Improved Early Detection of Drug-Induced Liver Injury by Integrating Predicted *in vivo* and *in vitro* Data

Srijit Seal[1,2], *Dominic P. Williams*[3,4], *Layla Hosseini-Gerami*[5], *Ola Spjuth**[6], *Andreas Bender**[1]

[1]Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Rd, CB2 1EW, Cambridge, United Kingdom

[2]Imaging Platform, Broad Institute of MIT and Harvard, US

[3] Safety Innovation, Clinical Pharmacology and Safety Sciences, AstraZeneca, Cambridge CB4 0FZ, United Kingdom

[4]Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge CB4 0FZ, United Kingdom

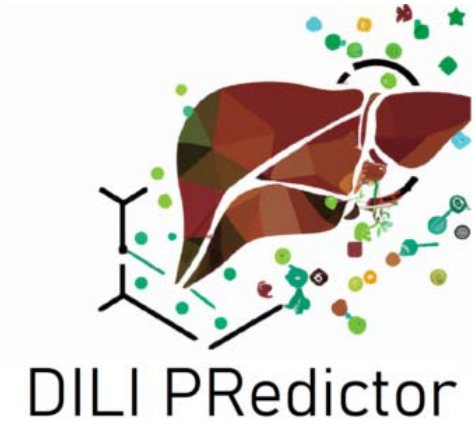
[5]Ignota Labs, County Hall, Westminster Bridge Rd, SE1 7PB, London, United Kingdom

[6]Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, Box 591, SE-75124, Uppsala, Sweden

* Email: ab454@cam.ac.uk, ola.spjuth@uu.se

Machine Learning, Toxicity, DILI, *in vivo*, *in vitro*, Drug-Induced Liver Injury

GRAPHICAL ABSTRACT



ABSTRACT

Drug-induced liver injury (DILI) presents a significant challenge in drug discovery, often leading to clinical trial failures and necessitating drug withdrawals. In this study, we introduce a novel method for DILI prediction that first predicts eleven proxy-DILI labels and then uses them as features in addition to chemical structural features to predict DILI. The features include *in vitro* (e.g., mitochondrial toxicity, bile salt export pump inhibition) data, *in vivo* (e.g., preclinical rat hepatotoxicity studies) data, pharmacokinetic parameters of maximum concentration, structural fingerprints, and physicochemical parameters. We trained DILI-prediction models on 1020 compounds from the DILIst dataset and tested on a held-out external test set of 255 compounds from DILIst dataset. The best model, DILIPredictor, attained a balanced accuracy of 70% and an LR+ score of 7.21. This model enabled the early detection of 26 toxic compounds compared to models using only structural features (4.62 LR+ score). Using feature interpretation from DILIPredictor, we were able to identify the chemical substructures causing DILI as well as differentiate cases DILI is caused by compounds in animals but not in humans. For example, DILIPredictor correctly recognized 2-butoxyethanol as non-toxic in humans despite its hepatotoxicity in mice models. Overall, the DILIPredictor model improves the detection of compounds causing DILI with an improved differentiation between animal and human sensitivity as well as the potential for mechanism evaluation. DILIPredictor is publicly available at https://broad.io/DILIPredictor for use *via* web interface and with all code available for download.

Toxicity Prediction; DILI; Drug-Induced Liver Injury

INTRODUCTION

The liver is a major organ of drug metabolism in the human body and thus it is vulnerable to not just drugs but also their reactive metabolites.[1,2] Drug-induced liver injury (DILI) is a leading cause of acute liver failure[3] (causing over 50% of such cases[4]) and accounts for a significant proportion of drug-related adverse events. DILI is detectable generally in phase III clinical trials and is also a leading cause of post-market drug withdrawals.[5] Two common types of DILI are intrinsic and idiosyncratic.[6] While intrinsic DILI is generally dose-dependent and predictable, idiosyncratic DILI is unpredictable with a variable onset time and several phenotypes and is generally not dependent on dosage. The mechanisms underlying DILI are multifactorial[7] and not completely understood. These include cellular toxicities such as mitochondrial impairment[8], inhibition of biliary efflux[9], oxidative stress[10], and more. Additionally, DILI can be influenced by dose variations, pharmacokinetics (PK), and biological variations, such as variations in cytochrome P450 (CYP) expression[11]. Thus, predicting DILI is a challenging task that necessitates consideration of various factors and novel methods to aid in the early detection of DILI and reduce related drug failures.

In the drug discovery pipeline, hepatotoxicity assessment encompasses a variety of *in vitro* and *in vivo* experimental models as well as *in silico* models**.** Several *in vitro* models for liver toxicity testing employ proxy endpoints (hepatotoxicity assays) with liver slices and cell lines such as primary animal and human hepatocytes[12] or even three-dimensional systems with the dynamic flow for the primary cell and/or stem cell cultures.[13] However, the ideal hepatocyte-like cell model system depends on the evaluation of particular cellular functions given there are substantial differences among various human liver-derived single-cell culture models as previously explored in the context of drug disposition, bioactivation, and detoxification.[14] The agreement between *in vitro* data and human *in vivo* data is also low.[15] For example,

methapyrilene is known to cause changes to the level of iron metabolism in the human hepatic HepaRG cell line[16] and oxidative stress, and mitochondrial dysfunction in rats[17] but has not been reported to cause hepatotoxicity in humans[18,19]. On the other hand, *in vivo* animal models also have low concordance as shown by recent studies using the eTOX database where organ toxicities were rarely concordant between species.[20] The concordance between animal and human data for liver toxicity, specifically, is often low (with some studies indicating rates as low as 40%[21] and others in the range of 39−44%[22]) which makes extrapolating safety assessments from animals to humans a challenging endeavour.[23,24] For example, 2-butoxyethanol causes hepatic toxicity in mice *via* an oxidative stress mechanism but not in humans given humans have higher levels of liver vitamin E (and a high resistance to iron accumulation) compared to mice.[25] Overall, this leads to a great need for improved DILI prediction from available data.

DILIst[26] and DILIrank[27] are lists of compounds that have been classified as inducing DILI or not and were developed from FDA-approved drug labels. Binary classification from labelling documents is challenging and this is evident in the fact that many DILIrank compounds are labelled ambiguous although the DILI for some of these compounds has been reported in literature. For *in silico* models, these ambiguous compounds are generally removed. Generally, *in silico* models rely on identifying chemical structural alerts[28] or use a range of chemical or physicochemical features. Ye et al. employed Random Forest algorithms and Morgan fingerprints for DILI prediction, achieving an AUC of 0.75 with random splitting (70% training, 30% testing).[29] Liu et al. utilized Support Vector Machines and obtained a 76% balanced accuracy on an external test set using Morgan Fingerprints; however, their predicted protein target descriptors provided less accurate predictions (balanced accuracy of 59%) but offered better interpretability.[30] Mora et al. employed QuBiLS-MAS 0–2.5D molecular descriptors to predict DILI (labels from various sources) on

an external test set comprising 554 compounds, achieving a 77% balanced accuracy.[31] Predicting organ-level toxicity solely based on chemical structure is challenging and the use of biological data helps improve toxicity prediction.[32,33] More recently, predicted off-target effects have also been considered to improve DILI prediction.[34] Chavan et al. integrated high-content imaging features with chemical features for DILI label prediction, resulting in a 0.74 AUC.[35] Previously this, the authors of this work explored this in the case of mitochondrial toxicity[36] (which at high doses is one of the mechanisms known to cause DILI) as well as cardiotoxicity[37].

In this study, we significantly extended the use of different data sources to several *in vivo* and *in vitro* data types in developing the DILIPredictor model presented here. We identified liver injury endpoints such as human hepatotoxicity[38], preclinical hepatotoxicity and animal hepatotoxicity[38,39,40] and DILI datasets compiled by various studies[31,41,42] (Table 1) These datasets provide the *in vivo* labels for DILI for different species at various stages of the drug discovery pipeline, from pre-clinical to post-market withdrawals. We identified three *in vitro* assays that could be indicative of liver toxicity and with public data: mitochondrial toxicity[43], bile salt export pump inhibition (BSEP)[44] and the formation of reactive metabolites[45].[7] Mitochondria accounts for 13-20% of the liver, and mitochondrial dysfunction can impact ATP synthesis, increase ROS generation and trigger liver injury.[46] The majority of the mitochondrial toxicity data in Hemmerich et al. originates from a Tox21 assay assessing mitochondrial membrane depolarization in HepG2 cells (which provides a distinct perspective compared to *in vitro* data derived from primary hepatocytes) thereby introducing additional biological information. When BSEP function is inhibited, bile salts accumulate within liver cells, causing hepatocyte injury and a risk of liver failure.[47] Metabolic processes can form reactive metabolites that bind covalently to hepatic proteins, altering their function and leading to damage in liver tissues.[48] Overall, in this study, we hypothesized that these

proxy-DILI labels along with chemical structure and physicochemical parameters would lead to enhanced predictivity in identifying potential liver injury endpoints while differentiating between sensitivities observed in human and animal proxy-DILI labels, allowing for interpretations of hepatotoxicity data across species. Finally, by including *in vitro* proxy-DILI labels, the models developed in this study have the potential for mechanistic evaluation and facilitating a comprehensive understanding of the underlying biochemical and cellular processes associated with drug-induced liver injuries.

MATERIALS AND METHODS

The workflow followed in this study is shown in Figure 1 and described in more detail in the following.

Drug-Induced Liver Toxicity Datasets: DILIst and DILIrank

The human *in vivo* dataset for liver toxicity was collected by combining DILIst[26] (714 toxic and 440 non-toxic compounds) and DILIrank[27] (268 toxic and 76 nontoxic compounds from Chavan et al[32]) datasets. The DILIst dataset classifies compounds into two classes based on their potential for causing DILI. The DILIrank dataset was released by the FDA prior to DILIst. This dataset analysed the hepatotoxic descriptions from FDA-approved drugs and assessed causality evidence from literature and classified compounds into four groups: ^{v}Most-, ^{v}Less, ^{v}No-DILI concern and Ambiguous-DILI-concern drugs. For the DILIrank dataset, we retrieved data from Chavan et al[32]. We treated ^{v}Most- and ^{v}Less as DILI Positive and those labelled with ^{v}No-DILI-concern as DILI Negative. Ambiguous-DILI-concern drugs were removed. Together these datasets form the largest drug list with DILI classification to date.

Proxy-DILI datasets: *in vivo* and *in vitro* assays

The first dataset we considered was the Liver Toxicity Knowledge Base (LTKB) benchmark dataset which was one of the earliest datasets developed from FDA-approved drug labels by NCTR and comprised drugs with the potential to cause DILI.[49] We determined one proxy-DILI label from studies on human hepatotoxicity[38], and two proxy-DILI labels from animal hepatotoxicity studies (animal hepatotoxicity A and B, and preclinical hepatotoxicity as detailed in Table 1)[38,39,40]. Animal hepatotoxicity datasets mentioned above consisted of data compiled by the authors from ToxRefDB[50], ORAD[40] and HESS[51] as well as hepatic histopathologic effects. Three diverse DILI datasets contain heterogeneous data

collected by other studies[31,41,42] (Divese DILI A, B and C as detailed in Table 1). These datasets consisted of data from the US drug-induced liver injury network, acute liver failure, withdrawn or suspended in the U.S. or compounds from the European markets, PubMed and FDA's MedWatch as well as compounds from the U.S. FDA Orange Book and Micromedex. We included three *in vitro* assays related to proposed or known mechanisms of liver injury, namely, mitochondrial toxicity[43], bile salt export pump inhibition (BSEP)[44] and the formation of reactive metabolites[45] (as detailed in Table 1). The labels for these *in vitro* datasets were the assay hit calls defined by the original studies. Previous studies indicated that mitochondrial toxicity and BSEP are reasonable predictors for cholestatic and mitochondrial toxins, however, they fail when applied to a wider chemical space for drugs with different mechanisms.[52] Many assay hits screened from chemical libraries often have unfavourable drug metabolism and pharmacokinetics presenting development challenges.[53] Thus, we considered pharmacokinetics as one of the proxy-DILI labels and compiled pharmacokinetic parameters of maximum concentration (Cmax) from Smith et al.[54] This dataset contains maximum unbound concentration in plasma for 534 compounds and maximum total concentration in plasma for 749 compounds. Together, as shown in Table 1, we obtained eleven *in vivo* and *in vitro* assays (proxy-DILI labels) related to liver injury and two pharmacokinetic parameters.

Dataset pre-processing and Compound Standardisation

Compound SMILES were standardized using the MolVS standardizer[56] (implemented using RDKit[55] v.2022.09.5) which included tautomer standardization and canonicalization to the parent molecule. Standardization settings involved sanitization, normalization, largest fragment chooser, charge fixed and reionised such that the strongest acids ionize first tautomer enumeration, and canonicalization as described in the MolVS standardizer guide.[56]

Among the 1,317 compounds, in cases where DILIrank and DILIst contained the same compound SMILES with conflicting toxicity labels, we retained compounds in DILIst (which is an updated version of DILIrank). Finally, we obtained a dataset of 1,275 unique compounds and associated DILI labels (820 toxic and 455 non-toxic compounds). This dataset is henceforth referred to as the gold standard DILI dataset.

For the proxy-DILI dataset, in case of any compounds with conflicting toxicity labels within a particular dataset after SMILES standardisation, we retained the compound as toxic/active (hence preferring the evidence of toxicity/activity which is a usual practice in drug discovery) resulting in a dataset of 18,679 compounds. For each of the eleven labels (as detailed in Table 1), if a compound was already present in the gold standard DILI dataset above, we removed the compound from the proxy-DILI dataset. This was done to avoid any information leaks in the models developed in this study. Finally, we obtained a dataset of 15,080 compounds in total for eleven proxy-DILI labels which are henceforth called the proxy-DILI dataset in this study.

Assay Concordance with Experimental Values

To evaluate the concordance of the eleven proxy-DILI labels and the gold standard DILI dataset with each other, we used all 18,679 compounds in the proxy-DILI dataset and compared them to the 1,275 compounds in the gold standard DILI dataset. To evaluate concordance, we used Cohen's kappa (as defined in scikit-learn v1.1.1[57]) to measure the level of agreement between activity values for each pair of labels which were present in the dataset.

Exploring the Physicochemical Space

Physicochemical space was explored using six characteristic physicochemical descriptors of molecular weight, TPSA, number of rotatable bonds, number of H donors, number of H acceptors and log P, (as implemented in RDKit[55] v.2022.09.5). We used a t-distributed stochastic neighbour embedding (t-SNE from scikit-learn v1.1.1[57]) to obtain a map of the physicochemical space for all compounds in the gold standard DILI dataset and proxy-DILI dataset with a high explained variance (PCA: 86.53% using two components).

Structural fingerprints, Mordred, and Physicochemical Descriptors

We used Morgan Fingerprints[58] of radius 2 and 2048 bits and 166-bit MACCS Keys[59], as implemented in RDKit[55] (v2022.09.5), as structural features for all compounds in the DILI dataset and proxy-DILI dataset. This resulted in 2,214-bit vector structural fingerprints.

We used molecular descriptors (as implemented in the Mordred[60] python package) and physicochemical properties (such as topological polar surface area TPSA, partition coefficient log P etc. as implemented in RDKit[55] v2022.09.5) for all compounds in the DILI dataset and proxy-DILI dataset. We dropped descriptors with missing values which resulted in 610 molecular descriptors for each compound.

