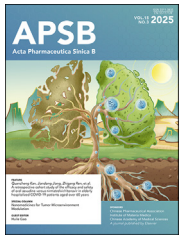
Acta Pharmaceutica Sinica B 2025;15(3):1344−1358

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



TOOLS

Establishment of interpretable cytotoxicity prediction models using machine learning analysis of transcriptome features

You Wu[a,b,†], Ke Tang[a,b,†], Chunzheng Wang[a,b], Hao Song[a,b], Fanfan Zhou[a,b], Ying Guo[a,b,*]

[a]*Beijing Key Laboratory of New Drug Mechanisms and Pharmacological Evaluation Study, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China*
[b]*Department of Pharmacology, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China*

Received 26 June 2024; received in revised form 13 September 2024; accepted 14 September 2024

KEY WORDS

Interpretable model;
Drug safety;
Cell viability;
Weak cytotoxicity;
Machine learning;
Transcriptome;
Cytotoxicity Signature genes;
Narrow therapeutic index drugs

Abstract Cytotoxicity, usually represented by cell viability, is a crucial parameter for evaluating drug safety *in vitro*. Accurate prediction of cell viability/cytotoxicity could accelerate drug development in the early stage. In this study, by integrating cellular transcriptome and cell viability data using four machine learning algorithms (support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM)) and two ensemble algorithms (voting and stacking), highly accurate prediction models of 50% and 80% cell viability were developed with area under the receiver operating characteristic curve (AUROC) of 0.90 and 0.84, respectively; these models also showed good performance when utilized for diverse cell lines. Concerning the characterization of the employed Feature Genes, the models were interpreted, and the mechanisms of bioactive compounds with a narrow therapeutic index (NTI) can also be analyzed. In summary, the models established in this research exhibit superior capacity to those of previous studies; these models enable accurate high-safety substance screening *via* cytotoxicity prediction across cell lines. Moreover, for the first time, Cytotoxicity Signature (CTS) genes were identified, which could provide additional clues for further study of mechanisms of action (MOA), especially for NTI compounds.

*Corresponding author.
E-mail address: yingguo6@imm.ac.cn (Ying Guo).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.02.009
2211-3835 © 2025 Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

© 2025 Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Drug safety is an important issue in drug development and causes approximately one-third of drug failures[1−3]. *In vitro* toxicity studies need to be conducted in the early stage of drug discovery. Cell viability, typically employed to ascertain the extent of cell proliferation, survival, or death after exposure to drugs or compounds, serves as a classic metric for evaluating drug safety *in vitro*[4−8]. Many cell viability assays have been developed that are more convenient than *in vivo* assays but are still costly. The development of an accessible and reliable cell viability prediction model will provide a valuable *in silico* tool for toxicity evaluation.

The accumulation of omics data has had a great impact on pharmaceutical R&D[9,10]. Cellular chemical perturbation transcriptomic signatures, which depict comprehensive transcriptomic alterations in cells after exposure to compounds, provide valuable clues for drug development[11−13], including the mechanism and toxicity of drug candidates. The signature of cytotoxicity identification and quantification would contribute greatly to compound safety prediction, which will promote the attainment of highly safe substances from the transcriptomic data of many chemical perturbations and may also help to understand the molecular mechanisms of toxicity and activity. As a powerful tool for data mining, machine learning has been applied to reveal valuable information from omics data in drug development[14−16], including cell viability prediction. Several cell viability prediction models were generated previously by correlating transcriptomic data and toxicity data in which the data collection conditions were roughly matched due to differences in database resources and leaded to relative performance of the models[17,18]. In addition, most of the reported models lack interpretability and need to be implemented for a deep understanding of the molecular mechanisms of cytotoxicity.

In this study, high-quality datasets were generated and utilized for modeling *via* machine learning algorithms. In addition to setting 50% cell viability as the threshold as in other studies, 80% cell viability was applied as a threshold for the prediction of weakly cytotoxic compounds. The models in this study are interpretable by quantification of Feature Genes on cell viability/cytotoxicity, which provides not only the basis for model prediction but could also be used for the analysis of the molecular mechanism of cytotoxicity caused by individual compounds. Notably, the pharmacodynamic mechanism of a compound with a narrow therapeutic index is usually affected by the cytotoxicity of the compound, and the list of Cytotoxicity Signature (CTS) genes obtained in this study could be utilized to eliminate the confounding effect to achieve the mechanism of action (MOA) of the substance.

2. Materials and methods

2.1. Cell lines and compounds

A549 (Cat. No. CCL-185) and HepG2 (Cat. No. HB-8065) cells were obtained from the American Type Culture Collection (Manassas, VA, USA); HT-29 (Cat. No. 1101HUM-PUMC000109) and HEK293 (Cat. No. 1101HUM-PUMC000010) cells were obtained from the Cell Resource Center of Peking Union Medical College (Beijing, China). A549 and HepG2 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 IU/mL penicillin and 100 μg/mL streptomycin; HEK293 cells were cultured in Minimum Essential Medium supplemented with 10% FBS, 100 IU/mL penicillin and 100 μg/mL streptomycin; HT-29 cells were cultured in DMEM/F-12 supplemented with 5% FBS, 100 IU/mL penicillin and 100 μg/mL streptomycin. DMEM (Cat. No. 11965500BT), Minimum Essential Medium (Cat. No. C11095500BT), FBS (Cat. No. 10099141C), penicillin streptomycin (Cat. No. 15140122), and DMEM/F-12 (Cat. No. 11320033) were purchased from Gibco (Waltham, MA, USA).

A total of 1243 compounds (purity >95%) were obtained from the compound libraries (Cat. Nos. L1000, L3400, L6000, and L4000), and the powders of the compounds, cyclosporine A (Cat. No. T0945; purity = 98%) and triptolide (Cat. No. T2179; purity = 98%), were purchased from TargetMol (Boston, MA, USA). The structure and purity of all the compounds were confirmed by nuclear magnetic resonance and high-performance liquid chromatography. The compounds used in this study were dissolved in dimethyl sulfoxide (DMSO, Cat. No. 34943; Sigma−Aldrich, St. Louis, MO, USA) and stored at −20 °C.

2.2. Cell viability assay

A549, HEK293, HepG2, and HT-29 cells were seeded in 96-well plates and treated with the test compounds for 48 h. Cell viability was determined *via* the CellTiter-Glo Assay (Cat. No. G7571; Promega, Madison, WI, USA) by measuring the relative light units (RLUs) according to the manufacturer's protocol. Cells treated with DMSO (1‰, *v*/*v*) served as the vehicle control. Cell viability was calculated using Eq. (1) [19]:

$$\text{Cell viability (\%)} = \text{RLUs}_{\text{compound}}/\text{RLUs}_{\text{DMSO}} \times 100 \quad (1)$$

All the experiments were performed in duplicate, and the mean values subsequently were utilized for classifier label determination.

2.3. RNA extraction and RNA sequencing

A549, HepG2, and HT-29 cells were seeded in 6-well plates and incubated with the test compounds or DMSO (1‰, *v*/*v*). Briefly, HT-29 cells were treated with cyclosporine A at a final concentration of 10 μmol/L; A549 and HepG2 cells were treated with triptolide at final concentrations of 30 and 10 nmol/L, respectively. Total RNA was extracted using TRIzol reagent (Cat. No. 15596026; Gibco, Waltham, MA, USA) according to the manufacturer's protocol. RNA quality assessment, cDNA library construction and RNA sequencing were conducted by OE Biotech Co., Ltd. (Shanghai, China). Briefly, RNA integrity was assessed by an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The cDNA library was constructed using a VAHTS Universal V6 RNA-seq Library Prep Kit (Vazyme, Nanjing, China) according to the manufacturer's instructions and

sequenced on Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA). The raw reads were first processed using fastp[20] to obtain the clean reads, which were subsequently mapped to the reference genome using HISAT2[21]. The FPKM[22] value of each gene was calculated, and the read counts of each gene were obtained *via* HTSeq-count[23]. Differential expression analysis was performed using DESeq2[24], and the significant differential expressed genes (DEGs) were identified by using the following thresholds: q value < 0.05 and fold change >2 or < 0.5.

2.4. Retrieval and processing of data

The transcriptome data utilized in this study were downloaded from the Library of Integrated Network-based Cellular Signatures (LINCS, https://clue.io/data/CMap2020#LINCS2020)[25] and "sig_id" information of the files is provided in Supporting Information Table S1. The level 3 and level 5 matrixes of 1243 compounds were extracted using the parse function of the cmapPy python package[26]. Subsequently, the signatures obtained under the same conditions were averaged using the Moderated z score (MODZ) procedure[25] as follows: the pairwise Spearman correlation values of the profiles were calculated for the weight of each signature, and the values with negative correlations were set as 0.01; then the weighted average of each signature was subsequently calculated for multiple expression profiles to obtain the profile for modeling. To obtain a high-level view of the distribution of the compounds in modeling data, t-SNE analysis[27] was performed on the 1243 signatures for non-linear dimensionality reduction and visualization.

According to the A549 cell viability, the 1243 compounds were divided into three groups according to the following ranges: less than 50%, 50%−80%, and greater than 80%. The training set and test set of this study were generated by randomly stratifying the data at a ratio of 8:2 by using the train_test_split function.

2.5. Feature selection

2.5.1. Identification of variable genes

Variable genes were obtained by calculating the SD and maximum fold change (MFC) of each gene among the training set on their level 3 matrix with the thresholds SD >1 or MFC >2.

2.5.2. Weighted correlation network analysis for feature selection

The modules of co-expressed genes among variable genes were identified by using the wgcna[28] R package, and each module eigengene was calculated for the correlation analysis between the module and cell viability to identify modules with the highest positive and negative correlation scores. The genes in the modules composed the feature set of weighted correlation network analysis (WGCNA). The module membership of each gene was subsequently assessed *via* correlation value calculations between the gene expression and the module eigengene.

2.5.3. Minimum redundancy maximum relevance for feature selection

The minimum redundancy maximum relevance (mRMR) method ranks features based on maximizing the correlation between the feature and the label and minimizing the redundancy across features. In this study, by using the mRMRe[29] package, the mRMR score of each variable gene was calculated; this score represents the correlation between the gene and the label minus the average redundancy score of the gene across all variable genes. The genes with a threshold of mRMR >0 were collected as the feature set of mRMR.

2.5.4. Gene enrichment analysis for feature selection

The samples in the training set with cell viability values less than 50% or 80% were used for enrichment analysis. For each sample, the variable genes were ranked by MODZ score, and enrichment analysis was subsequently applied to the variable genes of each sample by using the fgsea[30] R package based on biological processes (BP) terms to obtain the terms with false discovery rate <0.05. Leading-edge genes were also recorded. The C value of each gene was calculated using Eq. (2):

$$C = n / N \tag{2}$$

where n is the total frequency of the gene recorded and N is the number of samples for analysis. The feature set of gene set enrichment (GSEA) was composed of genes with $C > 0$.

