Toxicology Letters 340 (2021) 4–14

Contents lists available at ScienceDirect

Toxicology Letters

journal homepage: www.elsevier.com/locate/toxlet

Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints

Huawei Feng[a,1], Li Zhang[a,b,c,1], Shimeng Li[a], Lili Liu[a], Tianzhou Yang[a], Pengyu Yang[d], Jian Zhao[a], Isaiah Tuvia Arkin[e], Hongsheng Liu[b,c,f,*]

[a] School of Life Science, Liaoning University, Shenyang, 110036, China
[b] Technology Innovation Center for Computer Simulating and Information Processing of Bio-macromolecules of Shenyang, Shenyang, 110036, China
[c] Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning, Liaoning University, Shenyang, 110036, China
[d] School of Information, Liaoning University, Shenyang, 110036, China
[e] Department of Biological Chemistry, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat-Ram, Jerusalem, 91904, Israel
[f] School of Pharmaceutical Science, Liaoning University, Shenyang, 110036, China

HIGHLIGHTS

- Ensemble learning models were built for predicting chemical reproductive toxicity.
- The best model achieved an average accuracy of 86.33 % with 5-fold cross-validation.
- In 5-fold cross-validation, an area under the curve (AUC) of the best model was 0.937.
- Some substructures related to the reproductive toxicity were achieved.

ARTICLE INFO

Article history:
Received 11 July 2020
Received in revised form 29 October 2020
Accepted 3 January 2021
Available online 6 January 2021

Keywords:
Reproductive toxicity
Molecular fingerprint
Machine learning
Ensemble
Prediction models

ABSTRACT

Reproductive toxicity endpoints are a significant safety concern in the assessment of the adverse effects of chemicals in drug discovery. Computational models that can accurately predict a chemical's toxic potential are increasingly pursued to replace traditional animal experiments. Thus, ensemble learning models were built to predict the reproductive toxicity of compounds. Our ensemble models were developed using support vector machine, random forest, and extreme gradient boosting methods and 9 molecular fingerprints calculated for a dataset containing 1823 chemicals. The best prediction performance was achieved by the Ensemble-Top12 model, with an accuracy (ACC) of 86.33 %, a sensitivity (SEN) of 82.02 %, a specificity (SPE) of 90.19 %, and an area under the receiver operating characteristic curve (AUC) of 0.937 in 5-fold cross-validation and ACC, SEN, SPE, and AUC values of 84.38 %, 86.90 %, 90.67 %, and 0.920, respectively, in external validation. We also defined the applicability domain (AD) of the ensemble model by calculating the Tanimoto distance of the training set. Compared with models in existing literature, our ensemble model achieves relatively high ACC, SPE and AUC values. We also identified several fingerprint features related to chemical reproductive toxicity. Considering the performance of model, we recommend using the Ensemble-Top12 model to predict reproductive toxicity in early drug development.

1. Introduction

Drug-induced toxicity is a major reason for the failure of drug research and development (Andy et al., 2020; Munos, 2009). In particular, among the 14 types of toxicity associated with drug withdrawal, reproductive toxicity (herein more generally referred to as reprotoxicity) accounts for 3% of instances of drug withdrawal/discontinuation (Siramshetty et al., 2015) and more than 10 % of preclinical toxicology-related attrition (Guengerich

* Corresponding author at: School of Life Science, Liaoning University, Address: No. 66, Chongshan Zhonglu, Shenyang, Liaoning, 110036, China.
E-mail address: liuhongsheng@lnu.edu.cn (H. Liu).
[1] These co-first authors contributed equally to this work.

http://dx.doi.org/10.1016/j.toxlet.2021.01.002
0378-4274/

and Macdonald, 2007). Therefore, early assessment of the toxic properties of chemical structures is important in the field of drug development (Brannen et al., 2016). Reprotoxicity is not a singular endpoint but rather a range of endpoints mainly including reproductive toxicity and developmental toxicity (Marzo and Benfenati, 2018). Reprotoxicity is related to impairment of male and female reproductive capacities and the induction of non-genetically harmful effects on offspring (Lo Piparo and Worth, 2010). For reprotoxicity assessments, traditional experimental testing for chemical toxicity profiles is very expensive both financially and in terms of animal usage and requires a long time (Marzo and Benfenati, 2018). It has been estimated that toxicity tests in animal models account for 54 % of the testing costs related to REACH compliance (Scialli, 2008). Moreover, although far more animals are used in reprotoxicity testing than in other types of testing, the results of many of these reprotoxicity tests may not be highly correlated with effects observed in humans (Höfer et al., 2004). Another factor that increases the difficulty of toxicity endpoint assessments is that experimental data are not always easy to explain; usually, whether a compound is toxic to reproduction directly or because it causes general systemic toxicity that also affects reproductivity is not always clear. For all these reasons, the results of animal experiments cannot easily reflect the human body's response to new drugs or provide any risk exemption.

To develop effective and accurate alternatives (non-testing methods), conducting high-throughput computer toxicity predictions is an attractive strategy. One dominant and highly developed toxicity-prediction approach is Quantitative Structure-Activity Relationships (QSARs) based on chemical structural properties (Lu et al., 2018; Chen-Lo et al., 2018; Satpathy, 2019). This method aims to mathematically describe the contributions of one or more physicochemical properties to bioactivity (e.g., reprotoxicity) (Benfenati et al., 2019; Cherkasov et al., 2014). Additionally, QSARs that yield true quantitative predictions are useful as these quantitative outputs are more useful in risk assessments. To date, many QSAR models have been developed using machine learning methods to predict chemical reprotoxicity (Basant et al., 2016; Jiang et al., 2019; Ghorbanzadeh et al., 2016; Gunturi and Ramamurthi, 2014; Marzo and Benfenati, 2018). However, most of these models are local models based on only one chemical class, several chemical classes with only one or a few toxicity endpoints, or small datasets (fewer than 500 compounds). Thus, these models may have poor generalization ability in large external datasets. One QSAR global model has been established (Jiang et al., 2019), but its prediction performance requires further improvement to reduce the number of candidate chemicals to be tested with in vivo experiments. Notably, some models were produced through a simple fine-tuning process and have not yet been assessed through appropriate cross-validation (Zhang et al., 2017a; Marzo and Benfenati, 2018).

Compared to molecular descriptors, molecular fingerprints with binary features directly link to chemical features and are more interpretable because each bit corresponds to a specific substructure (Yang et al., 2018). Generally, these molecular fingerprints are typed based on transformation to a bit-string (Kwon et al., 2019). The main types are substructure key-based fingerprints, circular fingerprints, and topological or path-based fingerprints (Cereto-Massagué et al., 2015). Structure key-based fingerprints (e.g., MACCS and PubChem) encode molecular structures based on the presence of substructures or features and have higher performance than hash fingerprints in toxicity prediction (Webb et al., 2014), although the choice of descriptors is not a major factor in model performance. ECFP is a commonly used circular fingerprint that encoded molecular structures based on hashing fragments up to a specific radius (Kwon et al., 2019). Notably, however, an ECFP fingerprint is usually difficult to unravel. In general, current research shows that combining different types of fingerprints (especially structure-based methods) with different machine learning methods is an optimal strategy that can improve the diversity of the model and, thus, its prediction performance (Cereto-Massagué et al., 2015; Zhang et al., 2017b). In terms of model development, ensemble modelling can sufficiently fuse model predictions together. Ensemble learning usually produces higher accuracy than individual models because it can manage the strengths and weaknesses of each base learner (Ai et al., 2018; Mora et al., 2020; Kwon et al., 2019; Ballabio et al., 2019). More recently, ensemble learning has been applied to predict chemical toxicity and may provide encouraging results. In this respect, Furxhi et al. built voting ensemble models for predicting nanoparticle toxicity *in vitro* by using the Copeland Index to select the optimal classifiers. Liu et al. (Liu et al., 2020) used three machine learning methods to combine molecular fingerprints and molecular descriptors to build a predictive cardiotoxicity ensemble model with an accuracy of 84.9 % and an AUC of 0.887. Yin et al. (Yin et al., 2019) used MACCS fingerprints to construct ensemble models for predicting cytotoxicity, the best of which achieved an AUC of 0.852. One should be aware, however, that because machine learning methods are difficult to explain, sometimes combining them renders such methods increasingly challenging.

In this work, we developed in silico predictive models to predict reprotoxicity with machine learning methods using 1823 organic chemicals. To this end, multiple basic classifiers were first constructed using 9 types of molecular fingerprints and 3 machine learning algorithms. Then, several ensemble models were built by combining multiple subsets of the basic classifiers based on different fingerprint types and machine learning algorithms. An applicability domain (AD) is also defined to analyse the accuracy and robustness of the ensemble model. Finally, structural features associated with reprotoxicity were identified. The predictive capability of the models developed in this study was tested using both 5-fold cross-validation and an external test set.

