# ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning Approaches

Tailong Lei,[†] Huiyong Sun,[†] Yu Kang,[†] Feng Zhu,[†] Hui Liu,[†] Wenfang Zhou,[†] Zhe Wang,[†] Dan Li,[†] Youyong Li,[§] and Tingjun Hou*,[†,‡]

[†]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, P. R. China

[‡]State Key Lab of CAD&CG, Zhejiang University, Hangzhou, Zhejiang 310058, P. R. China

[§]Institute of Functional Nano and Soft Materials (FUNSOM), Soochow University, Suzhou, Jiangsu 215123, P. R. China

S Supporting Information

ABSTRACT: Xenobiotic chemicals and their metabolites are mainly excreted out of our bodies by the urinary tract through the urine. Chemical-induced urinary tract toxicity is one of the main reasons that cause failure during drug development, and it is a common adverse event for medications, natural supplements, and environmental chemicals. Despite its importance, there are only a few *in silico* models for assessing urinary tract toxicity for a large number of compounds with diverse chemical structures. Here, we developed a series of qualitative and quantitative structure—activity relationship (QSAR) models for predicting urinary tract toxicity. In our study, the recursive feature elimination method incorporated with random forests (RFE-RF) was used for dimension reduction, and then eight machine learning approaches were used for QSAR modeling, i.e., relevance vector machine (RVM), support vector machine (SVM), regularized random forest (RRF), C5.0 trees, eXtreme gradient boosting (XGBoost), AdaBoost.M1, SVM boosting (SVMBoost), and RVM boosting (RVMBoost). For building classification models, the synthetic minority oversampling technique was used to handle the imbalance data set problem. Among all the machine learning approaches, SVMBoost based on the RBF kernel achieves both the best quantitative ($q^2_{ext} = 0.845$) and qualitative predictions for the test set (MCC of 0.787, AUC of 0.893, sensitivity of 89.6%, specificity of 94.1%, and global accuracy of 90.8%). The application domains were then analyzed, and all of the tested chemicals fall within the application domain coverage. We also examined the structure features of the chemicals with large prediction errors. In brief, both the regression and classification models developed by the SVMBoost approach have reliable prediction capability for assessing chemical-induced urinary tract toxicity.

KEYWORDS: quantitative structure—activity relationship, urinary tract toxicity, nephrotoxicity, imbalanced classification, boosting, support vector machine, ensembles, machine learning
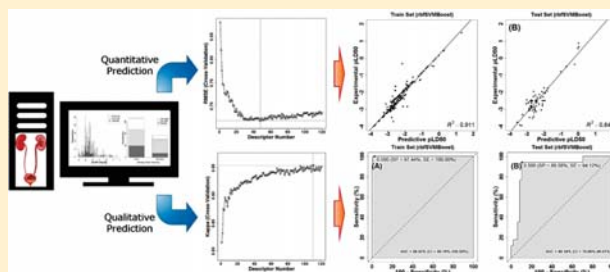
■ INTRODUCTIONS

Preclinical and clinical safety issue is the prominent cause of drug development failures,[1−3] and in the earlier phase of the drug development, more safety-related attritions occurred.[2] Urinary tract toxicity, especially nephrotoxicity, plays a main role in safety-related failures at all phases of drug development,[4−6] and it is common after medications due to the metabolism and excretion functions of the urinary system. The primary symptom of urinary tract toxicity is the impairment of the normal functions of the urinary system. The major targets of urinary tract toxicants are glomerulus, proximal tubule, and renal interstitium, and the main mechanisms involved in urinary tract toxicity include oxidative stress and immune-mediated reactions.[7,8] The clinical manifestations vary from a mild reduction in renal function to even severe acute renal failure.[7] In addition, some drugs can disturb renal perfusion and induce changes of filtration capacity.[7,9] Drugs that can cause urinary tract toxicity include antibiotics, antiretroviral agents, antihypertensive agents, anticoagulants, thrombolytic agents, chemotherapeutic agents, radiocontrast agents, immunosuppressive agents, diuretics, nonsteroidal anti-inflammatory drugs, designer drugs, etc.[7,10,11] Besides, traditional natural drugs and supplements are also a large source of urinary tract toxic substances.[12] Aristolochia and aconite are two famous examples.

Although its high incidence,[13] the detection of urinary tract toxicity is often delayed until an overt change in renal function occurs.[7] Usually when we detect urinary tract toxicity definitively,

Table 1. Reported Classification Models for Nephrotoxicity

No.	method[a]	number of chemicals[b]	source of data	descriptors	validation	sensitivity	specificity	concordance	ref
1	RF	41 chemicals (29/12)	cell-based in vitro testing	Interleukin 6 and 8 gene expression levels (in vitro)	10-fold CV	HPTC: 99.9% (train), 89.3% (test); HPTC-like: 99.7% (train), 89.0% (test)	HPTC: 99.8% (train), 74.7% (test); HPTC-like: 100.0% (train), 85.0% (test)	HPTC: 99.9% (train), 82.0% (test); HPTC-like: 99.8% (train), 87.0% (test)	19
2	SARpy	202 chemicals with rat LOAEL (89/113)	HESS database	no data	no data	no data	no data	no data	20
3	RF	44 chemicals (24/20)	published literatures	Haralick's texture features of cell images (in vivo)	10 × 10-fold CV	HPTC-A: 99.4–99.7% (train), 75.2–81.2% (test); HPTC-B: 99.9% (train), 83.7% (test); HK-2: 100.0% (train), 98.8% (test)	HPTC-A: 99.1–99.7% (train), 70.5–80.5% (test); HPTC-B: 99.6% (train), 79.5% (test); HK-2: 99.8% (train), 79.0% (test)	HPTC-A: 99.3–99.6% (train), 75.8–78.3% (test); HPTC-B: 99.7% (train), 81.6% (test); HK-2: 99.9% (train), 88.9% (test)	21
4	NB, k-NN, SVM, RF	41 chemicals	published literature	interleukin 6 and 8 gene expression levels	3-fold CV	NB: 62.0%, k-NN: 74.0%, SVM: 78.7%, RF: 89.4%	NB: 78.4%, k-NN: 74.2%, SVM: 84.2%, RF: 85.9%	NB: 70.2%, k-NN: 74.1%, SVM: 81.6%, RF: 87.8%	22
5	k-NN based ensemble classifier	131 chemicals	TG-GATEs	kidney pathological symptoms (in vivo)	3-fold CV	73% (individual), 85% (integrative)	88% (individual), 84% (integrative)	no data	23
6	SVM (radial kernel)	170 chemicals (136/34)	TG-GATEs	Conserved early cytotoxicity signatures (in vivo)	10-fold CV	No data	No data	80–97% (test)	24
7	SVM	172 parent chemicals of tubular necrosis (138/34)	PharmaPendium database human	8 fingerprints by PaDEL	10-fold CV	64–86% (train), 56–76% (test)	49–79% (train), 51–73% (test)	71–77% (train), 54–72% (test)	25
	SVM	176 parent chemicals of interstitial nephritis (141/35)	PharmaPendium database	8 fingerprints by PaDEL	10-fold CV	60–86% (train), 51–89% (test)	66–80% (train), 61–83% (test)	69–76% (train), 64–79% (test)	
	SVM	90 parent chemicals of tubulo-interstitial nephritis (72/18)	PharmaPendium database	8 fingerprints by PaDEL	10-fold CV	43–85% (train), 17–78% (test)	82–92% (train), 71–94% (test)	75–83% (train), 71–89% (test)	
	SVM	215 metabolite chemicals of tubular necrosis (172/43)	PharmaPendium database	8 fingerprints by PaDEL	10-fold CV	64–81% (train), 60–84% (test)	70–87% (train), 72–85% (test)	70–83% (train), 68–84% (test)	
	SVM	220 metabolite chemicals of interstitial nephritis (176/44)	PharmaPendium database	8 fingerprints by PaDEL	10-fold CV	53–85% (train), 48–86% (test)	76–90% (train), 79–96% (test)	72–84% (train), 73–85% (test)	
	SVM	112 metabolite chemicals of tubulo-interstitial nephritis (90/22)	PharmaPendium database	8 fingerprints by PaDEL	10-fold CV	48–80% (train), 36–82% (test)	85–95% (train), 83–100% (test)	79–84% (train), 72–83% (test)	
8	SVM	40 chemicals (22/19)	histopathology and microarray assays	24-h gene expression profiles of 19 genes	5-fold CV	93% (test)	no data	90% (test)	26
9	ChemTree (RP)	1001 chemicals of nephrotoxicity (847/154)	MetaTox Consortium	augmented atom pairs	10-fold CV	90% (test)	84% (test)	87% (test)	27
	ChemTree (RP)	263 chemicals of kidney necrosis (221/42)	MetaTox Consortium	augmented atom pairs	10-fold CV	96% (test)	100% (test)	98% (test)	
	ChemTree (RP)	289 chemicals of kidney relative weight gain (240/49)	MetaTox Consortium	augmented atom pairs	10-fold CV	95% (test)	100% (test)	98% (test)	

Table 1. continued

No.	method[a]	number of chemicals[b]	source of data	descriptors	validation	sensitivity	specificity	concordance	ref
	ChemTree (RP)	707 chemicals of nephron injury (598/109)	MetaTox Consortium	augmented atom pairs	10-fold CV	91% (test)	100% (test)	96% (test)	
10	CHAID, CART	507 chemicals with nephrotoxic ADRs	University of Alberta's DrugBank database	descriptors by ChemAxon and CDK	CV	no data	no data	CHAID: 84.71%, CART: 88.69%	28
11	MultiCASE (MC4PC)	1595 chemicals of kidney disorders	AERS derived databases	no data	Leave-10%-out CV	80.4%	42.6%	no data	29
	MultiCASE (MC4PC)	1595 chemicals of kidney function tests	AERS derived databases	no data	Leave-10%-out CV	87.7%	43.3%	no data	
	MultiCASE (MC4PC)	1595 chemicals of nephropathies	AERS derived databases	no data	Leave-10%-out CV	81.5%	49.5%	no data	
	MultiCASE (MC4PC)	1595 chemicals of hematuria	AERS derived databases	no data	Leave-10%-out CV	80.7%	55.2%	no data	
	MultiCASE (MC4PC)	1595 chemicals of urolithiasis	AERS derived databases	no data	Leave-10%-out CV	84.2%	31.3%	no data	
	MultiCASE (MC4PC)	1595 chemicals of bladder disorders	AERS derived databases	no data	Leave-10%-out CV	87.6%	36.4%	no data	
12	SVM (four categories)	148 treatment of 5 chemicals (120/28)	histopathology and microarray assays	gene expression profiles	no data	no data	82.1%	no data	30

[a]Abbreviations: RF, random forest; SVM, support vector machine; k-NN, k-nearest neighbors; NB, Naïve Bayes; RP, recursive partitioning; CHAID, chi-squared automatic interaction detection; CART, classification and regression trees; CV, cross validation; HESS, Hazard Evaluation Support System; TG-GATEs, Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System; HPTC, human primary renal proximal tubular cells; LOAEL, lowest observed adverse effect level; ADR, adverse drug reaction; AERS, FDA Adverse Events Reporting System. [b]The front/back numbers in the brackets refer to the number of the compounds in the training/test sets, respectively. If there is no bracket, it means that the corresponding literature did not report the size of the training and test sets.

patients have been extremely sick. Thus, identification of the biomarkers for early injuries has attracted great attention,[14,15] but most of the reported biomarkers do not show reliable predictions.[16] For this reason, serum blood urea nitrogen and creatinine are still the only two recognized biomarker used in preclinical and clinical examinations, although they are difficult to detect acute or subtle injuries.

