

Cite this: *Toxicol. Res.*, 2018, 7, 211

## *In silico* prediction of chemical genotoxicity using machine learning methods and structural alerts†

Defang Fan, Hongbin Yang, Fuxing Li, Lixia Sun, Peiwen Di, Weihua Li, Yun Tang \* and Guixia Liu 

Genotoxicity tests can detect compounds that have an adverse effect on the process of heredity. The *in vivo* micronucleus assay, a genotoxicity test method, has been widely used to evaluate the presence and extent of chromosomal damage in human beings. Due to the high cost and laboriousness of experimental tests, computational approaches for predicting genotoxicity based on chemical structures and properties are recognized as an alternative. In this study, a dataset containing 641 diverse chemicals was collected and the molecules were represented by both fingerprints and molecular descriptors. Then classification models were constructed by six machine learning methods, including the support vector machine (SVM), naïve Bayes (NB), k-nearest neighbor (kNN), C4.5 decision tree (DT), random forest (RF) and artificial neural network (ANN). The performance of the models was estimated by five-fold cross-validation and an external validation set. The top ten models showed excellent performance for the external validation with accuracies ranging from 0.846 to 0.938, among which models Pubchem\_SVM and MACCS\_RF showed a more reliable predictive ability. The applicability domain was also defined to distinguish favorable predictions from unfavorable ones. Finally, ten structural fragments which can be used to assess the genotoxicity potential of a chemical were identified by using information gain and structural fragment frequency analysis. Our models might be helpful for the initial screening of potential genotoxic compounds.

Received 29th September 2017,  
Accepted 14th December 2017

DOI: 10.1039/c7tx00259a

rsc.li/toxicology-research

## Introduction

In daily life, human beings are exposed to diverse chemicals, especially drugs, pesticides, food additives and cosmetic ingredients. It is necessary to evaluate these chemicals by means of toxicity tests, such as genotoxicity tests.<sup>1</sup> Genetic toxicology is the study of the adverse effects of chemical and physical agents on the process of heredity. Several methods can be used to measure the genetic toxicity of compounds, such as the comet assay,<sup>2</sup> micronucleus assay (MN),<sup>3</sup> chromosomal aberration assay,<sup>4</sup> bacterial reverse mutation test,<sup>5</sup> and sister chromatid exchange assay.<sup>6</sup> Among them, the *in vivo* micronucleus assay is the most commonly used method to investigate the *in vivo* genotoxic potential of chemicals, and it has been successfully performed for following the testing of positive *in vitro* results.<sup>7</sup> Negative results *in vitro* are usually considered sufficient to indicate the lack of mutagenicity, whereas a positive result is not considered sufficient to indicate that the

chemical represents a mutagenic hazard, which could be a false positive. The *in vivo* micronucleus assay can detect chemicals with the ability to disrupt the process of mitosis and form a “micronucleus”. It is morphologically similar to a normal nucleus but it has a smaller size. Since animal model assays are time-consuming, highly expensive and unethical, *in silico* methods are gradually developed as alternatives to experimental tests. In addition, REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) has provisions that facilitate the use of data generated by non-testing methods, specifically of the (quantitative) structure–activity relationship.<sup>8</sup>

With the development of computer science and cheminformatics, *in silico* modeling has become a powerful tool to assist the drug discovery and predict pharmacokinetic and toxic properties.<sup>9</sup> Over the past few decades, many structure–activity models<sup>10–12</sup> have been constructed to assess the genotoxicity of chemicals, such as the chemical Ames mutagenicity, carcinogenicity and chromosome aberration test. Most of the models were built with molecular fingerprints or descriptors, and showed excellent predictive power. For the micronucleus assay, there are few models for predicting *in vivo* micronucleus results. Romualdo Benigni *et al.*<sup>13</sup> have studied structural alerts (SAs) based on a database. However, in their research SAs were applied to an unbalanced database with a large pro-

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China.

E-mail: gxliu@ecust.edu.cn, ytang234@ecust.edu.cn; Fax: +86-21-64251033;

Tel: +86-21-64250811

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c7tx00259a

portion of micronucleus negative chemicals. Meanwhile, the SAs had a low frequency in the micronucleus positive chemicals. In view of the lack of available structure–activity relationship models (SAR) for predicting chemical genotoxicity, models for predicting chemical genotoxicity based on *in vivo* micronucleus assay results were built to fill this gap.

In this study, a large dataset containing 641 diverse compounds was used for modeling. On the basis of this dataset, binary classification models were built by using six kinds of machine learning methods. In order to validate the effectiveness of the models, five-fold cross validation and an external validation set were applied, and they showed excellent predictive ability. Meanwhile, high-frequency structural fragments in micronucleus positive chemicals were identified using information gain and structural fragment frequency analysis. Furthermore, in line with the organization for economic co-operation and development (OECD) principle (<http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsar-models.htm>), we defined the application domain to improve the predictive reliability of our models. This study may provide a useful tool to identify the genetic toxicity of chemicals.