2. Materials and methods

2.1. Toxicity datasets

The reprotoxicity dataset used in this study was acquired from a previous publication (Jiang et al., 2019) and contains 2487 compounds drugs collected from the ECHA-C & LInventory and OECD-eChemPortal, public databases. In the dataset, chemicals are labelled as positive (reprotoxicants) or negative (non-reprotoxicants) as described by Jiang et al. (Jiang et al., 2019). First, the chemical structure of each compound was carefully examined. Subsequently, to ensure the uniqueness and high quality of the data, all simplified molecular input line entry specifications (SMILES) were formatted to the canonical SMILES format, and some inorganic compounds and complex compounds were removed from the analysis (Jiang et al., 2019). Finally, 1823 selected compounds were randomly divided into a training set (cross-validation set) and an external validation set with a ratio of 8:2. The lists of chemicals in both the training set and the external set used for model building are shown in Table S1 and Table S2.

2.2. Calculation of molecular fingerprints

Nine types of molecular fingerprints for QSAR modelling were calculated for the compounds in the datasets using PaDEL-Descriptor (Yap, 2011): the Estate fingerprint, the MACCS fingerprint, the PubChem fingerprint, the Substructure (FP4) fingerprint, the Substructure Count (FP4C) fingerprint, the Klekota-Roth (KR) fingerprint, the Klekota-Roth Count (KRC) fingerprint, the 2D Atom

Pairs (AP2D) fingerprint and the 2D Atom Pairs Count (AP2DC) fingerprint (Table S3). Detailed descriptions of these fingerprints can be found in the original studies (Yap, 2011).

2.3. Feature selection

Feature selection provides an effective way to process high-dimensional data while removing irrelevant and redundant data to reduce the computation time, improve the learning accuracy, and gain a better understanding of machine learning models or data (Jie et al., 2018). In the current study, two methods were used for feature selection: low-variation feature filtering and high-correlation feature filtering. The nearZeroVar function in the R package caret (version 6.0–84) (Kuhn, 2008) was used to filter out identical or almost identical features from all samples. Since a molecular fingerprint consists of binary variables that take only values of 0 and 1, the Tanimoto coefficient (Tc) is more suitable than the Pearson correlation coefficient for computing the similarity between features for high-correlation feature filtering (Bajusz et al., 2015; Chen-Lo et al., 2018). Therefore, in this study, the Tc was used for high-correlation feature filtering. Various threshold values (0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, and 1.0 (no deletion)) were tested to identify the optimal Tc threshold based on the performance of a random forest model in 5-fold cross-validation. The Tc between two features (bits) A and B can be computed as follows:

$$T_c = \frac{c}{a + b - c} \tag{1}$$

where a is the number of features in fingerprint A, b is the number of features in B, and c is the number of features common to a and b.

2.4. Model construction

Ensemble models formed by fusing the results of various basic classifiers usually produce more accurate results than any individual model. In this study, three machine learning algorithms, namely, the support vector machine (SVM) algorithm, the random forest (RF) algorithm and the extreme gradient boosting (XGBoost) algorithm, were used for model building. These algorithms were implemented in R software (version 3.3.1). Machine learning packages such as the kernlab (version 0.9–25) package (Karatzoglou et al., 2004), the randomForest (version 4.6–12) package (Liaw and Wiener, 2002), and the xgboost (version 0.4–4) package (Chen and Guestrin, 2016) were used for model implementation. The way in which the ensemble models were built has been described fully in our previously published studies (Zhang et al., 2017b; Liu et al., 2020). Nine basic classifiers were constructed for each machine learning algorithm by applying the nine molecular fingerprints to the corresponding machine learning algorithms. Here, a total of 27 basic classifiers were generated, which were used to further develop the ensemble model. We then developed 27 new ensemble models by combining a subset of 27 basic classifiers via averaging their predictive probabilities. To do this, we first ranked all the ensemble models based on their AUC scores. Among them, the Ensemble-Top1 model was the best basic classifier with the highest AUC, and the Ensemble-Top2 model was built by merged the first two basic classifiers. The Ensemble-Top3 to Ensemble-Top27 models were also created by combining the first 3–27 basic classifiers. Last, cross-validation was performed to evaluate the predictive performance of all ensemble models. The best one with the highest AUC score was selected as the most appropriate ensemble model for future toxicity prediction.

2.4.1. Support vector machine

SVM is a popular intelligent learning algorithm based on the principle of structural risk minimization (SRM) and is generally applied to classification and regression problems. This method is well known for its high-performance prediction capability and low risk of overfitting (Cover and Hart, 1967). An SVM classifier operates by finding an optimal hyperplane (linear separator) in a multidimensional space to separate a set of points into two classes: positive and negative. In the current work, we adopted the radial basis function (RBF) kernel when constructing the SVM classifiers to map the input data to an infinite-dimensional space. We also applied the random search method (Bergstra and Bengio, 2012) using the caret package to optimize two specific SVM parameters: the regularization parameter C and the γ parameter of the RBF kernel.

2.4.2. Random forest

RF is an effective ensemble machine learning algorithm that operates by constructing many independent decision trees using randomly selected subsets of the training samples and features and then collecting the results of these decision trees. To devise a prediction for a new input, that input is run through every decision tree, and the results are treated as votes on how the input should be classified. The performance of an RF classifier is considered to be better than that of a single decision tree (DT) classifier, as it reduces the risk of overfitting when dealing with a large number of features and provides higher accuracy at the levels of individual trees and paths (Gandhi et al., 2018). In the current study, we used the randomForest function to build RF classifiers, where the two main parameters, i.e., the number of classification trees in the forest and the number of variables randomly selected for each node split, were set to default values of ntree = 500 and mtry = the square root of the number of features in the dataset, respectively. The relative importance of each type of fingerprint feature was also calculated using the importance function (randomForest package).

2.4.3. Extreme gradient boosting

XGBoost is another tree-based ensemble learning method. It uses subtle penalization of the individual trees, thus allowing the trees to have different numbers of terminal nodes (Zaslavskiy et al., 2019). XGBoost possesses the merits of ease of use, ease of parallelization and high predictive accuracy; consequently, it has achieved superior outcomes compared with many other algorithms in several machine learning competitions (Sheridan et al., 2016). In this study, the caret package was used to optimize the following four main parameters: the step size shrinkage (eta), the maximum tree depth (max. depth), the minimum sum of the instance weights (min. child. weight), and the maximum number of iterations (nrounds).

2.5. Model evaluation

The performance of the constructed models was evaluated using 5-fold cross validation and external validation. To reduce the influence of the randomness of prediction and achieve a robust performance evaluation, 5-fold cross-validation was repeated 100 times, resulting in the generation of 500 sets of performance indicators. The accuracy (ACC), sensitivity (SEN), specificity (SPE), and AUC were calculated by the following formulas:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \times 100 \tag{2}$$

$$SEN = \frac{TP}{TP + FN} \times 100 \tag{3}$$

$$SPE = \frac{TN}{TN + FP} \times 100 \tag{4}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The receiver operating characteristic (ROC) curve is a graphical plot of the TP rate (SEN) against the FP rate (SPE) for the different possible cut-off points of a diagnostic test. Compared to the SEN and SPE, which reflect a model's performance at a single cut-off, ROC analysis can better support a multi-threshold comparison of different algorithms and provide more global and unbiased evaluation results (He et al., 2019). The AUC ranges in value from 0 to 1, where '1' indicates perfect discrimination and '0.5' represents a random model (Catherine et al., 2013).

2.6. Analysis of the applicability domain

Because the training set of the classification models cannot cover the entire chemical space, the predictive power and explanatory power of model for unknown chemicals may be limited (Jiang et al., 2020). Therefore, the application domain (AD) of the model must be predefined. For binary fingerprints, the Tanimoto distance is a suitable method for calculating the minimum distance of compounds between training and test sets (Roy et al., 2015; González-Medina et al., 2017). The Tanimoto distance is a distance measure transformed from Tc with values ranging from 0 to 1. The Tanimoto distance is exactly 0 when two compounds are identical; otherwise, the distance is 1.

$$Tanimoto\ \ distance = 1 - Tc \tag{5}$$

Thus, we applied the Tanimoto distance to identify outliers and compounds residing outside the AD. We evaluated the effect of different Tanimoto distance values (0.1, 0.2, 0.3, 0.4, and 0.5) on model performance to identify a suitable distance threshold.