In recent years, urinary tract toxicity of environmental chemicals has been concerned.[17] They are exposed to people in a limited dose during everyday life and may cause chronic poisoning. Many notorious contaminants, such as phthalates, dioxins, furans, bisphenol A, polycyclic aromatic hydrocarbon, polychlorinated biphenyl, and perfluoroalkyl acid, are urinary tract toxic substances.[17] Although environmental chemicals have very different administration routes, they bear similar mechanisms to drugs with urinary tract toxicity. Unfortunately, there are yet no effective biomarkers to detect the early urinary tract toxicity caused by environmental chemicals promptly.[17]

Due to the severe limitations of experimental determination of urinary tract toxicity, the development of *in silico* models as alternative approaches to reliable assessment of drug candidates or environmental chemicals without animal testing is quite demanding. Compared with other toxicological end points, there are only a few QSAR models for urinary tract toxicity prediction. The variety and complexity of the symptoms and mechanisms of urinary tract toxicity bring more difficulty to model development.[18] Furthermore, it is difficult to obtain large high-quality data sets of urinary tract toxicity. All the reported models related to urinary tract toxicity are summarized in Table 1.[19−30] Among them, only five models were established based on theoretical descriptors, while the others based on biomarker descriptors. All the models established based on theoretical descriptors used the data sets of drugs in various development phases, most of which were collected from commercial databases. However, these data sets are hard to be verified and rectified because of the idiosyncratic reactions in human. The idiosyncratic urinary tract toxicity, which is caused by immune-mediated reactions related to individual constitution differences,[8] is not dose-related, and it might be not suitable for modeling. Therefore, in our study we used the dose-related data set of mouse intraperitoneal urinary tract toxicity in order to avoid the noise brought by idiosyncratic data.

In our study, the mouse intraperitoneal urinary tract toxicity data set of 258 chemicals from a public database was finally used for modeling. After data collation and dimension reduction, eight machine learning approaches were used to develop the regression and classification models for assessing urinary tract toxicity in mouse. These machine learning approaches include relevance vector machine (RVM), support vector machine (SVM), regularized random forest (RRF), C5.0 trees, eXtreme gradient boosting (XGBoost), AdaBoost.M1, SVM boosting (SVMBoost), and RVM boosting (RVMBoost). The model performance was validated by the internal and external validations. Moreover, the structural features and important fragments of the chemicals with large prediction errors were analyzed.

■ METHODS AND MATERIALS

Data Preparation and Preprocessing. In our study, the mouse intraperitoneal urinary tract toxicity data of 279 compounds were collected from the ChemIDplus public database,[31] where the effect option was defined as "KIDNEY, URETER,

AND BLADDER". All the toxicity data were expressed as LD_{50} values, which is the dose killing half of total treated animals. The data quality was carefully verified. We checked whether the downloaded entries are the same as the online entries and whether the SMILES in the entries are the same as the literatures they cited, and the chemical structures with wrong stereoisomers identified by manual verification were corrected. Then, inorganic chemicals, polymers, and mixtures were deleted with the sdwash and sdfilter functions in the molecular operating environment (MOE) 2009 molecular simulation package.[32] The final data set includes 258 organic compounds with their toxicity data.

The SMILES representations of the total 258 compounds in the final data set were converted to 3-D structures and prepared by using the Prepare Ligands module in the Discovery Studio 2.5 (DS2.5) molecular simulation package.[33] And then they were optimized by the Energy Minimize module based on the MMFF94x force field and the default settings in MOE 2009.[32] Then, the total 334 descriptors available in MOE 2009[32] were calculated, and they characterize the physicochemical, structural and drug-like properties of the studied molecules. The descriptors that have all zero values or zero variance were removed. Then, the correlations across all pairs of descriptors were calculated, and the descriptors with the correlation (r) to any descriptor higher than the specified threshold (0.90) were regarded to be redundant. Either descriptor in each redundant pair was removed. Finally, 181 retained descriptors were centralized and scaled for the next step.

For model development and validation, the whole final data set was randomly split into a training set with 193 compounds (74.8%) and an external test set with 65 compounds (25.2%) for regression and classification modeling, respectively. For regression model development, the toxicity data were transformed into the negative logarithm form as $-\log[mg/kg]$ (or pLD_{50}). For classification model development, according to the regulation of U.S. Environmental Protection Agency,[34] the compounds were categorized into two classes: urinary tract toxicants ($LD_{50} \leq 500$ mg/kg) and urinary tract nontoxicants ($LD_{50} > 500$ mg/kg). The modeling workflow is illustrated in Figure 1.

Dimension Reduction. Dimension reduction is necessary to select the appropriate feature subset from high-dimensional data for model building. The aim of dimension reduction is to remove redundant or irrelevant features without sacrificing too much information.

In our study, the Cramer's V coefficients derived from the chi-squared test was used as importance indices to prefilter 120 descriptors. Then the recursive feature elimination (RFE) method incorporated with random forests (RF) was applied to select an optimal descriptor subset for quantitative and qualitative modeling of urinary tract toxicity, respectively. RFE is a frequently used wrapper method that has high capability to search informative features. This algorithm needs to calculate and update the importance ranks and eliminate the least important feature. It was first proposed by Guyon *et al.* with SVM as its subfunction,[35] and then RF was used afterward.[36,37] Herein it is implemented as follows: (1) train a 10-fold cross validated RF model; (2) compute the permutated feature importance; (3) retain the most important variables; (4) repeat steps 1 to 3 until reaching the best performance. Finally, the subset of descriptors that gives the best prediction performance is selected.
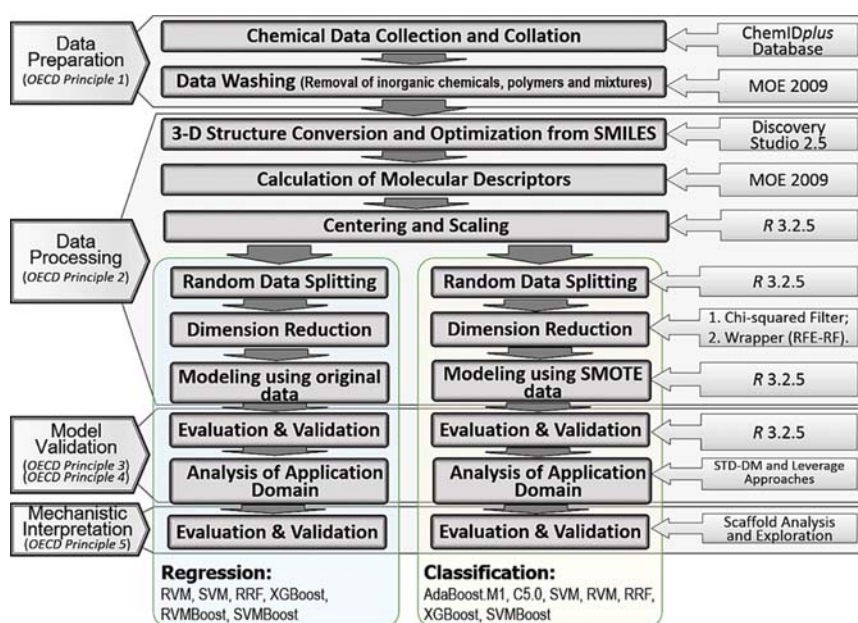
Figure 1. Workflow of the QSAR modeling for urinary tract toxicity.

Table 2. Optimal Hyperparameters Determined by the Self-Adaptive Differential Evolution Optimization Method for Different Machine Learning Approaches

models		optimal hyperparameters
regression	**rbfRVM**	The kernel width σ = 0.010693699.
	lpRVM	The kernel width σ = 0.092687008.
	rbfSVM	The kernel width σ = 0.029557736, the penalty parameter C = 29.88530174, and ε in the loss function = 0.172233249.
	lpSVM	The kernel width σ = 0.132737538, the penalty parameter C = 3.400185995, and ε in the loss function = 0.020162324.
	RRF	The number of predictors at each split = 34, the number of trees = 455, regularization value = 0.698064506, importance coefficient t = 0.777393866.
	XGBoost	The max number of boosting iterations = 90, maximum tree depth = 10, step size shrinkage = 0.2341923, minimum loss reduction = 0.3402731, subsample ratio of columns = 0.3144332, minimum sum of instance weight = 0.9968423.
	rbfSVMBoost	The kernel width σ = 0.022323451, the penalty parameter C = 30, ε in the loss function = 0.17, the number of learners = 8, learning rate = 0.062452257.
	lpSVMBoost	The kernel width σ = 36.23156771, the penalty parameter C = 4.555863298, ε in the loss function = 0.00001, the number of learners = 3, learning rate = 0.645761914.
	rbfRVMBoost	The kernel width σ = 0.027005144, the maximum number of iterations = 144, the number of learners = 100, learning rate = 0.469168167.
classification	**AdaBoost.M1**	The number of trees = 48, maximum tree depth = 11, coefficient type = "Freund".
	C5.0	Boosting iterations = 11, model type = "tree", winnow = FALSE.
	rbfSVM	The kernel width σ = 0.012, the penalty parameter C = 21.28, and ε in the loss function = 0.174.
	lpSVM	The kernel width σ = 0.136421758, the penalty parameter C = 6.869650869, and ε in the loss function = 0.186770399.
	rbfRVM	The kernel width σ = 0.010749923, the maximum number of iterations before termination = 127, the maximum number of iterations before termination of solver = 13
	RRF	The number of predictors at each split = 82, the number of trees = 537, regularization value = 0.668050052, importance coefficient t = 0.727091545.
	XGBoost	The max number of boosting iterations = 104, maximum tree depth = 16, step size shrinkage = 0.423371555, minimum loss reduction = 0.114881872, subsample ratio of columns = 0.318983358, minimum sum of instance weight = 0.864082756.
	rbfSVMBoost	The kernel width σ = 0.025811397, the penalty parameter C = 21.171073799, ε in the loss function = 0.173772599, the number of learners = 25, learning rate = 0.613803997.
	lpSVMBoost	The kernel width σ = 0.091274, the penalty parameter C = 4, ε in the loss function = 0.201532, the number of learners = 40, learning rate = 1.0.

QSAR Modeling by Machine Learning Approaches. Various machine learning approaches have been applied to QSAR modeling and ADME/T predictions, and the performances of different machine learning approaches on the prediction of urinary tract toxicity should be explored in order to develop reliable quantitative and qualitative models.[38] In our study, eight machine learning approaches − RVM, SVM, RRF, C5.0, XGBoost, AdaBoost.M1, SVMBoost and RVMBoost—were

employed for model building. The optimal parameters were determined by the self-adaptive differential evolution optimization method,[39] which is implemented by the DEoptimR package in R (version 3.2.5 ×64).

