Predicting drug-induced liver injury in human with Naïve Bayes classifier approach

Hui Zhang[1,2] · Lan Ding[1] · Yi Zou[1] · Shui-Qing Hu[1] · Hai-Guo Huang[1] · Wei-Bao Kong[1,3] · Ji Zhang[1,3]

Received: 28 May 2016 / Accepted: 12 September 2016 / Published online: 17 September 2016

Abstract Drug-induced liver injury (DILI) is one of the major safety concerns in drug development. Although various toxicological studies assessing DILI risk have been developed, these methods were not sufficient in predicting DILI in humans. Thus, developing new tools and approaches to better predict DILI risk in humans has become an important and urgent task. In this study, we aimed to develop a computational model for assessment of the DILI risk with using a larger scale human dataset and Naïve Bayes classifier. The established Naïve Bayes prediction model was evaluated by 5-fold cross validation and an external test set. For the training set, the overall prediction accuracy of the 5-fold cross validation was 94.0 %. The sensitivity, specificity, positive predictive value and negative predictive value were 97.1, 89.2, 93.5 and 95.1 %, respectively. The test set with the concordance of 72.6 %, sensitivity of 72.5 %, specificity of 72.7 %, positive predictive value of 80.4 %, negative predictive value of 63.2 %. Furthermore, some important molecular descriptors related to DILI risk and some toxic/non-toxic fragments were identified. Thus, we hope the prediction model established here would be employed for the assessment of human DILI risk, and the obtained molecular descriptors and substructures should be taken into consideration in the design of new candidate compounds to help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

Keywords Drug-induced liver injury · *In silico* prediction · Naïve Bayes classifier · Molecular descriptors · Extended connectivity fingerprints (ECFP_6)

Introduction

Drug-induced liver injury (DILI) contributes to about 5–10 % adverse drug events [1] and is an important reason why drugs fail during clinical trials and are withdrawn from the market post-approval [2–4]. For example, more than 700 drugs recently have been found to be associated with liver injury, with over 50 drugs have been withdrawn from the global market in the past 50 years due to serious hepatic adverse effects [5–7]. Moreover, a number of marketed drugs have been labeled with“black box” hepatotoxicity warnings. Presently, various preclinical testing strategies have been developed to assess the DILI risk, including in vivo toxicity studies [8–10] and in vitro assays [11–13]. However, these current experimental approaches for assessing the DILI are very expensive, time consuming, and even the results show poor correlation to human observed effects. Thus, developing new tools and approaches to better predict DILI risk in humans has become an important and urgent task. Computational methods have a number of advantages such as being cheaper and faster to

Electronic supplementary material The online version of this article (doi:10.1007/s10822-016-9972-6) contains supplementary material, which is available to authorized users.

✉ Hui Zhang
zhanghui123gansu@163.com

[1] College of Life Science, Northwest Normal University, Lanzhou 730070, Gansu, People's Republic of China

[2] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, Sichuan, People's Republic of China

[3] Bioactive Products Engineering Research Center for Gansu Distinctive Plants, Northwest Normal University, Lanzhou 730070, Gansu, People's Republic of China

generate results, and great efforts have been focused on the development of in silico approaches for DILI prediction.

Presently, various computational prediction approaches for assessing DILI risk have been reported [14, 15], and which can be roughly categorized as knowledge-based models and machine learning methods [16–25]. For example, Greene et al. [16] created the Derek for Windows method with 23 structure alerts, which was test with 626 Pfizer compounds with 56 % overall accuracy, 46 % sensitivity and 73 % specificity. Ekins et al. [17] developed a Bayesian model with several simple descriptors and ECFC_6 fingerprint, which gave concordance of 57–59 % for training set, and 60 % for external test set. Fourches et al. [18] constructed a support vector machine (SVM) prediction model based on 2D fragments and dragon molecular descriptors, which was evaluated by cross-validation method with the concordance of 62–68 %, and external test set containing 18 agents with concordance of 78 %. Liew et al. [19] presented an ensemble model [SVM, k-Nearest Neighbors (kNN)] used 1087 compounds. The model was tested with 120 compounds with an accuracy of 75 %. Recently, Chen et al. [20] performed a classification model of DILI by using decision forest and Molds chemical descriptors, which displayed 70 % for the training set, and 62–69 % for the test sets. Xu et al. [21] developed DILI prediction models using deep learning architectures, and the best model gave 86.9 % accuracy for the external validation set. Furthermore, it was worth mentioning that Matthews et al. [22–24] and Mulliner et al. [25] developed computational prediction models based on the different pathologies of DILI, and these established prediction models can be used for the estimation of the mechanism of action of hepatotoxicity. Obviously, the predictive capacities of previous reported models for DILI prediction were unsatisfactory (lower than 70 %). In addition, most of previous developed classification models of DILI are based on rat data. All of this suggests that creating and developing new computational method based on human derived DILI data with an reasonable accuracy is important and necessary. Thus, in this study, the Naïve Bayes (NB) classifier approach was considered to assess the DILI risk. The Naïve Bayes (NB) classification model based on the Bayes' theorem with the conditional independence assumptions [26, 27], in which each variable can be independently estimated as a one dimensional variable. Compared with other machine learning methods, the Naïve Bayes (NB) method have the following advantages [28]: (1) its ability for handling noisy data and safety with respect to over-fitting; (2) the important features related to activity will significantly influence the performance of model, and the redundant and unimportant descriptors cannot lead to over-fitting during learning; (3) the number of active and inactive compounds used in the training set need not to be balanced; (4) the conditional independence assumption. Due to these advantages, the Naïve Bayes (NB) classifier has been widely applied for the prediction of drug adverse effects, and displayed surprisingly well [29, 30].

