure 13](image3) | 4.762 | PubchemFP | 2.221 | 2.610 | 2.278 | 2.255 | 2.129 | 2.719 | 2.606 | 2.403 |
|     |           |            | SubFP       | 2.861 | 2.498 | 2.692 | 2.854 | 2.758 | 2.642 | 2.563 | 2.695 |
| 14  | ![Structure 14](image4) | 4.538 | PubchemFP | 2.284 | 2.576 | 2.483 | 2.533 | 2.575 | 2.921 | 2.165 | 2.505 |
|     |           |            | SubFP       | 2.055 | 2.570 | 2.747 | 2.492 | 2.433 | 2.818 | 2.512 | 2.518 |
| 15  | ![Structure 15](image5) | 5.006 | PubchemFP | 2.336 | 3.003 | 2.825 | 2.895 | 3.046 | 3.151 | 2.701 | 2.851 |
|     |           |            | SubFP       | 2.889 | 2.945 | 3.324 | 2.832 | 2.789 | 3.357 | 2.920 | 3.008 |
| 16  | ![Structure 16](image6) | 5.225 | PubchemFP | 3.292 | 3.740 | 3.101 | 3.310 | 3.369 | 3.088 | 3.369 | 3.324 |
|     |           |            | SubFP       | 3.580 | 3.144 | 3.396 | 3.200 | 3.322 | 2.991 | 3.153 | 3.255 |
| 17  | ![Structure 17](image7) | 1.740 | PubchemFP | 3.764 | 3.228 | 2.744 | 3.348 | 3.435 | 2.901 | 3.180 | 3.229 |
|     |           |            | SubFP       | 4.029 | 3.657 | 3.312 | 3.659 | 4.052 | 2.756 | 4.250 | 3.674 |
| 18  | ![Structure 18](image8) | 2.140 | PubchemFP | 4.421 | 3.509 | 3.397 | 3.555 | 3.886 | 3.466 | 3.482 | 3.674 |
|     |           |            | SubFP       | 4.914 | 3.898 | 3.355 | 3.919 | 4.375 | 3.246 | 3.606 | 3.902 |
| 19  | ![Structure 19](image9) | 0.291 | PubchemFP | 2.156 | 2.179 | 1.847 | 1.606 | 2.095 | 2.113 | 2.216 | 2.030 |
|     |           |            | SubFP       | 1.923 | 2.230 | 1.509 | 1.760 | 2.088 | 2.053 | 2.192 | 1.965 |
| 20  | ![Structure 20](image10) | 1.163 | PubchemFP | 3.159 | 2.766 | 2.636 | 2.635 | 3.018 | 2.793 | 1.973 | 2.711 |
|     |           |            | SubFP       | 3.743 | 2.757 | 2.547 | 2.829 | 3.033 | 2.669 | 2.047 | 2.804 |

\(^a\) Experimental LD\(_{50}\)

\(^b\) Consensus prediction
Analysis of important descriptors and fragments given by RVM regression models

One-dimensional sensitivity analysis was employed to evaluate the importance of the molecular descriptors and fragments for QSAR modeling, and the important descriptors were summarized in Table 8 [72]. The rings descriptor in the RF model had the highest sensitivity (0.075), and a_ICM, E_ele, vsa_pol, opr_brigid, E_nb, dipole, logS, MW, SlogP and vdw_vol were also important, indicated by relatively high sensitivity. The KierFlex descriptor in the RVM model have the highest sensitivity (0.028), and a_nF, MNDO_dipole, pmi, E_stb, E_oob, wienerPath, vsa_pol, a_acc and MW were also important, indicated by relatively high sensitivity. After examining the molecules descriptors shown in Table 8, we found that the molecular descriptors related to molecular polarity, van der Waals surface, molecular flexibility, partial charge distribution and solubility might have more contributions than the other descriptors. Furthermore, the numbers of the descriptors related to frontier molecular orbitals were relatively high, suggesting...
that rat oral acute toxicity might be related to molecular reactivity and intra-molecular interactions.

