OXFORD Toxicology Research

Toxicology Research, 2020, **9**, 164–172

doi: 10.1093/toxres/tfaa016
Advance Access Publication Date: 29 April 2020
Paper

PAPER

In silico prediction of chemical neurotoxicity using machine learning

Changsheng Jiang, Piaopiao Zhao, Weihua Li, Yun Tang* and Guixia Liu*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Rd, Xuhui District, Shanghai 200237, China

Correspondence address. Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Rd, Xuhui District, Shanghai 200237, China. Tel: +86-21-64250811; Fax: +86-21-64251033; E-mail: ytang234@ecust.edu.cn, gxliu@ecust.edu.cn

Abstract

Neurotoxicity is one of the main causes of drug withdrawal, and the biological experimental methods of detecting neurotoxic toxicity are time-consuming and laborious. In addition, the existing computational prediction models of neurotoxicity still have some shortcomings. In response to these shortcomings, we collected a large number of data set of neurotoxicity and used PyBioMed molecular descriptors and eight machine learning algorithms to construct regression prediction models of chemical neurotoxicity. Through the cross-validation and test set validation of the models, it was found that the extra-trees regressor model had the best predictive effect on neurotoxicity ($q^2_{test} = 0.784$). In addition, we get the applicability domain of the models by calculating the standard deviation distance and the lever distance of the training set. We also found that some molecular descriptors are closely related to neurotoxicity by calculating the contribution of the molecular descriptors to the models. Considering the accuracy of the regression models, we recommend using the extra-trees regressor model to predict the chemical autonomic neurotoxicity.

Key words: computational toxicology, machine learning, neurotoxicity, applicability domain

Introduction

According to the survey, the main reasons for the failure of clinical drugs into market are the safety and efficacy [1], which makes the cost of the expensive new drug development more expensive. In addition, some listed drugs still have the risk of withdrawal due to some side effects or adverse reactions. According to recent research reports [2], neurotoxicity is one of the main causes of drug withdrawal. Therefore, a risk assessment of the neurotoxicity of a drug or compound is necessary.

Drugs cause neurotoxicity mainly by affecting mitochondrial respiration, immunemediated response and inhibition of neurogenic activity [3–5]. Drugs with neurotoxicity can be broadly classified into three categories, including antibacterials, antifungals and antidepressants. For example, the antibacterial drug hexachlorophene [3] can cause the white matter edema and cavernous changes, resulting in neurotoxicity. The antifungal drug clioquinol can cause subacute bone marrow optic neuropathy [5]. The antidepressant zimelidine can cause Guillain-Barre syndrome [6], and its clinical side effect is acute sympathetic slow limb paralysis. All of these withdrawal drugs caused by neurotoxicity can be found in the WITHDRAWN database [2].

In addition to the neurotoxicity caused by drugs, there are more compounds on the industrial market that can cause neurotoxicity. For many industrial countries, neurotoxicity is a relatively prevalent occupational disease and environmental problem [7], which has attracted the attention of the countries concerned. Workers from industries such as the heavy metal smelting industry [8], pesticides [9] and the pharmaceutical industry [10] are vulnerable to neurotoxic compounds. Among many compounds that cause occupational exposure to

Received: 1 November 2019; **Revised:** 1 March 2020; **Accepted:** 18 March 2020

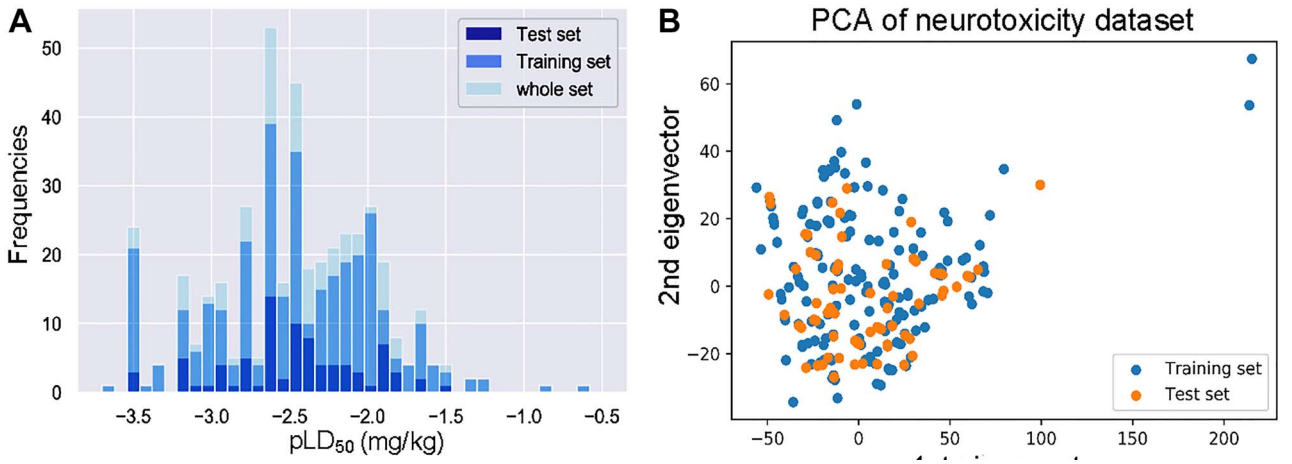

Figure 1: (**A**) Distributions of the experimental pLD_{50} values for the whole data set ($n = 422$, light blue bars), training set ($n = 339$, royal blue bars) and test set ($n = 83$, navy bule bars); (**B**) the PCA plot of the training set and test set.

neurological diseases, dimethylmercury and diethylmercury are the most common neurotoxins [11]. In addition, neurotoxins have been used as chemical weapons, exacerbating the dangers of war to the lives of people. One of the most typical chemical weapons is Tabun [12]. It is an organophosphorus pesticide insecticide whose insecticidal principle is to significantly inhibit the activity of acetylcholinesterase resulting in a large accumulation of acetylcholine, which leads to central and peripheral cholinergic neurological dysfunction. Through a series of efforts, neurotoxic substances have been included in a public database by the German Occupational Safety Supervision Agency [13].