2.5.5. Robust rank aggregation

The genes in the three Gene sets obtained by WGCNA, mRMR, and fgsea were ranked in descending order according to their module membership score, mRMR score, and C value, respectively. The rank of each gene in the three lists was used as the feature value, and the three lists were subsequently aggregated by robust rank aggregation (RRA) analysis using the RobustRankAggreg[31] R package to determine the significance of each gene. The genes with exact $P < 0.05$ were considered as Ensemble Features in the Ensemble Feature Set.

2.6. Model training

Four machine learning algorithms, namely, support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM), were used to train the cell viability prediction models, including the classification and the regression models. The classification models were developed by using the svm.SVC function, the RandomForestClassifier function, the xgboost python package and the lightgbm python package from the scikit-learn library, while regression models were developed by using the svm.SVR function, the RandomForestRegressor function, the xgboost python package, and the lightgbm python package. In addition, voting and stacking were applied to fuse the four classification models or regression models, as the voting model was developed by a soft voting strategy (votingClassifier), and the stacking model (StackingClassifier) was developed by using the four models as base learners, and the logistic regression model and LinearRegression model were the meta-learners for the classification models and regression models, respectively. The hyperparameter tuning of each model was developed by using the bayes_opt python package in 5-fold cross-validation. Additionally, deep learning algorithms were also used to train cell viability prediction models: the multilayer perceptron and convolutional neural network were developed by using the Keras package with the backend of TensorFlow[32]. The TabResnet model and TabNet model were developed by using the pytorch_widedeep package[33] and pytorch_tabnet package[34], respectively.

2.7. Classifier performance evaluation

The metrics, including accuracy, precision, recall, specificity, and area under the receiver operating characteristic curve (AUROC),

were calculated by sklearn.metrics and used to evaluate the performance of the classifiers. The metrics were defined using Eqs. (3)−(6):

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$\text{Precision} = TP / (TP + FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$\text{Specificity} = TN / (TN + FP) \quad (6)$$

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

2.8. *Interpretation analysis*

SHapley Additive exPlanations (SHAP) analysis was used for interpretation analysis[35]. The SHAP value of each gene in individual samples in the training set was calculated by the KernelShap function, and the global feature explanation for the model were achieved as the mean absolute SHAP value of each gene. For interpretation of individual sample predictions, local explanations were conducted by using the shap.summary_plot function for quantification of the local contribution of each gene.

2.9. *Acquirement of CTS genes*

The feature importance score of an individual feature on the 80% cell viability prediction model was represented by its mean absolute SHAP value across all samples in the training set. The features with feature importance score greater than the average feature importance score were acquired. Subsequently, Pearson correlation (ρ) was calculated by using the pandas.corr function between each important feature gene and other variable genes to obtain a comprehensive gene set associated with cell viability ($\rho > 0.5$), which was defined as the CTS genes in this study.

2.10. *MOA prediction*

The intersection of the DEGs (or the DEGs with CTS genes removal) of the compound with the genes detected in the query platform was obtained. The top 150 up- and downregulated genes were selected according to the $\log_2$ (fold change) and used for MOA prediction *via* the Query platform (https://clue.io/query)[25]. The top 10 MOA terms of the corresponding cell line based on the norm connectivity score were obtained with the gene sets before and after the removal of the CTS genes used as inputs, respectively.

2.11. *Pathway enrichment analysis*

Pathway enrichment analysis was performed by using the Metascape platform (http://metascape.org)[36] based on the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes pathways, and reactome databases, with the following thresholds: $P < 0.01$, a minimum count of 3, and an enrichment factor >1.5.

2.12. *Data and material availability*

All data needed to evaluate the conclusions of the paper are present in the paper and/or Supporting Information The RNA sequencing (RNA-seq) dataset generated in this manuscript has been deposited in the NCBI Gene Expression Omnibus database (GSE252529). All code for model training associated with the current submission is available at https://github.com/WuYou-imm/Interpretable-cytotoxicity-prediction-model.

3. Results

3.1. *Data collection and splitting*

Cell viability prediction models were constructed in this study (Fig. 1) by analyzing a total of 1243 bioactive samples with diverse MOAs that act on 431 targets and all samples were well-distributed across the training set and test set (Supporting Information Table S1 and Fig. S1A). The transcriptomic data of A549 cells treated with the compounds and the cell viabilities under the same conditions were correlated. The transcriptomic data were obtained by screening level 5 of the LINCS database, and the multiple expression profiles of the corresponding compounds were combined by using the MODZ (Fig. S1B). The cell viability data (Table S1 and Fig. S1C) of A549 cells treated with the test compounds were determined under the corresponding conditions.

Two thresholds, 50% and 80% cell viability, were applied in this study to construct classifiers. The 50% cell viability was commonly used as the cytotoxicity threshold, and the corresponding concentration, the 50% cytotoxic concentration, was often used to represent or compare the toxicity of compounds. In addition, given that compounds with weak toxicity have limitations in drug development, 80% cell viability was also used in this study as the threshold for distinguishing weakly toxic substances. Among the 1243 compounds, 991 compounds with cell viability greater than 50% (Fig. S1D) and 781 compounds with cell viability greater than 80% (Fig. S1E) were defined as positive samples for 50% and 80% cell viability classifiers, respectively. The 1243 compounds were subsequently divided into three groups according to cell viability, with percentages ranging from less than 50%, 50%−80%, and greater than 80%. The training set ($n = 994$) and test set ($n = 249$) of this study were acquired by random stratification at a ratio of 8:2 (Fig. S1C).

3.2. *Identification of cell viability Feature Genes*

Given the high dimensionality and redundancy of the transcriptome data with 12,328 gene expression values, an ensemble feature selection pipeline was used to obtain the Ensemble Feature Set (Fig. 2A). First, the MFC and SD of all genes among the samples in the training set were calculated to eliminate genes with invariable expression levels, and a total of 5135 variable genes (SD > 1 or MFC >2; Supporting Information Table S2) were obtained for subsequent analysis. Then, WGCNA, mRMR, and GSEA were applied to 5135 variable genes. By using WGCNA (Supporting Information Fig. S2), the expressed gene modules and the correlation coefficients between the modules and cell viability were quantified. The module with the highest positive coefficient of 0.58 ($P = 1 \times 10^{-91}$; the blue module in Fig. S2C) and the module with the highest negative coefficient of −0.36 ($P = 3 \times 10^{-32}$; the turquoise module in Fig. S2C) were obtained, and the genes in these two modules formed Gene set 1 ($n = 3430$) (Supporting Information Table S3). The mRMR algorithm was used to obtain the feature set with the greatest correlation with the label and

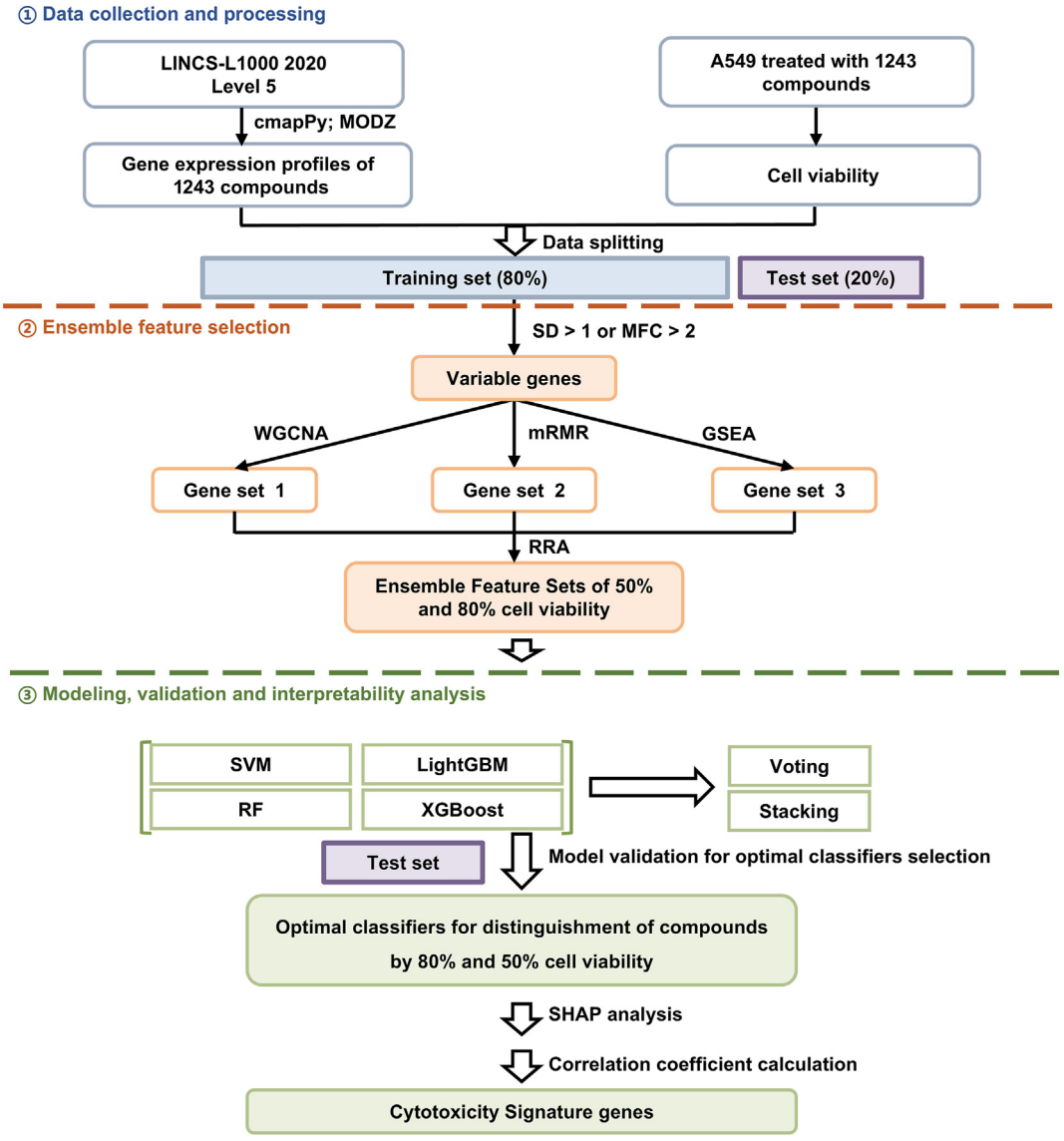


Figure 1 The schematic workflow of this study. The following models were developed for 50% and 80% cell viability prediction according to the basis of quantitative analysis of the relationship between gene signature and cell viability: (A) transcriptomic data from A549 cells treated with 1243 compounds from LINCS and cell viability data collected under the same conditions used for transcriptome data collection; (B) ensemble feature selection by using SD, MFC, WGCNA, GSEA, mRMR, and RRA algorithms; (C) and the optimal classifiers for 50% and 80% cell viability prediction were achieved by comparing various machine learning algorithms, including SVM, RF, XGBoost, LightGBM and two ensemble strategies (voting and stacking). The SHAP[35] analysis was used for interpretable analysis.