3. Results and discussion

3.1. Feature selection

In this study, 1823 compounds were collected from the reproductive toxicity dataset developed by Jiang et al. for use as the training set (1458) and the external validation set (365) for the establishment of reprotoxicity prediction models. We then calculated 9 types of fingerprints for the compounds in the dataset using the PaDEL-Descriptor software. As shown in Table S2, both low-variation feature filtering and high-correlation feature filtering can effectively eliminate a large number of redundant features. To select the best threshold for the Tc, we used an RF prediction model to evaluate the effects of several different

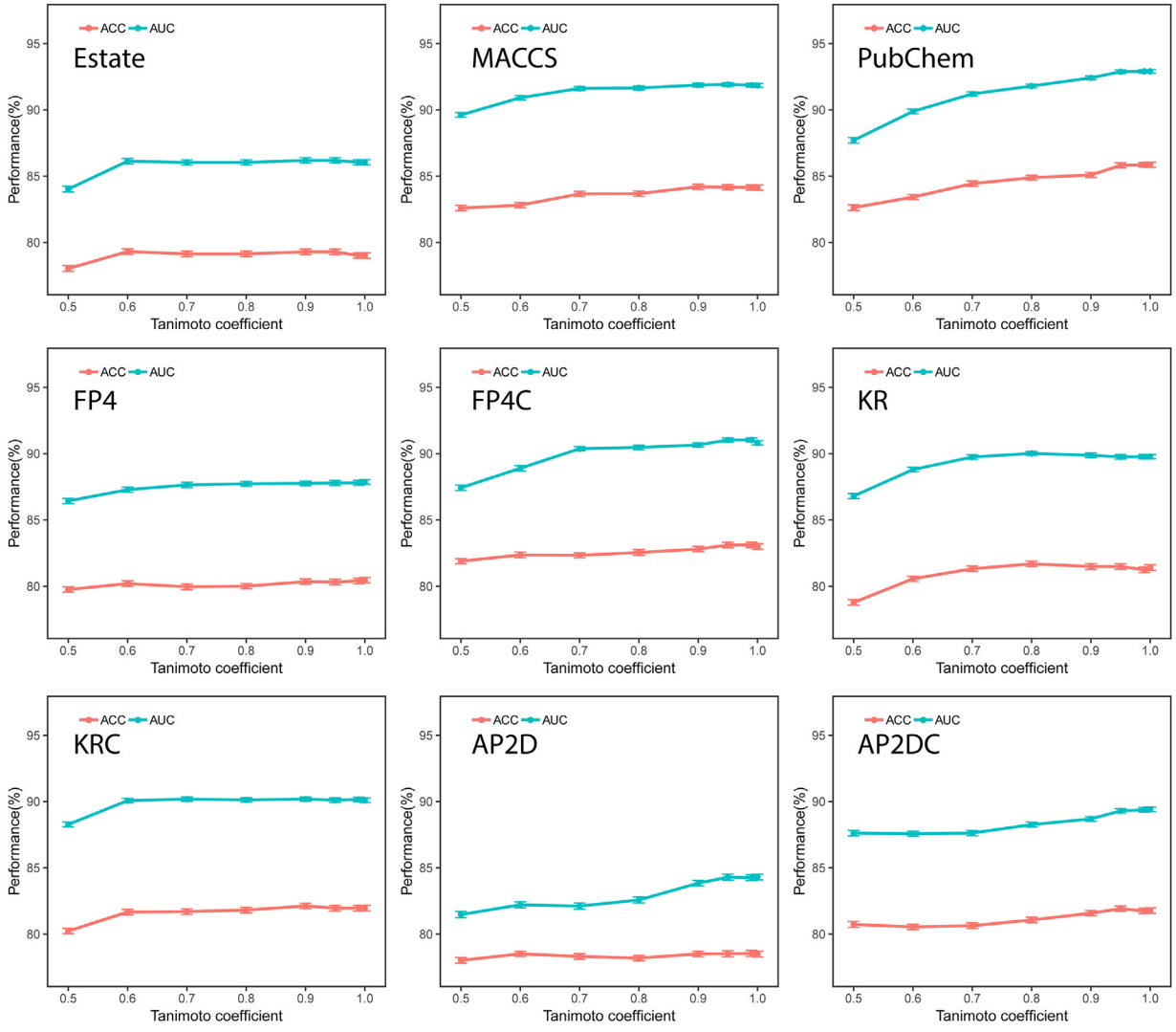

Fig. 1. Relationships between the prediction performance of an RF model and the Tc threshold for the nine different fingerprint types. The error bars represent the standard errors of the performance indexes obtained via 5-fold cross-validation with 100 repetitions.

thresholds on the high-correlation feature filtering of the molecular fingerprints. The evaluation results indicated that a threshold of 0.95 resulted in the best model prediction performance (Fig. 1).

3.2. Prediction performance of the models

After performing feature selection, we used the SVM, RF and XGBoost algorithms to generate 27 basic classifiers based on the resulting fingerprints. These machine learning models were assessed via 5-fold cross-validation, and their performances are presented in Table S4. The ACC values of these basic classifiers ranged from 77.1%–85.9%, the SEN values ranged from 69.2%–83.3%, the SPE values ranged from 80.4%–90.0%, and the AUC values ranged from 0.831 to 0.929. Among these models, the SVM model based on PubChem fingerprints yielded the most accurate results, and the RF model based on PubChem fingerprints achieved the highest AUC value. Overall, the basic classifiers showed good prediction performance, indicating that the machine learning algorithms and molecular fingerprints used in this study are effective for predicting the reproductive toxicity of chemicals.

To obtain a higher-performance prediction model for the reproductive toxicity of compounds, we established 27 new ensemble models by combining subsets of the basic models. The validation results are listed in Table 1 and Fig. 2. Consistent with expectations, the ACC and AUC values of the best ensemble learning models were much higher than those of any of the single models. Increased AUC values were found through 5-fold cross-validation as the number of basic classifiers included in the ensemble learning models increased. For example, the performance of Ensemble-Top3, generated by combining the three best-performing basic classifiers in terms of the AUC, was significantly higher than that of Ensemble-Top1 (equivalent to the basic classifier with the highest AUC value), with an ACC of 85.76 % ± 0.08 % and an AUC of 0.931 ± 0.001. Moreover, almost all ensemble learning models showed slightly higher SEN and SPE values than the basic models. Among all ensemble models, the performance of the Ensemble-Top12 model was the most satisfactory; this model was constructed from 12 basic classifiers, namely, PubChem + RF, MACCS + RF, FP4C + RF, KRC + RF, PubChem + SVM, MACCS + SVM, PubChem + XGBoost, MACCS + XGBoost, FP4C + XGBoost, APC2D + XGBoost, KRC + XGBoost, and KR + XGBoost. The ACC, SEN, SPE, and AUC values of this model were 86.33 %±0.08 %, 82.02 %±0.15 %, 90.19 %±0.11 %, and 0.937 ± 0.001, respectively. We also found that the AUC of the Ensemble-Top12 model was increased by 0.8 % (t-test, $p < 0.0001$) compared to that of the PubChem + RF classifier, which was also the basic model with the highest AUC value, and the SEN of the Ensemble-Top12 model was increased by 1.28 % (t-test, $p < 0.0001$) compared to that of the PubChem + SVM classifier, which was also the most sensitive basic model. Although no significant differences were observed between the two models, we observed that the ACC of the Ensemble-Top12 model was increased by 0.43 % (t-test, $p > 0.05$) compared to that of the most accurate basic model (PubChem + SVM) and that the SPE of the Ensemble-Top12 model was increased by 0.19 % (t-test, $p > 0.05$) compared to that of the most specific basic model (PubChem + RF). These results indicate that our ensemble model was able to effectively achieve improved performance in reprotoxicity prediction. This may be because the ensemble models combined diverse independent models, thus allowing them to achieve better prediction performance and more stable prediction results than the basic classifiers.

An external validation set containing 365 compounds was used to further validate the performance of the 27 ensemble models, and the prediction performance results are shown in Table S5. In the external validation, the ensemble models achieved ACC values of 84.11 %–84.66 %, SEN values of 73.26 %–77.91 %, SPE values of 87.05 %–92.23 %, and AUC values of 0.907−0.920. Compared with the performance of the 27 basic classifiers (Table S6), most ensemble models achieved higher ACC and AUC values than almost all individual classifiers. For example, for the basic model established using XGBoost and MACCS molecular fingerprinting, the ACC and AUC values in the external validation were 82.7 % and 0.906, respectively. Fig. 2 shows that the statistical performance

Table 1
Predictive performance of the 27 ensemble models with the training set, as assessed via 5-fold cross-validation with 100 repetitions. The statistical values of the performance indicators are expressed as the mean ± standard error of the mean (SEM).