When developing the classification models, the synthetic minority oversampling technique (SMOTE)[40] was used to handle the obvious class imbalance in the training set, where the nontoxic chemicals are minority. The technique herein synthesized
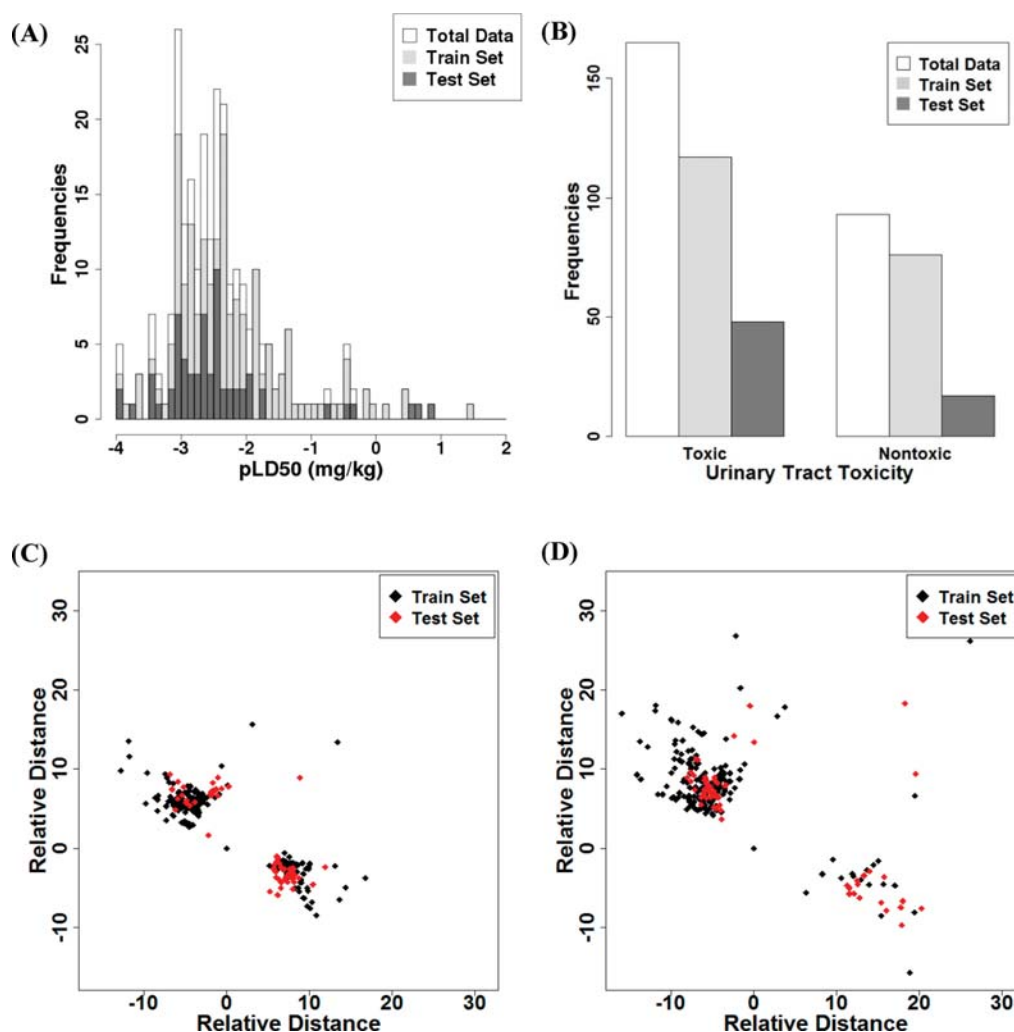
Figure 2. Diversity distribution of the training set ($n = 193$) and the external test set ($n = 65$). (A) Comparison of the toxicity value distribution in different data sets. (B) Comparison of the toxicity class distribution in different data sets. (C) Relative Euclidean distance of the compounds in the training set (black diamond) and the test set (red diamond) for regression. (D) Relative Euclidean distance of the compounds in the training set (black diamond) and the test set (red diamond) for classification.

samples through random linear interpolation of ten nearest neighbors of the observations of the minor class. A main integrative package in R (version 3.2.5 ×64), caret,[41] provides generic and object-oriented interfaces to implement all the following modeling algorithms with good scalability. The important hyperparameters of all the models are listed in Table 2.

Relevance Vector Machine (RVM). The RVM algorithm is a sparse Bayesian implementation of the standard SVM for regression and probabilistic classification.[42−44] In our study, the RVM based on the Gaussian radial basis (RBF) or Laplacian kernel is referred to as rbfRVM or lpRVM, respectively.

Support Vector Machine (SVM). SVM is one of the best supervised learning algorithm within the Vapnik-Chervonenkis framework[45] for QSAR modeling.[46,47] It was originally developed for classification, and then can also be used for regression. In our study, the SVM based on the RBF or Laplacian kernel was referred to as rbfSVM or lpSVM, respectively.

Regularized Random Forest (RRF). Random forest is a widely used ensemble method assembled by multiple decision trees and outputs the consensus predictions from individual trees.[48,49] RRF introduces a tree regularization framework into

random forest so that every tree in the forest would be guaranteed to possess a set of informative, but nonredundant features.[50]

C5.0 Trees (C5.0). C5.0 is a decision tree based algorithm developed by Ross Quinlan.[51−53] In our study, the winnowing function and the case weight attribute were ignored for speed raising.

eXtreme Gradient Boosting (XGBoost). Gradient boosting algorithm is a meta-algorithm to construct an ensemble strong learner from weak learners such as decision trees,[54,55] XGBoost is an efficient and distributed system to improve the gradient tree boosting algorithms.[56,57] In XGBoost, the cost function is expanded into two order Taylor's expansion, while the L1 and L2 regularizations are introduced. The regularization of the leaf nodes and column subsampling are used to balance the decline of the cost function and the model complexity in order to avoid overfitting. The step shrinkage multiplied to the leaf weights after each iteration can reduce the effects of trees to expand learning space. It retrofits gradient tree boosting algorithm for handling sparse data, raises a weighted quantile sketch for approximate optimization, and designs a column block structure for parallelization.[56]

AdaBoost.M1. AdaBoost, short for "adaptive boosting", is an iterative algorithm that aims at achieving a weighted sum of the boosted weak classifiers, especially decision trees.[58,59] It was widely used in the QSAR studies.[60,61] AdaBoost focuses the harder-to-classify samples so that it is sensitive to noisy data and outliers. In the original Freund's AdaBoost.M1 method, the negative logit coefficient α is calculated as $\ln((1-\varepsilon)/\varepsilon)$, which represents the importance of the weak classifiers in the final model. The AdaBoost.M1 algorithm trends to give higher weights to the individual trees of lower error rates to improve the predictive capability of the model.

SVM Boosting (SVMBoost) and RVM Boosting (RVMBoost). In our study, some of the extracted descriptors may have redundant information or be too sensitive to noise or background. Boosting could be a good way to solve the problems by aggregating the predictions of multiple weak learners to produce a powerful assembly,[62] and the AdaBoost framework is a popular solution. Here the AdaBoost.SAMME framework that was adjusted and expanded to deal with probabilistic multiclass classification cases by Zhu *et al.*[63] was used to build classification models, and the AdaBoost.R framework proposed by Drucker[64] was used to build regression models. In our study, SVMs and RVMs were used as the weak learners of the AdaBoost framework, and they were referred to as SVMBoost and RVMBoost. SVMBoost has illustrated improved performance over SVM in many fields for classification[65−69] and regression[70,71] modeling. However, RVMBoost has only been used in classification.[72,73] Here, SVMBoost based on the RBF or Laplacian kernel is referred to as rbfSVMBoost or lpSVMBoost, respectively. But for RVMBoost, only the regression based on the RBF kernel (rbfRVMBoost) could come to convergence in our study.

Evaluation and Validation of the QSAR Models. According to the Organization for Economic Co-operation and Development (OECD) guideline,[74] rigorous internal and external validations should be used to evaluate the reliability and predictivity of a QSAR model. The leave-one-out (LOO) cross-validation was applied as the internal validation, and the predictions to an independent test set was applied as the external validation. The goodness of fit of each regression model was assessed by adjusted R^2 (R^2_{adj}) and cross-validation R^2 (q^2) on the training set. Meanwhile, the adjusted R^2 on the test set (q^2_{ext}) was used to evaluate the external predictive power of each model. The acceptability thresholds of q^2 for the training set and q^2_{ext} for the test set were both set to ≥ 0.5. A model is overfitted when the difference between R^2_{adj} and q^2_{ext} is higher than 0.3.[75,76]

Moreover, other two statistics, mean absolute error (MAE) and root-mean-square error (RMSE), were used to evaluate the quality of each regression model.[44,77]

Each classifier was evaluated by the following statistics based on the confusion matrix:[77−82] sensitivity (SE), specificity (SP), concordance or global accuracy (GA), balanced accuracy (BA), precision or positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR), false discovery rate (FDR), false negative rate (FNR), detection rate (DR), F-measure (F), G-means (G), Cohen's kappa coefficient (κ) and Matthews correlation coefficient (MCC). More detailed description can be found in Supporting Information.

Among them, MCC and κ are mainly used to measure the classification quality for an imbalanced data set. They both range from −1 to 1, and a perfect classification gives a value of 1 while a random classification gives a value of 0. In addition, the classification capability was measured by the area under the receive operating characteristic (ROC) curve (AUC), which is a graphical plot to illustrate the classification performance by changing its discrimination threshold.

The chemicals with large prediction errors (MAE > 0.6) and the misclassified chemicals were examined after modeling.

Analysis of Application Domain (AD). Because the training set for QSAR modeling probably does not cover the entire chemical space, the query chemicals may be unsuitable to be predicted and interpreted using a developed model, and therefore the AD for a model should be defined.[83] As a result, only a certain fraction of the query chemicals fall within the AD or coverage of the interpolation space, and only the predictions for such chemicals are reliable. In our study, the standard deviation distance to model (STD-DM) approach was applied to estimate the AD of each regression model. More detailed description of the STD-DM algorithm to define AD can be found in Supporting Information. The margin range of AD is defined as three times of the STD-DM value.[84] When a chemical is outside the AD, its STD-DM value is higher than the margin range.

Meanwhile, the Hotelling's test and the related leverage statistics were applied to determine the AD.[85,86] The leverage value h_i measures the distance between the *i*th compound and the centroid of its training set, and it is between 0 and 1. A threshold value (h^*) is generally fixed at $3(p+1)/n$, where p is the number of descriptors, and n is the compound number of the training set. A leverage value higher than h^* suggests this chemical may be outside the AD.

The Williams plot is a plot of standardized residuals versus leverage values, and it was used to intuitively confirm a visual image of the outliers. Among the outliers, the response outliers (Y outliers) show the standardized residuals higher than 3.0. The structurally influential outliers (X outliers) show the leverage values higher than the threshold value ($h_i > h^*$) and relatively low standard deviation. The Cook's distance calculated from the Williams plot was applied to estimate the influence of a single observation to the model. The calculation details are provided in Supporting Information. The cutoff of the Cook's distance is defined as $4/(n-p-1)$, and the compounds with Cook's distance higher than the cutoff value are marked as the highly influential points of the model. All the AD definition in our study is based on the regression analysis, so they are only suitable for the quantitative models. The AD analysis was also implemented in R software (version 3.2.5 ×64).

Analysis of Structural Alerts by SARpy. The structural alerts for urinary tract toxicity were extracted and analyzed by using the SARpy software (version 1.0).[87−89] The urinary tract toxicants were assigned as ACTIVE compounds, and the urinary tract nontoxicants were assigned as INACTIVE compounds. Then the analysis was produced using ACTIVE only mode. The substructures for each chemical were generated by simulating the bond breakage directly from the SMILES representations of the trained chemicals. The generated fragments were rearranged as the SMARTS representations. Then evaluation and exception induction were performed iteratively. The likelihood ratio of each substructure architecture was computed and summarized. After that, the rule set was extracted. Finally, the rule set was used to predict the query data set, and the confusion matrix of the predictions was computed.

■ RESULTS AND DISCUSSIONS

Property Analysis of Mouse Intraperitoneal Urinary Tract Toxicity Data. In our study, 258 organic compounds

with their urinary tract toxicity data were collected from the ChemIDplus database.[31] The concrete toxicological end points include changes in tubules (including acute renal failure, acute tubular necrosis), inflammation, necrosis or scarring of bladder, changes in bladder weight, changes in both tubules and glomeruli, changes primarily in glomeruli, hematuria, incontinence, urine volume decreased, urine volume increased, proteinuria, other changes in urine composition, and other changes.