## Materials and methods

### Data collection and preparation

In this study, the dataset was collected from a report<sup>14</sup> and the webserver of eChemPortal which is a part of OECD. The compounds were labeled as negative or positive according to the *in vivo* micronucleus assay results. If a compound shows the ability to induce chromosomal damage or disrupt the cell division, it would be labeled as positive. Conversely, it is labeled as negative. All the data were prepared by the following steps by using Pipeline Pilot 7.5. Firstly, entries containing inorganic compounds, organometals, and mixtures were removed. Secondly, salts were converted into the corresponding acids or bases, and water molecules were removed from the hydrates. Thirdly, chemicals that were duplicated and contradictory in different sources were removed. As a result, a total of 641 chemicals (including 264 micronucleus positive chemicals and 377 micronucleus negative chemicals) were obtained. The dataset is comparatively balanced compared with the previous publication.<sup>13</sup> Finally, the dataset was randomly split into a training set and external validation set in the ratio of 9 : 1. The detailed statistical description of the datasets is listed in Table 1. The names, SMILES and CAS number of all compounds are available in the ESI (Table SI1†).

**Table 1** Statistical data of chemicals used in the training set and the external validation set

Datasets	Positive	Negative	Total
Training set	237	339	576
External validation set	27	38	65
Total	264	377	641

### Calculation of molecular fingerprints

Molecular fingerprints are a string representation of chemical structures designed to enhance the efficiency of chemical database searching and analysis. A molecule is described as a binary string in which each bit corresponds to a particular structural fragment. Six commonly used fingerprints were calculated for all molecules by PaDEL-Descriptor, respectively.<sup>15</sup> They are CDK fingerprint (FP, 1024 bits), CDK Extended fingerprint (Ext, 1024 bits), Estate fingerprint (Estate, 79 bits), MACCS fingerprint (MACCS, 166 bits), Pubchem fingerprint (Pubchem, 881 bits) and Substructure fingerprint (Sub, 307 bits).

### Calculation of molecular descriptors and feature selection

In this study, five feature groups of molecular descriptors (constitutional descriptors, Basak descriptors, Burden descriptors, CATS descriptors and MOE-type descriptors) were calculated by ChemsAR.<sup>16</sup> 325 features, which are related to the physico-chemical and structural properties of the studied molecules, were contained in these five groups.

The selection of molecular descriptors is very important for model building.<sup>17</sup> And the selection of features was only applied to the training set. Four methods were used to select molecular descriptors. Firstly, if the values of the descriptor are all zero or with a variance lower than the threshold of 0.05, it would be removed. Then, correlations across all pairs of descriptors were calculated, and any two descriptors with correlation values higher than 0.95 were regarded as redundant, between which the less correlative one was abandoned. Besides, the importance of each feature was calculated by tree-based estimators which in turn can be used to remove irrelevant descriptors. Furthermore, recursive feature elimination (RFE) with the linear kernel support vector machine was performed in a cross-validation loop to select an optimal number of features. This algorithm can calculate and update the importance ranks, and eliminate the least important feature accordingly. Finally, the subset of descriptors that showed the best prediction performance was selected.

### Model building by machine learning methods

Machine learning is widely used in building models. Six different machine learning methods were used to build models implemented by Orange 2.0 (freely available at <https://orange.biolab.si/>). They are the support vector machine (SVM),<sup>18</sup> naïve Bayes (NB),<sup>19</sup> k-nearest neighbor (kNN),<sup>20</sup> C4.5 decision tree (DT),<sup>21</sup> random forest (RF)<sup>22</sup> and artificial neural network (ANN).<sup>23</sup> The SVM algorithm was operated in the open source LIBSVM 3.2 package,<sup>24</sup> and the Gaussian radial basis function (RBF) was used as the kernel function. Meanwhile, its parameters were optimized through a python script in the LIBSVM 3.2 package. The parameters of the other five machine learning methods were optimized through a grid search approach to find the highest area under the receiver operation characteristic curve (AUC) value based on a five-fold

cross-validation. The detailed grid parameters for different machine learning methods are listed in Table S12.†

### Assessment of the models

The classification models are assessed by five-fold cross-validation and a diverse external validation set.<sup>25</sup> All the models are assessed by the counts of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Accuracy (CA) is the total percentage of chemicals that were correctly predicted. Specificity (SP) denotes the predictive accuracy of the micronucleus negative chemicals. Sensitivity (SE) means the predictive accuracy of the micronucleus positive chemicals and the AUC value is the area under the receiver operating characteristic curve (ROC). The equations of SE, SP and CA are shown as follows.

$$CA = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$SP = TN / (TN + FP) \quad (2)$$

$$SE = TP / (TP + FN) \quad (3)$$

### Definition of the applicability domain

It is well known that the predictive reliability is an important issue in the consideration of the fact that any QSAR/QSTR model is characterized by its applicability domain (AD). Moreover, the OECD has recommended a strict requirement to define the applicability domain.<sup>26</sup> A similarity-based applicability domain analysis was conducted by comparing the similarity between a query chemical and the chemicals in the training set.<sup>27</sup> If there is a high similarity between the query chemical and the molecules in the training set, the molecule would be reliably predicted by our models. The Tanimoto coefficient ( $T$ ) was used to assess the similarity between two molecules with the MACCS fingerprint. The average Tanimoto coefficient ( $\bar{T}$ ) represents the average similarity between a query chemical and each compound in the training set. The Tanimoto coefficient and the average Tanimoto coefficient are defined as follows:

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

$$\bar{T}(x) = \frac{1}{n} \sum_{i=1}^n T(x, x_i) \quad (5)$$

where  $A$  and  $B$  are the fingerprints of two chemicals.  $x$  represents the MACCS fingerprint of a query chemical, and  $x_i$  represents the fingerprint of the  $i$ -th chemical in the training set. The higher the  $\bar{T}$  value is, the more similar the molecule is to the training set.