Moreover, after examining importance of the substructure fingerprints in the RVM models described in the Tables 3 and 5, we found that that 9 PubchemFP fragments and 4 SubFP fragments were responsible for rat oral acute toxicity. The $R_{\text{adj}}^2$ change in the stepwise regression and Cramer’s V coefficient were used to evaluate the importance of the fragments [23, 73]. The numbers of the molecules with each fragment were counted. If the number of the molecules with $\text{pLD}_{50} \geq 3$ was more than that with $\text{pLD}_{50} < 3$, this fragment was considered to have positive contribution to high $\text{pLD}_{50}$ and vice versa. Four PubchemFP fragments had negative contributions, and they were PubchemFP15 (the counts of nitrogen atoms $\geq 2$), PubchemFP442 (N-ethylimino group), PubchemFP418 (carbon–nitrogen double bond) and PubchemFP14 (the counts of nitrogen atoms $\geq 1$). Meanwhile, five PubchemFP fragments gave positive contributions, and they were PubchemFP400 (aromatic C–NH–C bond), PubchemFP359 (aromatic 1,3-diazacyclo group with 2-carbonous substituent), PubchemFP770 (ortho aryl nitrogen) PubchemFP833 (ortho alicyclic nitrogen) and PubchemFP527 (aromatic C–C–NH bond). On the other hand, all the four SubFP fragments had positive contributions, and they were trifluoromethyl, alkylfluoride, hetero N basic H and heterocyclic. The structures of these representative fragments were summarized in Tables 9 and 10.

### Conclusions

In this study, on the basis of a comprehensive data set of rat oral acute toxicity, the relationships between eight important molecular properties and acute toxicity QSAR modeling.

| Molecular descriptors | Number of descriptors | 120 (Descriptor + PubchemFP) | 120 (Descriptor + SubFP) | 150 (Descriptor + PubchemFP) | 150 (Descriptor + SubFP) |
|-----------------------|-----------------------|-------------------------------|----------------------------|-------------------------------|----------------------------|
| 2D                    |                       |                               |                             |                               |                             |
| Physical properties   | 6                     | 7                             | 7                           | 7                             |                             |
| Subdivided surface areas | 8                     | 9                             | 10                          | 11                            |                             |
| Atom counts and bond counts | 10                    | 10                            | 10                          | 11                            |                             |
| Kier&Hall connectivity and kappa shape indices | 7 | 7 | 8 | 8 |
| Adjacency and distance matrix descriptors | 11 | 10 | 13 | 14 |
| Pharmacophore feature descriptors | 4 | 4 | 5 | 5 |
| Partial charge descriptors | 19 | 20 | 25 | 27 |
| 3D                    |                       |                               |                             |                               |                             |
| Potential energy descriptors | 2 | 1 | 5 | 4 |
| Mopac descriptors | 15                     | 15                            | 15                          | 15                            |                             |
| Surface area, volume and shape descriptors | 30 | 30 | 37 | 38 |
| Conformation dependent charge descriptors | 4 | 5 | 6 | 6 |
| Fingerprints (PubchemFP) | 4 | - | 9 | - |
| Fingerprints (SubFP) | -                      | 2                             | -                           | 4                             |                             |

The PubchemFP fragments found in the models are relatively small, but they might be important components for toxicophores that were not defined in the fingerprint dictionary. In the SubFP fragment alerts, trifluoromethyl and alkylfluoride were often constituent parts of toxic substances, but hetero N and heterocycle might be only the background noise of models, or they may be parts of some toxic substructures not defined in the fingerprint dictionary. As been mentioned in the previous literature [14, 74, 75], some toxic chemicals contained trifluoromethyl and alkylfluoride fragments such as 2-(trifluoromethyl)-benzimidazole, which were not defined in the fingerprint dictionary and were substructures of many antitumor drugs, antibiotics, antiparasitics and ionic liquids [76–80]. In addition, some important substructures in toxicophores, such as organophosphates, organochlorines and norbornene derivates, did not exist in the PubchemFP dictionary. The phosphonic groups could be found in the SubFP dictionary, but they were only found in limited molecules and therefore disappeared through dimension reduction. Our calculations suggested that more specific and diverse fingerprints were essential and important for toxicity QSAR modeling.
toxicity were examined. We observed that rat oral toxicity could not be reliably predicted by a single or several molecular properties. Then, seven machine learning approaches were used to establish the QSAR models for oral acute toxicity. Considering the overall prediction accuracy for the test set, the RF and RVM methods outperformed the others. The consensus model by integrating the outputs from multiple individual models demonstrated better predictivity ($q^2_{\text{ext}} = 0.669–0.689$) than any individual model for the test set. Our study also demonstrated that QSAR modeling based on structure fingerprints could afford potential important substructural fragments as toxicity alerts, but a proper and enough large fingerprint