minimum redundancy across features by calculating the mutual information score between features and the label, as well as the redundancy of features. The features of 50% (n = 1432) (Supporting Information Table S4) and 80% cell viability (n = 636) (Supporting Information Table S5) with a score >0 serving as the threshold were obtained for Gene set 2. In addition, the GSEA method was applied based on the BP terms in the GO database to obtain Gene set 3 by analyzing the frequency of genes associated with toxic compounds (*i.e.*, cell viability less than 50% or 80%) since more frequently, genes enriched in compounds associated with toxicity are more likely to be highly correlated with cell viability. In Gene set 3, 50% (n = 4736) (Supporting Information Table S6) and 80% cell viability (n = 4747) (Supporting Information Table S7) were obtained with $C > 0$ as the threshold. The RRA method was subsequently used to aggregate the three Gene sets to obtain Ensemble Feature Sets of 50% and 80% cell viability, which included 292 (Supporting Information Table S8) and 226 genes (Supporting Information Table S9), respectively. Among the features, 197 genes overlapped (Fig. 2B), implying that the two Ensemble Feature Sets were mostly involved in the cell cycle (dots in Steel Blue, Fig. 2C and D) and protein metabolism (dots in Light Sea Green, Fig. 2C and D); moreover, the difference between the two Ensemble Feature Sets was mainly reflected in the presence of cell death pathways associated with 50% cell viability, and cell proliferation genes were associated with 80% cell viability (Fig. 2C and D). Toxic compounds could suppress cell proliferation, affect metabolism, and/or induce cell death. In the event of weak toxicity, the cellular oxidative stress response is triggered as a means to

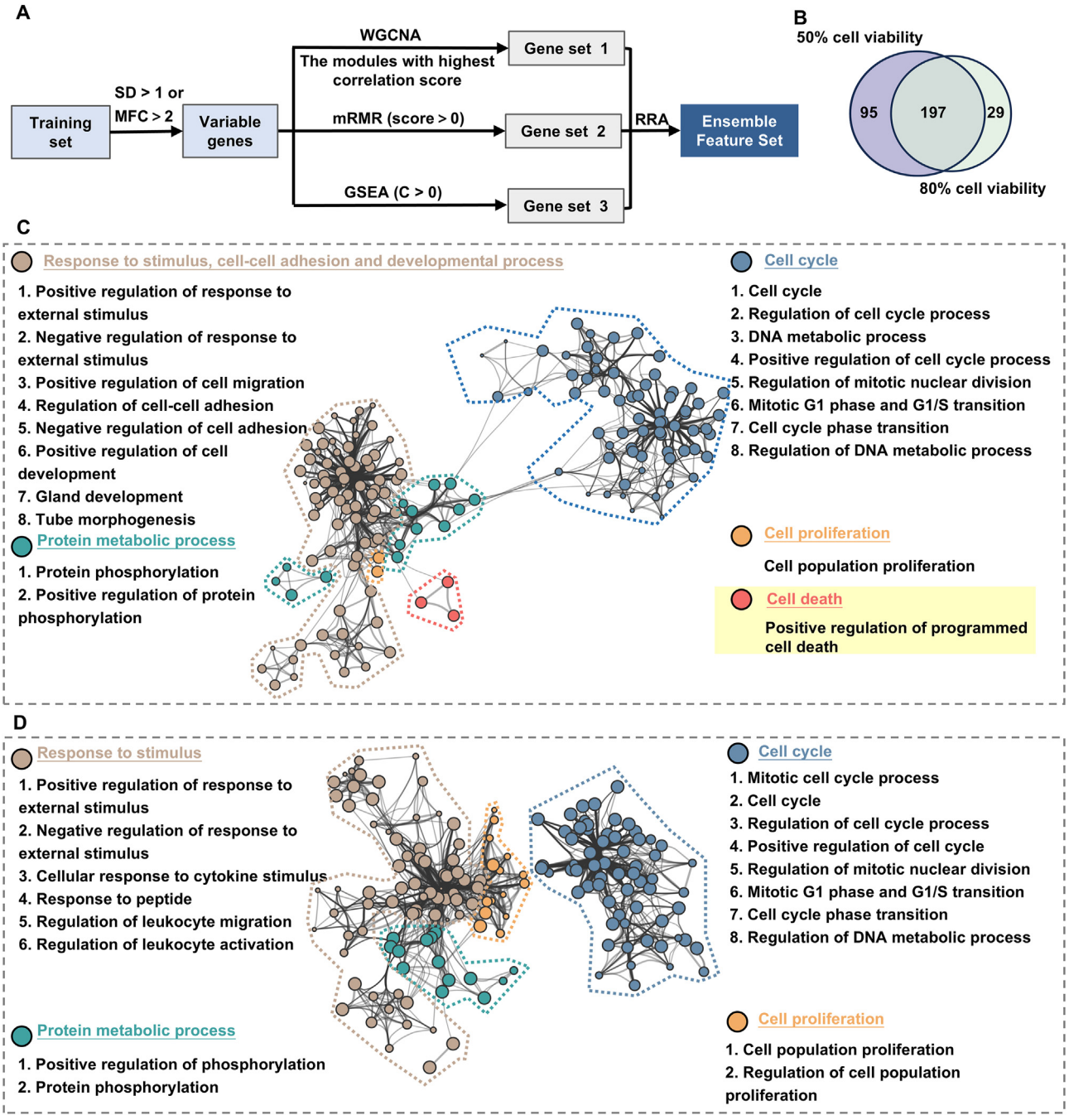


Figure 2 Ensemble feature selection. (A) The pipeline of ensemble feature selection. (B) The overlap of Feature Genes in Ensemble Feature Sets of 50% and 80% cell viability. (C, D) Pathways enriched by genes in the Ensemble Feature Sets of 50% (C) and 80% (D) cell viability according to the Metascape platform (http://Metascape.org)[36] based on BP terms of GO, Kyoto Encyclopedia of Genes and Genomes pathway, and Reactome databases, with the following thresholds: $P < 0.01$, a minimum count of 3, and an enrichment factor >1.5. The node represents the enriched pathway, the size of the node represents the number of enriched genes, and the thickness of the line between the nodes represents the similarity of the pathway. The nodes were divided into different clusters according to their similarity, and the most significant pathway in each cluster was used as the label. The clusters with relevant topics were subsequently grouped and are indicated by different colors.

counteract cellular oxidative imbalance, which is largely reversible and serves as an adaptive stress response[37–39]. However, severe toxic exposure significantly affects metabolic function and leads to heightened cell damage, inflammation, and cell death[37–39]. According to the results of the enrichment analysis of the two Ensemble Feature Sets, the majority of the pathways were in common, but some specific pathways were different due to the distinct toxicity events of the compounds, which had varying levels of toxicity.

3.3. *Construction of cell viability prediction models*

To construct models with a cell viability threshold of 50% or 80%, a sample with a cell viability greater than the threshold was

considered a positive sample. Four models, SVM, RF, XGBoost, and LightGBM, were trained *via* hyperparameter optimization through Bayesian optimization *via* 5-fold cross-validation (Supporting Information Tables S10 and S11); subsequently, two ensemble models, voting and stacking, were constructed by combining the four models. Finally, the performance of the six prediction models for each threshold was evaluated on the test set, with the AUROC score serving as the primary metric. The results showed that both the 50% and 80% cell viability prediction models exhibited superior performance in predicting positive samples, with AUROC values ranged from 0.8823 to 0.9011 (Supporting Information Table S12) and 0.8198 to 0.8408 (Supporting Information Table S13), respectively (Fig. 3). The 50% cell viability is the most common indicator for comparing differences in compound toxicity or tumor sensitivity. Among the 50% cell viability prediction models, the voting model had the highest AUROC value (0.9011) and was regarded as the optimal model for 50% cell viability prediction (Fig. 3A). The accuracy, precision, recall, and specificity of the voting model were 0.8755, 0.9239, 0.9192, and 0.6923, respectively (Fig. 3A), and the overall performance of the model exceeded that of previously reported models[17,18,40] (Supporting Information Table S14). In addition to the 50% cell viability threshold, 80% cell viability prediction models were developed to identify compounds with weak toxicity. The RF model performed the best, with an AUROC of 0.8408 and accuracy, precision, recall, and specificity metrics of 0.8112, 0.8225, 0.8910, and 0.7875, respectively (Fig. 3B). Identification of highly safe substances from massive samples in the early stage of drug discovery could significantly reduce subsequent costs in further development, which can be accessed either by choosing nontoxic samples or by excluding toxic substances. Among the metrics, "precision" indicates the proportion of the predicted nontoxic substances that are truly nontoxic, *i.e.*, how likely the chosen nontoxic substance is. The "specificity" represents the ratio of truly predicted toxic samples to total toxic samples, *i.e.*, the ability of the model to exclude toxic substances from all samples. Both optimal models, the voting model for 50% cell viability prediction and the RF model for 80% cell viability prediction, performed well in cell viability prediction, with precision scores of 0.9239 and 0.8225, and specificity scores of 0.6923 and 0.7875, respectively. Besides classical machine learning algorithms, deep learning algorithms are capable of extracting features

