ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

pDILI_v1: A Web-Based Machine Learning Tool for Predicting Drug-Induced Liver Injury (DILI) Integrating Chemical Space Analysis and Molecular Fingerprints

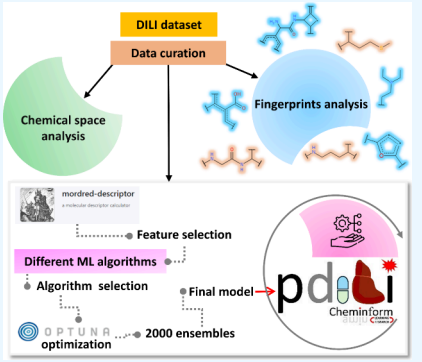
Sk Abdul Amin,* Supratik Kar,* and Stefano Piotto

Cite This: *ACS Omega* 2025, 10, 13502−13514

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Drug-induced liver injury (DILI) represents a critical safety concern for drug development, regulatory oversight, and clinical practice, with substantial economic and public health implications. While predicting DILI risk in humans has garnered significant attention, the associated chemical space has remained insufficiently explored. This study addresses this gap through a comprehensive computational approach, leveraging machine learning (ML) to investigate structural determinants of DILI risk systematically. The study focuses on three key objectives: (i) exploring the chemical space and scaffold diversity associated with DILI; (ii) employing fragment-based approaches to identify structural alerts (SAs) that influence DILI risk; and (iii) developing supervised ML models to not only predict DILI risk but also elucidate the structural significance of molecular fingerprints. To broaden accessibility, we introduce pDILI_v1, a Python-based web application available at https://pdiliv1web.streamlit.app/. This user-friendly platform facilitates the prediction and visualization of DILI risk, enabling both experts and nonexperts to screen compounds effectively. Additional formats, including a Google Colab notebook and a graphical user interface (GUI) for Windows, ensure flexibility for diverse user needs. The proposed models demonstrate the potential for early identification of hepatotoxic risks in drug candidates, providing critical insights into drug discovery and development. By integrating ML-driven predictions with chemical space analysis, this research advances the field of drug safety evaluation, contributing to the development of safer pharmaceuticals and mitigating the risks of DILI.



1. INTRODUCTION

Drug-induced liver injury (DILI) represents a critical challenge in pharmaceutical research and drug development, posing significant clinical and economic risks.[1,2] As a leading cause of acute liver failure, DILI manifests through two primary mechanisms: intrinsic reactions, which are dose-dependent and predictable, and idiosyncratic responses, characterized by their unpredictability and independence from dosage.[3] The complexity of DILI is underscored by its substantial impact on drug development. Thousands of individuals are affected annually, with DILI emerging as a primary catalyst for postmarket drug withdrawals and a major obstacle in successful clinical drug candidate progression. The high attrition rate of potential pharmaceutical compounds, frequently attributed to unforeseen hepatotoxicity, highlights the urgent need for advanced predictive methodologies.[4,5]

Advancements in machine learning (ML), coupled with the growing availability of high-quality, open-access experimental data, have significantly improved our ability to predict potential DILI risks. These computational models enable the systematic analysis of chemical and structural data, facilitating early identification of DILI in novel or untested drugs. This approach is especially valuable for lead compounds under consideration for preclinical and clinical trials, helping to mitigate the risk of late-stage drug failures, where the cost and impact are particularly high. Vall et al.[6] summarized AI and ML approaches for DILI prediction, including random forests and deep learning. It emphasized the challenges due to limited data availability and highlighted future directions involving advanced modalities such as 3D spheroids to improve annotations and model performance. In another comprehensive review, Shin et al.[7] highlighted advancements in silico models incorporating ML and adverse outcome pathways (AOPs) for DILI prediction. It emphasized the combination of in vitro data and structural information to enhance model accuracy and interpretability. These methods also extend to predicting herb-induced liver injury (HILI), showcasing versatility.

Seal et al.[8] created DILIPredictor, an ML model that integrates in vitro, in vivo, and structural data achieving prediction performance of AUC-PR of 0.79 by leveraging nine

Received: January 3, 2025
Revised: March 6, 2025
Accepted: March 18, 2025
Published: March 25, 2025

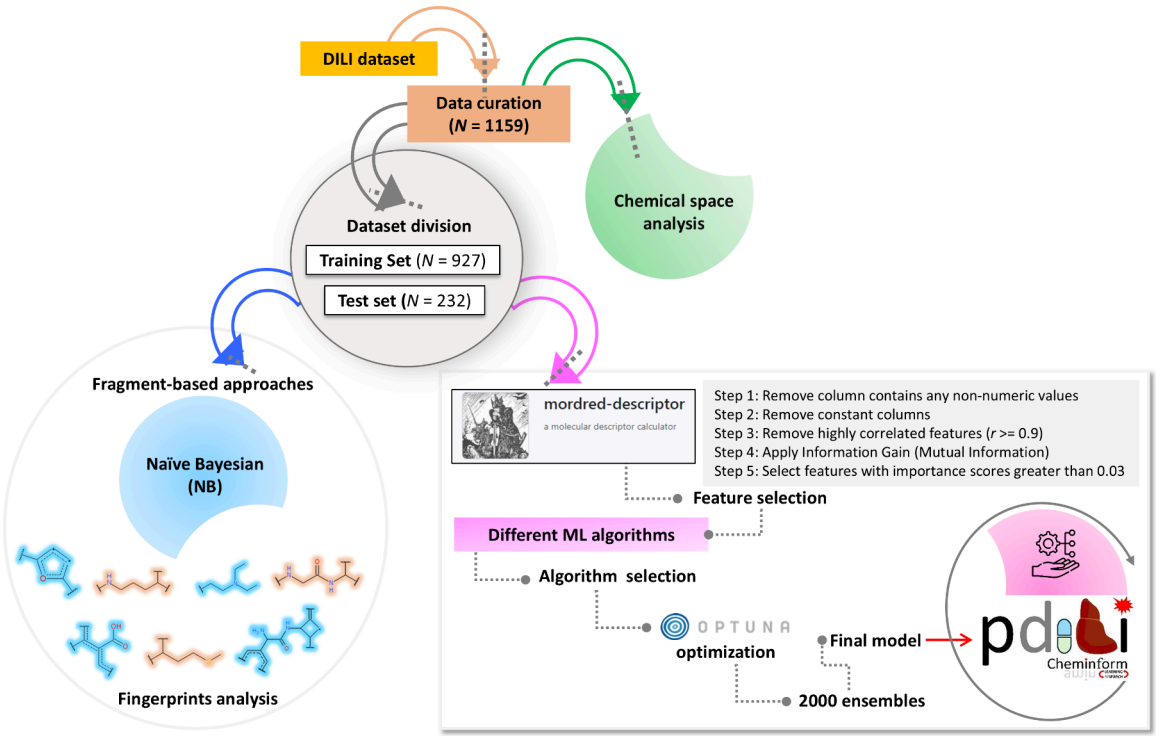


Figure 1. Workflow of the current study involves different approaches such as (i) chemical space analysis, (ii) the fragment-based approach, and (iii) ML-based QSAR.

proxy-DILI labels as features, allowing differentiation between animal and human sensitivities to DILI. The developed model identifies chemical substructures contributing to DILI and is accessible via a web interface. Lee and Yoo[9] proposed interpretable DILI prediction models using ML with permutation feature importance and attention mechanisms. By employing substructure and physicochemical descriptors, the models identified molecular features linked to DILI risks, achieving AUROC values of 0.88−0.97. This approach provides both predictive performance and mechanistic insights. Ye et al.[10] compared chemical structure-based models and in vitro assay data for DILI prediction. The best chemical structure-based models achieved an AUC-ROC of 0.75, while assay data alone showed only moderate predictive power. The combination of chemical and assay data did not significantly enhance the prediction accuracy, underscoring limitations in assay coverage. Shin et al.[11] introduced ToxSTAR, a web-based tool using ML models to predict four DILI subtypes: cholestasis, cirrhosis, hepatitis, and steatosis. The models leverage structural similarity and molecular descriptors, providing a user-friendly interface for researchers to assess DILI risks.

While several studies have employed quantitative structure−activity relationships (QSARs) and ML modeling to predict DILI, our approach introduces a novel combination of comprehensive chemical space analysis, fingerprint-based structural alerts (SAs) identification, and an accessible, user-friendly prediction tool called "pDILI_v1". Our research introduces a comprehensive ML-based strategy for DILI risk assessment, focusing on molecular fingerprint analysis. The study's key objectives include:

a. Developing predictive models to identify structural contributors to hepatotoxicity.
b. Creating an open-access tool, pDILI_v1.0, for molecular DILI risk screening that enables nonexperts to screen molecules for DILI risk using SMILES input.
c. Providing insights into the structural determinants that influence liver toxicity.

The undertaken study is illustrated with a flow diagram in Figure 1. By leveraging advanced computational techniques, this work aims to transform our understanding of DILI mechanisms and provide researchers with a scalable, accessible tool for early stage risk mitigation in drug development. This integration of chemical informatics, ML optimization, and accessibility not only advances predictive accuracy but also enhances the usability and impact of DILI prediction tools in drug development.

2. MATERIAL AND METHODS

2.1. Data Set. A large data set of drugs that are divided into binary classes (1: *Toxic*, 0: *Nontoxic*) according to their potential for causing DILI was investigated. The experimental data was taken from publicly available sources.[12] The list of all the drugs together with their class (1: DILI *Toxic*, 0: DILI *Nontoxic*) is shown in Table S1. Since the dependent variables (class 1: DILI *Toxic*, class 0: DILI *Nontoxic*) were binary data (Figure 2), the classification modeling process was considered.

