Hindawi
Journal of Chemistry
Volume 2021, Article ID 2293871, 20 pages
https://doi.org/10.1155/2021/2293871

Research Article

Joint Decision-Making Model Based on Consensus Modeling Technology for the Prediction of Drug-Induced Liver Injury

Yukun Wang[1,2] **and Xuebo Chen**[2]

[1]*School of Chemical Engineering, University of Science and Technology Liaoning, No. 185, Qianshan, Anshan 114051, Liaoning, China*
[2]*School of Electronic and Information Engineering, University of Science and Technology Liaoning, No. 185, Qianshan, Anshan 114051, Liaoning, China*

Correspondence should be addressed to Xuebo Chen; xuebochen@126.com

Received 27 August 2021; Revised 21 November 2021; Accepted 10 December 2021; Published 27 December 2021

Academic Editor: Robert Zalesny

Drug-induced liver injury (DILI) is the major cause of clinical trial failure and postmarketing withdrawals of approved drugs. It is very expensive and time-consuming to evaluate hepatotoxicity using animal or cell-based experiments in the early stage of drug development. In this study, an in silico model based on the joint decision-making strategy was developed for DILI assessment using a relatively large dataset of 2608 compounds. Five consensus models were developed with PaDEL descriptors and PubChem, Substructure, Estate, and Klekota–Roth fingerprints, respectively. Submodels for each consensus model were obtained through joint optimization. The parameters and features of each submodel were optimized jointly based on the hybrid quantum particle swarm optimization (HQPSO) algorithm. The application domain (AD) based on the frequency-weighted and distance (FWD)-based method and Tanimoto similarity index showed the wide AD of the qualified consensus models. A joint decision-making model was integrated by the qualified consensus models, and the overwhelming majority principle was used to improve the performance of consensus models. The application scope narrowing caused by the overwhelming majority principle was successfully solved by joint decision-making. The proposed model successfully predicted 99.2% of the compounds in the test set, with an accuracy of 80.0%, a sensitivity of 83.9, and a specificity of 73.3%. For an external validation set containing 390 compounds collected from DILIrank, 98.2% of the compounds were successfully predicted with an accuracy of 79.9%, a sensitivity of 97.1%, and a specificity of 66.0%. Furthermore, 25 privileged substructures responsible for DILI were identified from Substructure, PubChem, and Klekota–Roth fingerprints. These privileged substructures can be regarded as structural alerts in hepatotoxicity evaluation. Compared with the main published studies, our method exhibits certain advantage in data size, transparency, and standardization of the modeling process and accuracy and credibility of prediction results. It is a promising tool for virtual screening in the early stage of drug development.

1. Introduction

Bringing a new drug to market is a time-consuming and expensive process. Lots of candidate drugs fail to become drugs or withdraw from market mainly because of their safety and lack of efficacy [1]. Drug toxicity evaluation is an essential process of drug development as it is reportedly responsible for the attrition of approximately 30% of drug candidates [2]. Human adverse effects of drugs (AEDs) cost upward of $3.6 billion each year and constitute one of the top 10 causes of death in the United States [3]. Hepatotoxicity is a serious adverse effect, which seriously threatens the safety of patients and is also an important cause for drug withdrawal from the market [4, 5]. At least a quarter of the drugs are prematurely terminated or withdrawn from the market due to liver-related liabilities [6]. Currently, more than 50 drugs, such as troglitazone, pemoline, tienilic acid, and benoxaprofen, have been withdrawn worldwide due to severe hepatotoxicity [7]. It causes unnecessary huge loss of financial, energy, and time expenditure [8]. Hence, early

assessment of the drug-induced liver injury (DILI) potential of drug candidates is important and very useful for improving the efficiency of drug development. Commonly used techniques for DILI evaluation, such as animal studies, reactive metabolites, various cell cultures, human liver microsomes, and recombinant enzymes, are time-consuming and expensive [9–12]. Due to the lack of efficient laboratory tests for DILI assessment during drug design, the hepatotoxicity of some drugs is often observed when the population is exposed to drug postmarketing [13]. For some patients with chronic diseases, drug-induced hepatotoxicity or other toxicity (e.g., nephrotoxicity) is likely to be the ultimate killer, rather than the disease itself [14]. Therefore, the development of safe and effective drugs can better benefit mankind. Developing fast and accurate experimental and computational approaches to assess the risk of DILI, to understand the biological pathways of DILI, and to search novel biomarkers for liver injury has aroused a great deal of interest worldwide [15, 16]. In silico techniques, especially (quantitative) structure-activity relationship ((Q)SAR), can directly employ hepatotoxicity data to infer the toxicity of other candidate drugs and avoid the limitations of traditional animal or cell-based experiments [4, 8]. Compared with experimental methods, (Q)SAR models are applicable to virtual molecules even before they are isolated or synthesized, so as to conduct virtual screening in the early stage of drug development [17]. (Q)SAR can also identify key fragments of chemical structures that have an impact on hepatotoxicity, enabling pharmaceutical chemists to select a suitable structure for drug molecules. With low attrition of time and money, and a fast readout, (Q)SAR models have increasingly become good tools for predicting hepatotoxicity in the early stage of drug development [18, 19].

In the past decades, DILI data have been systematically recorded in numerous public databases (such as LiverTox, Hepatox, and Liver Toxicity Knowledge Base) and some published studies [20–22]. These data have become valuable resources in the study of DILI, and many classification models have been developed with these datasets. As far as we know, the reported models for DILI predictions were quite different in terms of modeling technologies, data size, data sources, and features. In terms of modeling technologies, various algorithms, such as support vector machine (SVM) [5], random forest (RF) [8], Bayesian [23], deep learning methods [24, 25], AdaBoost decision trees [26], k-nearest neighbor [27], and artificial neural networks (ANNs) [28–30], have been used to develop DILI prediction models. In terms of data size, the smallest dataset consists of only 74 molecules [29], and the largest dataset consists of 3712 molecules [31]. In terms of data source, hepatotoxicity data from clinical trials [10, 11, 31] and cell-based [24, 32] and animal (such as rodents and nonrodents [15]) experiments have been used to develop DILI prediction models. These data are mainly from a public hepatotoxicity database or published papers or mixed sources. In terms of the types and numbers of features, there were various descriptors or fingerprints (such as PubChem fingerprint (PubchemFP), SiRMS, MDs, ECFP6, Estate fingerprint (EstateFP), CDK fingerprint, CDK extended fingerprint, Klekota–Roth fingerprint (KRFP), MACCS keys, MOE, PaDEL, and Dragon) used in these models [17, 21, 32–37]. The minimum number of features selected in these models is 12, and the maximum number of selected features exceeds 1000. Although most of DILI prediction models have relatively reasonable accuracy, they still have some shortcomings, mainly in the following aspects: (1) Some studies used less than 500 compounds to develop their DILI models [23, 28, 29, 37]. Although these models have relatively high prediction accuracy, some of them may have little practical value due to their small datasets. It is difficult to collect sufficient and diverse molecular information for them to develop a classification model with a wide AD. (2) During the modeling process, not all models conform to the principles of the Organisation for Economic Co-operation and Development (OECD). Some models were not cross-verified [9, 30, 38], some models lacked an AD [31, 39], and some models lacked mechanistic interpretation [9, 40]. In addition, the modeling process of some models is not transparent enough [11, 17–21, 30], which leads to poor repeatability and reproducibility. (3) Imbalanced chemical data are very common in DILI studies. It is difficult to achieve a well-balanced sensitivity and specificity of a machine learning model trained on imbalanced chemical data, but only Bajželj [30] used technical means to balance DILI negatives and DILI positives. (4) Optimization technology has been widely used in (Q)SAR research to improve the performance of (Q)SAR models [41]. But as far as we know, only Mulliner et al. [31] applied genetic algorithm (GA) optimization technology to optimize their DILI models.

These shortcomings limit the widespread applications of these models in drug discovery. Taking into consideration the above-mentioned issues, we attempt to develop a joint decision-making model based on a relatively large and chemically diverse DILI dataset to achieve better performance and credibility, as well as a wide AD. To achieve this goal, a total of 2608 compounds (1643 DILI positives/965 DILI negatives) were first manually collected. Then, molecular features of all compounds were characterized by PaDEL descriptors, Substructure fingerprint (SubFP), PubchemFP, EstateFP, and KRFP, respectively. Five consensus models were constructed based on support vector machine (SVM) and hybrid quantum particle swarm optimization (HQPSO) algorithm [42] and consensus modeling technology. Four qualified consensus models were integrated into a joint decision-making model to improve the accuracy and credibility of prediction results. Furthermore, the contributions of PaDEL descriptors to DILI were analyzed, and some special substructures associated with DILI were identified as privileged substructures for structural alerts. We hope the model could be a useful tool for DILI assessment of drug candidates in the early stage of drug discovery, and the privileged substructures will be helpful for pharmaceutical chemists to design an appropriate structure for drug molecules.

2. Experimental Section

2.1. Modeling Overview. A workflow for the modeling process is shown in Figure 1. In the modeling process, principal component analysis (PCA) and self-organizing mapping (SOM) neural network were used to split the dataset into diverse training and test sets, and then the training set was divided into a fivefold cross-validation set for subsequent study. In feature selection of PaDEL descriptors, the mean decrease impurity (MDI) method [43] was used to evaluate the importance of descriptors. Each consensus model is composed of 20 submodels, each of which is constructed by SVM and joint optimization strategy. The HQPSO algorithm and fivefold cross-validation were employed to jointly optimize the parameters and features of submodels. The joint decision-making model was integrated with the qualified consensus models whose accuracy of fivefold cross-validation (ACC_{cv5}) is great than 0.7. Finally, the test set and independent external validation set were used to validate the performance of the proposed joint decision-making model. Specific details of each modeling step are provided in subsequent sections.

2.2. Assessment of Model Performance. A clear performance evaluation index is the premise of (Q)SAR research. In this study, several statistical parameters were used for assessment of the performance of constructed models, including sensitivity (SE), specificity (SP), prediction accuracy (ACC), Matthews correlation coefficient (MCC), and geometric mean (GMEN), which are defined, respectively, as follows:

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{3}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \tag{4}$$

$$\mathrm{GMEN} = \sqrt{\mathrm{SE} \times \mathrm{SP}}, \tag{5}$$

where TP, TN, FP, and FN are the numbers of true DILI positives, true DILI negatives, false DILI positives, and false DILI negatives, respectively, and SE and SP stand for the prediction accuracy of DILI positives and DILI negatives, respectively. The MCC value is generally regarded as a balanced measure, which can be used even if the classes have very different sizes. It returns a value between −1 and 1. A coefficient of 1 represents a perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation [5]. The GMEN can indicate the balance between specificity and sensitivity of a machine learning model [44].

Additionally, the receiver-operating characteristic (ROC) curve was employed to assess the performance of constructed models. The ROC curve is plotted by varying the threshold values of prediction probabilities of each consensus model to show the separation ability of these binary classification models. Above the threshold value, the output of the model is predicted as positive; otherwise, it is negative. The curve is a graphical plot of sensitivity against the false-positive rate (1 − specificity). The values of the area under the ROC curve (AUC) were also computed to quantitatively describe the ROC curve. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The AUC value ranges from 0.5 to 1.0. AUC = 1.0 means the classifier is perfect, and AUC = 0.5 indicates the classifier is useless and random [39].

2.3. Data Collection. It is well known that the more the chemically diverse compounds used for (Q)SAR modeling, the wider the AD of the obtained model. Merging datasets is an effective way to increase the amount of data and expand the application domain of DILI models [9]. In our study, a relatively large DILI dataset was extracted from five publications [15, 24, 27, 34, 39]. To improve the quality and reliability of data, we coped with them as follows:

(1) Inorganic compounds, mixtures, and compounds without SMILES string were removed.

(2) Compounds with no clear DILI effect were removed.

(3) All confused DILI effects were checked carefully. For the same compound listed in different publications, if the number of DILI-positive results and DILI-negative results is the same, we delete the compound. If the number of DILI-positive results is greater than the number of DILI-negative results, we label it as DILI positive, and vice versa.

(4) Duplicates in the dataset were removed.

2.4. Description and Descriptor Pruning. In this study, to realize the joint decision-making strategy, and construct a robust and convincing (Q)SAR model, commonly used PaDEL descriptors and four types of fingerprints were calculated. These fingerprints include PubchemFP (881 bits), SubFP (307 bits), KRFP (4860 bits), and EstateFP (79 bits). The descriptors and fingerprints were calculated through the online website (http://www.scbdd.com).