Table 9 Nine PubchemFP fragment alerts and representative structures

| No. | Fingerprint | Fragment | Description | Bit substructure | $R^2_{\text{adj}}$ change | Cramer’s V | Representative structure |
|-----|-------------|----------|-------------|------------------|---------------------------|------------|--------------------------|
| **Positive fragment alerts** | | | | | | | |
| 1 | PubchemFP400 | | Detailed atom neighborhoods | N(~ H)(C)(C) | 0.00086 | 0.15810 | |
| 2 | PubchemFP359 | | Simple atom nearest neighbors | C(~ C)(N)(N) | 0.00162 | 0.15641 | |
| 4 | PubchemFP770 | | Complex SMARTS patterns | Nc1c(N)cccc1 | 0.00167 | 0.14722 | |
| 5 | PubchemFP833 | | Complex SMARTS patterns | Nc1c(N)cccc1 | 0.00100 | 0.14504 | |
| 6 | PubchemFP527 | | Simple SMARTS patterns | C:C:N-[#1] | 0.00036 | 0.13989 | |
| **Negative fragment alerts** | | | | | | | |
| 1 | PubchemFP15 | Counts of N ≥ 2 | Hierarchic element counts | ≥2 N | 0.00174 | 0.16108 | |
| 2 | PubchemFP442 | | Detailed atom neighborhoods | C(~ C)(≡N) | 0.00024 | 0.15474 | |
| 3 | PubchemFP418 | | Simple SMARTS patterns | C≡N | 0.00094 | 0.13903 | |
| | PubchemFP14 | Counts of N ≥ 1 | Hierarchic element counts | ≥1 N | 0.00002 | 0.13650 | |
dictionary should be adopted. By scaffold analysis, we found that quite limited numbers of molecules with certain scaffolds in the training set would reduce the prediction accuracy of the models. According to the results of this study, we believed that the successful modeling methods used here could be employed for other toxicity endpoints.

Authors’ contributions
TL and TH conceived and designed the experiments. TL and DL performed the simulations. TL, YS, DL, HS and YL analyzed the data. TL, HS, YL and TH wrote the manuscript. All authors read and approved the manuscript.

Author details
1 College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, People’s Republic of China. 2 State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, People’s Republic of China. 3 Institute of Functional Nano and Soft Materials (FUNSOM), Soochow University, Suzhou 215123, Jiangsu, People’s Republic of China. 4 Department of Medicinal Chemistry, School of Pharmacy, Second Military Medical University, Shanghai 200433, People’s Republic of China.

Acknowledgements
This study was supported by the National Science Foundation of China (21575128, 81302679), the ‘Construction of Shanghai Municipal NCB medical rescue system’ Project of Shanghai Municipal Commission of Health and Family Planning, and the Special Program for National Basic Work on Science and Technology (2015FY111400) of Ministry of Science and Technology of China. We would like to thank Dr. Hao Zhu and Alexander Tropsha for valuable dataset of rat oral LD50.

Competing interests
The authors declare that they have no competing interests.

Received: 5 November 2015 Accepted: 20 January 2016 Published online: 01 February 2016

Table 10 Four SubFP fragment alerts and representative structures

| No. | Fingerprint | Fragment | Description | SMILES | $R^2_{adj}$ | Cramer’s V |
|-----|-------------|----------|-------------|--------|-------------|------------|
| 1   | SubFP294    | F        | Trifluoromethyl | [F$X1$][C$X4$][S($H0$)[C,Br,I]]$S(F(F))[F(F)]$([F$X1$])[F$X1$]] | 0.00173 | 0.15737 |
| 2   | SubFP9      | R$_1$    | Alkylfluoride | [F$X1$][C$X4$] | 0.00024 | 0.15386 |
| 3   | SubFP179    | R$_2$    | Hetero N basic H | [n$X3H1+0$] | 0.00161 | 0.14669 |
| 4   | SubFP275    | C        | Heterocyclic | [n$6$] | 0.00002 | 0.14306 |