Feature Selection

We first used feature selection on the compounds in the proxy-DILI dataset using a variance threshold (as implemented in scikit-learn v1.1.1[57]) to filter features (Figure 1 Step 1). We used a low variance threshold of 0.05 for Morgan fingerprints resulting in 93 selected bits, a threshold of 0.10 for MACCS keys resulting in 100 selected keys, and a threshold of 0.10 for Mordred descriptors resulting in 346 selected descriptors. Lower thresholds for variance ensured strict selection criteria, leading to fewer selected features to strike a balance between the length of all fingerprints and physicochemical parameters. An additional fifteen

calculated physicochemical parameters (as implemented in RDKit[55] v2022.09.5: topological polar surface area, hydrogen bond acceptors and donors, fraction of sp3 carbons, log P, and the number of rotatable bonds, rings, assembled rings, aromatic rings, hetero atoms, stereocenters, positive and negatively charged atoms, and the counts of NHOH and NO) were also added. This resulted in 193 bit-vector structural fingerprints and 361 molecular descriptors for each compound in the proxy-DILI dataset. The same selected features were used for the gold standard DILI dataset (Figure 1 Step 4) to avoid any information leaks.

Evaluation of predictions from individual proxy-DILI models

First, we trained individual models for each of the eleven proxy-DILI endpoints for all of the other proxy -DILI endpoints. For each proxy-DILI endpoint, we trained individual Random Forest models (Figure 1 Step 2) with a 5-fold stratified cross-validation and random halving search hyperparameter optimisation (as implemented in scikit-learn v1.1.1[57] with hyperparameter space given in Supplementary Table 1). We used this hyperparameter-optimised model to obtain predicted probabilities for all compounds for the other proxy-DILI endpoints for every 11×11 combination. For each model built on a proxy-DILI endpoint, we chose an optimal decision threshold based on the J-statistic value (see released code for implementation) by comparing the predicted probabilities to the true values. We obtained final binary predictions using this threshold thereby choosing the best-case scenario where the balanced accuracy is optimised from the AUC-ROC curve. Next, we compared how well each proxy-DILI model was at predicting other proxy-DILI labels by comparing the F1 Score and Likelihood Ratios.

Evaluating predictivity of individual proxy-DILI models for the gold standard DILI dataset

To train and evaluate models for DILI, we first split our gold-standard DILI dataset (containing 1,275 unique compounds) using a stratified split (Figure 1 Step 3) into training DILI data (of 1,020 unique compounds) and a held-out DILI test set (of 255 unique compounds). We evaluated the performance of individual models built on each of the eleven proxy-DILI endpoints on the held-out DILI test set (255 compounds). First, for each of the eleven individual models, we obtained out-of-fold predicted probabilities on the DILI training data (1,020 compounds) using cross-validation with a 5-fold stratified split. We used these out-of-fold predicted probabilities and true values to obtain an optimal decision threshold based on the J-statistic value. Finally, we used each of the individual models and the corresponding optimal decision threshold to obtain predictions of the held-out DILI test set. We used the Jaccard similarity coefficient score (as implemented in scikit-learn v1.1.1[57]) to compare the similarity of predictions, that is, the predicted DILI vectors from each model. The Jaccard similarity coefficient measures the similarity between two sets of data counting mutual presence (positives/toxic) as matches but not the absences.

Models for prediction of Cmax

Next, we trained two Random Forest regressor models to predict the median pMolar unbound plasma concentration and median pMolar total plasma concentration for 534 and 749 compounds respectively (Figure 1 Step 2). We used the selected 193 bit-vector structural fingerprints and 361 molecular descriptors as features to train the models with a 5-fold stratified cross-validation and random halving search hyperparameter optimisation as described above. The best estimator was refit on the entire dataset and the final model was used to generate predictions for compounds and these predicted features were used for training DILI models.

Models for prediction of DILI

In this study, we built models (Figure 1 Step 5) using (a) selected 193-bit structural fingerprints, (b) selected 361 molecular descriptors, (c) a combination of selected 193-bit structural fingerprints and selected 361 molecular descriptors, (d) predicted eleven proxy-DILI labels and two predicted pharmacokinetic parameters , and (e) a combination of all three features spaces.

For each feature space, we used repeated nested cross-validation. First, the DILI training data was split into 5-folds. One of these folds was used as a validation set while the data from the remaining 4 folds were used to train and hyperparameter optimise a Random Forest Classifier (as implemented in scikit-learn v1.1.1[57]). We optimised the classifier model using a random halving search (as implemented in scikit-learn v1.1.1[57]) and 4-fold cross-validation (see Supplementary Table S1 for hyperparameter space). Once hyperparameters were optimised, we then used the fitted model to generate 4-fold cross-validated estimates for each compound in the fitted data. These predicted probabilities along with the real data were used to generate an optimal threshold using the J statistic value (see released code for implementation). Finally, we predicted the DILI endpoint for the validation set and used the optimal threshold to determine the DILI toxicity. The process was repeated 5 times in total until all 1020 compounds in the DILI training data were used as a validation set. This entire nested-cross validation set-up was repeated eleven times with different splits. The model with the highest AUC was fit on the entire DILI training data and we obtained the optimal threshold using the J statistic value on the 4-fold cross-validated estimates for each of these compounds. Finally, this threshold was used to evaluate our models (Figure 1 Step 6) on the held-out DILI test set (255 unique compounds). Thus, for each model using a feature space (or the combination), we obtained evaluation metrics on (a) the nested cross-validation (on training data), and (b) the held-out test set. The best-performing model (Figure 1 Step 7), as shown in the Results section, was the combination of all three feature spaces. This model was

retrained (Figure 1 Step 8) on the complete gold-standard DILI dataset consisting of 1,275 distinct compounds. This model, DILIPredictor, can be accessed through a web application https://broad.io/DILIPredictor and have all code available for local use on GitHub at https://github.com/srijitseal/DILI.

To calculate the structural similarity of the held-out test to training data, we first calculated pairwise Tanimoto similarity (using 2048-bit Morgan fingerprint, see released code for implementation) for each test compound to each training compound. Finally, we calculated the mean of the three highest Tanimoto similarities (that is the three nearest neighbours) which was used to define the structural similarity of the particular test compound.

Evaluation Metrics

All predictions (nested-cross validation and held-out test set) were evaluated using sensitivity, specificity, balanced accuracy (BA), Mathew's correlation constant (MCC), F1 scores, positive predictive value (PPV), likelihood ratio (LR+)[61], average precision score (AP), Area Under Curve-Receiver Operating Characteristic (AUC-ROC) as implemented in scikit-learn v1.1.1[57].

Feature importance measures to understand the chemistry and biological mechanisms for common DILI compounds

For the final model released publicly that used a combination of all feature spaces, we used SHAP values (as implemented in the shap python package[62]) to obtain feature importance for each input compound. This included proxy-DILI data, pharmacokinetic parameters, physicochemical features as well as MACCS key substructures that contributed to DILI toxicity/safety. Further, we show how DILIPredictor can be used to eluate the causes of DILI, both in chemistry and via mechanisms on the biological level using the importance measures

on proxy-DILI labels. For this purpose, we chose a total of 16 compounds that were not present in the training data of these models. As shown in Table 4, several of toxic compounds were related to the study by Chang et al who compiled compounds causing DILI in patients undergoing chemotherapy.[63] We also included two pairs of compounds studied by Chen et al such as doxycycline/minocycline and moxifloxacin/trovafloxacin; these pairs were defined by a similar chemical structure and mechanism of action but differed in their liver toxicity effects.[64] Overall, fourteen of these compounds are known to cause DILI in humans while two compounds did not cause DILI in humans, namely, 2-butoxyethanol and astaxanthin.

Statistics and Reproducibility

We have released the datasets used in this proof-of-concept study which are publicly available at https://broad.io/DILIPredictor. We released the Python code for the models which are publicly available on GitHub at https://github.com/srijitseal/DILI.

RESULTS AND DISCUSSION

In this work, we trained models on each of eleven proxy-DILI endpoints related to liver toxicity. We used these models to obtain predicted proxy-DILI labels for 1,275 compounds in the gold standard DILI dataset (as defined in Methods) none of which overlapped with the proxy-DILI dataset. We then trained new models using those predicted proxy-DILI labels as inputs, together with the compounds' structural fingerprints, physicochemical properties, and a combination thereof, for 1,020 compounds the gold standard DILI datasets. We then evaluated the models on a held-out test set of 255 compounds.

Comparing chemical spaces for the proxy-DILI and gold standard DILI datasets

We first examined the diversity and representation of compounds in the proxy-DILI and gold standard DILI datasets, to ensure the evaluation would be reasonable. The distribution of compounds in each of the eleven labels of the proxy-DILI dataset covers a diverse range of physicochemical parameters as shown in Supplementary Figure S1. Gold standard DILI compounds effectively capture the diversity and representativeness of the compounds in the proxy-DILI dataset as shown in Supplementary Figure S2 for the physicochemical space of the 1,275 compounds in the gold standard DILI dataset compared to 15,080 compounds in the proxy-DILI dataset. Further, the held-out DILI test set (255 compounds) was also representative of the training DILI data (1,020 compounds) as shown in Supplementary Figure S3. The main caveat to consider is that the six characteristic physicochemical descriptors capture the variability of physicochemical space only to a certain extent. Overall, we conclude that the chemical space covered by the datasets is sufficiently similar for our evaluation to be reliable.

Concordance of Proxy-DILI dataset and DILI compounds

Next, we aimed to evaluate the concordance of labels in the proxy-DILI dataset with the gold standard DILI dataset. To do so, we compared all 18,679 compounds in the proxy-DILI dataset to the 1,275 compounds in the gold standard DILI dataset. It is important to note that these compounds (that overlapped between the proxy-DILI and gold standard DILI dataset) were only used to analyse concordance in this section and not in training the models, because that would leak information. As depicted in Figure 2, we observed a strong concordance between the data sourced from Liver Toxicity Knowledge Base and the human hepatotoxicity dataset (Cohen's Kappa = 0.50), preclinical data (0.54), and the three diverse DILI datasets (0.70, 0.80 and 0.82) used in this study. The lack of perfect concordance is reasonable given these datasets are primarily derived from human-related data, as opposed to animal data or *in vitro* assays. Note, concordance between DILI and proxy-DILI labels may be affected as the proxy-DILI dataset used here includes some of the DILI compounds (these overlapped compounds were removed later when training models).

Individual proxy-DILI models are complementary to each other and distinct in their prediction for DILI compounds

We next used the individual models built on the eleven proxy-DILI labels to predict the other proxy-DILI labels (with evaluation metrics as shown in Supplementary Table S2). As shown in Figure 3, we observed the Liver Toxicity Knowledge Base (the label most similar to the gold standard DILI label) was well predicted using human hepatotoxicity (LR+ = 4.56, F1=0.87) and preclinical hepatotoxicity (LR+ = 5.92, F1=0.89), as well as by the diverse DILI datasets (mean LR+ = 5.52, mean F1=0.85). Bile salt export pump inhibition (BESP) and mitochondrial toxicity were strongly predictive of each other (LR+ = 2.85, F1= 0.38 when using BSEP to predict mitotox and LR+ = 3.56, F1=0.73 when using mitotox to predict BSEP). Overall, the assays in the proxy-DILI dataset can be used to train individual models

to generate predicted proxy-DILI labels which then provide a complementary source of information.

We next analysed the eleven individual proxy-DILI models and a model built on the two PK parameters (Cmax unbound and total) for their predictions on the 255 compounds in the held-out compounds of the gold standard DILI dataset. As shown in Figure 4 (further details in Supplementary Table S3), the best-performing models were the model built on the preclinical animal hepatotoxicity (AUC=0.67, LR+ = 2.04), the model built on the LTKB dataset (AUC=0.67, LR+=1.88), and the model built on PK parameters (AUC=0.68, LR+ = 1.30), and. Although the LTKB dataset contained only 103 compounds, this dataset was used to inform the labels selected by the FDA in its gold standard DILI dataset, which explains its high performance. However, other biological labels have compounds covering a wider biological and chemical space coverage which warrants their inclusion in our study as shown by Jaccard similarity for predictions on the held-out DILI dataset. Predictions from models built on animal hepatotoxicity labels were not similar to predictions from models built on human hepatotoxicity labels (Figure 5; mean Jaccard similarity of 0.03). We found that predictions from models built on human-related labels were similar (e.g., predictions from the preclinical hepatotoxicity model and LTKB models have a Jaccard similarity of 0.69). However, predictions from human-related labels were dissimilar to predictions from *in vitro* assays (e.g., predictions from the preclinical hepatotoxicity model had only a 0.02 Jaccard similarity to predictions from the mitotox model and 0.03 Jaccard similarity to predictions from the reactive metabolite formation model). Overall, we conclude that each model built on a proxy-DILI label and the PK parameters was distinctive in its prediction, thus providing complementary information on compounds' potential for DILI.

Models combining chemical structure, physicochemical properties, PK parameters and predicted proxy-DILI data outperform individual models

We next compared models built on combinations of proxy-DILI labels (including PK parameters), chemical structure, and physicochemical properties including Mordred descriptors (Table 2). When comparing results from 55 held-out test sets from the repeated nested cross-validation (as shown in Figure 6 with the comparison of differences in distribution using a paired t-test.), the models combining structural fingerprints, physicochemical properties, Mordred descriptors, PK parameters and predicted proxy-DILI labels achieved a mean AUC = 0.71 (mean LR+ = 2.02) compared to the second-best models using only physicochemical properties and Mordred descriptors with a mean AUC=0.70 (mean LR+ = 1.93) and models using structural fingerprints, physicochemical properties and Mordred descriptors which also achieved a mean AUC of 0.70 (mean LR+ = 1.93). Models using only structural fingerprints achieved a mean AUC of 0.69 (mean LR+ = 1.87) while models using only predicted proxy-DILI labels and PK parameters as features achieved a mean AUC of 0.69 (mean LR+ = 2.09) in the nested cross-validation. Supplementary Figure S4 compares the distribution of balanced accuracy for all model combinations using all feature sets (predicted proxy-DILI labels and PK parameters, structural features, and Mordred physicochemical descriptors).

We next retrained all hyperparameter-optimised models on the DILI training data (1020 compounds) and evaluated the final models on the held-out DILI test set (255 compounds). The DILIPredictor model (combining all predicted proxy-DILI labels and PK parameters, structural features, and Mordred physicochemical descriptors) achieved an AUC = 0.74 (LR+ = 2.50) (Table 2). This was quite successful compared to other models such as AUC = 0.74 (LR+ = 1.92) for the model using both structural, Mordred and physicochemical descriptors,

AUC = 0.73 (LR+ = 1.67) for structural models, AUC = 0.74 (LR+ = 1.96) for the model using only proxy-DILI and PK parameters, and AUC = 0.72 (LR+ = 2.29) for the model using Mordred and physicochemical descriptors.

One metric relevant in predictive safety/toxicology is the positive likelihood ratio[61] and the detection of toxic compounds with a lower false positive rate. We next analysed the performance of models along the AUC-ROC curve from a false positive rate of 0 to 1. When predicting the first 26 compounds correctly as true positives (or approximately 10% of the 255 compounds in the held-out test set), DILIPredictor achieved the highest LR+ score of 7.21 (26 toxic compounds correctly predicted out of 28 compounds, PPV = 0.93) compared to the structural model which achieved an LR+ score of 4.62 (25 toxic compounds correctly predicted out of 28 compounds, PPV = 0.89). This improvement is mainly from being able to detect compounds at a wider range of structural similarity to training data (as shown in Supplementary Figure S5 using the distribution of the top true positives detected early with low false positive rates for each model). Overall, this shows that using all feature types in DILIPredictor allows for the early detection of a greater number of toxic compounds with a low false positive rate.