AD was optimized in order to not only guarantee the reliability of the models but also to cover a wide range of compounds. Several thresholds were used to divide the chemicals into in-domain (ID) and out-of-domain (OD). The AUC values were used to assess the performance of our models.<sup>28</sup> The AUC

values of different models for predicting ID and OD chemicals in the external validation set were calculated.

### Privileged substructure analysis

Structural alerts (SAs) are defined as chemical substructures, whose presence shows the relationship with the capability of a substance to cause certain adverse effects on organs.<sup>29</sup> Several methods can be used to identify privileged substructures, such as MoSS (graph-based)<sup>30</sup> and SARpy (fragment-based).<sup>31</sup> For the identification of SAs, our previous publication has reported that SARpy could detect highly accurate substructures.<sup>32</sup> So in this study, SARpy was used to identify the high-frequency fragment in the micronucleus positive chemicals.<sup>33</sup> The privileged substructures were assessed by information gain (IG) and substructure fragment analysis. IG measures the information entropy of a classification system obtained for class prediction by knowing the presence or absence of a pattern in a molecule.<sup>34</sup> If a substructure showed a high frequency in the micronucleus positive molecules, this substructure would be defined as a privileged substructure. Their appearance in a molecule can alert researchers to pay more attention to this molecule. The frequency of a fragment is defined as follows:

$$\text{Frequency of a fragment} = \frac{N_{\text{fragment\_class}} \times N_{\text{total}}}{N_{\text{fragment\_total}} \times N_{\text{class}}} \quad (6)$$

where  $N_{\text{fragment\_class}}$  is the number of chemicals containing the fragments in micronucleus positive chemicals;  $N_{\text{total}}$  represents the total number of chemicals;  $N_{\text{fragment\_total}}$  is the total number of chemicals containing the fragments and  $N_{\text{class}}$  represents the number of micronucleus positive chemicals.

## Results

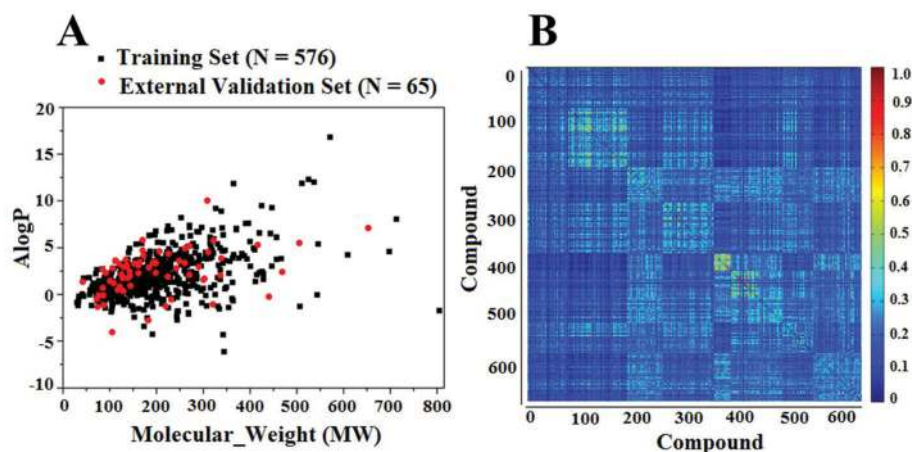
### Dataset analysis

A total of 641 non-duplicated chemicals were collected to build the models. In order to investigate the chemical space distribution, we calculated the molecular weight (MW) and Ghose-Crippen LogKow ( $A \log P$ ) for the training set and the external validation set. The distribution scatter diagram is presented in Fig. 1(A). The MW of the training set was distributed from 30 to 804 (mainly from 50 to 500), and  $A \log P$  ranged from  $-6$  to  $16$  (mainly from  $-2$  to  $5$ ). It can be observed that the chemical space of the external validation set shared a similar chemical space and it was within the scope of the training set.

To further explore the chemical diversity of the dataset, the Tanimoto coefficient was calculated for the whole dataset by using the MACCS fingerprint. The average Tanimoto index is  $0.19$ . As shown in Fig. 1(B), red points indicate the high similarity between two compounds and blue points indicate the low similarity. It can be concluded that the structure of the compounds in the dataset was diverse.

### Selection of molecular descriptors

The aim of the selection of descriptors is to remove redundant and irrelevant descriptors. Approximately 325 descriptors were



**Fig. 1** (A) Diversity distribution of the training set and the external validation set. The chemical space is defined by molecular weight (MW) as the X-axis, and A log P as the Y-axis. *N* represents the number of chemicals in different datasets. The training set is shown as black squares, while the external validation set is in red circles. (B) Heat map of the molecular similarity plotted by the Tanimoto similarity using the MACCS fingerprint.

calculated for each compound by using the feature groups of constitutional descriptors, Basak descriptors, Burden descriptors, CATS descriptors and MOE-type descriptors.<sup>35</sup> After a stream of descriptors' analysis procedure, 49 representative molecular descriptors were selected for modeling. Among them, 20 constitutional and CAST descriptors, such as the number of H-bond donors, the count of heteroatoms, the number of rings, lipophilicity and the average molecular weight descriptors, described the molecular structural information and physicochemical properties. 14 Basak descriptors reflected the molecular polarity and polarizability.<sup>36</sup> Besides, 6 Moe-Type descriptors are related to the van der Waals surface areas, such as MRVSA6, MRVSA8, MRVSA9, VSAEstate7, VSAEstate8 and PEOEVSA12. Furthermore, 9 Burden descriptors characterized the atomic masses, atomic van der Waals volumes and atomic Sanderson electronegativities. Molecular

polarity, electronegativity and the number of heteroatoms are important characteristics, which promote the insertion of chemicals into DNA, and thus destroy the structure of DNA.<sup>37</sup> The detailed information of these descriptors can be found in the ESI (Table SI3†). And the heat map of these descriptors' inter-correlation is shown in Fig. SI1.† The average correlation coefficient is 0.26, which showed a low correlation for the descriptors.