Whether the neurotoxicity is caused by drugs or caused by industrial compounds, it poses great health risks and harms to society. The current standard test rules for testing the neurotoxicity of compounds are OECD TG 418, TG 419 and TG 424. Although these methods for testing the toxicity of compounds can provide very detailed and valuable toxicity information, they are still long tests and expensive tests. They also need to sacrifice a lot of test animals and exist other deficiencies. Moreover, these defects are contrary to the 3R principle advocated by REACH [14], and cannot meet the demand for chemical toxicity testing brought by the rapid increase in the number of industrial compounds. Therefore, there is a need to develop more rapid and convenient toxicity testing methods to predict the neurotoxicity of chemicals. REACH advocates the use of machine learning-based methods such as QSAR to predict the neurotoxicity, reproductive toxicity and carcinogenicity of chemicals. The QSAR prediction models of neurotoxicity that have been reported so far are extremely rare [15–18]. Although some QSAR models can provide accurate predictions of neurotoxicity, these QSAR models generally have disadvantages, such as a small data set of chemicals and sparse chemical species. To some extent, the scope of application of these models is limited. Therefore, in order to develop a more reliable neurotoxicity prediction model with larger applicability domains, it is necessary to collect more data sets with larger diversity of chemicals to construct prediction models of neurotoxicity.

In this study, we collected a large number of highly diverse data sets (422 compounds) for constructing regression prediction models of chemical neurotoxicity. After descriptor calculation and descriptor screening, eight machine learning methods were applied to construct regression prediction models to predict the neurotoxicity caused by intraperitoneal injection in mice. The eight machine learning methods include bagging regressor (BGR), extra-trees regressor (ETR), Gaussian process regression (GPR), k- nearest neighbors regression (KNN), multi-layer perceptron regressor (MLPR), Nu- support vector regression (Nu-SVR), random forest regressor (RF) and epsilon-support vector regression (SVR). The reliability of the model is evaluated by internal validation and external validation. In addition, the applicability domain of the model was evaluated and compounds with large prediction errors were analyzed.

Materials and Methods

Data collection and preprocess

Because of the sensitivity to the neurotoxic chemicals and the convenience for the experiment observation, mice are often used as animal models for testing neurotoxic chemicals. Therefore, in this study, we collected a mouse neurotoxicity data set of 495 chemicals from ChemIDplus (https://chem.nlm.nih.gov/chemidplus/chemidlite.jsp), a sub-database of TOXNET official toxicity data. The following keywords were used to search on the ChemIDplus database: the test species were mouse, the route of administration was intraperitoneal administration, the test index was LD_{50} and the effect type was autonomic nervous system toxicity. The LD_{50} values of these compounds represent that half of the test mice died at this dose due to neurotoxicity. The inorganic compounds, metalorganic compounds, mixtures, salts and duplicated compounds were eliminated by using PipeLine Pilot software, and finally the SMILES of the compounds were canonicalized. Finally, 422 organic compounds (Supplementary Table S1 in the supporting information) were obtained and divided into training set and test set in the ratio of 8:2 by using Discovery Studio software. Among these chemicals, there are two drugs including Promazine and Propranolol, and the other chemicals are not drugs. Besides, the chemicals consist of seven types of substances in which the substructures piperazine, carbamate carbanilate, propyltrimethylene ester, propanediol, phospholene oxide, 2-propanol and ammonium were separately included. The data distribution are shown in Fig. 1. Regression prediction models of autonomic neurotoxicity induced by intraperitoneal administration of mice will be constructed. The experimental values LD_{50} were converted to pLD_{50} values before model development.

Table 1: Optimal parameters identified by the grid search for QSAR modeling

Models	Optimal parameters
BGR	n_estimators = 300
ETR	n_estimators = 300, random_state = 14
GPR	Kernel = RationalQuadratic (alpha = 1) + WhiteKernel (noise_level = 10)
KNN	n_neighbors = 5, leaf_size = 30, p = 2, weights = 'distance'
MLPR	Alpha = 0.001, hidden_layer_sizes = 250, max_iter = 10000
NuSVR	Nu = 0.5, C = 15, gamma = 0.25
RF	n_estimators = 200, random_state = 100
SVR	C = 15, gamma = 0.2

BGR, bagging regressor; ETR, extra-trees regressor; GPR, Gaussian process regression; KNN, k-nearest neighbors regression; MLPR, multi-layer perceptron regressor; NuSVR, Nu support vector regression; RF, random forest regressor; SVR, epsilon-support vector regression.

Calculation of molecular descriptors

For the compiled neurotoxicity data, we used PyBioMed software to calculate the molecular descriptors of the compounds. The PyBioMed software has 765 features, including constitutional descriptors, topological descriptors, kappa shape indices, connectivity indices, Burden descriptors, E-state indices, charge descriptors, Basak information indices, autocorrelation descriptors, molecular properties, MOE-type descriptors and pharmacophore descriptors.

Feature reduction

The feature reduction is a critical step for the model building. The purpose of feature reduction is to remove redundant or unrelated features without losing a lot of useful information about the compounds and affecting the predictive power of the constructed models. In the feature reduction step, we used the scikit-learn package to create a script for feature reduction. The concrete feature reduction process can be divided into five steps. (i) Remove features with eigenvalues of 0 or variance of 0, (ii) descriptors with the correlation (r) higher than the predefined threshold (0.90) to any descriptor were also removed, (iii) then scale and regularize the remaining features using the QuantileTransformer module in scikit-learn, (iv) then use the SelectPercentile module of scikit-learn for the selection of the features that are highly correlated with the experimental LD_{50}, (v) finally uses the recursive feature eliminate module in the scikit-learn to select features. The QuantileTransformer method mentioned here uses a quantile information to transform feature, it converts features into features that follow a uniform distribution or a normal distribution, so that for a given feature, this transformation tends to disperse the most common values. In addition, it also reduces the impact of outliers, so this is a powerful preprocessing scheme. This transformation will be applied independently to each feature, the cumulative distribution function of the features is used to project the original values, and the eigenvalues of the new unseen data below or above the fit range will be mapped to the boundaries of the output distribution. It is worth noting that this transformation is non-linear. It may distort the linear correlation between variables measured at the same scale, but it will make the variables measured at different scales more directly comparable. In addition, the SelectPercentile method mentioned here first determines the percentage of features retained as a percentage of the overall feature, then calculates the F score and P-value score of each feature and observation, and sorts them to select the previous set.

