Applications of Machine Learning Methods in Drug Toxicity Prediction

Li Zhang\textsuperscript{a,\#}, Hui Zhang\textsuperscript{a,\#}, Haixin Ai\textsuperscript{a,b,c}, Huan Hu\textsuperscript{a}, Shimeng Li\textsuperscript{a}, Jian Zhao\textsuperscript{a}, and Hongsheng Liu\textsuperscript{a,b,c,*}

\textsuperscript{a}School of Life Science, Liaoning University, Shenyang, China; \textsuperscript{b}Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning Province, Liaoning University, Shenyang, China; \textsuperscript{c}Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning, Liaoning University, Shenyang, China

Abstract: Toxicity evaluation is an important part of the preclinical safety assessment of new drugs, which is directly related to human health and the fate of drugs. It is of importance to study how to evaluate drug toxicity accurately and economically. The traditional in vitro and in vivo toxicity tests are laborious, time-consuming, highly expensive, and even involve animal welfare issues. Computational methods developed for drug toxicity prediction can compensate for the shortcomings of traditional methods and have been considered useful in the early stages of drug development. Numerous drug toxicity prediction models have been developed using a variety of computational methods. With the advance of the theory of machine learning and molecular representation, more and more drug toxicity prediction models are developed using a variety of machine learning methods, such as support vector machine, random forest, naïve Bayesian, back propagation neural network. A variety of machine learning and molecular representation methods and have been considered useful in the early stages of drug development. Numerous drug toxicity prediction models have been developed using a variety of computational methods. With the advance of the theory of machine learning and molecular representation, more and more drug toxicity prediction models are developed using a variety of machine learning methods, such as support vector machine, random forest, naïve Bayesian, back propagation neural network. And significant advances have been made in many toxicity endpoints, such as carcinogenicity, mutagenicity, and hepatotoxicity.

Keywords: Drug toxicity prediction, Machine learning, QSAR, Molecular descriptors, Carcinogenicity prediction, Mutagenicity prediction, Hepatotoxicity prediction

1. INTRODUCTION

Drug discovery aims to provide therapeutic compounds that are safe and effective and that can relieve the pain and improve the quality of life of patients. Despite the current drug development technology has made great progress, it is estimated that in recent years, about 90.4% of compounds entering the clinical trial phase are reported to be failures, of which, 24% are due to lack of safety [1, 2]. Since the toxicity of drugs can cause great harm to the patient's body and even can cause a serious adverse social impact. The issue of drug safety has attracted the attention of the pharmaceutical administrations, pharmaceutical manufacturers and academia all over the world.

Traditional experimental toxicity testing usually requires the use of animal models. However, these animal tests are not only time-consuming but also costly, and even involve animal welfare issues. Computational methods have been shown to be effective in many aspects of drug discovery [3-6]. In recent decades, researchers and drug discovery companies have used various experimental models to determine various toxicological characteristics of a large number of organic compounds, making it possible to predict the toxicity of organic compounds using computational methods. Given the large-scale toxicological screening of compounds using computer models prior to conducting in vitro and in vivo tests, the probability of candidate compounds to be toxic could be effectively reduced. Drug toxicity prediction using computational methods only requires fewer efforts but can provide very useful information for the early stages of drug design. In particular, computational models based on physicochemical and structural properties of compounds can even predict compounds before their synthesis, greatly increasing their efficiency and making them more and more popular [7].

A variety of computational methods, such as machine learning, structural alerts, read-across, molecular modeling, have been used to develop drug toxicity prediction models [8]. With the continuous improvement of computing power and the rapid development of machine learning theory, more
and more researchers have applied machine learning method to the development of drug toxicity prediction model. A large number of outstanding drug toxicity prediction models based on machine learning have been reported in the literature. Therefore, in this review, we will focus on the studies that developed drug toxicity prediction models using machine learning methods.

Machine learning methods build classification or regression models to describe the complex relationships between the chemical structure of drug molecules and their toxicity endpoints based on the knowledge obtained from experimental data. In general, the process that develops drug toxicity prediction models based on machine learning methods includes the following five major steps: (1) collecting biological data that contains the structure of chemicals and the experimental measurements of certain toxicity endpoint; (2) calculating molecular descriptors, molecular fingerprints or other molecular representations; (3) model building; (4) evaluation and verification of the model [9-11]. The main content of this review includes the first three steps. The main topics are as follows: introducing publicly available data sets for developing drug toxicity prediction models, summarizing the recent progress of molecular representation methods, introducing machine learning methods that are commonly applied to drug toxicity prediction, and introducing some representative studies that developed toxicity prediction models using machine learning methods.

2. PUBLICLY AVAILABLE DATA SETS

Machine learning methods mine the relationships between drug structures and their toxic endpoints from known experimental data. Therefore, the first step in developing machine learning-based drug toxicity prediction models is to obtain high-quality experimental data. Many studies have collated data on drug toxicity from a variety of data sources and established databases for researchers. Table 1 lists publicly available datasets for drug toxicity prediction.