Firstly, PaDEL descriptors were calculated. To construct a stable DILI model, we only calculated 1D and 2D PaDEL descriptors to avoid uncertainty caused by molecular structure optimization when calculating 3D descriptors. After that, we examined each compound carefully. If a compound is with null descriptor and the variance of this descriptor in other compounds is greater than 0.3, we removed this compound to avoid deleting important descriptors in subsequent steps. To simplify the structure of constructed models and reduce redundancy, one pretreatment was performed to delete some

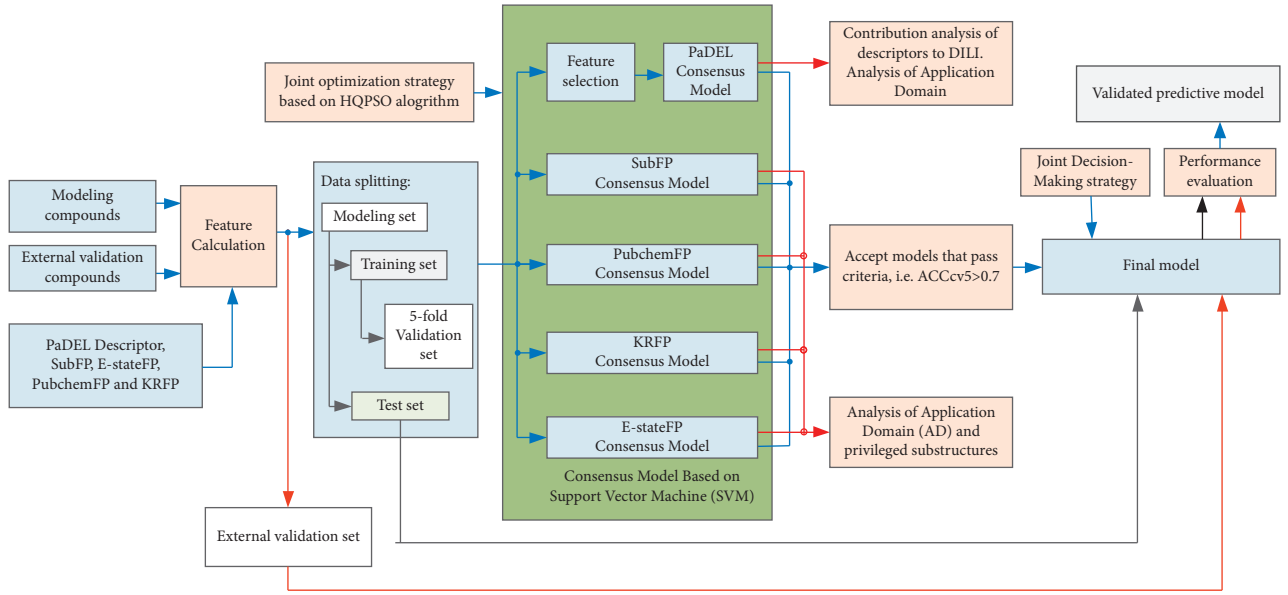

Figure 1: Flowchart of the modeling process of the joint decision-making model.

uninformative or redundant descriptors before further selection: descriptors with constant or null values were excluded and if two descriptors were found to be correlated pairwise (greater than 0.85), one of them was deleted randomly [43]. After that, 2608 compounds (1643 DILI positives/965 DILI negatives) and 298 descriptors were reserved. The 2608 compounds used in this study are listed in the Supplementary Materials (DILI_DataSet.pdf).

Secondly, fingerprints (PubchemFP, SubFP, EstateFP, and KRFP) of the 2608 compounds were calculated.

2.5. Data Splitting Based on PaDEL Descriptors. The real predictive ability of a DILI model must be evaluated by the predictive accuracy of compounds not used in model development. This type of assessment requires the use of a test set. To obtain chemically diverse training and test sets, and keep the balance of DILI positives and DILI negatives in the training and test sets, PCA and SOM neural network were employed to split the modeling data. A workflow for data splitting is shown in Figure 2.

Firstly, the PCA method was used to deal with the input variables (characterized by 298 PaDEL descriptors) of the DILI-positive and DILI-negative datasets, respectively. Then, two SOM neural networks were constructed to split the DILI-positive and DILI-negative compounds, respectively. For each SOM neural network, the first K principal components whose cumulative contribution rate is more than 90% (for the DILI-positive dataset, $K = 46$, and for the DILI-negative dataset, $K = 41$) were selected as the input of the SOM neural network.

After that, each SOM divided the DILI positives and DILI negatives into nine groups, respectively, so that each group of data has structural similarity. The clustering results of DILI positives and DILI negatives are shown in Figures 3(a) and 3(b), respectively.

Then, for each group of data, we randomly selected 80% of the data to constitute the training set and assigned the remaining 20% to the test set. Finally, we obtained a training set containing 2080 compounds (1311 DILI positives/769 DILI negatives) and a test set containing 528 compounds (332 DILI positives/196 DILI negatives). A detailed classification is listed in the Supplementary Materials (DILI_DataSet.pdf).

In addition, to facilitate optimization and obtain a robust DILI model in subsequent studies, a fivefold cross-validation dataset was constructed using the DILI-positive and DILI-negative training sets for each kind of descriptor and fingerprint. The flow of constructing the fivefold cross-validation dataset is shown in Figure 4.

Firstly, DILI-positive and DILI-negative sets were clustered to nine SOM nodes by the PCA-SOM method, respectively. And then the compounds in each node were randomly divided into five subsets of the same size and assigned to onefold to fivefold. In this way, each fold in the fivefold cross-validation set was as balanced as possible in terms of chemical diversity and the ratio of DILI positives and DILI negatives.

2.6. Model Building. As can be seen from Figure 1, the joint decision-making model consists of some qualified consensus models based on descriptors and fingerprints. Consensus modeling technology has been proved to be effective in improving the overall performance of the (Q)SAR model [38, 45, 46]. In our study, each consensus model is composed of 20 submodels constructed by SVM. SVM is an excellent kernel-based tool for classification or regression introduced by Vapnik et al [47]. The algorithm employs a kernel function (e.g., linear, polynomial, and radial) to maximize the decision boundary between classes and find an optimal hyperplane able to best discriminate the classes [48]. It is an algorithm suitable for processing a large number of features

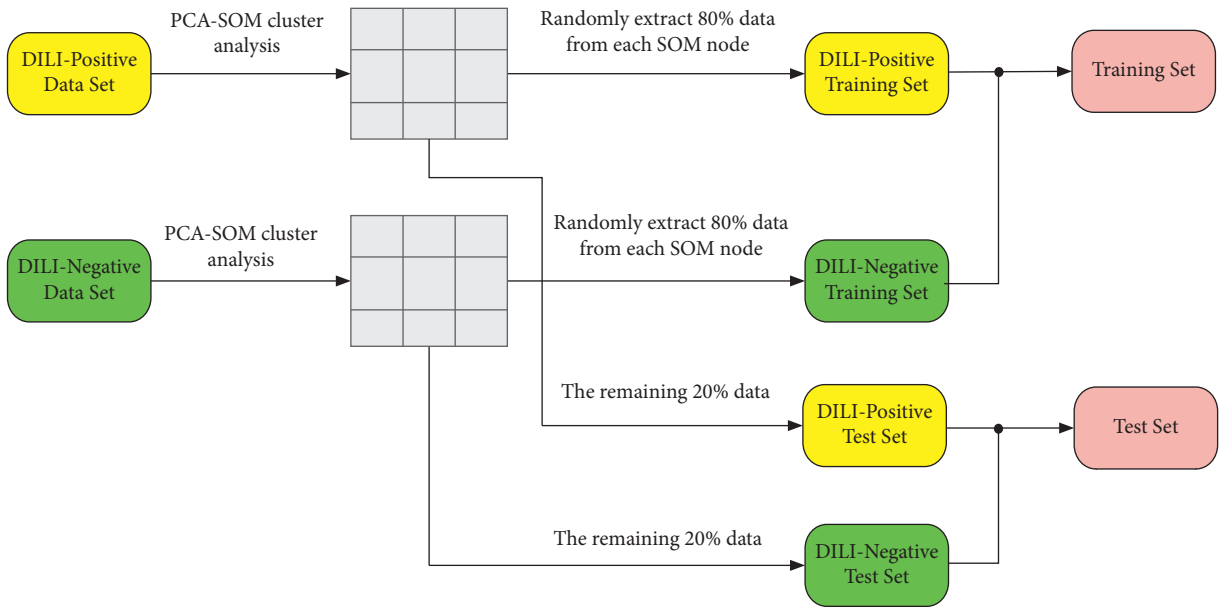

Figure 2: Flowchart of data splitting.

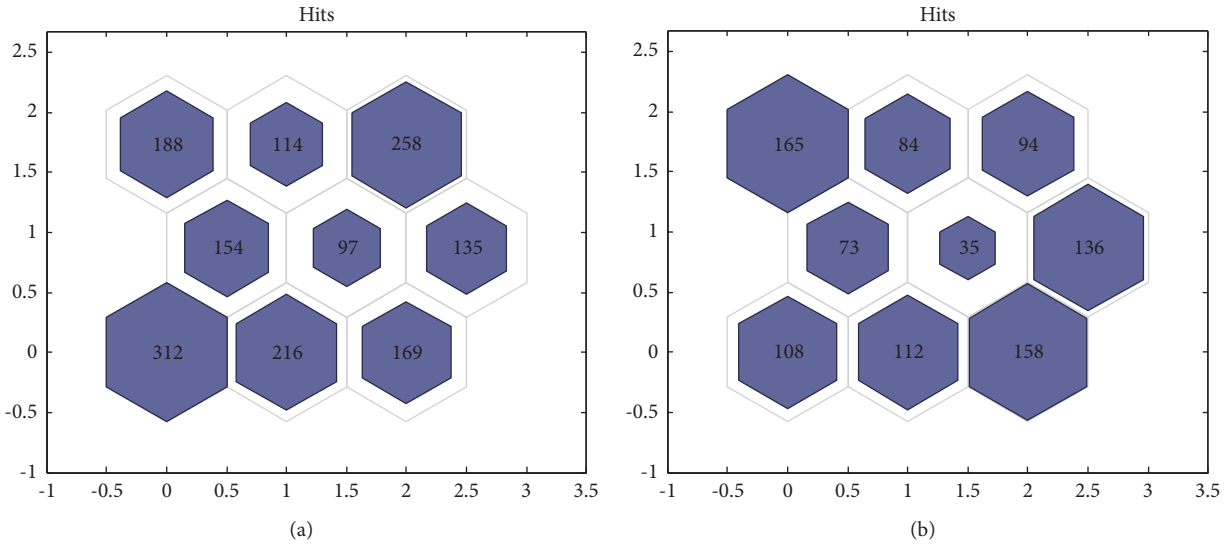

Figure 3: Clustering results of modeling data with the self-organizing mapping (SOM) neural network. (The number in each node represents the number of compounds contained in the node.) (a) Clustering results of DILI positives. (b) Clustering results of DILI negatives.

and is naturally good at processing classification tasks. SVM has been used for hepatotoxicity predictions in some publications and has achieved better performance than most other algorithms (such as k-nearest neighbor (kNN), naive Bayes (NB), decision tree, random forest (RF), AdaBoost decision trees, and artificial neural network (ANN)) [5, 21, 31, 33, 45, 49]. So, SVM was also employed to carry out our research in this study. When using descriptors to construct the (Q)SAR model, it is necessary to remove the useless descriptors and keep the important ones according to the importance of descriptors, so as to simplify the structure, avoid overfitting, improve the performance, and facilitate mechanistic interpretation [40, 43]. Molecular descriptors quantitatively characterize the physical and chemical properties, topological structure, and electronic properties of compounds. While molecular fingerprints do not use quantitative description, only 1 and 0 indicate the presence and absence of specific molecular fragments or substructures, so it is difficult to evaluate the importance of these fragments or substructures and delete some unimportant bits according to their importance. Therefore, the modeling approach using descriptors is different from that using molecular fingerprints, which will be described separately in subsequent sections.

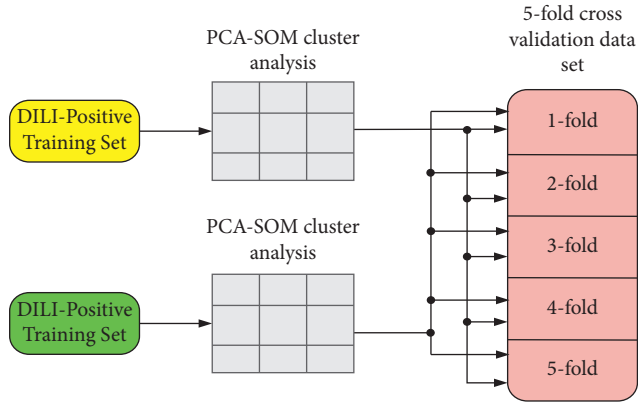

Figure 4: Flowchart of constructing a fivefold cross-validation set.

2.6.1. Modeling with PaDEL Descriptors

(1) Preliminary Selection of Descriptors. To delete the unimportant descriptors, we implemented the importance evaluation of descriptors based on the mean decrease impurity (MDI) method. The method has been successfully applied to evaluate the importance of descriptors in our previous study [43].