Model Name	ACC (%)	SEN (%)	SPE (%)	AUC
Ensemble-Top1	85.78 ± 0.08	81.13 ± 0.15	89.96 ± 0.12	0.929 ± 0.001
Ensemble-Top2	85.09 ± 0.08	81.05 ± 0.14	88.72 ± 0.12	0.930 ± 0.001
Ensemble-Top3	85.76 ± 0.08	82.08 ± 0.14	89.06 ± 0.12	0.931 ± 0.001
Ensemble-Top4	86.05 ± 0.08	82.49 ± 0.14	89.25 ± 0.12	0.934 ± 0.001
Ensemble-Top5	86.09 ± 0.08	82.60 ± 0.14	89.23 ± 0.12	0.934 ± 0.001
Ensemble-Top6	86.05 ± 0.08	82.73 ± 0.14	89.04 ± 0.11	0.933 ± 0.001
Ensemble-Top7	86.02 ± 0.08	82.36 ± 0.14	89.3 ± 0.12	0.935 ± 0.001
Ensemble-Top8	86.01 ± 0.08	82.19 ± 0.14	89.44 ± 0.12	0.935 ± 0.001
Ensemble-Top9	86.08 ± 0.08	81.99 ± 0.14	89.76 ± 0.11	0.936 ± 0.001
Ensemble-Top10	86.26 ± 0.08	82.04 ± 0.15	90.05 ± 0.11	0.937 ± 0.001
Ensemble-Top11	86.33 ± 0.08	82.01 ± 0.15	90.22 ± 0.11	0.937 ± 0.001
Ensemble-Top12	86.33 ± 0.08	82.02 ± 0.15	90.19 ± 0.11	0.937 ± 0.001
Ensemble-Top13	86.35 ± 0.08	81.92 ± 0.15	90.34 ± 0.11	0.937 ± 0.001
Ensemble-Top14	86.25 ± 0.08	81.84 ± 0.15	90.21 ± 0.12	0.936 ± 0.001
Ensemble-Top15	86.27 ± 0.08	81.71 ± 0.15	90.37 ± 0.12	0.936 ± 0.001
Ensemble-Top16	86.4 ± 0.08	81.73 ± 0.15	90.59 ± 0.11	0.936 ± 0.001
Ensemble-Top17	86.27 ± 0.08	81.48 ± 0.15	90.58 ± 0.11	0.936 ± 0.001
Ensemble-Top18	86.28 ± 0.08	81.44 ± 0.15	90.62 ± 0.11	0.936 ± 0.001
Ensemble-Top19	86.23 ± 0.08	81.11 ± 0.15	90.82 ± 0.11	0.935 ± 0.001
Ensemble-Top20	86.18 ± 0.08	80.77 ± 0.15	91.04 ± 0.11	0.935 ± 0.001
Ensemble-Top21	86.03 ± 0.08	80.51 ± 0.15	90.99 ± 0.11	0.934 ± 0.001
Ensemble-Top22	85.88 ± 0.08	80.39 ± 0.15	90.81 ± 0.11	0.934 ± 0.001
Ensemble-Top23	85.71 ± 0.08	80.15 ± 0.15	90.70 ± 0.11	0.933 ± 0.001
Ensemble-Top24	85.54 ± 0.08	79.85 ± 0.15	90.64 ± 0.11	0.932 ± 0.001
Ensemble-Top25	85.50 ± 0.08	79.69 ± 0.16	90.72 ± 0.11	0.932 ± 0.001
Ensemble-Top26	85.38 ± 0.08	79.25 ± 0.16	90.90 ± 0.11	0.931 ± 0.001
Ensemble-Top27	85.29 ± 0.08	79.02 ± 0.16	90.92 ± 0.11	0.931 ± 0.001

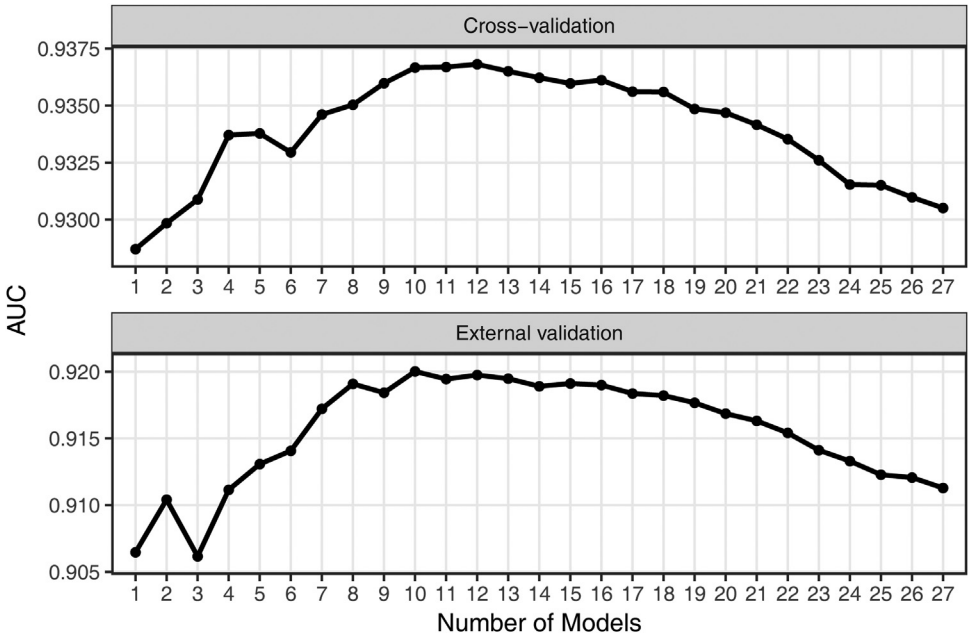

Fig. 2. Relationships between the AUC value of an ensemble model and the number of basic classifiers included in the model, as assessed via cross-validation and external validation.

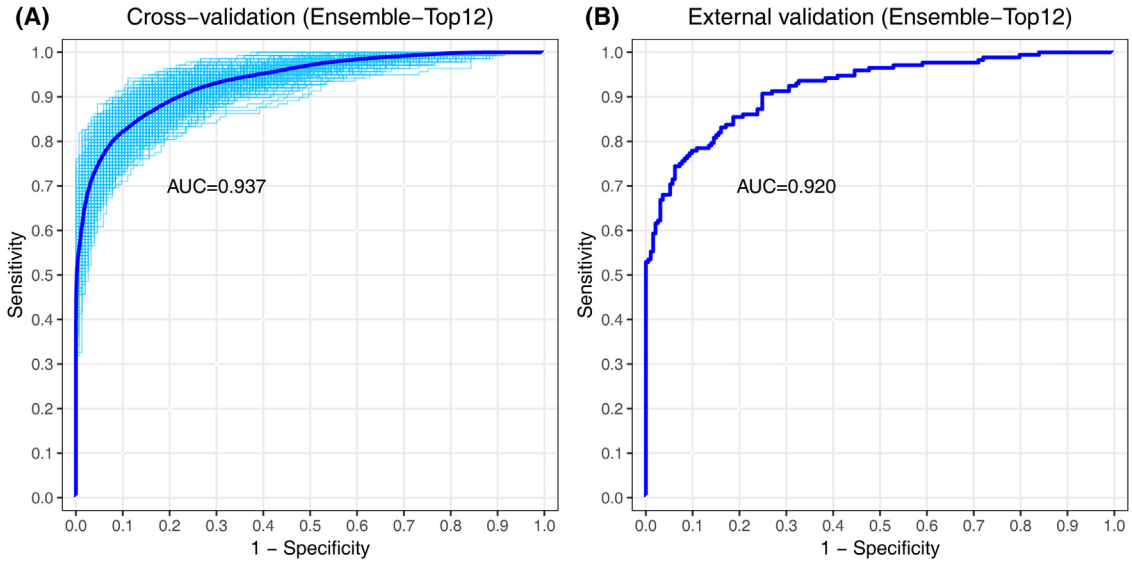

Fig. 3. ROC curves of the Ensemble-Top12 model obtained from cross-validation and external validation. A, ROC curves with the training set. The thin, light-blue lines represent the individual ROC curves of Ensemble-Top12 for each of the 100 repetitions of cross-validation, and the thick blue line represents the average ROC curve. B, ROC curve for the external validation.

Table 2
Prediction results for chemicals outside the domain (OD) and inside the domain (ID) for the Ensemble-Top12 model with corresponding Tanimoto distance values.