The goal of this investigation is to develop a novel computer prediction model of DILI risk by using a larger scale human dataset and Naïve Bayes (NB) classifier, and identify some important molecular descriptors and substructures associated with DILI. The generated prediction model will be validated by 5-fold cross validation and an external test set. We hope the established computational model should be employed for the prediction of DILI in humans at early stage of drug development, and the molecular descriptors and substructures associated with DILI should be taken into consideration in the design of new candidate compounds to help medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

Materials and methods

Dataset selection

Jennings et al. [31] reviewed comprehensive mechanisms of different classes of hepatotoxins. The liver pathologies were classified as two broad categories: cytotoxicity and lipid disorders. Hepatotoxins induced mitochondrial impairment, oxidative stress and apoptosis are classified as the category of cytotoxicity, while steatosis, cholestasis, and phospholipidosis are recognized as the category of lipid dysregulation. The mechanisms of DILI are very complicated and diverse, and even difficult to be elucidated. Especially, some drug could cause severe liver injury in humans but do not induce hepatotoxicity in animals. Thus, in this research, those reported to cause DILI in humans according to the FDA-approved prescription drug labels were extracted from the Chen et al. [3] and Zhu et al. [32]. Chen et al. [3] used FDA-approved drug labeling to generate a benchmark dataset with 287 drugs including 137 drugs of most concern, 85 of less concern and 65 with no concern of causing DILI. Thus, the 137 drugs of most concern were defined as DILI positives and the 65 drugs of no concern were assigned to be DILI negatives. Furthermore, Zhu et al. [32] collected an calibration set, which was approved by US and European toxicity registries. The calibration set contained 177 DILI positives and 105 DILI negatives. The two datasets were then integrated as one dataset, and some duplicate agents were deleted. Finally, 420 agents, including 257 DILI positives and 163 DILI negatives, remained. The structures of 420 compounds were optimized using Discovery Studio (DS) 3.1 software (http://accelrys.com/products/discovery-studio/), and randomly divided as training set (336 compounds, 80 % of the data) and test set (84 compounds, 20 % of the data) (Table 1).

Molecular descriptors

All the molecular descriptors were calculated by Discovery Studio (DS) 3.1 software (http://accelrys.com/products/discovery-studio/). In this investigation, 201 molecular descriptors were initially calculated, mainly including the following classes: 1D descriptors, AlogP, molecular properties, molecular property counts, surface area and volume and topological descriptors. Then, the initial features with too many zero or same values were eliminated. Finally, 91 molecular descriptors were retained and used in the construction of the prediction model.

ECFP_6 fingerprint descriptor

The extended connectivity fingerprints (ECFPs), a class of topological fingerprints for molecular characterization, are derived using a variant of the Morgan algorithm [33]. The ECFPs are designed to capture molecular features relevant to molecular activity, and recently have been applied to substructure searching, drug activity predicting, similarity searching, clustering, and virtual screening [34]. In this study, the ECFP_6 fingerprints were used to analyze the structural features of hepatotoxic/non-hepatotoxic compounds.

Naïve Bayes (NB) classifier

The Naïve Bayes (NB) methods use knowledge of probability and statistics based on applying Bayes' theorem [26, 27]. The Bayes' theorem is described as:

$$P(c|F) = \frac{P(F|c)P(c)}{P(F)} \quad (1)$$

here the parameter c indicates the class variable ("+", positive class and "−", negative class.), $F = (f_1, f_2, \ldots, f_n)$ stands for the object and the $(f_1, f_2, \ldots, f_n)$ represents the feature variables (molecular descriptors) of a object. $P(c)$ is prior probability or marginal probability, $P(F)$ is constant for all classes, $P(c|F)$ and $P(F|c)$ denotes the posterior probability and conditional probability, respectively.

Table 1 Training set and test set used

	Training set	Test set	Sum
Hepatotoxicants	206	51	257
Non-hepatotoxicants	130	33	163
Total	336	84	420

In Naïve Bayes (NB) classifier, all attributes (molecular descriptors) are independent given the value of the class variable, such as:

$$P(F|c) = P(f_1, f_2, \ldots f_n|c) = \prod_{i}^{n} P(f_i|c) \quad (2)$$

Then, the Naïve Bayes (NB) classifier was defined as:

$$f_{nb}(F) = \frac{P(c=+)}{P(c=-)} \prod_{i}^{n} \frac{P(f_i|c=+)}{P(f_i|c=-)} \quad (3)$$

In this investigation, the Naïve Bayes (NB) classifiers were developed by using Discovery Studio (DS) version 3.1 (http://accelrys.com/products/discovery-studio/). The internal validation method for the training set was set as 5-fold cross validation. The "number of Bins" was changed from 10 to 2500 systematically in order to pick appropriate bin sizes. Selection of the number of bins appeared in the histogram, which was used to divide the entire range of observed values for the variable into a series of intervals, and then count how many values fall into each interval [35]. The bin size critically influences the performance of the Naïve Bayes (NB) model. The "Learn Options" was selected as Track Property Ranges, Validate Models, Ignore uninformative Bins and Equipopulate Bins. The FCFP_6 was picked as "Model Domain Fingerprint".