#### Performance of cross-validation

In this study, the models were built by six kinds of machine learning methods combined with six fingerprints and 49 molecular descriptors. On the basis of their accuracy, the top ten models built from fingerprints and the five models resulting from descriptors are presented in Table 2, the CA of the fifteen models ranged from 0.821 to 0.889; the SE of the fifteen

**Table 2** Performance of the classification models (top ten models developed by fingerprints and five models developed by descriptors) for the five-fold cross validation using different fingerprints as well as descriptors

Molecule representation	Method <sup>a</sup>	CA	SE	SP	AUC
Fingerprint	Pubchem_SVM	0.889	0.923	0.840	0.948
	MACCS_RF	0.882	0.841	0.910	0.947
	FP_SVM	0.877	0.917	0.819	0.928
	Estate_SVM	0.872	0.929	0.789	0.930
	Pubchem_ANN	0.872	0.906	0.823	0.938
	MACCS_SVM	0.872	0.897	0.835	0.931
	Estate_ANN	0.865	0.912	0.797	0.927
	Ext_SVM	0.865	0.917	0.789	0.926
	Sub_SVM	0.863	0.909	0.797	0.925
	MACCS_NB	0.852	0.885	0.806	0.904
Descriptor	SVM	0.882	0.853	0.901	0.952
	RF	0.861	0.845	0.872	0.933
	kNN	0.859	0.797	0.901	0.932
	NB	0.842	0.819	0.858	0.900
	DT	0.821	0.754	0.866	0.810

<sup>a</sup> ANN: artificial neural network, DT: C4.5 decision tree, kNN: k-nearest neighbor, NB: naïve Bayes, RF: random forest, SVM: support vector machine; Estate: Estate fingerprint, Ext: CDK extended fingerprint, FP: CDK fingerprint, MACCS: MACCS fingerprint, Pubchem: Pubchem fingerprint, and Sub: Substructure fingerprint.

models ranged from 0.754 to 0.929; the SP ranged from 0.789 to 0.910 and AUC values ranged from 0.810 to 0.952. Based on the value of CA, the top three models are Pubchem\_SVM (CA = 0.889, SE = 0.923, SP = 0.840, AUC = 0.948), MACCS\_RF (CA = 0.882, SE = 0.841, SP = 0.910, AUC = 0.947) and descriptors\_SVM (CA = 0.882, SE = 0.853, SP = 0.901, AUC = 0.952).

According to the results of the five-fold cross-validation, the Pubchem fingerprint and MACCS fingerprint combined with SVM and RF have higher accuracy, which is consistent with our previous study.<sup>10</sup> Meanwhile, the descriptors combined with SVM and RF outperformed other machine learning methods. Generally, compared with the descriptor used in this study, a fingerprint is a better way to represent molecules and has a better performance. The detailed performance of the five-fold cross-validation for all models can be found in the ESI (Table S14†).

Furthermore, Y-randomization was carried out to validate the robustness of the models.<sup>38</sup> Keeping the original independent variable constant, the dependent variable vector was randomly shuffled, and then a new model was developed. This procedure was repeated three times. The AUC values of the three new models for external validation are shown in Fig. 2. As was expected, the three new models had low AUC values, which indicated that the results in our original models are not accidental.

### Performance of the external validation set

The external validation set was used to validate the performance of the fifteen models mentioned above. The results of the fifteen models for the external validation set are shown in

Fig. 3. Considering the values of CA, all the models showed excellent predictive ability with a CA higher than 0.840. The top four models were MACCS\_RF (CA = 0.938, SE = 0.947, SP = 0.926, AUC = 0.963), Pubchem\_ANN (CA = 0.923, SE = 0.921, SP = 0.926, AUC = 0.963), Descriptor\_RF (CA = 0.922, SE = 0.903, SP = 0.939, AUC = 0.974) and Pubchem\_SVM (CA = 0.908, SE = 0.895, SP = 0.926, AUC = 0.980) with the CA value over 0.900. The best model was MACCS\_RF, which showed an accuracy of 0.937 and ranked second in the cross-validation. The accuracy of Pubchem\_SVM is 0.908, which ranked fourth in the external validation results and performed best in the cross-validation. It can be concluded that the fingerprints of Pubchem and MACCS in SVM, and the RF algorithm performed better, which is consistent with the five-fold cross-validation. Meanwhile, the SP and SE of the models were relatively balanced, which may be ascribed to the comparatively balanced dataset.

### Analysis of the structural alerts

To investigate privileged structural fragments in micronucleus positive chemicals, substructure frequency analysis and information gain were performed for all the datasets. Substructure frequency analysis was implemented by SARpy. Ten structural fragments were identified, which showed a higher frequency in micronucleus positive chemicals than in micronucleus negative chemicals.