<table>
<thead>
<tr>
<th>Database Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOXNET</td>
<td><a href="https://toxnet.nlm.nih.gov/">https://toxnet.nlm.nih.gov/</a></td>
</tr>
<tr>
<td>ToxCast</td>
<td><a href="https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data">https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data</a></td>
</tr>
<tr>
<td>admetSAR</td>
<td><a href="http://lnmd.ecust.edu.cn/admetsar1/">http://lnmd.ecust.edu.cn/admetsar1/</a></td>
</tr>
<tr>
<td>Ames Mutagenicity Benchmark Data Set</td>
<td><a href="http://doc.ml.tu-berlin.de/toxbenchmark/">http://doc.ml.tu-berlin.de/toxbenchmark/</a></td>
</tr>
<tr>
<td>ChEMBL</td>
<td><a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a></td>
</tr>
<tr>
<td>BindingDB</td>
<td><a href="http://www.bindingdb.org/bind/index.jsp">http://www.bindingdb.org/bind/index.jsp</a></td>
</tr>
<tr>
<td>T3DB</td>
<td><a href="http://www.t3db.ca/">http://www.t3db.ca/</a></td>
</tr>
</tbody>
</table>

3.2. Molecular Fingerprints

Molecular fingerprints is another commonly used molecular representation method, also known as molecular descriptors in some literature. The most common type of molecular fingerprints is “Structural fingerprints”, which encode the molecular structure information into binary strings (strings of 0 and 1). The position of each digit in this string corresponds to a specific chemical structure fragment. If the molecule has a specific fragment, the corresponding position is set to 1, otherwise to 0. There are many different ways to design a molecular fingerprint, depending on which fragments are included in the fingerprint definition. Different researchers have designed many kinds of molecular fingerprints according to different needs. The most commonly used molecular fingerprints include PubChem fingerprint, MACCS fingerprint, Klekota-Roth fingerprint [23], Estate fingerprint [24] and so on.

Another major type of commonly used molecular fingerprint is “hash fingerprints”, in which, no pre-defined structural fragment was used to encode the molecule. Instead,
hash fingerprints generate patterns from the molecule itself, by enumerating all the linear substructures of the molecule within a certain length range and use a hash function to turn the values of the substructures to a set of bits and then add the bits to the fingerprints. Hence, hash fingerprints have advantages over structural fingerprints. Representative hash fingerprints include Daylight fingerprints, CDK fingerprints [25], GraphOnly fingerprints, RDK fingerprints.

Extended Connectivity Fingerprint (ECFP) is a new type of molecular representation method [26], which was specifically developed for structure-activity relationship modeling. ECFP is a refinement of Morgan algorithm [27, 28], and has a number of useful qualities, that are not pre-defined, and can represent any number of different molecular features with easier interpretation.

3.3. Graph Convolutions

The latest research has developed a type of novel molecular representation method called graph convolutions [29, 30]. In graph convolutions, molecules are treated as undirected graphs, and the molecular fingerprints with unfixed-length were generated using deep neural networks in the process of building predictive models. Graph convolutions have achieved higher performance in many test datasets [29] and an acute oral toxicity prediction model [31]. The graph convolution fingerprints offer several advantages, such as higher predictive performance, shorter fingerprint length, and higher interpretability. Drug toxicity prediction models with higher performance could be developed using more advanced molecular representation method.

4. MACHINE LEARNING ALGORITHMS

After molecular representation has been completed, a model describing the relationship between molecules and toxicological properties can be established using machine learning algorithms. When building a machine learning model, each molecular descriptor and each bit of the fingerprints will serve as an independent variable or called a feature in the scope of machine learning. Some useful programming tools, such as R, Weka, Python, and some useful QSAR modeling software, such as KNIME, RDKit, provide implementations of the machine learning algorithms that are widely used to model drug toxicity prediction models. Here we briefly introduce the basic principles of the following several commonly used machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), Naive Bayesian (NB), Neural Network (NN) and Ensemble Learning (EL).

4.1. Support Vector Machine

SVM (Fig. 1) is an effective supervised machine learning method based on the principle of structural risk minimization. The Linear SVM constructing an optimal hyperplane separates the two categories of eigenvectors at maximum intervals. This hyperplane is constructed by seeking a vector \( \mathbf{w} \) and a parameter \( b \) that minimize \( \| \mathbf{w} \| \) and satisfy \( \mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \) (positive class) and \( \mathbf{w} \cdot \mathbf{x}_i - b \leq 1 \) (negative class), where \( \mathbf{x}_i \) is the eigenvector, \( \mathbf{w} \) is the vector vertical to the hyperplane. \( \frac{b}{||w||} \) is the vertical distance from the hyperplane to the origin, \( ||w|| \) is an euclidean norm of \( \mathbf{w} \) [32]. Non-linear svm was used a kernel function (such as gaussian radial basis function, homogeneous polynomial function) to project the eigenvector into high dimensional space in order to find the hyperplane. After determining the value of \( \mathbf{w} \) and \( b \), for a given feature vector \( \mathbf{x} \), we could use \( f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{n} a_i^p y_i k(\mathbf{x}, \mathbf{x}_i) + b) \) for classification, where the coefficients \( a_i^p \) and \( b \) were determined by maximizing the lagrange function \( \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \) under the condition of \( a_i \geq 0, \sum_{i=1}^{n} a_i y_i = 0 \). The positive and negative of the results of \( f(\mathbf{x}) \) represent eigenvectors \( \mathbf{x} \) that belongs to the positive class or the negative class. SVM had a good learning performance in case of small samples, and was applied to the datasets with the nonlinear problem, and the obtained model had better generalization performance. Due to these advantages, SVM had been widely used in bioinformatics studies and computer-aided drug design [19].