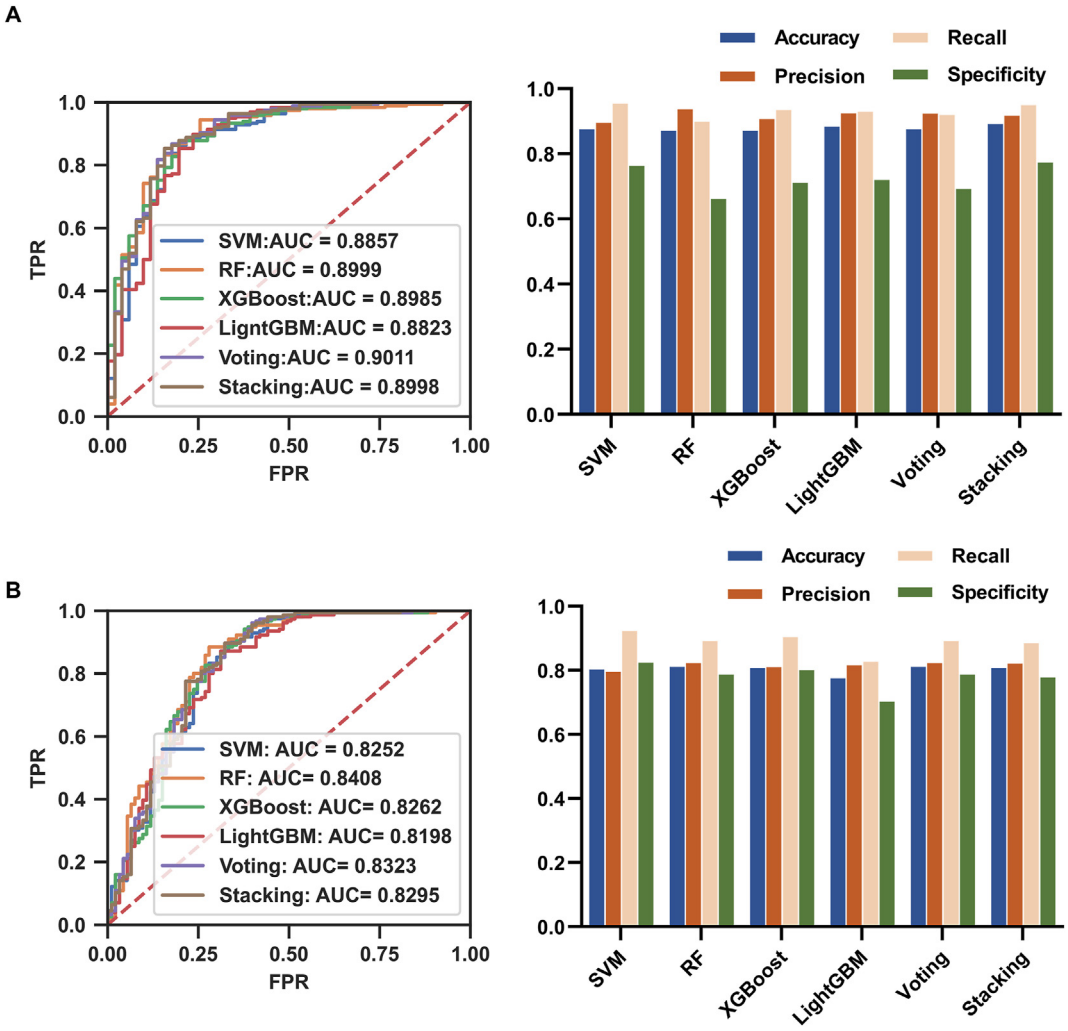


Figure 3 Performance evaluation of models using A549 cell line. Receiver operating characteristic curve plot (left) and histogram (right) of the accuracy, precision, recall, and specificity of the models achieved by the six machine learning algorithms for 50% cell viability prediction (A) and 80% cell viability prediction (B). AUC, area under the curve; FPR, false positive rate; TPR, true positive rate.

directly from raw data to enable end-to-end prediction and offer flexible architectures for stable and scalable training with large datasets. Therefore, four DL algorithms, including multilayer perceptron, convolutional neural network TabResnet, and TabNet, were also utilized to develop prediction models and the AUROCs of the four models ranged from 0.4851 to 0.8751 (Supporting Information Table S15). The unfavorable performance of DL models might be due to the high-dimensional data (over 5000 features) and relatively small size of samples (1243 compounds), which made the models overfit.

The generalizability of the optimal models across cell lines was evaluated in the HEK293 (kidney), HepG2 (liver), and HT-29 (colon) cell lines. Gene expression profile data for the cells treated with the test compounds and relevant cell viability data (Supporting Information Tables S16−S18 and Fig. S3) were used for this evaluation. Both the voting model (50% cell viability) and the RF model (80% cell viability) exhibited good performance (Supporting Information Table S19), with AUROC values ranging from 0.7898 to 0.9318 and 0.7622 to 0.8146, respectively, across the three cell lines (Fig. 4), indicating the two optimal models could be used for cell viability prediction across cell lines. In addition to AUROCs, "precision" and "recall" are also important in this case as they represent the ability to select nontoxic substances, with "precision" representing the proportion of the predicted nontoxic substances that are truly nontoxic, and "recall" representing the ratio of truly predicted nontoxic samples to the total number of actual nontoxic samples. As shown in Table S19, the precision and recall scores of two optimal models ranged from 0.6969 to 0.9217 and from 0.8023 to 0.9833, respectively, indicating that the models have a strong ability to predict nontoxic substances and could contribute to the identification of "safe" substances for drug development. Additionally, regression models also were developed, and their Pearson correlation coefficients ranged from 0.7025 to 0.7422 (Supporting Information Fig. S4). In summary, these findings indicate that, by using the models built in this study, cytotoxicity at different thresholds can be accurately predicted in diverse cell lines, which could aid in primary cytotoxicity screening across cell lines from massive amounts of transcriptome data.

3.4. *Interpretability analysis of the optimal models*

Interpretability is pivotal because it provides information on the decision-making process of the model and uncovers elemental features. To determine the importance of the Feature Genes for cytotoxicity classification, two optimal models for 50% and 80% cell viability prediction, the voting model and the RF model were interpreted using SHAP analysis. SHAP is a feature attribution

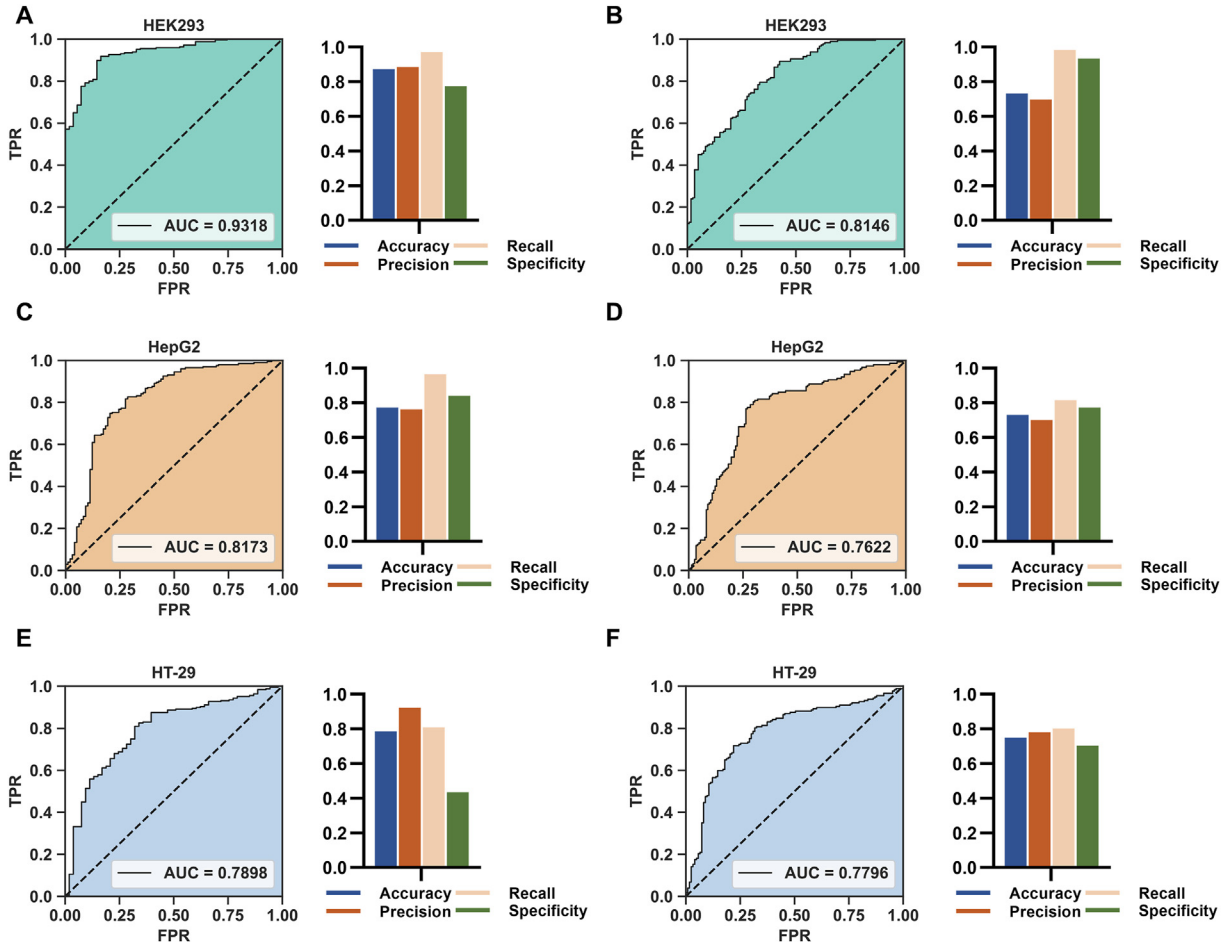


Figure 4 Performance of the optimal models in HEK293, HepG2, and HT-29 cells. (A, C, E) Performance of the voting model for 50% cell viability prediction in HEK293, HepG2, and HT-29 cells. (B, D, F) Performance of the RF model for 80% cell viability prediction in HEK293, HepG2, and HT-29 cells.

method that estimates the contribution of each feature to the prediction. The absolute value of SHAP represents the extent of feature attribution, while positive and negative SHAP values indicate the contributions of the feature to the prediction of non-cytotoxicity and cytotoxicity, respectively. In this study, the SHAP values among samples in the training set were calculated (Supporting Information Tables S20 and S21), and the average absolute SHAP values across samples were used to measure the global importance of features in the model (Fig. 5A and B). Additionally, the distributions of the SHAP values of the Ensemble Features among the samples in the training set were displayed, which revealed the contribution of each Feature Gene to a certain sample (Fig. 5C and D). As expected, for both the voting model (50% cell viability prediction) and the RF model (80% cell viability prediction), the top 10 features with the highest feature importance were involved in cell cycle events (*CCNA2*, *CDK1*, and *RUVBL1* in the voting model; *TOP2A*, *CCNA2*, *CDK1*, *POLE2*, *MCM10*, *TMPO*, and *CENPX* in the RF model; Fig. 5A and B). Of the remaining differences between the two models due to cytotoxicity level variance, the top 10 genes of the voting model (50% cell viability prediction) were involved in metabolic and immune pathways (*ATP8B4*, *DERA*), and the top 10 genes of the RF model (80% cell viability prediction) were involved in the cellular stress response (*TSC22D3*, *HES1*) (Fig. 5A and B). In addition, the distributions of individual gene expression patterns and SHAP values among the samples were analyzed. As shown in Fig. 5E and F, the expression level of the top 10 features of the two optimal models, the voting model for 50% cell viability prediction and the RF model for 80% cell viability prediction was correlated with the direction of cytotoxicity/viability. The results showed that the interpretation of the prediction models provided insights into cytotoxicity, including the quantification of feature importance and the direction of regulation. As important features made vital contributions to models, two re-fit models (the voting model for 50% cell viability prediction and RF for 80% cell viability prediction) were developed with their top 10 features, and the