We subsequently compared our models to those reported in earlier publications. Table 3 presents a selection of recent DILI prediction models that employ chemical features and biological data to predict liver toxicity. Since most previous studies did not emphasize early detection or likelihood ratios, it is not possible to compare LR+ scores; therefore, we can only make comparisons within the models developed in this study. It is important to note that the size, source, and consequently the quality of training and test datasets vary across previous literature, rendering direct comparisons infeasible. In our study, the final DILIPredictor

model achieved a high AUC-ROC of 0.74 on the held-out gold standard DILI dataset (255 compounds), which aligns with the average AUC (0.74) from prior studies.

Feature Interpretation

We next used feature interpretation to analyse the chemistry and biological mechanisms for compounds known to cause DILI. We chose sixteen compounds (Table 4) of which fourteen were known for their DILI[63] and two compounds that do not cause DILI in humans (namely, 2-butoxyethanol and astaxanthin). DILIPredictor could detect structural information relevant to causing DILI (six compounds shown in Figure 7 and further in Table 4 and Supplementary Figure S6). The MACCS features most contributing to the toxicity for gemcitabine (a nucleoside analogue drug) was related to the presence of a ribose ring, which is key to the attachment of the three phosphates and then works as a faulty base in DNA synthesis, causing cell death[65] (Figure 7). Further, DILIPredictor correctly predicted compounds such as 2-butoxyethanol and astaxanthin to be non-toxic in humans even though they cause hepatic injury in animal models[17,25] (Figure 7). In these compounds, proxy-DILI features associated with either animal hepatotoxicity or preclinical hepatotoxicity contributed to predicting toxicity in humans, however, the proxy-DILI indicators related to human hepatotoxicity ultimately led to the prediction of non-toxicity.

Finally, among structurally similar pairs of compounds, acitretin was correctly predicted as toxic while astaxanthin was correctly predicted to be non-toxic. For acitretin, the preclinical hepatotoxicity label contributed to the toxicity prediction. Conversely, labels associated with human hepatotoxicity contributed to correctly predicting astaxanthin as non-toxic. Among tetracyclines, pairs of compounds doxycycline (prediction scores = 0.72) and minocycline (0.73), and among fluoroquinolones, pairs of compounds moxifloxacin (0.84) and trovafloxacin (0.86) were correctly predicted toxic. Furthermore, the prediction scores

obtained from DILIPredictor were in agreement with the less-toxic or more-toxic DILI annotations collated by Chen et al.[64] Among compounds withdrawn from market, sitaxentan, trovafloxacin and ximelagatran were flagged with prediction scores above 0.80 threshold; however many compounds currently on the market such as docetaxel and paclitaxel were also flagged in the same threshold as being DILI-toxic. Overall, DILIPredictor combined chemical structures and biological data to correctly predict DILI in humans.

Limitations of DILIPredictor

The primary focus of this study was the generation of binary classification models for drug-induced liver injury. Besides using predicted Cmax (unbound ant total), we did not incorporate factors such as dose or time point into this study due to its scarcity in available public data. Labelling schemes are not always binary but sometimes include an "ambiguous" class (such as in the DILIrank dataset) and these compounds are hence not included in this study. While *in vitro* data can provide valuable insights into drug toxicities, they are still proxy endpoints for the *in vivo* effects. Toxic compounds detected in *in vitro* assays can often cause corresponding toxicity *in vivo*, but compounds that appear to be safe in *in vitro* are not necessarily safe in humans.[23,24]

CONCLUSIONS

In this work, we trained models to predict drug-induced liver injury (DILI) using not only chemical data but also heterogeneous biological *in vivo* (human and animal) and *in vitro* data from various sources. We found a strong concordance in observed data between compounds with the proxy-DILI labels and DILI compounds. The eleven proxy-DILI models were not predictive of each other- this complementarity suggests that they could be combined to predict drug-induced liver injury. Random Forest models that combined different types of input data - structural fingerprints, physicochemical properties, PK properties and proxy-DILI labels - improved predictive performance, especially in early detection (with low false positive rates), with the highest LR+ score of 7.21 (26 toxic compounds with PPV=0.93). DILIPredictor accurately predicted the toxicity of various compounds known to cause DILI, including fourteen notorious DILI-inducing compounds, by recognizing chemical structure as well as biological mechanisms. DILIPredictor was further able to differentiate between animal and human sensitivity for DILI and exhibited a potential for mechanism evaluation for these compounds. Overall, the study demonstrated that incorporating all complementary sources of information can significantly improve the accuracy of DILI prediction models. Furthermore, the availability of larger, high-quality, and standardized datasets for DILI in the public domain can greatly enhance the development of predictive models for drug-induced liver injury such as from the Omics for Assessing Signatures for Integrated Safety Consortium (OASIS).[66] We released our final interpretable models at (with all code available for download at GitHub at https://github.com/srijitseal/DILI) and datasets used in this study at https://broad.io/DILIPredictor.

ACKNOWLEDGEMENTS

S.S. acknowledges funding from the Cambridge Commonwealth, European and International Trust, Boak Student Support Fund (Clare Hall), Jawaharlal Nehru Memorial Fund, Allen, Meek and Read Fund, and Trinity Henry Barlow (Trinity College). S.S. would like to thank Dr Anne E. Carpenter (Broad Institute of MIT and Harvard) for her valuable feedback on this work. O.S. acknowledges funding from the Swedish Research Council (grants 2020-03731 and 2020-01865), FORMAS (grant 2022-00940), Swedish Cancer Foundation (22 2412 Pj), and Horizon Europe grant agreement #101057014 (PARC) and #101057442 (REMEDI4ALL). This work was performed using resources provided by the Cambridge Service for Data-Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

ASSOCIATED CONTENT

Supplemental Information. The Supporting Information is available. Supporting Information (PDF). We released the Python code for our models which are publicly available at https://github.com/srijitseal/DILI

AUTHOR INFORMATION

Corresponding Authors

Prof Andreas Bender

ab454@cam.ac.uk

Prof Ola Spjuth

ola.spjuth@uu.se

All authors have approved the final version of the manuscript.

Contributions

SS designed and performed data analysis, implemented, and trained the models. SS analysed the interpretation of features and the results of models. SS wrote the manuscript with extensive discussions with OS and AB and who supervised the project. All the authors (SS, DPW, LHG, OS, and AB) reviewed, edited, contributed to discussions, and approved the final version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

(1) Andrade, R. J.; Chalasani, N.; Björnsson, E. S.; Suzuki, A.; Kullak-Ublick, G. A.; Watkins, P. B.; Devarbhavi, H.; Merz, M.; Lucena, M. I.; Kaplowitz, N.; Aithal, G. P. Drug-Induced Liver Injury. Nat. Rev. Dis. Prim. 2019, 5 (1), 1–22. https://doi.org/10.1038/s41572-019-0105-0.

(2) Remmer, H. The Role of the Liver in Drug Metabolism. Am. J. Med. 1970, 49 (5), 617–629. https://doi.org/10.1016/S0002-9343(70)80129-2.

(3) Licata, A. Adverse Drug Reactions and Organ Damage: The Liver. Eur. J. Intern. Med. 2016, 28, 9–16. https://doi.org/10.1016/j.ejim.2015.12.017.

(4) Ostapowicz, G.; Fontana, R. J.; Schiødt, F. V.; Larson, A.; Davern, T. J.; Han, S. H. B.; McCashland, T. M.; Shakil, A. O.; Hay, J. E.; Hynan, L.; Crippin, J. S.; Blei, A. T.; Samuel, G.; Reisch, J.; Lee, W. M.; Santyanarayana, R.; Caldwell, C.; Shick, L.; Bass, N.; Rouillard, S.; Atillasoy, E.; Flamm, S.; Benner, K. G.; Rosen, H. R.; Martin, P.; Stribling, R.; Schiff, E. R.; Torres, M. B.; Navarro, V.; McGuire, B.; Chung, R.; Abraczinskas, D.; Dienstag, J. Results of a Prospective Study of Acute Liver Failure at 17 Tertiary Care Centers in the United States. Ann. Intern. Med. 2002, 137 (12), 947–954. https://doi.org/10.7326/0003-4819-137-12-200212170-00007.

(5) Onakpoya, I. J.; Heneghan, C. J.; Aronson, J. K. Post-Marketing Withdrawal of 462 Medicinal Products Because of Adverse Drug Reactions: A Systematic Review of the World Literature. BMC Med. 2016, 14 (1), 1–11. https://doi.org/10.1186/s12916-016-0553-2.

(6) Raschi, E.; De Ponti, F. Strategies for Early Prediction and Timely Recognition of Drug-Induced Liver Injury: The Case of Cyclin-Dependent Kinase 4/6 Inhibitors. Front. Pharmacol. 2019, 10 (OCT), 1235. https://doi.org/10.3389/fphar.2019.01235.

(7) Weaver, R. J.; Blomme, E. A.; Chadwick, A. E.; Copple, I. M.; Gerets, H. H. J.; Goldring, C. E.; Guillouzo, A.; Hewitt, P. G.; Ingelman-Sundberg, M.; Jensen, K. G.; Juhila, S.; Klingmüller, U.; Labbe, G.; Liguori, M. J.; Lovatt, C. A.; Morgan, P.; Naisbitt, D. J.; Pieters, R. H. H.; Snoeys, J.; van de Water, B.; Williams, D. P.; Park, B. K. Managing the Challenge of Drug-Induced Liver Injury: A Roadmap for the Development and Deployment of Preclinical Predictive Models. Nat. Rev. Drug Discov. 2020, 19 (2), 131–148. https://doi.org/10.1038/s41573-019-0048-x.

(8) Mihajlovic, M.; Vinken, M. Mitochondria as the Target of Hepatotoxicity and Drug-Induced Liver Injury: Molecular Mechanisms and Detection Methods. Int. J. Mol. Sci. 2022, 23 (6). https://doi.org/10.3390/ijms23063315.

(9) Kenna, J. G.; Taskar, K. S.; Battista, C.; Bourdet, D. L.; Brouwer, K. L. R.; Brouwer, K. R.; Dai, D.; Funk, C.; Hafey, M. J.; Lai, Y.; Maher, J.; Pak, Y. A.; Pedersen, J. M.; Polli, J. W.; Rodrigues, A. D.; Watkins, P. B.; Yang, K.; Yucha, R. W. Can Bile Salt Export Pump Inhibition Testing in Drug Discovery and Development Reduce Liver Injury Risk? An International Transporter Consortium Perspective. Clin. Pharmacol. Ther. 2018, 104 (5), 916–932. https://doi.org/10.1002/cpt.1222.

(10) Villanueva-Paz, M.; Morán, L.; López-Alcántara, N.; Freixo, C.; Andrade, R. J.; Lucena, M. I.; Cubero, F. J. Oxidative Stress in Drug-Induced Liver Injury (DILI): From Mechanisms

to Biomarkers for Use in Clinical Practice. Antioxidants 2021, 10 (3), 1–35. https://doi.org/10.3390/antiox10030390.

(11) Bao, Y.; Wang, P.; Shao, X.; Zhu, J.; Xiao, J.; Shi, J.; Zhang, L.; Zhu, H. J.; Ma, X.; Manautou, J. E.; Zhong, X. B. Acetaminophen-Induced Liver Injury Alters Expression and Activities of Cytochrome P450 Enzymes in an Age-Dependent Manner in Mouse Liver. Drug Metab. Dispos. 2020, 48 (5), 326–336. https://doi.org/10.1124/DMD.119.089557.

(12) Soldatow, V. Y.; Lecluyse, E. L.; Griffith, L. G.; Rusyn, I. *In vitro* Models for Liver Toxicity Testing. Toxicol. Res. (Camb). 2013, 2 (1), 23–39. https://doi.org/10.1039/c2tx20051a.

(13) Meng, Q. Three-Dimensional Culture of Hepatocytes for Prediction of Drug-Induced Hepatotoxicity. Expert Opin. Drug Metab. Toxicol. 2010, 6 (6), 733–746. https://doi.org/10.1517/17425251003674356.

(14) Sison-Young, R. L. C.; Mitsa, D.; Jenkins, R. E.; Mottram, D.; Alexandre, E.; Richert, L.; Aerts, H.; Weaver, R. J.; Jones, R. P.; Johann, E.; Hewitt, P. G.; Ingelman-Sundberg, M.; Goldring, C. E. P.; Kitteringham, N. R.; Park, B. K. Comparative Proteomic Characterization of 4 Human Liver-Derived Single Cell Culture Models Reveals Significant Variation in the Capacity for Drug Disposition, Bioactivation, and Detoxication. Toxicol. Sci. 2015, 147 (2), 412–424. https://doi.org/10.1093/toxsci/kfv136.

(15) Steger-Hartmann, T.; Raschke, M. Translating in Vitro to in Vivo and Animal to Human. Curr. Opin. Toxicol. 2020, 23–24, 6–10. https://doi.org/10.1016/j.cotox.2020.02.003.

(16) Kindrat, I.; Dreval, K.; Shpyleva, S.; Tryndyak, V.; de Conti, A.; Mudalige, T. K.; Chen, T.; Erstenyuk, A. M.; Beland, F. A.; Pogribny, I. P. Effect of Methapyrilene Hydrochloride on Hepatic Intracellular Iron Metabolism *in vivo* and *in vitro*. Toxicol. Lett. 2017, 281, 65–73. https://doi.org/10.1016/j.toxlet.2017.09.011.

(17) Graham, E. E.; Walsh, R. J.; Hirst, C. M.; Maggs, J. L.; Martin, S.; Wild, M. J.; Wilson, I. D.; Harding, J. R.; Kenna, J. G.; Peter, R. M.; Williams, D. P.; Park, B. K. Identification of the Thiophene Ring of Methapyrilene as a Novel Bioactivation-Dependent Hepatic Toxicophore. J. Pharmacol. Exp. Ther. 2008, 326 (2), 657–671. https://doi.org/10.1124/jpet.107.135483.

(18) Hamadeh, H. K.; Knight, B. L.; Haugen, A. C.; Sieber, S.; Amin, R. P.; Bushel, P. R.; Stoll, R.; Blanchard, K.; Jayadev, S.; Tennant, R. W.; Cunningham, M. L.; Afshari, C. A.; Paules, R. S. Methapyrilene Toxicity: Anchorage of Pathologic Observations to Gene Expression Alterations. Toxicol. Pathol. 2002, 30 (4), 470–482. https://doi.org/10.1080/01926230290105712.

(19) Mirsalis, J. C. Genotoxicity, Toxicity, and Carcinogenicity of the Antihistamine Methapyrilene (MTR 07216). Mutat. Res. Genet. Toxicol. 1987, 185 (3), 309–317. https://doi.org/10.1016/0165-1110(87)90022-4.

(20) Wright, P. S. R.; Briggs, K. A.; Thomas, R.; Smith, G. F.; Maglennon, G.; Mikulskis, P.; Chapman, M.; Greene, N.; Phillips, B. U.; Bender, A. Statistical Analysis of Preclinical Inter-Species Concordance of Histopathological Findings in the ETOX Database. Regul. Toxicol. Pharmacol. 2023, 138, 105308. https://doi.org/10.1016/j.yrtph.2022.105308.