Tanimoto distance	ID				OD			
	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
0.10	93.46%	89.47 %	98.00 %	0.991	80.62%	71.30 %	88.11 %	0.872
0.20	89.29%	82.42 %	95.24 %	0.953	78.70%	71.60 %	85.23 %	0.857
0.30	86.67 %	78.69 %	92.64 %	0.935	76.25%	74.00 %	80.00 %	0.815
0.40	85.09%	78.34 %	90.81 %	0.923	73.91%	66.67 %	87.50 %	0.850
0.50	84.76%	77.51 %	91.15 %	0.921	50.00 %	66.67 %	0.00 %	0.000
No AD analysis	84.38 %	77.33 %	90.67 %	0.920	–	–	–	–

increase for the ensemble model is even more pronounced in external validation. Meanwhile, we also observed that in both the cross-validation and the external validation, the AUC values of the ensemble learning models formed an inverted U-curve when plotted against the number of basic classifiers included in each ensemble model (Fig. 2). This finding indicates that the predictive power decreases when too many basic classifiers are combined, which may be due to the inclusion of many poorly performing basic classifiers. In addition, the ROC curves for the Ensemble-Top12 model on the training set and the external validation set are shown in Fig. 3. These graphs show large AUC values for both cross-validation and external validation, and the difference in the AUCs between the two types of validation is only 1.7 %, indicating that the level of overfitting in the ensemble model is quite low. Based on the above results, we can conclude that the Ensemble-Top12 model is a good classifier that can effectively and stably predict the reproductive toxicity of compounds.

3.3. The applicability domain of the best model

The applicability domain (AD) reflects the coverage of the prediction model. In this study, we used the Tanimoto distance method based on AD2D, EState, KR, MACCS, Pubchem, and FP4 fingerprint to analyse the boundaries of the AD of the training set. As shown in Figure S1, the number of compounds inside the AD (ID) increases with increasing Tanimoto distance as more test compounds were considered structurally similar to the training set. Then, we investigated the predictability of the best Ensemble-Top12 model by analysing whether the predefined AD can accurately identify query samples from the external validation set. As shown in Table 2, when the Tanimoto distance is set too small (e.g., 0.1, 0.2), the test compounds inside the AD (ID) have higher performance because the AD contains more highly similar compounds. Similarly, the compounds outside the AD (OD) also have higher performance because a distance that is too small will lead to excessive extrapolation of intermediately similar compounds. However, if the Tanimoto distance is set too large (e.g., 0.4, 0.5), too many dissimilar compounds are included in the compounds ID, resulting in a decrease in model performance. Notably, a large distance also leads to extrapolation of too few compounds, which cannot be used for model validation, resulting in lower performance. To balance the predictability of the model, we recommend using 0.3 as the AD threshold. Of course, the AD threshold can also be adjusted according to the needs of users. For example, a lower threshold can be selected when a more accurate model is needed. Under the recommended AD threshold (0.3), the count of test compounds ID was 78.08 % (285/365) of all the compounds in the external validation set; compounds ID had an AUC of 0.935, ACC of 86.67 %, SEN of 78.69 %, and SPE of 92.64 %, which were significantly higher than the values for the performance of compounds OD. Compared with the external validation set without AD analysis (Table S5), the performance of compounds ID with a Tanimoto distance of 0.3 achieved higher ACC, SEN, SPE, and AUC values. From the above results, the use of the AD significantly improved the prediction performance of the ensemble model.

3.4. Comparison with other models

Recently, a few computational models have been developed for the prediction of chemical reprotoxicity. The prediction performance estimated using the classical validation method (where the dataset consists of two separate parts, a training set and a test set) may be biased by the single split of the data. Thus, to ensure the accuracy of our comparative analysis, we present comparisons only with other (Q)SAR methods that have been subjected to appropriate cross-validation evaluations. Comparisons between our Ensemble-Top12 model and other models for reprotoxicity prediction reported in the literature are presented in Table 3. We can draw the following conclusions from Tables 1 and 3. (i) The ACC of the Ensemble-Top12 model is higher than those of the models previously reported in the literature. In fact, the ACC values of the NB model (Zhang et al., 2019) and the CAESAR AFP model (Cassano et al., 2010) are higher than those of Ensemble-Top12. However, these models consider a limited number of compounds, and a risk of overfitting exists. For example, significant overfitting exists in the NB model (Zhang et al., 2019). In addition, the ACC values of the other models are relatively low. (ii) The Ensemble-Top12 model achieves the highest AUC value among the models for which the AUC indicator was used for performance evaluation. (iii) Ensemble-Top12 achieves the highest predictive SPE of 90.19 %. (iv) Recently, Jiang et al. (Jiang et al., 2019) developed binary classification models for predicting reprotoxicity using 6 machine learning methods and 9 molecular fingerprints based on a larger data set with multiple reprotoxicity endpoints. The performance of the established best MACCSFP-SVM model was superior to that obtained in previous studies, which also indicates that it will omit fewer reproductively toxic chemicals because it can be used to predict the toxic chemicals causing sperm reduction, gonadal dysgenesis, abnormal ovulation, teratogenicity, infertility and

Table 3
Performance indexes and evaluation methods for multiple reprotoxicity classification models previously reported in the literature.

Model Name	Endpoint(s)	Data-set	Validation Method (s)	ACC (%)	SEN (%)	SPE (%)	AUC	Ref.
MultiCASE	Drosophila SLRL	377	LMO	81.60	73.90	88.10	–	(Jensen et al., 2008)
CAESAR RF	Developmental toxicity	292	10-fold CV	84.00	95.00	59.00	–	(Cassano et al., 2010)
CAESAR AFP	Developmental toxicity	292	LOOCV	88.00	90.00	82.00	–	(Cassano et al., 2010)
LDA	Reproductive LOAEL	206	5-fold CV	80.0	81.0	80.0	–	(Martin et al., 2011)
NB	Developmental toxicity	284	5-fold CV	91.11	93.00	90.00	–	(Zhang et al., 2019)
			Test set	83.9	87.2	76.5	–	
MACCSFP-SVM	Sperm reduction, gonadal dysgenesis, abnormal ovulation, teratogenicity and infertility growth retardation	1823	10-fold CV	84.90	82.30	87.20	0.915	(Jiang et al., 2019)
			Test set	83.60	78.50	88.10	0.900	
Ensemble-Top12	Sperm reduction, gonadal dysgenesis, abnormal ovulation, teratogenicity and infertility growth retardation	1823	5-fold CV	86.33	82.02	90.19	0.937	Present study
			Test set	84.38	77.33	90.67	0.920	

LMO, Leave-many-out; 10-fold CV, Ten-fold cross-validation; 5-fold CV, Five-fold cross-validation; LOOCV, Leave-one-out cross-validation; RF, Random forest; AFP, Adaptive fuzzy partition; LDA, Linear discriminant analysis; NB, Naïve Bayesian; SVM, Support vector machine; LOAEL, Lowest-observed adverse effect levels; ACC, Accuracy; SEN, Sensitivity; SPE, Specificity; AUC, Area under the receiver operating characteristic curve.

Table 4
Classification results for different classes of compounds by our EnsembleTop12 model.

Category	No. of test compounds	No. of misclassified compounds	ACC (%) of each category	No. of positive compounds	No. of misclassified positive compounds	No. of negative compounds	No. of misclassified negative compounds
Heterocyclic compounds	113	9	85.25	78	2	35	7
Alicyclic compounds	34	3	83.78	24	3	10	0
Aromatic ring compounds	86	14	72.00	37	8	49	6
Open-chain compounds	125	26	65.56	29	23	96	3
Inorganic compounds	7	5	16.67	4	3	3	2

delayed growth. However, it should be noted that the major limitation of this model was that it was unable to recognize some compounds, including non-toxic compounds that can cause reprotoxicity through metabolic activation *in vivo* and compounds eliminated in data pre-processing, such as inorganic compounds, metal organic compounds, mixtures and salts. It is encouraging to see that our Ensemble-Top12 model achieved better performance than the MACCSFP-SVM model. The ACC, SEN, SPE and AUC values obtained by the Ensemble-Top12 model were 1.43 %, 0.03 %, 2.99 % and 2.2 % higher, respectively, than those of the MACCSFP-SVM model. As stated above, the results of these comparisons show that Ensemble-Top12 is an outstanding classification model for predicting the reprotoxicity of compounds.