After data splitting, for both regression and classification modeling, the training and test sets include 193 and 65 compounds, respectively (Figure 2A and B). As shown in the relative Euclidean distance plot of the compounds calculated from the selected molecular descriptors for both regression and classification modeling (Figure 2C and D), the compounds in the test set are basically distributed within the chemical space of the training set, and therefore it is reasonable to assess the prediction performance and generalization ability of the QSAR models using the test set. In addition, as shown in Figure 2B, the class distributions of the training and test sets are imbalanced with less urinary tract nontoxicants. So the SMOTE method was employed to mitigate the imbalance problem in the classification modeling. In our study, the 120 molecular descriptors identified by the chi-squared test were used for dimension reduction using the RFE-RF method. As shown in Figure 3, the RMSE of the training set decreases gradually and
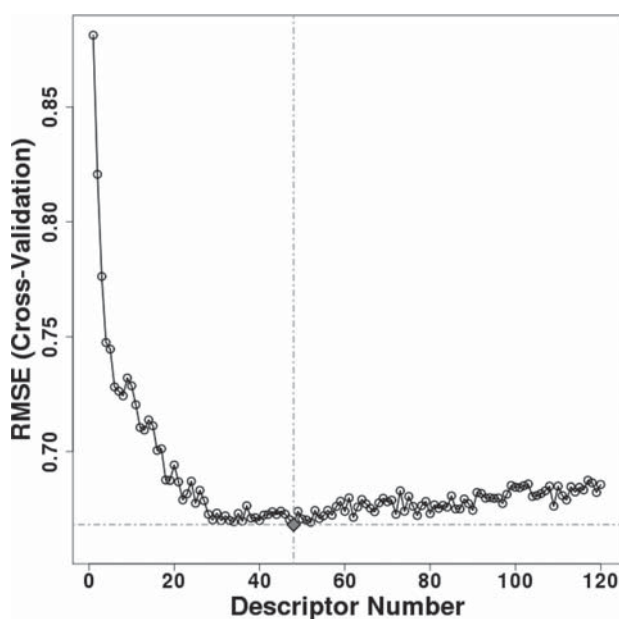
Figure 3. Dimension reduction result of the regression modeling based on the RFE-RF method (10-fold cross validation).

reaches stability at 48 descriptors, and therefore these 48 representative descriptors were selected for regression modeling. The definitions of the 48 representative molecular descriptors are summarized in Table S1. Among them, 20 descriptors are related to the atomic distributions of partial charges, such as PEOE_RPC +, FASA-, Q_VSA_PNEG, PEOE_VSA-1, PEOE_VSA_FPOL, etc. Six descriptors characterize molecular refractivity and polarizability, including SMR_VSA7, BCUT_SMR_0, BCUT_SMR_3, GCUT_SMR_0, GCUT_SMR_1, and GCUT_SMR_2. Eight descriptors describe hydrophilicity or solubility, including logP(o/w), SlogP, SlogP_VSA9, GCUT_SLOGP_0, GCUT_-SLOGP_3, BCUT_SLOGP_0, GCUT_SLOGP_1, and logS.

Besides, 14 descriptors are related to the van der Waals surface areas for different properties, such as SMR_VSA7, SlogP_VSA9, FASA-, Q_VSA_PNEG, PEOE_VSA-1, etc. Furthermore, number of some atoms (a_nN and a_nCl), number and fraction of rotatable bonds (opr_nrot and b_1rotR), bond potential energy (E_stb and E_str), solvation energy (E_sol), principal moment of inertia (pmiZ), atomic connectivity and molecular shape (petitjeanSC, radius, KierA2, and KierA3) are also relatively important. In summary, partial charges, molecular refractivity and polarizability, hydrophilicity and molecular shape and flexibility contribute largely to urinary tract toxicity, implying that urinary tract toxicity may be caused by the redox reactivity[8,90−93] and membrane permeability[9,94,95] of chemicals. It appears that all the 48 representative descriptors show varying degrees of importance contributing to urinary tract toxicity, as shown in Figure 4.
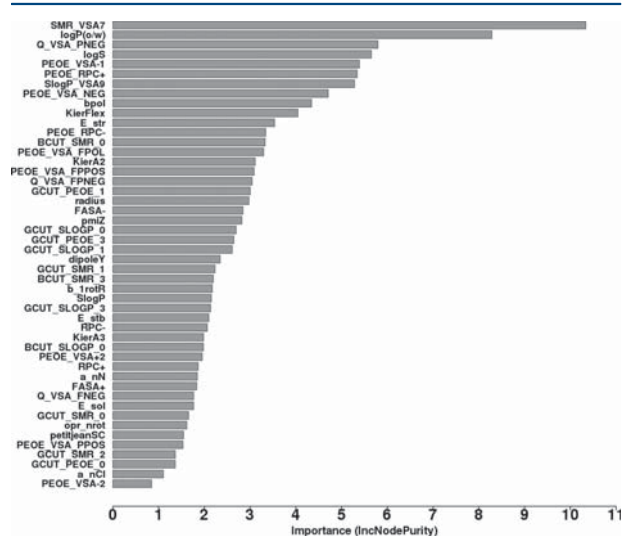

Figure 4. Importance of the representative molecular descriptors for regression selected by RFE-RF based the decrease of node impurity (mean squared error).

Similar to the 48 descriptors selected for regression modeling, the same procedure was conducted to determine the 110 descriptor subset for classification as shown in Figure 5. The importance of the 110 descriptors is shown in Figure S1. The top three important descriptors are Q_VSA_PNEG, logP(o/w) and PEOE_VSA_NEG. A large portion of the selected descriptors is closely related to atomic distribution of partial charges, molecular refractivity, hydrophilicity, and bond energies. Of course, some of the selected descriptors contribute to atomic connectivity and molecular shape. Briefly speaking, atomic partial charges of molecules, molecular refractivity and polarizability, solubility, and molecular flexibility also play a vital role in urinary tract toxicity. This also indicates the important contributions of redox reactivity[8,90−93] and membrane permeability[9,94,95] to urinary tract toxicity of chemicals.

Comparison of Various Regression Models for Urinary Tract Toxicity. After dimension reduction, a subset of 48 molecular descriptors was applied to develop the regression models using six machine learning approaches, i.e., RVM, SVM, RRF, XGBoost, RVMBoost, and SVMBoost. The statistical results for the training and test sets given by the optimal regression models are listed in Table 3.
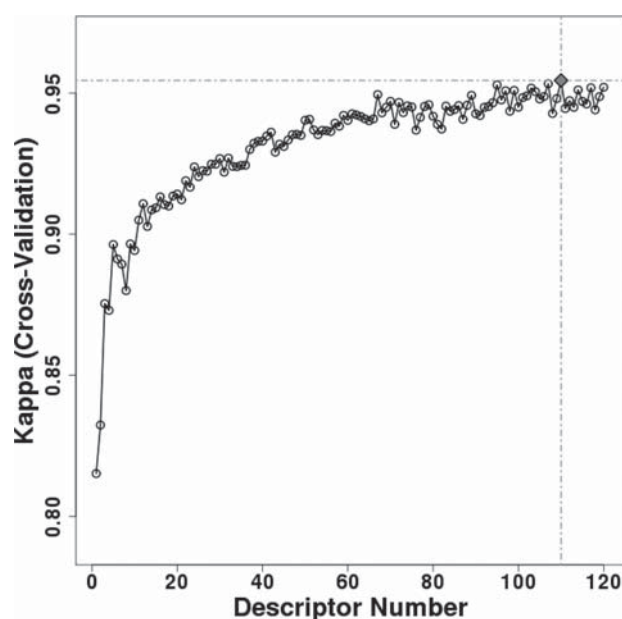
Figure 5. Dimension reduction result of the classification modeling based on the RFE-RF method (10-fold cross validation).

The performances of the six statistical approaches are quite diverse. According to the external validations, the rbfSVMBoost model achieves the best performance (q_{ext}^2 = 0.845) among all the regression models. The ranks of the prediction capability of these models are rbfSVMBoost > rbfRVMBoost > lpSVMBoost > XGBoost > rbfSVM > lpSVM > lpRVM > RRF > rbfRVM. The RMSE and MAE show the similar tends to q_{ext}^2. The prediction capability of the rbfRVMBoost model (q_{ext}^2 = 0.819) is the second best among the three AdaBoost based ensemble models, and is much better than that of the model built by rbfRVM (q_{ext}^2 = 0.678). The other two AdaBoost based ensemble models show less improvement than the rbfRVMBoost model. It seems that the weaker the individual RBF kernel learner is, the larger the boosting trick improves them. From the performances of the SVM and RVM models, we can see that the selection of the kernel function has apparent influence on the prediction capability of the developed models. According to the performances of the XGBoost and RRF models that are both decision tree based ensembles, we can find that the boosting trick may lead to larger improvement than the bagging (bootstrap aggregation) and random subspace methods. In boosting, the individual learners are trained sequentially. When observations are predicted with more deviations in the former learners, they

are more likely to appear in the training set of the subsequent learners and result in higher weights. Thus, boosting can present good generalization to the tested data, and avoid over-fitting. In summary, considering the overall statistics and prediction accuracy, the rbfSVMBoost approach is the best choice to develop the regression model for predicting mouse intra-peritoneal urinary tract toxicity. The scatter plots of the experimental pLD_{50} versus the predicted values given by the rbfSVMBoost model for the training and test sets are presented in Figure 6.

The AD coverages defined by STD-DM approach for all the regression models are 100%. The full AD coverage for the test set indicates the property distributions of the test set fall within the scope of the training set. In our study, the AD was also determined by the leverage approach, and the Williams plot and Cook's distance plot are shown in Figure 7. According to the Williams plot, three toxic chemicals can be considered as the response outliers because their leverage values are higher than the warning limit (h^* = 0.762). They are two organic platinum salts (CAS No. 20647-46-5, 89497-83-6) and an intermediate chemical (CAS No. 1192-76-3). There are four influential outliers outside the predicted residual threshold, and they are two urinary tract toxic chemicals with false residuals and two urinary tract nontoxic chemicals with positive residuals. The two toxic chemicals are cylindrospermopsin from cyanobacteria (CAS No. 143545-90-8) and reumycin from actinomyces (CAS No. 5016-18-2), and the two nontoxic chemicals are a sodium salt of saccharic acid (CAS No. 92413-99-5) and a platinum salt (CAS No. 75880-85-2). Otherwise, according to the Cook's distances of the chemicals in the training set, ten highly influential chemicals may greatly distort the regression. The top three highly influential chemicals are cylindrospermopsin (CAS No. 143545-90-8), 1,1′-sulphinylbis(aziridine) (CAS No. 1192-76-3), and cis-bis(2-norbornaneammine)dichloro-platinum(II) (CAS No. 75880-85-2).

As far as we know, the regression models for predicting urinary tract toxicity have never been reported. Therefore, our study is the first attempt to build the regression models for urinary tract toxicity, and the optimal model achieves excellent performance.

Comparison of Various Classification Models for Urinary Tract Toxicity. Besides the regression models, seven machine learning approaches were employed to develop the classification models based on the 110 finally selected descriptors, i.e. RVM, SVM, RRF, C5.0, XGBoost, AdaBoost.M1, and SVMBoost. The prediction statistics for the training and test sets given by the optimal classification models are summarized in Table 4.