2.2. Chemical Space Exploration. Chemical space is crucial in chemical and biological research, especially in medicinal chemistry.[13] Initially, the chemical space of the compounds was analyzed by the frequency distribution of eight molecular properties such as the number of aromatic rings (*nAR*), number of rotatable bonds (*nRB*), lipophilicity (*LogP*), molecular weight (*MW*), number of rings (*nRings*), topological polar surface area (*TPSA*), and number of hydrogen bond acceptors (*nHBA*) and donors (*nHBD*). Further to demonstrate the structural diversity of the data set, a frequency

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

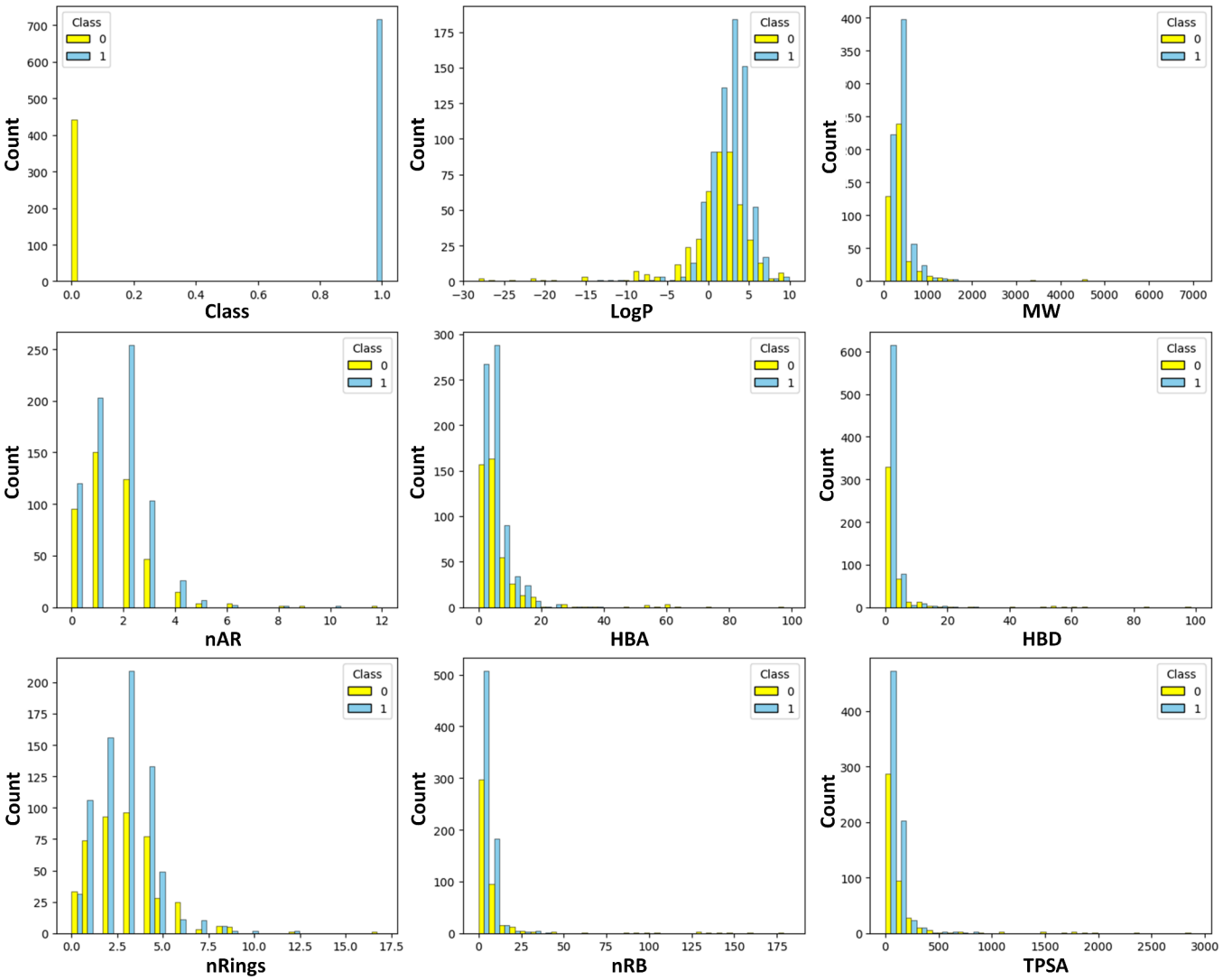


Figure 2. Bin plots of each feature (*Class*, *LogP*, *MW*, *nAR*, *HBA*, *HBD*, *nRings*, *nRB*, and *TPSA*), colored by DILI *Toxic* (1) and DILI *Nontoxic* (0).

distribution of similarity values has been plotted. Tanimoto coefficients (T_c)[14] were calculated using the Morgan fingerprint[15,16] to explore the molecular similarity of these 1159 compounds. The Tanimoto coefficient typically ranges between 0 (no similarity) and 1 (perfect similarity). The formula [n*(n-1)]/2 calculates the number of ways to choose 2 molecules from a set of "*n*" molecules. For 1159 molecules, we calculated the T_c for every possible pair of molecules. This analysis was performed by using *in house* Python code[13,17] with the RDKit module.[15,18]

2.3. Fragment-Based Analysis. *2.3.1. Data Set Division.* Primarily, the data set was divided using the k-means clustering method[19] which divided the data set of 1159 molecules into multiple clusters by the maximum dissimilarity approach based on earlier discussed molecular properties. Test set molecules were then selected from each cluster to ensure a balanced representation of molecular diversity.[20] Consequently, the PCA method was applied not only to visualize the chemical distribution of the 1159 compounds but also to understand whether the distribution of the test set compounds truly represents the training set or not.[21]

2.3.2. Construction of the Laplacian-Corrected Bayesian Model. Bayesian classification study[22,23] is a statistical technique primarily based on Bayes' theorem as depicted in eq 1.

$$P\left(\frac{h}{d}\right) = \frac{P\left(\frac{d}{h}\right)P\left(h\right)}{P(d)} \quad (1)$$

Where, (h/d) = Posterior probability, where h (hypothesis) and d (observed data), (d/h) = Likelihood, $P(h)$ = Prior belief, and $P(d)$ = Evidenced data

Here, the Bayesian classification model was developed.[22,23] Since our focus was to understand the critical substructure/fingerprint features regulating DILI, a topological fingerprint descriptor namely extended connectivity fingerprints of diameter 6 (*ECFP-6*)[24] was considered. The Bayesian model was developed on the training set molecules, and subsequently, validated on test set molecules.[25] Moreover, the predictive quality of the Bayesian model was analyzed by a 5-fold cross-validation technique including receiver operating characteristics (ROC) and other statistical parameters mentioned earlier.[25]

2.4. ML-Based QSAR Study. *2.4.1. Calculation of Descriptors and Feature Selection.* Features were calculated by using the Mordred calculator.[26] Then descriptors exhibiting

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

Table 1. Statistical Parameters of the Laplacian-Corrected Bayesian Classification Model[a]

set	TP	FN	FP	TN	Se	Ac	Pr	F1	FDR	FOR
train	474	98	27	328	0.829	0.865	0.946	0.884	0.053	0.230
test	114	31	36	51	0.786	0.711	0.760	0.773	0.240	0.378

[a]True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN).

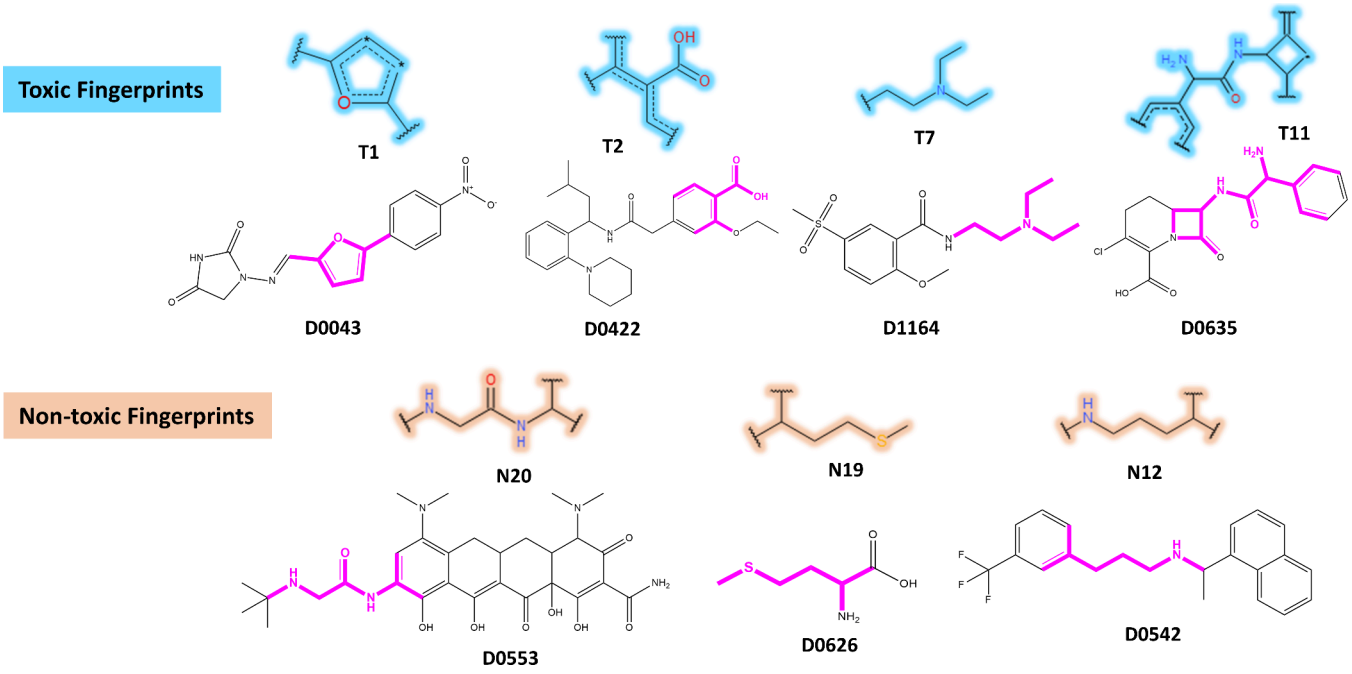


Figure 3. Representative drug compounds with DILI *toxic* (T) and DILI *nontoxic* (N) fingerprints/substructural features. DILI *toxic* fragments promote DILI risk, while DILI *nontoxic* fragments hinder DILI risk. These substructural features were produced by the *ECFP-6* fingerprint descriptor.