The basic idea is to scramble the values of each descriptor in turn and observe its influence on the model's accuracy. Variables that have a significant impact on the accuracy of the model are also important. The detailed implementation process of MDI method can be found in the literature [43]. In our study, SVM was employed to implement the classification model in MDI. We only changed the evaluation index *Ir (i)* to adapt to our DILI classification model. The changed *Ir* (*i*) is defined as

$$Ir(i) = \left| \frac{ACC_0^2 - ACC_1^2(i)}{ACC_0^2} \right|, \quad (6)$$

where ACC_0 is the accuracy of the established SVM model with original data and $ACC_1(i)$ is the accuracy of the model with the values of the ith descriptor being scrambled.

To avoid the influence of SVM parameters on the importance evaluation of the descriptors, we randomly selected 100 sets of different parameters for the SVM model ($C \in [10^6, 10^8]$ and $g \in [1, 100]$) and repeated the above evaluation method 100 times. The mean value of the 100 evaluations was used to judge the importance of each descriptor. After evaluating the importance of descriptors, the score of each descriptor is graphically shown in Figure 5.

As can be seen from Figure 5, when the number of descriptors is greater than 150, its importance score is less than 0.01. In other words, the impact of descriptors behind 150 descriptors on the accuracy of the model is less than 1%. In order to simplify the structure of the model to be constructed, we selected 150 relatively important descriptors for further research.

(2) PaDEL Submodels Based on Joint Optimization Strategy. An optimized submodel with carefully selected descriptors is more conducive to improving the performance of the consensus model. MDI is essentially a single factor analysis method. The obtained 150 descriptors may still have multicollinearity. The MDI method only removes most unnecessary descriptors but does not fully achieve the optimization of descriptors. Only the joint optimization of the descriptors and model parameters can guarantee that the most appropriate molecular descriptors are selected under the optimal model parameters [43]. In this paper, the HQPSO algorithm was employed to implement the joint optimization strategy. Herein, Gaussian radial basis function was selected as the kernel function of SVM to process the high-dimensional nonlinear modeling data, and the SVC_C-type [50] SVM classification model was implemented. In this case, the penalty factor C and the width of radial basis function g are the parameters that affect the performance of SVM.

When the parameters and input variables (descriptors) of SVM models are jointly optimized, the value of C (C is a big positive integer) ranges from 10^6 to 10^8, and the value of g ranges from 1 to 100. The 150 descriptors are coded into 15 integers between 0 and 1023 according to the coding method of our previous studies [43], and each integer represents 10 descriptors. A value of 0 (expressed in binary as "0000000000") means that no descriptor was selected, and 1023 (expressed in binary as "1111111111") means that the 10 descriptors were all selected [43]. Therefore, the vector to be optimized has 17 dimensions. The first two dimensions represent the parameters C and g, and the last 15 dimensions represent the descriptors.

The parameters C, g and descriptors were sought by the HQPSO algorithm based on fivefold cross-validation. The parameters of the HQPSO algorithm are set as follows: the population size is 30, the number of maximum iterations is 1000, and the internal parameters are $\lambda = 1$ and $L = 10$ (the values of λ and L are selected according to the study [42]). The fitness function of the HQPSO algorithm used for the optimization of SVM is defined as

$$\text{fitness} = \frac{1}{GMEN_0 \times GMEN_{cv5}}, \quad (7)$$

where $GMEN_0$ and $GMEN_{cv5}$ are defined by equation (5) and are the classification ability indexes of the SVM model on the

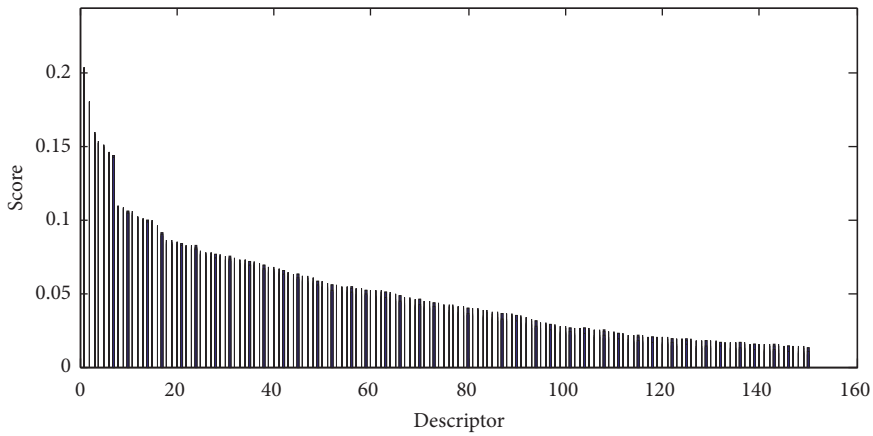

Figure 5: Scores of PaDEL descriptors calculated by the MDI method.

training set and fivefold cross-validation dataset, respectively. $GMEN_0$ and $GMEN_{cv5}$ will help to strike a balance between sensitivity and specificity of the SVM models trained on an imbalanced chemical dataset. At the same time, the fitness function can effectively avoid overfitting or underfitting, that is, the model has strong classification ability on the training set, but the fivefold cross-validation accuracy is poor, or the fivefold cross-validation accuracy is good, but the model has poor classification ability on the training set.

After joint optimization, 20 optimized SVM-based classification models for DILI prediction were obtained with the training set. The optimized parameters of each SVM model are listed in Table 1, and the performance of each SVM model on the training set, fivefold cross-validation set, and test set is also listed in Table 1, respectively.

Descriptors selected for model no. 1 are sorted in descending order according to the importance evaluated by the MDI method under the optimized model parameters and shown in Figure 6. The descriptors selected for other models are graphically shown in the Supplementary Materials (Figures S1–S19 in Supplementary.pdf).

Using structurally diverse (Q)SAR models as submodels to construct a consensus model will help to achieve a better performance [51]. It is easy to obtain structurally diverse (Q)SAR models with different parameters and descriptors. From Table 1, Figure 6, and Figures S1–S19, we can see that these models use different combinations of parameters and descriptors, and they have relatively good predictive performance. From Table 1, we can see that the accuracy of each model is all greater than 69% in the fivefold cross-validation set, greater than 71% in the test set, and greater than 92% in the training set. These models reached a relatively good balance between sensitivity and specificity in the training set, fivefold validation set, and test set, respectively. The relatively high accuracy of fivefold cross-validation indicates that these models are robust. In a word, the joint optimization strategy and data balancing method are helpful to obtain structurally diverse classification models with good stability and predictive performance.

2.6.2. Modeling with Fingerprints. In this paper, fingerprints are also employed to construct consensus models to implement a joint decision-making strategy. Due to the particularity characterization method of molecular fingerprints, it is difficult to evaluate the importance of substructures in fingerprints. Therefore, we cannot delete unimportant fingerprints according to the importance of each molecular fingerprint. We only delete the substructures that appeared less than 10 times or more than 2598 times in the modeling dataset. The joint optimization of the SubFP-, PubchemFP-, EstateFP-, and KRFP-based submodels is the same as that of the above models constructed by PaDEL descriptors.

In (Q)SAR modeling, the rule of thumb condition (that is, Topliss ratio [52]) requires that the chemical number over the number of selected variables should be at least 5 to avoid overfitting. KRFP has 4860 bits, and the number of substructures is obviously more than that of compounds. So, when the KRFP-based models are optimized, we limit the number of selected variables to no more than 400 to fulfill the Topliss ratio. 20 submodels for each kind of fingerprint were constructed, and the performance of optimized models is listed in the Supplementary Materials (Tables S1–S4 in Supplementary.pdf).

2.6.3. Consensus Modeling. The idea of consensus modeling is to integrate several weak learners into a strong learner, which can improve the stability and generalization performance of the (Q)SAR model to some extent [53]. In this study, five consensus models were integrated with submodels constructed by using descriptors and fingerprints, respectively. To fairly compare the performance of these consensus models, we set the threshold T to 50% according to the commonly used relative majority principle [33, 38]. That is to say, for each consensus model, if the number of DILI-positive outputs in its submodels is greater than that of DILI-negative outputs, we define the consensus output as DILI positive or otherwise DILI negative. The comparison results are listed in Table 2.

Table 1: Parameters and performance of the 20 PaDEL submodels.

No.	Parameters		Performance (%)											
			Training set				Fivefold cross-validation set				Test set			
	C	g	ACC	SE	SP	MCC	ACC	SE	SP	MCC	ACC	SE	SP	MCC
1	2.10E + 06	11.0	96.4	96.4	96.5	92.4	70.5	72.2	67.6	38.8	74.1	75.3	71.9	46.2
2	6.45E + 07	4.3	97.8	98.6	96.4	95.2	71.3	73.0	68.3	40.3	71.2	72.6	68.9	40.5
3	6.00E + 06	14.4	95.5	95.7	95.2	90.4	70.0	69.9	70.0	38.7	74.4	77.4	69.4	46.2
4	6.58E + 07	41.9	92.1	92.8	91.0	83.2	70.6	71.9	68.5	39.4	72.0	75.0	66.8	41.2
5	4.87E + 07	10.4	97.7	97.9	97.4	95.1	69.9	69.9	69.8	38.6	74.2	74.4	74.0	47.1
6	9.46E + 07	34.8	94.4	94.4	94.4	88.1	71.2	69.2	74.6	42.4	71.8	72.6	70.4	41.9
7	3.25E + 07	22.3	96.3	96.0	96.6	92.0	69.6	69.0	70.7	38.5	72.0	73.2	69.9	42.1
8	8.80E + 07	32.3	94.0	94.3	93.6	87.3	70.0	70.7	68.8	38.4	72.2	74.7	67.9	41.8
9	1.67E + 07	17.3	94.5	94.2	94.9	88.3	70.5	69.4	72.3	40.4	72.3	72.3	72.4	43.5
10	8.97E + 07	29.6	95.6	95.7	95.6	90.7	70.0	69.6	70.5	38.9	73.1	74.1	71.4	44.5
11	2.18E + 07	13.3	97.5	98.0	96.7	94.7	70.0	70.0	69.8	38.7	72.5	73.8	70.4	43.2
12	2.39E + 07	23.8	94.9	95.0	94.7	89.1	70.9	70.2	72.2	41.1	71.2	72.0	69.9	40.8
13	6.63E + 07	27.7	95.0	95.0	95.1	89.5	70.1	70.3	69.8	38.9	72.5	73.8	70.4	43.2
14	6.33E + 07	36.5	93.7	93.9	93.2	86.5	70.1	69.5	71.1	39.4	72.0	72.9	70.4	42.2
15	9.85E + 07	41.0	94.2	94.1	94.3	87.7	70.1	69.9	70.6	39.3	72.3	72.6	71.9	43.3
16	3.88E + 07	24.1	95.0	95.1	94.7	89.3	71.0	69.7	73.1	41.4	73.7	74.7	71.9	45.6
17	6.27E + 07	28.3	95.2	95.6	94.7	89.8	70.3	69.6	71.7	39.9	72.2	73.5	69.9	42.4
18	4.33E + 07	26.9	95.9	96.4	94.9	91.2	70.3	70.5	70.0	39.3	73.7	73.8	73.5	46.0
19	2.47E + 07	26.5	94.0	94.5	93.2	87.3	70.5	69.5	72.3	40.5	73.7	73.5	74.0	46.2
20	6.56E + 07	37.1	92.7	93.3	91.7	84.4	69.7	69.6	69.7	38.2	73.9	73.2	75.0	46.8

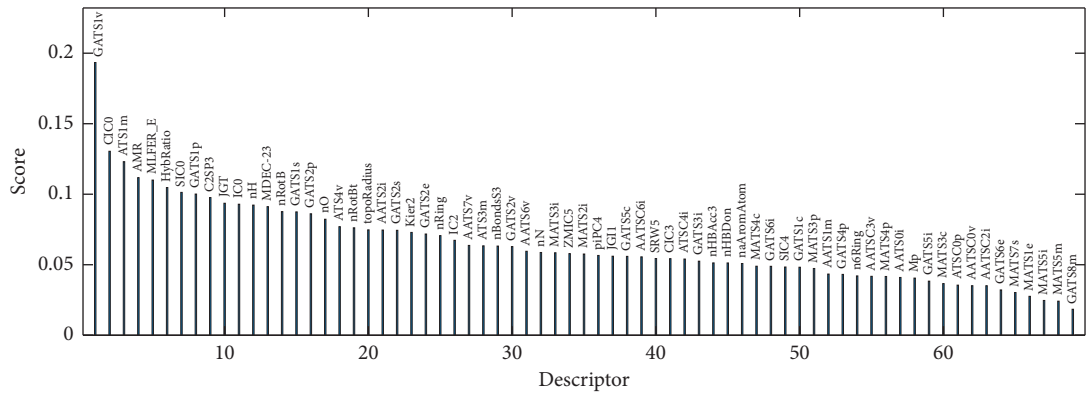

Figure 6: Descriptors selected for submodel no. 1 of the PaDEL consensus model.

Table 2: Comparison results of five consensus models.