4.2. Random Forest

RF (Fig. 2) is a non-linear multi-class machine learning method. RF uses bootstrap sampling to extract a number of training samples and subsets of features from the original training set, establishes a plurality of un-pruned decision trees, and then combines the decision trees to form a random forest model. RF introduces randomness to each generated decision tree at two levels. First, a certain proportion of training samples was randomly selected from the total sample set for each of the decision trees. And, when the node of the decision tree is bifurcated, the feature subset is selected randomly from the feature set, and then the optimal feature is selected from the feature subset for bifurcation. As a result, the randomness increases the diversity of decision trees and makes the resulting integrated model have better predictive performance. One of the main advantages of RF is that it can handle a large number of features without overfitting and is
insensitive to data with missing values and noise. RF can estimate the importance of features, which further expand its scope of application [33].

4.3. K-Nearest Neighbor

The core idea of kNN (Fig. 3) algorithm is that if the majority of the k nearest neighbor samples in a feature space belong to a certain category, the sample to be predicted also has the similar characteristics of the sample in this category and may belong to this category. The method determines the sample to be classified according to the category of k nearest samples in the training sample sets. Because the kNN method mainly depends on the limited neighboring samples rather than the discriminant domain, the kNN method is more efficient than other methods for the cross-over or overlapping sample set.

4.4. Naive Bayesian (NB)

Naive Bayesian is a simple probabilistic classifier based on Bayesian theorem with the assumption that the conditional independence components of the feature vector are relatively independent of the decision variables. The whole naive Bayesian classification can be divided into three stages as follows (Fig. 4): First stage - preparation stage, the task at this stage is to prepare the naive Bayesian classification and to make necessary preparations for naive Bayesian classification. The input of this stage is all the data to be classified, while the output is the characteristic attribute and training sample. The second stage - the classifier training stage: the task of this stage is to generate the classifier, the main job is to calculate the frequency of occurrence of each category in the training sample and the classification of each feature attribute for each category of conditional probability estimates. The third stage - application stage: the task of this stage is to use the classifier to classify the classification items.

4.5. Neural Network

Artificial neural networks, also known as neural networks, are computational models that mimic the structure and function of biological neural networks. A neural network generally consists of three layers of artificial neurons: the input layer, the hidden layer, and the output layer, which are

![Image](image_url)
interconnected by modifiable weights (Fig. 5). The network learning process consists of two processes: the forward propagation of the signal and the reverse propagation of the error. When propagating forward, the pattern acts on the input layer, through the hidden layer to the output layer. If the output layer fails to obtain the desired output, it goes into the reverse propagation phase of the error, through the hidden layer to the input layer. So as to obtain the reference error of each layer unit, as a basis for modifying the weights of each unit, this process is continued until the error of the network output is gradually reduced to an acceptable level or until the set number of learning times has been reached.

Fig. (4). Diagram representing the basic principles of Naive Bayesian.

Fig. (5). Diagram representing the basic principles of neural network.

4.6. Ensemble Learning

Ensemble learning models can be built by voting or averaging a series of simple basic classifiers (Fig. 6). Ensemble learning methods often produce models that are more accurate than their constituent models. Dietterich [34] pointed out that the accuracy of the integrated learning model will exceed that of all base classifiers when the base classifier is accurate (the error rate is lower than the random guess) and the base classifiers have significant diversity (the prediction results are not the same). Supposing that there exists an ensemble model that consists of three classifiers \( \{h_1, h_2, h_3\} \) and uses this integrated model to predict a new sample \( x \). If the prediction results of the three classification models are uncorrelated, when \( h_1(x) \) is wrong, \( h_2(x) \) and \( h_3(x) \) may be correct, so by voting, \( x \) can be correctly predicted. To be more precise, if the error rates of L classifiers \( \{h_1, h_2, ..., h_L\} \) are all smaller than 1/2, and the errors of the classifiers are all independent, the probability that these models will vote incorrectly will be equivalent to the probability that more than L/2 classifiers are wrong. Researchers have developed a variety of highly efficient machine learning algorithms based on ensemble learning theory, such as Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Extreme gradient boosting (XGBOOST). The ensemble learning methods are shown to be effective in recent studies [35-37].

Fig. (6). Diagram representing the basic principles of ensemble learning.