missing values, non-numeric entries, or quasi-constant behavior were eliminated from the study. A descriptor was treated as quasi-constant when a single value was observed in >98% of the samples. Figure 1 describes a typical data pretreatment and feature selection process in an ML workflow. The goal is to clean the data and reduce the number of features while keeping only the most informative ones, which helps to improve model performance, interpretability, and efficiency.[27,28]

2.4.2. ML Model Development. Seven ML algorithms, including Logistic Regression (LR), k-Nearest Neighbors (k-NN), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Quadratic Discriminant Analysis (QDA), and Multilayer perception (MLP) classifiers, were implemented to find the best one. The hyperparameters were optimized with the *SearchCV* method.[29,30] To select the best algorithm for the investigated data set, the variation in the *accuracy* and *precision* values was examined. Subsequently, the best ML algorithm is further considered for hyperparameter optimization by *Optuna*.[31,32] *Optuna* is a framework for hyperparameter optimization that automates the trial-and-error process of finding the optimal values for *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*. The developed models were validated by the statistical matrices as discussed earlier.[23]

3. RESULTS AND DISCUSSION

The focus (i) chemical space exploration, (ii) fragment-based analysis, and (iii) ML-based QSAR approach were considered to explore the structure−property relationship within the data set. A fragment-based as well as structure−property relationship analysis was performed to gain emergent knowledge that could be used to predict and screen the potential of a molecule for causing DILI.

3.1. Analysis of the Chemical Space. Chemical space is crucial in chemical and biological research, especially in medicinal chemistry.[13] Meanwhile, Figure 2 also shows the frequency of the distribution of the different molecular properties (*LogP*, *MW*, *nAR*, *HBA*, *HBD*, *nRings*, *nRB*, *TPSA*) in the data set.

The mean *LogP* values across all data set molecules is 1.923, suggesting most molecules are moderately lipophilic. The highest *LogP* value is found to be 9.908, indicating the most lipophilic molecule (Probucol) in the data set. The compound D0309 (Ecallantide) with *LogP* value −28.144 is the most hydrophilic molecule in this data set. This compound (*MW* = 7053.952) is also the largest molecule by weight, whereas ethanol (*MW* = 46.069) is the smallest molecule by weight. The average molecular weight (419.726) of molecules indicates that most of them are small to medium-sized. On average, molecules have one or two aromatic rings. There are 215 molecules without aromatic rings, 353 molecules with one aromatic ring, 378 molecules with two aromatic rings, and 41 molecules with three aromatic rings. Notably, most molecules have around 6 hydrogen bond acceptors. Molecules such as Mitotane, Bromobenzene, Lindane, Carbon tetrachloride, Perflutren, and Halothane are found with 0 *HBA*, suggesting that they cannot accept hydrogen bonds. These mentioned molecules are nonpolar. However, most molecules of the data

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

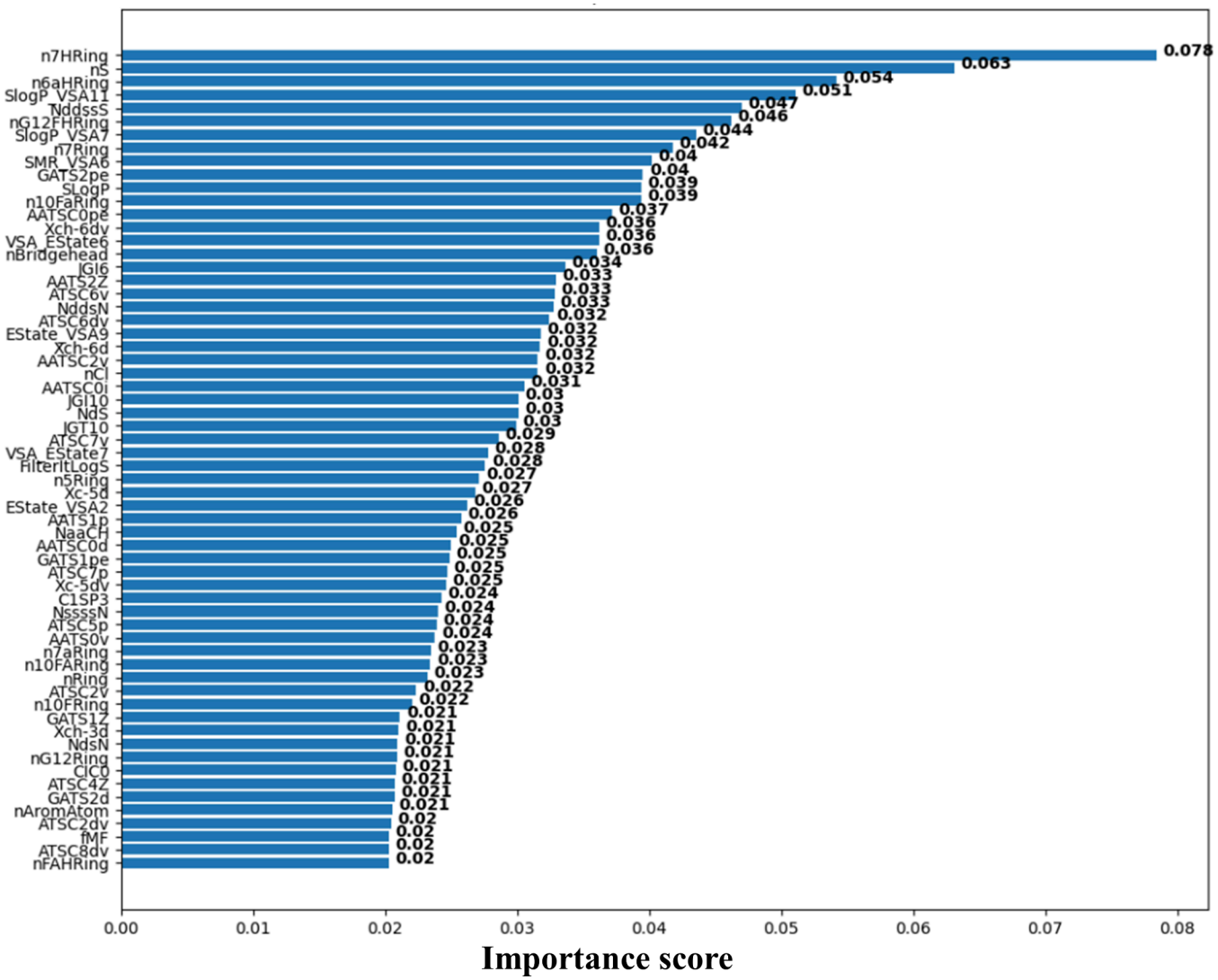


Figure 4. 68 descriptors with their importance score.

set exhibit moderate polarity. On average, molecules have about 3 hydrogen bond donors. Molecules typically have around 7 rotatable bonds, indicating moderate flexibility. In a nutshell, the wide ranges in molecular properties, particularly *MW*, *LogP*, and *TPSA*, suggest diverse chemical structures in the data set, ranging from small organic compounds to large, some macromolecular, as well as highly flexible molecules.

Further to demonstrate the structural diversity of the data set, a frequency distribution of similarity values has been plotted. Tanimoto coefficients (T_c)[14] were calculated by using the Morgan fingerprint[15] to understand the molecular similarity of these 1159 compounds. In our case, this results in 671060 unique pairs (or observations), where each observation represents a T_c calculation between a pair of molecules. In the 671060 observations (pairwise Tanimoto coefficient calculations) for 1159 molecules, only a few observations share T_c values of more than 0.81−1 (Figure S2). From this detailed observation, we can suggest that most of the compounds are dissimilar and unique. Thus, now we will analyze the critical fingerprints of the investigated molecules to understand the DILI toxic and nontoxic SAs.

3.2. Fingerprint Analysis. Out of 1159 compounds, 927 molecules were used as the training set to train the Laplacian-corrected Bayesian model. Figure S2 shows how these compounds are spread out in the PCA space. The distribution of the test set compounds in the PCA space validates the proper division of the data sets.

3.2.1. Construction of the Laplacian-Corrected Bayesian Model. The Laplacian-corrected Bayesian classifier,[22,23] a robust ML technique, was employed to develop classification models for distinguishing 717 DILI *toxic* (1) and 442 DILI *nontoxic* (0). Internal 5-fold cross-validation was performed to assess the stability of the developed model.[25] The statistical outcomes of the 5-fold cross-validation for the training and the test set are presented in Table 1.