Statistical analysis

The following parameters were used to assess the predictive performance of the classification models: the overall prediction accuracy [Q (Eq. 4)]; Sensitivity [SE (Eq. 5)], the prediction accuracy for the hepatotoxicants); Specificity, [SP (Eq. 6)], the prediction accuracy for the non-hepatotoxicants); Positive predictive value [PPV (Eq. 7)]; Negative predictive value [NPV (Eq. 8)]

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\,\% \quad (4)$$

$$SE = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\,\% \quad (5)$$

$$SP = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\,\% \quad (6)$$

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\,\% \quad (7)$$

$$NPV = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \times 100\,\% \quad (8)$$

where TP (true positives) is the hepatotoxicants that are correctly identified; TN (true negatives) is the non-hepatotoxicants that are correctly recognized; FP (false positives) is the non-hepatotoxicants that are wrongly predicted as hepatotoxicants; FN (false negatives) is the hepatotoxicants that are wrongly classified as non-hepatotoxicants. In addition, the ROC (Receiver Operating Characteristic) score is also widely used to measure the discriminatory power of a classifier system. The value of ROC score is between 1 and 0. The ROC score value of 1 represents the model has a perfect prediction performance, and below 0.5 indicates the model has no discriminative ability.

Results and discussion

Molecular important for the drug-induced liver injury

The mechanisms of DILI are very complicated and diverse, and even difficult to be elucidated. However, the properties of compounds are closely related to their molecular structures. Thus, elimination of these redundant and unimportant descriptors, and identification of most relevant molecular descriptors to the DILI prediction, is an emphasis of this work.

The 91 molecular descriptors selected from 201 initial descriptors together with ECFP_6 were employed for NB-1 construction. The detailed prediction results of NB-1 were shown in Table 2. For the training set, the overall accuracy (*Q*) was 91.1 %. The sensitivity (SE) and specificity (SP) were 98.1 and 86.1 %, respectively. The positive predictive value (PPV) was 91.8 %, the negative predictive value (NPV) was 96.6 %. Furthermore, the NB-1 was evaluated by external test set with the concordance (*Q*) of 70.2 %, sensitivity (SE) of 70.6 %, specificity (SP) of 69.7 %, positive predictive value (PPV) of 78.3 %, negative predictive value (NPV) of 60.5 %. In order to find the most relevant molecular descriptors to the DILI prediction, the following strategy was applied: (1) All of the selected 91 descriptors were deleted one by one to construct the models. (2) The prediction performance of the Naïve Bayes (NB) classification model improved or kept unchanged when one descriptor was deleted from the initial molecular features, which indicated the feature was redundant and unimportant for DILI prediction. Otherwise, the prediction performance reduced when one feature was removed, which represented the descriptor was important for DILI prediction. (3) This processes were repeated many times. Finally, 18 descriptors were identified as the most prominent features for distinguishing DILI positive from DILI negative compounds, such as ALogP, Apol, logD, molecular solubility, molecular weight, number of aromatic rings, number of H acceptors, number of H donors, number of rings, molecular fractional polar SASA, molecular fractional polar surface area, molecular polar SASA, molecular polar surface area, molecular SASA, molecular SAVol, molecular surface area, Wiener and Zagreb. By careful analysis of the selected 18 molecular descriptors, we found which can be roughly classified as molecular structure related descriptors (number of aromatic rings, number of H acceptors, number of H donors and number of rings), molecular properties related descriptors (ALogP, Apol, logD, molecular solubility and molecular weight), molecular surface area and volume related descriptors (molecular fractional polar SASA, molecular fractional polar surface area, molecular polar SASA, molecular polar surface area, molecular SASA, molecular SAVol and molecular surface area) and topological descriptors (Wiener and Zagreb).

Table 2 The prediction results for the training set and test set. NB-1 (91 molecular descriptors + ECFP_6); NB-2 (18 molecular descriptors + ECFP_6); NB-3 (17 molecular descriptors + ECFP_6); NB-4 (18 molecular descriptors); NB-5 (ECFP_6 fingerprint)

Model name		ROC score	TP	FN	TN	FP	SE (%)	SP (%)	PPV (%)	NPV (%)	*Q* (%)
NB-1	Training	0.668	202	4	112	18	98.1	86.1	91.8	96.6	93.5
	Test	0.797	36	15	23	10	70.6	69.7	78.3	60.5	70.2
NB-2	Training	0.714	200	6	116	14	97.1	89.2	93.5	95.1	94.0
	Test	0.805	37	14	24	9	72.5	72.7	80.4	63.2	72.6
NB-3	Training	0.714	200	6	116	14	97.1	89.2	93.5	95.1	94.0
	Test	0.798	37	14	21	12	72.5	63.6	75.5	58.3	69.0
NB-4	Training	0.678	194	12	118	12	94.2	90.8	94.2	90.7	92.9
	Test	0.720	19	32	28	5	37.3	84.8	79.2	46.7	56.0
NB-5	Training	0.693	192	14	119	11	93.2	91.5	94.6	89.5	92.6
	Test	0.782	42	9	20	13	82.4	60.6	76.4	70.0	73.8

Establishment of Naïve Bayes classification model of drug-induced liver injury

Presently, various computational prediction methods of DILI have been extensively reported. However, the prediction performances of these reported models were unsatisfactory, and the prediction accuracies usually lower than 70 %. In this investigation, the Naïve Bayes (NB) classifier approach was selected to predict the DILI. The Naïve Bayes (NB) classification model of DILI was successfully established based on the training set containing 206 hepatotoxicants and 130 non-hepatotoxicants, in which 18 molecular descriptors (ALogP, Apol, logD, molecular solubility, molecular weight, number of aromatic rings, number of H acceptors, number of H donors, number of rings, molecular fractional polar SASA, molecular fractional polar surface area, molecular polar SASA, molecular polar surface area, molecular SASA, molecular SAVol, molecular surface area, Wiener and Zagreb) together with ECFP_6 fingerprint were used. The parameter of number of Bins was defined as 2000. The 5-fold cross validated ROC score for the model built with 18 molecular descriptors and ECFP_6 was 0.714.