5. DRUG TOXICITY PREDICTION

There are many applications of these machine learning methods in drug toxicity prediction, including various toxicity endpoints, such as carcinogenicity, mutagenicity, hepatotoxicity, acute oral toxicity, and h\(_{\text{ERG}}\) (human ether-a-go-go-related gene) inhibition.

5.1. Carcinogenicity

Carcinogenic compounds are chemicals that can cause tumors or increase the incidence of tumors. These compounds are widely found in the environment in which we depend on, and they are one of the decisive factors leading to cancer. The carcinogenicity of compounds is generally measured using animal experiments, such as two-year rodent experiments and 26-week Tg-rash2 mouse carcinogenicity experiments. However, in the early stages of drug development, it is not possible to evaluate a large number of new compounds to determine if they are carcinogenic. Therefore, using computer methods to predict the carcinogenicity before experimental validation has become the focus of research in recent years [8].

Many carcinogenicity prediction models only target specific compound classes (homologous compounds such as amines, nitro compounds [38], aromatic amines [39], polycyclic aromatic hydrocarbons [40], and polychlorinated biphenyls [41]) and have achieved very high accuracy. However, these models can only be applied to specific homologous compounds, that is to say, these models have very lim-
In recent years, researchers have developed models that can be used to predict the carcinogenicity of non-homologous compounds based on larger databases. The training datasets for these models include different types of compounds, with more complex and diverse structures. The covered chemical structure is more extensive and can, therefore, be used to predict the carcinogenicity of more compounds. Using the Lois Gold Carcinogenic Potency Database (CPDB), which contains findings of 2-year rodent carcinogenicity study for 1481 chemicals with diverse chemical structures, as the training dataset, Helma et al. developed a rodent carcinogenicity prediction model (named lazer) for predicting carcinogenicity of diverse chemicals using a modified k-nearest-neighbour (knn) algorithm [42]. The accuracy of Lazer in predicting of rodent carcinogenicity is 86% in leave-one-out cross-validation LOOCV for structures within the applicability domain. Using the same database, and the twenty-seven two-dimensional MDL descriptors as molecular representation, Fjodorova et al. developed counter propagation artificial neural network (CP ANN) models with an overall accuracy of 68% in a test set [43]. In order to make models in accordance with the principles of the European Commission (EC) funded project CAESAR (Computer Assisted Evaluation of industrial chemical Substances According to Regulation), Fjodorova et al. [44] developed two new public carcinogenicity prediction models using eight MDL descriptors and twelve Dragon descriptors based on Counter Propagation Artificial Neural Network (CP ANN) algorithm. Moreover, the accuracy of the two models was 66% and 62%, respectively, when using the evaluation method of five-fold cross-validation. Probabilistic Neural Network (PNN) and Generalized Regression Neural Network (GRNN) models were also used to construct carcinogenicity prediction models [45].

Support vector machine is the algorithm used by many carcinogenicity prediction models [46-48]. Tanabe et al. [46] used the substructure grouping method to divide the compound database into a series of mutually overlapping subgroups and trained a specific optimized SVM model for each subgroup, and the predicted results are voted by these models. The resulted model predicts diverse chemicals in the compound database with an overall accuracy of approximately 80%. In 2013, Tanabe [47] et al. applied a new Sensitivity Analysis (SA) method, which selects the optimum sets of effective molecular descriptors, to support vector machine algorithm, and obtained carcinogenicity prediction models with higher performance. But the predictive performances of the models developed by Tanabe et al. were not validated by any of the cross-validation methods. Naive Bayes algorithm combined with ECFP fingerprints was also used to develop classification models for predicting the carcinogenicity of diverse chemicals [49], and the model gave an overall prediction accuracy of 68% in five-fold cross-validation.

Zhang et al. [36] developed a carcinogenicity prediction model with higher performance using ensemble learning method. The ensemble model was built by fusing a series of machine learning models built by support vector machine, random forest, and extreme gradient boosting algorithms and 12 types of molecular fingerprints. The ensemble model was evaluated using five-fold cross-validation and external validation, and the overall accuracy was 70.1% and 70.0% respectively. And these models have been integrated into a web server http://cspib.lnu.edu.cn/toxicity/CarcinPred-EL/. The investigation of chemical groups that are related to carcinogenicity by means of machine learning algorithms (such as RF [36] and NB [49]) also provides valuable information about rational drug design.

Recently, organ-specific chemical carcinogenicity prediction model was developed by Lagunin et al. [50] using a bayesian-like approach and MNA (Multilevel Neighborhoods of Atoms) descriptors. This QSAR model can predict whether the compound can cause cancer in specific organs. More detailed predictions could be an important future direction of the development of drug toxicity prediction model.

As a summary, Table 2 lists the performances of some machine learning-based drug carcinogenicity prediction models published in the literature in recent years. Since it is difficult to accurately estimate the prediction performance of the model without using cross-validation [51-53], we only list the cross-validated models.