Descriptor or fingerprints	Performance (%)												
	Training set				Fivefold cross validation set				Test set				
	ACC	SE	SP	MCC	ACC	SE	SP	MCC	ACC	SE	SP	MCC	AUC
PaDEL	96.2	96.1	96.2	91.8	72.1	71.4	73.2	43.3	75.4	74.1	77.6	50.2	81.8
PubchemFP	97.2	98.0	95.8	94.0	73.2	78.3	64.4	43.6	78.4	82.8	70.9	53.8	81.4
KRFP	97.4	98.1	96.1	94.3	70.8	77.0	60.1	37.2	78.4	83.7	69.4	53.5	82.9
SubFP	93.0	95.3	89.2	85.0	71.2	75.9	63.2	38.8	75.2	78.0	70.4	47.8	81.8
EstateFP	86.4	91.6	77.4	70.4	68.1	76.7	53.6	30.7	69.9	78.9	54.6	34.3	75.2

It can be seen from Table 1, Tables Tables S1–S4 in the Supplementary Materials, and Table 2 that the overall performance of these consensus models is all better than that of their submodels. Therefore, the consensus modeling strategy is beneficial to improving the performance of classification models. From Table 2, we can also see that the performance of consensus models constructed by different features is different in *ACC*, *SE*, *SP*, and *MCC*. The ROC curves on the test set in Figure 7 also show the difference. The values of the area under the ROC curve (AUC) are also listed in Table 2.

In Table 2, consensus modes based on PaDEL, PubchemFP, KRFP, and SubFP passed the criteria, i.e., $ACC_{cv5} > 0.7$. Therefore, the four models are used to integrate our joint decision-making model.

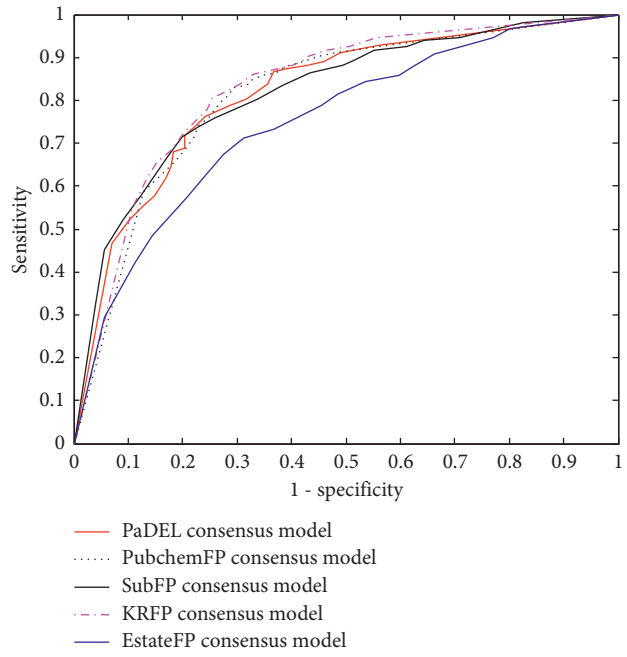

Figure 7: ROC curves of consensus models.

2.6.4. AD Analysis of Consensus Models. According to the OECD principle, a standard (Q)SAR model must give a defined domain of application. The AD indicates the applicable scope of a (Q)SAR model. The analysis of AD is a guarantee for (Q)SAR models in predicting newly synthesized compounds accurately and reasonably [54]. Therefore, the ADs of these consensus models were analyzed before constructing the joint decision-making model.

(1) AD of PaDEL Consensus Model. In the PaDEL consensus model, the frequency of each descriptor used in the submodels is different, which indicates that different descriptors have different contributions to the output of the model. So, the frequency-weighted and distance (FWD)-based method proposed in our previous study [43] was employed to define the AD of the consensus model. The essential difference between the FWD method and the traditional leverage method [55] is that it considers the importance of descriptors, and it can more reasonably define the AD of the consensus model. The detailed information of definition and implementation steps of the FWD method can be found in [43]. Figure 8 shows the AD of the PaDEL consensus model defined by the FWD method.

In Figure 8, black circles represent compounds in the training set and blue crosses represent compounds in the test set. The vertical red line represents a warning $h* = 3.0776$. The compounds outside the AD are plotted on the right side of the vertical red line, and the black and blue numbers are the serial number of the compounds in the training set and test set, respectively.

(2) *AD of Fingerprint-Based Consensus Model.* Fingerprints are different from descriptors. The FWD method and traditional leverage method are not suitable for fingerprints. In our study, the AD defined method based on the Tanimoto similarity index [39] is proposed. The definition and implementation steps of the method are as follows:

Step 1: calculate the mean value and standard deviation of Tanimoto similarity index of the training set characterized by fingerprints and record them as T_{av} and δ.

Step 2: for the ith ($i = 1,2,3, \ldots, N$, where N is the number of compounds in the training set) compound in the training set, calculate its Tanimoto similarity index $T_{tr}(i,j)$ ($i \neq j$) with all other compounds and record the mean value as $T_{trav}(i)$. If $T_{trav}(i) < T_{\text{av}} - 3\delta$, consider the compound is outside the AD. Otherwise, it is inside the AD. For the ith ($i = 1,2,3, \ldots, N_1$, where N_1 is the number of compounds in the test set) compound in the test set, calculate its Tanimoto similarity index $T_{\text{te}}(i)$ with each compound in the training set and record the mean value as $T_{\text{teav}}(i)$. If $T_{teav}(i) < T_{\text{av}} - 3\delta$, consider the compound is outside the AD. Otherwise, it is inside the AD.

Figure 9 shows the AD of the PubchemFP consensus model.

In Figure 9, black circles represent compounds in the training set and blue crosses represent compounds in the test set. The vertical red line represents a warning $h* = 0.91722$ ($h* = 1-(T_{\text{av}} - 3\delta)$). The compounds outside the AD are plotted on the right side of the vertical red line, and the black and blue numbers are the serial number of the compounds in the training set and test set, respectively. The AD plot of the SubFP and KRFP consensus models is graphically shown in the Supplementary Materials (Figures S20 and S21 in Supplementary.pdf).

The AD coverage of the PaDEL consensus model on training and test sets is 0.9731 and 0.9716, and the AD coverage of the PubchemFP consensus model on training and test sets is 0.9817 and 0.9811, respectively. The AD coverage of the KRFP consensus model on training and test sets is 0.9817 and 0.9867, and the AD coverage of the SubFP consensus model on training and test sets is 0.9880 and 0.9867, respectively. That is to say, for each consensus model, most of the compounds are inside its AD. So, these consensus models have a wide AD.

As can be seen from Figures 8, 9 and Figures S20, S21, test compounds 225 (ethylene dichloride), 226 (chloroform), 232 (ethylene dibromide), 254 (iodoform), and 335 (perfluoroether) are outside the AD in at least three consensus models. The structures of these compounds are shown in Figure 10.

As can be seen from Figure 10, the molecular structure of these compounds is relatively simple, and all contain halogen elements. They have poor structural similarity to most compounds in the training set. This indicates that the above AD defined methods can detect structurally abnormal compounds from the dataset effectively, although different features and AD definition methods are used. So, the AD defined methods based on the FWD method and Tanimoto similarity index are suitable for the above models.

2.6.5. Joint Decision-Making Model Based on Overwhelming Majority Principle. As is known to all, in both business decisions and political elections, the adoption of

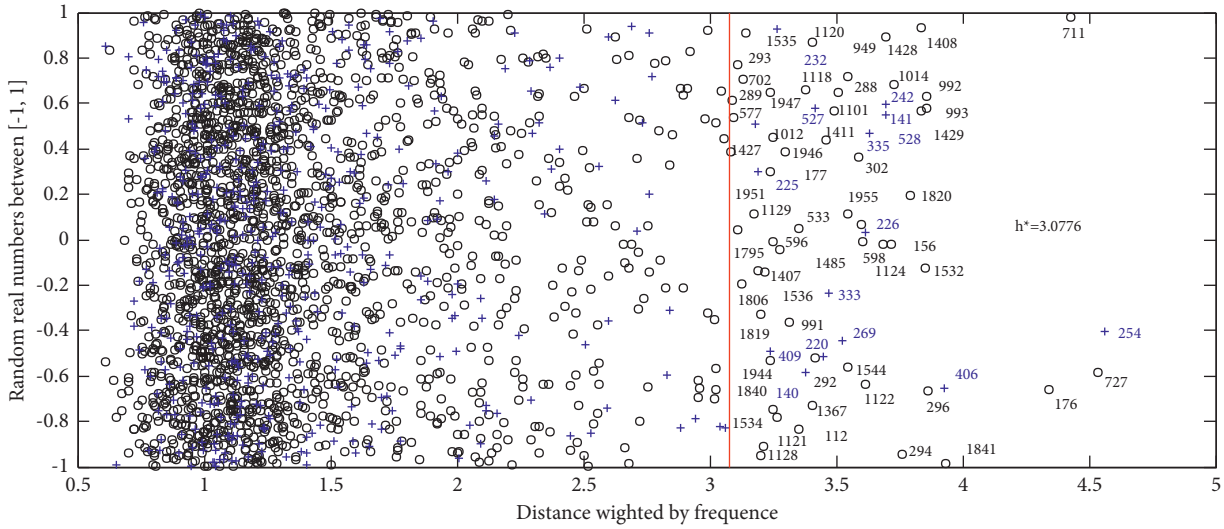

Figure 8: Application domain (AD) plot of the PaDEL consensus model.

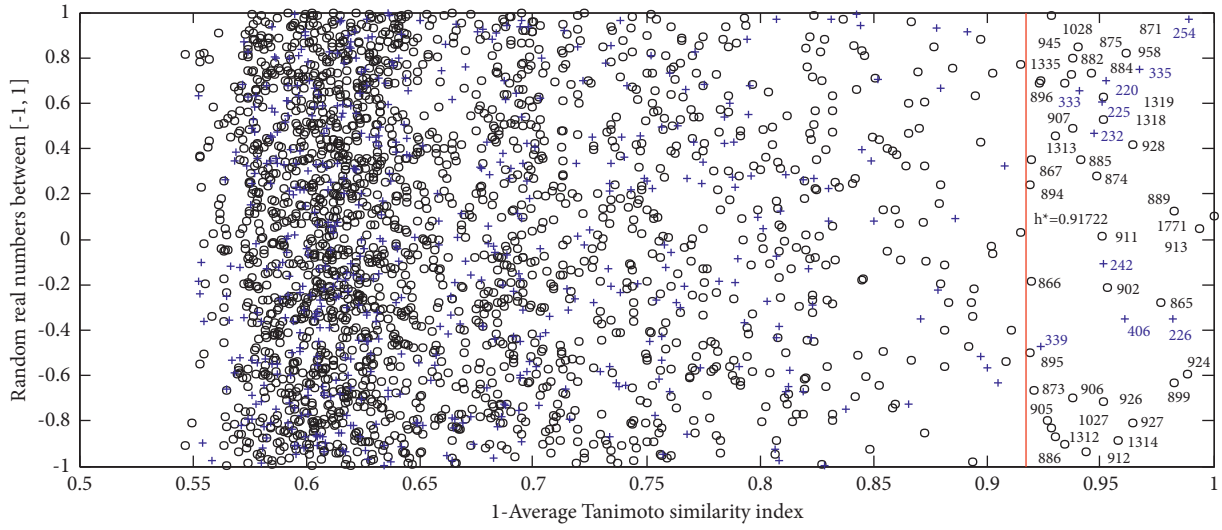

Figure 9: Application domain (AD) plot of the PubchemFP consensus model.

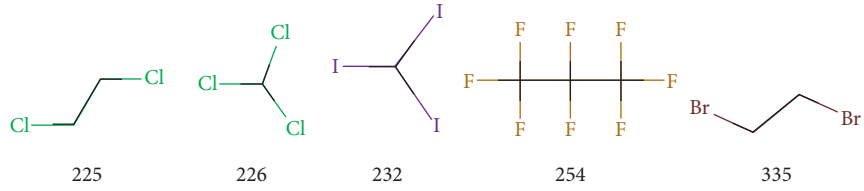

Figure 10: Molecular structures of the compounds are outside the AD in at least three consensus models.

overwhelming majority principle is all conducive to improving the credibility of decision-making, although sometimes the decision-making results may not be obtained. That is to say, a decision that cannot be passed by the majority is invalid. To improve the credibility and accuracy of the prediction results of these consensus models, we adopt the overwhelming majority principle instead of the commonly used relative majority principle (threshold $T = 50\%$) to determine the output of consensus models. To solve the problem of invalid decisions, the output of the consensus model is defined as DILI positive, DILI negative, and DILI uncertain. We set the values of threshold T to 70%, 80%, and 90%, respectively, to reveal the influence of overwhelming majority principle on the performance of

consensus models. For a compound, if T or more of the predicted results are DILI positive or DILI negative, we regard the prediction results of the consensus model as DILI positive or DILI negative or otherwise DILI uncertain. The performance of consensus models on the test set that increases with T is listed in Table 3.