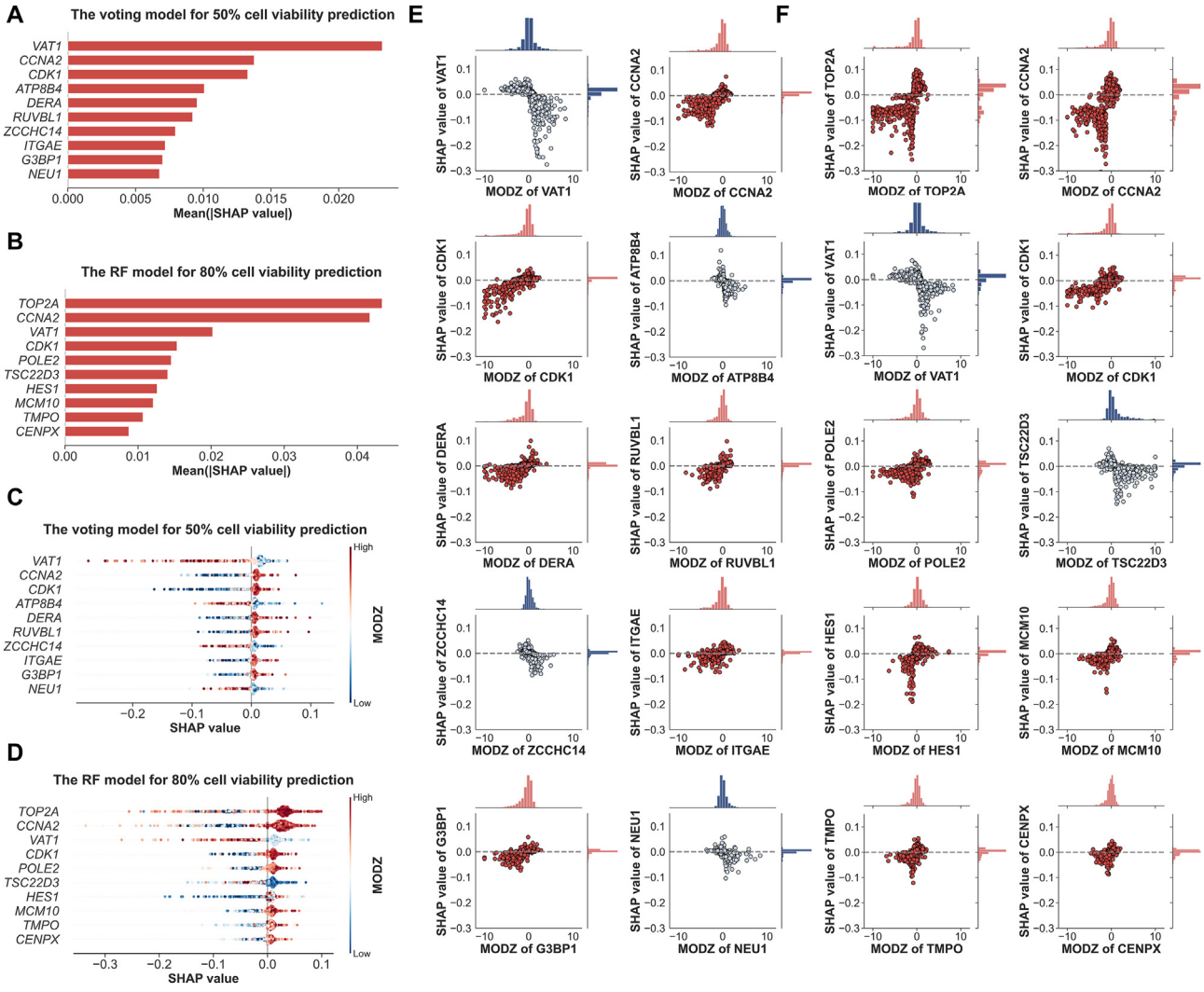


Figure 5 Interpretability analysis of the optimal cell viability prediction models. SHAP was applied to the training set (n = 994) for interpretability analysis. (A, B) Bar plots of the mean absolute SHAP values of the top 10 Feature Genes of the voting model (A) and the RF model (B). (C, D) Bee swarm plots of the SHAP values of the top 10 Feature Genes in the voting model (C) and the RF model (D). Each dot in the plot represents a sample; the positive and negative values indicate the feature contributions to the prediction of noncytotoxicity and cytotoxicity, respectively. (E, F) The sample scatter plots of the MODZ and SHAP values of the top 10 Feature Genes with the highest absolute SHAP values in the voting model (E) and the RF model (F) are displayed.

AUROCs of the two re-fit models across four cell lines are shown in Fig. 6. Compared with the AUROCs of the models with all ensemble features, the re-fit models also exhibited good performance (0.77 to 0.87 by re-fit voting *vs.* 0.79 to 0.93 by voting model; 0.79 to 0.84 by re-fit RF *vs.* 0.76 to 0.81 by RF model), indicating that the important features obtained by SHAP analysis had substantial impacts on cytotoxicity/viability prediction. Notably, the good performance and generalization of the 10-features models across cell lines may provide an opportunity for cytotoxicity/safety prediction with minimal cellular transcription information, which will make the models widely applicable.

3.5. Mechanistic analysis of bioactive compounds with a NTI by removal of CTS genes

As high-dimensional data, cellular transcriptomic data represent molecular changes that embody valuable information for compound pharmacological mechanism studies. However, for NTI compounds, the median therapeutic concentration is close to or even overlaps with the cytotoxic concentration, resulting in the co-occurrence of biofunctional activities and toxic effects that obscure the primary mechanisms involved. To exclude confounding effects, the CTS genes, which represent the major gene

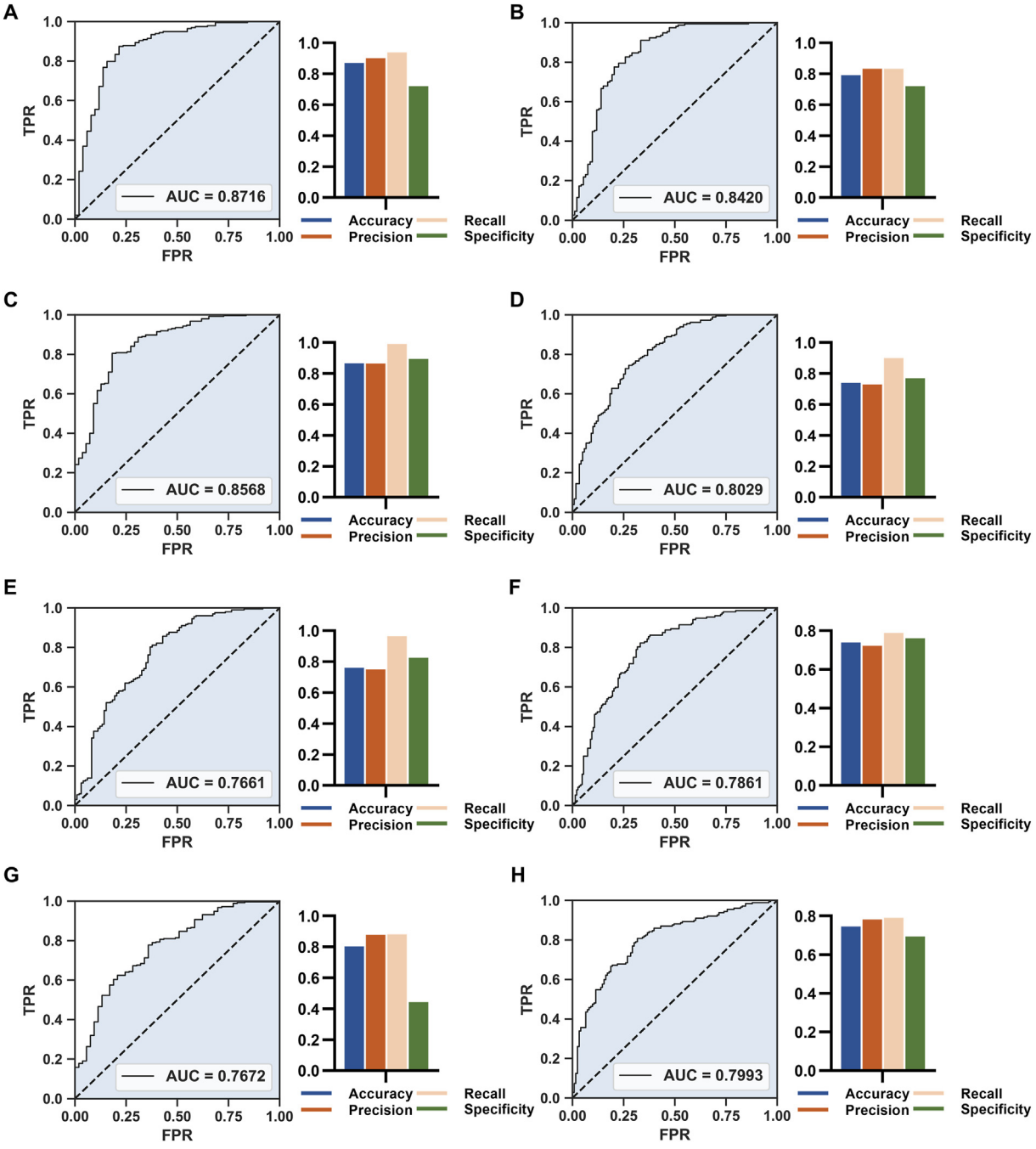


Figure 6 Performance of the re-fit models in A549, HEK293, HepG2, and HT-29 cells. (A, C, E, G) Performance of the re-fit voting model for 50% cell viability prediction in A549, HEK293, HepG2 and HT-29 cells. (B, D, F, H) Performance of the re-fit RF model for 80% cell viability prediction in A549, HEK293, HepG2 and HT-29 cells.

group causing cytotoxicity, were measured and defined. First, a total of 41 genes, which had higher mean absolute SHAP values than the average absolute SHAP value of 228 Feature Genes in the RF model, were acquired as the most relevant subset of genes associated with cytotoxicity (Fig. 7A; Supporting Information Table S22). The variable genes correlated with each gene in the subset were subsequently measured with a correlation coefficient greater than 0.5 as the threshold, and 545 genes (Fig. 7A; Supporting Information Table S23) were defined as CTS genes. Subsequently, to investigate the application of CTS genes in the MOA study, two known NTI compounds, cyclosporine A and triptolide, were analyzed by using transcriptome data with or without CTS gene removal to verify the contribution of these CTS genes to the MOA study of NTI compounds.