The privileged fragments and corresponding representative chemicals are shown in Table 3. These structural fragments are aromatic nitro compounds, benzimidazole, benzidine, aniline, aziridine, epoxy propane, thiophosphate, aromatic diazo, formamide or thioformamide and cyanide. A new compound containing one or more structural fragments has a high

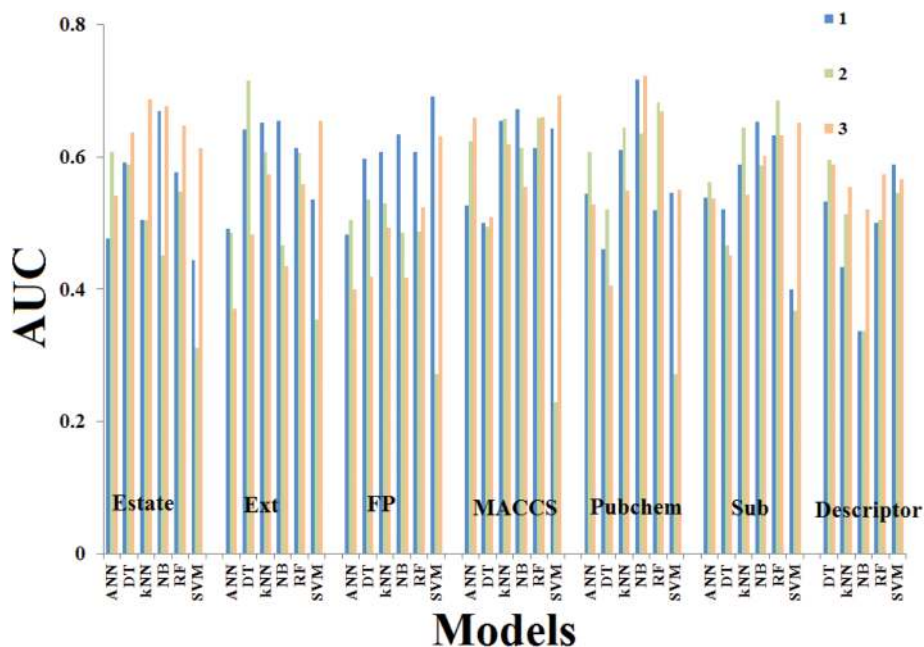
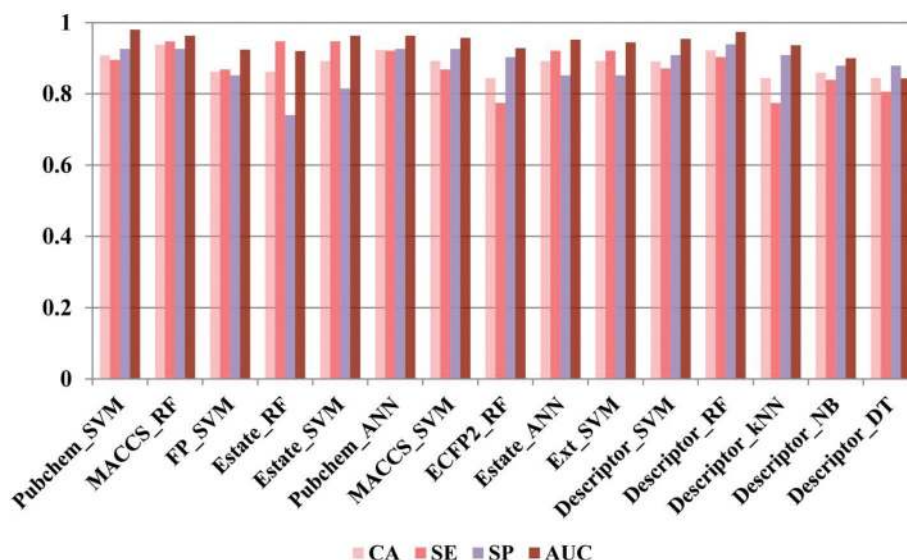


Fig. 2 AUC values of different models after Y-randomization tests three times. Blue, green and orange represent three Y-randomization tests, respectively.



**Fig. 3** Performance of the fifteen models for the external validation set using different algorithms combined with different fingerprints or descriptors.

**Table 3** Representative structural alerts and the corresponding value of information gain (IG) and the frequency value of privileged substructures

No.	Description	General structure	Representative structures	IG	$F_m$
1	Aromatic nitro compounds			0.086	2.428
2	Benzimidazole			0.040	2.428
3	Benzidine			0.020	2.428
4	Aniline			0.138	2.150
5	Aziridine			0.012	2.428
6	Epoxy propane			0.008	1.820
7	Thiophosphate			0.012	2.428
8	Cyanide			0.020	2.428
9	Formamide or thioformamide			0.054	1.970
10	Aromatic diazo			0.047	2.428

possibility to be micronucleus positive. The higher the information gain value, the more important the substructure.

Six of these fragments were covered in Toxtree, in which four were different. The first structural fragment is benzimida-

zole. Many herbicides belong to the benzimidazole derivative,<sup>39</sup> and 20 benzimidazole derivatives were shown in the micronucleus positive chemicals. The second fragment is formamide or thioformamide. A total of 69 chemicals contained

the fragment of formamide or thioformamide. Among them, 56 chemicals are micronucleus positive. The third is thiophosphate. For the thiophosphate, six micronucleus positive chemicals contain the fragment. The last is cyanide, which can cause toxicity *via* multiple routes, including inhalation, ingestion, parenteral administration and dermal or conjunctival contact.<sup>40</sup> There are 10 types of cyanide in the micronucleus positive chemicals.