In the final recursive feature elimination step, the machine learning method we use is random forest. This method was originally proposed by Guyon *et al*. [19], in which the machine learning method used was SVM, and later the RF algorithm [20] was introduced for recursive feature elimination. The RFE–RF process can be described by the following steps: (i) train a random forest model, (ii) compute the permutated feature importance criterion, (iii) clear the least relevant feature, (iv) repeating steps 1–3 until there are no remaining features. After completing these processes, a subset of features that determine and give the best prediction accuracy are determined.

QSAR modeling methods

In this study, we applied eight machine learning algorithms to construct regression prediction models for chemical neurotoxicity prediction. These eight machine learning methods are BGR, ETR, GPR, KNN, MLPR, Nu-SVR, RF and SVR. The optimized parameters of these machine learning algorithms are shown in Table 1. The optimization method is realized by the grid search module in the scikit-learn package.

Validation and evaluation of QSAR models

The reliability and predictive power of the regression models are assessed by internal validation and external validation. Internal validation is achieved by leave-one-out cross-validation (LOO-CV). The external validation is to predict the neurotoxicity of the test set by constructed models, and obtain a series of predicted pLD_{50} values. The predictive power of each model was evaluated by the cross-validated R^2 coefficient (q^2) and the adjusted R^2 (R^2_{adj}). The calculation formulas of R^2_{adj} and q^2 are as follows:

$$q^2\left(R^2\right) = 1 - \frac{\mathrm{PRESS}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2} \quad (1)$$

$$R^2_{adj} = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-p-1} \quad (2)$$

Here, R^2 is the square of the Pearson correlation coefficient, p is the number of variables in the regression equation, SST is the sum of the squares of the deviation of the dependent variable value and its mean value, PRESS is the sum of the squares of the prediction residuals and n is the number of the compounds.

The traditional decision coefficient R^2 (q^2_{test}) is used to evaluate the predictive power of each model on the test set. The acceptable thresholds for q^2_{train} of the training set and q^2_{test} of the test set are both set to $\geq$0.5. When the difference between R^2_{adj} and q^2_{test} is >0.3 [21], the model is considered to be overfitting.

Besides, the Y-randomization test [22] was performed to support the reliability of our models.

Analysis of applicability domain

According to the OECD's principle for the QSAR models, the QSAR models must have a defined applicability domain [23]. Since the training set of the QSAR models cannot cover the entire chemical space, the predictive power and explanatory power of any model for unknown compounds are limited. Therefore, any model needs to predefine an applicability domain. The prediction of the model for unknown compounds distributed in the applicability domain can be considered relatively reliable, while the prediction of unknown compounds distributed outside the applicability domain is judged to be less reliable [24]. In this study, we used the standard deviation distance (SDD) and leverage distance methods to evaluate the applicability domain of each model. The formula for calculating the SDD is as follows:

$$\mathrm{SDD(i)} = \sqrt{\frac{\sum \left(y(i) - \bar{y}(i)\right)^2}{n-1}} \tag{3}$$

where i represents the ith compound, y(i) represents the quantitative predictive value of the ith compound of the model, n represents the number of compounds in the training set and the applied boundary is defined as a 3 × SDD value [25]. If a compound is outside the applicability domain, its SDD value is higher than the 3 × SDD value.

In addition, we use the Hotelling's test and related leverage statistics [26, 27] to measure the applicability domain of the model. The leverage distance h_i measures the distance between the observed independent variable and other observations. The formula for calculating the lever distance is as follows:

$$H = X(X^TX)^{-1}X^T \tag{4}$$

$$h_i = [H]_{ii} \tag{5}$$

where H is a hat matrix, X represents a matrix of multiple features of multiple compounds in the training set, X^T represents the transposed matrix of the X matrix and the matrix of $\left(X^TX\right)^{-1}$ represents the inverse matrix of $\left(X^TX\right)$, h_i is the leverage distance of the ith compound and $\left[H\right]_{ii}$ represents the diagonal value of the hat matrix H. The leverage value is between 0 to 1. The threshold h^*[28] for defining the leverage distance is generally $3p/n$, where p is the number of the training set descriptors and n is the number of training set compounds. If the leverage distance of a compound is greater than the threshold h^*, the prediction of the compound is considered to be unreliable.

After calculating these SDDs and lever distances, we also use Williams' plot to plot the distribution in two dimensions, which can visually describe the scope of an applicability domain. In addition, the effect of each compound on the model was evaluated using Cook's distance. The Cook's distance is defined as follows:

$$D_i = \frac{r_i^2}{p \times \mathrm{MSE}}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{6}$$

where r_i represents the residual of the ith compound, h_i represents the leverage distance of the ith compound, p represents the number of training set descriptors used for modeling and MSE represents the mean square error. The threshold of the Cook's distance is defined as $4/(n\text{-}p\text{-}1)$, and if the Cook's distance of the compound is greater than the threshold, it is considered to have a large interference with the model prediction ability. Analysis of all of these applicability domains is done through Python scripts.

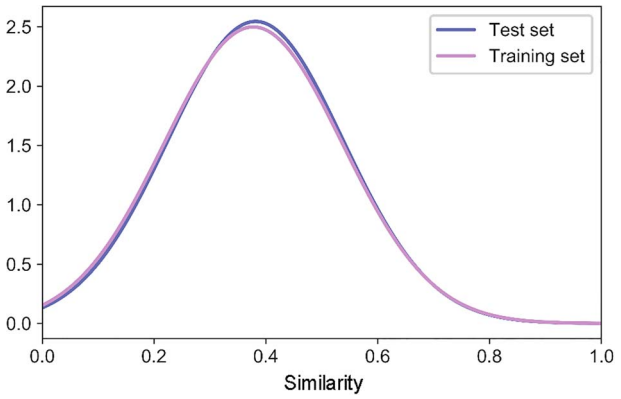

Figure 2: Distribution of pair-based Tanimoto coefficients of mouse neurotoxicity data (FP = MACCSFP).