3.5.1. *Case No. 1 Identification of cyclosporine A MOAs by excluding CTS genes*

Cyclosporine A (CsA) is an immunosuppressant with a NTI[41]. CsA binds directly to intracellular cyclophilin A to form the CsA-cyclophilin A complex, which subsequently binds firmly to both the catalytic and regulatory subunits of calcineurin[42,43]. This binding leads to the inhibition of calcineurin phosphatase activity and consequent events, including the dephosphorylation of nuclear factor of activated T cells and the transcription of cytokines[42,43]. This results in an immunosuppressive effect that is widely utilized in organ transplantation to prevent liver, kidney, intestine, lung, and heart rejection[42,43]. In addition, CsA can increase butyrate uptake by inhibiting calcineurin activity and upregulating the expression of monocarboxylate transporter-1, ultimately reducing intestinal epithelial damage[44,45]. In addition, several clinical studies have shown that CsA can alleviate the clinical symptoms of hormone-refractory severe inflammatory bowel disease[46,47] and decrease the probability of colectomy in patients with acute severe ulcerative colitis when used as the last line of treatment[48].

In this study, RNA-seq was performed on HT-29 cells (colon) that were treated with CsA (10 μmol/L, 48 h); the results showed low cytotoxicity with cell viability of 75.5% (Supporting Information Table S24). The DEGs were assessed (Supporting Information Table S25) and the top 10 MOA terms were enriched according to all DEGs or DEGs with the CTS genes removal. As shown in Fig. 7B, the ATPase and AKT inhibitors were both enriched according to DEGs or DEGs associated with CTS gene removal, implying that the mechanism of CsA function in the protein kinase B (AKT)/mechanistic target of rapamycin (mTOR)/nuclear factor-kappa B (NF-κB) pathway[49] and in Na^+/K^+-ATPase activity[50] involves direct action on cyclophilin. However, after excluding CTS-related genes for enrichment, CsA appeared to be positively correlated with calcineurin inhibition, which was consistent with its immunosuppressive and intestinal epithelial protective effects (Fig. 7B; Supporting Information Table S26), indicating that the exclusion of these genes would facilitate the discovery of the MOA of CsA at a concentration with low cytotoxicity.

3.5.2. *Case No. 2 Triptolide MOAs analysis of the DEGs associated with CTS gene removal*

Triptolide, a natural product isolated from the herb *Tripterygium wilfordii* Hook. F. suis has multiple biological activities[51,52], and its therapeutic index is narrow, ranging from 1 to 4[53−55]. As reported, triptolide can directly bind to several proteins, including the transcription factor IIH (TFIIH) XPB subunit[56], TAK1-binding protein (TAB1)[57], and polycystin-2[34], and it can exert various biological effects, such as anti-pulmonary fibrosis[58−61], liver protection[62,63], and liver fibrosis effects[64]. In this study, the effects of triptolide on lung cells (A549) and liver cells (HepG2) were investigated. After triptolide treatment for 48 h, the A549 cells (48.4% cell viability; Supporting Information Table S24) and the HepG2 cells (49.3 % cell viability; Table S24) were subjected to RNA-seq, and the DEGs were assessed (Supporting Information Tables S27 and S28).

The MOA of triptolide on lung cells was analyzed based on the DEGs associated with triptolide in A549 cells and the DEGs associated with CTS gene removal (Fig. 7C, Table S26). As reported, TFIIH, an important target of triptolide, consists of multiple subunits with various enzymatic activities and is involved in regulating transcriptional activity[65]. Triptolide can covalently bind to XPB[56], a subunit of the transcription factor TFIIH, which leads to the inhibition of ATPase activity and RNA polymerase II-mediated transcription. As shown in Fig. 7C, the DEGs associated with triptolide were positively correlated with protein kinase inhibitors, ATPase inhibitors and NF-κB pathway inhibitors before and after the removal of the CTS genes. This finding was consistent with the observed inhibition of transcription and ATPase activity mentioned above. Additionally, it was reported that the expression of heat shock protein 90 (HSP90) can be repressed when TFIIH is inhibited by triptolide[52,66], and the repression of HSP90 blocks the TGF-β signaling pathway, reduces the extracellular matrix, inhibits epithelial−mesenchymal transition, and alleviates pulmonary fibrosis[67]. As shown in Fig. 7C, after removing the CTS genes from the triptolide/A549 DEGs, the term "HSP inhibitor" was enriched. These findings provide important information for investigating the mechanism of action of triptolide in preventing pulmonary fibrosis. In addition, glucocorticoids are widely used in the clinical treatment of lung diseases[68,69], and studies have shown that miR-142-5p and miR-181a downregulate glucocorticoid receptor expression[70] and that triptolide upregulates glucocorticoid receptor expression by inhibiting miR-142-5p and miR-181a transcription[71]. As shown in Fig. 7C, the results revealed enrichment of glucocorticoid receptor agonists after removal of the CTS genes, which was consistent with the findings of previous studies.

Triptolide also plays roles in liver protection[62,63] and protection against liver fibrosis[64]. Previous studies demonstrated that triptolide could repress transforming growth factor β-activated kinase 1 (TAK1) activity by directly binding to TAB1 and interfering with the formation of the TAK1−TAB1 complex. The TAK1−TAB1 complex[72] plays a critical role in regulating inflammation by phosphorylating IκB kinase (IKK) and mitogen-activated protein kinase (MAPK) kinase[73], leading to the activation of NF-κB[73] and the phosphorylation of c-Jun N-terminal kinase (JNK) and P38, respectively. Subsequently, the transcription factors c-fos and c-Jun are phosphorylated by JNK; thus, their expression[74] and activity increase. Activation of NF-κB and AP-1 promotes the expression of proinflammatory cytokines in hepatocytes[75], which leads to liver damage and fibrosis[76]. By binding to TAB1, triptolide inhibits transcriptionally regulated NF-κB and AP-1, thereby exerting hepatoprotective and anti-liver fibrosis effects. In this study, the MOA of triptolide on liver cells was also investigated by using the transcriptome data of triptolide-treated HepG2 cells. Treatment with triptolide strongly correlated with transcriptional inhibition (a topoisomerase inhibitor) based on DEGs or DEG data excluding the CTS genes (Fig. 7C). After removal of the CTS genes from the DEGs, the terms "JNK inhibitor" and "IKK inhibitor" were enriched, suggesting that the

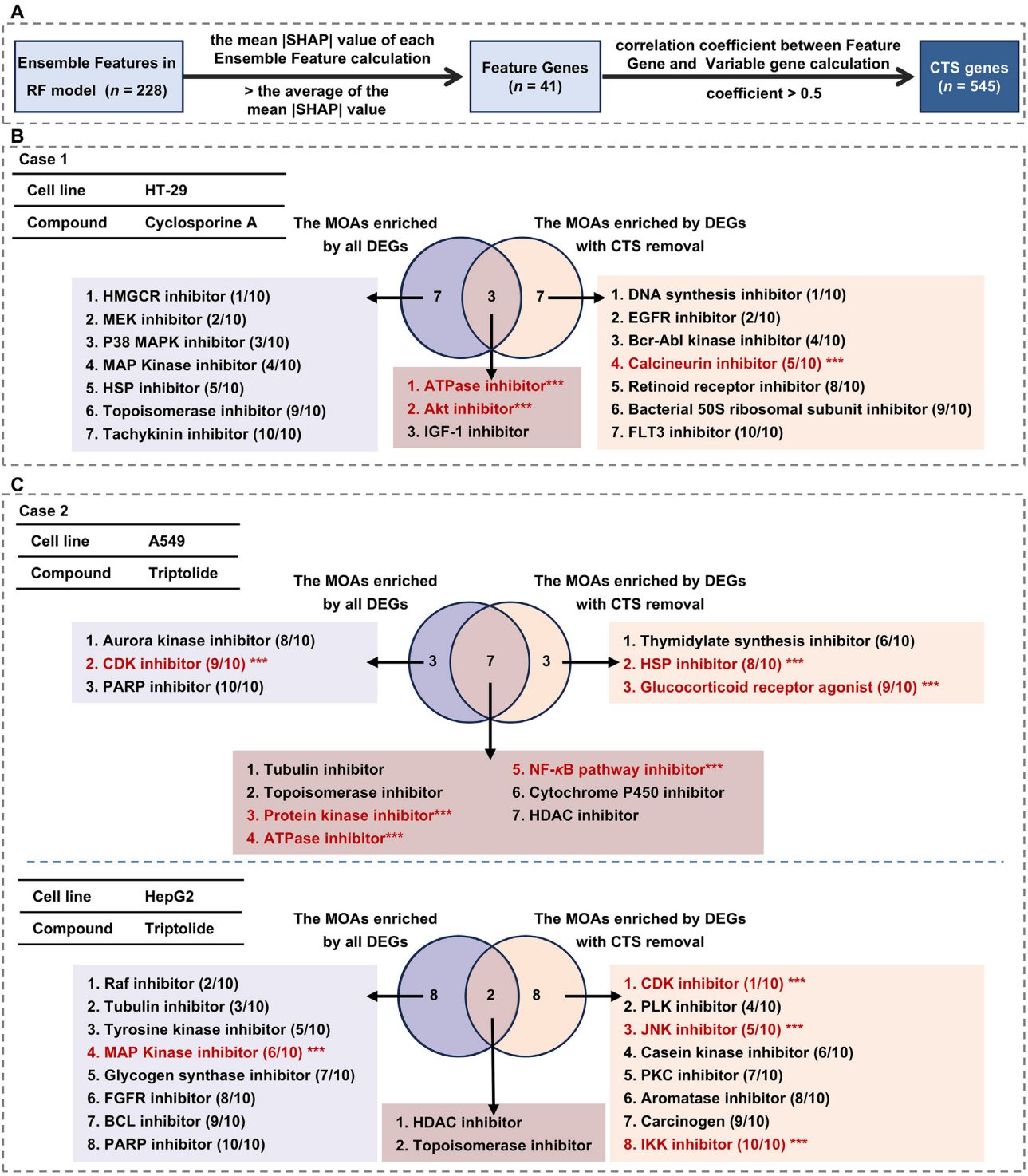


Figure 7 CTS genes facilitating the MOA study of the compounds with NTI. (A) The pipeline for identifying CTS genes in the RF models. (B, C) The MOAs for cyclosporine A (B) and triptolide (C) were predicted using DEGs of the compound and the DEGs with removal of CTS genes. The two sets of top 10 MOAs identified by using the compound DEGs (in lavender gray) and the DEGs associated with CTS gene removal (in apricot) are shown in the Venn diagram. The target-related terms are shown in red and marked with asterisks (***). The ranking for each MOA term is indicated in parentheses. The MOA enrichment was analyzed using the Query platform (https://clue.io/query). BCL, B cell lymphoma; CDK, cyclin-dependent kinase; EGFR, epidermal growth factor receptor; FGFR, fibroblast growth factor receptor; FLT3, FMS-like tyrosine kinase 3; HDAC, histone deacetylase; HMGCR, 3-hydroxy-3-methylglutaryl-coenzyme A reductase; IGF-1, insulin-like growth factor 1; MEK, mitogen-activated extracellular signal-regulated kinase; PARP, poly(ADP-ribose) polymerase; PKC, protein kinase C.

hepatoprotective mechanism of triptolide occurs through TAK1-TAB1 inhibition. According to the results of the MOA analysis of two well-known NTI compounds, cyclosporine A and triptolide, the removal of CTS genes from the DEGs indeed provided additional clues to the primary target of the active substances with NTIs.