For the training set, the false discovery rate (FDR) and false omission rate (FOR) were 0.053 and 0.230, respectively, resulting in a sensitivity of 82.9%. Among the test set compounds, 114 of 232 compounds were correctly classified as true positive, resulting in a sensitivity of 78.6%. For the test set, the FDR and FOR values were 0.240 and 0.378, respectively. Overall results suggested that the classification model could achieve satisfactory discrimination capacity.

3.2.2. Analysis of SAs Produced by the ECFP-6 Fingerprint Descriptor. The fingerprints/features, produced by *ECFP-6* fingerprints,[24] are important for drug-induced liver toxicity

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

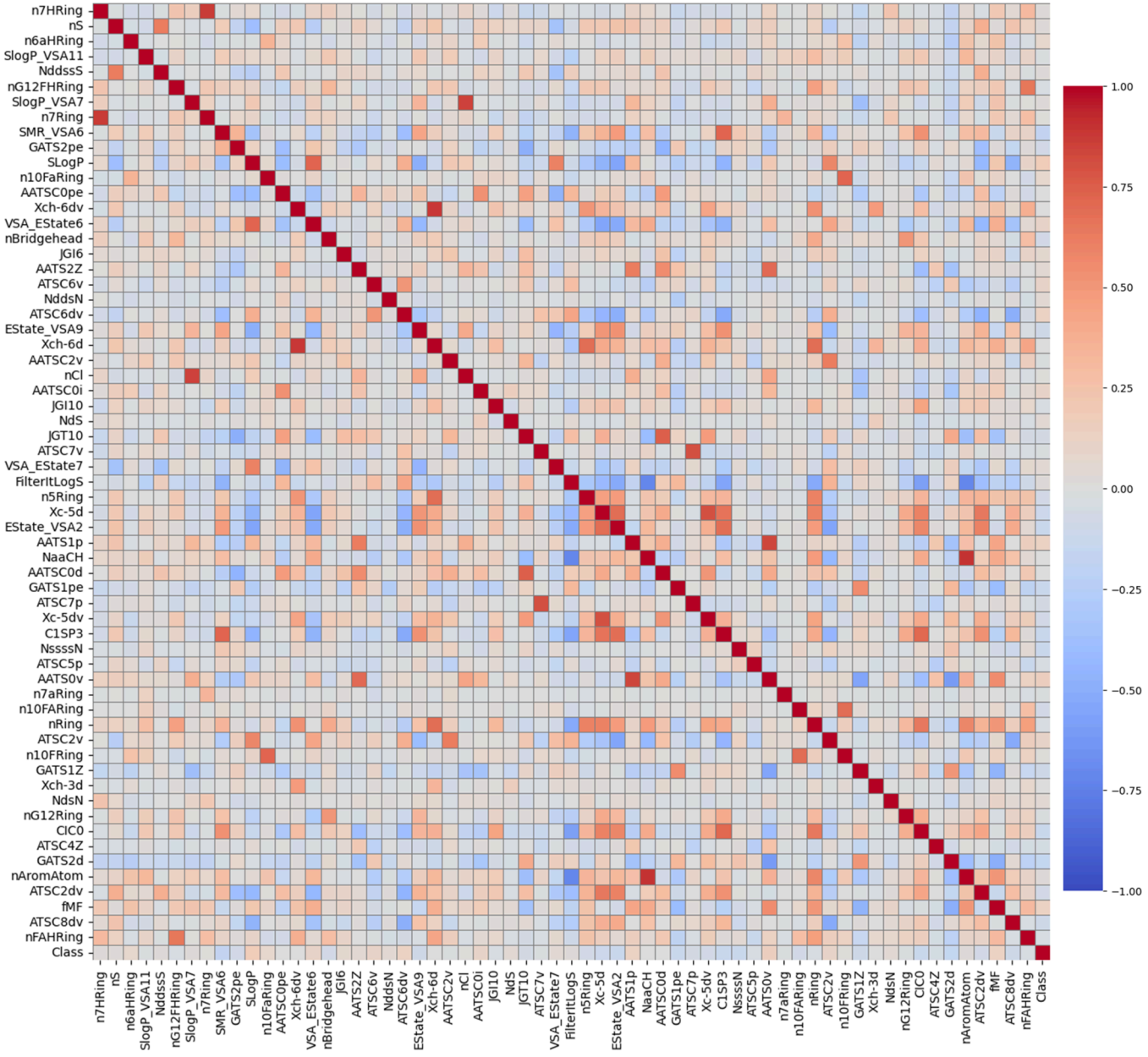


Figure 5. Heatmap of the correlation matrix of the selected descriptors.

prediction as suggested by the Bayesian analysis. The substructural features that increase the chance of drug-induced liver toxicity or are directly associated with DILI are considered as DILI *toxic* (T), whereas the substructural features not associated with DILI may be treated as DILI *nontoxic* (N). The DILI toxic and nontoxic SAs ranked by the Bayesian scores are shown in Figure 3. Upon carefully comparing the fragments, it was evident that no common substructure was shared between the toxic and nontoxic features. Further examination of the structural characteristics of the toxic SAs (shown in Figure 3) revealed that the majority of the toxic fingerprints contain aromatic acids, furans, and substituted triethylamines. These fingerprints are present in most of the hepatotoxic compounds.

The analysis further reveals the presence of a furan ring, highlighted in fingerprint T1, as seen in compound D0043 attributing to DILI *toxic*. Moreover, the aromatic acid functionality (represented by fingerprint T2) in compounds D0422 is identified as risky for DILI. Additionally, the substructural features T7 elucidate the impact of the substituted triethylamine functionality to promote DILI risk. For instance, D1164 renders *toxic* due to the presence of substituted triethylamines. On the other hand, the fingerprint N19 suggests that the $-CH_2CH_2SCH_3$ functionality impedes the DILI risk of compounds, as observed in D0626. Likewise, analogs containing $-NHCH_2CH_2CH_2$ (represented by fingerprint N12) demonstrate nonrisky DILI properties. It can be postulated that a new compound containing one or more of these toxic SAs is likely to pose a high risk of inducing liver toxicity in humans. Consequently, the toxic substructures identified in this study may be recognized as critical SAs for liver toxicity. These SAs should be considered during structural modification and optimization to mitigate the risk of hepatotoxicity.

3.3. ML Studies. Mordred descriptors were calculated in the Python environment.[26] A pool of 1614 descriptors were

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

considered for Feature selection prior to the ML model development.

3.3.1. Data Pretreatment and Feature Selection. Feature selection in ML is an important technique used to preprocess the steps to enhance model performance.[27] First, columns containing any non-numeric values were removed. Non-numeric data, such as categorical or text data, can add complexity, especially if not encoded properly. Followed by constant columns (identical across samples) are deleted to reduce the unnecessary noise and dimensionality since such features do not contribute to distinguishing between classes or outcomes. Since high correlations between features can cause instability in model training and can make the model overly complex without adding much value. By removing one feature from each highly correlated pair, redundancy can be reduced while preserving essential information. Next, Information gain (or mutual information)[33] measures are used to identify the most relevant features for predicting the target (*i.e.*, toxicity value). Only features with higher mutual information scores are retained. Finally, features with importance scores greater than 0.02 were selected to improve accuracy and reduce computational cost. From a set of 1614 descriptors, 68 descriptors that have an importance score greater than 0.02 are selected (Figure 4). In summary, these steps prepare a data set by removing uninformative, redundant, or irrelevant features.

Finally, 68 descriptors were selected for ML studies. These descriptors provide diverse information to capture various aspects of investigated molecules and their properties to correlate the association of DILI. These descriptors (Figure 5) represent a range of physicochemical, structural (topological, constitutional), and electronic properties to characterize the investigated compounds.[34] The descriptors *FilterItLogS*, *SLogP*, and *SMR_VSA6* are related to physicochemical properties. *FilterItLogS* and *SLogP* denote solubility and partition coefficients of molecules, whereas *SMR_VSA6* is a surface area descriptor weighted by molar refractivity.

Topological descriptors (*JGI6*, *JGI10*, *JGT10*, *CIC0*, *Xch-6dv*, *Xch-6d*, *Xc-5d*, *Xc-5dv*, and *Xch-3d*) quantify the topological features (connectivity and overall shape of the molecular graph) of a molecule based on its 2D structure. In particular, JGI descriptors (e.g., *JGI6*, *JGI10*) measure the degree of branching in a molecule, whereas the Xch descriptors describe the connectivity or eccentricity indices with specific orders (e.g., 6d, 5dv). CIC0 denotes the connectivity index of zero-order. Several constitutional descriptors (*n7Ring*, *n7HRing*, *n6aHRing*, *n10FARing*, *n10FRing*, *n5Ring*, *nAromAtom*, *nBridgehead*, *nS*, *nCl*, *NdsN*, *NddssS*, *NddddN*, *NssssN*, and *NdS*) have also been identified by the Information gain approach. These descriptors describe the counts of molecular substructures or specific atoms. For instance, *n7Ring* (number of 7-membered rings), *n10Fring* (number of fluorinated rings), *nAromAtom* (number of aromatic atoms), *nBridgehead* (number of bridgehead atoms), *nS* (number of sulfur atoms), *nCl* (number of chlorine atoms), and specific counts of nitrogen atoms with bonding environments (*NdsN*, *NddddN*), nFARing (number of fluorinated aromatic rings), and fMF (functional group-based molecular fingerprint).