### Table 2. The performances of machine learning-based drug carcinogenicity prediction models (only the cross-validated models). Q: Accuracy; SE: Sensitivity; SP: Specificity.

<table>
<thead>
<tr>
<th>Name</th>
<th>Q(%)</th>
<th>SE(%)</th>
<th>SP(%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC4PC</td>
<td>66.5</td>
<td>61.4</td>
<td>70.9</td>
<td>[54]</td>
</tr>
<tr>
<td>MDL-QSAR</td>
<td>69.2</td>
<td>62.8</td>
<td>74.8</td>
<td>[54]</td>
</tr>
<tr>
<td>lazer</td>
<td>66.9</td>
<td>59.9</td>
<td>73.4</td>
<td>[55]</td>
</tr>
<tr>
<td>CP ANN MDL</td>
<td>66</td>
<td>-</td>
<td>-</td>
<td>[56]</td>
</tr>
<tr>
<td>CP ANN Dragon</td>
<td>62</td>
<td>-</td>
<td>-</td>
<td>[56]</td>
</tr>
<tr>
<td>Naïve Bayesian</td>
<td>68</td>
<td>57</td>
<td>79</td>
<td>[49]</td>
</tr>
<tr>
<td>Ensemble XGBoost</td>
<td>70.1</td>
<td>67.0</td>
<td>73.1</td>
<td>[36]</td>
</tr>
</tbody>
</table>

5.2 Mutagenicity

Mutagens are physical processes or chemical reagents that alter the body's genetic material (DNA), making the organism's mutation rate higher than the natural background level. Many mutagens can cause cancer. Therefore, the assessment of the mutagenicity of compounds is an important part of toxicity evaluation in the process of drug design. The mutagenicity of chemicals can be detected using the Ames test designed by Bruce Ames. The Ames test was designed based on the fact that the exposure of Salmonella typhimurium and Escherichia coli auxotrophic strains (for S.
typhimurium is histidine and for E. coli is tryptophan) to mutagenic compounds can restore the synthesis of essential amino acids. Therefore, this experiment is also called a bacterial reverse mutation test. It is estimated that the repetition rate of Ames test between laboratories is about 80-85%. Therefore, it can be considered that the upper limit of the accuracy of the Ames mutagenicity prediction models should also be in this interval.

The Ames mutagenicity benchmark dataset developed by Hansen et al. is the most commonly used training dataset for the development of predictive compound mutagenicity models [57]. This data set was established in 2009. It comprehensively collected the result of Ames testing from CCRIS, GeneTox, VITIC databases and three papers [58-60]. The benchmark dataset contains 6512 compounds, of which, 2503 are positive, i.e. mutagenic compounds, and 3009 are negative, i.e. non-mutagenic compounds.

Many studies have applied machine learning algorithms to predict compound mutagenicity [61-65]. For example, Xu et al. [62] collected data from a number of papers and formed a data set containing 7,617 compounds. And then, using this as a training set, machine learning algorithms such as support vector machines, k-nearest neighbors, naive Bayes, artificial neural networks and decision trees are used to establish a series of mutagenicity prediction models. For example, The accuracy of the best five models in the five-fold cross-validation ranged from 80.8% to 84.1%, and the accuracy in the external test set reached 90.4% to 98.0%. Although the accuracy of the model in this study is very high, the accuracy of the estimated model may be biased due to the fact that it does not repeat the five-fold cross-validation for building a large number of models. Recent studies have conducted more rigorous assessments. For example, Zhang et al. established a mutagenicity prediction model using naive Bayes algorithm combined with ECFP fingerprints [63]. The accuracy of this model reached 77.3% in five-fold cross-validation and 90.9% in predicting 446 drugs that are already in the market. Table 3 lists the performance of some machine learning-based drug mutagenicity prediction models published in the literature in recent years.

Researchers have conducted a comprehensive assessment of mutagenicity prediction models reported in the literature. In 2011, Hillebrecht et al. [66] tested the predictive performance of four commonly used mutagenicity prediction software (Derek [67], Toxtree [68], MC4PC, and Leadscope MA) by collecting a dataset containing public mutagenicity data and Roche's proprietary data. The results show that the accuracy of these models is between 66.4-75.4% for public data and 73.1-85.5% for data from Roche, but its sensitivity is very low (17.4-43.4%) and specificity is high (77.5-93.9%). These machine-learning models have high specificity and low sensitivity. Bakhtyari et al. [69] and Yang et al. [70] also compared the performance of many mutagenic prediction models. It can be seen that the predictive power of machine learning models still needs to be improved. Although the current mutagenicity prediction models have obtained relatively high accuracy rate, there is still a gap between the theoretical maximum value, 85%, and more effective prediction tools need to be developed.

Table 3. The performances of machine learning-based drug mutagenicity prediction models (only the cross-validated models). Q: Accuracy; SE: Sensitivity; SP: Specificity.