It can be seen from Table 3 that, with the increase of T, the performance of each consensus model has obviously improved, but the coverage (ratio of compounds that can give definite prediction results) is significantly reduced. In other words, the performance of the model is improved, but the application scope is narrowed. In this paper, the application scope narrowing caused by the overwhelming majority principle will be solved by the joint decision-making strategy. First, the four qualified consensus models are sorted in descending order according to the values of $ACC_{cv5}(T = 50\%)$, and then they are concatenated to form a joint decision-making model. For a compound to be predicted, the working flow of the joint decision-making model is shown in Figure 11.

As shown in Figure 11, for a compound to be predicted, if the PubchemFP consensus model cannot give a clear prediction result, the compound will enter the PaDEL consensus model for prediction, and so on. The final output of the joint decision-making model is DILI positive, DILI negative, or DILI uncertain.

3. Results and Discussion

3.1. Diversity Analysis of Modeling Dataset. In practical application, the high diversity of the modeling data is helpful to obtain a (Q)SAR model with a wide AD. To illustrate the diversity of our dataset, we applied the radar chart to explore the chemical space of the 2608 compounds. The PCA method was employed to transform the modeling data characterized by 150 selected PaDEL descriptors, and then the first 10 principal components were used to draw the radar map. The radar chart is shown in Figure 12. It can be seen from Figure 12 that the 2608 compounds in the modeling dataset covered a sufficiently large chemical space.

To further explore the chemical diversity of the dataset, the Tanimoto similarity index of the modeling dataset was also calculated using PubchemFP, SubFP, and KRFP, respectively. The heat map of the Tanimoto similarity index based on PubchemFP is shown in Figure 13. The average Tanimoto similarity index calculated from PubchemFP, SubFP, and KRFP is 0.335, 0.372, and 0.165, respectively. A lower Tanimoto similarity index means a more diverse dataset [56]. So, the chemical diversity of our modeling set is significant.

In this study, the chemical space was analyzed using PCA to check the quality of data splitting on training and test sets [57]. As shown in the PCA plot of the compounds based on 150 selected PaDEL descriptors (Figure 14), the compounds in the test set were basically distributed within the chemical space of the training set.

Therefore, the use of the chemically diverse modeling set as well as the balanced training set and the test set in this study is conducive to the construction of DILI models with a wide AD and the achievement of fair test performance.

3.2. Performance Analysis of Joint Decision-Making Model. In order to reveal the ability of the joint decision-making model in solving the problem of application scope narrowing caused by the overwhelming majority principle, and to obtain the performance of the joint decision-making model changing with the threshold T, we set T as 70%, 80%, and 90%, respectively. The performance of the joint decision-making model on the test set is listed in Table 4. As can be seen from Table 4, when T is 70%, the clear prediction results can be obtained for 99.24% of the compounds in the test set, and the ACC of the predicted results is 80.0%.

From Tables 2, 3, and 4, we can see that ACC with $T = 70\%$ is obviously better than the consensus model with $T = 50\%$. The AD with $T = 70\%$ is obviously better than consensus models with $T = 70\%$ and even better than consensus models with $T = 50\%$. Therefore, the combination of overwhelming majority principle and joint decision-making strategy can effectively improve the performance of the consensus models. However, with the increase of T, the accuracy of the joint decision-making model does not improve significantly, but the coverage of the model is significantly reduced. Therefore, we believe that $T = 70\%$ is appropriate in this study.

3.3. Application to External Validation Set. To further test the performance of the joint decision-making model, an independent external validation set containing 390 compounds was collected from DILIrank (the detailed information of the 390 compounds is listed in the Supplementary Materials: External_validation_Set.pdf). The DILIrank dataset was developed by Chen et al. [58], and it includes 1036 marketed drugs approved by the US FDA before 2010. In the dataset, the DILI risk of the drugs was marked as ^{v}No-DILI concern, ^{v}Less-DILI concern, ^{v}Most-DILI concern, and ^{v}Ambiguous-DILI concern. First, we removed the compounds overlapping with our training set or classified as ^{v}Ambiguous-DILI concern. The ^{v}Most-DILI concern and ^{v}Less-DILI concern compounds were marked as DILI positive, and ^{v}No-DILI concern compounds were marked as DILI negative. Next, we screened the remaining 390 compounds (172 DILI positives/218 DILI negatives) using the joint decision-making model. Finally, for the independent external dataset, the clear prediction results are obtained for 98.20% of the compounds. The joint decision-making model achieves a prediction accuracy with $ACC = 79.9\%$, $SE = 97.1\%$, $SP = 66.0\%$, and $MCC = 64.7\%$.

The performance on the external validation dataset indicates that the model we constructed has good performance and a wide AD.

3.4. Contributions of Selected PaDEL Descriptors to DILI. In order to reveal the association between PaDEL descriptors and DILI, we calculated the mean value of each descriptor in

Table 3: Performance of consensus models on the test set that increases with T.

Name	Performance (%)											
	$T = 70\%$				$T = 80\%$				$T = 90\%$			
	ACC	SE	SP	Coverage	ACC	SE	SP	Coverage	ACC	SE	SP	Coverage
PaDEL	79.6	80.7	77.8	83.7	81.3	83.0	78.4	75.8	82.6	83.6	80.9	66.5
PubchemFP	79.8	84.6	71.5	91.9	81.2	86.0	72.7	85.4	82.2	87.2	73.5	76.7
SubFP	81.2	85.9	71.7	78.4	83.7	88.2	74.6	67.4	86.5	89.6	79.6	53.2
KRFP	81.2	86.9	70.9	87.7	82.9	88.0	73.7	80.7	84.8	88.6	77.8	72.2

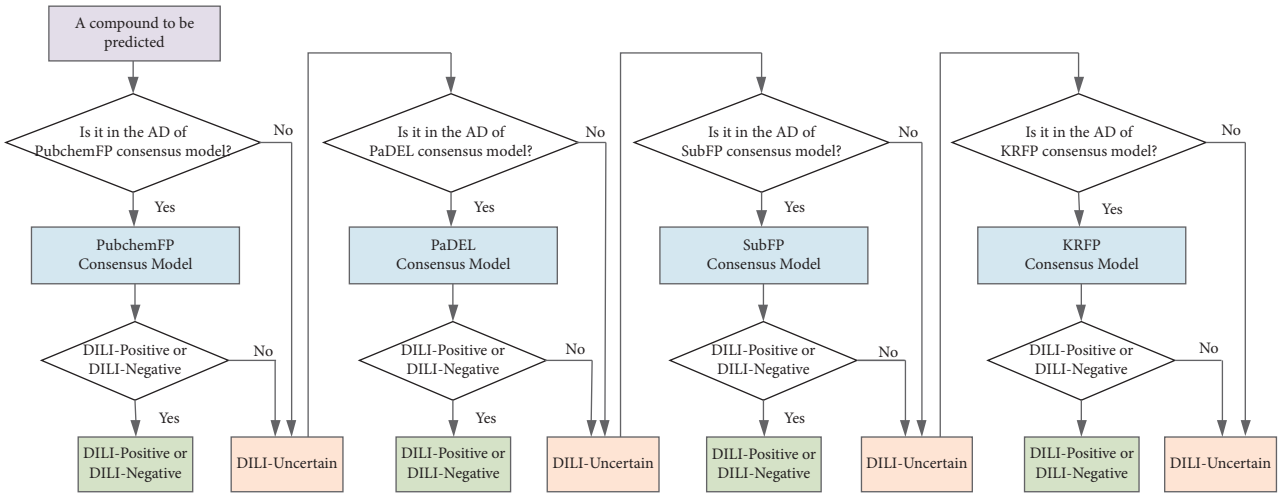

Figure 11: Working flow of the joint decision-making model for a compound to be predicted.

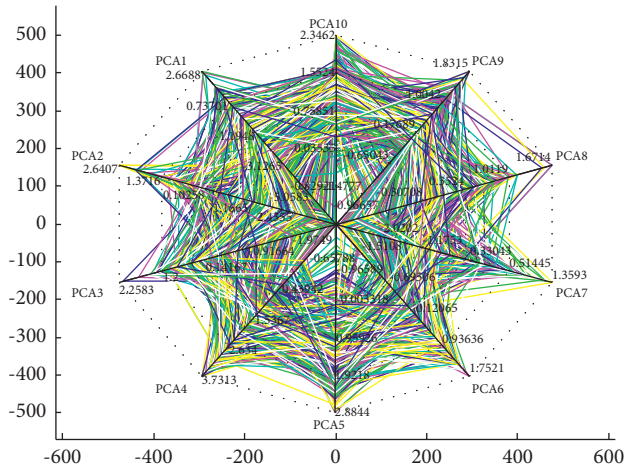

Figure 12: Radar chart of the first 10 principal components for the modeling dataset. Each color line represents a compound.

DILI positives and DILI negatives, respectively, and defined the difference index as

$$\mathrm{DI}(i) = \left| \frac{V_{PM}(i) - V_{NM}(i)}{V_M(i)} \right|, \tag{8}$$

where DI(i) is the difference index of the ith descriptor and $V_{PM}(i)$, $V_{NM}(i)$, and $V_M(i)$ are the average value of the ith descriptor in DILI positives, DILI negatives, and the whole modeling dataset, respectively.

The top 10 descriptors with a significant difference between DILI positives and DILI negatives are listed in Table 5.

It can be seen from Table 5 that compounds with higher values of ALogP, MATS2c, MATS7s, AATSC6v, and ATSC4i are more likely to be hepatotoxic positive. Compounds with lower values of MATS3c, ATSC1p, MATS2i, AATSC2v, and AATSC1p are more likely to be hepatotoxic positive. Among these descriptors, only MATS3c, ALogP, and ATSC1p had significant differences between hepatotoxic-positive and -negative compounds (their DI values are greater than 1). Among them, MATS3c is related to the charge state of compound molecules, and the difference between hepatotoxic-negative and -positive compounds is the most significant. ALogP represents the lipid water partition coefficient of the compound. The greater its value, the stronger the lipophilicity of the compound and the easier it is to present hepatotoxicity. ATSC1p is related to the polarizability of compound molecules. The stronger the polarizability, the greater the toxicity of molecules.

We plot the values of "MATS3c" and "ALogP" in DILI positives and DILI negatives in Figures 15(a) and 15(b), respectively.

Although for the two descriptors, the average value of DILI positives and DILI negatives is significantly different, there is still no clear boundary between them. This indicates that these descriptors have a weak differentiating effect on DILI potential and the DILI potential cannot be simply classified by individual or several simple descriptors. It is well known that DILI has complicated mechanisms, and

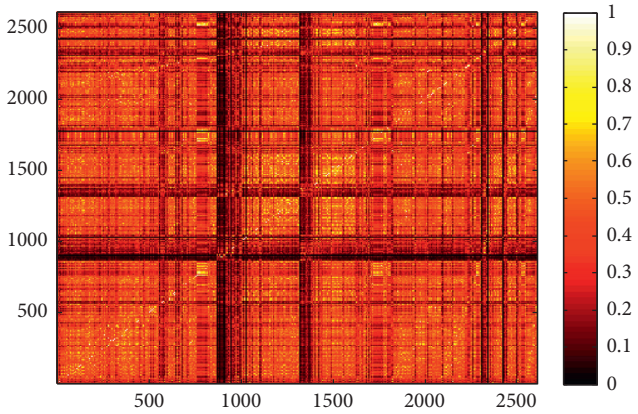

Figure 13: Heat map of molecular similarity plotted by the Tanimoto similarity index based on PubchemFP. (The x-axis and y-axis represent the 2608 compounds.).

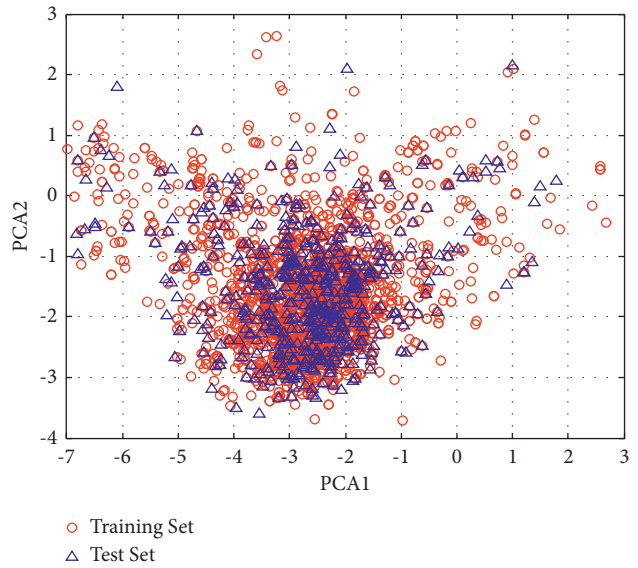

Figure 14: Chemical space analyzed using principal component analysis (PCA) technology.

Table 4: Performance of the joint decision-making model changing with the increase of T.

Threshold (%)	Performance(%)				
	ACC	SE	SP	MCC	Coverage
70	80.0	83.9	73.3	57.2	99.24
80	80.3	85.0	72.3	57.6	96.02
90	80.8	84.8	73.9	58.6	90.5

most mechanisms involve chemical reactions with proteins and other macromolecules in liver cells [44]. Therefore, it is necessary to adopt the nonlinear modeling method to establish the DILI model.