References
1. Parasuraman S (2011) Toxicological screening. J Pharmacol Pharmacother 2(2):74
2. Nicolotti O, Benfenati E, Carotti A, Gadaleta D, Gissi A, Mangiatordi GF, Novellino E (2014) REACH and in silico methods: an attractive opportunity for medicinal chemists. Drug Discov Today 19(11):1757–1768
3. Benz RD (2007) Toxicological and clinical computational analysis and the US FDA/CDER. Expert Opin Drug Metab Toxicol 3(1):109–124
4. Creton S, Dewhurst IC, Earl JK, Gehen SC, Guest RL, Hotchkiss JA, Indans J, Woolfizer MP, Billington R (2009) Acute toxicity testing of chemicals—opportunities to avoid redundant testing and use alternative approaches. Crit Rev Toxicol 40(1):50–83
5. Cheng F, Li W, Liu G, Tang Y (2013) In silico ADMET prediction: recent advances, current challenges and future trends. Curr Top Med Chem 13(11):1273–1289
6. Benz RD (2007) Computational toxicology—a tool for early safety evaluation. Drug Discov Today 15(1–2):16–22
7. Kruhlak NL, Benz RD, Zhou H, Colatsky TJ (2012) (Q)SAR modeling and safety assessment in regulatory review. Clin Pharmacol Ther 91(3):529–534
8. Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K (2014) Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. Chem Res Toxicol 27(10):1643–1651
9. Diaz RG, Manganelli S, Esposito A, Roncaglioni A, Manganaro A, Benfenati E (2015) Comparison of in silico tools for evaluating rat oral acute toxicity. SAR QSAR Environ Res 26(1):1–27
10. Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A (2009) Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. Chem Res Toxicol 22(12):1913–1921
11. Xu C, Cheng F, Chen L, Du Z, Liu W, Lee PW, Tang Y (2012) In silico prediction of chemical Ames mutagenicity. J Chem Inf Model 52(11):2840–2847
12. Zang Q, Retoff DM, Judson RS (2013) Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. J Chem Inf Model 53(12):3244–3261
13. Raveisky OA, Grigor’Ev VJ, Modina EA, Worth AP (2010) Prediction of acute toxicity to mice by the Arithmetic Mean Toxicity (AMT) modelling approach. SAR QSAR Environ Res 21(3–4):265–275
63. Funar-Timofei S, Ionescu D, Suzuki T (2010) A tentative quantitative structure-toxicity relationship study of benzodiazepine drugs. Toxicol In Vitro 24(1):184–200
64. Zhu J, Wang J, Yu H, Li Y, Hou T (2011) Recent developments of in silico predictions of oral bioavailability. Comb Chem High Throughput Screen 14(5):362–374
65. Hou T, Li Y, Zhang W, Wang J (2009) Recent developments of in silico predictions of intestinal absorption and oral bioavailability. Comb Chem High Throughput Screen 12(5):497–506
66. Chen B, Sheridan RP, Hornak V, Voigt JH (2012) Comparison of random forest and Pipeline Pilot Naive Bayes in prospective QSAR predictions. J Chem Inf Model 52(3):792–803
67. Eklund M, Norinder U, Boyer S, Carlsson L (2014) Choosing feature selection and learning algorithms in QSAR. J Chem Inf Model 54(3):837–843
68. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20(3):318–331
69. Lei B, Li J, Yao X (2013) A novel strategy of structural similarity based consensus modeling. Mol Inform 32(7):599–608
70. Lei B, Xi L, Li J, Liu H, Yao X (2009) Global, local and novel consensus quantitative structure-activity relationship studies of 4-(phenylamino)methylene isoquinoline-1, 3 (2H, 4H)-diones as potent inhibitors of the cyclin-dependent kinase 4. Anal Chim Acta 644(1):17–24
71. Li J, Lei B, Liu H, Li S, Yao X, Liu M, Gramatica P (2008) QSAR study of malonyl-CoA decarboxylase inhibitors using GA-MLR and a new strategy of consensus modeling. J Comput Chem 29(16):2636–2647
72. Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. Inform Sci (N Y) 225:1–17
73. Oh DS, Troester MA, Usary J, Hu Z, He X, Fan C, Wu J, Carey LA, Perou CM (2006) Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol 24(1):1656–1664
74. Li X, Chen L, Cheng F, Wu Z, Bao H, Xu C, Li W, Liu G, Shen X, Tang Y (2014) In silico prediction of chemical acute oral toxicity using multi-classification methods. J Chem Inf Model 54(4):1061–1069
75. Bhattachar B, Gramatica P (2011) Oral LD50 toxicity modeling and prediction of per- and polyfluorinated chemicals on rat and mouse. Mol Divers 15(2):467–476
76. Andrzejewska M, Yepez-Mulia L, Cedillo-Rivera R, Tapia A, Vilpo L, Vilpo J, Kazimierczuk Z (2002) Synthesis, antiprotozoal and anticancer activity of substituted 2-trifluoromethyl- and 2-pentafluoroethylbenzimidazoles. Eur J Med Chem 37(12):973–978
77. Kazimierczuk Z, Andrzejewska M, Kaustova J, Klimesova V (2005) Synthesis and antimycobacterial activity of 2-substituted halogenobenzimidazoles. Eur J Med Chem 40(2):203–208
78. Navarrete-Vazquez G, Rojano-Vilchis MM, Yepez-Mulia L, Melendez V, Genena I, Hernandez-Campos A, Castillo R, Hernandez-Luis F (2006) Synthesis and antiprotozoal activity of some 2-(trifluoromethyl)-1H-benzimidazole bioisosteres. J Med Chem 49(1):135–141
79. Perez-Villanueva J, Santos R, Hernandez-Campos A, Giulianiotti MA, Castillo R, Medina-Franco JL (2011) Structure–activity relationships of benzimidazole derivatives as antiparasitic agents: dual-activity difference (DAD) maps. MedChemComm 2(1):44–49
80. Paterno A, D’Anna F, Musumarrara G, Noto R, Scire S (2014) A multivariate insight into ionic liquids toxicities. RSC Adv 4(46):23985–24000