## Discussion

### Comparison of different methods

Six algorithms combined with six kinds of fingerprints were utilized to build models. Comparing the six machine learning methods, SVM and RF algorithms performed better than others when combined with fingerprints. Meanwhile, they also showed excellent predictive ability combined with descriptors, which indicates that SVM and RF algorithms are more suitable to predict chemical genotoxicity. As for the RF method, it performed best when combined with the MACCS fingerprint, which indicated that the RF method is more suitable for the MACCS fingerprint.

As is well known, random forest is currently considered one of the best QSAR methods available for prediction.<sup>41</sup> The SVM algorithm is widely used in building QSAR models with a great power to fit the nonlinear relationship.<sup>10</sup> Many experiments

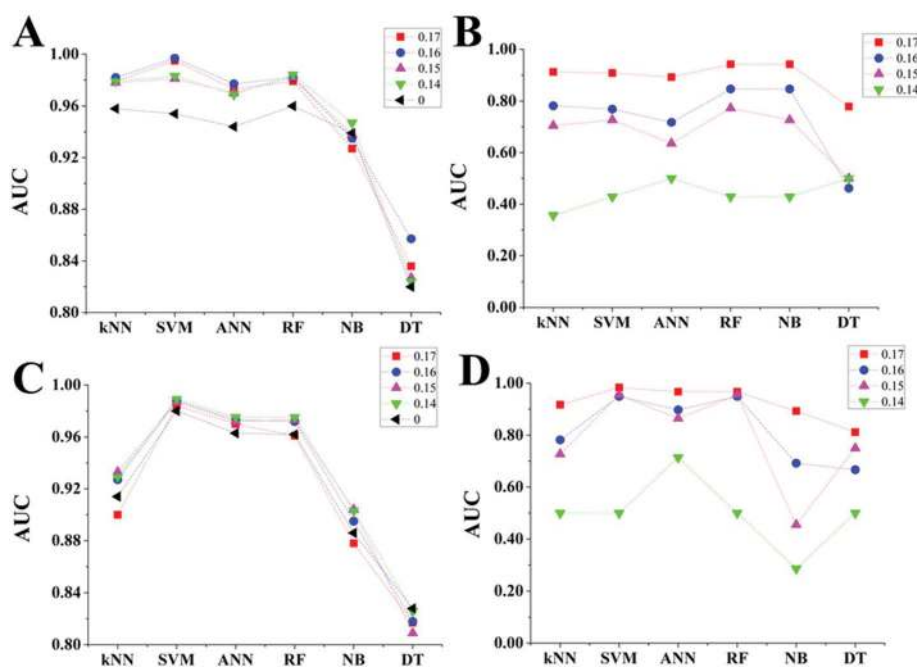
have validated that the SVM algorithm can result in a higher accuracy than other algorithms.<sup>42</sup>

### Comparison of the models built by fingerprints and descriptors

The fingerprint consists of structural fragments. It can reflect the structural characteristics of a molecule, while molecular descriptors represent the physicochemical and topological properties of a molecule. In this study, the fingerprints and the descriptors were used to represent molecules to build the classification models. The results indicate that both molecular fingerprints and descriptors can build excellent models (Tables 2 and SI4†). For some algorithms, using molecular descriptors showed a higher CA than using fingerprints. For example, models built by molecular descriptors with kNN algorithms had better performance than models built by fingerprints with the same algorithms. However, comparing

**Table 4** The number of chemicals determined to be in-domain and out-of-domain in the validation sets using application domain assessment methods with different thresholds

Threshold	In-domain (ID)	Out-of-domain (OD)
0	65	0
0.17	42	23
0.16	49	16
0.15	52	13
0.14	57	8



**Fig. 4** The performance of different models to predict in-domain (ID) and out-of-domain (OD) chemicals. (A) The AUC values of ID chemicals in different models with the MACCS fingerprint. (B) The AUC values of OD chemicals in different models with the MACCS fingerprint. (C) The AUC values of ID chemicals in different models with the Pubchem fingerprint. (D) The AUC values of OD chemicals in different models with the Pubchem fingerprint. The black line represents the undivided external validation set, the red line represents the threshold of 0.17, the blue line represents the threshold of 0.16, magenta represents the threshold of 0.15, and the green line represents the threshold of 0.14.

the top models, we found that using molecular fingerprints is more suitable to build the classification models for genotoxicity. For instance, for the RF algorithm, a model built by the MACCS fingerprint outperformed the model constructed by descriptors.

### Comparison with the Toxtree software

Toxtree is a flexible and user-friendly predictive toxicology open-source application (<https://sourceforge.net/projects/toxtree/>). It places chemicals into categories and predicts various kinds of toxic effects by applying decision tree approaches. For the micronucleus assay, 36 structural alerts were used to identify *in vivo* micronucleus positive chemicals in Toxtree.<sup>13</sup> In our study, 10 structural alerts were identified. Six of these structural alerts were also covered in Toxtree. These structural alerts have a higher positive prediction ability, which may be ascribed to a larger proportion of positive chemicals in our data.

### Analysis of the applicability domain

In order to find the optimized AD, four different thresholds (ranging from 0.17 to 0.14 with a step size of 0.01) were determined according to the average Tanimoto coefficient of the training set (0.18). The chemical is treated as the ID chemical in case the average Tanimoto coefficient is higher than the threshold. Otherwise, it is considered as the OD chemical. As shown in Table 4, with the decrease of the threshold, more test compounds were considered as structurally similar to the training set. Models built by MACCS and Pubchem fingerprints were used to validate the optimum threshold in consideration of the better performance in the five-fold cross-validation.