In order to further reveal the relationship between important descriptors and hepatotoxicity, we plot the first 50 important descriptors and their importance scores among the 150 descriptors used in the PaDEL models in Figure 16. From Figure 16, we made the following inferences:

(1) "Mv" (mean atomic van der Waals volumes (scaled on a carbon atom)) is the most important descriptor that affects the performance of the model. This shows that van der Waals volumes will affect the hepatotoxicity of the compound.

(2) "AMR" is the second important descriptor that affects the performance of the model. "AMR" is molar refractivity. AMR can be used as a measure of electron polarizability in molecules. Previous studies have shown that [15, 17], for aquatic organisms, the stronger the polarizability, the greater the toxicity of the molecules. As an important descriptor, "MP" is also related to polarizability.

(3) "MLogP" is a relatively important descriptor in the model, which predicts the oil-water partition coefficient of compounds. The greater the MLogP value, the stronger the fat solubility of the compound and the more likely it is to cause hepatotoxicity.

Table 5: Top 10 descriptors with a significant difference between DILI positives and DILI negatives.

No.	Name	Description	V_{PM}	V_{NM}	DI
1	MATS3c	Moran autocorrelation—lag 3/weighted by charges	−3.72e-3	6.88e-3	52.32
2	ALogP	Ghose–Crippen LogKow	3.74e-3	−0.435	2.77
3	ATSC1p	Centered Broto–Moreau autocorrelation—lag 1/weighted by polarizabilities	−0.068	8.86e-3	1.94
4	MATS2c	Moran autocorrelation—lag 2/weighted by charges	0.021	6.62e-3	0.92
5	MATS2i	Moran autocorrelation—lag 2/weighted by first ionization potential	0.021	0.037	0.60
6	MATS7s	Moran autocorrelation—lag 7/weighted by I-state	−0.035	−0.06	0.56
7	AATSC6v	Average centered Broto–Moreau autocorrelation—lag 6/weighted by van der Waals volumes	−0.954	−1.474	0.45
8	AATSC2v	Average centered Broto–Moreau autocorrelation—lag 2/weighted by van der Waals volumes	1.262	1.900	0.43
9	ATSC4i	Average centered Broto–Moreau autocorrelation—lag 4/weighted by first ionization potential	−9.774	−13.831	0.36
10	AATSC1p	Average centered Broto–Moreau autocorrelation—lag 1/weighted by polarizabilities	−2.00e-3	−1.51e-3	0.32

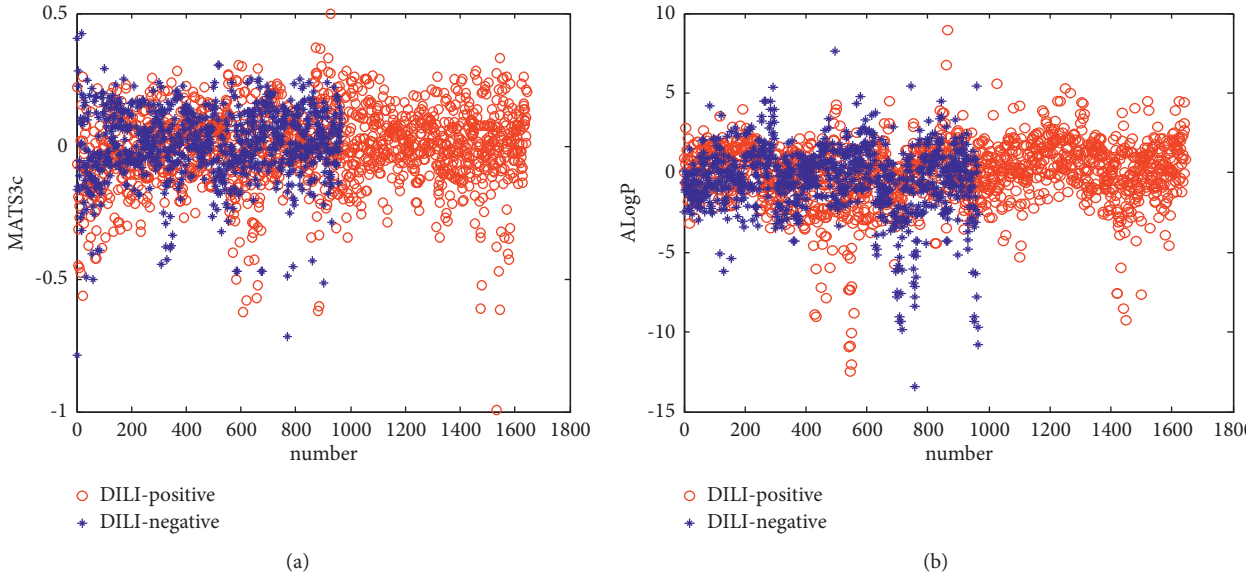

Figure 15: Values of descriptors in DILI positives and DILI negatives. (a) Values of MATS3c. (b) Values of ALogP.

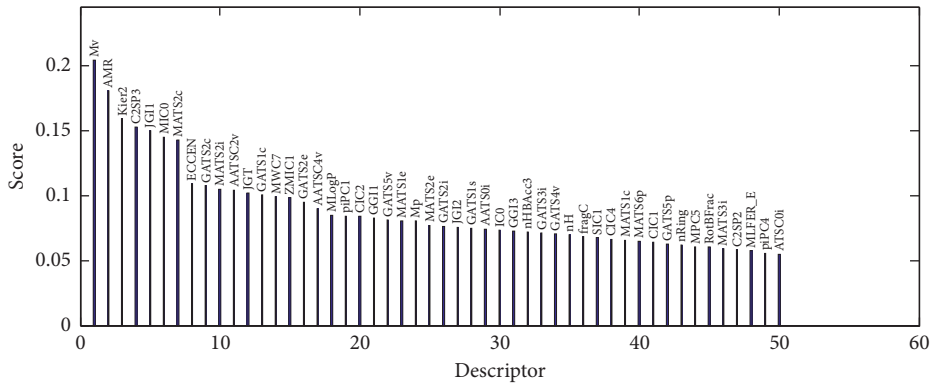

Figure 16: The first 50 important descriptors and their importance scores used in the PaDEL models.

(4) "MATS2i" is also a relatively important descriptor. It is related to the first ionization potential. The first ionization potential is the energy required for a gaseous atom in the ground state to lose one electron in its outermost layer. The larger the initial ionization energy, the harder it is for an atom to lose an electron

and the less likely the compound will both react and produce toxicity.

3.5. Structural Alert Identification. To reveal the chemical structures related to DILI, several privileged substructures responsible for DILI positives were identified by frequency analysis [39] from PubchemFP, SubFP, EstateFP, and KRFP. The DILI-positive and DILI-negative frequencies of each substructure in the modeling dataset were calculated by

$$F_{DP}(i) = \frac{N_{SP}(i) \times N_T}{\left(N_{SP}(i) + N_{SN}(i)\right) \times N_P}, \quad (9)$$

$$F_{DN}(i) = \frac{N_{SN}(i) \times N_T}{\left(N_{SP}(i) + N_{SN}(i)\right) \times N_N}, \quad (10)$$

where $F_{DP}(i)$ and $F_{DN}(i)$ are the frequency of the ith substructure in DILI positives and DILI negatives, respectively; $N_{SP}(i)$ and $N_{SN}(i)$ are the number of DILI positives and DILI negatives containing the ith substructure, respectively; N_T is the number of compounds in the modeling dataset; and N_P and N_N are the number of DILI positives and DILI negatives in the modeling dataset, respectively.

A substructure appearing in DILI positives far more frequently than DILI negatives could be considered a privileged substructure responsible for DILI positives. In this study, only the substructures present in 10 or more DILI positives and $F_{DN}/F_{DP}<0.2$ were analyzed. Finally, 25 substructures were identified as privileged substructures and are listed in Table 6.

The detailed information of these privileged substructures is listed in the Supplementary Materials (Table S5 in Supplementary.pdf). The chemical structures of privileged substructures or representative drugs containing privileged substructures listed in Table 6 are graphically shown in Figure 17. All the representative drugs containing privileged substructures listed in Table 6 have certain hepatotoxicity. Halothane (No. 1) is a general inhalation anesthetic, and it is forbidden for patients with liver dysfunction and biliary tract diseases. Nimodipine (No. 7) is a peripheral vasodilator and hypotensive. It can lead to hepatitis and jaundice in individual patients and elevated alkaline phosphatase and alanine aminotransferase. Trovafloxacin (No. 12), a fluoroquinolone antibiotic, has high anti-Gram-positive bacteria activity, especially against *Streptococcus pneumoniae*. It has significant hepatotoxicity, which can lead to elevated transaminase, hepatitis symptoms, severe liver damage, and even liver failure. Due to severe hepatotoxicity, it has been withdrawn from the US market. Fluvastatin (No. 14) is the first synthetic cholesterol-lowering drug. It can lead to continuous increase of alanine aminotransferase (ALT) or aspartate aminotransferase (AST). Minocycline (No. 16) is a kind of tetracycline antibiotic with broad-spectrum antibacterial property. It can cause nausea, vomiting, jaundice, fatty liver, elevated serum aminotransferase, hematemesis, hematochezia, etc. In severe cases, it can cause coma and death. Pazufloxacin (No. 17) is a fluoroquinolone antibiotic. It can cause liver dysfunction and jaundice. Mithramycin (No. 24) is the only indication for testicular embryonal cell carcinoma, but it has considerable toxicity. The liver and kidney function damage is more prominent, and the liver and kidney function should be checked frequently in the treatment. Information about these drugs can be found in the FDA database (https://www.fda.gov/Drugs) or online drug database (https://www.druges.com).

Form Table 6 and Figure 17, we can see that fluorine- or bromine-containing groups and amine or nitro derivatives were identified as privileged substructures. Some fluorine-containing groups and amine or nitro derivatives were also identified as privileged substructures in the literature [5, 39]. Although different datasets and identification indexes were used to identify the privileged substructures in this study and the literature [5], 10 identical KRFP substructures (marked with $*$ in Table 5) were simultaneously identified as privileged substructures. This indicates that our privileged substructure identification method is credible. To a large extent, these 10 privileged substructures can be regarded as structural alerts responsible for the prediction of DILI-positive compounds in drug discovery, and other newly discovery privileged substructures listed in Table 6 should also be alert in DILI prediction.

In these privileged substructures, six substructures are fluorine-containing groups. The carbon-fluorine bond is commonly present in about one-fifth of pharmaceuticals, and it is metabolically stable in general. Fluorine acts as a bioisostere of the hydrogen atom in pharmaceuticals, and the lipophilicity of drugs would increase extremely with the introduction of fluorine atoms, thus enriching the intracellular concentration of hepatotoxic drugs [39]. Three substructures are bromine-containing groups. Bromine and fluorine are halogen functional groups with similar chemical properties. Specific fragment structures containing bromine atoms may also increase the risk of hepatotoxicity. 15 substructures are amine or nitro derivatives, and this class of chemicals could bind to proteins through covalent bonds with cysteine residues via Michael addition reactions and result in hepatotoxicity [59].

3.6. Comparison with Other Published Models. To better illustrate the superiority of the joint decision-making model, we developed a detailed comparative analysis between the proposed model and the main published DILI models. Although many models have been developed for hepatotoxicity prediction, our interest is only focused on models constructed with a dataset larger than half of our dataset ($N_m > 1304$) because models with small datasets make it difficult to collect sufficient and diverse molecular information to develop a superior model for practical application. The results of this comparison can be found in Table 7.

As can be seen from Table 7, different models use different types and quantities of descriptors. However, most models use descriptors related to lipophilicity and molecular polarizability, which shows that there is a great relationship between them and hepatotoxicity.

The main purpose of comparative analysis does not aim to rank the performance of these models but to give readers a

Table 6: Privileged substructures responsible for DILI positives.