4. Discussion

Drug safety is of the utmost importance throughout the drug development and is the primary issue causing drug failure[1−3]. The early application of cytotoxicity evaluation *in vitro* could provide information on human-specific toxicity, which is important for drug development[77]. Chemical structural characteristics have been utilized for toxicity prediction[18,77,78]. Nevertheless, the assessment of substance toxicity encompasses intricate factors, and structural information is insufficient to adequately capture the toxic concentration of a substance or the variances in its cytotoxicity in different cell lines. Furthermore, the existence of "activity cliffs" (pairs of compounds that are structurally similar but display significant variations in toxicity) and other variables restrict the ability of toxicity models to predict toxicity based on structural information[78−80]. Compared with structural information, omics data, which are rarely applied in toxicity prediction, reveal alterations in cellular gene expression levels in response to compound exposure, precisely capturing the comprehensive impacts of certain concentrations on distinct tissues/cells[81]. Additionally, the omics-based prediction model can be applied to omics profiling of any substances, including macromolecules and mixtures. Furthermore, the transcriptomic features enable the elucidation of the mechanism of action of toxic substances, including the initiating bio-events and the cascade of events, as well as the identification of toxicity-related biomarkers[82,83]. This study constructed a good cytotoxicity prediction model by using transcriptome data; thus, this model provides a novel tool for drug safety screening through gene transcription evaluation. Notably, most available toxicity prediction models are structure-based and use the chemical properties of the substance as features[84−87]. The models in this study offer predictions from biological perspective of cells, *i.e.*, the features of cells following their perturbation by substances. Thus, coupling the biological feature-based model established in this research and a structural feature-based model for prediction, or a model based on integrating the two types of features, should greatly improve the identification of highly safe substances.

Data quality is also a vital issue in machine learning model development and model performance[88]. Currently, most prediction models are trained by using open-source toxicity data of thousands of chemical compounds from various assays; these models inevitably include false positive and false negative data, decreasing the accuracy of the models. To construct highly accurate models in this study, cell viability data were measured and collected under the corresponding conditions of compound transcriptomic data determination to minimize label noise and the influence of data heterogeneity on model construction. Consequently, the optimal model achieved in this study exhibited superior performance (AUROC = 0.90) comparing to the reported transcriptome-based toxicity classification models, which had AUROCs ranging from 0.72 to 0.80[17,18,40]. In addition, all the reported studies have shown the importance of the 50% toxicity threshold for these compounds; however, weak toxicity limits drug development to a great extent. Therefore, in addition to 50%, 80% cell viability was used as the threshold for distinguishing weakly toxic substances in this study. The optimal model for determining 80% cell viability also displayed good performance (AUROC = 0.84) and could be used as a valuable tool for identifying nontoxic substances. Furthermore, the good prediction performance of the optimal models across cell lines sourced from different tissues (Table S19) suggests that the models have the potential to be applied across diverse cell lines. In addition, for the first time, this study implements interpretability analysis of the models to identify the critical features by SHAP analysis iterating through all the transcriptomic data of the compounds, which provides guidance for toxicity and bioactivity mechanism analysis of individual substances and can be conducted as MOAs analysis of compounds with a narrow therapeutic window.

In summary, by utilizing compound transcriptomic and cell viability data, the toxicity prediction models developed in this study could be valuable tools for efficiently predicting the cytotoxicity of numerous substances in different cell lines with high accuracy and providing information on the potential mechanisms of both toxicity and bioactivity.

5. Conclusions

Cytotoxicity prediction models were constructed in this study which provided accurate tools not only for distinguishing substances with 50% cytotoxicity as the threshold but also for identifying weakly toxic substances with a threshold of 20% cytotoxicity. In addition, these models can be utilized in diverse cell lines for all substances, including small molecules, biological macromolecules, nanomaterials, and mixtures. Moreover, the models in this study were interpreted, and for the first time, a CTS gene list was generated which could facilitate the MOA study of the substances with a narrow therapeutic index.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (82204472), the CAMS Innovation Fund for Medical Sciences (2022-I2M-1-014, 2021-I2M-1-028, and 2022-I2M-2-002, China).

Author contributions

You Wu: Writing − review & editing, Writing − original draft, Software, Methodology, Investigation, Formal analysis, Data curation. Ke Tang: Writing − review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation. Chunzheng Wang: Writing − review & editing, Investigation. Hao Song: Writing − review & editing, Investigation. Fanfan Zhou: Writing − review & editing, Investigation. Ying Guo: Writing − review & editing, Writing − original draft, Supervision, Data curation, Conceptualization.

Conflicts of interest

The authors declare that they have no competing interests.

Appendix A. Supplementary information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2025.02.009.

References

1. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun* 2018;**11**:156−64.
2. Guengerich FP. Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug Metab Pharmacokinet* 2011;**26**:3−14.
3. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it?. *Acta Pharm Sin B* 2022;**12**:3049−62.
4. Adan A, Kiraz Y, Baran Y. Cell proliferation and cytotoxicity assays. *Curr Pharm Biotechnol* 2016;**17**:1213−21.
5. Johnson S, Nguyen V, Coder D. Assessment of cell viability. *Curr Protoc Cytom* 2013;**64**:9.2.1−9.2.26.
6. Aslantürk ÖS. *In vitro* cytotoxicity and cell viability assays: principles, advantages, and disadvantages. In: *Genotoxicity-A predictable risk to our actual world*, **2**. Rijeka: InTech Open; 2018. p. 64−80.
7. Benbow JW, Aubrecht J, Banker MJ, Nettleton D, Aleo MD. Predicting safety toleration of pharmaceutical chemical leads: cytotoxicity correlations to exploratory toxicity studies. *Toxicol Lett* 2010;**197**:175−82.
8. Ramirez CN, Antczak C, Djaballah H. Cell viability assessment: toward content-rich platforms. *Expert Opin Drug Discov* 2010;**5**:223−33.
9. Nguyen N, Jennen D, Kleinjans J. Omics technologies to understand drug toxicity mechanisms. *Drug Discov Today* 2022;**27**:103348.
10. D'Adamo GL, Widdop JT, Giles EM. The future is now? Clinical and translational aspects of "Omics" technologies. *Immunol Cell Biol* 2021;**99**:168−76.
11. Xin Y, Chen S, Tang K, Wu Y, Guo Y. Identification of nifurtimox and chrysin as anti-influenza virus agents by clinical transcriptome signature reversion. *Int J Mol Sci* 2022;**23**:2372.
12. Chen S, Xin Y, Tang K, Wu Y, Guo Y. Nardosinone and aurantio-obtusin, two medicine food homology natural compounds, are anti-influenza agents as indicated by transcriptome signature reversion. *Phytomedicine* 2023;**108**:154515.
13. Tang K, Wu Y, Chen S, Xin Y, Guo Y. Discovery of the anti-influenza A virus activity of SB216763 and cyclosporine A by mining infected cells and compound cellular signatures. *Chin Chem Lett* 2022;**33**:2541−4.
14. Eisenstein M. Machine learning powers biobank-driven drug discovery. *Nat Biotechnol* 2022;**40**:1303−5.
15. Eisenstein M. Active machine learning helps drug hunters tackle biology. *Nat Biotechnol* 2020;**38**:512−4.
16. Wang W, Feng S, Ye Z, Gao H, Lin J, Ouyang D. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. *Acta Pharm Sin B* 2022;**12**:2950−62.
17. Lu J, Chen M, Qin Y. Drug-induced cell viability prediction from LINCS-L1000 through WRFEN-XGBoost algorithm. *BMC bioinformatics* 2021;**22**:13.
18. Hao Y, Romano JD, Moore JH. Knowledge-guided deep learning models of drug toxicity improve interpretation. *Patterns (N Y)* 2022;**3**:100565.
19. Tang K, Zhang X, Guo Y. Identification of the dietary supplement capsaicin as an inhibitor of Lassa virus entry. *Acta Pharm Sin B* 2020;**10**:789−98.
20. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884−90.
21. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357−60.
22. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* 2011;**12**:R22.
23. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**:166−9.
24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
25. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**. 1437-52.e17.
26. Enache OM, Lahr DL, Natoli TE, Litichevskiy L, Wadden D, Flynn C, et al. The GCTx format and cmap {Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices. *Bioinformatics* 2019;**35**:1427−9.
27. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579−605.
28. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
29. De Jay N, Papillon Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics* 2013;**29**:2365−8.
30. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. *bioRxiv* 2016. https://doi.org/10.1101/060012.
31. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**:573−80.
32. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for largescale machine learning. In: *Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI '16)*. Savannah: USENIX Association; 2016. p. 265−83.
33. Zaurin JR, Mulinka P. Pytorch-widedeep: a flexible package for multimodal deep learning. *JOSS* 2023;**8**:5027.
34. Arik SÖ, Pfister T. TabNet: attentive interpretable tabular learning. In: *Proceedings of the 35th AAAI conference on artificial intelligence (AAAI 2021)*, **35**; 2021. p. 6679−87. Available from: https://arxiv.org/pdf/1908.07442v5.
35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan SVN, et al., editors. *Advances in neural information processing systems 30 (NIPS 2017)*. New York: Curran Associates; 2017. p. 4765−74.
36. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019;**10**:1523.
37. Fulda S, Gorman AM, Hori O, Samali A. Cellular stress responses: cell survival and cell death. *Int J Cell Biol* 2010;**2010**:214074.
38. Eisner V, Picard M, Hajnóczky G. Mitochondrial dynamics in adaptive and maladaptive cellular stress responses. *Nat Cell Biol* 2018;**20**:755−65.
39. Kohonen P, Parkkinen JA, Willighagen EL, Ceder R, Wennerberg K, Kaski S, et al. A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat Commun* 2017;**8**:15932.
40. Szalai B, Subramanian V, Holland CH, Alföldi R, Puskás LG, Saez Rodriguez J. Signatures of cell death and proliferation in perturbation transcriptomics data-from confounding factor to effective prediction. *Nucleic Acids Res* 2019;**47**:10010−26.
41. Tapia C, Nessel TA, Zito PM. Cyclosporine. In: *StatPearls. Treasure Island, FL*. StatPearls Publishing; 2023.
42. Barbarino JM, Staatz CE, Venkataramanan R, Klein TE, Altman RB. PharmGKB summary: cyclosporine and tacrolimus pathways. *Pharmacogenet Genomics* 2013;**23**:563−85.
43. Kapturczak MH, Meier Kriesche HU, Kaplan B. Pharmacology of calcineurin antagonists. *Transpl Proc* 2004;**36**:25S−32S.
44. Ota S, Sakuraba H, Hiraga H, Yoshida S, Satake M, Akemoto Y, et al. Cyclosporine protects from intestinal epithelial injury by modulating butyrate uptake *via* upregulation of membrane monocarboxylate transporter 1 levels. *Biochem Biophys Rep* 2020;**24**:100811.
45. Suwa M, Nakano H, Kumagai S. Inhibition of calcineurin increases monocarboxylate transporters 1 and 4 protein and glycolytic enzyme activities in rat soleus muscle. *Clin Exp Pharmacol Physiol* 2005;**32**:218−23.
46. Lichtiger S, Present DH, Kornbluth A, Gelernt I, Bauer J, Galler G, et al. Cyclosporine in severe ulcerative colitis refractory to steroid therapy. *N Engl J Med* 1994;**330**:1841−5.