(21) Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; Dorato, M.; Van Deun, K.; Smith, P.; Berger, B.; Heller, A. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. Regul. Toxicol. Pharmacol. 2000, 32 (1), 56–67. https://doi.org/10.1006/rtph.2000.1399.

(22) Fourches, D.; Barnes, J. C.; Day, N. C.; Bradley, P.; Reed, J. Z.; Tropsha, A. Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species. Chem. Res. Toxicol. 2010, 23 (1), 171–183. https://doi.org/10.1021/tx900326k.

(23) Bender, A.; Cortés-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1: Ways to Make an Impact, and Why We Are Not There Yet. Drug Discovery Today. Elsevier Ltd February 1, 2021, pp 511–524. https://doi.org/10.1016/j.drudis.2020.12.009.

(24) Van Norman, G. A. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials: Is It Time to Rethink Our Current Approach? JACC Basic to Transl. Sci. 2019, 4 (7), 845–854. https://doi.org/10.1016/j.jacbts.2019.10.008.

(25) Cunningham, M. L. A Mouse Is Not a Rat Is Not a Human: Species Differences Exist. Toxicol. Sci. 2002, 70 (2), 157–158. https://doi.org/10.1093/toxsci/70.2.157.

(26) Thakkar, S.; Li, T.; Liu, Z.; Wu, L.; Roberts, R.; Tong, W. Drug-Induced Liver Injury Severity and Toxicity (DILIst): Binary Classification of 1279 Drugs by Human Hepatotoxicity. Drug Discov. Today 2020, 25 (1), 201–208. https://doi.org/10.1016/J.DRUDIS.2019.09.022.

(27) Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans. Drug Discov. Today 2016, 21 (4), 648–653. https://doi.org/10.1016/J.DRUDIS.2016.02.015.

(28) Hewitt, M.; Enoch, S. J.; Madden, J. C.; Przybylak, K. R.; Cronin, M. T. D. Hepatotoxicity: A Scheme for Generating Chemical Categories for Read-across, Structural Alerts and Insights into Mechanism(s) of Action. Crit. Rev. Toxicol. 2013, 43 (7), 537–558. https://doi.org/10.3109/10408444.2013.811215.

(29) Ye, L.; Ngan, D. K.; Xu, T.; Liu, Z.; Zhao, J.; Sakamuru, S.; Zhang, L.; Zhao, T.; Xia, M.; Simeonov, A.; Huang, R. Prediction of Drug-Induced Liver Injury and Cardiotoxicity Using Chemical Structure and *in vitro* Assay Data. Toxicol. Appl. Pharmacol. 2022, 454. https://doi.org/10.1016/J.TAAP.2022.116250.

(30) Liu, A.; Walter, M.; Wright, P.; Bartosik, A.; Dolciami, D.; Elbasir, A.; Yang, H.; Bender, A. Prediction and Mechanistic Analysis of Drug-Induced Liver Injury (DILI) Based on Chemical Structure. Biol. Direct 2021, 16 (1), 1–15. https://doi.org/10.1186/s13062-020-00285-0.

(31) Mora, J. R.; Marrero-Ponce, Y.; García-Jacas, C. R.; Suarez Causado, A. Ensemble Models Based on QuBiLS-MAS Features and Shallow Learning for the Prediction of Drug-Induced Liver Toxicity: Improving Deep Learning and Traditional Approaches. Chem. Res. Toxicol. 2020, 33 (7), 1855–1873. https://doi.org/10.1021/ACS.CHEMRESTOX.0C00030/ASSET/IMAGES/LARGE/TX0C00030_0007.JPEG.

(32) Liu, A.; Seal, S.; Yang, H.; Bender, A. Using Chemical and Biological Data to Predict Drug Toxicity. SLAS Discov. 2023. https://doi.org/10.1016/J.SLASD.2022.12.003.

(33) Seal, S.; Yang, H.; Trapotsi, M.-A.; Singh, S.; Carreras-Puigvert, J.; Spjuth, O.; Bender, A. Merging Bioactivity Predictions from Cell Morphology and Chemical Fingerprint Models Using Similarity to Training Data. J. Cheminform. 2023, 15 (1), 56. https://doi.org/10.1186/s13321-023-00723-x.

(34) Rao, M.; Nassiri, V.; Alhambra, C.; Snoeys, J.; Van Goethem, F.; Irrechukwu, O.; Aleo, M. D.; Geys, H.; Mitra, K.; Will, Y. AI/ML Models to Predict the Severity of Drug-Induced Liver Injury for Small Molecules. Chem. Res. Toxicol. 2023, 36, 1129–1139. https://doi.org/10.1021/acs.chemrestox.3c00098.

(35) Chavan, S.; Scherbak, N.; Engwall, M.; Repsilber, D. Predicting Chemical-Induced Liver Toxicity Using High-Content Imaging Phenotypes and Chemical Descriptors: A Random Forest Approach. Chem. Res. Toxicol. 2020, 33 (9), 2261–2275. https://doi.org/10.1021/ACS.CHEMRESTOX.9B00459/SUPPL_FILE/TX9B00459_SI_003.TXT.

(36) Seal, S.; Carreras-Puigvert, J.; Trapotsi, M. A.; Yang, H.; Spjuth, O.; Bender, A. Integrating Cell Morphology with Gene Expression and Chemical Structure to Aid Mitochondrial Toxicity Detection. Commun. Biol. 2022, 5 (1), 858. https://doi.org/10.1038/s42003-022-03763-5.

(37) Seal, S.; Spjuth, O.; Hosseini-Gerami, L.; Garcia-Ortegon, M.; Singh, S.; Bender, A.; Carpenter, A. E. Insights into Drug Cardiotoxicity from Biological and Chemical Data: The

First Public Classifiers for FDA DICTrank. bioRxiv 2023, 2023.10.15.562398. https://doi.org/10.1101/2023.10.15.562398.

(38) Mulliner, D.; Schmidt, F.; Stolte, M.; Spirkl, H. P.; Czich, A.; Amberg, A. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope. Chem. Res. Toxicol. 2016, 29 (5), 757–767. https://doi.org/10.1021/ACS.CHEMRESTOX.5B00465

(39) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast *in vitro* Bioactivity and Chemical Structure. Chem. Res. Toxicol. 2015, 28 (4), 738–751. https://doi.org/10.1021/TX500501H.

(40) Ambe, K.; Ishihara, K.; Ochibe, T.; Ohya, K.; Tamura, S.; Inoue, K.; Yoshida, M.; Tohkin, M. In Silico Prediction of Chemical-Induced Hepatocellular Hypertrophy Using Molecular Descriptors. Toxicol. Sci. 2018, 162 (2), 667–675. https://doi.org/10.1093/TOXSCI/KFX287.

(41) Zhu, X. W.; Sedykh, A.; Liu, S. S. Hybrid in Silico Models for Drug-Induced Liver Injury Using Chemical Descriptors and *in vitro* Cell-Imaging Information. J. Appl. Toxicol. 2014, 34 (3), 281–288. https://doi.org/10.1002/jat.2879.

(42) He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An In Silico Model for Predicting Drug-Induced Hepatotoxicity. Int. J. Mol. Sci. 2019, 20 (8). https://doi.org/10.3390/IJMS20081897.

(43) Hemmerich, J.; Troger, F.; Füzi, B.; F.Ecker, G. Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. Mol. Inform. 2020, 39 (5). https://doi.org/10.1002/minf.202000005.

(44) McLoughlin, K. S.; Jeong, C. G.; Sweitzer, T. D.; Minnich, A. J.; Tse, M. J.; Bennion, B. J.; Allen, J. E.; Calad-Thomson, S.; Rush, T. S.; Brase, J. M. Machine Learning Models to Predict Inhibition of the Bile Salt Export Pump. J. Chem. Inf. Model. 2021, 61 (2), 587–602. https://doi.org/10.1021/ACS.JCIM.0C00950.

(45) Mazzolari, A.; Vistoli, G.; Testa, B.; Pedretti, A. Prediction of the Formation of Reactive Metabolites by A Novel Classifier Approach Based on Enrichment Factor Optimization (EFO) as Implemented in the VEGA Program. Molecules 2018, 23 (11). https://doi.org/10.3390/MOLECULES23112955.

(46) Zhang, C.; Zhao, Y.; Yu, M.; Qin, J.; Ye, B.; Wang, Q. Mitochondrial Dysfunction and Chronic Liver Disease. Curr. Issues Mol. Biol. 2022, 44 (7), 3156–3165. https://doi.org/10.3390/cimb44070218.

(47) Kenna, J. G.; Taskar, K. S.; Battista, C.; Bourdet, D. L.; Brouwer, K. L. R.; Brouwer, K. R.; Dai, D.; Funk, C.; Hafey, M. J.; Lai, Y.; Maher, J.; Pak, Y. A.; Pedersen, J. M.; Polli, J. W.; Rodrigues, A. D.; Watkins, P. B.; Yang, K.; Yucha, R. W. Can Bile Salt Export Pump Inhibition Testing in Drug Discovery and Development Reduce Liver Injury Risk? An International Transporter Consortium Perspective. Clin. Pharmacol. Ther. 2018, 104 (5), 916–932. https://doi.org/10.1002/cpt.1222.

(48) Attia, S. M. Deleterious Effects of Reactive Metabolites. Oxid. Med. Cell. Longev. 2010, 3 (4), 238–253. https://doi.org/10.4161/oxim.3.4.13246.

(49) Thakkar, S.; chen, M.; Fang, H.; Liu, Z.; Roberts, R.; Tong, W. The Liver Toxicity Knowledge Base (LKTB) and Drug-Induced Liver Injury (DILI) Classification for Assessment of Human Liver Injury. Expert Rev. Gastroenterol. Hepatol. 2018, 12 (1), 31–38. https://doi.org/10.1080/17474124.2018.1383154.

(50) Martin, M. T. and R. Judson. U.S. Environmental Protection Agency, Washington, DC, 2010. Release User-Friendly Web-Based Tool for Mining ToxRefDB.

(51) Sakuratani, Y.; Zhang, H. Q.; Nishikawa, S.; Yamazaki, K.; Yamada, T.; Yamada, J.; Gerova, K.; Chankov, G.; Mekenyan, O.; Hayashi, M. Hazard Evaluation Support System (HESS) for Predicting Repeated Dose Toxicity Using Toxicological Categories. SAR QSAR Environ. Res. 2013, 24 (5), 351–363. https://doi.org/10.1080/1062936X.2013.773375.

(52) Williams, D. P.; Lazic, S. E.; Foster, A. J.; Semenova, E.; Morgan, P. Predicting Drug-Induced Liver Injury with Bayesian Machine Learning. Chem. Res. Toxicol. 2020, 33 (1), 239–248. https://doi.org/10.1021/acs.chemrestox.9b00264.

(53) Horne, R.I.; Wilson-Godber, J.; Díaz, A. G.; Brotzakis, Z. F.; Seal, S.; Gregory, R. C.; Possenti, A.; Chia, S.; Vendruscolo, M. Using Generative Modeling to Endow with Potency Initially Inert Compounds with Good Bioavailability and Low Toxicity. J. Chem. Inf. Model. 2024 (in press). https://doi.org/10.1021/acs.jcim.3c01777.

(54) Smit, I. A.; Afzal, A. M.; Allen, C. H. G.; Svensson, F.; Hanser, T.; Bender, A. Systematic Analysis of Protein Targets Associated with Adverse Events of Drugs from

Clinical Trials and Postmarketing Reports. Chem. Res. Toxicol. 2021, 34 (2), 365–384. https://doi.org/10.1021/ACS.CHEMRESTOX.0C00294/ASSET/IMAGES/LARGE/TX0C00294_0009.JPEG.

(55) RDKit (v2022.09.5) https://www.rdkit.org/ (accessed Apr 18, 2023).

(56) Swain, M. MolVS: Molecule Validation and Standardization https://molvs.readthedocs.io/en/latest/ (accessed Apr 15, 2021).

(57) scikit-learn: machine learning in Python — scikit-learn 1.2.0 documentation https://scikit-learn.org/stable/index.html (accessed Jan 3, 2023).

(58) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50 (5), 742–754. https://doi.org/10.1021/ci100050t.

(59) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. J. Chem. Inf. Comput. Sci. 2002, 42 (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(60) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. J. Cheminform. 2018, 10 (1), 1–14. https://doi.org/10.1186/s13321-018-0258-y.

(61) Fierz, W.; Bossuyt, X. Likelihood Ratio Approach and Clinical Interpretation of Laboratory Tests. Front. Immunol. 2021, 12, 1170. https://doi.org/10.3389/fimmu.2021.655262.

(62) SHAP. Welcome to the SHAP documentation — SHAP latest documentation https://shap.readthedocs.io/en/latest/index.html (accessed May 17, 2023).

(63) A.; Chang, P.; Devuni, D.; Bichoupan, K.; Kesar, V.; Branch, A. D.; Oh, W. K.; D. Galsky, M.; Ahmad, J.; A. Odin, J. Real World Experience of Drug Induced Liver Injury in Patients Undergoing Chemotherapy. J. Clin. Gastroenterol. Hepatol. 2018, 02 (03). https://doi.org/10.21767/2575-7733.1000047.

(64) Chen, M.; Borlak, J.; Tong, W. A Model to Predict Severity of Drug-Induced Liver Injury in Humans. Hepatology 2016, 64 (3), 931–940. https://doi.org/10.1002/hep.28678.

(65) Schäfer, A.; Schomacher, L.; Barreto, G.; Döderlein, G.; Niehrs, C. Gemcitabine Functions Epigenetically by Inhibiting Repair Mediated DNA Demethylation. PLoS One 2010, 5 (11). https://doi.org/10.1371/journal.pone.0014060.

(66) HESI Insights - February 2023 - HESI - Health and Environmental Sciences Institute https://hesiglobal.org/hesi-insights-february-2023/ (accessed May 29, 2023).

TABLES

Table 1. Sources of Liver-safety and Toxicity Data Used in this study.