Results

Data distribution and chemical structure diversity

In this study, we collected 422 organic compounds from the ChemIDplus database to construct regression prediction models for autonomic neurotoxicity induced by intraperitoneal administration in mouse. We used Discover Studio software to randomly divide the 422 compounds into training set and test sets by 8:2 (Fig. 1). There are 339 compounds in the training set, and 83 compounds in the test set. The distribution of these data is shown in Fig. 1. The experimental pLD_{50} values for the entire data were mainly distributed between −3.2 and −1.5. In addition, the PCA space of the descriptors filtered by the feature selection process is shown in Fig. 1B. The training set and the test set data are basically covered in the same spatial distribution. Therefore, it is relatively reliable to use the prediction results of the test set to evaluate the prediction models constructed by the training set. The Tanimoto similarity coefficient of the training set and the test set was calculated by using MACSSFP fingerprint, and Python was used for statistical drawing. In this study, the Tanimoto index is pair-based. As shown in Fig. 2, the average Tanimoto index of the training set was 0.381 and it was 0.377 for the test set. To a certain extent, this indicates that the structural diversity of our compounds is high, so the applicability domain of the prediction models we constructed is relatively large.

Feature reduction results

In this study, we performed feature reduction on 596 features with removed eigenvalues of 0 or covariance of 0. The result was shown in Fig. 3. Through a series of feature reduction steps, 19 most suitable descriptors for modeling were finally selected. The distribution of these 19 features is shown in Fig. 3. The chemical space of the test set and the training set is basically covered, so it is reliable to use the test set validation to verify the prediction ability and generalization ability of the QSAR models. The 19 molecular descriptors are MATSp2, bcutv10, MRVSA5, GATSe2, Rpc, EstateVSA1, Geto, Smax15, MTPSA, bcute2, J, Chiv10, Chiv9, mChi1, Smin8, Hy, Smin32, MATSv3 and MATSe3.

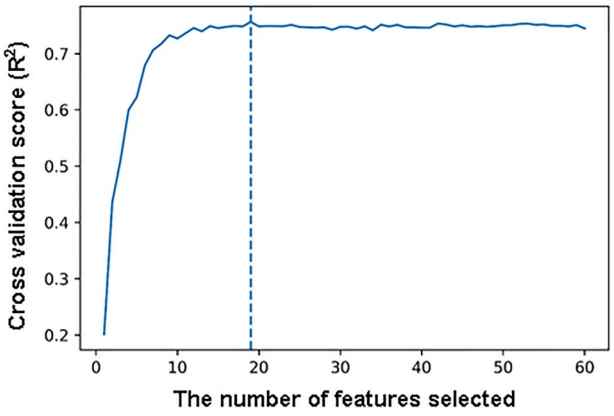

Figure 3: Feature dimension plot of the recursive feature elimination incorporated with random forest (RFE–RF) method (10-fold cross validation).

These descriptors mainly include autocorrelation molecular descriptors, hydrophilicity indices, charge descriptors, topological property descriptors, molecular connection indices, etc. In addition, we have regularized the 19 molecular descriptors that were filtered out using the QuantileTransformer module in scikit-learn. As shown in Fig. 4, before feature scaling, these feature values have a wide range of sizes and large differences in distribution space. These features were scaled to normal distribution or uniform distribution, and were distributed in between 0 and 1, the variance of the scaled features was small. The reason why we performed feature scaling was that if the variance of a feature was one or more orders of magnitude larger than the other variances, it was likely to dominate the objective function in machine learning and prevent the machine learning method from learning other features as expected correctly.

Internal and external validation results of the models

In total, 19 molecular descriptors were selected by feature selection, and then 8 machine learning algorithms were used to construct regression prediction models of mouse autonomic neurotoxicity. The statistical results of the prediction models constructed based on these 19 descriptors and using the optimized machine learning algorithms were shown in Table 2. Based on the internal validation of the training set and the predictions of the test set, the ability of these eight machine learning methods to predict neurotoxicity data is different. Among these models, the GPR model has the highest predictive power for the internal validation results of the training set, and the ETR model also shows high predictive ability. Interestingly, we found that the best predictive power for the test set was not the GPR model, but the ETR model (q^2_{test}= 0.784). Comparatively, the two models BGR (q^2_{test}= 0.702) and MLPR (q^2_{test} = 0.662) have the worst prediction ability for the test set. The scatter plot of the predicted value of the best model ETR and the experimental value pLD_{50} is shown in Fig. 5. From the figure, we can see that the prediction results of the ETR model are basically cover on the diagonal or near the diagonal, and the prediction error is relatively small. The statistical data $RMSE_{test}$ is 0.201, which also illustrates this point. Therefore, we propose to use the ETR model as a prediction model for autonomic neurotoxicity in mouse. Besides, the Y-randomization test was performed to support the reliability of our models. The results of Y-randomization test were listed in Table 3. The low q^2_{test} values showed that the good results of our models were not due to a chance correlation or structural dependency of the training set.

Applicability domain of models

According to the principles of the OECD validation for QSAR models, each QSAR prediction model must have a clear applicability domain. In this study, we used the SDD and the lever distance to describe the boundaries of the applicability domain of the training set. The distribution map of the applicability domain is shown in Fig. 6A. The coverage of the test set in the applicability domain using the SDD and the lever distance is shown in Table 3 For each model, 97.6% of its test set is in the applicability domain. In addition, the Cook's distance describing the effect of each compound on the model is shown in Fig. 6B. Since the leverage distance is greater than the applicability domain threshold ($h*$ = 0.168), six compounds in the training set were considered to be outside the applicability domain. In addition, according to the threshold of the Cook's distance (cutoff = 0.0125), 28 compounds in the training set were considered to have a large interference effect on the prediction of the model, and the prediction ability of the model was destroyed.