Evaluation of Naïve Bayes classification model of drug-induced liver injury

Generally, the internal cross validation method and external validation set were widely applied to evaluate the prediction capability of the classification models. In this study, the internal 5-fold cross validation method for the training set was performed to demonstrate the predictive performance and stability of the established model, and the external validation data set with 84 compounds was used to assess the model's predictive power. The detailed prediction results of NB-2 were given in Table 2. As it can be seen from the Table 2, among these 206 hepatotoxicants, 200 agents were correctly classified as true positives and 6 agents were wrongly defined as negatives. The sensitivity (SE) was 97.1 %. Of these 130 non- hepatotoxicants, 116 agents were distinguished as true negatives and 14 agents were wrongly recognized as positives. The specificity (SP) was 89.2 %. The positive and negative predictive values (PPV and NPV) were 93.5 and 95.1 %, respectively. The concordance (*Q*) of the training set was 94.0 %. The above results represented the Naïve Bayes (NB) prediction model of DILI (NB-2) generated in this investigation have better and stable predictive performance.

In addition, the external validation data set with 84 compounds including 51 hepatotoxicants and 33 non-hepatotoxicants was used to assess the model's predictive power. The detailed prediction results were displayed in Table 2. For these 51 hepatotoxicants, 37 agents were correctly predicted. The sensitive (SE) of the test set was 72.5 %. For these 33 non-hepatotoxicants, 24 were correctly identified as negatives. The specificity (SP) of the test set was 72.7 %. The positive predictive value (PPV) was 80.4 %, the negative predictive value (NPV) was 63.2 %. The overall accuracy (*Q*) and ROC score of the test set were 72.6 % and 0.805, respectively. Obviously, the prediction results of the test set were lower than that of the training set, but the prediction model of the NB-2 could successfully discriminate agents as positives or negatives using some molecular descriptors and ECFP_6 fingerprints.

Moreover, random elimination of one descriptor from the selected 18 descriptors, and using the remaining 17 descriptors combined with ECFP_6 to construct other Naïve Bayes (NB) model was also run. For example, the 17 descriptors together with ECFP_6 fingerprints were used to build other Naïve Bayes models (NB-3, where the descriptor of molecular weight was deleted from the 18 descriptors). The detailed prediction results of the training set and test set for NB-3 were presented in Table 2. The ROC score in the training set and test set were 0.714 and 0.798, respectively. The sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) for the training set were 97.1, 89.2, 93.5 and 95.1 %, respectively. The concordance (*Q*) was 94.0 %. The NB-3 was assessed by the test set with overall accuracy (*Q*) of 69.0 %, sensitivity (SE) of 72.5 %, specificity (SP) of 63.6 %, positive predictive value (PPV) of 75.5 %, and negative predictive value (NPV) of 58.3 %. Clearly, among the NB-1, NB-2 and NB-3, the NB-2 performed best. This illustrated that these selected 18 descriptors were important for DILI prediction.

Analysis of the hepatotoxic/non-hepatotoxic fragments produced by the ECFP_6 fingerprint descriptors

Fingerprint descriptors have been used to capture molecular features relevant to molecular activity. In order to evaluate whether the ECFP_6 fingerprint descriptors are important for the DILI prediction, the NB-4 was also constructed, in which the ECFP_6 fingerprint descriptor was removed, and only the 18 molecular descriptors were applied. The detailed prediction results of NB-4 were listed in Table 2. For the training set of NB-4, the sensitivity (SE) was 94.2 %, the specificity (SP) was 90.8 %. The positive and negative predictive values ((PPV and NPV) were 94.2 and 90.7 %, respectively. The overall accuracy (Q) was 92.9 %. For the test set, the sensitivity (SE) was 37.3 %, the specificity (SP) was 84.8 %, the positive predictive value (PPV) was 79.2 %, the negative predictive value (NPV) was 46.7 %, and the overall accuracy (Q) was 56.0 %. The ROC score values in the training set and test set were 0.678 and 0.720, respectively. Obviously, the prediction performance of NB-4, especially for the test set, significantly reduced compared with that of NB-2. Furthermore, in order to directly decide whether the ECFP_6 fingerprint descriptors are important for DILI prediction, only ECFP_6 fingerprint descriptors without the 18 selected descriptors were used to construct the prediction model (NB-5). The detailed prediction results of NB-5 were given in Table 2. For the training set of NB-5, the sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) were 93.2, 91.5, 94.6 and 89.5 %, respectively. And the overall accuracy (*Q*) was 92.6 %. For the test set of NB-5, the overall accuracy (*Q*) was 73.8 %. The sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) were 82.4, 60.6, 76.4 and 70.0 %, respectively. The ROC scores of the training set and test set for NB-5 were 0.693 and 0.782, respectively. Compared the prediction accuracies and ROC scores given by NB-2, NB-4 and NB-5, the results indicated the ECFP_6 fingerprints played important role in the prediction of DILI.