Table 3. Statistical Results of the Regression Models for the Training and Test Sets

	R_{adj}^2 [a]	q_{LOO}^2 [b]	q_{ext}^2 *	$RMSE_{train}$	$RMSE_{test}$	MAE_{train}	MAE_{test}
rbfRVM	0.775	0.767	0.678	0.412	0.559	0.318	0.446
lpRVM	0.928	0.926	0.749	0.236	0.507	0.185	0.411
rbfSVM	0.976	0.975	0.798	0.148	0.437	0.139	0.347
lpSVM	0.988	0.986	0.790	0.098	0.478	0.035	0.386
RRF	0.957	0.943	0.740	0.256	0.557	0.192	0.427
XGBoost	0.969	0.948	0.811	0.173	0.448	0.141	0.364
rbfSVMBoost	0.911	0.904	0.845	0.270	0.396	0.196	0.319
lpSVMBoost	0.937	0.907	0.812	0.220	0.412	0.089	0.335
rbfRVMBoost	0.822	0.819	0.819	0.276	0.370	0.404	0.245

[a] q_{LOO}^2 refers to the cross-validation R^2 on the training set. [b] q_{ext}^2 refers to the adjusted R^2 on the test set.
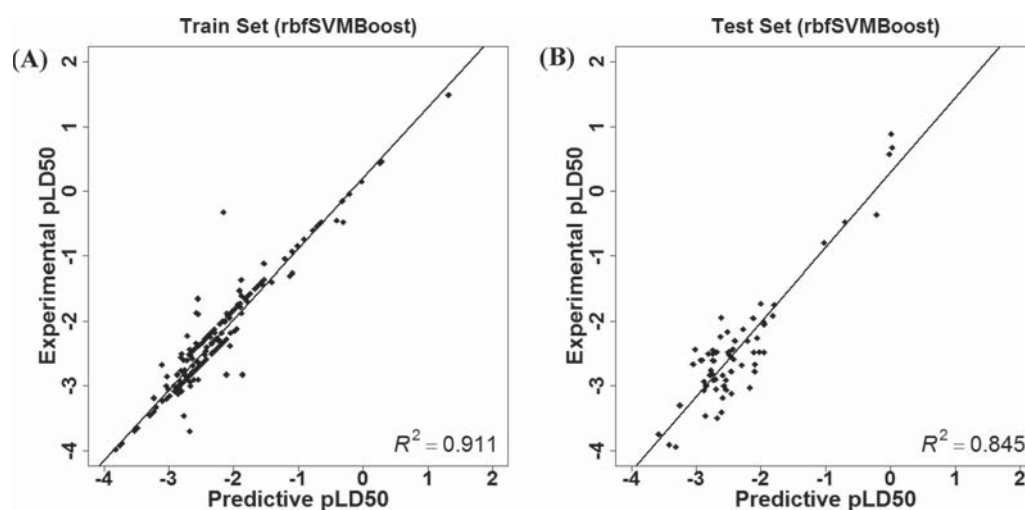
Figure 6. Scatter plots of the experimental pLD$_{50}$ values versus the predicted values for the chemicals in the (A) training and (B) test sets given by the rbfSVMBoost model in Table 3.
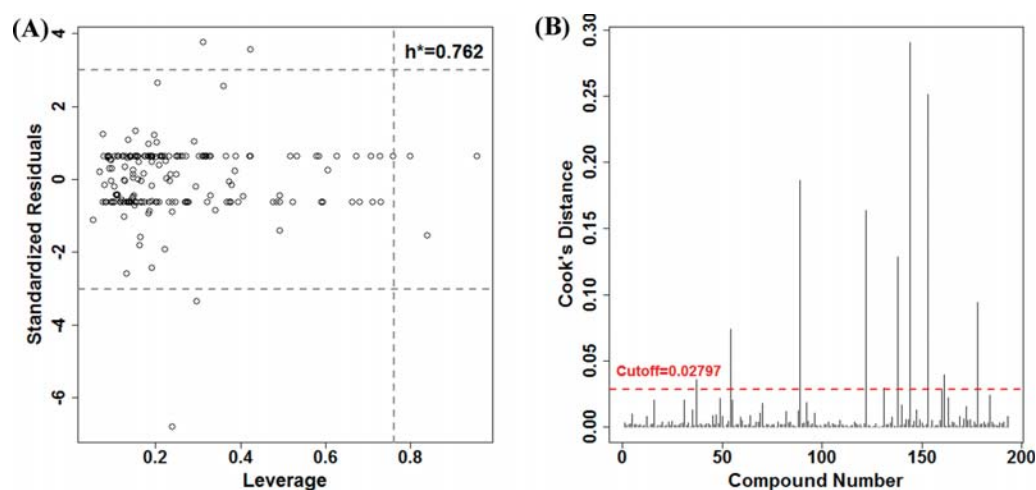

Figure 7. Application domain defined in our study. (A) Williams plot and (B) Cook's distance plot (B) were given by the leverage approach.

According to the statistical results, the rbfSVMBoost approach performs best, and the RRF approach performs worst. Based on the MCC values in combination with other evaluation measures, the predictive performances of all the classification models are ranked from the best to the worst as rbfSVMBoost > lpSVMBoost ≈ rbfSVM > lpSVM > AdaBoost.M1 > rbfRVM > XGBoost > C5.0 > RRF. All the models were LOO cross-validated. Among all the classification models, the rbfSVMBoost model gives the best prediction for the test set (MCC = 0.787). This model also achieves the second highest sensitivity of 89.58%, the highest specificity of 94.12%, and the highest global accuracy of 90.77% for the test set. It gives high AUC measures for both the training (1.000) and test sets (0.893). The F1 measures, G-means, and Cohen's κ of the model show the same trends as well. The prediction capabilities of the lpSVMBoost and rbfSVM models are similar, and they have opposite performance on the urinary tract toxicants and nontoxicants for the test set. The lpSVMBoost model shows more accurate predictions for the urinary tract toxicants, and the rbfSVM model shows more accurate predictions for the urinary tract non-toxicants. Besides, the rbfSVMBoost and lpSVMBoost approaches both show obvious improvement than their individual learners,

suggesting that the boosting strategy brings large improvement to SVM. In addition, the boosting approach exhibits the best prediction performance among the decision tree based approaches. Therefore, boosting is a viable approach to reduce the prediction errors and improve weak learners. In conclusion, the rbfSVMBoost model yields lower false detection rate and false omission rate than the other models, and therefore it is the best choice.

Among all the classification models, SVM, AdaBoost.M1 and SVMBoost approaches show relatively strong prediction capabilities, but RVM, XGBoost, C5.0 and RRF give bad performance. This indicates that SVM and AdaBoost based approaches are suitable for classifying complicated toxicological toxicants, although the kernel function of SVM should be chosen carefully. Boosting or other better ensemble methods show much improved performance over individual learners significantly in many cases.[67−69,96]

In summary, the rbfSVMBoost approach is recommended to develop the classification models for the prediction of mouse intraperitoneal urinary tract toxicity. The ROC curves of sensitivities versus false positive rates given by the rbfSVMBoost model (Table 4) for the training and test sets are shown

Table 4. Statistical Results of the Classification Models for the Training and Test Sets

models[a]		SE	SP	GA	BA	PPV	NPV	FPR	FDR	FNR	DR	F1	G	κ	MCC	AUC
AdaBoost.M1	TR	100.00%	97.44%	98.45%	98.72%	96.20%	100.00%	2.56%	3.80%	0.00%	39.38%	0.981	0.981	0.968	0.968	0.994
	TE	88.24%	81.25%	83.08%	84.74%	62.50%	95.12%	18.75%	37.50%	11.76%	23.08%	0.732	0.743	0.613	0.633	0.848
C5.0	TR	100.00%	96.58%	97.93%	98.29%	95.00%	100.00%	3.42%	5.00%	0.00%	39.38%	0.974	0.975	0.957	0.958	0.998
	TE	82.35%	75.00%	76.92%	78.68%	53.85%	92.31%	25.00%	46.15%	17.65%	21.54%	0.651	0.666	0.490	0.514	0.815
rbfSVM	TR	98.68%	98.29%	98.45%	98.49%	97.40%	99.14%	1.71%	2.60%	1.32%	38.86%	0.980	0.980	0.968	0.968	0.992
	TE	82.35%	89.58%	87.69%	85.97%	73.68%	93.48%	10.42%	26.32%	17.65%	21.54%	0.778	0.779	0.693	0.695	0.868
lpSVM	TR	98.68%	98.29%	98.45%	98.49%	97.40%	99.14%	1.71%	2.60%	1.32%	38.86%	0.980	0.980	0.968	0.968	0.999
	TE	94.12%	81.25%	84.62%	87.68%	64.00%	97.50%	18.75%	36.00%	5.88%	24.62%	0.762	0.776	0.654	0.681	0.876
rbfRVM	TR	85.86%	82.89%	84.59%	84.38%	87.00%	17.11%	13.00%	81.47%	14.14%	49.06%	0.864	0.864	0.686	0.686	0.920
	TE	89.58%	70.59%	84.62%	80.09%	89.58%	29.41%	10.42%	70.59%	10.42%	66.15%	0.896	0.896	0.602	0.602	0.857
RRF	TR	100.00%	98.29%	98.96%	99.15%	97.44%	100.00%	1.71%	2.56%	0.00%	39.38%	0.987	0.987	0.978	0.979	0.993
	TE	64.71%	83.33%	78.46%	74.02%	57.89%	86.96%	16.67%	42.11%	35.29%	16.92%	0.611	0.612	0.463	0.464	0.692
XGBoost	TR	100.00%	96.58%	97.93%	98.29%	95.00%	100.00%	3.42%	5.00%	0.00%	39.38%	0.974	0.975	0.957	0.958	1.000
	TE	76.47%	85.42%	83.08%	80.94%	65.00%	91.11%	14.58%	35.00%	23.53%	20.00%	0.703	0.705	0.586	0.589	0.853
rbfSVMBoost	TR	97.44%	100.00%	98.45%	98.72%	100.00%	96.20%	0.00%	0.00%	2.56%	59.07%	0.987	0.987	0.968	0.968	1.000
	TE	89.58%	94.12%	90.77%	91.85%	97.73%	76.19%	5.88%	2.27%	10.42%	66.15%	0.935	0.936	0.778	0.787	0.893
lpSVMBoost	TR	98.29%	100.00%	98.96%	99.15%	100.00%	97.44%	0.00%	0.00%	1.71%	59.59%	0.991	0.991	0.978	1.000	1.000
	TE	89.58%	82.35%	87.69%	85.97%	93.48%	73.68%	17.65%	6.52%	10.42%	66.15%	0.915	0.915	0.693	0.695	0.846

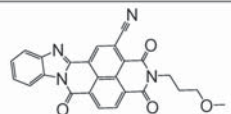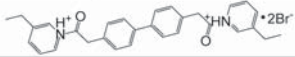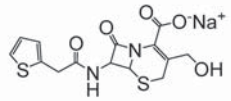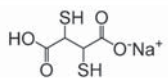[a]TR refers to the training set, and TE refers to the test set.
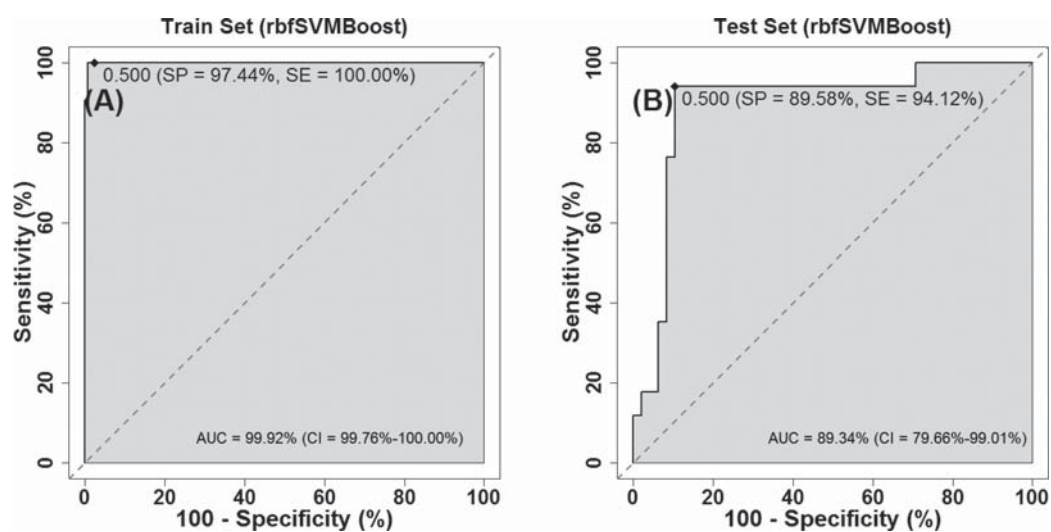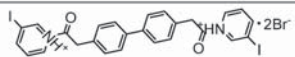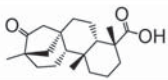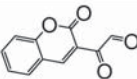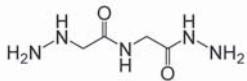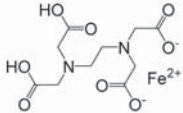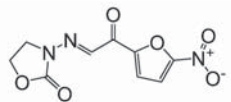
Figure 8. ROC curves for the (C) training and (D) test sets given by the rbfSVMBoot model in Table 4.