Data Source	Assay Type	Cell line used	Used in this study	Total number of compounds	Number of compounds in Active Class	Description	Reference (data retrieved from)
Liver Toxicity Knowledge Base	DILI	N/A	Training Data (Liv)	103	77	Liver Toxicity Knowledge Base (LTKB), prescription drugs, hepatotoxicity,	Chen et al
Human hepatotoxicity	Human hepatotoxicity		Training Data (Liv)	1582	933	Human hepatotoxicity	Mulliner et al
Animal hepatotoxicity A	Animal hepatotoxicity		Training Data (Liv)	602	203	Chronic oral administration, Hepatic histopathologic effects, ToxRefDB	Liu et al
Animal hepatotoxicity B	Animal hepatotoxicity		Training Data (Liv)	738	412	Hepatocellular hypertrophy, rats, ORAD, HESS,	Ambe et al
Preclinical hepatotoxicity	Animal hepatotoxicity		Training Data (Liv)	2750	2112	Preclinical hepatotoxicity	Mulliner et al
Diverse DILI A	Heterogenous Data	N/A	Training Data (Liv)	1230	474	Large-scale and diverse DILI dataset,	He et al

Diverse DILI B	Heterogenous Data	N/A	Training Data (Liv)	143	71	US drug-induced liver injury network, acute liver failure, withdrawn or suspended in US or European markets, PubMed and FDA's MedWatch.	Zhu et al
Diverse DILI C	Heterogenous Data	N/A	Training Data (Liv)	556	291	Transient liver function abnormalities, adverse hepatic effects, U.S. FDA Orange Book, Micromedex	Mora et al
BESP	Mechanisms of Liver Toxicity		Training Data (Liv)	763	354	Bile Salt Export Pump Inhibition	McLoughlin et al
Mitotox	Mechanisms of Liver Toxicity		Training Data (Liv)	6111	874	Mitochondrial Toxicity	Hemmerich et al
Reactive Metabolite	Mechanisms of Liver Toxicity		Training Data (Liv)	502	118	Reactive Metabolite	Mazzolari et al
Cmax (total)	Pharmacokinetic Properties		Predicted Property	749	N/A	Maximum total concentration in plasma	Smith et al
Cmax (unbound)	Pharmacokinetic Properties		Predicted Property	534	N/A	Maximum unbound	Smith et al

						concentration in plasma	
DILIst	DILI	N/A	Test Data (DILI)	1123	698	DILIst Classification	Tong et al
DILIrank	DILI	N/A	Test Data (DILI)	152	122	DILIrank dataset	Chen et al, Chavan et al

Table 2: Performance of combination models from (a) 55 held-out test sets from repeated nested cross validation and (b) the 255 compounds in the held-out DILI dataset. rNCV: repeated nested cross validation

Model	**Features Used**	**Evaluation data**	**Balanced Accuracy (BA)**	**Mathew's correlation constant (MCC)**	**Area Under Curve-Receiver Operating Characteristic (AUC-ROC)**	**Sensitivity**	**Specificity**	**F1 Score**	**Likelihood ratio (LR+)**	**Positive predictive value (PPV)**	**Average precision score (AP)**
proxy-DILI data only	11 *in vitro* and *in vivo* labels	rNCV (mean)	0.64	0.28	0.69	0.59	0.69	0.72	2.09	0.76	0.79
		External Test	0.67	0.32	0.74	0.68	0.65	0.71	1.96	0.79	0.85
Chemical Structure only	Morgan Fingerprints and MACCS Keys	rNCV (mean)	0.63	0.26	0.69	0.60	0.66	0.70	1.87	0.75	0.78
		External Test	0.64	0.28	0.73	0.71	0.57	0.66	1.67	0.78	0.82
Physicochemical properties	Mordred Descriptors and Physicochemical parameters	rNCV (mean)	0.65	0.28	0.70	0.63	0.66	0.70	1.93	0.77	0.79
		External Test	0.68	0.37	0.72	0.67	0.70	0.75	2.29	0.79	0.79
Chemical Structure and Physicochemica	Morgan Fingerprints, MACCS Keys, Mordred Descriptors, Physicochemical	rNCV (mean)	0.65	0.29	0.70	0.64	0.66	0.71	1.93	0.77	0.80

l Properties	parameters	External Test	0.68	0.34	0.74	0.74	0.62	0.70	1.92	0.81	0.82
DILIPredictor	Morgan Fingerprints, MACCS Keys, Mordred Descriptors, Physicochemical parameters and 11 *in vitro* and *in vivo* labels	rNCV (mean)	0.66	0.31	0.71	0.64	0.67	0.72	2.02	0.78	0.80
		External Test	0.70	0.39	0.74	0.67	0.73	0.76	2.50	0.80	0.81

Table 3: Previously published models used in the evaluation of hepatotoxicity/liver injury (for test sets only)

Model	Features	Compounds in the Train set	Compounds in the Test set	Balanced Accuracy	AUC-ROC	Source
Ensemble of RF and SVM	Molecular fingerprints	1241	286	0.82	0.9	Ai et al. (2018)
Random Forests	Imaging Phenotypes and Chemical Descriptors	346	41	0.52	0.74	Chavan et al (2020)
Ensemble Models	Molecular features, physicochemical properties	1254	204	0.72	0.73	He et al (2019)
Random Forests	2D molecular descriptors	996	341	0.67	0.71	Kotsampasakou et al. (2017)
Random Forests	2D molecular descriptors	996	921	0.57	0.59	Kotsampasakou et al. (2017)
SVM	Morgan Fingerprints	923	49	0.67	-	Liu et al (2021)
SVM	Predicted protein targets	923	49	0.59	-	Liu et al (2021)
Random Forests	0−2.5D molecular descriptors	1075	554	0.77	0.81	Mora et al (2020)
GA-SVM	2D and 3D molecular descriptors	3712	375	0.75	0.73	Mulliner et al (2016)
Random Forests	Morgan Fingerprints	845	362	-	0.75	Ye et al (2022)
Naïve Bayes	Morgan Fingerprints	336	84	0.73	0.81	Zhang et al (2016)
SVM	MACCS keys	1317	88	0.68	0.62	Zhang et al. (2016)
Average				0.68	0.74	

Random Forests	**Structural, Physicochemical, *in vitro* and *in vivo***	**1020**	**225**	0.7	0.74	**Present Study**

Table 4: Table 4: DILI predictions for 14 compounds known to cause DILI and 2 compounds which do not cause DILI in humans (not used in training models in this study) and top 3 proxy-DILI labels positively and negatively for contributing to the prediction.

Compound Name	DILI (Literature)	DILI Prediciton	DILI Probability	Hepatotoxic Potential (Chen et al)	Remarks	Most Contribution proxy-DILI endpoints to Prediction		
						Ranked 1	Ranked 2	Ranked 3
2-Butoxyethanol	Not Toxic	Not Toxic	0.53		Known DILI in mice, not in human	Liver Toxicity Knowledge Base	Human hepatotoxicity	Diverse DILI B
Acitretin	Toxic	Toxic	0.88		Vitamin A derivative; Retinoid that causes DILI	Preclinical hepatotoxicity	Mitotox	BESP
Astaxanthin	Not Toxic	Not Toxic	0.56		Vitamin A derivative; High structural similarity to retinoid but does not cause DILI	Human hepatotoxicity	Diverse DILI C	Liver Toxicity Knowledge Base
Cabazitaxel	Toxic	Toxic	0.86			Diverse DILI B	Mitotox	BESP
Clopidogrel	Toxic	Toxic	0.8	Less toxic	Thienopyridines ; Adverse Reaction	Preclinical hepatotoxicity	Liver Toxicity Knowledge Base	Human hepatotoxicity
Docetaxel	Toxic	Toxic	0.86		In Market	Diverse DILI B	Mitotox	BESP
Doxycycline	Toxic	Toxic	0.72	Less toxic	Tetracyclines; Adverse Reaction	Diverse DILI B	Diverse DILI A	Liver Toxicity Knowledge Base

Entacapone	Toxic	Toxic	0.82	Less toxic	Catechol-O-methyl transferase inhibitors	Preclinical hepatotoxicity	Liver Toxicity Knowledge Base	Mitotox
Enzalutamide	Toxic	Toxic	0.83			Preclinical hepatotoxicity	Diverse DILI B	Liver Toxicity Knowledge Base
Gemcitabine	Toxic	Toxic	0.77		In Market	Liver Toxicity Knowledge Base	Diverse DILI A	Diverse DILI B
Minocycline	Toxic	Toxic	0.73	More toxic	Tetracyclines, Wanring/Precaution	Diverse DILI B	Diverse DILI A	Liver Toxicity Knowledge Base
Moxifloxacin	Toxic	Toxic	0.84	Less toxic	Fluoroquinolones, Adverse Reaction	Preclinical hepatotoxicity	Diverse DILI B	Liver Toxicity Knowledge Base
Paclitaxel	Toxic	Toxic	0.87			Diverse DILI B	Mitotox	BESP
Sitaxentan	Toxic	Toxic	0.82		Withdrawn	Preclinical hepatotoxicity	Diverse DILI B	Liver Toxicity Knowledge Base
Trovafloxacin	Toxic	Toxic	0.86	More toxic	Fluoroquinolones, Withdrawn	Preclinical hepatotoxicity	Diverse DILI B	Liver Toxicity Knowledge Base
Ximelagatran	Toxic	Toxic	0.82		Withdrawn	Diverse DILI B	Liver Toxicity Knowledge Base	Human hepatotoxicity

FIGURES

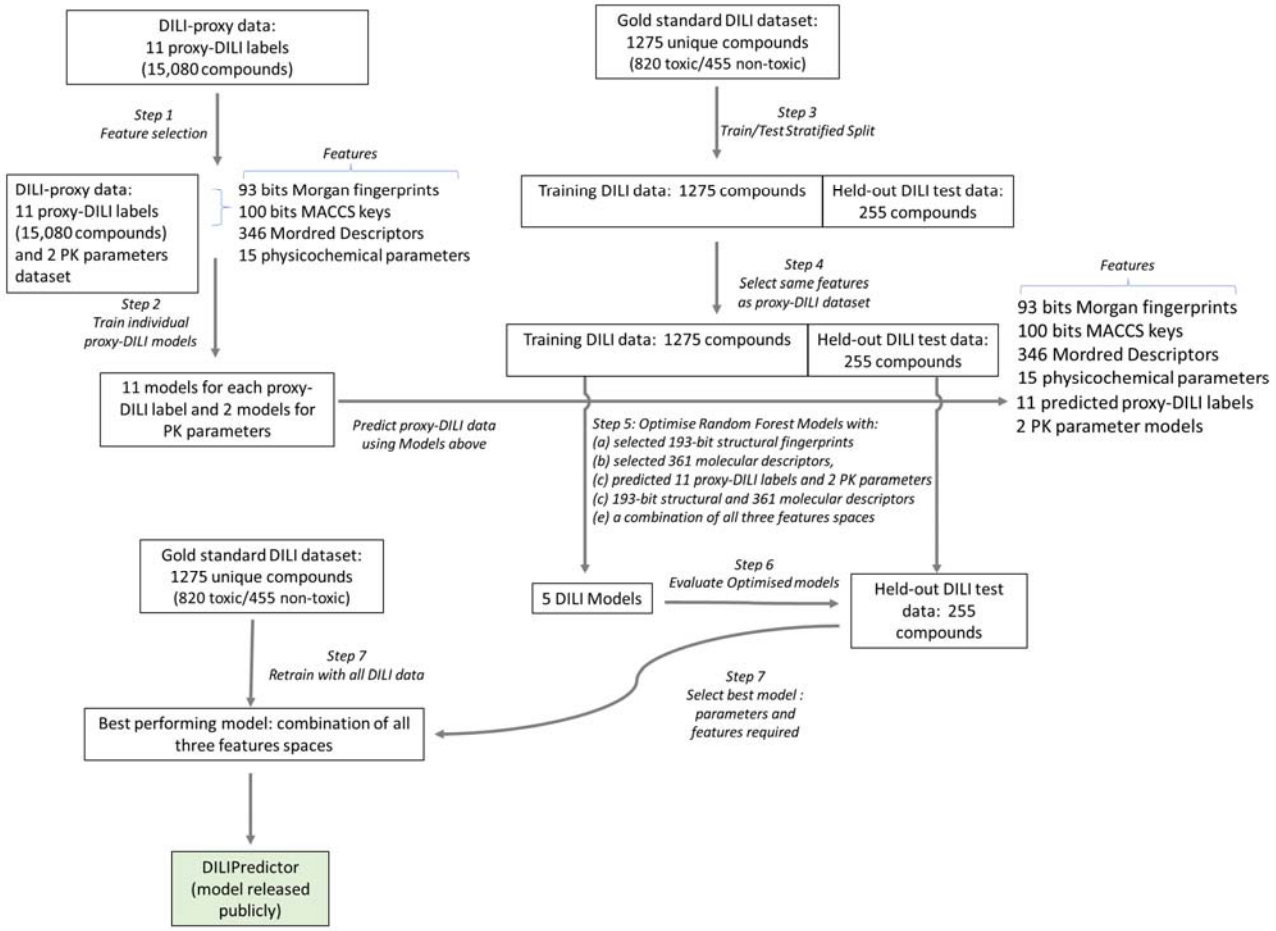


Figure 1: Workflow of the current study. Individual models for eleven *in vivo* and *in vitro* assays in the Proxy-DILI dataset and 2 PK parameters were used to predict these endpoints for compounds in the gold standard DILI dataset. A combination of these predictions along with chemical structure and molecular descriptors were then used to train and evaluate the models on DILI compounds.

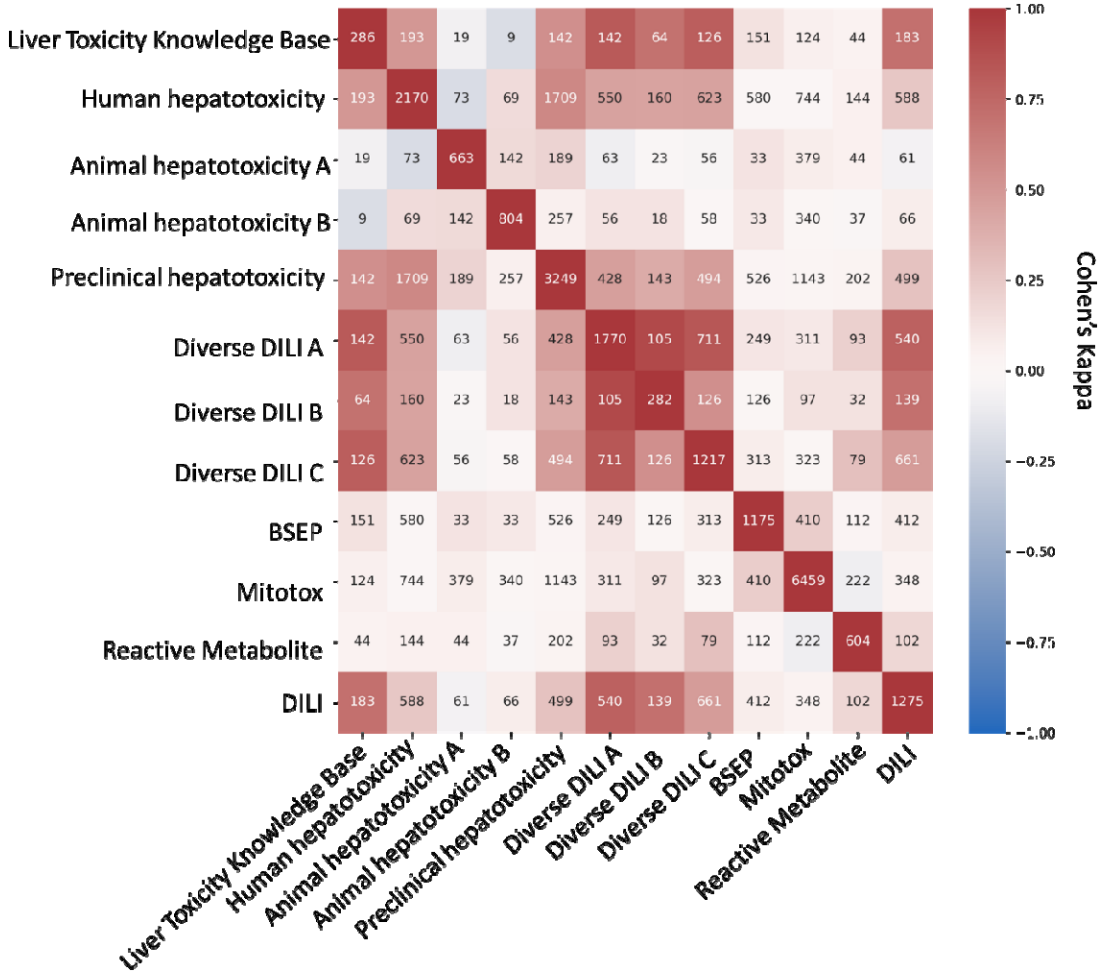


Figure 2: Concordance of compounds overlapping in-between eleven labels in the proxy-DILI dataset (18,679 compounds) including compounds that overlapped with DILI data (1,275 compounds). Concordance is given using Cohen's kappa (and the number of overlapping compounds given as annotations). Overall, the human-related proxy-DILI labels and diverse heterogenous DILI labels showed high concordance with DILI compounds and among each other.