Thus, in order to better understand the contribution of particular structural features to DILI prediction, the molecular fragments among the 257 hepatotoxic agents and 163 non-hepatotoxic compounds were produced by using ECFP_6 fingerprints (Fig. 1). The top 20 good features associated with hepatotoxicity displayed in Fig. 1a and the top 20 bad features not associated with hepatotoxicity presented in Fig. 1b. As it can be seen from Fig. 1a, each panel shows the naming convention for each fragment, the number of molecules that occur in hepatotoxic agents, and the Bayesian score for the fragment. The Bayesian score is a measure of how different this is from the hit rate as a whole (the ratio that would be expected if the feature was occurring randomly across the hepatotoxicants and non-hepatotoxicants). The score takes the total number of occurrences of the feature into account, ensuring more weight is placed on those more frequent occurrences of the feature, and little weight is placed on the feature with very few occurrences. By analyzing the fingerprints generated in hepatotoxic and non-hepatotoxic agents, we observed that there was no common substructure shown in the hepatotoxic compounds (Fig. 1a) and in the non-hepatotoxic agents (Fig. 1b). Some fragments, such as these fragments containing prop-1-en-2-ylbenzene (G1, G16), furan (G2), N-methylacetamide (G3), trimethylamine groups (G4, G5, G8), azetidine (G6, G18, G20), methyl acetate group (G7), isoxazole (G10), sulfane (G11), Fluorine (G12) or Bromum (G17) and pyridine (G15) group, only appeared in these hepatotoxicants. In addition, the 2-methylbut-1-ene (G9, G13, G14, G19) toxic groups occurred in these non-hepatotoxicants, but which appeared more often in the hepatotoxic active compounds than in the inactive ones. In addition, structural alerts (SAs) for hepatotoxicity were also developed by other research groups. For example, Hewitt et al. [36] displayed 16 structural alerts associated with observed human hepatotoxicity, and investigated the mechanism(s) of DILI. Comparing some representative substructures displayed in Fig. 1a with SAs for hepatotoxicity obtained by Hewitt et al. [36], we found some fragments listed in Fig. 1a occurred in these 16 SAs. For example, The prop-1-en-2-ylbenzene (G1) was the moiety of Structural alert 1 (Tamoxifen-like antioestrogen), which can interact with biological membranes, and disrupt many processes within cells. The pyridine group (G15) was part of the Structural alert 2 (Adenosine-based reverse transcriptase inhibitors), which can induce liver injury through mitochondrial dysfunction. The groups of G6, G18 and G20 were the Structural alert 4 (β-lactam substructure), which can induce minor liver injury characterized by liver enzyme elevations. Moreover, it was worth mentioning that some new hepatotoxic fragments, such as furan, N-methylacetamide, trimethylamine, methyl acetate, isoxazole, sulfane Fluorine or Bromum and 2-methylbut-1-ene, were identified in this research, which should also be taken into consideration in hepatotoxicity screening, chemical risk assessment and drug design, and helping medicinal chemists rationally select the chemicals with the best prospects to be effective and safe.

Comparison with other prediction models of drug-induced liver injury

Developing in silico approaches to predict DILI risk has become a research focus in recent years. Presently, various computational classification models have been developed to assess the DILI risk, including Derek for Windows, Bayesian model, SVM, Decision Forest and Deep Learning Architectures. For comparison, the detailed classification accuracies of these recently reported models were summarized in Table 3. It is worth mentioning that direct comparison of our results with previous studies was inappropriate, because of the application of different sets of chemicals, number of molecular descriptors, and validation methods. However, a simple comparison could provide some basic information about the accuracy of various prediction methodologies. By carefully comparing the prediction performance of these methods displayed in Table 3, it was found that the prediction accuracies of the training set for Naïve Bayes (NB) prediction model established in this investigation were significantly better than that of other methods. Furthermore, the prediction results of the test set obtained in this study were lower than that of the Deep Learning Architectures, but better than that of other methods. This suggests that the Naïve Bayes (NB) prediction model of DILI established here could give