Discussion

The influence of different features on the regression models

The feature is a detailed description of the physical and chemical properties of the compounds, and the contribution of

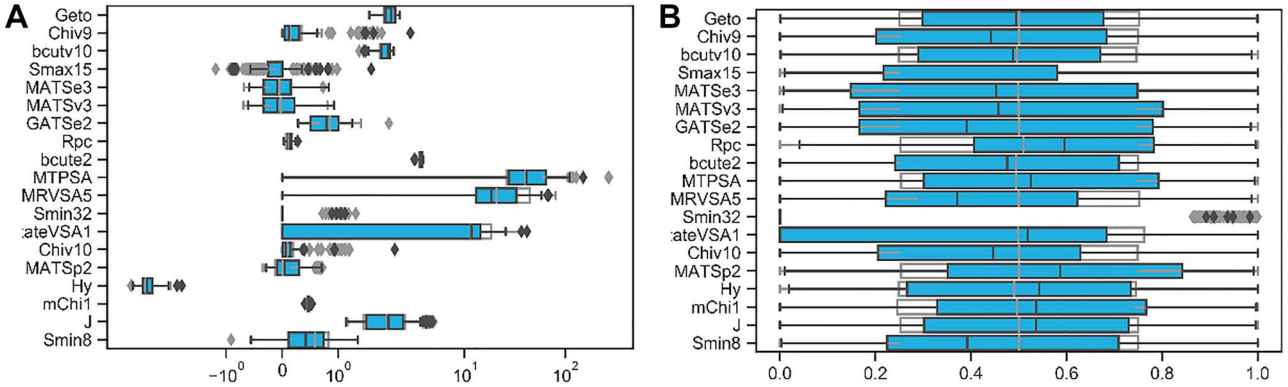

Figure 4: Comparison of the original (**A**) and normalized (**B**) distributions of the final selected descriptors in different data sets in this study. Gray boxplot stands for the training set and blue boxplot stands for the test set.

Table 2: Statistical results for the regression models based on 19 descriptors for the training set and test set

Models	R^2_{adj}	q^2_{LOO}	q^2_{test}	MAE_{train}	$RMSE_{train}$	MAE_{test}	$RMSE_{test}$	AD coverage (%)
GPR	0.822	0.832	0.742	0.159	0.211	0.174	0.220	97.6
ETR	0.818	0.828	0.784	0.158	0.214	0.159	0.201	97.6
NuSVM	0.810	0.821	0.706	0.163	0.218	0.189	0.235	97.6
kNN	0.805	0.816	0.720	0.158	0.222	0.168	0.228	97.6
SVM	0.791	0.803	0.720	0.173	0.229	0.183	0.229	97.6
RF	0.773	0.785	0.703	0.167	0.239	0.175	0.236	97.6
BGR	0.767	0.780	0.702	0.168	0.242	0.175	0.236	97.6
MLPR	0.711	0.727	0.662	0.207	0.270	0.192	0.252	97.6

LOO, leave one out.

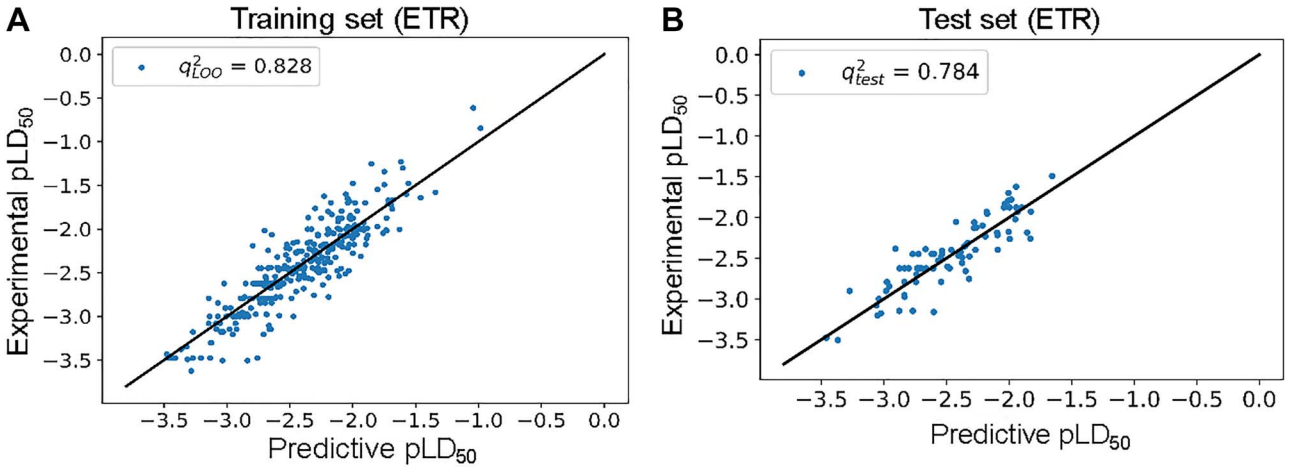

Figure 5: Predicted results of the best regression model (ETR). Scatter plots of the experimental pLD_{50} values versus the predicted values for the compounds in the (**A**) training set and (**B**) test set.

Table 3: Y-randomization results (q^2_{test}) of the eight models

Iteration	GPR	NuSVM	SVM	kNN	MLPR	RF	BGR	ETR
1	0.063	0.123	0.085	0.077	0.072	0.078	0.073	0.085
2	0.067	0.119	0.094	0.110	0.064	0.072	0.102	0.092
3	0.077	0.114	0.084	0.090	0.077	0.081	0.097	0.072
4	0.113	0.107	0.110	0.114	0.084	0.119	0.070	0.092
5	0.095	0.135	0.065	0.085	0.081	0.099	0.078	0.059
6	0.080	0.109	0.087	0.072	0.069	0.107	0.045	0.106
7	0.056	0.140	0.094	0.072	0.091	0.069	0.088	0.084
8	0.057	0.144	0.032	0.095	0.095	0.081	0.041	0.063
9	0.073	0.094	0.127	0.088	0.072	0.126	0.061	0.086
10	0.090	0.058	0.065	0.103	0.071	0.104	0.100	0.063

different features to the constructed regression models is different. As shown in Fig. 7, the top five features of importance are MATSe3, MATSv3, Smin32, Hy and Smin8. Among them, MATSe3 and MATSv3 belong to the Morgan autocorrelation descriptors, Smin32 and Smin8 belong to the E-state descriptors, and Hy belongs to the hydrophilic index.