3.5. Analysis of classification results of compounds in the test set

The structure of the 365 compounds in the test set was further analysed to understand the predictive power of our ensemble model for different classes of compounds (Table S7). We found that 39 out of 172 positive compounds were falsely identified as negative compounds, and 18 out of 193 negative compounds were misclassified as positive compounds. Overall, our model shows excellent predictive ability for cyclic compounds, including heterocyclic compounds, carbocyclic compounds, and aromatic ring compounds, but it has relatively weak predictive ability for open-chain compounds and inorganic compounds (Table 4). Among the heterocyclic compounds, imidazole, pyridine, piperidine and pyrrole, all belonging to the nitrogen-containing heteroaromatic compounds, have been reported to be related to and are often used as structural alerts (SAs) of reproductive toxicity (Jiang et al., 2019; Chinaza et al., 2014). Consistent with this, our model also successfully recognized the imidazoles (e.g., imidazole, nocodazole, 2-heptadecyl-3-hydroxyethylimidazoline), pyrroles (e.g., 1-ethyl-2-pyrrolidinone), pyridines (e.g., thiacloprid), and piperidines (e.g., haloperidol) among the positive compounds (Fig. 4A). Compared with alicyclic compounds with simple rings (e.g., gabapentin, oxaliplatin, dicyclopentadiene), alicyclic compounds with multiple rings seemed to be more easily classified correctly (Fig. 4B). For example, the models successfully identified steroid compounds, such as hydroxyprogesterone acetate,

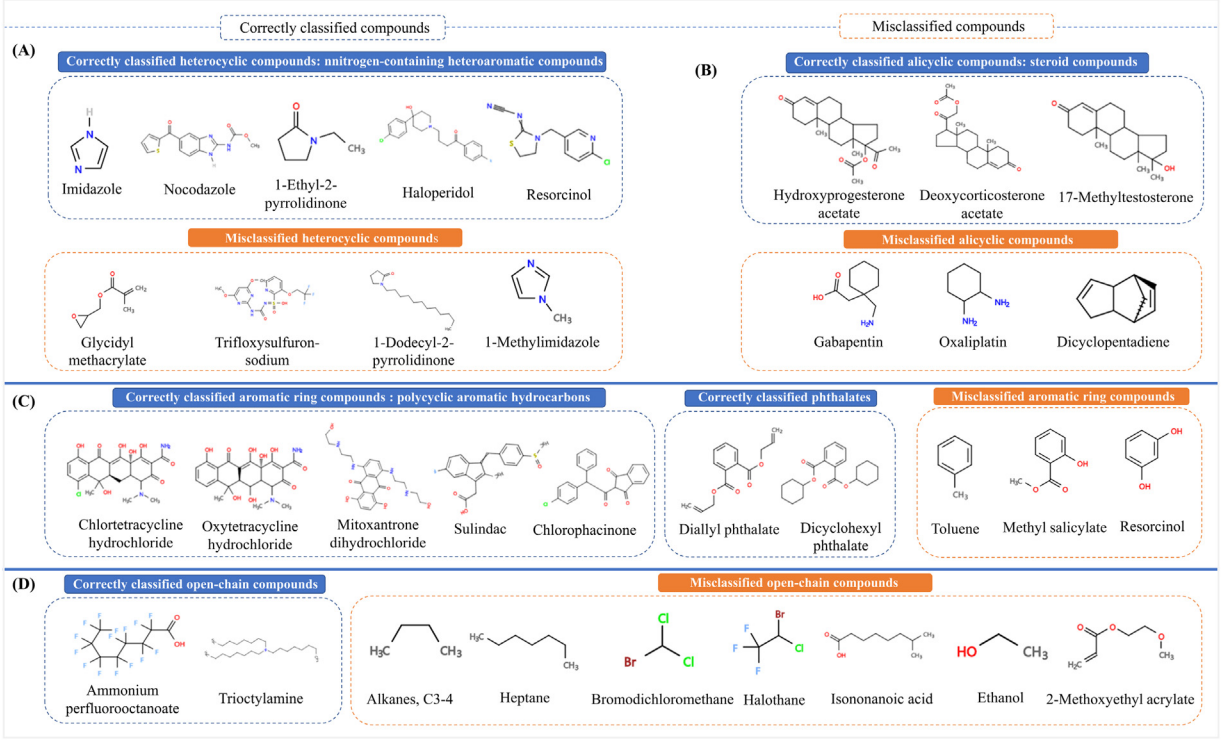

Fig. 4. Correctly classified compounds with complex structures and misclassified compounds with simple structures, including heterocyclic compounds (A), alicyclic compounds (B), aromatic ring compounds (C) and open chain compounds (D).

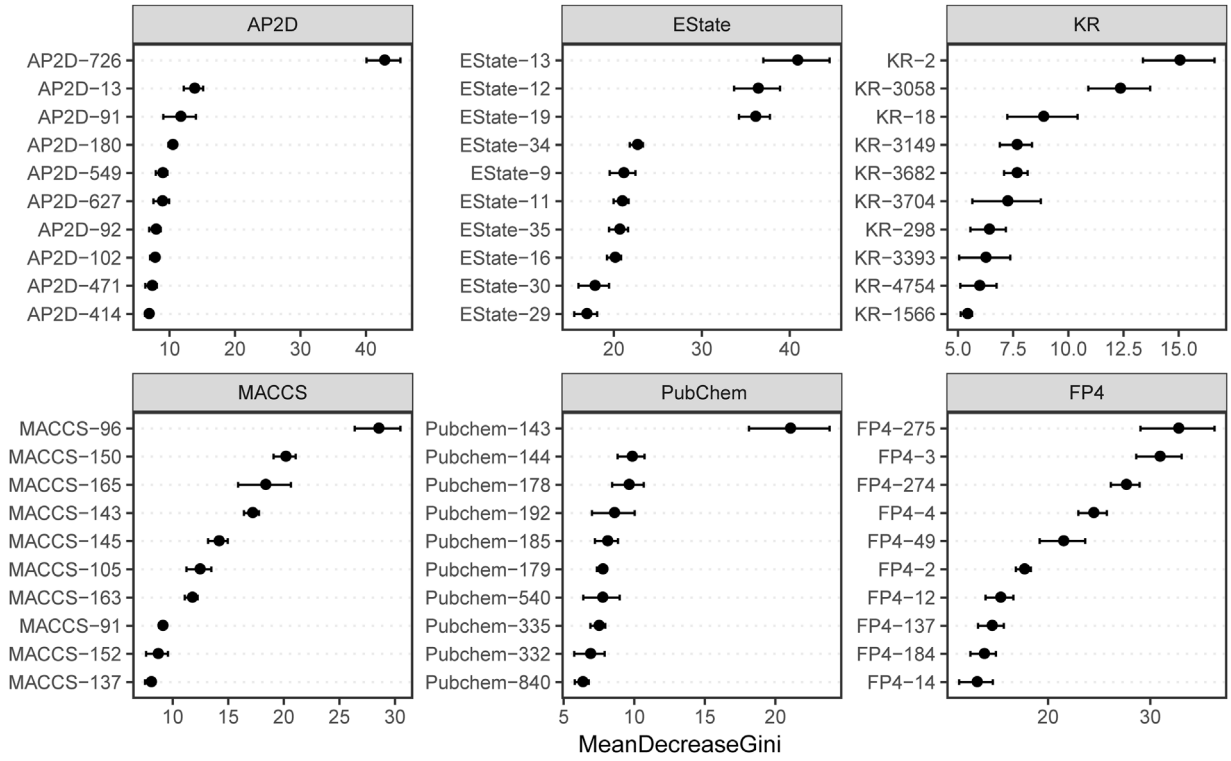

Fig. 5. The 10 most important fingerprint features from each RF model trained on six structure-key fingerprints. The statistical values of the MDG index are expressed as the mean ± standard deviation (SD).

Table 5
Descriptions of the 11 top-ranked fingerprint keys and their occurrence among reprotoxicants and non-reprotoxicants.

Fingerprint Key	Description	SMARTS Pattern	Present in Reprotoxicants	Present in Non-reprotoxicants	MDG
AP2D-726	O-O at topological distance 10		253	57	42.7
EState-13	sssCH	[CD3H](-*)(-*)-*	423	237	40.8
EState-12	aaCH	[C,c;D2H](:*):*	381	243	36.3
EState-19	ssssC	[CD4H0](-*)(-*)(-*)-*	289	134	36.0
KR-2		[!#1][CH]([!#1])[CH]([!#1])[!#1]	262	22	15.0
KR3058		C1CC1	280	48	12.3
MACCS-96	5 M ring	*1~*~*~*~*~1	364	66	28.4
Pubchem-143	Presence of any ring size 5		364	66	21.0
FP4−275	Heterocyclic	[!#6;!R0]	294	120	32.7
FP4−3	Tertiary carbon	[CX4H1]([#6])([#6])[#6]	312	118	30.8
FP4−274	Aromatic	a	387	247	27.5

deoxycorticosterone acetate, and 17-methyltestosterone (Fig. 4B). Polycyclic aromatic hydrocarbons and phthalates are common environmental toxic aromatic ring compounds that affect human reproductive health and cause damage to reproduction, development, and hormone secretion (Ramesh et al., 2017; Hlisníková et al., 2020). Here, our model not only successfully predicted polycyclic aromatic hydrocarbons as positive compounds, including naphthacene (e.g., chlortetracycline hydrochloride, oxytetracycline hydrochloride), anthraquinone (e.g., mitoxantrone dihydrochloride), and indene (e.g., sulindac, chlorophacinone) but also recognized phthalates (e.g., ciallyl phthalate, dicyclohexyl phthalate) (Fig. 4C). In contrast, monocyclic aromatic hydrocarbon compounds (e.g., toluene, methyl salicylate) and phenol (e.g., resorcinol) cannot be correctly predicted (Fig. 4C). Moreover, 23 out of 29 positive compounds with open-chain structures were misclassified as negative compounds, such as alkanes (e.g., C3−4, heptane), halogenated alkanes (e.g., bromodichloromethane, halothane), saturated monocarboxylic acids (e.g., isononanoic acid), monohydric alcohols (e.g., ethanol) and esters (e.g., 2-methoxyethyl acrylate) (Fig. 4D). The above results indicate that compounds with complex structures (e.g., cyclic compounds with multiple rings) were more easily predicted by the model, while compounds with simple structures were more likely to be misclassified because they may lack unique structural features. Therefore, the predictive performance of our model for these simple structure open-chain compounds remains unsatisfactory. Our future research will focus on improving the predictive performance of these compounds. Finally, it should be noted that our model cannot recognize inorganic compounds because we removed inorganic salts during data preprocessing.