Electrotopological descriptors found important for this data set are *EState_VSA2*, *EState_VSA9*, *VSA_EState6*, and *VSA_EState7*. They provide combined electronic and topological information. EState descriptors encode electronic states of atoms considering their connectivity and environment, while VSA_EState descriptors combine the van der Waals surface area (VSA) with EState indices. In addition, *AATSC0pe*, *AATSC2v*, *ATSC6v*, *ATSC7p*, and *ATSC8dv* are autocorrelation descriptors. They convey information in terms of the geometric or spatial properties of the molecules. They also encode atomic contributions over a defined topological distance. Similarly, the geometrical autocorrelation descriptors (*GATS2pe*, *GATS1pe*, and *GATS1Z*) suggest the involvement of weighted atomic properties (e.g., pe for polarizability, Z for atomic number) of a molecule.

3.3.2. ML Model Development. The primary aim of this study is to devise a proficient algorithm that can discern and classify input data into designated output categories with a remarkable level of *Accuracy*.[25] Seven ML algorithms, including Logistic Regression (LR), k-Nearest Neighbors (k-NN), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Quadratic Discriminant Analysis (QDA), Multilayer perception (MLP) Classifier, were implemented to find the best one. First, the hyperparameters were optimized by the *SearchCV* method with StratifiedKFold in the training set. This ensures that parameter optimization is performed solely on the training data and thereby avoiding data leakage. After selecting the best model configuration from the cross-validation, the optimized classifier was then evaluated on the test set. This two-step step ensures that the performance metrics (Table 2) reflect both the tuning phase of the model and its generalization capability on unseen data. From Table 2, it can be seen that the RF model demonstrated reliable discriminatory power and consistent and robust performance. Figure S3 shows the confusion matrix and the ROC plot of the developed LR, k-NN, NB, RF, DT, QDA, and MLP classifier algorithms. Based on these findings, RF exhibited the best performance over other algorithms used in this study in terms of predicting the association of DILI of molecules. Therefore, the RF algorithm[35] is considered to further *Optuna* hyperparameter optimization[36] to meticulously select the hyperparameters for the final model. Hyperparameter optimization is critical for determining the output and overall effectiveness of an ML model. Optuna optimization provides details of the RF model, search space, and optimal combination achieved.

Table 2. Result of the Different ML Models

model	best parameters	accuracy (%)	precision (%)
LR	'solver': 'liblinear', 'penalty': 'l2'	66.38	68.31
k-NN	'weights': 'distance', 'n_neighbors': 13, 'leaf_size': 1, 'algorithm': 'kd_tree'	59.48	63.93
NB	'priors': None	65.09	66.33
RF	'n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 20, 'criterion': 'gini', 'bootstrap': False	**71.98**	**72.22**
DT	'splitter': 'random', 'min_samples_split': 10, 'min_samples_leaf': 10, 'max_features': 'log2', 'max_depth': 3, 'criterion': 'entropy'	62.93	62.77
QDA	'reg_param': 0.1, 'priors': None	65.95	66.67
MLP classifier	'solver': 'sgd', 'max_iter': 500, 'learning_rate': 'adaptive', 'hidden_layer_sizes': (50,), 'alpha': 0.001, 'activation': 'logistic'	58.19	62.77

3.3.3. Results of Random Forest (RF) Model. The result of the *Optuna* optimized RF model[36] is encouraging. The best accuracy score for the test set is found to be 0.7716. Notably, accuracy is calculated as the proportion of correctly predicted instances out of the total instances and is a measure of the model's overall performance. This is the highest accuracy score

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

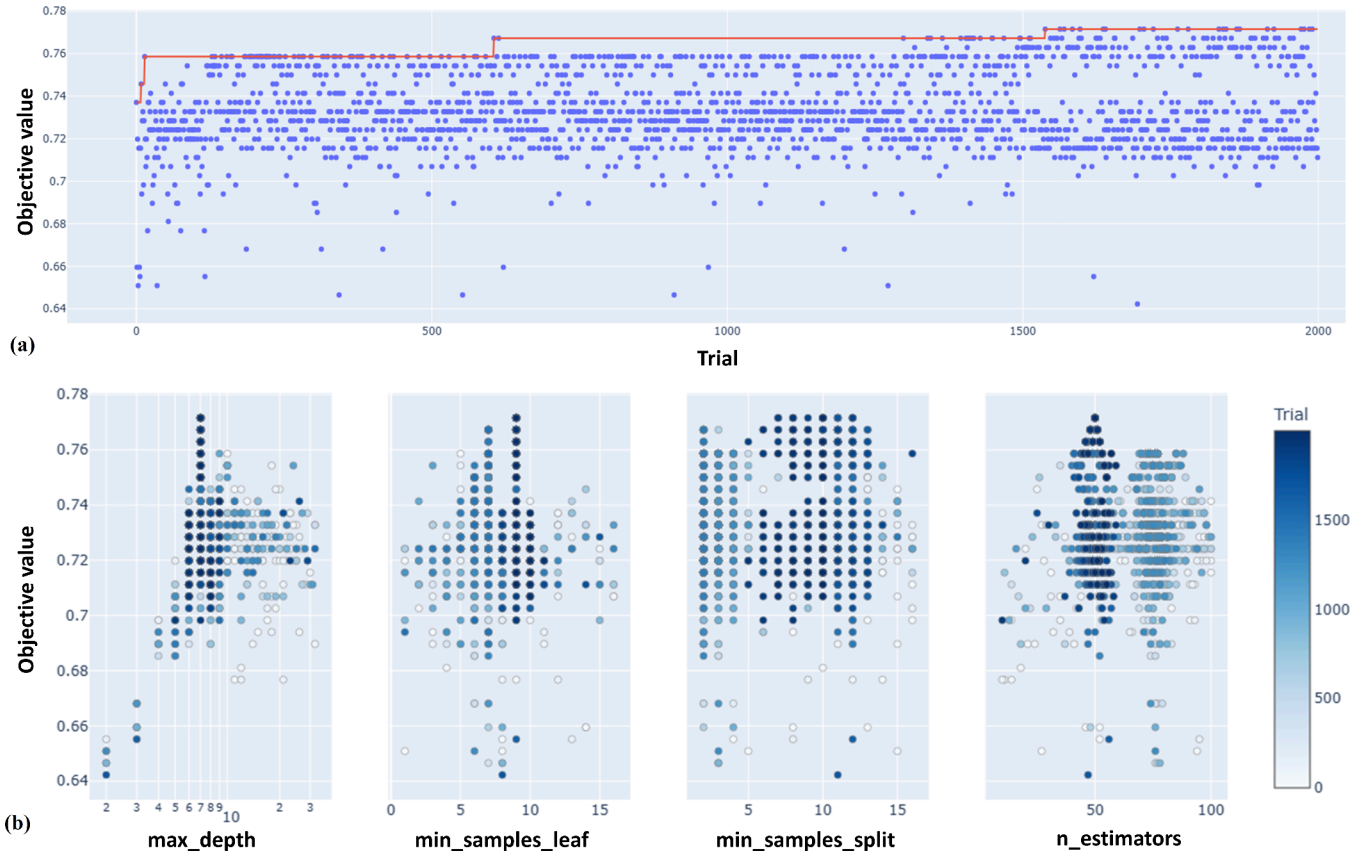


Figure 6. (a) Optimization history plot of the hyperparameter optimization process, (b) Slice plot of specific hyperparameters (*max_depth*, *min_samples_leaf*, *min_samples_split*, and *n_estimators*) with respect to the objective value.

achieved during the hyperparameter tuning process (2000 runs), indicating that the model correctly classified approximately 77.16% of the test or validation data. The other metrics are Precision: 75.27%, Recall: 94.48%, and F1 Score: 83.79%. The optimal hyperparameters found during tuning are *n_estimators*: 50, *max_depth*: 07, *min_samples_split*: 10, *min_samples_leaf*: 09. A value of *n_estimators* (50) indicates that the model performs best with 50 individual trees (Figure 6a). More trees can improve performance but also increase computation time; therefore, this number strikes a balance between *accuracy* and *efficiency*.

The next parameter, *max_depth* (07), controls the maximum depth of each tree. With *max_depth* set to 07, each tree in the ensemble is allowed to grow up to 07 levels deep, capturing more complex patterns in the data (Figure 6b). In addition, a value of *min_samples_split* (minimum number of samples required to split a node) ensures that a node must have at least 10 samples to be split, helping to prevent very deep branches that could capture noise rather than meaningful patterns, reducing the risk of overfitting. Furthermore, with the *min_samples_leaf* (minimum number of samples that a leaf node must have) of 09, each leaf node (end of a branch) must contain at least nine samples, which prevent nodes from representing a single sample, further reducing the potential for overfitting. Taken together, these hyperparameters create a model that can effectively capture the patterns in the data with reasonable complexity, achieving a balance between accuracy and generalizability.

3.3.4. Partial Dependence Plot (PDP) and Features Impact Interpretations. A PDP illustrates the marginal effect of a feature or a combination of features on the predictions made by the RF model (Figure 7). PDPs effectively depict the pattern of variations in a specific feature or set of features that induce an impact on the predictions of the model while averaging out the influence of all other features.

SLogP (Logarithm of the Octanol–Water Partition Coefficient) represents the logarithm of the partition coefficient between octanol and water of the compound, calculated using an atomic contribution approach (Figure 7a). *GATS 2d* (Geary Autocorrelation Descriptor, Lag 2, Distance-Based) is a 2D molecular descriptor based on the Geary autocorrelation function (Figure 7b). Another descriptor, *AATS0v* (Average Atom-Type Electrotopological Descriptor, 0 Lag, Valence-Based), is a valence-based average descriptor calculated using atom-type electrotopological indices at a lag of 0 (immediate atomic environment) (Figure 7c). Figure 8 reveals the influence of chemical descriptors *SLogP vs GATS 2d* to a binary classification outcome, such as determining whether a molecule is likely to be *toxic* (DILI-risk) or *nontoxic*.