No.	Type	Bit	SMARTS	Representative drugs
1	SubFP	10	[BrX1][CX4]	Halothane
2		62	[NX3v3,SX2,OX2;!$(∗ C=[#7,#8,#15,#16])][CX4;!$(C([N,S,O]) ([N,S,O])[!#6])][FX1,ClX1,BrX1,IX1]	Desflurane
3		69	[FX1][CX3]=[CX3]	Trovafloxacin
4		197	[NX3;$([H2]),$([H1][#6]),$([H0]([#6])[#6]);!$(NC=[O,N,S])][NX3;$([H2]),$([H1][#6]),$([H0]([#6])[#6]);!$(NC=[O,N,S])]	Hydralazine
5	PubchemFP	329	C(~Br)(~H)	Bromisoval
6		648	O=N-C:C-N	Cinitapride
7		40	[!#1][CH]([!#1])c1[cH][cH][cH]c([cH]1)[N+](=O)[O-]	Nimodipine
8		748	[!#1][NH][CH]=[CH][!#1]	Cefpirome
9		1575	[!#1]c1[cH][cH][cH][cH]c1C(=O)[OH]	Flufenamic acid
10		1597	[!#1]c1[cH][cH][cH]c([cH]1)[N+](=O)[O-]	Benidipine
11		1756	[!#1]c1[cH][cH]c([CH3])c([!#1])[cH]1	Gemfibrozil
12	KRFP	1799 ∗	[!#1]c1[cH][cH]c(F)[cH]c1F	Trovafloxacin
13		3182 ∗	c1cnoc1	Flucloxacillin
14		3524	Cc1cc2ccccc2n1C	Fluvastatin
15		3953 ∗	CNNC(=O)C	Phenylbutazone
16		4232 ∗	NC=C1C(=O)CCCC1=O	Minocycline
17		4252 ∗	Nc1ccc(F)cc1	Pazufloxacin
18		4387	O=C1Cc2ccccc2N1	Sunitinib
19		4556	O=CNCCCCNC=O	Thalidomide
20		4651 ∗	OC(=O)C1CCCN1	Fosinopril
21		4689 ∗	OC(CC=C)CC=C	Doxycycline
22		4692 ∗	OC(F)F	Isoflurane
23		4708	Oc1cc(O)cc(O)c1	Griseofulvin
24		4778 ∗	OCC(=O)c1ccccc1	Mithramycin
25		4808 ∗	OCC(F)(F)F	Silodosin

∗ The substructures were also identified as privileged substructures in [5].

general overview of the advantages of the joint decision-making model because all the previous models were constructed under different conditions (dataset, algorithms, features, and so on).

First, we compare and analyze whether these models conform to OECD principles. The widely accepted OECD principles for ensuring the effectiveness of (Q)SAR model state that a compliant (Q)SAR model must meet the following five conditions: (1) a defined endpoint, (2) an unambiguous algorithm, (3) a definite AD of applicability, (4) appropriate measures of goodness-of-fit, robustness, and predictability of the model, and (5) a reasonable mechanism explanation, if possible [60]. In Table 6, the study [40] and our model are consistent with the principles of OECD, but there is no mechanism explanation in [40]. Among these models, only our model gives specific model parameters, which makes the modeling process more transparent. There is no cross-validation in [9, 31, 39]. Although cross-validation was carried out in [5] for determining the parameters of their model, the accuracy of cross-validation is not given clearly. So, the robustness of these models cannot be well explained. Studies [5, 9, 31, 39] do not give clear ADs, which limits the practical application of the model.

Second, we compare the modeling dataset of these models. Compared with the modeling dataset, our dataset contains 2608 compounds, which is more than that in other models in Table 6. Although 3712 compounds were used in [31], the largest model constructed in [31] only used 2171 compounds. We confirmed the diversity of our modeling dataset through radar chart and Tanimoto similarity index, and the AD coverage in test and external validation sets shows that the model constructed by a diverse dataset really obtained a wide AD. The modeling set was dealt with the PCA-SOM method to obtain diverse and balanced training and test sets. Therefore, fairer test accuracy can be obtained in our model. For other models in Table 6, the study [39] carried out data diversity analysis, but it did not balance the training and test sets with any technical means. In [39], the specificity on the test set and external validation set is 34.38% and 37.93%, respectively. Such specificity makes it difficult for practical application, although the model has good sensitivity.

Finally, we compare the performance of these models. The study [5] and joint decision-making model proposed in this paper have good performance on the test set and external validation set, and the performance of other models is obviously worse than them. Although the study [40] has a good sensitivity on the external validation set (88.9%), the sensitivity comes from an external validation set of only 18 compounds, and the specificity is not given. Meanwhile, a serious imbalance between sensitivity ($39.2 \pm 2.6\%$) and specificity ($87.1 \pm 2.6\%$) of cross-validation indicates that the model developed in [40] is difficult to be applied in practice. Although the study [5] has better prediction ability on test and external validation sets than our model, the accuracy of cross-validation, AD coverage, and parameters of the model

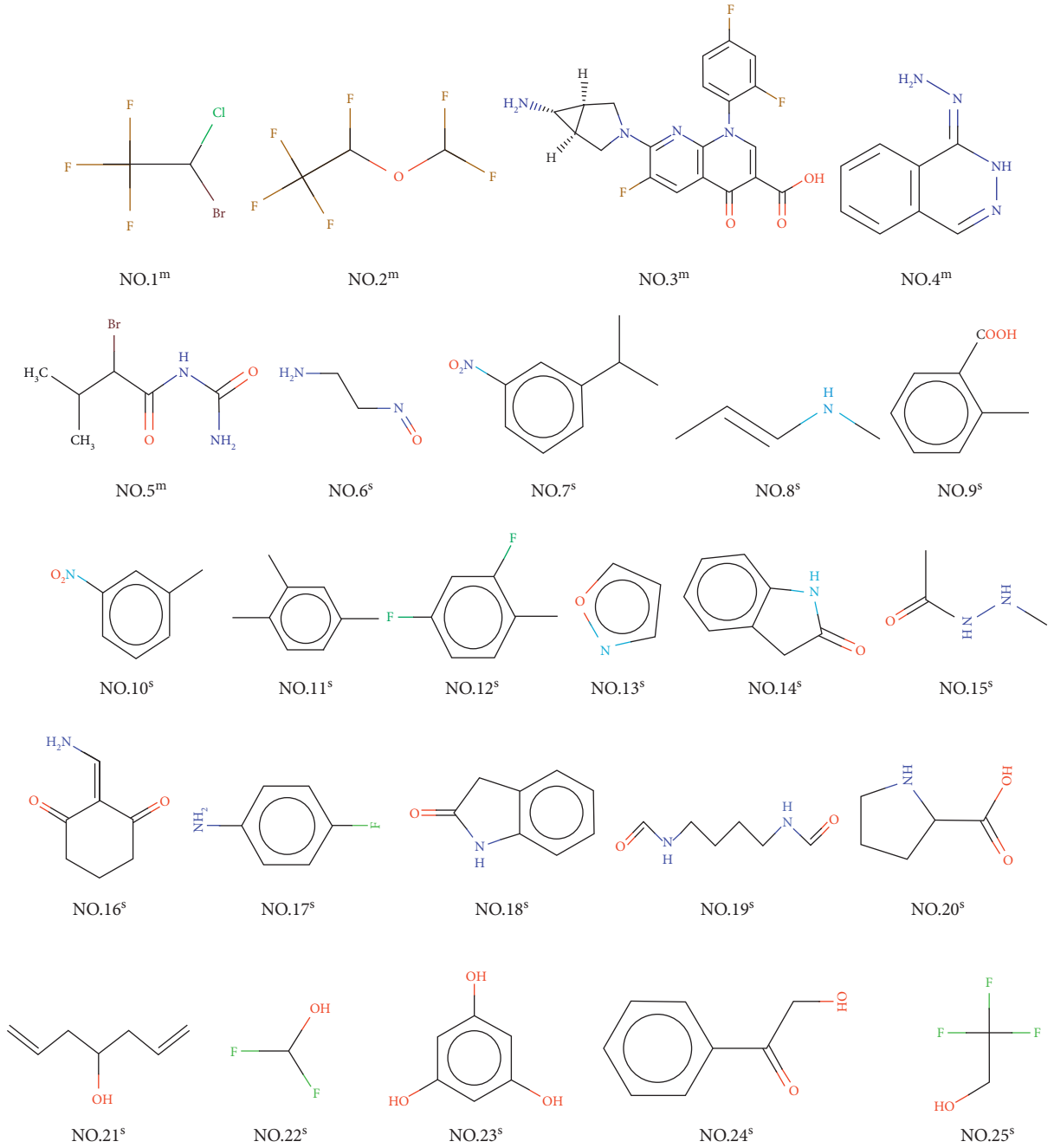

Figure 17: Privileged substructures or representative drugs containing privileged substructures. [m]: representative drugs containing corresponding privileged substructures. [S]: privileged substructures.

are not given clearly. Furthermore, the number of compounds in the test set and external validation set used in [5] is 413 and 151, respectively, significantly less than that in this study. In the (Q)SAR model, a larger validation set is easier to obtain a reliable evaluation result, so we believe that the two models cannot be compared fairly. Considering the predictive ability and AD coverage of the model, as well as the standardization of the modeling process, we believe that the joint decision-making model proposed in this paper is superior to other models in Table 6.

Overall, there is compelling evidence that we built a normative and robust DILI model with a wide AD and got a better performance than most of comparative studies. The modeling process is normative and transparent, which is in line with OECD principles, and the model can be replicated by other researchers. However, our study still has some limitations, which also exist in the most of previous studies. For example, like most studies, we only constructed the two-class DILI model to distinguish whether the inquiry compound has hepatotoxicity or not, and the effects of dose and

Table 7: Comparison of the present model with the main published models (Nm > 1304).

Ref.	Method	N_m	N_{de}	Performance	Coverage	Mechanism analysis
[5]	SVM	2295	166	TV: 80.39% ACC, 88.15% SE, 65.73% SP EV: 82.78% ACC, 93.18% SE, 68.25% SP	—	Yes
[9]	RF	2513	206	68.7 ± 1.7% ACC, 81.4 ± 1.6% SE, 50.8 ± 4.6% SP	—	No
[31]	GA-SVM	2171	674	ACC: 75%, TV: 68% SE, 95% SP	—	Yes
[40]	MC4PC, MDL-QSAR, BioEpisteme, PDM[a]	1608	Not given	CV: 39.2 ± 2.6% SE, 87.1 ± 2.6% SP EV: 88.9% SE	93.5 ± 3.8%	No
[39]	SVM, NB, kNN, CT, RF	1317	307	TV: 65.74% ACC, 85.16% SE, 34.38% SP EV: 75% ACC, 93.22% SE, 37.93% SP	—	Yes
Present	Joint decision-making based on SVM	2608	150	CSM(T = 50%): CV5 (70.8%–73.2% ACC) JDM(T = 70%): TV (80.0% ACC, 83.9% SE, 73.3% SE); EV (79.8% ACC, 96.5% SE, 66.8% SP)	JDM(T = 70%) TV: 99.24% EV: 98.20%	Yes

SVM: support vector machine; GA-SVM: genetic algorithm-support vector machine; RF: random forest; [a]detailed information of these methods can be found in [40]; NB: naive Bayes; kNN: k-nearest neighbor; CT: classification tree; N_m is the number of compounds in the modeling dataset; N_{de} is the number of descriptors/fingerprints used in (Q)SAR models; CV: cross-validation; CV5: fivefold cross-validation; EV: external validation; TV: validation on the test set; CSM: consensus model; JDM: joint decision-making model.

metabolites on DILI were not considered in our study. The reason is that we lack adequate, reliable DILI data with rich information, such as dose, severity, and type of hepatotoxicity. If there are more detailed data sources available, all the limitations will be broken, and we will be able to construct a dose-dependent and metabolite-related multiclass DILI model that takes into account the severity and type of hepatotoxicity with the method proposed in this study.

4. Conclusions

In this study, we developed a new joint decision-making model for DILI prediction under the OECD principle. A relatively large DILI dataset containing 2608 compounds was collected and characterized by PaDEL, PubchemFP, SubFP, EstateFP, and KRFP for modeling. Consensus models for joint decision-making were constructed with structurally diverse submodels optimized by the joint optimization method. The overwhelming majority principle obviously improved the accuracy and credibility of the consensus model's outputs. The application scope narrowing caused by the overwhelming majority principle was successfully solved by the joint decision-making strategy. The joint decision-making model performed well on both 528 test compounds and 390 independent external validation compounds. Furthermore, the contributions of top 10 PaDEL descriptors with a significant difference between DILI-positive and DILI-negative compounds were analyzed. 25 privileged substructures responsible for DILI positives were identified from SubFP, PubchemFP, and KRFP. We hope these structural alerts can provide some useful warning information for pharmaceutical chemists. The comparison results show that the joint decision-making model has certain advantages in data size, standardization, and transparency of the modeling process, coverage of the AD, and performance on test and external verification sets. As far as we know, the overwhelming majority principle combined with the joint decision-making strategy for (Q)SAR modeling has not been used in (Q)SAR study. Under the OECD principle, this is the best (Q)SAR model for DILI prediction with such a large data size. It will be a promising tool for virtual screening in the early stage of drug discovery.

Data Availability

The data used to support the findings of this study are available from this article and supplementary information file.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Grant nos. 71571091 and 71771112), the University of Science and Technology Liaoning Talent Project (Grant no. 601011507-03), and the scientific research project of Liaoning Provincial Department of Education (Grant no. LJKZ0320).

Supplementary Materials

Supplementary.pdf: Figures S1–S21 and Tables S1–S5. DILI_DataSet.csv: DILI data used for (Q)SAR study in this paper. External_validation_Set.csv: external validation dataset used for external validation in this study. (*Supplementary Materials*)

References

[1] X. Huang, F. Tang, Y. Hua, and L. Xiao, "In silico prediction of drug-induced toxicity using machine learning and deep learning methods," *Chemical Biology & Drug Design*, vol. 98, no. 2, pp. 248–257, 2021.

[2] Y. Hua, Y. Shi, X. Cui, and X. Li, "In silico prediction of chemical-induced hematotoxicity with machine learning and deep learning methods," *Molecular Diversity*, vol. 25, no. 3, pp. 1585–1596, 2021.