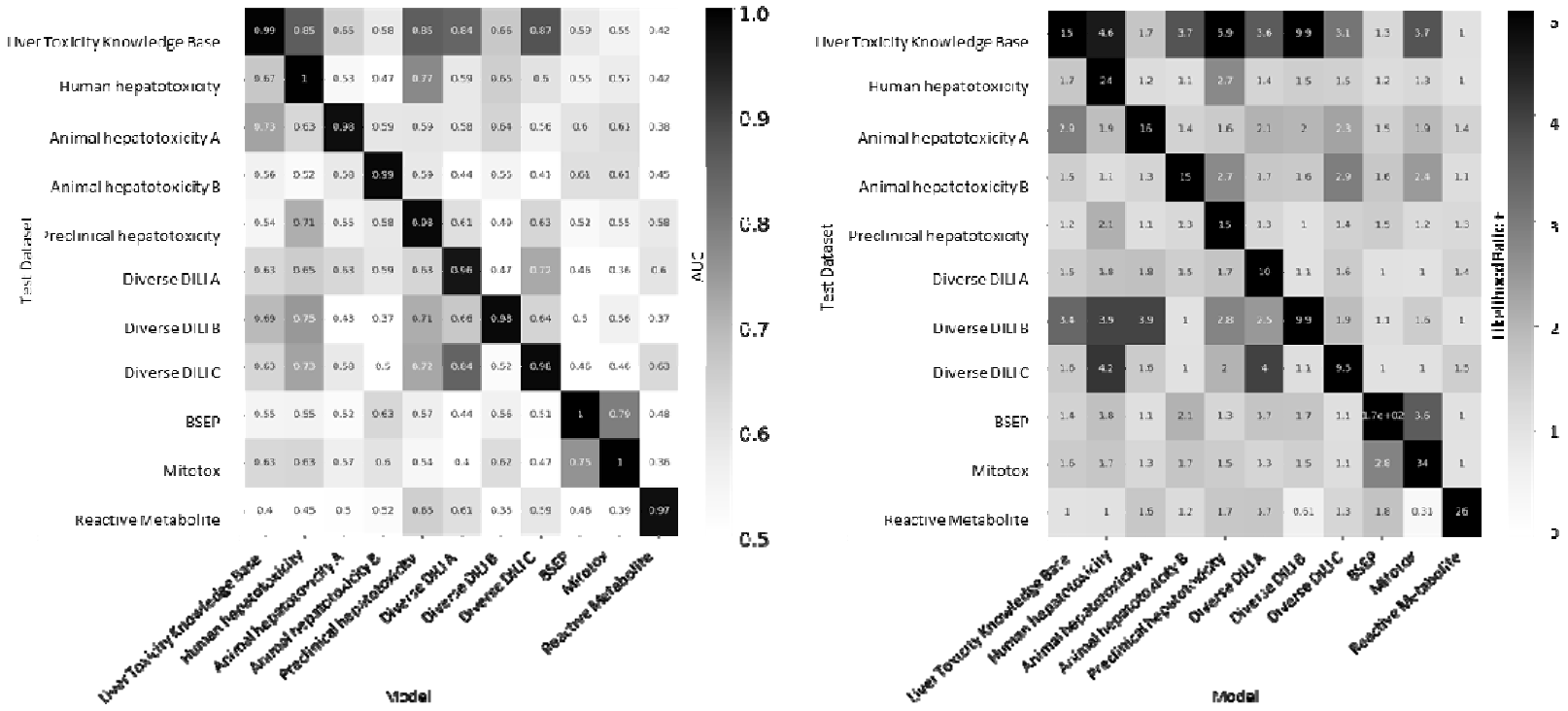


Figure 3. Performance metrics for models built on eleven proxy-DILI labels when predicting labels for the other proxy-DILI in the model were evaluated using (a) AUC-ROC and (b) Likelihood Ratio (LR+).

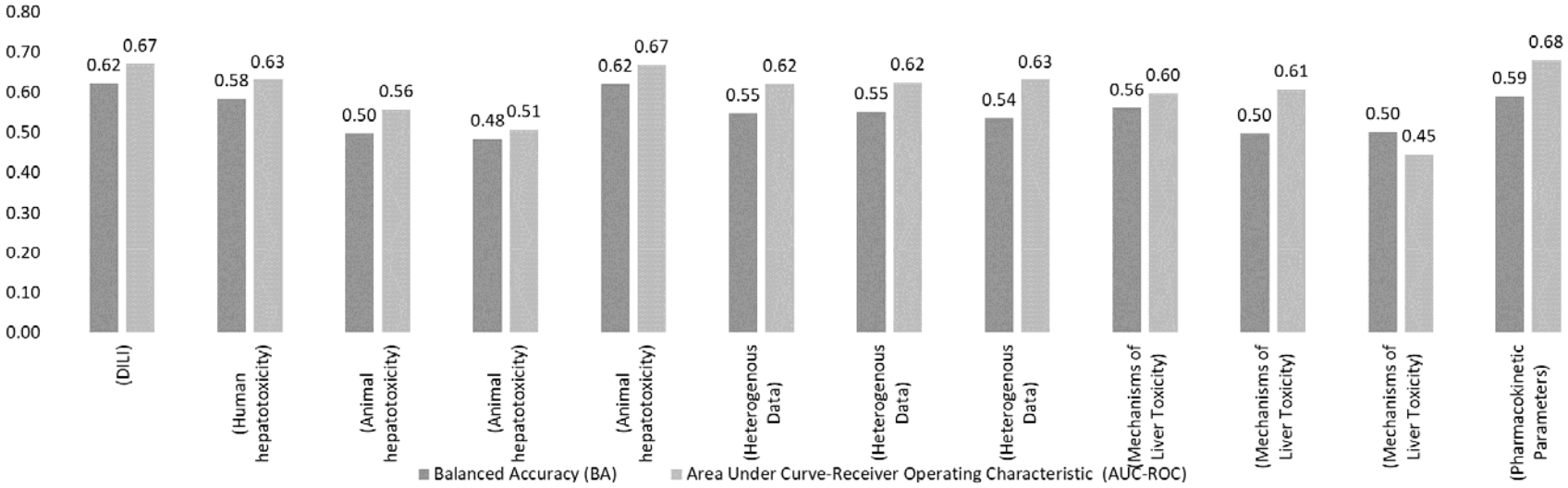


Figure 4. Performance Metrics AUC-ROC and Balanced Accuracy achieved by each of eleven individual models built on the proxy-DILI labels and a model built on two pharmacokinetic parameters (Cmax total and unbound) when tested on the 255 compounds in the held-out DILI dataset.

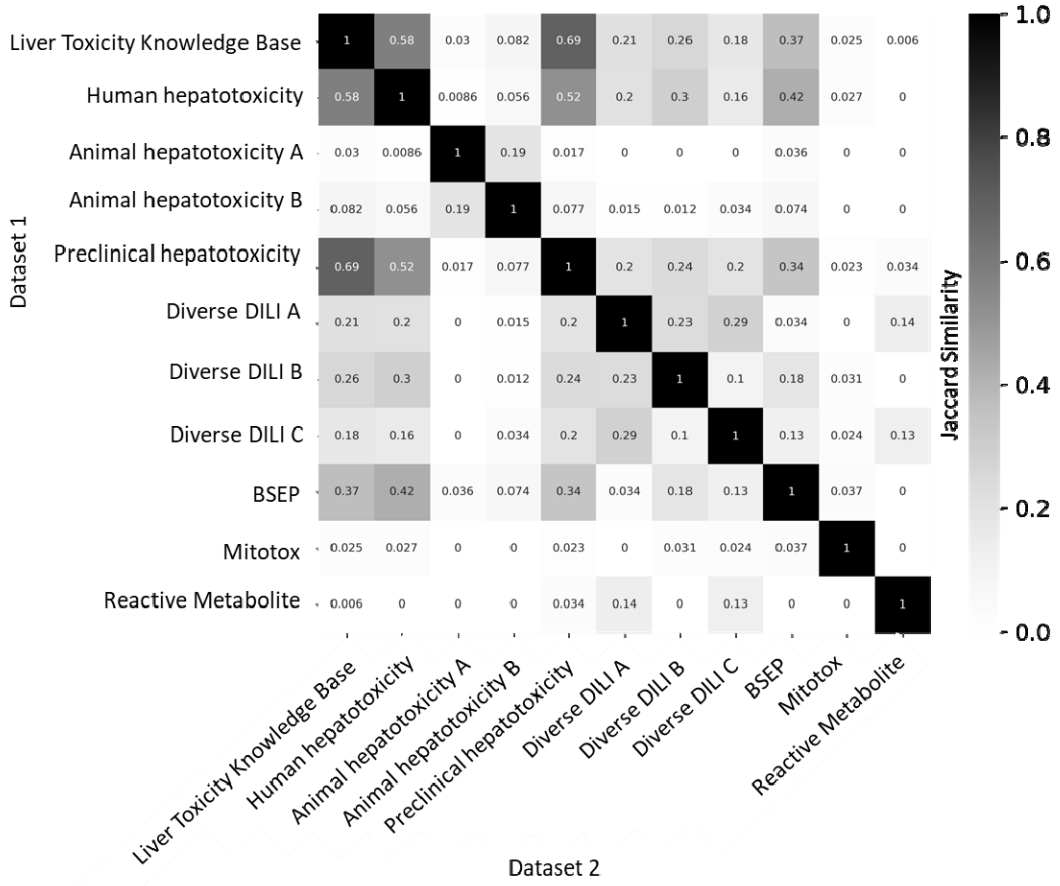


Figure 5: Jaccard Similarity of predictions on the held-out DILI dataset (255 compounds) for individual models built on eleven proxy-DILI labels in the proxy-DILI dataset.

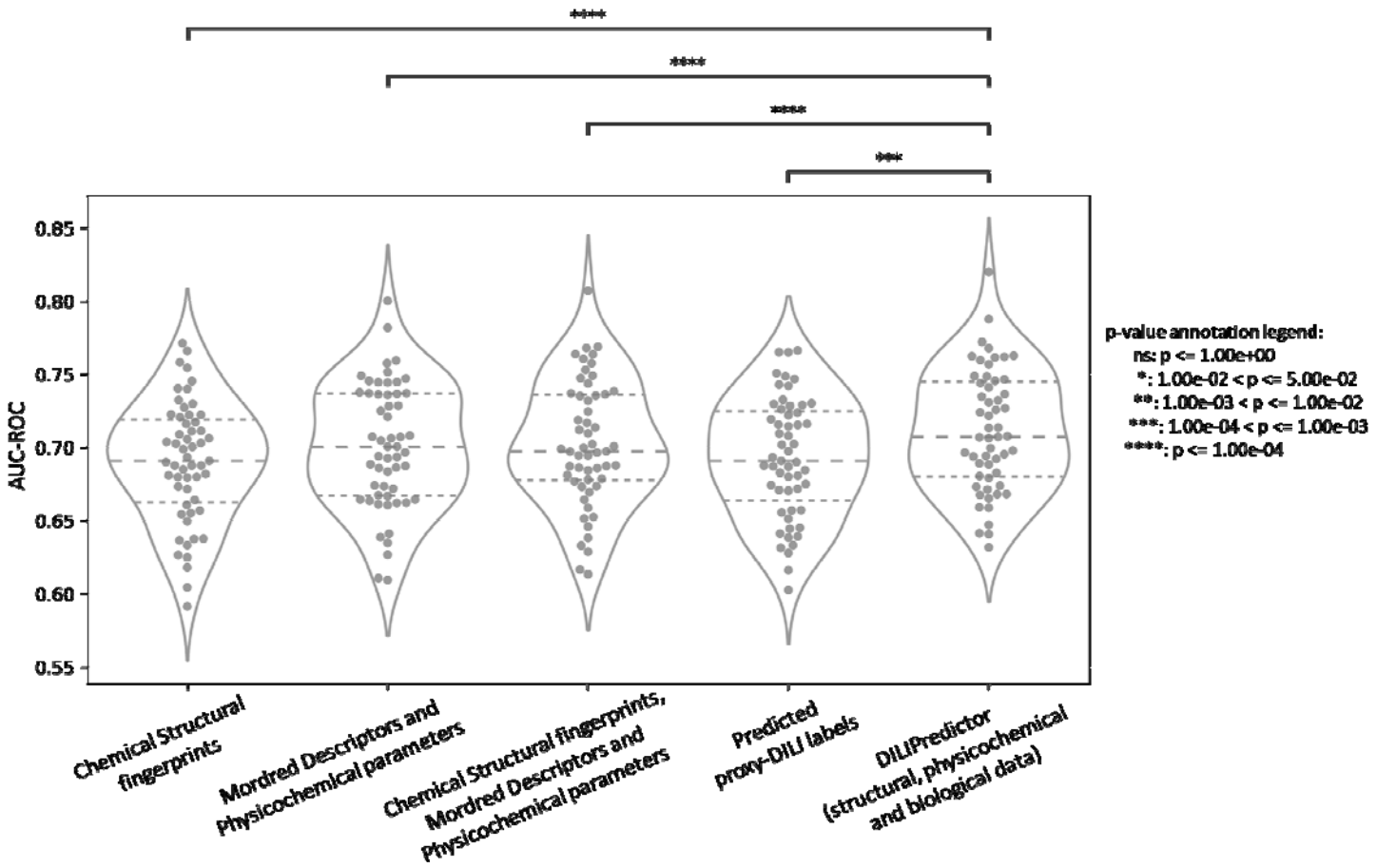


Figure 6. Performance metrics AUC-ROC for combination models from 55 held-out test sets from repeated nested cross-validation using (a) selected 193-bit structural fingerprints, (b) selected 361 molecular descriptors, (c) selected 193-bit structural fingerprints and selected 361 molecular descriptors, (d) predicted eleven proxy-DILI labels and 2 PK parameters, and (e) a combination of all three features spaces compared with a paired t-test.