Table 5. Experimental and Predicted LD$_{50}$ Values for the Eight Chemicals with Large Prediction Errors in the Training Set Predicted by the rbfSVMBoost Regression Model (MAE > 0.6)

No.	Structure	Experimental pLD$_{50}$	Predicted pLD$_{50}$	MAE
1		-0.322	-2.157	1.834
2		-3.699	-2.677	1.022
3		-2.826	-1.859	0.967
4		-1.650	-2.552	0.901
5		-2.826	-2.107	0.719
6		-3.462	-2.765	0.698
7	(trans-erythro-)	-1.875	-2.571	0.696
8		-1.892	-2.545	0.653

Table 6. Experimental and Predicted LD$_{50}$ Values for the Ten Chemicals with Large Prediction Errors in the Test Set Predicted by the rbfSVMBoost Regression Model (MAE > 0.6)

No.	Structure	Experimental pLD$_{50}$	Predicted pLD$_{50}$	MAE
1		0.886	0.011	0.875
2		-3.029	-2.170	0.859
3		-3.498	-2.676	0.822
4		-3.411	-2.608	0.803
5		-2.778	-2.095	0.683
6		-1.944	-2.612	0.667
7		-3.121	-2.455	0.665
8		0.678	0.028	0.650
9		-3.942	-3.314	0.628
10		-3.462	-2.858	0.604
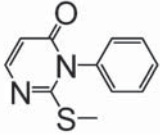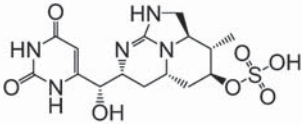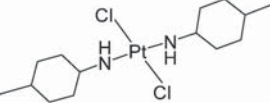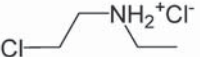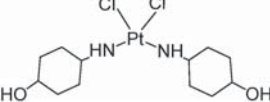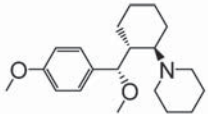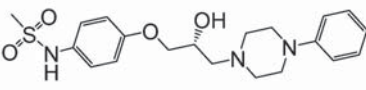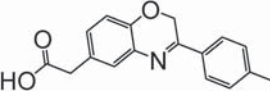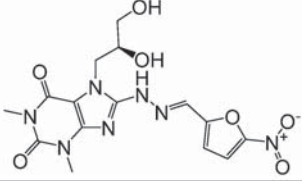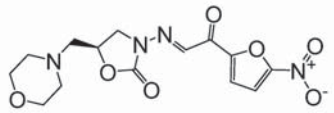
in Figure 8, respectively. The plots show considerably large areas under the curves, suggesting that the optimal rbfSVMBoost model is an ideal classifier.

Analysis of Chemicals with Large Prediction Errors. As stated in the regression analysis, the rbfSVMBoost model generates the best prediction performance for the test set. But some chemicals in the test set still cannot be well predicted. Although the Williams plot and Cook's distance plot were employed to detect the outliers and highly influential points, the MAE statistics was also used to analyze the outlying degree of these outliers. When MAE > 0.6 was defined as the threshold, the MAEs of the chemicals with large prediction errors predicted by the rbfSVMBoost model in Table 3 range from 0.604 to 0.875 for the test set, and those range from 0.653 to 1.834 for the training set. In total, eight chemicals in the training set (Table 5) and ten chemicals in the test set (Table 6) cannot be well predicted by the rbfSVMBoost model in Table 3. Among the chemicals with large prediction errors in the training and test sets (Tables 5 and 6), many of them form salts, and have a lot of heterocyclic amines, amino bonds and carboxyl groups. These charged chemicals are easy to be metabolized and excreted *in vivo*, and their toxicity may occur in these processes and be influenced by the atomic charges distributed

on the molecular surface. In addition, it seems that some molecules with symmetrical organic components cannot be well predicted. For example, No. 3 and No. 6 chemicals in Table 5, as well as No. 1, No. 4, No. 5, No. 8, and No. 10 chemicals in Table 6, contain symmetrical organic components. Moreover, we discovered that the top six chemicals in Table 5 are all the influential chemicals determined by the Cook's distance. So the Cook's distance could be used to check how accurately the query chemicals are predicted by the rbfSVMBoost regression model.

Among the 14 chemicals in the test set with large prediction errors predicted by the XGBoost model (Table S2), four chemicals have the hemicholinium-3 derived scaffold 2,2′-([1,1′-biphenyl]-4,4′-diyl)bis(1-(1λ^4-pyridin-1-yl)ethan-1-one). So the chemicals with this scaffold may be hard to be predicted accurately. Moreover, we found that more chemicals with large prediction errors predicted by the XGBoost regression model tend to have the same scaffold than those predicted by the rbfSVMBoost regression model, which may be caused by allocating similar observations to the same learner nodes in every iteration. Thus, SVMs used as the weak learners in boosting could give better generalization to molecular scaffolds than decision trees.

Table 7. Misclassified Chemicals in the Training and Test Sets Predicted by the rbfSVMBoost Classification Model
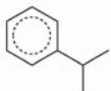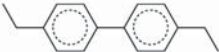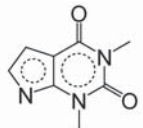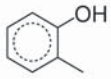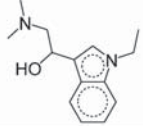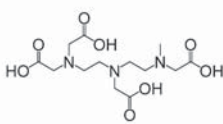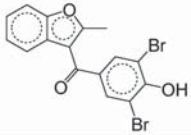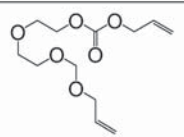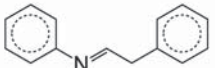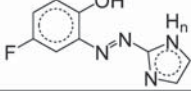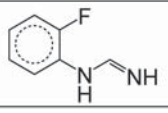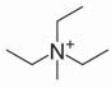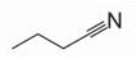
No.	Molecules	Name	CAS No.	LD$_{50}$ (mg/kg)	Effects	Classification Results
		Training Set				
1		2-(methylthio)-3-phenyl-4(3H)-pyrimidinone	89069-18-1	400	Urine volume increased	FN
2		cylindrospermopsin	14354 5-90-8	2.1	Other changes	FN
3		trans-bis(4-methylcyclohexylammine)-dichloro-platinum (II)	56815-77-1	360	Other changes	FN
		Test Set				
1		2-chloro-diethylamine hydrochloride	4535-87-9	1120	Other changes	FP
2		cis-bis(4-hydroxycyclohexylammine)-dichloro-platinum (II)	62905-79-7	23	Other changes	FN
3		trans-erythro-1-(2-(p,alpha-dimethoxybenzyl)cyclohexyl)-piperidine	13724-66-8	75	Urine volume increased	FN
4		N-(4-(2-hydroxy-3-(4-phenyl-1-piperazinyl)propoxy)phenyl)-methanesulfonamide	64511-28-0	37.5	Urine volume increased	FN
5		3-(4-methylphenyl)-2H-1,4-benzoxazine-6-acetic acid	86818-28-2	316	Urine volume increased	FN
6		5-nitro-2-furaldehyde-(7-(2,3-dihydroxypropyl)-1,3-dimethyl-2,6-dioxo-1,2,3,6-tetrahydropurin-8-yl)hydrazone	78960-58-4	221	Urine volume increased	FN
7		5-(morpholinomethyl)-3-(((5-nitro-2-furoyl)methylene)amino)-2-oxazolidinone	92297-09-1	64	Other changes in urine composition	FN

Analysis of Misclassified Chemicals by rbfSVMBoost. The chemicals misclassified (false positives and false negatives) by the best rbfSVMBoost classifier were analyzed and summarized in Table 7. There is only one false positive in the test set, and the others are all false negatives. This is because that

the lack of enough urinary tract nontoxicants in the training set results in the bias. Two chemicals (CAS No. 143545-90-8 and 13724-66-8) can be found in both Tables 6 and 7. The misclassification of CAS No. 143545-90-8 may be caused by the triazaacenaphthylene group, and that of CAS No. 13724−66−8

Table 8. Structural Alerts for Urinary Tract Toxicity Analyzed by SARpy

No.	SMARTS	Structural Alerts	Number of Chemicals in the Total Data		No.	SMARTS	Structural Alerts	Number of Chemicals in the Total Data	
			Toxic	Non-toxic				Toxic	Non-toxic
1	C([NH](C)CC)	NH	11	0	2	CC(C)(c1ccc(cc1))		12	0
3	CCc1ccc(cc1)c1ccc(cc1)CC		13	1	4	n1ccc2c1n(c(=O)n(c2=O)C)C		8	2
5	NN=CC	H2N	10	1	6	CSC	S	9	3
7	NCN	H2N NH2	29	11	8	C=NN	H2N N	18	7
9	c1ccccn1	N	4	2	10	C(C(Cc1c(ccc1)))(C))		9	3
11	n1(c(nccc1=O))	N N O	20	5	12	Cc1cc(c(cc1O))	OH	8	6
13	Cl	Cl	59	20	14	c12c(c(C(CN(C)C)O)cn1CC)cccc2	N HO	2	0
15	C(N(CCN(CC(=O)O)CC(=O)O)CC(=O)O)CN(CC(=O)O)C	HO O OH N N OH O	2	0	16	c1(c2c(oc1C)ccc2)C(=O)c1cc(c(c(c1)Br)O)Br	O Br OH Br	1	0
17	O(C(=O)OCC=C)CCOCCOCOCC=C	O O O O O	1	0	18	[Fe]12345(C6(C(=O)C)C1C2C3C46C1(C(=O)C)C5CCC1	O O	1	0
19	c1(ccccc1)CC(=Nc1ccccc1)	N	1	0	20	Oc1ccc(cc1N=Nc1[nH]ccn1)F	OH F N N N Hn	1	0
21	C(=N)Nc1ccccc1F	F NH H	1	2	22	[N+](CC)(CC)(CC)C	N+	3	0
23	C(CC#N)C	N	1	0	24	N1CC1	HN	4	0

may be caused its special spatial configuration that cannot be well characterized by regular molecular descriptors. Besides, all the chemicals except the first chemical in Table 7 have more than two rotatable bonds, which may lead to inaccurate descriptions of molecular surface.

Analysis of Structural Alerts of Urinary Tract Toxicity. At last, we used the SARpy software (version 1.0) to analyze the structural alerts for urinary tract toxicity. The confusion matrix is summarized in Table S3, and the corresponding structural alerts with their occurrence in the total data are summarized

in Table 8. We can see that there are many groups containing nitrogen atoms and benzene rings. There is a false structural alert in these substructures, i.e. N-(2-fluorophenyl)formamidine (No. 21 fragment), which exists in one toxicants and two nontoxicants. In these structural alerts, some fragments have been proved to be nephrotoxic. For example, No.1 fragment N,N-dimethylethylamine is an aliphatic triamine that may induce kidney edema and other adverse events.[97−99] In addition, many fragments shown in Table 8 contain amine. No. 2 fragment cumene is a potential carcinogen.[90,100] No.3 fragment

4,4′-diethylbiphenyl is part of hemicholinium-like neuromuscular blocker that may cause renal vasoconstriction and urinary volume increase.[101,102] No. 6 fragment thioether shows nephrotoxicity according to the previous studies.[25,103,104] No. 12 fragment o-cresol is a well-known uremic toxicant.[105] With these structural alerts, we can quickly identify potential chemicals that may cause urinary tract toxicity.

CONCLUSIONS

In our study, nine regression models and nine classification models were developed based on an extensive data set of mouse intraperitoneal urinary tract toxicity. The RFE-RF method was applied to determine an optimal subset of 48 molecular descriptors for regression and an optimal subset of 110 molecular descriptors for classification, respectively. Then, six machine learning approaches were employed to build the regression models for urinary tract toxicity. Considering the overall prediction accuracy for the test set, the rbfSVMBoost approach achieves the best prediction, and reaches R_{adj}^2 of 0.911 for the training set and q_{ext}^2 of 0.845 for the test set. Furthermore, seven machine learning approaches were employed to develop the classification models for urinary tract toxicity. Taking account of MCC statistics in combination with other confusion matrix metrics, the rbfSVMBoost model achieves better prediction than the others. The optimal rbfSVMBoost model gives MCC of 0.787, AUC of 0.893, sensitivity of 89.58%, specificity of 94.12%, and global accuracy of 90.77% for the test set. We also analyzed the chemicals with large prediction errors in the regression analysis and the misclassified chemicals. At last, we used the SARpy software to extract the structural alerts for urinary tract toxicity. According to the results of this study, the two successful models, the rbfSVMBoost regression model and the rbfSVMBoost classification model, can be used as reliable prediction tools for the assessment of chemical-induced urinary tract toxicity.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.molpharmaceut.7b00631.