Methotrexate is used in cancer and autoimmune diseases and can lead to chronic liver toxicity. Ketoconazole is an antifungal medication that has been linked to liver injury. Both of these drugs fall in a significant region around *SLogP* of more than −1.5 and *GATS 2d* of more than 1.25 (Figure 7d). For optimal GATS 2d values between 1 and 1.3 and *SLogP* values higher than 3, both descriptors have a strong impact on DILI risk. This can be explained by a cholesterol-lowering agent, Cerivastatin (as seen in Figure 8), which can cause liver injury. Similarly, another liver toxic statin (e.g., Simvastatin) that is used to lower cholesterol falls in the region of *SLogP* of more

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

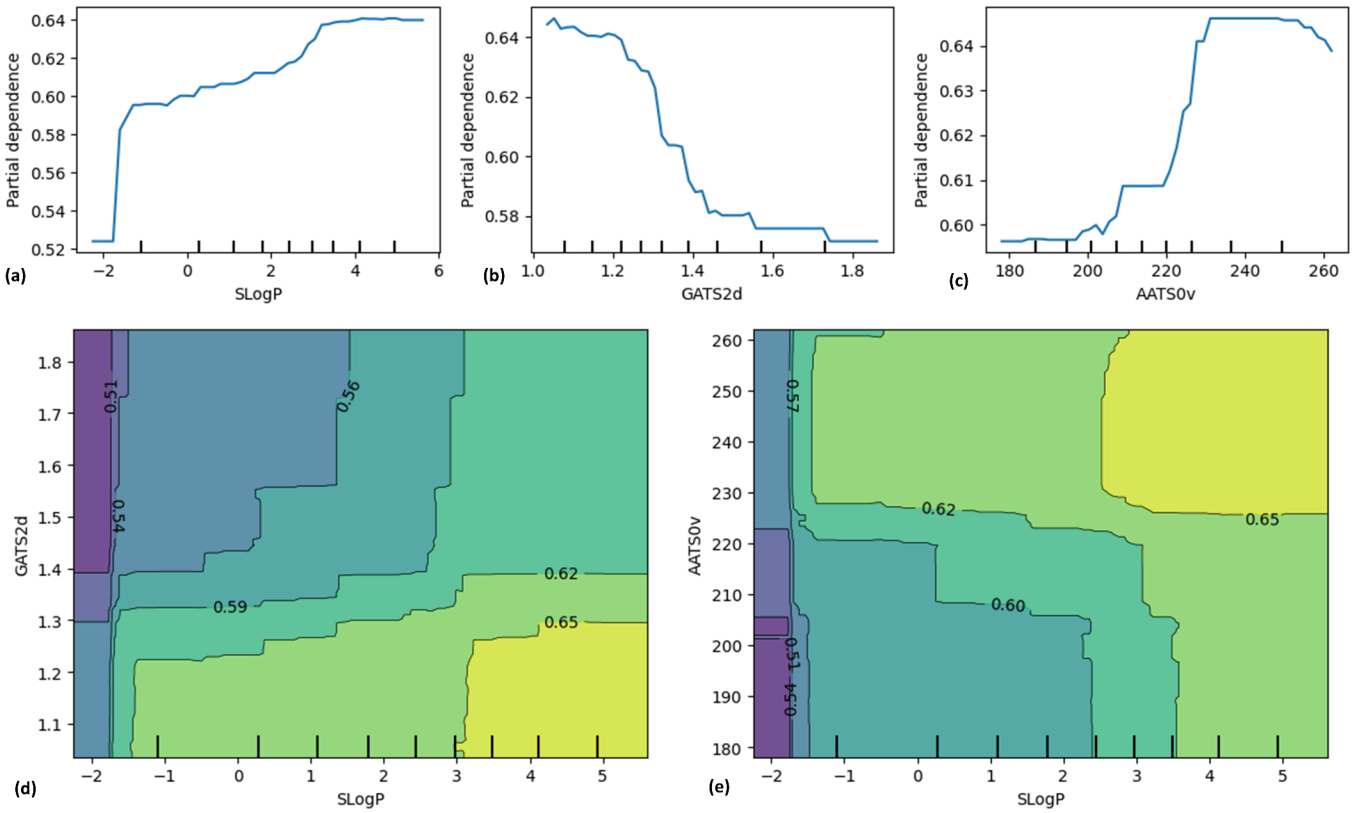


Figure 7. Partial dependence plot (PDP) of descriptors (a) *SLogP*, (b) *GATS 2d*, and (c) *AATS0v*. Two-variable PDP of chemical descriptors (d) *SLogP vs GATS 2d* and (e) *SLogP vs AATS0v*. The contour plot uses color gradients to represent regions corresponding to specific numerical values, as indicated by the contour levels. These values likely reflect a performance metric with the highest values appearing in the yellow-green regions and lower values in the darker blue and purple areas. The gradient transitions from darker to lighter shades, where lighter regions correspond to higher values.

than 4.5 and *GATS 2d* of more than 1.4. Hence, a high number of *GATS 2d* induces distance-based properties of atoms within a hydrophobic molecule (e.g., electronegativity or polarizability) to describe structural relationships. Acetaminophen (high doses can lead to acute liver failure, especially in cases of overdose) and Methyldopa (an antihypertensive agent that can lead to immune-mediated liver injury) also fall in this window. It can be seen that an interaction between the two features with an *SLogP* value higher than 3.5, mainly the *AATS0v* has an impact on the DILI risk (Figure 9).

AATS0v typically summarizes atomic-level information without considering bond distances. For instance, isoniazid, a tuberculosis treatment known to be associated with hepatotoxicity, falls within this window. Our analysis indicates that compounds with an *AATS0v* value above 230 and an *SLogP* value higher than 3.5 tend to have a higher risk of DILI. In this context, drugs such as the antiarrhythmic agent amiodarone (see Figure 7e) and the nonsteroidal anti-inflammatory drug (NSAID) diclofenac, both linked to liver injury, are also found in this region. These observations suggest that atomic contributions, particularly the valence states and connectivity of atoms, may play a crucial role in liver toxicity. It is important to note, however, that liver toxicity is influenced not only by molecular structural properties but also by factors such as dosage, duration of use, individual patient characteristics (e.g., genetics, pre-existing liver conditions), and drug interactions.

3.3.5. Applicability Domain (AD) Analysis. The applicability domain (AD) of molecules is important for assessing the uncertainty in predicting a specific molecule, as it depends on the molecule's similarity to the compounds used to construct the model.[37,38] According to Principle 3 of the Organization for Economic Co-operation and Development (OECD) guidelines,[39] it is essential to define the AD when applying validated models to predict new data points. The predictability of an ML model is considered reliable only if the compound being analyzed falls within the AD. The leverage approach is considered to define the X-outliers (training set) and identify the molecules that reside outside the AD (in the case of the test set).[40] The results suggest that the test set compound numbers D0987 (Leverage Value: 0.3028), D0407 (Leverage: 0.2011), D0456 (Leverage: 0.5476), Compound D0166 (Leverage: 0.3309), D0796 (Leverage: 0.4878), D0759 (Leverage: 0.2894), D0899 (Leverage: 0.2037), D0977 (Leverage: 0.2880), and D0581 (Leverage: 0.2210) are the outliers because the leverage values of these compounds are just under the threshold value (0.2006) (Figure 10).

4. PDILI_V1: A MULTIPLATFORM TOOL FOR PREDICTING DILI RISK

The **pDILI_v1** tool provides a versatile and robust framework for predicting the potential risk of DILI associated with small molecules. Leveraging a rigorously validated ML model, pDILI_v1 classifies compounds into two categories: *RISKy* (1) or *Non-RISKy* (0). This tool integrates chemical space analysis and molecular fingerprints to ensure accurate predictions and user-friendly visualization of the (a) structure of the query compound (Your Molecule as given in the