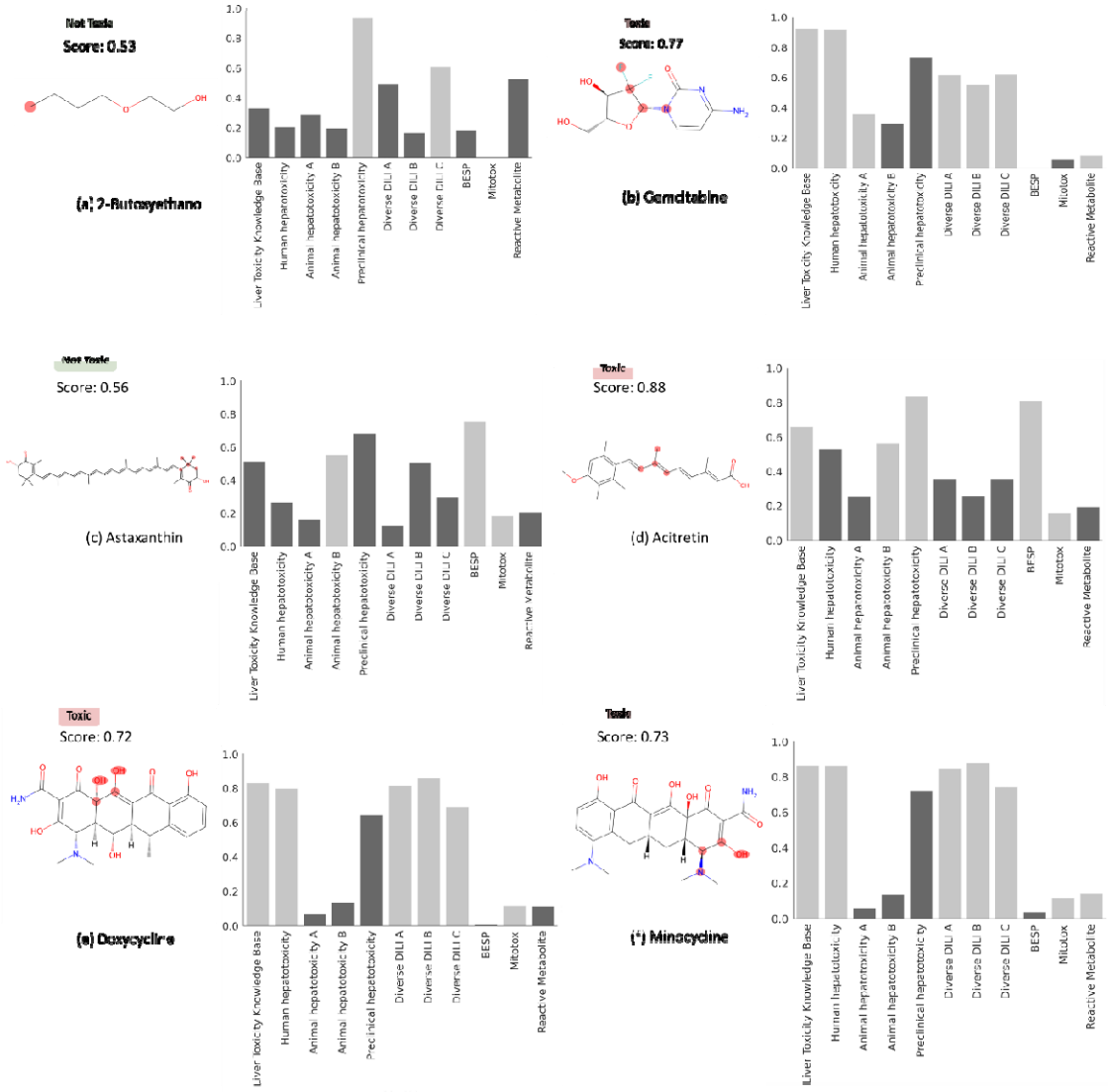
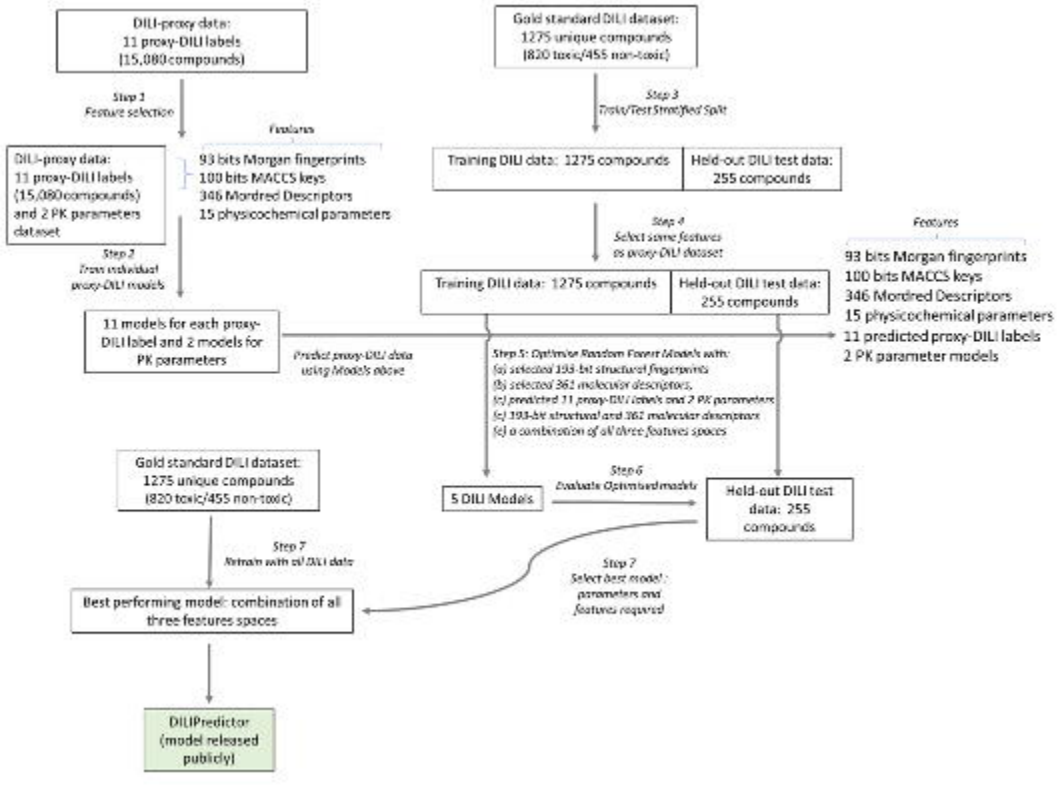
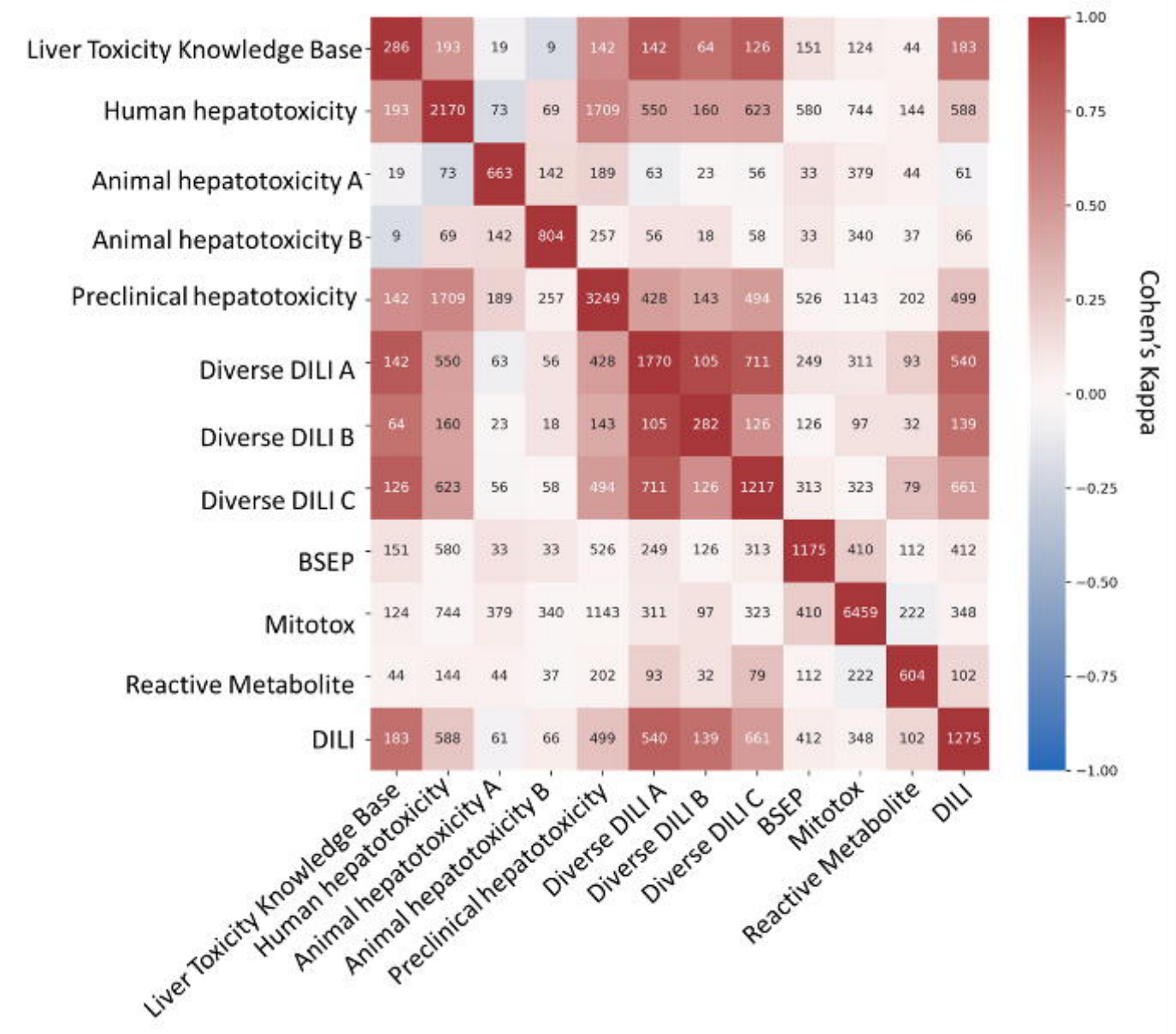
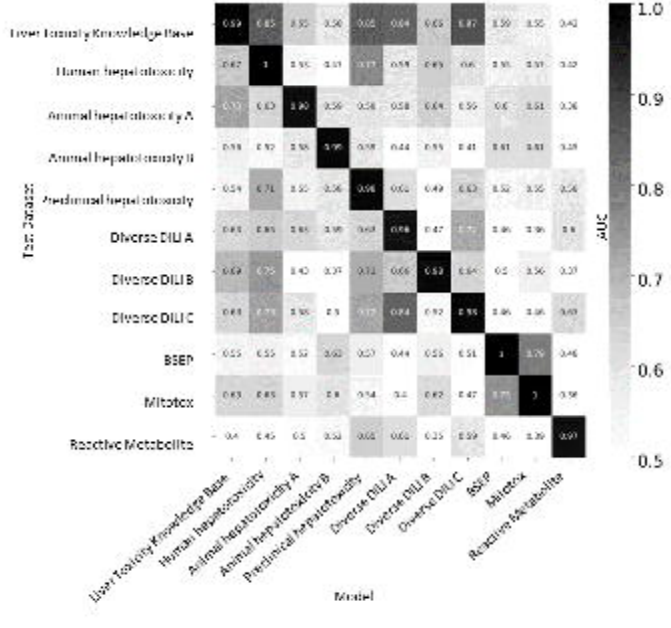
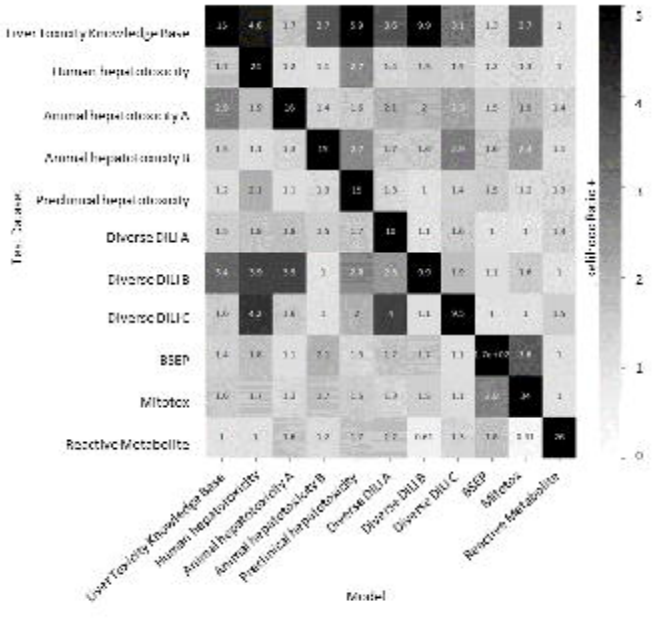


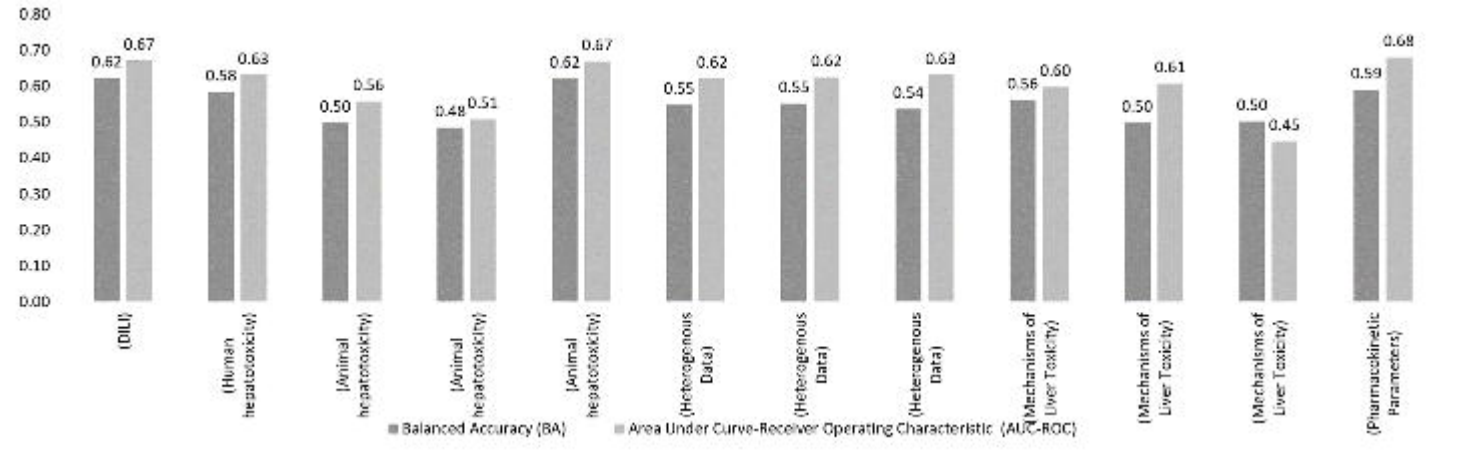
Figure 7: MACCS substructure (highlighted) and proxy-DILI labels positively contributing to DILI when using DILIPredictor for four compounds known to cause DILI and two compounds which do not cause DILI in humans (further details for another 10 compounds in Supplementary Figure S5). The highest positive contribution from the MACCS substructure is highlighted with the chemical structure.