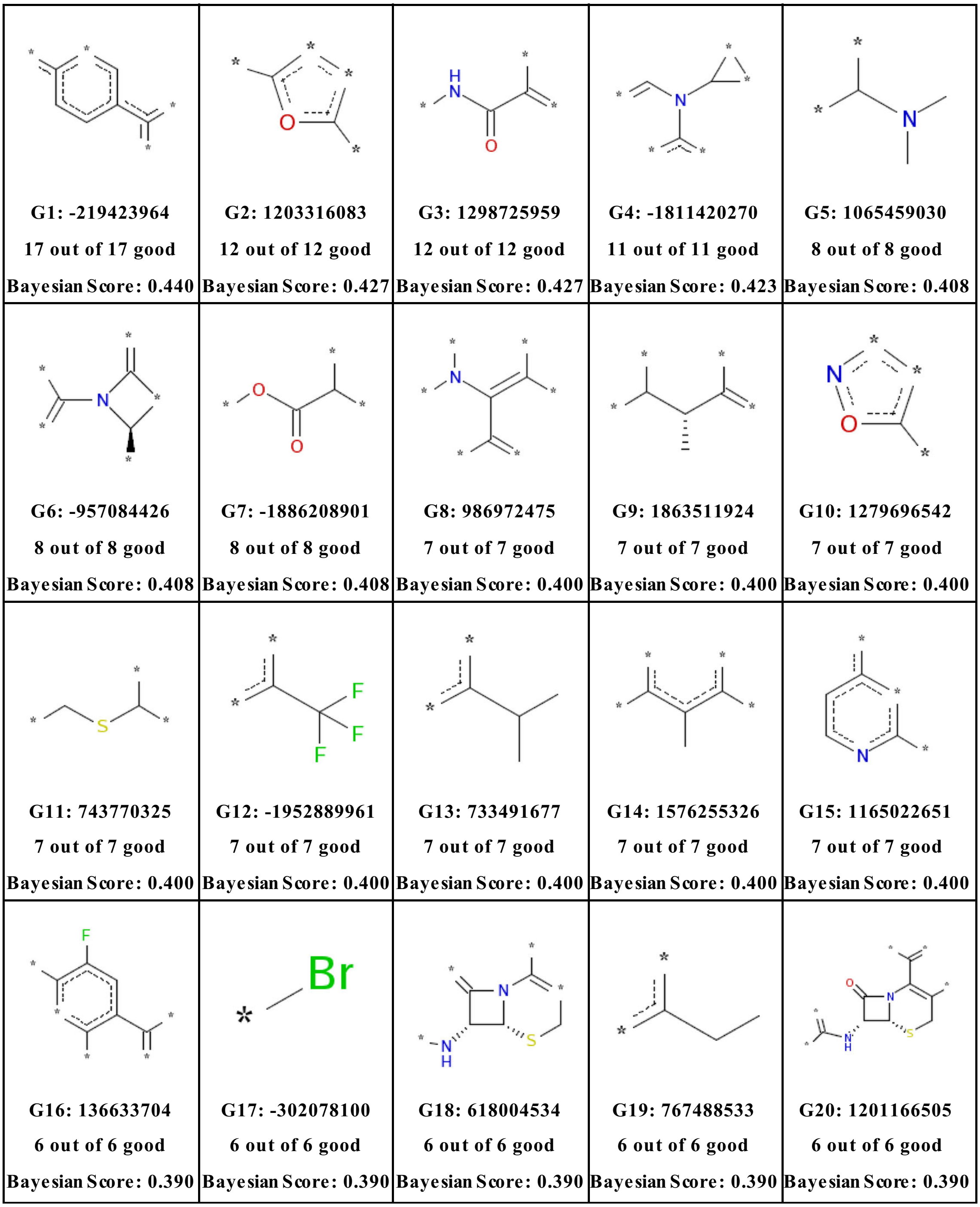
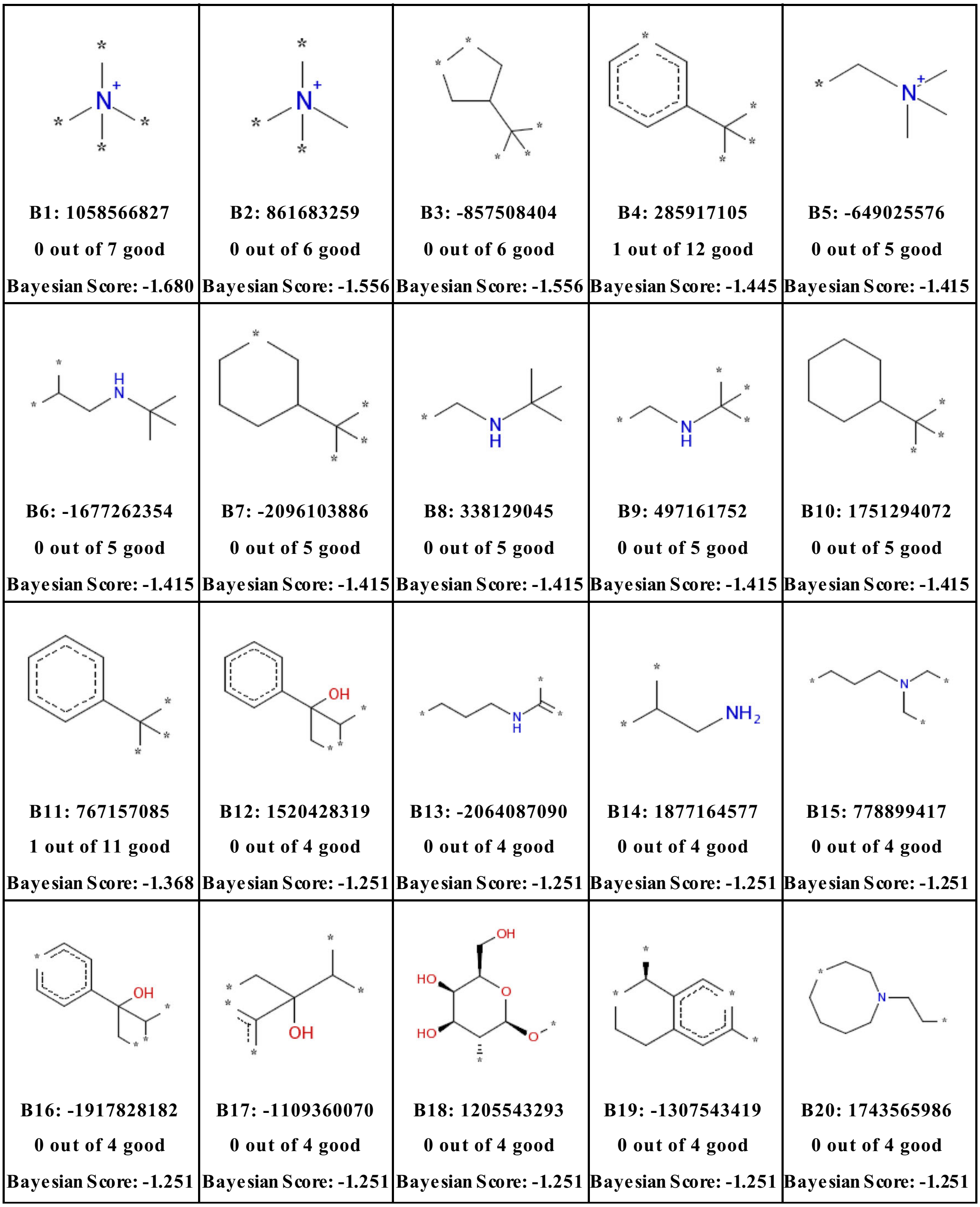