> QSAR modeling by machine learning approaches; evaluation and validation of the QSAR models; analysis of application domain (AD); distribution plot of the compounds in the training and test sets; importance of molecular descriptors of classification selected by RFE-RF based the decrease of node impurity (Gini index); definitions of the final selected 48 descriptors for regression; experimental and predicted LD_{50} values for the 14 tested chemicals with large prediction errors predicted by the XGBoost regression model (MAE > 0.6); and confusion matrix built by the prediction of SARpy (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: tingjunhou@zju.edu.cn or tingjunhou@hotmail.com; Phone: +86-571-88208412.

ORCID Ⓞ

Tailong Lei: 0000-0003-2067-1787
Huiyong Sun: 0000-0002-7107-7481
Youyong Li: 0000-0002-5248-2756
Tingjun Hou: 0000-0001-7227-2580

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

ABBREVIATIONS

LD_{50}, median lethal dose; QSAR, quantitative structure—activity relationship; OECD, the Organization for Economic Co-operation and Development; EPA, the U.S. Environmental Protection Agency; SMOTE, synthetic minority oversampling technique; LOO, leave-one-out; AD, application domain; MOE, molecular operating environment; RFE, recursive feature elimination; AdaBoost, adaptive boosting; RRF, regularized random forest; rbf, Gaussian radial basis kernel; lp, Laplacian kernel; SVM, support vector machine; RVM, relevance vector machine; XGBoost, extreme gradient boosting; R^2, coefficient of determination; MAE, mean absolute error; RMSE, root-mean-square error; TP, true positive; TN, true negative; FP, false positive; FN, false negative; SE, sensitivity; SP, specificity; GA, global accuracy or concordance; BA, balanced accuracy; PPV, positive predictive value or precision; NPV, negative predictive value; FPR, false positive rate; FDR, false discovery rate; FNR, false negative rate; DR, detection rate; F, F1-measure; G, G-means; κ, Cohen's kappa coefficient; MCC, Matthews correlation coefficient; ROC, the receive operating characteristic curve; AUC, the area under the ROC; STD-DM, standard deviation distance to model; ADMET, absorption, distribution, metabolism, excretion and toxicity; PCA, principal component analysis; SARPy, structure—activity relationships in python

REFERENCES

(1) Arrowsmith, J.; Miller, P. Trial Watch: Phase II and Phase III attrition rates 2011−2012. *Nat. Rev. Drug Discovery* **2013**, *12* (8), 569−569.

(2) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discovery* **2015**, *14* (7), 475−486.

(3) Harrison, R. K. Phase II and phase III failures: 2013−2015. *Nat. Rev. Drug Discovery* **2016**, *15* (12), 817−818.

(4) Pannu, N.; Nadim, M. K. An overview of drug-induced acute kidney injury. *Crit. Care Med.* **2008**, *36* (Suppl), S216−S223.

(5) Hornberg, J. J.; Laursen, M.; Brenden, N.; Persson, M.; Thougaard, A. V.; Toft, D. B.; Mow, T. Exploratory toxicology as an integrated part of drug discovery. Part I: Why and how. *Drug Discovery Today* **2014**, *19* (8), 1131−1136.

(6) Siramshetty, V. B.; Nickel, J.; Omieczynski, C.; Gohlke, B.-O.; Drwal, M. N.; Preissner, R. WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.* **2016**, *44* (D1), D1080−D1086.

(7) Choudhury, D.; Ahmed, Z. Drug-associated renal dysfunction and injury. *Nat. Clin. Pract. Nephrol.* **2006**, *2* (2), 80−91.

(8) Nolin, T. D.; Himmelfarb, J. Mechanisms of Drug-Induced Nephrotoxicity. In *Adverse Drug Reactions*; Uetrecht, J., Ed.; Springer: Berlin Heidelberg, 2010; pp 111−130.

(9) Kitterer, D.; Schwab, M.; Alscher, M. D.; Braun, N.; Latus, J. Drug-induced acid-base disorders. *Pediatr. Nephrol.* **2015**, *30* (9), 1407−1423.

(10) Kim, S.-Y.; Moon, A. R. Drug-induced nephrotoxicity and its biomarkers. *Biomol. Ther.* **2012**, *20* (3), 268−272.

(11) Luciano, R. L.; Perazella, M. A. Nephrotoxic effects of designer drugs: Synthetic is not better! *Nat. Rev. Nephrol.* **2014**, *10* (6), 314−324.

(12) Luyckx, V. A.; Naicker, S. Acute kidney injury associated with the use of traditional medicines. *Nat. Clin. Pract. Nephrol.* **2008**, *4* (12), 664−671.

(13) Taber, S. S.; Pasko, D. A. The epidemiology of drug-induced disorders: The kidney. *Expert Opin. Drug Saf.* **2008**, *7* (6), 679−690.

(14) Bonventre, J. V.; Vaidya, V. S.; Schmouder, R.; Feig, P.; Dieterle, F. Next-generation biomarkers for detecting kidney toxicity. *Nat. Biotechnol.* **2010**, *28* (5), 436−440.

(15) Gobe, G. C.; Coombes, J. S.; Fassett, R. G.; Endre, Z. H. Biomarkers of drug-induced acute kidney injury in the adult. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11* (11), 1683−1694.

(16) Fuhrman, D. Y.; Kellum, J. A. Biomarkers for diagnosis, prognosis and intervention in acute kidney injury. *Contrib. Nephrol.* **2016**, *187*, 47−54.

(17) Kataria, A.; Trasande, L.; Trachtman, H. The effects of environmental chemicals on renal function. *Nat. Rev. Nephrol.* **2015**, *11* (10), 610−625.

(18) Perazella, M. A. Renal vulnerability to drug toxicity. *Clin. J. Am. Soc. Nephrol.* **2009**, *4* (7), 1275−1283.

(19) Kandasamy, K.; Chuah, J. K. C.; Su, R.; Huang, P.; Eng, K. G.; Xiong, S.; Li, Y.; Chia, C. S.; Loo, L.-H.; Zink, D. Prediction of drug-induced nephrotoxicity and injury mechanisms with human induced pluripotent stem cell-derived cells and machine learning methods. *Sci. Rep.* **2015**, *5* (1), 12337.

(20) Pizzo, F.; Gadaleta, D.; Lombardo, A.; Nicolotti, O.; Benfenati, E. Identification of structural alerts for liver and kidney toxicity using repeated dose toxicity data. *Chem. Cent. J.* **2015**, *9* (1), 62.

(21) Su, R.; Xiong, S.; Zink, D.; Loo, L.-H. High-throughput imaging-based nephrotoxicity prediction for xenobiotics with diverse chemical structures. *Arch. Toxicol.* **2016**, *90* (11), 2793−2808.

(22) Su, R.; Li, Y.; Zink, D.; Loo, L.-H. Supervised prediction of drug-induced nephrotoxicity based on interleukin-6 and −8 expression levels. *BMC Bioinf.* **2014**, *15* (16), S16.

(23) Kim, J.; Shin, M. An integrative model of multi-organ drug-induced toxicity prediction using gene-expression data. *BMC Bioinf.* **2014**, *15* (16), S2.

(24) Zhang, J. D.; Berntenis, N.; Roth, A.; Ebeling, M. Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity. *Pharmacogenomics J.* **2014**, *14* (3), 208−216.

(25) Lee, S.; Kang, Y.-M.; Park, H.; Dong, M.-S.; Shin, J.-M.; No, K. T. Human nephrotoxicity prediction models for three types of kidney injury based on data sets of pharmacological compounds and their metabolites. *Chem. Res. Toxicol.* **2013**, *26* (11), 1652−1659.

(26) Minowa, Y.; Kondo, C.; Uehara, T.; Morikawa, Y.; Okuno, Y.; Nakatsu, N.; Ono, A.; Maruyama, T.; Kato, I.; Yamate, J.; Yamada, H.; Ohno, Y.; Urushidani, T. Toxicogenomic multigene biomarker for predicting the future onset of proximal tubular injury in rats. *Toxicology* **2012**, *297* (1−3), 47−56.

(27) Myshkin, E.; Brennan, R.; Khasanova, T.; Sitnik, T.; Serebriyskaya, T.; Litvinova, E.; Guryanov, A.; Nikolsky, Y.; Nikolskaya, T.; Bureeva, S. Prediction of organ toxicity endpoints by QSAR modeling based on precise chemical-histopathology annotations. *Chem. Biol. Drug Des.* **2012**, *80* (3), 406−416.

(28) Hammann, F.; Gutmann, H.; Vogt, N.; Helma, C.; Drewe, J. Prediction of adverse drug reactions using decision tree modeling. *Clin. Pharmacol. Ther.* **2010**, *88* (1), 52−59.

(29) Matthews, E. J.; Ursem, C. J.; Kruhlak, N. L.; Benz, R. D.; Sabaté, D. A.; Yang, C.; Klopman, G.; Contrera, J. F. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regul. Toxicol. Pharmacol.* **2009**, *54* (1), 23−42.

(30) Thukral, S. K.; Nordone, P. J.; Hu, R.; Sullivan, L.; Galambos, E.; Fitzpatrick, V. D.; Healy, L.; Bass, M. B.; Cosenza, M. E.; Afshari, C. A. Prediction of nephrotoxicant action and identification of candidate toxicity-related biomarkers. *Toxicol. Pathol.* **2005**, *33* (3), 343−355.

(31) U.S. National Library of Medicine *ChemIDplus database*. https://chem.nlm.nih.gov/chemidplus/ (accessed March 15, 2016).

(32) *MOE molecular simulation package*; Chemical Computing Group Inc.: Montreal, Candada, 2010.

(33) *Discovery Studio 2.5 Guide*; Accelrys Inc.: San Diego, CA, USA, 2009.

(34) U.S. EPA 40 CFR 156.64: Toxicity Category. http://www.ecfr.gov/cgi-bin/text-idx?SID=01a7e4bc780b0368240af36919256d1c&mc=true&node=se40.26.156_162&rgn=div84 (accessed March 5, 2016),.

(35) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389−422.

(36) Svetnik, V.; Liaw, A.; Tong, C.; Wang, T., Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In *MCS 2004: Multiple Classifier Systems*; Roli, F., Kittler, J., Windeatt, T., Eds.; Springer Berlin Heidelberg: 2004; pp 334−343.

(37) Granitto, P. M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.* **2006**, *83* (2), 83−90.

(38) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Choosing Feature Selection and Learning Algorithms in QSAR. *J. Chem. Inf. Model.* **2014**, *54* (3), 837−843.

(39) Brest, J.; Greiner, S.; Boskovic, B.; Mernik, M.; Zumer, V. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE T. Evolut. Comput.* **2006**, *10* (6), 646−657.

(40) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, *16* (1), 321−357.

(41) Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28* (5), 1−26.

(42) Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **2001**, *1* (3), 211−244.

(43) Burden, F. R.; Winkler, D. A. Relevance Vector Machines: Sparse Classification Methods for QSAR. *J. Chem. Inf. Model.* **2015**, *55* (8), 1529−1534.

(44) Lei, T.; Li, Y.; Song, Y.; Li, D.; Sun, H.; Hou, T. ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *J. Cheminf.* **2016**, *8* (1), 6.

(45) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20* (3), 273−297.