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

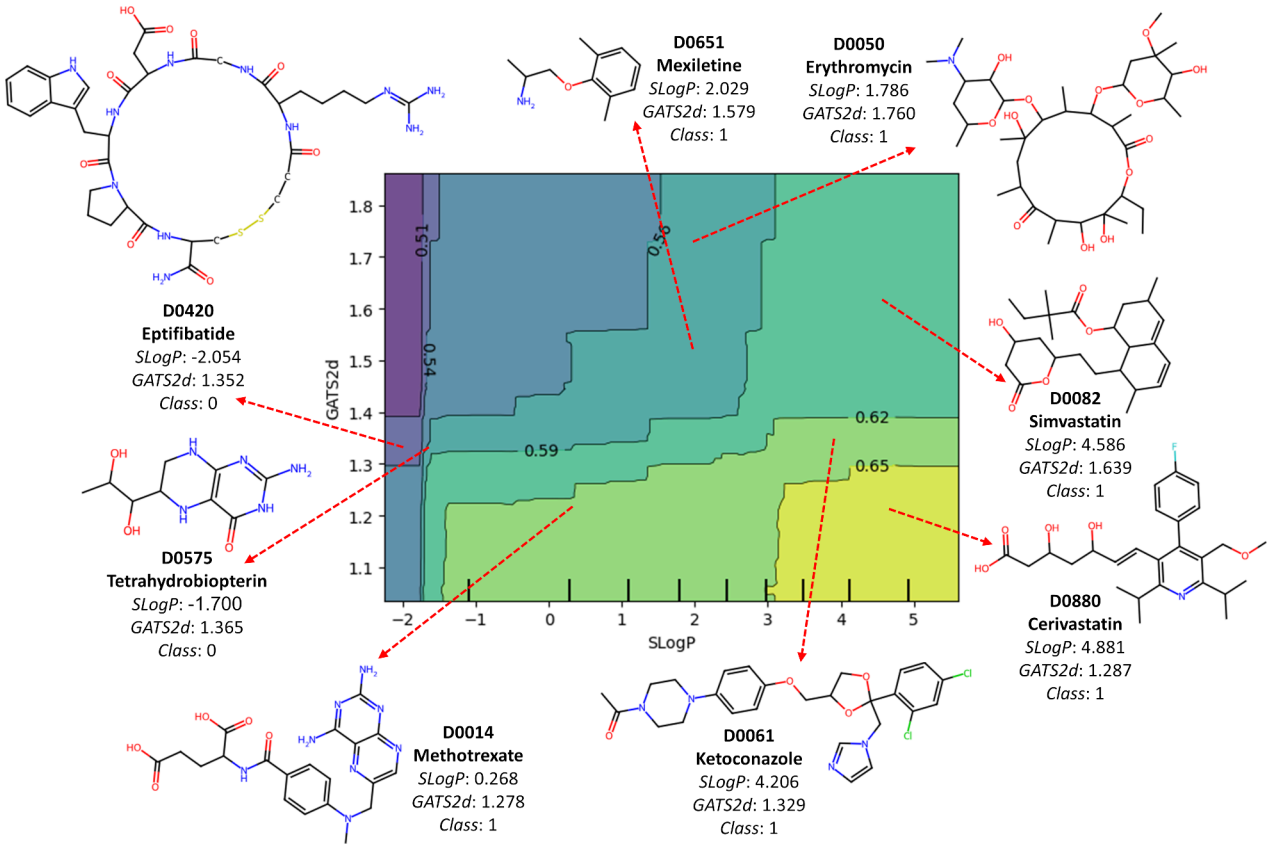


Figure 8. Mechanistic interpretations of the descriptors (*SLogP* and *GATS 2d*) and mathematical contributions to the ML model.

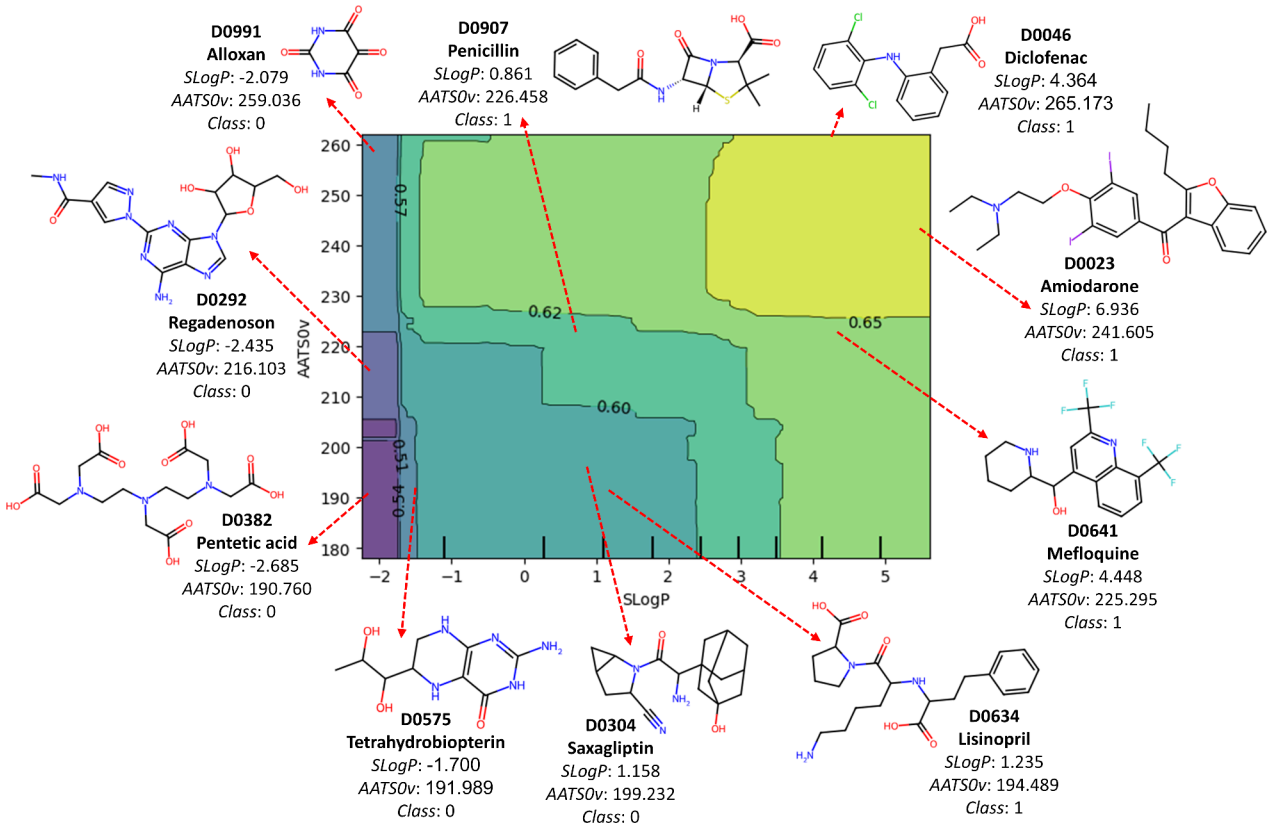


Figure 9. Mechanistic interpretations of the descriptors (*SLogP* and *AATS0v*) and mathematical contributions to the ML model.

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

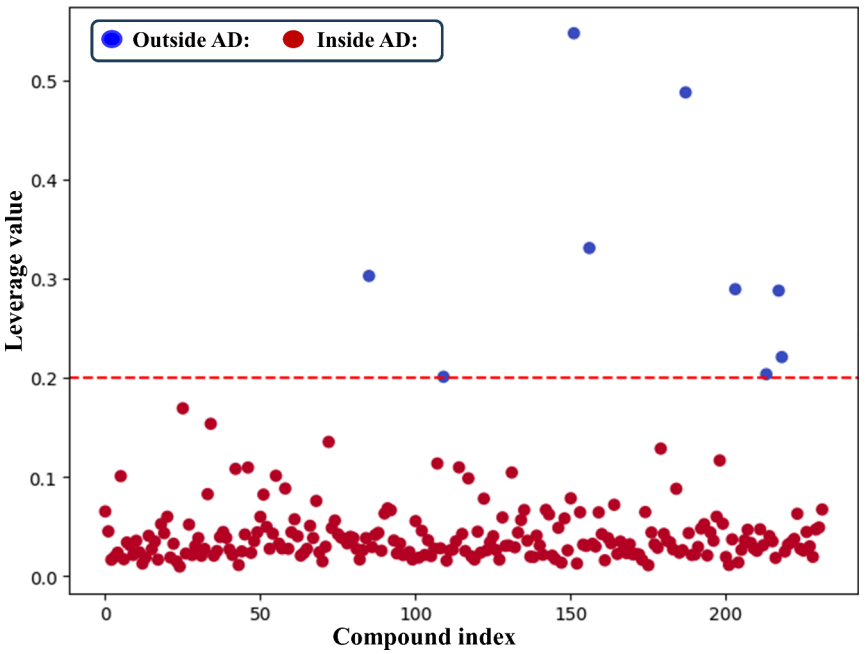


Figure 10. Applicability domain (AD) was based on the leverage approach. The outliers (those outside AD) identified by leverage are highlighted in blue circles.

SMILES) and (b) the position of the query compound in the AD.

Key Features and Accessibility:

1. *Python-Based Web Application:* The primary format is a Python-based web application accessible at https://pdiliv1web.streamlit.app. Users can input a SMILES string of their query molecule to predict its DILI risk and visualize the position of the query compound in the AD.
2. *Google Colab Notebook:* For users preferring an online notebook environment, pDILI_v1 is hosted on Google Colab. This format requires downloading and uploading specific training and test data sets ('1_train_pDILI.csv' and '2_test_pDILI.csv') into a designated directory on Google Drive accessible at GitHub [https://github.com/Amincheminfom/pDILI_v1]. Upon execution, the notebook predicts the DILI risk of the query compound and generates corresponding visualizations.
3. *Graphical User Interface (GUI):* A standalone GUI version of pDILI_v1 is available for Windows systems via the Anaconda environment. Users can install and configure the environment using detailed instructions provided at the GitHub repository [https://github.com/Amincheminfom/pDILI_v1]. This format is particularly suitable for users requiring offline access.

The multiplatform availability of pDILI_v1 makes it a comprehensive tool for DILI prediction, catering to a diverse user base with varying computational needs.

5. CONCLUSIONS

This study provides a comprehensive analysis of the structural determinants of DILI using advanced cheminformatics and ML techniques. Key findings highlight the critical role of specific substructural features such as aromatic acids, substituted sulfur chains, and heterocyclic scaffolds such as furans in contributing to hepatotoxicity. These SAs, identified through *ECFP-6* fingerprint descriptors, serve as predictive markers for DILI risk. Conversely, nontoxic fragments and their physicochemical profiles, such as low aromaticity and reduced lipophilicity, indicate safer structural alternatives.

The study also underscores the importance of molecular descriptors such as *SLogP*, *GATS 2d*, and *AATS0v* in understanding and predicting DILI risk. These descriptors reveal that excessive lipophilicity, high hydrophobicity, and specific electronic and topological configurations are strongly associated with DILI. This insight is pivotal in drug design, as structural optimization strategies can aim to reduce or avoid these high-risk features to minimize the hepatotoxic potential.