Fig. 1 Some molecular fragments that important for DILI were produced by using the ECFP_6 fingerprint descriptors. **a** The top 20 hepatotoxic substructures generated as good features. Each panel shows the naming convention for each fragment, the numbers of molecules it is present in that are hepatotoxic agents, and the Bayesian score for the fragment. **b** The top 20 non-hepatotoxic substructures produced as bad features. Each panel shows the naming convention for each fragment, the numbers of molecules it is present in that are hepatotoxic compounds, and the Bayesian score for the fragment. The *dashed lines* means conjugated double bond, *asterisk* represents the site can be replaced by different atoms

reasonable prediction accuracies. In addition, compared with previous studies, the other advantages of this research was that the 18 important molecular descriptors related to DILI prediction and 20 toxic/non-toxic fragments were identified, which could provide guidance for medicinal chemists working in drug discovery and lead optimization, and avoid the DILI occurrence in the later phase of drug development.



B1: 1058566827 0 out of 7 good Bayesian Score: -1.680	B2: 861683259 0 out of 6 good Bayesian Score: -1.556	B3: -857508404 0 out of 6 good Bayesian Score: -1.556	B4: 285917105 1 out of 12 good Bayesian Score: -1.445	B5: -649025576 0 out of 5 good Bayesian Score: -1.415
B6: -1677262354 0 out of 5 good Bayesian Score: -1.415	B7: -2096103886 0 out of 5 good Bayesian Score: -1.415	B8: 338129045 0 out of 5 good Bayesian Score: -1.415	B9: 497161752 0 out of 5 good Bayesian Score: -1.415	B10: 1751294072 0 out of 5 good Bayesian Score: -1.415
B11: 767157085 1 out of 11 good Bayesian Score: -1.368	B12: 1520428319 0 out of 4 good Bayesian Score: -1.251	B13: -2064087090 0 out of 4 good Bayesian Score: -1.251	B14: 1877164577 0 out of 4 good Bayesian Score: -1.251	B15: 778899417 0 out of 4 good Bayesian Score: -1.251
B16: -1917828182 0 out of 4 good Bayesian Score: -1.251	B17: -1109360070 0 out of 4 good Bayesian Score: -1.251	B18: 1205543293 0 out of 4 good Bayesian Score: -1.251	B19: -1307543419 0 out of 4 good Bayesian Score: -1.251	B20: 1743565986 0 out of 4 good Bayesian Score: -1.251

Fig. 1 continued

Conclusion

In this investigation, the prediction model of DILI has been successfully developed by using a larger scale human dataset and Naïve Bayes classifier. The established Naïve Bayes prediction model was evaluated by 5-fold cross validation and an external test set. For the training set, the sensitivity, specificity, positive predictive value and negative predictive value were 97.1, 89.2, 93.5 and 95.1 %, respectively. The overall prediction accuracy was 94.0 %. The test set with the concordance of 72.6 %, sensitivity of 72.5 %, specificity of 72.7 %, positive predictive value of 80.4 %, negative predictive value of 63.2 %. Furthermore, 18 important molecular descriptors related to DILI and some toxic/non-toxic fragments were identified. Thus, we hope the prediction model established here would be

Table 3 Previous classification accuracies of DILI

Model name		Sensitivity SE (%)	Specificity SP (%)	*Q* (%)	Reference
Derek for windows		46	73	56	Greene et al. [16]
Bayesian model	Training set	57.6–58.5	61.8–65.4	57.6–59.2	Ekins et al. [17]
	Test set	56	67	60	
SVM	Training set			61.9–67.5	Fourches et al. [18]
	Test set			78	
Decision forest	Training set	57.8	77.9	69.7	Chen et al. [20]
	Test set	58.4–66.3	66.1–71.6	61.6–68.9	
Deep learning architectures	Training set	89.9	87.0	88.4	Xu et al. [21]
	Test set	82.5	92.9	86.9	
Novel Naïve Bayes	Training set	97.1	89.2	94.0	Present study
	Test set	72.5	72.7	72.6	

employed for the assessment of DILI, and the molecular descriptors and fragments should be taken into consideration in the design of new candidate compounds to produce safer and more effective drugs, and finally reduce DILI occurrence in later stages of drug development. The prediction model (NB-2) was supplied in the supplementary material.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 81660589) and the Project for Enhancing the Research Capability of Young Teachers in Northwest Normal University (NWNU-LKQN-12-7).