[3] P. B. Fontanarosa, D. Rennie, and C. D. DeAngelis, "Post-marketing Surveillance-lack of vigilance, lack of trust," *Jama*, vol. 292, no. 21, pp. 2647–2650, 2004.
[4] M. Fung, A. Thornton, K. Mybeck, J. H. H. Wu, K. Hornbuckle, and E. Muniz, "Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets-1960 to 1999," *Drug Information Journal*, vol. 35, no. 1, pp. 293–317, 2001.
[5] X. Li, Y. Chen, X. Song, Y. Zhang, H. Li, and Y. Zhao, "The development and application of in silico models for drug induced liver injury," *RSC Advances*, vol. 8, no. 15, pp. 8101–8111, 2018.
[6] D. Schuster, C. Laggner, and T. Langer, "Why drugs fail—a study on side effects in new chemical entities," *Current Pharmaceutical Design*, vol. 11, no. 27, pp. 3545–3559, 2005.
[7] M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong, "FDA-approved drug labeling for the study of drug-induced liver injury," *Drug Discovery Today*, vol. 16, no. 15, pp. 697–703, 2011.
[8] L. Goldkind and L. Laine, "A systematic review of NSAIDs withdrawn from the market due to hepatotoxicity: lessons learned from the bromfenac experience," *Pharmacoepidemiology and Drug Safety*, vol. 15, no. 4, pp. 213–220, 2006.
[9] L. Liu, L. Fu, J. W. Zhang et al., "Three-level hepatotoxicity prediction system based on adverse hepatic effects," *Molecular Pharmaceutics*, vol. 16, no. 1, pp. 393–408, 2019.
[10] M. D. Aleo, F. Shah, S. Allen et al., "Moving beyond binary predictions of human drug-induced liver injury (DILI) toward contrasting relative risk potential," *Chemical Research in Toxicology*, vol. 33, no. 1, pp. 223–238, 2020.
[11] D. L. Mendrick, "Toxicogenomics and classic toxicology: how to improve prediction and mechanistic understanding of human toxicity," *Essential Concepts in Toxicogenomics*, vol. 460, pp. 1–22, 2008.
[12] R. M. Walker and T. F. McElligott, "Furosemide induced hepatotoxicity," *The Journal of Pathology*, vol. 135, no. 4, pp. 301–314, 1981.
[13] F. Ballet, "Hepatotoxicity in drug development: detection, significance and solutions," *Journal of Hepatology*, vol. 26, pp. 26–36, 1997, supp-s2.
[14] U. Spengler, M. Lichterfeld, and J. K. Rockstroh, "Antiretroviral drug toxicity—a challenge for the hepatologist?" *Journal of Hepatology*, vol. 36, no. 2, pp. 283–294, 2002.
[15] D. Fourches, J. C. Barnes, N. C. Day, P. Bradley, J. Z. Reed, and A. Tropsha, "Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species," *Chemical Research in Toxicology*, vol. 23, no. 1, pp. 171–183, 2010.
[16] P. B. Watkins, "Quantitative systems toxicology approaches to understand and predict drug-induced liver injury," *Clinics in Liver Disease*, vol. 24, no. 1, pp. 49–60, 2020.
[17] X. Li, Y. Zhang, H. Chen, H. Li, and Y. Zhao, "In silico prediction of chronic toxicity with chemical category approaches," *RSC Advances*, vol. 7, no. 66, pp. 41330–41338, 2017.
[18] K. R. Przybylak and M. T. Cronin, "In silicomodels for drug-induced liver injury—current status," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 8, no. 2, pp. 201–217, 2012.
[19] L. Kuna, I. Bozic, T. Kizivat et al., "Models of drug induced liver injury (DILI)—current issues and future perspectives," *Current Drug Metabolism*, vol. 19, no. 10, pp. 830–838, 2018.
[20] E. Kotsampasakou, F. Montanari, and G. F. Ecker, "Predicting drug-induced liver injury: the importance of data curation," *Toxicology*, vol. 389, pp. 139–145, 2017.
[21] Y. Low, T. Uehara, Y. Minowa et al., "Predicting drug-induced hepatotoxicity using qsar and toxicogenomics approaches," *Chemical Research in Toxicology*, vol. 24, no. 8, pp. 1251–1262, 2011.
[22] R. Chan and L. Z. Benet, "Evaluation of DILI predictive hypotheses in early drug development," *Chemical Research in Toxicology*, vol. 30, no. 4, pp. 1017–1029, 2017.
[23] Z. Liu, Q. Shi, D. Ding, R. Kelly, H. Fang, and W. Tong, "Translating clinical findings into knowledge in drug safety evaluation-drug induced liver injury prediction system (DILIps)," *PLoS Computational Biology*, vol. 7, no. 12, Article ID e1002310, 2017.
[24] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, and L. Lai, "Deep learning for drug-induced liver injury," *Journal of Chemical Information and Modeling*, vol. 55, no. 10, pp. 2085–2093, 2015.
[25] J. R. Mora, Y. Marrero-Ponce, C. R. García-Jacas, and A. Suarez Causado, "Ensemble models based on qubils-mas features and shallow learning for the prediction of drug-induced liver toxicity: improving deep learning and traditional approaches," *Chemical Research in Toxicology*, vol. 33, no. 7, pp. 1855–1873, 2020.
[26] D. P. Russo, K. M. Zorn, A. M. Clark, H. Zhu, and S. Ekins, "Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction," *Molecular Pharmaceutics*, vol. 15, no. 10, pp. 4361–4370, 2015.
[27] A. D. Rodgers, H. Zhu, D. Fourches, I. Rusyn, and A. Tropsha, "Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure–activity relationship method," *Chemical Research in Toxicology*, vol. 23, no. 4, pp. 724–732, 2010.
[28] V. Drgan and B. Bajelj, "Application of supervised SOM algorithms in predicting the hepatotoxic potential of drugs," *International Journal of Molecular Sciences*, vol. 22, no. 9, Article ID 4443, 2021.
[29] M. Cruz-Monteagudo, M. N. D. S. Cordeiro, and F. Borges, "Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity," *Journal of Computational Chemistry*, vol. 29, no. 4, pp. 533–549, 2008.
[30] B. Bajželj and V. Drgan, "Hepatotoxicity modeling using counter-propagation artifificial neural networks: handling an imbalanced classifification problem," *Molecules*, vol. 25, p. 481, 2020.
[31] D. Mulliner, F. Schmidt, M. Stolte, H.-P. Spirkl, A. Czich, and A. Amberg, "Computational models for human and animal hepatotoxicity with a global application scope," *Chemical Research in Toxicology*, vol. 29, no. 5, pp. 757–767, 2016.
[32] J. M. Choi, S. J. Oh, J.-Y. Lee et al., "Prediction of drug-induced liver injury in hepg2 cells cultured with human liver microsomes," *Chemical Research in Toxicology*, vol. 28, no. 5, pp. 872–885, 2015.
[33] E. Minerali, D. H. Foil, K. M. Zorn, T. R. Lane, and S. Ekins, "Comparing machine learning algorithms for predicting drug-induced liver injury (DILI)," *Molecular Pharmaceutics*, vol. 17, no. 7, pp. 2628–2637, 2020.
[34] N. Greene, L. Fisk, R. T. Naven, R. R. Note, M. L. Patel, and D. J. Pelletier, "Developing structure–activity relationships for the prediction of hepatotoxicity," *Chemical Research in Toxicology*, vol. 23, no. 7, pp. 1215–1222, 2010.
[35] C. Y. Liew, Y. C. Lim, and C. W. Yap, "Mixed learning algorithms and features ensemble in hepatotoxicity prediction," *Journal of Computer-Aided Molecular Design*, vol. 25, no. 9, pp. 855–871, 2010.

[36] R. Ancuceanu, M. V. Hovanet, A. L. Anghel, F. Furtunescu, M. N. C. Constantin, and M. Dinu, "Computational models using multiple machine learning algorithms for predicting drug hepatotoxicity with the DILIrank dataset," *International Journal of Molecular Sciences*, vol. 21, Article ID 2114, 2020.

[37] A. Cheng and S. L. Dixon, "In silico models for the prediction of dose-dependent human hepatotoxicity," *Journal of Computer-Aided Molecular Design*, vol. 17, no. 12, pp. 811–823, 2003.

[38] H. Ai, W. Chen, L. Zhang et al., "Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints," *Toxicological Sciences*, vol. 165, no. 1, pp. 100–107, 2018.

[39] C. Zhang, F. Cheng, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silico prediction of drug induced liver toxicity using substructure pattern recognition method," *Molecular Informatics*, vol. 35, no. 3-4, pp. 136–144, 2016.

[40] E. J. Matthews, C. J. Ursem, N. L. Kruhlak et al., "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," *Regulatory Toxicology and Pharmacology*, vol. 54, no. 1, pp. 23–42, 2009.

[41] Q. Shen, J. H. Jiang, J. C. Tao, G. L. Shen, and R. Q. Yu, "Modified Ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors," *Journal of Chemical Information and Modeling*, vol. 45, no. 4, pp. 1024–1029, 2005.

[42] Y. Wang and X. Chen, "Hybrid quantum particle swarm optimization algorithm and its application," *Science China Information Science*, vol. 63, no. 5, pp. 199–201, 2020.

[43] Y. Wang and X. Chen, "A joint optimization qsar model of fathead minnow acute toxicity based on a radial basis function neural network and its consensus modeling," *RSC Advances*, vol. 10, Article ID 21292, 2020.

[44] R. Batuwita and V. Palade, "Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 4, pp. 1250003–1251125, 2012.

[45] A. Cherkasov, E. N. Muratov, D. Fourches et al., "QSAR modeling: where have you been? where are you going to?" *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[46] J. R. Votano, M. Parham, L. H. Hall et al., "Three new consensus QSAR models for the prediction of ames genotoxicity," *Mutagenesis*, vol. 19, no. 5, pp. 365–377, 2004.

[47] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions On Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 2002.

[48] G. Idakwo, J. Luttrell, M. Chen et al., "A review on machine learning methods for in silico toxicity prediction," *Journal of Environmental Science and Health. Part C, Environmental Carcinogenesis & Ecotoxicology Reviews*, vol. 36, no. 1, pp. 169–191, 2018.

[49] J. Liu, K. Mansouri, R. S. Judson et al., "Predicting Hepatotoxicity using toxcast in vitro bioactivity and chemical structure," *Chemical Research in Toxicology*, vol. 28, no. 4, pp. 738–751, 2015.

[50] J. Alves, C. B. Henriques, and R. J. Poppi, "Classification of diesel pool refinery streams through near infrared spectroscopy and support vector machines using c-svc and ν-svc," *Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy*, vol. 117, pp. 389–396, 2004.

[51] P. Gramatica, P. Pilutti, and E. Papa, "Validated QSAR Prediction of OH tropospheric degradation of VOCs: splitting into training–test sets and consensus modeling," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1794–1802, 2004.

[52] C. Hansch and R. P. Verma, "20-(S)-camptothecin analogues as DNA topoisomerase I inhibitors: a QSAR study," *Journal of Medicinal Chemistry*, vol. 2, no. 12, pp. 1807–1813, 2008.

[53] P. Gramatica, E. Giani, and E. Papa, "Statistical external validation and consensus modeling: a qspr case study for koc prediction," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 6, pp. 755–766, 2007.

[54] Y. Wang and X. Chen, "QSPR Model for caco-2 cell permeability prediction using a combination of HQPSO and dual-RBF neural network," *RSC Advances*, vol. 10, no. 70, pp. 42938–42952, 2020.

[55] T. Öberg, "A QSAR for baseline toxicity:? validation, domain of application, and prediction," *Chemical Research in Toxicology*, vol. 17, no. 12, pp. 1630–1637, 2004.

[56] F. E. Önen Bayram, S. A. A. Alradhwani, G. Tugcu, and H. Sipahi, "Do we build similar molecules for comorbid diseases? tevarud in drug design, an analysis for depression and inflammation," *ACS Medicinal Chemistry Letters*, vol. 11, no. 2, pp. 147–153, 2020.

[57] Y. Wang, H. Liu, Y. Fan, X. Chen, and Y. Zhang, "In silico prediction of human intravenous pharmacokinetic parameters with improved accuracy," *Journal of Chemical Information and Modeling*, vol. 59, no. 9, pp. 3968–3980, 2020.

[58] M. Chen, A. Suzuki, S. Thakkar, K. Yu, C. Hu, and W. Tong, "DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans," *Drug Discovery Today*, vol. 21, no. 4, pp. 648–653, 2016.

[59] J. P. Sanderson, D. J. Naisbitt, J. Farrell et al., "Sulfamethoxazole and its metabolite nitroso sulfamethoxazole stimulate dendritic cell costimulatory signaling," *The Journal of Immunology*, vol. 178, no. 9, pp. 5533–5542, 2007.

[60] OECD, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q) SAR] Models*, Organisation for Economic Co-Operation and Development, Paris, France, 2007.