To investigate their respective impact on the predictive accuracy, their performance with different thresholds was compared. The AUC values of different models for ID and OD chemicals are summarized in Fig. 4. The threshold of zero represents the undivided external validation set, *i.e.*, all the compounds would be ID chemicals. On one hand, an improvement of AUC values can be observed after using the AD to pick out ID chemicals. On the other hand, as the threshold decreased from 0.17 to 0.14, fewer and fewer compounds were divided into OD. And the predictive power for OD chemicals declined, which indicated that our models had a relatively worse predictive ability for OD chemicals. Compared to the models built by the Pubchem fingerprint, models built by the MACCS fingerprint improved greatly after using AD to pick out ID chemicals, which may result from the fact that the MACCS fingerprint was used to calculate the Tanimoto coefficient. In conclusion, considering the prediction ability and application range of models, the threshold of 0.16 is more suitable.

## Conclusion

In this study, binary classification models were developed to predict chemical genotoxicity based on a diverse dataset. Five-fold cross-validation was utilized to validate the robustness of

our models and an external validation set was used to validate the predictive ability of the models. The models showed high predictive accuracy by using molecular fingerprints or descriptors as attributes. The Pubchem fingerprint combined with the SVM algorithm and the MACCS fingerprint combined with the RF algorithm performed better. Meanwhile, ten privileged structural fragments were identified by information gain and structural frequency analysis. If a molecule contains structural fragments, it has a high possibility of genotoxicity. These structural fragments can make researchers pay more attention to chemicals that contain structural fragments. The applicability domain was defined to improve the predictive accuracy of the models. Furthermore, all the tools used in this study are free and easy to access. This study provided a useful strategy for evaluating the genotoxicity property of chemicals. And the modeling methods used in this article can also be applied to other genotoxicity assay end points.

## Conflicts of interest

The authors declare that there are no conflicts of interest.

## Acknowledgements

We gratefully acknowledge the financial support from the National Natural Science Foundation of China (Grants 81273438, 81373329 and 81673356).

## References

- 1 M. Aiba née Kaneko, M. Hirota, H. Kouzuki and M. Mori, Prediction of genotoxic potential of cosmetic ingredients by an *in silico* battery system consisting of a combination of an expert rule-based system and a statistics-based system, *J. Toxicol. Sci.*, 2015, **40**, 77–98.
- 2 A. R. Collins, The comet assay for DNA damage and repair, *Mol. Biotechnol.*, 2004, **26**, 249–261.
- 3 G. Krishna and M. Horisberger, *In vivo* rodent micronucleus assay: protocol, conduct and data interpretation, *Mutat. Res.*, 2000, **455**, 155–166.
- 4 M. A. Bender, R. J. Preston, R. C. Leonard, B. E. Pyatt, P. C. Gooch and M. D. Shelby, Chromosomal aberration and sister-chromatid exchange frequencies in peripheral blood lymphocytes of a large human population sample, *Mutat. Res.*, 1988, **204**, 421–433.
- 5 R. Kanode, S. Chandra and S. Sharma, Application of bacterial reverse mutation assay for detection of non-genotoxic carcinogens, *Toxicol. Mech. Methods*, 2017, **25**, 376–381.
- 6 D. F. Deen, L. E. Kendall, L. J. Marton and P. J. Tofton, Prediction of human tumor cell chemosensitivity using the sister chromatid exchange assay, *Cancer Res.*, 1986, **46**, 1599–1602.
- 7 S. H. Kang, J. Y. Kwon, J. K. Lee and Y. R. Seo, Recent Advances in *In Vivo* Genotoxicity Testing: Prediction of