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Björnsson E (2006) Drug-induced liver injury: Hy's rule revisited. Clin Pharmacol Ther 79(521–528):1
2. Fung M, Thornton A, Mybeck K, Hsiao-Hui W, Hornbuckle K, Muniz E (2001) Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets 1960 to 1999. Drug Inf J 35:293–317
3. Chen MJ, Vijay V, Shi Q, Liu ZC, Fang H, Tong WD (2011) FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov Today 16:696–703
4. Oda S, Matsuo K, Nakajima A, Yokoi T (2016) A novel cell-based assay for the evaluation of immune- and inflammatory-related gene expression as biomarkers for the risk assessment of drug-induced liver injury. Toxicol Lett 241:60–70
5. Food and Drug Administration (2009) Guidance for industry drug-induced liver injury: premarketing clinical evaluation. Food and Drug Administration, Silver Spring, MD, pp 38035–38036
6. Hoofnagle JH, Serrano J, Knoben JE, Navarro VJ (2013) Livertox: a website on drug-induced liver injury. Hepatology 57:873–874
7. Assis DN, Navarro VJ (2009) Human drug hepatotoxicity: a contemporary clinical perspective. Expert Opin Drug Metab Toxicol 5:463–473
8. Mattes W, Davis K, Fabian E, Greenhaw J, Herold M, Loosere R, Mellert W, Groeters S, Marxfeld H, Moellerf N, Montoya-Parra G, Prokoudin A, van Ravenzwaay B, Strauss V, Walk T, Kamp H (2014) Detection of hepatotoxicity potential with metabolite profiling (metabolomics) of rat plasma. Toxicol Lett 230:467–478
9. Jennen D, Polman J, Bessem M, Coonen M, van Delft J, Kleinjans J (2014) Drug-induced liver injury classification model based on in vitro human transcriptomics and in vivo rat clinical chemistry data. Syst Biomed 2:63–70
10. Zhang M, Chen MJ, Tong WD (2012) Is toxicogenomics a more reliable and sensitive biomarker than conventional indicators from rats to predict drug-induced liver injury in humans? Chem Res Toxicol 25:122–129
11. Shah F, Greene N (2013) Analysis of Pfizer compounds in EPA's ToxCast chemicals-assay space. Chem Res Toxicol 27:86–98
12. Chen M, Tung C, Shi Q, Guo L, Shi L, Fang H, Borlak J, Tong W (2014) A testing strategy to predict risk for drug-induced liver injury in humans using high-content screen assays and the 'rule-of-two' model. Arch Toxicol 88:1439–1449
13. Tomida T, Okamura H, Satsukawa M, Yokoi T, Konno Y (2015) Multiparametric assay using HepaRG cells for predicting drug-induced liver injury. Toxicol Lett 236:16–24
14. Ekins S (2014) Progress in computational toxicology. J Pharmacol Toxicol Methods 69:115–140
15. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W (2016) DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discov Today 21:648–653
16. Greene N, Fisk L, Naven RT, Note RR, Patel ML, Pelletier DJ (2010) Developing structure–activity relationships for the prediction of hepatotoxicity. Chem Res Toxicol 23:1215–1222
17. Ekins S, Williams AJ, Xu JJ (2010) A predictive ligand-based Bayesian model for human drug induced liver injury. Drug Metab Dispos 38:2302–2308
18. Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A (2010) Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. Chem Res Toxicol 23:171–183

19. Liew CY, Lim YC, Yap CW (2011) Mixed learning algorithms and features ensemble in hepatotoxicity prediction. J Comput Aided Mol Des 25:855–871
20. Chen M, Hong H, Fang H, Kelly R, Zhou G, Borlak J, Tong W (2013) Quantitative structure–activity relationship models for predicting drug-induced liver injury based on FDA-approved drug labeling annotation and using a large collection of drugs. Toxicol Sci 136:242–249
21. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L (2015) Deep learning for drug-induced liver injury. J Chem Inf Model 55:2085–2093
22. Matthews EJ, Kruhlak NL, Benz RD, Aragonés Sabaté D, Marchant CA, Contrera JF (2009) Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: part C: use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities. Regul Toxicol Pharmacol 54(1):43–65
23. Matthews EJ, Ursem CJ, Kruhlak NL, Benz RD, Sabaté DA, Yang C, Klopman G, Contrera JF (2009) Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. Regul Toxicol Pharmacol 54:23–42
24. Ursem CJ, Kruhlak NL, Contrera JF, MacLaughlin PM, Benz RD, Matthews EJ (2009) Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans. Part A: use of FDA post-market reports to create a database of hepatobiliary and urinary tract toxicities. Regul Toxicol Pharmacol 54:1–22
25. Mulliner D, Schmidt F, Stolte M, Spirkl HP, Czich A, Amberg A (2016) Computational models for human and animal hepatotoxicity with a global application scope. Chem Res Toxicol 29:757–767
26. Berger JO (2013) Statistical decision theory and Bayesian analysis. Springer, Berlin
27. Box G, Tiao CC (2011) Bayesian inference in statistical analysis. Wiley, London
28. Langdon SR, Mulgrew J, Paolini GV, van Hoorn WP (2010) Predicting cytotoxicity from heterogeneous data sources with Bayesian learning. J Cheminform 2:11–29
29. Zhang H, Yu P, Zhang TG, Kang YL, Zhao X, Li YY, He JH, Zhang J (2015) In silico prediction of drug-induced myelotoxicity by using Naïve Bayes method. Mol Divers 19:945–953
30. Zhang H, Yu P, Xiang ML, Li XB, Kong WB, Ma JY, Wang JL, Zhang JP, Zhang J (2016) Prediction of drug-induced eosinophilia adverse effect by using SVM and Naïve Bayesian approaches. Med Biol Eng Comput 54:361–369
31. Jennings P, Schwarz M, Landesmann B, Maggioni S, Goumenou M, Bower D, Leonard MO, Wiseman JS (2014) SEURAT-1 liver gold reference compounds: a mechanism-based review. Arch Toxicol 88(2099–2133):32
32. Zhu X, Kruhlak NL (2014) Construction and analysis of a human hepatotoxicity database suitable for QSAR modeling using post-market safety data. Toxicology 321:62–72
33. Morgan HL (1965) The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. J Chem Doc 5:107–113
34. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
35. Shimazaki H, Shinomoto S (2007) A method for selecting the bin size of a time histogram. Neural Comput 19:1503–1527
36. Hewitt M, Enoch SJ, Madden JC, Przybylak KR, Cronin MT (2013) Hepatotoxicity: a scheme for generating chemical categories for read-across, structural alerts and insights into mechanism(s) of action. Crit Rev Toxicol 43:537–558