<table>
<thead>
<tr>
<th>Name</th>
<th>Q(%)</th>
<th>SE(%)</th>
<th>SQ(%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC4PC</td>
<td>79.0</td>
<td>79.8</td>
<td>78.1</td>
<td>[71]</td>
</tr>
<tr>
<td>DEREK</td>
<td>59.1</td>
<td>80.9</td>
<td>59.1</td>
<td>[66]</td>
</tr>
<tr>
<td>LSMA</td>
<td>66.4</td>
<td>67.8</td>
<td>63.8</td>
<td>[66]</td>
</tr>
<tr>
<td>Toxtree</td>
<td>74.6</td>
<td>85.2</td>
<td>53.1</td>
<td>[66]</td>
</tr>
<tr>
<td>SciQSAR</td>
<td>77</td>
<td>83</td>
<td>71</td>
<td>[72]</td>
</tr>
</tbody>
</table>

5.3. Hepatotoxicity

Drug-Induced Liver Injury (DILI) is one of the main causes of drug development failure and withdrawal [73, 74]. At present, various models have been reported for predicting hepatotoxicity of chemicals, most of which are developed by machine learning methods [75-83]. For example, Ekins et al. [75] developed a hepatotoxicity prediction model based on the Bayesian approach using ECFP molecular fingerprinting. The model was trained on a training set of 295 compounds and tested on a test set containing 237 compounds. The model has an accuracy of 60% in the test set. The Derek software developed by Greene et al. [76] achieved an accuracy of 55.5% in a test set containing 466 compounds provided by Pfizer, with a sensitivity of 46.5% and a specificity of 72.9%. Fourches et al. [77] developed a hepatotoxicity prediction model based on support vector machine. In the five-fold cross-validation, the accuracy of this model was 63.9%. Liew et al. [84] developed an ensemble model that integrates 617 base classifiers based on the support vector machine and k-nearest neighbor algorithm using a dataset containing 1087 compounds. The accuracy of this model is 63.8% in the five-fold cross-validation and 62.2% in an external verification set containing 120 compounds. Chen et al. [79] developed a hepatotoxicity prediction model based on the decision forest algorithm [85]. This model has achieved an accuracy of 69.7%, a sensitivity of 57.8%, and a specificity of 77.9% in 10-fold cross-validation. Zhang et al. [86] calculated three types of molecular fingerprints for 1229 compounds and obtained a series of molecular fingerprint subsets using different information gain thresholds. Using these molecular fingerprint subsets as features, a series of hepatotoxicity prediction models was established using support vector machine algorithm. The accuracy of the best model was 75% in the test set and 64.5% in the external test set. However, the specificity of these models was very low (about 35%). Zhang et al. [83] developed a hepatotoxicity prediction model using Naive Bayes algorithm. In the test set, the accuracy was 72.6%, sensitivity was 72.5%, and specificity was 72.7%. Table 4 lists the performance of some machine learning-based drug hepatotoxicity prediction models published in the literature in recent years. It can be seen that the accuracy of the hepatotoxicity prediction model is still unsatisfactory.
5.4. Acute Oral Toxicity

Another important toxic endpoint that is of great concern in drug design is acute toxicity, which describes the adverse effects occurring in a relatively short period of time (usually less than 24 hours) after administration of a drug [89]. Of all acute toxicities, acute oral toxicity is the most studied. Acute oral toxicity is generally determined through animal tests, and the results were expressed as the median lethal dose (LD$_{50}$). According to the median lethal dose, chemicals can be divided into four categories, namely categories I, II, III and IV, based on the classification criterion of the U.S. Environmental Protection Agency (EPA).

Currently, many QSAR models have been developed for acute oral toxicity prediction using various machine learning algorithms, such as k-nearest neighbor [90-92], decision tree [93], naive Bayes [93], support vector machine [92, 93], neural network [31, 93-95], random forest [90, 92, 93], and extreme gradient boosting [92]. These models can be roughly divided into two categories: regression models and classification models. For the regression model, the model predicts the LD$_{50}$ value of chemicals. For the classification model, the model predicts which of the four categories the compound belongs to. Here we briefly introduced several outstanding studies of recent years. Li et al. [93] developed several multiclassification models for acute oral toxicity prediction using five machine learning methods based on a data set containing 12204 diverse compounds with LD$_{50}$ values. The overall accuracy of the best model was 83.0% and 89.9% on two test sets. Lei et al. [92] employed relevance vector machine and other methods to build consensus models (ensemble models) for acute oral toxicity prediction. The consensus models yielded better performance ($R^2 = 0.669-0.689$) in the test set. Xu et al. [31] developed a regression model (deepAOT-R) and a multiclassification model (deepAOT-C) for acute oral toxicity prediction using an improved Molecular Graph Encoding Convolutional Neural Networks (MGE-CNN) architecture. The $R^2$ and Mean Absolute Error (MAE) of the regression model were 0.864 and 0.195, and the accuracy of the multiclassification model was 95.5% on a test dataset containing 1673 compounds. And these models have been integrated into a web server http://www.pkumdl.cn/DLAOT/DLAOThome.php.