According to the literature report [29], the hydrophily index Hy is inversely proportional to the neurotoxicity pLD_{50} value. It can also be seen from Fig. 8 that the hydrophilic index Hy of the compounds used for modeling is indeed inversely proportional to the neurotoxicity value pLD_{50}, which further corroborates the rationality and reliability of the models we built. Compounds with lower hydrophilic index have a higher partition coefficient of lipid water, and their ability to penetrate the biomembrane or blood–brain barrier is relatively strong, so their potential neurotoxicity is also greater.

Among them, the highest contribution to the model is the Morgan autocorrelation descriptor MATSe3, which is calculated by atomic Sanderson electronegativity weighting. From Fig. 9, the electronegativity value MATSe3 of the compounds are directly proportional to the neurotoxicity value pLD_{50} (the linear correlation coefficient of training set is 0.4, the linear correlation coefficient of test set is 0.41). Functional groups with a large electronegativity such as nitro (-NO_2), phenyl (-C_6H_5), aldehyde groups (-CHO) can be mutually absorbed with positively charged groups in the body, thereby enhancing toxicity.

Except that the hydrophily index Hy is inversely proportional to the neurotoxicity pLD_{50} value, the other four top 5 molecular

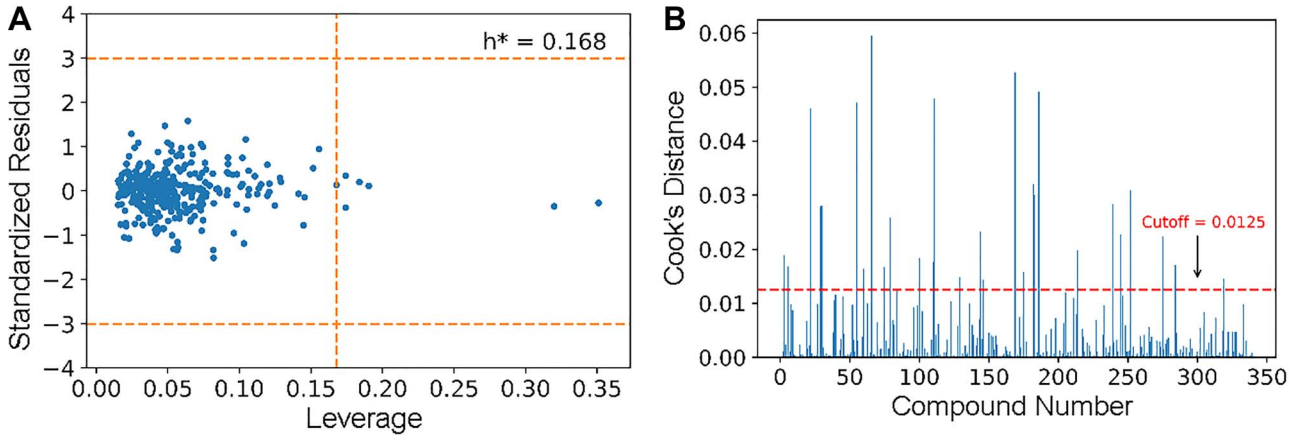

Figure 6: Applicability domain defined in this study. Williams plot (**A**) and Cook's distance plot (**B**) were given using the leverage approach.

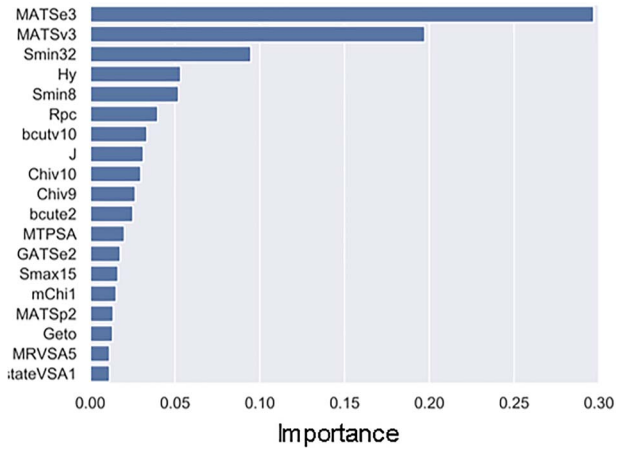

Figure 7: Importance of molecular descriptors selected by RFE–RF.

descriptors are directly proportional to neurotoxicity (the linear correlation coefficient ranges from 0.2 to 0.4).

Comparison of our model and previous models

Compared with previous prediction models (Table 4), it can be found that we built the regression models with larger data. Although the predictive power of our model is not the best, the chemicals we used to construct the models are rich in structural diversity. The chemicals of the models 1, 2 and 4 were mainly benzene, alkane, alcohol, ester, pyridine compounds, etc. The chemicals of model 3 were organophosphorus compounds. The chemicals used for constructing neurotoxicity regression models were relatively rare, so the application range of these models is relatively small. In contrast, we used more than dozens types of compounds to build the regression models. It can be found that the models we build have a wider applicability range for chemical neurotoxicity prediction with good performance in predicting neurotoxicity.

Effect of different machine learning algorithms on regression models

Different machine learning methods have a gap in the ability to learn neurotoxicity data, so the models we constructed have different predictive powers for unknown compounds. As shown in Table 3, considering the evaluation indexes of the models, the ranking of the model prediction ability is ETR > GPR > SVR > kNN > NuSVR>RF > BGR > MLPR. ETR is an extreme random tree regression learner, which is composed of an integrated method similar to random forest, but different. The random forest applies the Bagging model, while the ETR uses all the training samples to get each decision tree. In other words, each decision tree applies the same training samples. In addition, the random forest is the best bifurcation attribute in a random subset, and the ETR is completely random to obtain the bifurcation value, which implements bifurcation of the decision tree. For a decision tree, because its optimal bifurcation property is randomly selected, its prediction result is often inaccurate, but when multiple decision trees are combined, a good prediction effect can be achieved. It may be the reason why ETR performs excellently in model prediction. In addition to the ETR model, the GPR model also showed good results in the prediction of the test set (q^2_{test} = 0.742). GPR is a Gaussian process regression method, whose prediction interpolates the observations. In addition, the prediction is probabilistic (Gaussian) so that one can compute empirical confidence intervals and decide based on those if one should refit (online fitting, adaptive fitting) the prediction in some region of interest. The method also has the advantage of wide application. Different modeling situations provide different kernels to improve the predictive power of the model. In addition, except for the test set q^2 of the MLPR model, which is less than 0.7, the q^2 of the prediction results of other machine learning methods are greater than 0.7, indicating that these machine learning algorithms are superior in predicting chemical autonomic neurotoxicity. In summary, considering the performance of the model and the accuracy of the prediction, we recommend ETR as a regression prediction model for autonomic neurotoxicity in mouse.