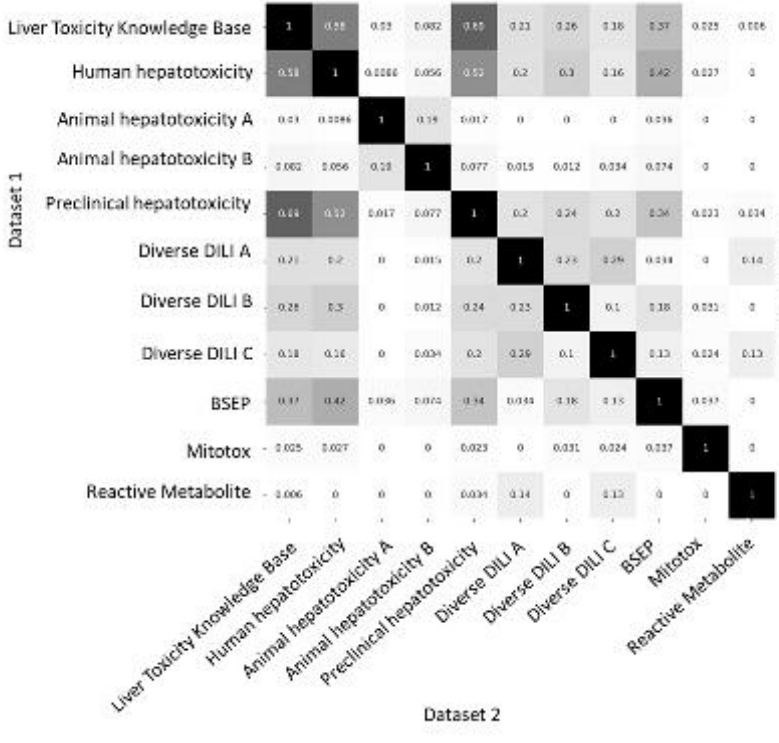

DILI-proxy data:
11 proxy-DILI labels
(15,080 compounds)
Step 1
Feature selection
Features
DILI-proxy data:
11 proxy-DILI labels
(15,080 compounds)
and 2 PK parameters
dataset
93 bits Morgan fingerprints
100 bits MACCS keys
346 Mordred Descriptors
15 physicochemical parameters
Step 2
Train individual
proxy-DILI models
11 models for each proxy-DILI label and 2 models for PK parameters
Predict proxy-DILI data using Models above
Gold standard DILI dataset:
1275 unique compounds
(820 toxic/455 non-toxic)
Step 3
Train/Test Stratified Split
Training DILI data: 1275 compounds
Held-out DILI test data: 255 compounds
Step 4
Select same features as proxy-DILI dataset
Training DILI data: 1275 compounds
Held-out DILI test data: 255 compounds
Features
93 bits Morgan fingerprints
100 bits MACCS keys
346 Mordred Descriptors
15 physicochemical parameters
11 predicted proxy-DILI labels
2 PK parameter models
Step 5: Optimize Random Forest Models with:
(a) selected 193-bit structural fingerprints
(b) selected 362 molecular descriptors,
(c) predicted 11 proxy-DILI labels and 2 PK parameters
(d) 193-bit structural and 362 molecular descriptors
(e) a combination of all three features spaces
Gold standard DILI dataset:
1275 unique compounds
(820 toxic/455 non-toxic)
5 DILI Models
Step 6
Evaluate Optimized models
Held-out DILI test data: 255 compounds
Step 7
Retrain with all DILI data
Step 7
Select best model parameters and features required
Best performing model: combination of all three features spaces
DILIPredictor
(model released publicly)

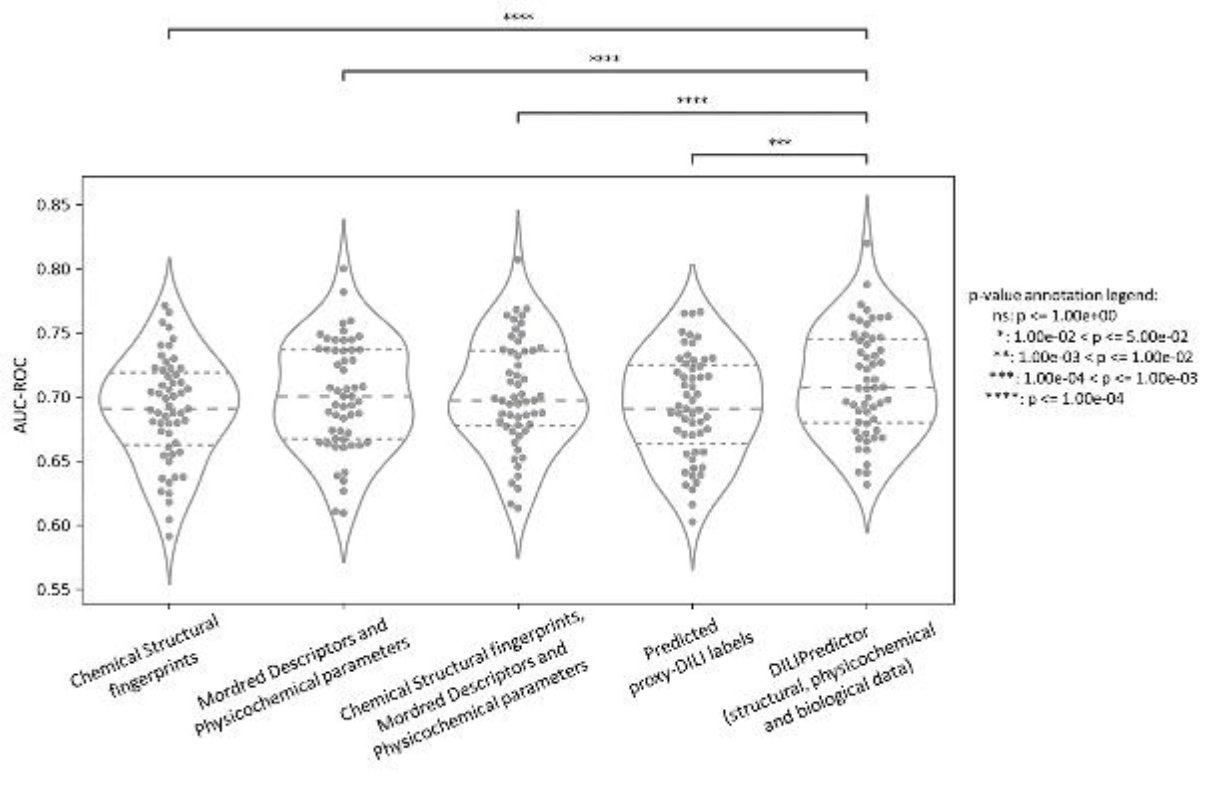

Liver Toxicity Knowledge Base
Human hepatotoxicity
Animal hepatotoxicity A
Animal hepatotoxicity B
Preclinical hepatotoxicity
Diverse DILI A
Diverse DILI B
Diverse DILI C
BSEP
Mitotox
Reactive Metabolite
DILI
286 193 19 9 142 142 64 126 151 124 44 183
193 2170 73 69 1709 550 160 623 580 744 144 588
19 73 663 142 189 63 23 56 33 379 44 61
9 69 142 804 257 56 18 58 33 340 37 66
142 1709 189 257 3249 428 143 494 526 1143 202 499
142 550 63 56 428 1770 105 711 249 311 93 540
64 160 23 18 143 105 282 126 126 97 32 139
126 623 56 58 494 711 126 1217 313 323 79 661
151 580 33 33 526 249 126 313 1175 410 112 412
124 744 379 340 1143 311 97 323 410 6459 222 348
44 144 44 37 202 93 32 79 112 222 604 102
183 588 61 66 499 540 139 661 412 348 102 1275
Liver Toxicity Knowledge Base
Human hepatotoxicity
Animal hepatotoxicity A
Animal hepatotoxicity B
Preclinical hepatotoxicity
Diverse DILI A
Diverse DILI B
Diverse DILI C
BSEP
Mitotox
Reactive Metabolite
DILI
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Cohen's Kappa

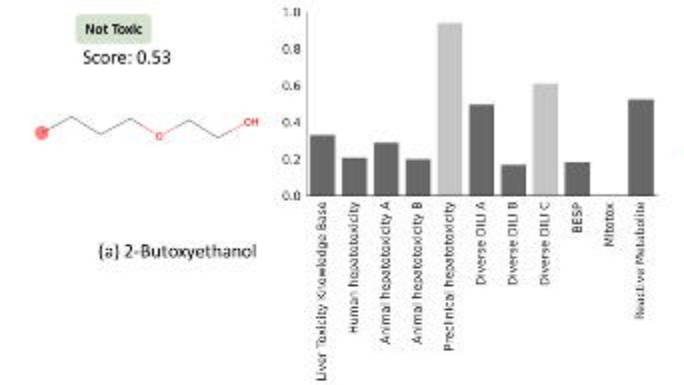

Liver Toxicity Knowledge Base
Human Hepatotoxicity
Animal Hepatotoxicity A
Animal Hepatotoxicity B
Preclinical Hepatotoxicity
Diverse DILI A
Diverse DILI B
Diverse DILI C
BSEP
Mitotox
Reactive Metabolite
Test Dataset
Model
AUC
1.0
0.9
0.8
0.7
0.6
0.5


Liver Toxicity Knowledge Base
Human Hepatotoxicity
Animal Hepatotoxicity A
Animal Hepatotoxicity B
Preclinical Hepatotoxicity
Diverse DILI A
Diverse DILI B
Diverse DILI C
BSEP
Mitotox
Reactive Metabolite
Test Dataset
Model
odds ratio for AUC
5
4
3
2
1
0

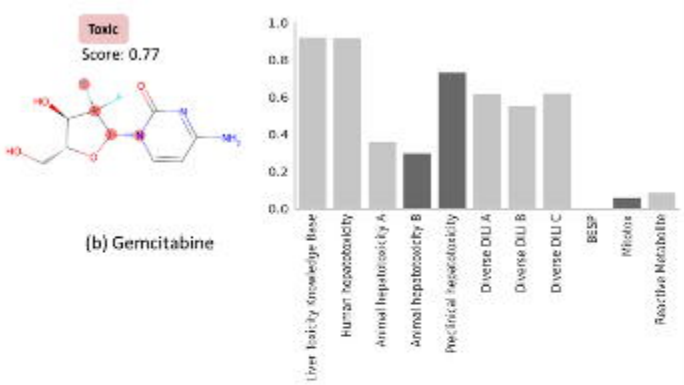

0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.62
0.67
0.58
0.63
0.50
0.56
0.48
0.51
0.62
0.67
0.55
0.62
0.55
0.62
0.54
0.63
0.56
0.60
0.50
0.61
0.50
0.45
0.59
0.68
(DILI)
(Human hepatotoxicity)
(Animal hepatotoxicity)
(Animal hepatotoxicity)
(Animal hepatotoxicity)
(Heterogeneous Data)
(Heterogeneous Data)
(Heterogeneous Data)
(Mechanisms of Liver Toxicity)
(Mechanisms of Liver Toxicity)
(Mechanisms of Liver Toxicity)
(Pharmacokinetic Parameters)
Balanced Accuracy (BA)
Area Under Curve-Receiver Operating Characteristic (AUC-ROC)


Liver Toxicity Knowledge Base
Human hepatotoxicity
Animal hepatotoxicity A
Animal hepatotoxicity B
Preclinical hepatotoxicity
Diverse DILI A
Diverse DILI B
Diverse DILI C
BSEP
Mitotox
Reactive Metabolite
Dataset 1
Dataset 2
Jaccard Similarity
1.0
0.8
0.6
0.4
0.2
0.0

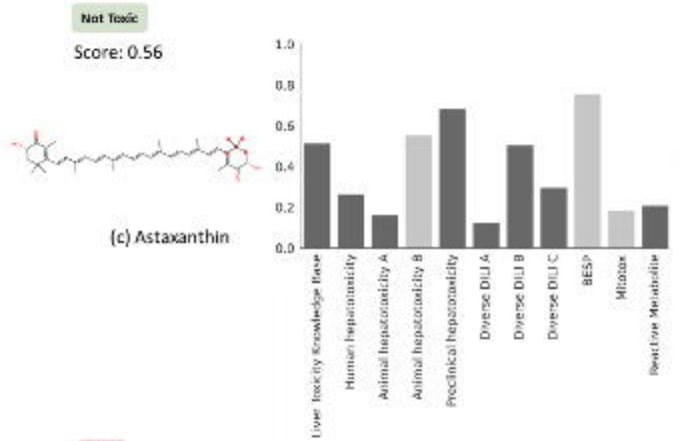

****
****
****
***
AUC-ROC
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Chemical Structural fingerprints
Mordred Descriptors and Physicochemical parameters
Chemical Structural fingerprints, Mordred Descriptors and Physicochemical parameters
Predicted proxy-DILI labels
DILIPredictor (structural, physicochemical and biological data)
p-value annotation legend:
ns: p <= 1.00e+00
*: 1.00e-02 < p <= 5.00e-02
**: 1.00e-03 < p <= 1.00e-02
***: 1.00e-04 < p <= 1.00e-03
****: p <= 1.00e-04



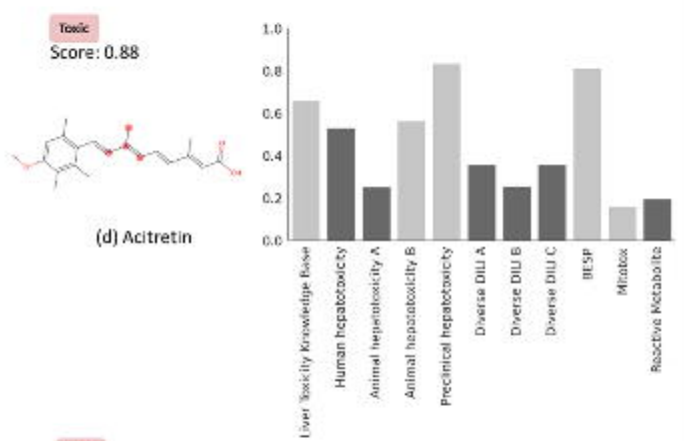
(a) 2-Butoxyethanol



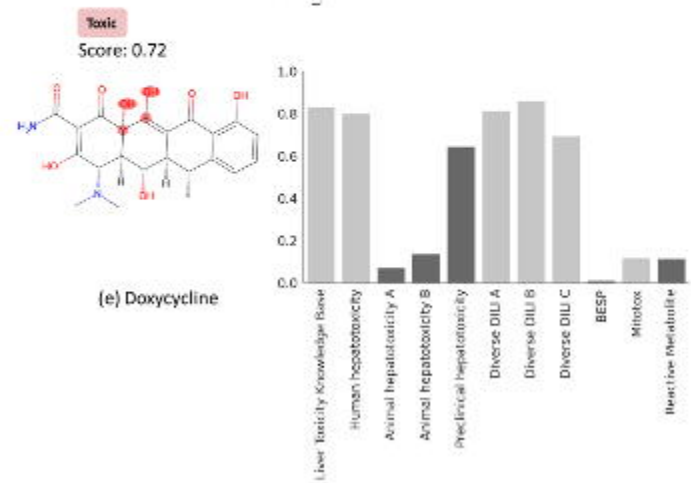
(b) Gemcitabine



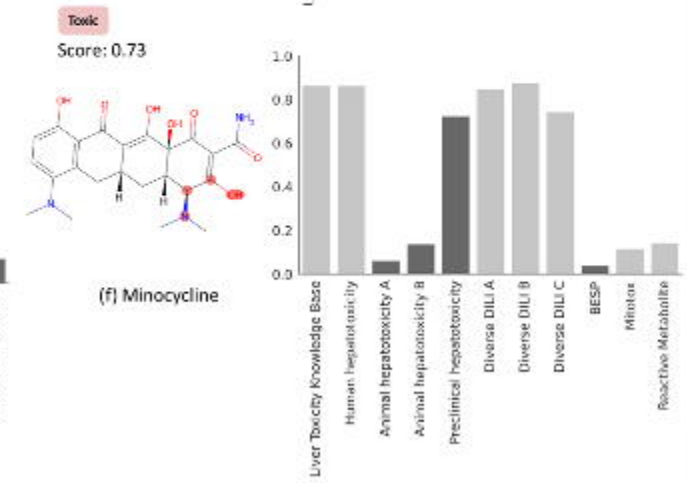
(c) Astaxanthin



(d) Acitretin



(e) Doxycycline



(f) Minocycline