The Random Forest model, optimized for high accuracy and precision, has emerged as a reliable predictive tool. Its application, coupled with PDP analysis, illustrates a mechanistic understanding of how individual and combined structural features influence DILI outcomes. Furthermore, to enhance accessibility for the broader scientific community, the release of the open-access "pDILI_v1" [link: https://github.com/Amincheminfom/pDILI_v1] tool aims to facilitate nonexpert use for screening the DILI risk of molecules effectively, supporting safer and more efficient drug development processes.

To advance safer drug design, it is imperative to actively exclude or modify identified toxic substructures during the lead optimization phase. Emphasis on reducing aromatic furan content, optimizing lipophilicity within acceptable ranges, and minimizing hydrophobic fragments can significantly lower the risk of DILI. Additionally, leveraging tools such as pDILI_v1 can help researchers to identify and address hepatotoxic risks at early stages, streamlining the path toward regulatory approval and market success.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c00075.

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514

List of all the drugs together with their class (1: DILI RISKy, 0: DILI Non-RISKy); frequency distribution of the similarity values (Tanimoto coefficient); principal component analysis (PCA) of data set molecules: training vs test sets; and confusion matrix and the ROC plots of the developed LR, k-NN, NB, RF, DT, QDA, and MLP classifier algorithms (PDF)

AUTHOR INFORMATION

Corresponding Authors

Sk Abdul Amin − *Department of Pharmacy, Universita degli Studi di Salerno, Fisciano 84084 Campania, Italy;* orcid.org/0000-0003-4799-7322; Email: pharmacist.amin@gmail.com

Supratik Kar − *Chemometrics and Molecular Modeling Laboratory, Department of Chemistry and Physics, Kean University, Union, New Jersey 07083, United States;* orcid.org/0000-0002-9411-2091; Email: skar@kean.edu

Author

Stefano Piotto − *Department of Pharmacy, Universita degli Studi di Salerno, Fisciano 84084 Campania, Italy;* orcid.org/0000-0002-3102-1918

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c00075

Author Contributions

S.A.A.: conceptualization, data curation, formal analysis, methodology, writing−original draft, review and editing. S.K.: conceptualization, formal analysis, resources, supervision, writing−original draft, review and editing. S.P.: resources, writing−review and editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.A.A. would like to sincerely thank Prof. Tarun Jha (Jadavpur University, Kolkata, India) and Dr. Lucia Sessa (Università degli Studi di Salerno, Italy) for their constant support during the manuscript preparation. S.K. wants to thank the administration of Dorothy and George Hennings College of Science, Mathematics, and Technology (HCSMT) of Kean University for providing research opportunities through research release time and resources.

REFERENCES

(1) Vaja, R.; Rana, M. Drugs and the liver. *Anaesthesia and Intensive Care Medicine* **2020**, *21* (10), 517−523.

(2) 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), 58.

(3) Licata, A. Adverse drug reactions and organ damage: The liver. *European Journal of Internal Medicine* **2016**, *28*, 9−16.

(4) 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), 10.

(5) 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*, 1235.

(6) Vall, A.; Sabnis, Y.; Shi, J.; Class, R.; Hochreiter, S.; Klambauer, G. The promise of AI for DILI prediction. *Frontiers in Artificial Intelligence* **2021**, *4*, No. 638410.

(7) Shin, H. K.; Huang, R.; Chen, M. In silico modeling-based new alternative methods to predict drug and herb-induced liver injury: A review. *Food Chem. Toxicol.* **2023**, *179*, No. 113948.

(8) Seal, S.; Williams, D.; Hosseini-Gerami, L.; Mahale, M.; Carpenter, A. E.; Spjuth, O.; Bender, A. Improved detection of drug-induced liver injury by integrating predicted in vivo and in vitro data. *Chem. Res. Toxicol.* **2024**, *37* (8), 1290−1305.

(9) Lee, S.; Yoo, S. InterDILI: Interpretable prediction of drug-induced liver injury through permutation feature importance and attention mechanism. *J. Cheminform.* **2024**, *16* (1), 1.

(10) Ye, L.; Ngan, D. K.; Xu, T.; Liu, Z.; Zhao, J.; Sakamuru, S.; Huang, R. Prediction of drug-induced liver injury and cardiotoxicity using chemical structure and in vitro assay data. *Toxicol. Appl. Pharmacol.* **2022**, *454*, No. 116250.

(11) Shin, H. K.; Chun, H. S.; Lee, S.; Park, S. M.; Park, D.; Kang, M. G.; Hwang, S.; Oh, J. H.; Han, H. Y.; Kim, W. K.; Yoon, S. ToxSTAR: Drug-induced liver injury prediction tool for the web environment. *Bioinformatics* **2022**, *38* (18), 4426−4427.

(12) U.S. Food and Drug Administration. *Drug-Induced Liver Injury Severity and Toxicity (DILIST) dataset*. Retrieved November 18, 2024, from https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-severity-and-toxicity-dilist-dataset.

(13) Banerjee, S.; Bhattacharya, A.; Dasgupta, I.; Gayen, S.; Amin, S. A. Exploring molecular fragments for fraction unbound in human plasma of chemicals: A fragment-based cheminformatics approach. *SAR and QSAR in Environmental Research* **2024**, *35* (9), 817−836.

(14) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20.

(15) RDKit. *Getting started in Python*. Retrieved November 18, 2024, from https://www.rdkit.org/docs/GettingStartedInPython.html.

(16) Pilgrim, M. *Dive into Python 3*; Springer: Berkeley, CA, 2009.

(17) Google. *Google Colaboratory*. Retrieved November 18, 2024, from https://colab.research.google.com.

(18) Morita, S. Chemometrics and related fields in Python. *Anal. Sci.* **2020**, *36* (1), 107−111.

(19) Likas, A.; Vlassis, N.; Verbeek, J. J. The global k-means clustering algorithm. *Pattern Recognit.* **2003**, *36* (2), 451−461.

(20) De, P.; Kar, S.; Ambure, P.; Roy, K. Prediction reliability of QSAR models: an overview of various validation tools. *Arch. Toxicol.* **2022**, *96*, 1279−1295.

(21) Kalian, A. D.; Benfenati, E.; Osborne, O. J.; Gott, D.; Potter, C.; Dorne, J. C. M.; Guo, M.; Hogstrand, C. Exploring dimensionality reduction techniques for deep learning-driven QSAR models of mutagenicity. *Toxics* **2023**, *11* (7), 572.

(22) Liu, L. L.; Lu, J.; Lu, Y.; Zheng, M. Y.; Luo, X. M.; Zhu, W. L.; Jiang, H. L.; Chen, K. X. Novel Bayesian classification models for predicting compounds blocking hERG potassium channels. *Acta Pharmacologica Sinica* **2014**, *35* (8), 1093−1102.

(23) Bhattacharya, A.; Amin, S. A.; Kumar, P.; Jha, T.; Gayen, S. Exploring structural requirements of HDAC10 inhibitors through comparative machine learning approaches. *Journal of Molecular Graphics and Modelling* **2023**, *123*, No. 108510.

(24) David, R.; Mathew, H. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(25) Roy, K.; Kar, S.; Das, R. N. QSAR/QSPR methods. In *A Primer on QSAR/QSPR Modeling*. SpringerBriefs in Molecular Science; Springer: Cham, 2015.

(26) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **2018**, *10* (4), 4.

(27) Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A. W.; O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics* **2022**, *2*, No. 927312.

(28) Theng, D.; Bhoyar, K. K. Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowledge and Information Systems* **2024**, *66*, 1575−1637.

(29) Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295−316.
(30) Li, W.; Huang, G.; Tang, N.; Lu, P.; Jiang, L.; Lv, J.; Qin, Y.; Lin, Y.; Xu, F.; Lei, D. Effects of heavy metal exposure on hypertension: A machine learning modeling approach. *Chemosphere* **2023**, *337*, No. 139435.
(31) Lai, J. P.; Lin, Y. L.; Lin, H. C.; Shih, C. Y.; Wang, Y. P.; Pai, P. F. Tree-based machine learning models with Optuna in predicting impedance values for circuit analysis. *Micromachines* **2023**, *14* (2), 265.
(32) Optuna. *A hyperparameter optimization framework*. Retrieved from https://optuna.readthedocs.io/en/stable, 2024.
(33) Scikit-learn. *Feature selection using mutual information*. Retrieved from https://scikit-learn.org/1.5/modules/generated/sklearn.feature_selection.mutual_info_regression.html, 2024.
(34) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, 2009.
(35) Pal, M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* **2005**, *26* (1), 217−222.
(36) Kaggle. *Optimization of random forest model using Optuna*. Retrieved from https://www.kaggle.com, 2024.
(37) Yang, S.; Kar, S. Applicability domain for trustable predictions. *Methods Mol. Biol.* **2025**, *2834*, 131−149.
(38) Kar, S.; Roy, K.; Leszczynski, J. Applicability domain: A step toward confident predictions and decidability for QSAR modeling. In *Computational Toxicology*; Humana Press: New York, NY, 2018; pp. 213−237.
(39) OECD. *Guidance document on the validation of QSAR models*. Retrieved from https://www.oecd.org/en/publications/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models, 2024.
(40) NumPy. *The fundamental package for numerical computing*. Retrieved from https://numpy.org/, 2024.

https://doi.org/10.1021/acsomega.5c00075
ACS Omega 2025, 10, 13502−13514