Conclusion

In this study, we used PyBioMed molecular descriptors in conjunction with eight machine learning methods to develop regression models for predicting potential neurotoxic compounds based on a chemical diverse mouse autonomic neurtoxicity data set. Using the feature dimension reduction method, 19 descriptors suitable for the neurotoxicity prediction model were selected, and 5 descriptors that had the greatest impact on the model were analyzed. Taking into account the evaluation

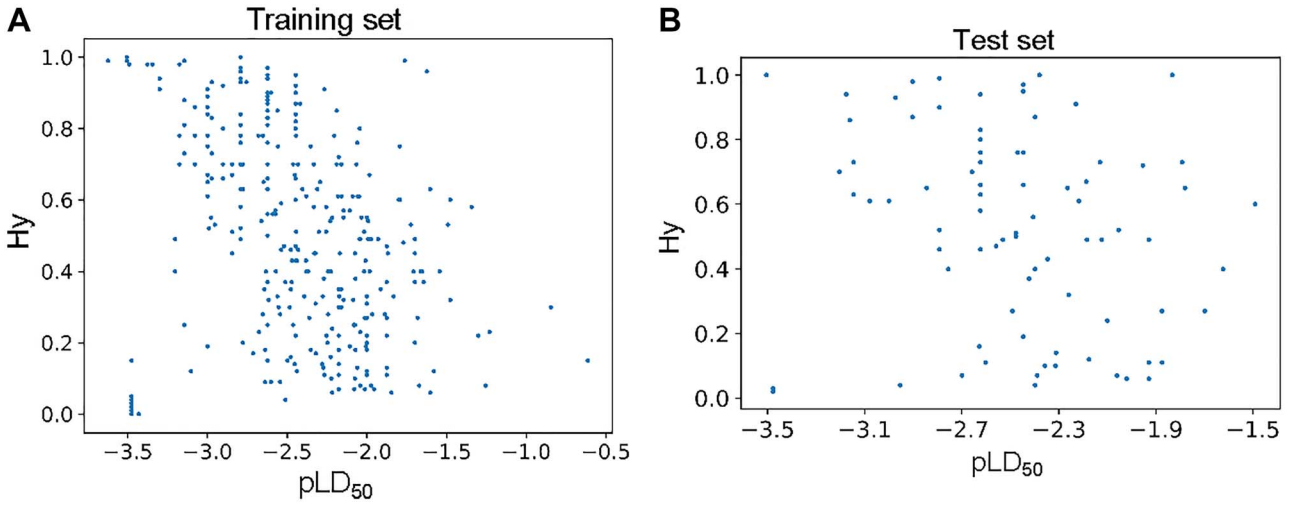

Figure 8: The scatter plot of Hy and pLD_{50} for the training set (**A**) and test set (**B**).

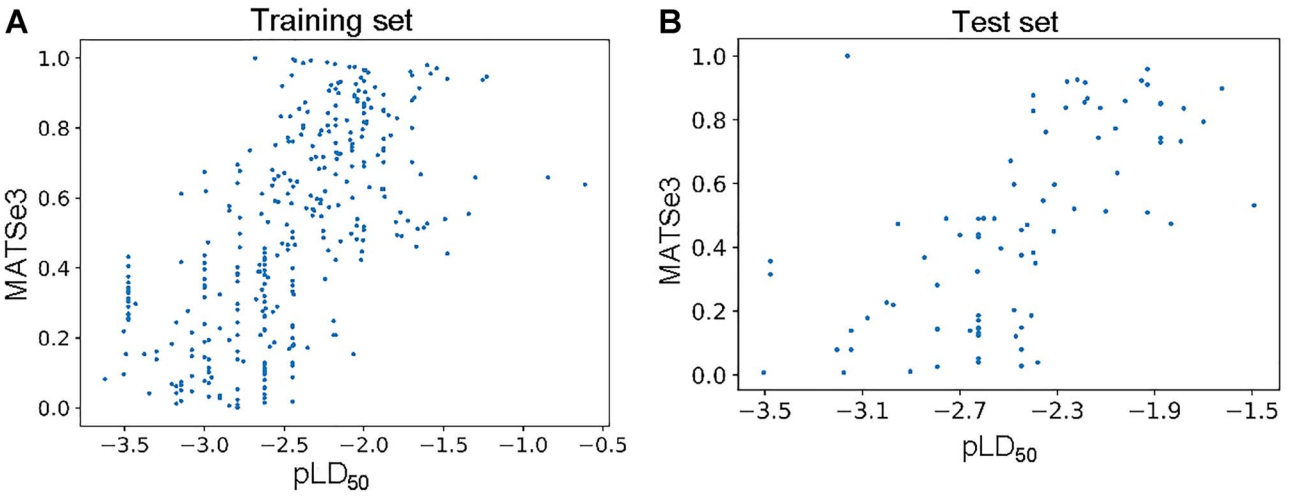

Figure 9: The scatter plot of MATSe3 and pLD_{50} for the training set (**A**) and test set (**B**).