3.6. Fingerprint features associated with reprotoxicity

To reveal the contributions of fingerprint features associated with reprotoxicity, the relative importance (measured as the Mean Decrease in the Gini coefficient, MDG) of each of these fingerprint features was evaluated by applying the RF algorithm. In the present study, 6 structure-key fingerprints, namely, AP2D, Estate, KR,

MACCS, PubChem and FP4, were used to analyse the importance of the structural features. We selected the 10 most important features based on the 10 highest MDG values for each fingerprint (Fig. 5). As shown in Fig. 5, a total of 11 features were found to have significantly higher MDG values, indicating that the substructures represented by these features may be closely associated with the reprotoxicity of compounds. As shown in Table 5, most of these features appear in nearly half of the compounds (1458 compounds), revealing which substructures are important in the prediction process. These fingerprint keys also occur more frequently in reprotoxicants than in non-reprotoxicants, indicating that reprotoxicants are structurally different from non-reprotoxicants and may have many distinctive substructures. Therefore, although the substructure patterns reported here are very simple and may not be suitable for the rule-based prediction of reprotoxicity for the generation of structural alerts (SAs), we suggest that these substructures should be considered in the early design of therapeutic compounds.

4. Conclusion

In this study, we used 1823 compounds based on multiple toxicity endpoints as a training set to establish reproductive toxicity classification prediction models based on ensemble learning. First, we used two feature selection methods, namely, low-variation feature filtering and high-correlation feature filtering with an optimal Tc threshold of 0.95, to remove redundant features from nine kinds of molecular fingerprints. Then, based on these 9 types of molecular fingerprints after feature selection, 27 basic prediction models for reproductive toxicity were constructed using 3 machine learning algorithms, namely, RF, SVM and XGBoost. These basic models showed high predictive performance, with ACC and AUC values ranging from 77.1 %–85.9 % and from 0.831 to 0.929, respectively. To further improve the prediction performance, 27 ensemble models were then developed by combining subsets of these basic models. Compared with the basic models, almost all of the ensemble models achieved higher ACC, SEN, SPE and AUC results, with the Ensemble-Top12 model in particular achieving an ACC of 86.33 %±0.08 %, a SEN of 82.02 %± 0.15 %, an SPE of 90.19 %±0.11 %, and an AUC of 0.937 ± 0.001 on the training set. The Ensemble-Top12 model also showed the best prediction performance on the external set, with an ACC of 84.38 %, a SEN of 77.33 %, an SPE of 90.67 %, and an AUC of 0.920. These findings indicate that ensemble learning can play a beneficial role in the prediction of reproductive toxicity and can significantly improve the ability to predict such toxic compounds. In addition, based on the Tanimoto distance, we define the applicability domain (AD) of the ensemble model. Compared with other models, the Ensemble-Top12 model shows relatively high ACC, SPE and AUC performance, and it could be used to effectively assess the risk of reprotoxicity of new drug candidates, especially in the early stages of drug discovery.

Author contributions

HWF, LZ and HSL conceived the project, developed the prediction method, designed, and implemented the experiments, analysed the result, and wrote the paper. SML, LLL, TZY, PYY implemented the experiments, analysed the result, and wrote the paper. JZ and ITA analysed the result. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Key R & D Program of Liaoning Province [Grant numbers 2019JH2/10300041, 2019JH5/10100041], Scientific Research Project from Department of Education of Liaoning Province [Grant number LQN201906], Youth Research Fund of Liaoning University [Grant number LDQN2019010], High-level innovation team foreign training project [Grant number 2018LNGXGJWPY-YB006], Excellent Chinese and Foreign Youth Exchange Plant Project from China Association for Science and Technology [Grant number 2018CASTQNJL50]. This project was supported by Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning and Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning Province.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.toxlet.2021.01.002.

References

Ai, H., Chen, W., Zhang, L., Huang, L., Yin, Z., Hu, H., Zhao, Q., Zhao, J., Liu, H., 2018. Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints. Toxicol. Sci. 165 (1), 100–107. doi:http://dx.doi.org/10.1093/toxsci/kfy121.

Andy, H.V., Terry, R.V., Rishi, R.G., Michael, J.L., Mohan, S.R., 2020. An overview of machine learning and big data for drug toxicity evaluation. Chem. Res. Toxicol. 33 (1), 20–37. doi:http://dx.doi.org/10.1021/acs.chemrestox.9b00227.

Bajusz, D., Rácz, A., Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminform. 7 (1), 20. doi:http://dx.doi.org/10.1186/s13321-015-0069-3.

Ballabio, D., Grisoni, F., Consonni, V., Todeschini, R., 2019. Integrated QSAR models to predict acute oral systemic toxicity. Mol. Inform. 38, 1800124 doi:http://dx.doi.org/10.1002/minf.201800124.

Basant, N., Gupta, S., Singh, K.P., 2016. QSAR modeling for predicting reproductive toxicity of chemicals in rats for regulatory purposes. Toxicol. Res. 5 (4), 1029–1038. doi:http://dx.doi.org/10.1039/c6tx00083e.

Benfenati, E., Chaudhry, Q., Gini, G., Dorne, J.L., 2019. Integrating in silico models and read-across methods for predicting toxicity of chemicals: a step-wise strategy. Environ. Int. 131 (2019), 105060. doi:http://dx.doi.org/10.1016/j.envint.2019.105060.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Brannen, K.C., Chapin, R.E., Jacobs, A.C., Green, M.L., 2016. Alternative models of developmental and reproductive toxicity in pharmaceutical risk assessment and the 3Rs. ILAR J. 57 (2), 144. doi:http://dx.doi.org/10.1093/ilar/ilw026.

Cassano, A., Manganaro, A., Martin, T., Young, D., Piclin, N., Pintore, M., Benfenati, E., 2010. CAESAR models for developmental toxicity. Chem. Cent. J. S4.

Catherine, M., Haslam, N.J., Holton, T.A., Gianluca, P., Shields, D.C., 2013. PeptideLocator: prediction of bioactive peptides in protein sequences. Bioinformatics 29 (9), 1120–1126. doi:http://dx.doi.org/10.1093/bioinformatics/btt103.

Cereto-massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-vallvé, S., Pujadas, G., 2015. Molecular fingerprint similarity search in virtual screening. Methods 71 (2015), 58–63. doi:http://dx.doi.org/10.1016/j.ymeth.2014.08.005.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, pp. 785–794.

Chen-lo, Y., Stefano, R.E., Torng, W., Altman, R.B., 2018. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today 23 (8), 1538–1546. doi:http://dx.doi.org/10.1016/j.drudis.2018.05.010.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Consonni, V., 2014. QSAR modeling: where have you been? Where are you going to?. J. Med. Chem. 57 (12), 4977–5010. doi:http://dx.doi.org/10.1021/jm4004285.

Chinaza, E., Jessica, L., Debashis, G., 2014. Mechanism of inhibition of estrogen biosynthesis by azole fungicides. Endocrinology (12), 4622–4628. doi:http://dx.doi.org/10.1210/en.2014-1561.

Cover, T.M., Hart, E.P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27. doi:http://dx.doi.org/10.1109/TIT.1967.1053964.

Gandhi, K., Schmidt, B., Ng, A.H., 2018. Towards data mining based decision support in manufacturing maintenance. Procedia CIRP 72 (2018), 261–265. doi:http://dx.doi.org/10.1016/j.procir.2018.03.076.

Ghorbanzadeh, M., Zhang, J., Andersson, P.L., 2016. Binary classification model to predict developmental toxicity of industrial chemicals in zebrafish. J. Chemom. 30 (6), 298–307. doi:http://dx.doi.org/10.1002/cem.2791.

González-medina, M., John, R.O., El-elimat, T., Cedric, J.P., Nicholas, H.O., Figueroa, M., Medina-franco, J.L., 2017. Scaffold diversity of fungal metabolites. Front. Pharmacol. 8, 180. doi:http://dx.doi.org/10.3389/fphar.2017.00180.