(46) Hou, T.; Wang, J.; Li, Y. ADME Evaluation in Drug Discovery. 8. The Prediction of Human Intestinal Absorption by a Support Vector Machine. *J. Chem. Inf. Model.* **2007**, *47* (6), 2408−2415.

(47) Zhou, S.; Li, G.-B.; Huang, L.-Y.; Xie, H.-Z.; Zhao, Y.-L.; Chen, Y.-Z.; Li, L.-L.; Yang, S.-Y. A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. *Comput. Biol. Med.* **2014**, *51*, 122−127.

(48) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.

(49) Sheridan, R. P. Using Random Forest To Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53* (11), 2837−2850.

(50) Deng, H.; Runger, G. Gene selection with guided regularized random forest. *Pattern Recogn.* **2013**, *46* (12), 3483−3489.

(51) Im, J.; Jensen, J. A change detection model based on neighborhood correlation image analysis and decision tree classification. *Remote Sens. Environ.* **2005**, *99* (3), 326−340.

(52) Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z.-H.; Steinbach, M.; Hand, D. J.; Steinberg, D. Top 10 algorithms in data mining. *Knowledge and Information Systems* **2008**, *14* (1), 1−37.

(53) Chen, X. Y.; Ma, L. Z.; Chu, N.; Zhou, M.; Hu, Y. Classification and progression based on CFS-GA and c5.0 boost decision tree of TCM Zheng in chronic hepatitis B. *Evid.-based Compl. Alt.* **2013**, *2013*, 1−9.

(54) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **2001**, *29* (5), 1189−1232.

(55) Singh, K. P.; Gupta, S. In silico prediction of toxicity of non-congeneric industrial chemicals using ensemble learning based modeling approaches. *Toxicol. Appl. Pharmacol.* **2014**, *275* (3), 198−212.

(56) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16*; Association for Computing Machinery (ACM): San Francisco, California, USA, 2016; pp 785−794.

(57) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J. S.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56* (12), 2353−2360.

(58) Freund, Y. Boosting a weak learning algorithm by majority. *Inform. Comput.* **1995**, *121* (2), 256−285.

(59) Freund, Y. S.; Schapire, R. E. In *Experiments with a new boosting algorithm*, Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, July 4, 1996; Morgan Kaufmann: Bari, Italy, pp 148−156.

(60) Li, H.; Sun, J.; Fan, X.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput.-Aided Mol. Des.* **2008**, *22* (11), 843−855.

(61) Niu, B.; Jin, Y.; Lu, W.; Li, G. Predicting toxic action mechanisms of phenols using AdaBoost Learner. *Chemom. Intell. Lab. Syst.* **2009**, *96* (1), 43−48.

(62) Freund, Y.; Schapire, R. E. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* **1999**, *14* (5), 771−780.

(63) Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Multi-class AdaBoost. *Stat. Interface* **2009**, *2* (3), 349−360.

(64) Drucker, H., Improving Regressors using Boosting Techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, Fisher, D. H., Ed.; Morgan Kaufmann Publishers Inc.: Nashville, Tennessee, USA, 1997; pp 107−115.

(65) Pavlov, D.; Mao, J.; Dom, B. In,*Scaling-up support vector machines using boosting algorithm*, Proceedings 15th International Conference on Pattern Recognition (ICPR-2000), Barcelona, Spain, September 3−7, 2000; Barcelona, Spain, 2002; pp 219−222.

(66) Chan, J. C.-W.C.; Huang, C.; DeFries, R. Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting. *IEEE T. Geosci. Remote* **2001**, *39* (3), 693−695.

(67) Wu, C.; Wang, X.; Zheng, C. In *Classifying HRRP by AdaBoostSVM*, 2008 9th International Conference on Signal Processing, Beijing, China, October 26−29, 2008; IEEE: Beijing, China, 2008; pp 1582−1585.

(68) Cheng, Z.; Zhang, Y.; Zhou, C.; Zhang, W.; Gao, S. Classification of 5-HT1A receptor ligands on the basis of their binding affinities by using PSO-Adaboost-SVM. *Int. J. Mol. Sci.* **2009**, *10* (8), 3316−3337.

(69) Yu, D.-J.; Hu, J.; Tang, Z.-M.; Shen, H.-B.; Yang, J.; Yang, J.-Y. Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing* **2013**, *104*, 180−190.

(70) Zhou, Y.-P.; Jiang, J.-H.; Wu, H.-L.; Shen, G.-L.; Yu, R.-Q.; Ozaki, Y. Dry film method with ytterbium as the internal standard for near infrared spectroscopic plasma glucose assay coupled with boosting support vector regression. *J. Chemom.* **2006**, *20* (1−2), 13−21.

(71) Zhou, Y.-P.; Jiang, J.-H.; Lin, W.-Q.; Zou, H.-Y.; Wu, H.-L.; Shen, G.-L.; Yu, R.-Q. Boosting support vector regression in QSAR studies of bioactivities of chemical compounds. *Eur. J. Pharm. Sci.* **2006**, *28* (4), 344−353.

(72) Silva, C.; Ribeiro, B.; Sung, A. H., Boosting, R. V. M. Classifiers for large data sets. In *Adaptive and Natural Computing Algorithms*: 8th International Conference, ICANNGA 2007, Warsaw, Poland, April 11−14, 2007, Proceedings, Part II; Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B., Eds.; Springer: Berlin, Heidelberg, 2007; pp 228−237.

(73) Tashk, A. R. B.; Sayadiyan, A.; Valiollahzadeh, S. In *Face detection using Adaboosted RVM-based component classifier*, Proceedings of the 5th International Symposium on Image and Signal Processing and Analysis (ISPA 2007), Istanbul, Turkey, Sept. 27−29, 2007; IEEE: Istanbul, Turkey, pp 351−355.

(74) Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. In *OECD Series on Testing and Assessment*; OECD Publishing: Paris, 2014; pp 1−154.

(75) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361−1375.

(76) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29* (6−7), 476−488.

(77) Roy, K.; Mitra, I. On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. *Comb. Chem. High Throughput Screening* **2011**, *14* (6), 450−474.

(78) Chang, C.-Y.; Hsu, M.-T.; Esposito, E. X.; Tseng, Y. J. Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods. *J. Chem. Inf. Model.* **2013**, *53* (4), 958−971.

(79) Das, R. N.; Roy, K. Predictive modeling studies for the ecotoxicity of ionic liquids towards the green algae Scenedesmus vacuolatus. *Chemosphere* **2014**, *104*, 170−176.

(80) Li, D.; Chen, L.; Li, Y.; Tian, S.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery. 13. Development ofin SilicoPrediction Models for P-Glycoprotein Substrates. *Mol. Pharmaceutics* **2014**, *11* (3), 716−726.

(81) Shi, H.; Tian, S.; Li, Y.; Li, D.; Yu, H.; Zhen, X.; Hou, T. Absorption, Distribution, Metabolism, Excretion, and Toxicity Evaluation in Drug Discovery. 14. Prediction of Human Pregnane X Receptor Activators by Using Naive Bayesian Classification Technique. *Chem. Res. Toxicol.* **2015**, *28* (1), 116−125.

(82) Lei, T.; Chen, F.; Liu, H.; Sun, H.; Kang, Y.; Li, D.; Li, Y.; Hou, T. ADMET evaluation in drug discovery. Part 17: Development of quantitative and qualitative prediction models for chemical-induced respiratory toxicity. *Mol. Pharmaceutics* **2017**, *14* (7), 2407−2421.

(83) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1315−1326.

(84) Kaneko, H.; Funatsu, K. Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses. *J. Chem. Inf. Model.* **2014**, *54* (9), 2469−2482.

(85) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* **2005**, *33* (5), 445−459.

(86) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26* (5), 694−701.

(87) Ferrari, T.; Gini, G.; Bakhtyari, N. G.; Benfenati, E. In *Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships*, 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, April 11−15, 2011; IEEE: Paris, France, 2011; pp 120−127.

(88) Ferrari, T.; Cattaneo, D.; Gini, G.; Bakhtyari, N. G.; Manganaro, A.; Benfenati, E. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. *SAR QSAR Environ. Res.* **2013**, *24* (5), 365−383.

(89) Lombardo, A.; Pizzo, F.; Benfenati, E.; Manganaro, A.; Ferrari, T.; Gini, G. A new in silico classification model for ready biodegradability, based on molecular fragments. *Chemosphere* **2014**, *108*, 10−16.

(90) Rush, G. F.; Smith, J. H.; Newton, J. F.; Hook, J. B.; Berndt, W. O. Chemically induced nephrotoxicity: Role of metabolic activation. *Crit. Rev. Toxicol.* **1984**, *13* (2), 99−160.

(91) Lau, S. S. Quinone-thioether-mediated nephrotoxicity. *Drug Metab. Rev.* **1995**, *27* (1−2), 125−141.

(92) Kovacic, P.; Sacman, A.; Wu-Weis, M. Nephrotoxins: Widespread role of oxidative stress and electron transfer. *Curr. Med. Chem.* **2002**, *9* (8), 823−847.

(93) Emma, F.; Montini, G.; Parikh, S. M.; Salviati, L. Mitochondrial dysfunction in inherited renal disease and acute kidney injury. *Nat. Rev. Nephrol.* **2016**, *12* (5), 267−280.

(94) Honda, N.; Hishida, A. Pathophysiology of experimental nonoliguric acute renal failure. *Kidney Int.* **1993**, *43* (3), 513−521.

(95) Detrenis, S.; Meschi, M.; Musini, S.; Savazzi, G. Lights and shadows on the pathogenesis of contrast-induced nephropathy: state of the art. *Nephrol., Dial., Transplant.* **2005**, *20* (8), 1542−1550.

(96) Wang, B. X.; Japkowicz, N. Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems* **2010**, *25* (1), 1−20.

(97) Duchowicz, P. R.; Ocsachoque, M. A. Quantitative structure-toxicity models for heterogeneous aliphatic compounds. *QSAR Comb. Sci.* **2009**, *28* (3), 281−295.

(98) Simenhoff, M. L. Metabolism and toxicity of aliphatic amines. *Kidney Int. Suppl.* **1975**, No. 3, 314−7.

(99) Levine, S.; Sowinski, R. T-Lymphocyte depletion and lesions of choroid plexus and kidney induced by tertiary amines in rats. *Toxicol. Appl. Pharmacol.* **1977**, *40* (1), 147−159.

(100) National Toxicology Program *Report on carcinogens monograph on cumene*; 2331−267X; National Toxicology Program: Research Triangle Park, North Carolina, USA, September 25, 2013; pp 1−166.

(101) Bove, F. C.; Haarstad, V. B. Pharmacological activities of acetal derivatives of hemicholinium No. 3. *Eur. J. Pharmacol.* **1979**, *57* (2−3), 149−163.

(102) Trimarchi, G. R.; Germanò, A.; Campo, G. M.; De Luca, R.; Caputi, A. P. Changes in urine volume and urinary electrolyte excretion after intracerebroventricular injection of arecoline and hemicholinium-3. *Life Sci.* **1991**, *48* (21), 2097−2107.

(103) Carvalho, M.; Hawksworth, G.; Milhazes, N.; Borges, F.; Monks, T. J.; Fernandes, E.; Carvalho, F.; Bastos, M. Role of metabolites in MDMA (ecstasy)-induced nephrotoxicity: an in vitro study using rat and human renal proximal tubular cells. *Arch. Toxicol.* **2002**, *76* (10), 581−588.

(104) Monks, T. J.; Jones, D. C. The metabolism and toxicity of quinones, quinonimines, quinone methides, and quinone-thioethers. *Curr. Drug Metab.* **2002**, *3* (4), 425−438.

(105) Chan, C.-P.; Yuan-Soon, H.; Wang, Y.-J.; Lan, W.-H.; Chen, L.-I.; Chen, Y.-J.; Lin, B.-R.; Chang, M.-C.; Jeng, J.-H. Inhibition of cyclooxygenase activity, platelet aggregation and thromboxane B2 production by two environmental toxicants: m- and o-cresol. *Toxicology* **2005**, *208* (1), 95−104.