Table 4: The statistical results of comparison of our model (ETR) with previous models

No.	Method	Data set	Descriptors	Validation	q^2_{train}	RMSE	Ref.
1	MLR	44	physicochemical descriptors	No data	0.667	0.280	[14]
2	MLR	45	TOPS-MODE	LOO	0.814	0.273	[15]
3	polynomial regression	7	Hansch constants	No data	0.994	0.087	[16]
4	PNN, GRNN	47	Chemopy	5-CV	0.999	0.01	[17]
5	ETR	422	PyBioMed	LOO	0.828	0.214	–

MLR, multiple linear regression; PNN, probabilistic neural networks; GRNN, generalized regression neural networks; ETR, extra-trees regressor; LOO, leave one out; 5-CV, 5-fold cross-validation.

indicators of R^2_{adj}, q^2_{test} and $RMSE_{test}$, the ETR model is optimal for the autonomic neurotoxicity endpoint, which provides a robust and reliable prediction of the autonomic neurotoxicity values of chemicals. In addition, based on the leverage distance, we define the applicability domain of the model. In summary, this study developed robust and reliable quantitative models for the prediction of neurotoxicity of compounds. These results can be used for the accurate quantitative prediction of chemical induced neurotoxicity and for the drug screening in early stage of drug discovery.

Supplementary data

Supplementary data is available at *TOXRES Journal* online.

Acknowledgments

We gratefully acknowledged the financial supports from the National Natural Science Foundation of China (Grants 81673356 and 81872800).

Conflict of interest statement

None declared.

References

1. Harrison RK. Phase II and phase III failures: 2013-2015. *Nat Rev Drug Discov* 2016;**15**:817–8.
2. Siramshetty VB, Nickel J, Omieczynski C *et al.* Preissner, WITHDRAWN-a resource for withdrawn and discontinued drugs. *Nucleic Acids Res* 2015;**44**:D1080–6.
3. Rose AL, Wiśniewski HM, Cammer W. Neurotoxicity of hexachlorophene: new pathological and biochemical observations. *J Neurol Sci* 1975;**24**:425–35.
4. Thomas C, Groten J, Kammüller M *et al.* Popliteal lymph node reactions in mice induced by the drug zimeldine. *Int J Immunopharmacol* 1989;**11**:693–702.
5. Fukui T, Asakura K, Hikichi C *et al.* Histone deacetylase inhibitor attenuates neurotoxicity of clioquinol in PC12 cells. *Toxicology* 2015;**331**:112–8.
6. Fagius J, Osterman P, Siden A *et al.* Guillain-Barré syndrome following zimeldine treatment. *J Neurol Neurosurg Psychiatry* 1985;**48**:65–9.
7. Meyer-Baron M, Kim EA, Nuwayhid I *et al.* Occupational exposure to neurotoxic substances in Asian countries-challenges and approaches. *Neurotoxicology* 2012;**33**:853–61.
8. Mason LH, Harp JP, Han DY. Pb neurotoxicity: neuropsychological effects of lead toxicity. *Biomed Res Int* 2014;**2014**:1–8.
9. Lein PJ, Bonner MR, Farahat FM *et al.* Experimental strategy for translational studies of organophosphorus pesticide neurotoxicity based on real-world occupational exposures to chlorpyrifos. *Neurotoxicology* 2012;**33**:660–8.
10. Keski-Säntti P, Kaukiainen A, Hyvärinen HK *et al.* Occupational chronic solvent encephalopathy in Finland 1995-2007: incidence and exposure. *Int Arch Occup Environ Health* 2010;**83**:703–12.
11. Lidsky TI, Schneider JS. Lead neurotoxicity in children: basic mechanisms and clinical correlates. *Brain* 2003;**126**:5–19.
12. Henderson JD, Higgins RJ, Dacre JC *et al.* Neurotoxicity of acute and repeated treatments of tabun, paraoxon, diisopropyl fluorophosphate and isofenphos to the hen. *Toxicology* 1992;**72**:117–29.
13. Breuer D. Analytical performance issues: GESTIS database: International limit values for chemical agents-a readily accessible source of occupational exposure limits (OELs). *J Occup Environ Hyg* 2010;**7**:D37–42.
14. Nicolotti O, Benfenati E, Carotti A *et al.* REACH and in silico methods: an attractive opportunity for medicinal chemists. *Drug Discov Today* 2014;**19**:1757–68.
15. Cronin M. Quantitative structure-activity relationship (QSAR) analysis of the acute sublethal neurotoxicity of solvents. *Toxicol In Vitro* 1996;**10**:103–10.
16. Estrada E, Molina E, Uriarte E. Quantitative structure-toxicity relationships using TOPS-MODE. 2. Neurotoxicity of a non-congeneric series of solvents. *SAR QSAR Environ Res* 2001;**12**:445–59.
17. Malygin VV, Sokolov VB, Richardson RJ *et al.* Quantitative structure-activity relationships predict the delayed neurotoxicity potential of a series of o-alkyl-o-methylchloroformimino phenylphosphonates. *J Toxicol Environ Health, Part A* 2003;**66**:611–25.
18. Basant N, Gupta S, Singh KP. Predicting the acute neurotoxicity of diverse organic solvents using probabilistic neural networks based QSTR modeling approaches. *Neurotoxicology* 2016;**53**:45–52.
19. Guyon I, Weston J, Barnhill S *et al.* Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;**46**:389–422.
20. Granitto PM, Furlanello C, Biasioli F *et al.* Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom Intel Lab Syst* 2006;**83**:83–90.
21. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inf* 2010;**29**:476–88.
22. Nicolotti O, Carotti A. QSAR and QSPR studies of a highly structured physicochemical domain. *J Chem Inf Model* 2006;**46**:264–76.
23. Chen Y, Cheng F, Sun L *et al.* Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. *Ecotox Environ Safe* 2014;**110**:280–7.
24. Gissi A, Gadaleta D, Floris M *et al.* An alternative QSAR-based approach for predicting the bioconcentration factor for regulatory purposes. *ALTEX-Altern Anim Ex* 2014;**31**:23–36.
25. Kaneko H, Funatsu K. Applicability domain based on ensemble learning in classification and regression analyses. *J Chem Inf Model* 2014;**54**:2469–82.
26. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *ATLA-Altern Lab Anim* 2005;**33**:445–59.
27. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 2007;**26**:694–701.
28. Sahigara F, Mansouri K, Ballabio D *et al.* Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 2012;**17**:4791–810.
29. Wible J, James H, Barco SJ *et al.* Neurotoxicity of non-ionic X-ray contrast media after intracisternal administration in rats. *Eur J Radiol* 1995;**19**:206–11.