### Table 4. The performances of machine learning-based drug hepatotoxicity prediction models (only the cross-validated models). Q: accuracy; SE: sensitivity; SP: specificity.

<table>
<thead>
<tr>
<th>Name</th>
<th>Q (%)</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>58.5</td>
<td>52.8</td>
<td>65.5</td>
<td>[87]</td>
</tr>
<tr>
<td>SVM-Dragon</td>
<td>63.9</td>
<td>-</td>
<td>-</td>
<td>[77]</td>
</tr>
<tr>
<td>Ensemble (617 base classifiers)</td>
<td>63.8</td>
<td>64.1</td>
<td>63.3</td>
<td>[84]</td>
</tr>
<tr>
<td>Decision Forest</td>
<td>69.7</td>
<td>57.8</td>
<td>77.9</td>
<td>[88]</td>
</tr>
</tbody>
</table>

5.5. hERG (Human Ether-a-go-go-related Gene) Inhibition

The hERG is a potassium ion channel protein. When this protein is inhibited, it can lead to a potentially fatal disease - QT prolongation syndrome [96]. The early assessment of hERG inhibition potency is essential to confidently advance novel drug candidates [97]. In recent years, many QSAR models have been developed for hERG inhibition prediction [98-103]. For example, Zhang et al. [102] collected a hERG blockage database containing 1570 compounds and developed several classification models for hERG inhibition prediction using five machine learning methods and molecular descriptors combining fingerprints. This study analyzes in detail the effects of different combinations of machine learning algorithms, molecular descriptors and molecular fingerprints on the accuracy of model prediction under different discrimination thresholds. Wang et al. [103] have built a more reliable hREG inhibition model by combining pharmacophore modeling and SVM. The best model achieved the prediction accuracies of 82.1% in an external test set.

**CONCLUSION AND PERSPECTIVES**

In summary, many machine learning models have been developed for effectively predicting various drug toxicity endpoints. In this review, we mainly focused on the carcinogenicity, mutagenicity, hepatotoxicity, acute oral toxicity, and hERG inhibition. SVM, kNN, and NN are the most commonly used algorithms, as their theory are mature and they are easy to implement. Some new algorithms have also been applied to drug toxicity predictions such as deep neural networks and ensemble learning. The use of machine learning methods has further improved the accuracy of drug toxicity predictions. Using machine learning methods to establish toxicity prediction models has become the first choice for most studies.

Although the application of various machine learning methods in recent years has made significant advances in the prediction of drug toxicity, there are still many challenges that need to be addressed. The reliability that the machine learning model can be achieved depending mainly on the reliability and diversity of the training data. Establishing benchmark datasets for all types of toxic endpoints will be a very important task. Currently, most machine learning-based drug toxicity prediction models have been developed based on combining many molecular descriptors or molecular fingerprints, which make the models become “black box” and difficult to interpret and difficult to be understood by researchers in the field of drug design. Therefore, more interpretive molecular features and effective feature selection algorithms should be used to refine the features used in the model. Compared with other molecular representation methods, graph convolutions have many practical advantages, and their effective use can help to develop better models for drug toxicity prediction.

**CONSENT FOR PUBLICATION**

Not applicable.
CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (no. 31570160), the Innovation Team Project (no. LT2015011) of the Education Department of Liaoning Province, the Important Scientific and Technical Achievements Transformation Project (no. Z17-5-078), the Large-scale Equipment Shared Services Project (no. F15165400) of the Science and Technology Bureau of Shenyang, the Applied Basic Research Project (nos F16205151, 17231104) of the Science and Technology Bureau of Shenyang. This work was supported by the Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning.

REFERENCES

machine learning techniques for the identification of mutagenicity
scriptors and statistical methods.

ta set for


Chen, X.; Huang, L. LRSSLMDA: Laplacian Regularized Sparse
Biol

specific carcinogenicity prediction.
Poroikov, V.; Wren, J. ROSC

Zhong, M.; Nie, X.; Yan, A.; Yuan, Q. Carcinogenicity prediction
based on substructure grouping and SVM modeling.

č
ko, M.; Novi

ANN method for regulatory uses.
Kuhne, R.; Schuurmann, G. Quantitative and quali
activity relationship (QSAR).

phenyls (PCBs) and their metabolites using quantitative structure

Zhong, M.; Nie, X.; Yan, A.; Yuan, Q. Carcinogenicity prediction

Zhang, H.; Cao, Z.X.; Li, M.; Li, Y.Z.; Peng, C. Novel naïve Bayes
classification models for predicting the carcinogenicity of chemi-

Logamin, A.; Rudik, A.; Druzhilovsky, D.; Filimonov, D.;
Porokov, V.; Wren, J. ROSC-Pred: web-service for rodent organ-

Liu, H.; Ren, G.; Hu, H.; Zhang, L.; Ai, H.; Zhang, W.; Zhao, Q. LPI-NRLMF. InCrNA-protein interaction prediction by neighbor-

Chen, X. PBMDA: A novel and effective path-based computational

Chen, X.; Huang, L. LRSSLMDA: Laplacian Regularized Sparse
Subspace Learning for MiRNA-Disease Association prediction.