Guengerich, F.P., Macdonald, J.S., 2007. Applying mechanisms of chemical toxicity to predict drug safety. Chem. Res. Toxicol. 20 (3), 344–369. doi:http://dx.doi.org/10.1021/tx600260a.

Gunturi, S.B., Ramamurthi, N., 2014. A novel approach to generate robust classification models to predict developmental toxicity from imbalanced datasets. SAR QSAR Environ. Res. 25 (9), 711–727. doi:http://dx.doi.org/10.1080/1062936X.2014.942357.

He, S., Ye, T., Wang, R., Zhang, C., Zhang, X., Sun, G., Sun, X., 2019. An in silico model for predicting drug-induced hepatotoxicity. Int. J. Mol. Sci. 20 (8), 1987. doi:http://dx.doi.org/10.3390/ijms20081897.

Hlisníková, H., Petrovičová, I., Kolena, B., Šidlovská, M., Sirotkin, A., 2020. Effects and mechanisms of phthalates' action on reproductive processes and reproductive health: a literature review. Int. J. Environ. Res. 17 (8), 6811. doi:http://dx.doi.org/10.3390/ijerph17186811.

Höfer, T., Gerner, I., Gundert-remy, U., Liebsch, M., Schulte, A., Spielmann, H., Wettig, K., 2004. Animal testing and alternative approaches for the human health risk assessment under the proposed new European chemicals regulation. Arch. Toxicol. 78 (10), 549–564. doi:http://dx.doi.org/10.1007/s00204-004-0577-9.

Jensen, G.E., Niemelä, J.R., Wedebye, E.B., Nikolov, N.G., 2008. QSAR models for reproductive toxicity and endocrine disruption in regulatory use–a preliminary investigation. SAR QSAR Environ. Res. 19 (7/8), 631–641. doi:http://dx.doi.org/10.1080/10629360802550473.

Jiang, C., Yang, H., Di, P., Li, W., Liu, G., 2019. In silico prediction of chemical reproductive toxicity using machine learning. J. Appl. Toxicol. 1–11. doi:http://dx.doi.org/10.1002/jat.3772.

Jiang, C.S., Zhao, P., Li, W.H., Yun, T., Liu, G.X., 2020. In silico prediction of chemical neurotoxicity using machine learning. Toxicol. Res. 1–9. doi:http://dx.doi.org/10.1093/toxres/tfaa016.

Jie, C., Luo, J., Wang, S., Sheng, Y., 2018. Feature selection in machine learning: a new perspective. Neurocomputing 300 (2018) S092523121 doi:http://dx.doi.org/10.1016/j.neucom.2017.11.077.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. Kernlab-an S4 package for kernel methods in R. J. Stat. Softw. 11 (9), 1–20. doi:http://dx.doi.org/10.18637/jss.v011.i09.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 26 (5), 1–26. doi:http://dx.doi.org/10.18637/jss.v028.i05.

Kwon, S., Bae, H., Jo, J., Yoon, S., 2019. Comprehensive ensemble in QSAR prediction for drug discovery. BMC Bioinformatics 20 (2019), 521. doi:http://dx.doi.org/10.1186/s12859-019-3135-4.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Liu, M., Zhang, L., Li, S.M., Yang, T.Z., Liu, L.L., Jian, Z., Liu, H.S., 2020. Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints. Toxicol. Lett. 332, 88–96. doi:http://dx.doi.org/10.1016/j.toxlet.2020.07.003.

Lo piparo, E., Worth, A., 2010. Review of QSAR models and software tools for predicting developmental and reproductive toxicity. JRC Scientific and Technical Reports EUR 24522, .

Lu, J., Lu, D., Fu, Z., Zheng, M., Luo, X., 2018. Machine learning-based modeling of drug toxicity. Computational Systems Biology, , pp. 247–264. doi:http://dx.doi.org/10.1007/978-1-4939-7717-8_15.

Martin, M.T., Knudsen, T.B., Reif, D.M., Houck, K.A., Judson, R.S., Kavlock, R.J., Dix, D.J., 2011. Predictive model of rat reproductive toxicity from ToxCast high throughput screening. Biol. Reprod. 85 (2), 327–339. doi:http://dx.doi.org/10.1095/biolreprod.111.090977.

Marzo, M., Benfenati, E., 2018. Classification of a Naïve Bayesian fingerprint model to predict reproductive toxicity. SAR QSAR Environ. Res. 29 (8), 631–645. doi:http://dx.doi.org/10.1080/1062936X.2018.1499125.

Mora, J.R., Marrero-ponce, Y., García-jacas, C.R., Causado, A.S., 2020. Ensemble models based on QuBiLS-MAS features and shallow learning for the prediction of drug-induced liver toxicity: improving deep learning and traditional approaches. Chem. Res. Toxicol. 33 (7), 1855–1873. doi:http://dx.doi.org/10.1021/acs.chemrestox.0c00030.

Munos, B., 2009. Lessons from 60 years of pharmaceutical innovation. Nat. Rev. Drug Discov. 8 (12), 959–968. doi:http://dx.doi.org/10.1038/nrd2961.

Ramesh, A., Kenneth, J.H., Archibong, A.E., 2017. Reproductive toxicity of polycyclic aromatic hydrocarbons. Reprod. Dev. Toxicol. 745–763. doi:http://dx.doi.org/10.1016/B978-0-12-382032-7.10043-8.

Roy, K., Kar, S., Das, R.N., 2015. Validation of QSAR model. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, , pp. 231–289. doi:http://dx.doi.org/10.1016/B978-0-12-801505-6.00007-7.

Satpathy, R., 2019. Quantitative structure–activity relationship methods for the prediction of the toxicity of pollutants. Environ. Chem. Lett. 17 (1), 123–128. doi:http://dx.doi.org/10.1007/s10311-018-0780-1.

Scialli, A.R., 2008. The challenge of reproductive and developmental toxicology under REACH. Regul. Toxicol. Pharmacol. 51 (2), 244–250. doi:http://dx.doi.org/10.1016/j.yrtph.2008.04.008.

Sheridan, R.P., Wang, W.M., Liaw, A., Ma, J., Gifford, E.M., 2016. Extreme gradient boosting as a method for quantitative structure–activity relationships. J. Chem. Inf. Model. 56 (12), 2353–2360. doi:http://dx.doi.org/10.1021/acs.jcim.6b00591.

Siramshetty, V.B., Nickel, J., Omieczynski, C., Gohlke, B.O., Drwal, M.N., Preissner, R., 2015. WITHDRAWN—a resource for withdrawn and discontinued drugs. Nucleic Acids Res. 44 (D1), DD108–D1080. doi:http://dx.doi.org/10.1093/nar/gkv1192.

Webb, S., Hanser, T., Howlin, B., Vessey, J., 2014. Feature combination networks for the interpretation of statistical machine learning models: application to Ames mutagenicity. J. Cheminform. 6 (2014), 8. doi:http://dx.doi.org/10.1186/1758-2946-6-8.

Yang, H., Sun, L., Li, W., Liu, G., Tang, Y., 2018. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front. Chem. 6, 30. doi:http://dx.doi.org/10.3389/fchem.2018.00030.

Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32 (7), 1466–1474. doi:http://dx.doi.org/10.1002/jcc.21707.

Yin, Z., Ai, H., Zhang, L., Ren, G., Wang, Y., Zhao, Q., Liu, H., 2019. Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints. J. Appl. Toxicol. 39 (10), 1366–1377. doi:http://dx.doi.org/10.1002/jat.3785.

Zaslavskiy, M., Jégou, S., Tramel, E.W., Wainrib, G., 2019. ToxicBlend: virtual screening of toxic compounds with ensemble predictors. Comput. Toxicol. 10, 81–88. doi:http://dx.doi.org/10.1016/j.comtox.2019.01.001.

Zhang, H., Ren, J.X., Kang, Y.L., Bo, P., Liang, J.Y., Ding, L., Zhang, J., 2017a. Development of novel in silico model for developmental toxicity assessment by using naive Bayes classifier method. Reprod. Toxicol. 71 (2017), 8–15. doi:http://dx.doi.org/10.1016/j.reprotox.2017.04.005.

Zhang, H., Mao, J., Qi, H.Z., Ding, L., 2019. In silico prediction of drug-induced developmental toxicity by using machine learning approaches. Mol. Divers. 1–10. doi:http://dx.doi.org/10.1007/s1103 0-019-09991-y.

Zhang, L., Ai, H., Chen, W., Yin, Z., Hu, H., Zhu, J., Liu, H., 2017b. CarcinoPred-EL: novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. Sci. Rep. 7 (1), 2118. doi:http://dx.doi.org/10.1038/s41598-017-02365-0.