47. Weissman S, Chris Olaiya A, Mehta TI, Aziz M, Alshati A, Berry R, et al. A novel player: cyclosporine therapy in the management of inflammatory bowel disease. *Transl Gastroenterol Hepatol* 2019;**4**:67.
48. Cohen RD. Intravenous cyclosporine in severe ulcerative colitis: ready to stand alone? *Gastroenterology* 2001;**120**:1541−3.
49. Xin S, Liu L, Li Y, Yang J, Zuo L, Cao P, et al. Cyclophilin A binds to AKT1 and facilitates the tumorigenicity of Epstein-Barr virus by mediating the activation of AKT/mTOR/NF-κB positive feedback loop. *Virol Sin* 2022;**37**:913−21.
50. Suñé G, Sarró E, Puigmulé M, López Hellín J, Zufferey M, Pertel T, et al. Cyclophilin B interacts with sodium-potassium ATPase and is required for pump activity in proximal tubule cells of the kidney. *PLoS One* 2010;**5**:e13930.
51. Zhou ZL, Yang YX, Ding J, Li YC, Miao ZH. Triptolide: structural modifications, structure−activity relationships, bioactivities, clinical development and mechanisms. *Nat Prod Rep* 2012;**29**:457−75.
52. Tong L, Zhao Q, Datan E, Lin GQ, Minn I, Pomper MG, et al. Triptolide: reflections on two decades of research and prospects for the future. *Nat Prod Rep* 2021;**38**:843−60.
53. Shamon LA, Pezzuto JM, Graves JM, Mehta RR, Wangcharoentrakul S, Sangsuwan R, et al. Evaluation of the mutagenic, cytotoxic, and antitumor potential of triptolide, a highly oxygenated diterpene isolated from *Tripterygium wilfordii*. *Cancer Lett* 1997;**112**:113−7.
54. Zheng J. Screening of active anti-inflammatory, immunosuppressive and antifertility components of *Tripterygium wilfordii*. III. A comparison of the antiinflammatory and immunosuppressive activities of 7 diterpene lactone epoxide compounds *in vivo*. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* 1991;**13**:391−7.
55. Li XJ, Jiang ZZ, Zhang LY. Triptolide: progress on research in pharmacodynamics and toxicology. *J Ethnopharmacol* 2014;**155**:67−79.
56. Titov DV, Gilman B, He QL, Bhat S, Low WK, Dang Y, et al. XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nat Chem Biol* 2011;**7**:182−8.
57. Lu Y, Zhang Y, Li L, Feng X, Ding S, Zheng W, et al. TAB1: a target of triptolide in macrophages. *Chem Biol* 2014;**21**:246−56.
58. Guo K, Chen J, Chen Z, Luo G, Yang S, Zhang M, et al. Triptolide alleviates radiation-induced pulmonary fibrosis *via* inhibiting IKKβ stimulated LOX production. *Biochem Biophys Res Commun* 2020;**527**:283−8.
59. Chen C, Yang S, Zhang M, Zhang Z, Zhang SB, Wu B, et al. Triptolide mitigates radiation-induced pneumonitis *via* inhibition of alveolar macrophages and related inflammatory molecules. *Oncotarget* 2017;**8**:45133−42.
60. Chen H, Chen Q, Jiang CM, Shi GY, Sui BW, Zhang W, et al. Triptolide suppresses paraquat induced idiopathic pulmonary fibrosis by inhibiting TGFβ1-dependent epithelial mesenchymal transition. *Toxicol Lett* 2018;**284**:1−9.
61. Wang L, Li S, Yao Y, Yin W, Ye T. The role of natural products in the prevention and treatment of pulmonary fibrosis: a review. *Food Funct* 2021;**12**:990−1007.
62. Lu Y, Bao X, Sun T, Xu J, Zheng W, Shen P. Triptolide attenuate the oxidative stress induced by LPS/D-GalN in mice. *J Cel Biochem* 2012;**113**:1022−33.
63. Yang YQ, Yan XT, Wang K, Tian RM, Lu ZY, Wu LL, et al. Triptriolide alleviates lipopolysaccharide-induced liver injury by Nrf2 and NF-κB signaling pathways. *Front Pharmacol* 2018;**9**:999.
64. Chong LW, Hsu YC, Chiu YT, Yang KC, Huang YT. Antifibrotic effects of triptolide on hepatic stellate cells and dimethylnitrosamine-intoxicated rats. *Phytother Res* 2011;**25**:990−9.
65. Kolesnikova O, Radu L, Poterszman A. TFIIH: a multi-subunit complex at the cross-roads of transcription and DNA repair. *Adv Protein Chem Struct Biol* 2019;**115**:21−67.
66. Zhang FZ, Ho DH, Wong RH. Triptolide, a HSP90 middle domain inhibitor, induces apoptosis in triple manner. *Oncotarget* 2018;**9**:22301−15.
67. Zhang X, Zhang X, Huang W, Ge X. The role of heat shock proteins in the regulation of fibrotic diseases. *Biomed Pharmacother* 2021;**135**:111067.
68. Ito K, Getting SJ, Charron CE. Mode of glucocorticoid actions in airway disease. *Sci World J* 2006;**6**:1750−69.
69. Niewoehner DE, Erbland ML, Deupree RH, Collins D, Gross NJ, Light RW, et al. Effect of systemic glucocorticoids on exacerbations of chronic obstructive pulmonary disease. *N Engl J Med* 1999;**340**:1941−7.
70. Vreugdenhil E, Verissimo CS, Mariman R, Kamphorst JT, Barbosa JS, Zweers T, et al. MicroRNA 18 and 124a down-regulate the glucocorticoid receptor: implications for glucocorticoid responsiveness in the brain. *Endocrinology* 2009;**150**:2220−8.
71. Huang X, Yang M, Jin J. Triptolide enhances the sensitivity of multiple myeloma cells to dexamethasone *via* microRNAs. *Leuk Lymphoma* 2012;**53**:1188−95.
72. Hirata Y, Takahashi M, Morishita T, Noguchi T, Matsuzawa A. Post-translational modifications of the TAK1−TAB complex. *Int J Mol Sci* 2017;**18**:205.
73. Xu YR, Lei CQ. TAK1−TABs complex: a central signalosome in inflammatory responses. *Front Immunol* 2020;**11**:608976.
74. Whitmarsh AJ, Davis RJ. Transcription factor AP-1 regulation by mitogen-activated protein kinase signal transduction pathways. *J Mol Med Berl* 1996;**74**:589−607.
75. Sun B, Karin M. NF-κB signaling, liver disease and hepatoprotective agents. *Oncogene* 2008;**27**:6228−44.
76. Luedde T, Schwabe RF. NF-κB in the liver—linking injury, fibrosis and hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* 2011;**8**:108−18.
77. Badwan BA, Liaropoulos G, Kyrodimos E, Skaltsas D, Tsirigos A, Gorgoulis VG. Machine learning approaches to predict drug efficacy and toxicity in oncology. *Cell Rep Methods* 2023;**3**:100413.
78. Tran TTV, Surya Wibowo A, Tayara H, Chong KT. Artificial intelligence in drug toxicity prediction: recent advances, challenges, and future perspectives. *J Chem Inf Model* 2023;**63**:2628−43.
79. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model* 2022;**62**:5938−51.
80. Sheridan RP, Karnachi P, Tudor M, Xu Y, Liaw A, Shah F, et al. Experimental error, kurtosis, activity cliffs, and methodology: what limits the predictivity of quantitative structure-activity relationship models?. *J Chem Inf Model* 2020;**60**:1969−82.
81. Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. *Nat Rev Drug Discov* 2023;**22**:895−916.
82. Zhang Y. Cell toxicity mechanism and biomarker. *Clin Transl Med* 2018;**7**:34.
83. Alexander-Dann B, Pruteanu LL, Oerton E, Sharma N, Berindan Neagoe I, Módos D, et al. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. *Mol Omics* 2018;**14**:218−36.
84. Fu L, Shi S, Yi J, Wang N, He Y, Wu Z, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Res* 2024;**52**:W422−31.
85. Yin Z, Ai H, Zhang L, Ren G, Wang Y, Zhao Q, et al. Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints. *J Appl Toxicol* 2019;**39**:1366−77.
86. Webel HE, Kimber TB, Radetzki S, Neuenschwander M, Nazaré M, Volkamer A. Revealing cytotoxic substructures in molecules using deep learning. *J Comput Aided Mol Des* 2020;**34**:731−46.
87. Schrey AK, Nickel Seeber J, Drwal MN, Zwicker P, Schultze N, Haertel B, et al. Computational prediction of immune cell cytotoxicity. *Food Chem Toxicol* 2017;**107**:150−66.
88. Gong Y, Liu G, Xue Y, Li R, Meng L. A survey on dataset quality in machine learning. *Inform Softw Techn* 2023;**162**:107268.