Contrera, J.F.; Kruhlak, N.L.; Matthews, E.J.; Benz, R.D. Compari-
sion of MC4PC and MDL-QSR rodent carcinogenicity predictions
and the enhancement of predictive performance by combining

Helma, C. Lazy structure-activity relationships (lazar) for the prediction
of rodent carcinogenicity and Salmonella mutagenicity.

Fjodorova, N.; Vračko, M.; Novič, M.; Roncaglioni, A.; Benfenati, E. New public QSR model for carcinogenicity. Chem. Central J.,
2010, 4(Supp 1), 1.

Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; Ter Laak, A.;
Steger-Hartmann, T.; Heinrich, N.; Mu Iler, K.-R. Benchmark da-
ta set for in silico prediction of Ames test mutagenicity. J. Chem. In-

Young, S. S. Predictive toxicology: Benchmarking molecular de-
2003, 43, 1463-1470.

Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and
machine learning techniques for the identification of mutagenicity
inducing substrates and structure activity relationships of non-
1402-1411. 

Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of
 toxicophores for mutagenicity prediction. J. Med. Chem., 2005, 48,
312-320.

Modi, S.; Li, J.; Malcomber, S.; Moore, C.; Scott, A.; White, A.;
Carmichael, P. Integrated in silico approaches for the prediction of
1017-1033.

Tang, Y. In silico prediction of chemical Ames mutagenity. J. Chem.

Zhang, H.; Kang, Y.-L.; Zhao, Y.-Y.; Zhao, K.-X.; Liang, J.-Y.;
Ding, L.; Zhang, T.-G.; Zhu, J. Novel naïve Bayes classification
models for predicting the chemical Ames mutagenicity. Toxicolo-

Webb, S. J.; Hanser, T.; Howlin, B.; Krause, P.; Vessey, J.D. Fea-
ture combination networks for the interpretation of statistical ma-

Fan, D.; Yang, H.; Li, F.; Sun, L.; Di, P.; Li, W.; Tang, Y.; Liu, G.
In silico prediction of chemical genotoxicity using machine learn-

Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.;
Singer, T. Comparative evaluation of in silico systems for Ames test
mutagenicity prediction: Scope and limitations. Chem. Res. Toxic-
ol., 2011, 24, 843-854.

Greene, N.; Judson, P.N.; Langowski, J.J.; Marchant, C.A.
Knowledge-based expert systems for toxicity and metabolism predic-
tion: DEREK, STAR and METEOR. SAR and QSAR Environ. Res.,
1999, 10, 299-314.

Benigri, R.; Bossa, C. Structure alerts for carcinogenicity, and the
Salmonella assay system: A novel insight through the chemical re-

Bakhtyari, N.G.; Raitano, G.; Benfenati, E.; Martin, T.; Young, D.
Comparison of in silico models for prediction of mutagenicity. J.

Yang, H.; Li, J.; Wu, Z.; Li, W.; Liu, G.; Tang, Y. Evaluation of
different methods for identification of structural alerts using chemi-
cal Ames mutagenicity data set as a benchmark. Chem. Res. Toxic-
ol., 2017, 30, 1355-1364.

Saiakhov, R.D.; Klopman, G. Benchmark performance of Multi-
2010, 50, 1521-1521.

Contrera, J.F. Validation of Toxtree and SciQSAR in silico predic-
tive software using a publicly available benchmark mutagenicity
database and their applicability for the qualification of impurities in

Segall, M.; Barber, C. Addressing toxicity risk when designing and

Chen, Q.; Zheng, B.; Shao, J.; Ai, N.; Wu, L.; Liu, Y.; Lu, X.; Fan,
X. Network-based assessment on chemical-induced cholestatic liver

Ekins, S.; Williams, A.J.; Xu, J.J. A predictive ligand-based Bayes-
ian model for human drug-induced liver injury. Drug Metab.
Dispos., 2010, 38, 2302-2308.

Greene, N.; Fisk, L.; Naven, R.T.; Note, R.R.; Patel, M.L.; Pelle-
tier, D.J. Developing structure-activity relationships for the predic-

Foucheux, D.; Barnes, J.C.; Day, N.C.; Bradley, P.; Reed, J.Z.;
Tropsch, A. Cheminformatics analysis of assertions mined from lit-
erature that describe drug-induced liver injury in different species.

Liew, C.Y.; Lim, Y.C.; Yap, C. In silico prediction of chemical

Zhang, C.; Chen, J.; Sechler, S.; Peng, C.; Ya, M. Development and
validation of a novel method for predicting chemical genotoxicity.

Tong, W. Quantitative structure-activity relationship models for
predicting drug-induced liver injury based on FDA-approved drug
tablet annotation data using a large collection of drugs. Toxicol.

Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P.W.; Yang, H.; Lang,
J.; Wei, C. In silico prediction of drug induced liver toxicity using
Applications of Machine Learning Methods in Drug Toxicity Prediction


DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.