- Carcinogenic Potential Using Comet and Micronucleus Assay in Animal Models, *J. Cancer Prev.*, 2013, **18**, 277–288.
- 8 P. Kamath, G. Raitano, A. Fernández, R. Rallo and E. Benfenati, In silico exploratory study using structure-activity relationship models and metabolic information for prediction of mutagenicity based on the Ames test and rodent micronucleus assay, *SAR QSAR Environ. Res.*, 2015, **26**, 1017–1031.
- 9 F. Cheng, W. Li, G. Liu and Y. Tang, In silico ADMET prediction: recent advances, current challenges and future trends, *Curr. Top. Med. Chem.*, 2013, **13**, 1273–1289.
- 10 C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee and Y. Tang, In silico prediction of chemical Ames mutagenicity, *J. Chem. Inf. Model.*, 2012, **52**, 2840–2847.
- 11 L. Zhang, H. Ai, W. Chen, Z. Yin, H. Hu, L. Zhu and L. Zhao, CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods, *Sci. Rep.*, 2017, **7**, 2118–2131.
- 12 J. Mohr, B. Jain, A. Sutter, A. T. Laak, T. Steger-Hartmann, N. Heinrich and K. Obermayer, A maximum common subgraph kernel method for predicting the chromosome aberration test, *J. Chem. Inf. Model.*, 2010, **50**, 1821–1838.
- 13 R. Benigni, C. Bossa and A. Worth, Structural analysis and predictive value of the rodent in vivo micronucleus assay results, *Mutagenesis*, 2010, **25**, 335–341.
- 14 D. Kirkland, E. Zeiger, F. Madia and R. Corvi, Can in vitro, mammalian cell genotoxicity test results be used to complement positive results in the Ames test and help predict carcinogenic or in vivo, genotoxic activity? II. Construction and analysis of a consolidated database, *Mutat. Res.*, 2014, **775–776**, 69–80.
- 15 C. W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.*, 2011, **32**, 1466–1474.
- 16 J. Dong, Z. Yao, M. Zhu, N. Wang, B. Lu, A. Chen, A. Lu, H. Miao, W. Zeng and D. Cao, ChemSAR: an online pipeline platform for molecular SAR modeling, *J. Cheminf.*, 2017, **9**, 1–13.
- 17 L. Sun, C. Zhang, Y. Chen, X. Li, S. Zhuang, W. Li, G. Liu, P. W. Lee and Y. Tang, In silico prediction of chemical aquatic toxicity with chemical category approaches and substructural alerts, *Toxicol. Res.*, 2015, **4**, 452–463.
- 18 W. S. Noble, What is a support vector machine?, *Nat. Biotechnol.*, 2006, **24**, 1565–1567.
- 19 H. Sun, A Naive Bayes Classifier for Prediction of Multidrug Resistance Reversal Activity on the Basis of Atom Typing, *J. Med. Chem.*, 2005, **48**, 4031–4039.
- 20 T. Denoeux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, *IEEE Trans. Syst. Man Cybern.*, 2008, **25**, 804–813.
- 21 S. L. Salzberg, C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. 1993, *Mach. Learn.*, 1994, **16**, 235–240.
- 22 V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. R. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, *J. Chem. Inf. Comput. Sci.*, 2016, **43**, 1947–1958.
- 23 S. Agatonovic-Kustrin and R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, *J. Pharm. Biomed. Anal.*, 2000, **22**, 717–727.
- 24 A. Abdiansah and R. Wardoyo, Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM, *Int. J. Comput. Appl.*, 2015, **128**, 975–8887.
- 25 F. Li, D. Fan, H. Wang, H. Yang, W. Li, Y. Tang and G. Liu, In silico prediction of pesticide aquatic toxicity with chemical category approaches, *Toxicol. Res.*, 2017, **6**, 831–842.
- 26 K. Roy, S. Kar and P. Ambure, On a simple approach for determining applicability domain of QSAR models, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22–29.
- 27 F. Cheng, Y. Yu, J. Shen, L. Yang, W. Li, G. Liu, P. W. Lee and Y. Tang, Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers, *J. Chem. Inf. Model.*, 2011, **51**, 996–1011.
- 28 J. A. Hanley and B. J. Mcneil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 1982, **143**, 29–36.
- 29 F. Pizzo, D. Gadaleta, A. Lombardo, O. Nicolotti and E. Benfenati, Identification of structural alerts for liver and kidney toxicity using repeated dose toxicity data, *Chem. Cent. J.*, 2015, **9**, 1–11.
- 30 C. Borgelt and M. R. Berthold, Mining Molecular Fragments: Finding Relevant Substructures of Molecules, *IEEE Int. Conf. Data Mining*, 2002, 51–58.
- 31 T. Ferrari, D. Cattaneo, G. Gini, N. G. Bakhtyari, A. Manganaro and E. Benfenati, Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction, *SAR QSAR Environ. Res.*, 2013, **24**, 365–383.
- 32 H. Yang, J. Li, Z. Rui, W. Li, G. Liu and Y. Tang, Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark, *Chem. Res. Toxicol.*, 2017, **30**, 1355–1364.
- 33 T. Ferrari, G. Gini, N. G. Bakhtyari and E. Benfenati, Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships, *Comput. Intell. Data Mining*, 2011, 120–127.
- 34 J. Shen, F. Cheng, Y. Xu, W. Li and Y. Tang, Estimation of ADME properties with substructure pattern recognition, *J. Chem. Inf. Model.*, 2010, **50**, 1034–1041.
- 35 D. Cao, Q. Xu, N. Hu and Y. Liang, ChemoPy: freely available python package for computational biology and chemoinformatics, *Bioinformatics*, 2013, **29**, 1092–1094.
- 36 H. Du, Z. Hu, A. Bazzoli and Y. Zhang, Prediction of Inhibitory Activity of Epidermal Growth Factor Receptor Inhibitors Using Grid Search-Projection Pursuit Regression Method, *PLoS One*, 2011, **6**, 1–8.
- 37 H. Li, C. Y. Ung, C. W. Yap, Y. Xue, Z. R. Li, Z. W. Cao and Y. Z. Chen, Prediction of genotoxicity of chemical compounds by statistical learning methods, *Chem. Res. Toxicol.*, 2005, **18**, 1071–1080.

- 38 T. Asadollahi, S. Dadfarnia, A. M. H. Shabani, J. B. Ghasemi and M. Sarkhosh, QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silico virtual screening, *Molecules*, 2011, **16**, 1928–1955.
- 39 L. C. Davidse, Benzimidazole Fungicides: Mechanism of Action and Biological Impact, *Annu. Rev. Phytopathol.*, 1986, **24**, 43–65.
- 40 L. Nelson, Acute Cyanide Toxicity: Mechanisms and Manifestations, *J. Emerg. Nurs.*, 2006, **32**, 8–11.
- 41 B. Chen, R. P. Sheridan, V. Hornak and J. H. Voigt, Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions, *J. Chem. Inf. Model.*, 2012, **52**, 792–803.
- 42 X. Li, L. Chen, F. Cheng, Z. Rui, H. Bian, C. Xu, W. Li, G. Liu, X. Shen and Y. Tang, In silico prediction of chemical acute oral toxicity using multi-classification methods, *J. Chem. Inf. Model.*, 2014, **54**, 1061–1069.